DOI: 10.22032/dbt.39955
DOI (proceedings): 10.22032/dbt.39936

vdt Verband Deutscher Tonmeister e.V.

Full Reviewed Paper at ICSA 2019

Presented * by VDT.

Enhanced Immersion for Binaural Audio Reproduction of Ambisonics in Six-Degrees-of-Freedom: The Effect of Added Distance Information

Axel Plinge, Sebastian J. Schlecht, Olli S. Rummukainen, and Emanuël A. P. Habets
International Audio Laboratories Erlangen†, Germany

Abstract

The immersion of the user is of key interest in the reproduction of acoustic scenes in virtual reality. It is enhanced when movement is possible in six degrees-of-freedom, i.e., three rotational plus three translational degrees. Further enhancement of immersion can be achieved when the user is not only able to move between distant sound sources, but can also move towards and behind close sources. In this paper, we employ a reproduction method for Ambisonics recordings from a single position that uses meta information on the distance of the sound sources in the recorded acoustic scene. A subjective study investigates the benefit of said distance information. Different spatial audio reproduction methods are compared with a multi-stimulus test. Two synthetic scenes are contrasted, one with close sources the user can walk around, and one with far away sources that can not be reached. We found that for close or distant sources, loudness changing with the distance enhances the experience. In case of close sources, the use of correct distance information was found to be important.

1. Introduction

Immersion is one of the key goals of virtual reality (VR) [1]. Along with covering the field of view with live graphics, realistic spatial sound is an important component. It was found that tracked rendering improves localization [2, 3], meaning it is also important for realistic and immersive reproduction [4]. Thus tracked rendering is an active research topic in the VR community: It is important to know how accurate the reproduction has to be and how to get to that accuracy.

It was shown in our recent paper [5] that by using distance information in addition to spatial recording from one location allows six degrees-of-freedom (6DoF) reproduction. The directions of arrival are estimated from the recording, so with known distance the sound sources can be correctly positioned. The question investigated here is in which scenarios the added distance information leads to a better immersion. For spatial audio reproduction of recorded scenes, the fundamental pipeline is as follows: The recording is done from a single or multiple spatially distributed positions with one or more microphone arrays; During reproduction, the listener's relative position is tracked and used to change the sound synthesis; The synthesis itself is done using loudspeakers or, in most cases, headphones. These three basic steps of the pipeline are briefly reviewed in the following sections.

1.1. Recording

Regardless of the recording apparatus, spatial sound is often encoded in the Ambisonics format. This format is a compact and well defined representation of the sound field [6]. It uses the spherical harmonic domain representation [7], which captures the directional information of a sound scene with respect to a specific point in space. The distance of the points of origin of the sounds is not described. It is theoretically possible to derive distance information using the knowledge of the exact microphone positioning. However it might be hardly practical to derive the location of sound sources from the Ambisonics signal alone. The simplest way of interpreting

† Audiolabs is a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS). *

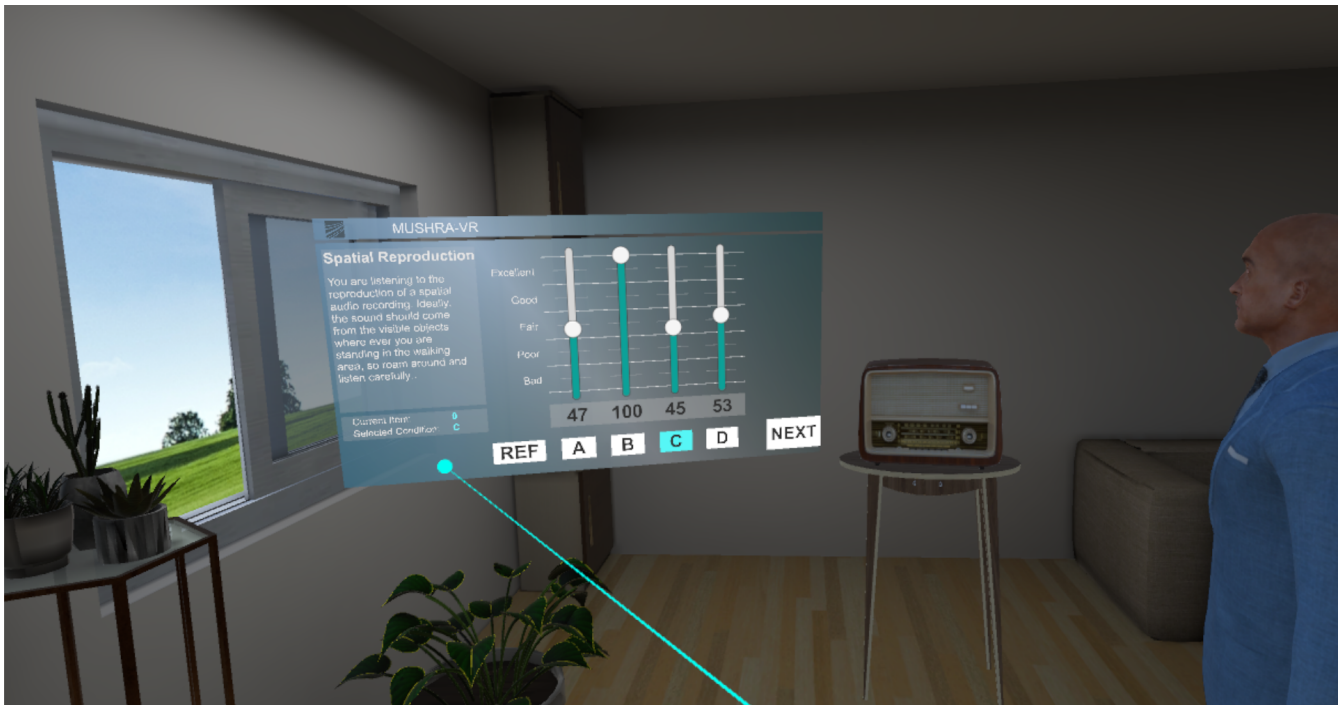

Fig. 1: VR scene 1, living room, with the MUSHRA panel. The sliders for rating the four conditions are in cyan, below them are the numerical value and a button each to select the condition. The active condition and a short descriptive text are shown on the left. The sound objects are the man on the left and the radio, which are inside the walking area of the listener. Bird sounds can be heard from the window.

the sound field for reproduction is to place the whole sound scene in the 'far-field', i.e. outside the reach of the listener.

The size of the recording apparatus has direct implications on the reproduction possibilities. Here we distinguish two cases: First, sound can be recorded either in a single location with compact array; Second, in multiple locations with distributed microphones or microphone arrays [8]. The first is typically the case for recordings done with consumer devices such as cameras and smartphone add-ons. Such a spatial recording from a single location is often used to reproduce a general spatial ambiance. Examples for the second case are dedicated recording sessions in sound or film studios, where multiple synchronized recording devices are employed. Such a recording allows for reproduction of a complex sound scene. It is possible to reproduce sound sources enclosed in the recording area by interpolation between the microphone arrays.

1.2. Spatial Processing

Next, we investigate some common processing methods for the reproduction of the spatial characteristics to a tracked listener. When the listener is moving, both the angle of arrival of the sound and its relative loudness change. Ideally, this is to be reproduced perfectly with perfect knowledge of the exact geometric location of each sound source and the acoustic absorption and reverberation effects of the environment.

When reproducing a recording of a single compact microphone array, all sources are often placed in the far-field. Here, the relative distance of the listener in reproduction to the recorded sources is of limited consequence and the reproduction is often done in three degrees-of-freedom (3DoF) only. This means that only head rotation is applied. This rotation can be computed directly inside the spherical harmonic domain [9].

It is, however, technically possible to extend the freedom of movement even if the recording was done in only a single location. This is facilitated by using parametric sound processing [10]. Examples of such manipulations are the so called 'acoustic zoom' techniques that allow the user to close in on a far-field sources in one direction [11, 12].

The novel method first presented in [5] brings the concept of extending the reproduction possibilities even one step further. It is also based on a single location of recording and parametric encoding of the recorded sound is employed to facilitate the necessary manipulations. By adding distance information for the sound source in each direction, it is possible to virtually place the sound sources in the room, allowing for full 6DoF. It thus is possible to walk around the sources in reproduction, cf. **Fig. 1**. This is a strong effect for immersion enhancement that will be investigated more thoroughly in this paper.

When multiple microphone arrays or Ambisonics microphones are used in a spatially distributed setup, a complex sound scene can be reproduced more easily. In order to enable the listener to walk around sound sources in the scene, the relative positioning of the recording devices has to be mapped to the reproduction scene. Then, the relative distance of the sound sources can be incorporated and the reproduction can be performed in full 6DoF. Different methods for interpolation of their signals have been proposed [13–16]. It is also possible to apply source separation techniques for isolating sound sources in the far-field, especially if the

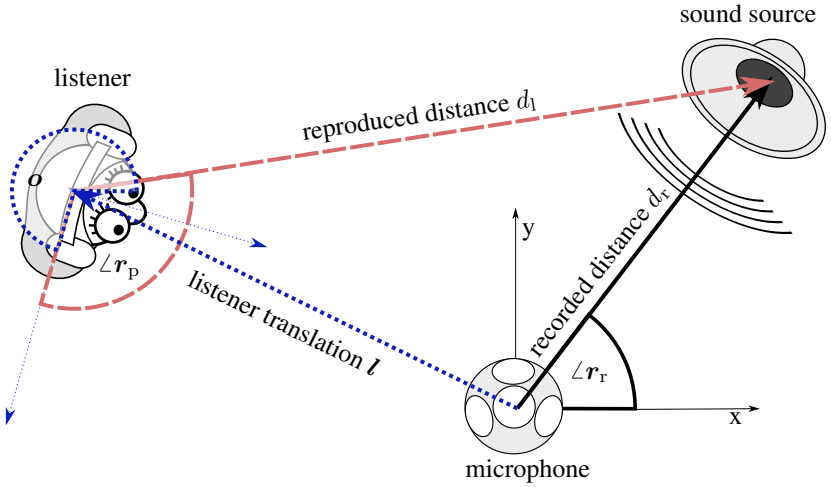

Fig. 2: The 6DoF reproduction of spatial audio. A sound source is recorded by a microphone with the direction of arrival (DOA) $\boldsymbol{r}_\mathrm{r}$ in the distance d_r relative to the microphones position and orientation (black line and arc). It has to be reproduced relative to the moving listener with the DOA $\boldsymbol{r}_\mathrm{l}$ and distance d_l (red dashed). This has to consider the listeners translation $\boldsymbol{l}$ and rotation $\boldsymbol{o}$ (blue dotted).

recording is done with many microphones and then encoded in higher-order Ambisonics (HOA) [17].

1.3. Reproduction

Finally, we describe the actual reproduction. This can be implemented with loudspeakers or headphones.

Sound produced by loudspeakers is usually targeted at the so called 'sweet spot', a small area where the directions and loudness of all speakers are matched. The listener has to stay in this area in order to experience the intended spatial impression, with the head oriented in the intended direction. Some techniques allow for widening this area [18] or adjusting to head rotations [19]. The spatial reproduction ability is often limited by the loudspeakers' arrangement, unless a large number of them is used.

The use of headphones is common in VR applications. The user is often wearing a head-mounted display (HMD), thus adding headphones is quite practical. When using headphones, head-related transfer functions (HRTFs) are applied in order to spatialize the sound [20, 21]. HRTFs allow to reproduce the effect of the shape of torso, head, and outer ears on the sound depending on its direction of origin. Often virtual loudspeakers are placed around the user. These virtual loudspeaker signals are then binauralized [22]. It is possible to binauralize Ambisonics by applying HRTFs directly in the spherical harmonic domain, however the limited Ambisonics order can lead to unwanted filtering effects in practice [23].

1.4. Research Question

The question this paper looks to answer is how important the distance information of sound sources is in VR reproduction from a recording at a single location. This will be done by a subjective listening experiment. Recently multi-stimulus testing has been employed in listening tests for VR. In several studies the multiple stimuli with hidden reference and anchor (MUSHRA) paradigm, common in general non-interactive audio testing, was applied [16,24]. The test subject is immersed in a VR scene, typically rendered as computer generated imaging (CGI). Different audio renderings can be switched from inside the scene and rated on a scoreboard. In our experiment, such a MUSHRA test is performed in a virtual indoor scene as in our previous paper [5]. In this scene the listener can walk around the sources. This is contrasted with a second virtual outdoor scene, where the distance information is of less importance as the sources are placed outside the walking area. This way it is investigated when the use of distance information actually enhances the experience in a meaningful way. By keeping the processing pipeline the same for all conditions, the timbre is as close as possible. The only difference in conditions is the application of the tracking data to the sound processing. Angular and distance effects are separated to allow judging their individual importance.

The rest of this paper is organized as follows: First, the method introduced in [5] is described in Section 2. The individual implementation of the sound rendering and the scenes used are explained in Section 3. Section 4 gives the listening test results followed by a short conclusion in the final Section 5.

2. Method

In this paper, a binaural signal is produced at the listener's position given the signal at a single recording position and information about the distances of sound sources from that recording position. Given a scene of limited size, the physical sources are assumed to be separable by their angle towards the recording position.

The position of the recording microphone is used as the origin of the reference coordinate system. The listener is tracked in 6DoF, cf. **Fig. 2**. At a given time, the listener is at a position $\boldsymbol{l} \in \mathbb{R}^3$ relative to the microphone and has a rotation $\boldsymbol{o} \in \mathbb{R}^3$ relative to the microphones' coordinates system. We deliberately choose the recording position as the origin of our coordinate system to simplify the notation. The sound is reproduced with a different distance d_l, leading to a changed signal level, and a different DOA $\boldsymbol{r}_\mathrm{p}$ that is the result of both translation and subsequent rotation.

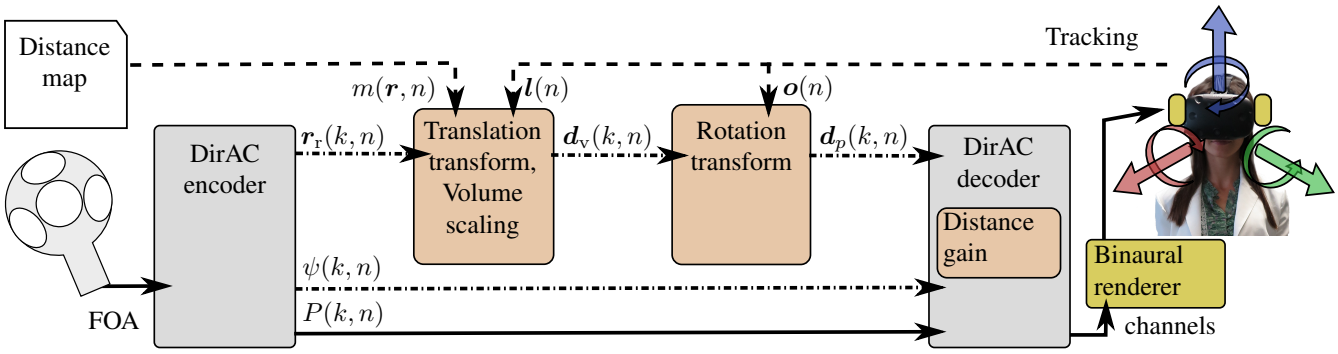

Fig. 3: Proposed method of 6DoF reproduction. The recorded first-order Ambisonics (FOA) signal in B-format is processed by a directional audio coding (DirAC) encoder that computes direction and diffuseness values for each time-frequency bin of the complex spectrum. The direction vector is then transformed by the listener's tracked position and according to the distance information given in a distance map. The resulting direction vector is then rotated according to the head rotation. Finally, signals for 8+4+4 virtual loudspeaker channels are synthesized in the DirAC decoder. These are then binauralized.

The reproduction pipeline introduced in [5] is sketched in **Fig. 3**. An Ambisonics recording and a distance map are used as input signal for rendering. In the parametric DirAC representation [25], geometric transformations are applied. Then a channel signal for virtual loudspeakers is binauralized to headphones.

The Ambisonics input signal, in this case FOA, is decomposed into a parametric DirAC representation [25]. This consists of a complex spectrum $P(k,n)$, where k denotes the frequency bin and n the time frame. For each time frequency bin, a diffuseness ψ and unit length direction vector $\boldsymbol{r}_r$ are estimated. The sound scene is thereby decomposed into a diffuse and direct component, cf. [26]. The directed sound is complemented by distance information for each time frame. It is formulated as distance to the closest potential sound source in a spherical coordinate system. The mapping function $m(\boldsymbol{r}, n)$ returns a distance in meters for each direction vector $\boldsymbol{r}$ and time frame n.

The direction vector then undergoes different transformation steps. First, the distance according to the distance map is added by multiplying the unit direction vectors with the corresponding distance map entry:

$$\boldsymbol{d}_{\mathrm{r}}\left(k,n\right) = \boldsymbol{r}_{\mathrm{r}}\left(k,n\right) m\left(\boldsymbol{r}_{\mathrm{r}}\left(k,n\right),n\right), \tag{1}$$

then the translation by the listener position $\boldsymbol{l}(n) = [l_{\mathrm{X}}(n), l_{\mathrm{Y}}(n), l_{\mathrm{Z}}(n)]^T$ is accounted for by subtracting it from each direction vector:

$$\boldsymbol{d}_{\mathrm{l}}\left(k,n\right) = \boldsymbol{d}_{\mathrm{r}}\left(k,n\right) - \boldsymbol{l}(n). \tag{2}$$

Additionally, the distance vector's length is compensated to map the level change with respect to the closest source given by the distance map:

$$\boldsymbol{d}_{\mathrm{v}}\left(k,n\right) = \frac{\boldsymbol{d}_{\mathrm{l}}\left(k,n\right)}{\|\boldsymbol{d}_{\mathrm{r}}\left(k,n\right)\|}. \tag{3}$$

The resulting distance vector $\boldsymbol{d}_v(k,n)$ is then rotated according to the listeners orientation $\boldsymbol{o}(n)$. It can be written as vector composed of the pitch, yaw, and roll $\boldsymbol{o}(n) = [o_{\mathrm{X}}(n), o_{\mathrm{Z}}(n), o_{\mathrm{Y}}(n)]^T$, which allows implementing the transformation using 2D rotation matrices, cf. eqn. (23) in [9]:

$$\boldsymbol{d}_{\mathrm{p}}\left(k,n\right) = \boldsymbol{R}_Y(o_{\mathrm{Y}}(n))\boldsymbol{R}_Z(o_{\mathrm{Z}}(n))\boldsymbol{R}_X(o_{\mathrm{X}}(n))\boldsymbol{d}_{\mathrm{v}}\left(k,n\right). \tag{4}$$

The parametric representation is then decoded into virtual loudspeaker signals following an edge fading amplitude panning (EFAP) panning scheme [27] with the DirAC method. The angle of the unit vector $\boldsymbol{r}_{\mathrm{p}}$ is used for the panning, and the length of the vector $||\boldsymbol{d}_{\mathrm{p}}||$ is used for a distance dependent gain. The diffuse sound component is reproduced equally to all loudspeakers in order to provide an undirected ambiance. Cf. [5] and references therein for more mathematical detail.

The channel signals are binauralized by convolving each virtual loudspeaker signal of the 8+4+4 setup with a HRTF for left and right ear. The distance of all speakers is fixed and no additional loudness change is added.

3. Experiments

The experiments follow a MUSHRA-like paradigm adopted for VR [24]. The different methods are compared and rated on a scale of 0 to 100 points, with 0 being the worst and 100 being perfect. There is a reference condition, which can be selected explicitly in addition. It is also one of the presented choices, this hidden reference is to be rated with 100 points. There is one clearly bad condition, the so called anchor. In order to do so in VR, the MUSHRA panel can be opened any time by the subject. They can then switch and rate the different renderings at will.

3.1. Conditions

The four randomized conditions in our experiments were:

REF Object-based rendering. This is the reference condition. The B-format is generated on the fly for the listener's current position and then rendered via the virtual loudspeakers.

C1 3DoF reproduction. The listener position is ignored, i.e. $\boldsymbol{l}(n) = \boldsymbol{0}$, but the head rotation $\boldsymbol{o}(n)$ is still applied. The

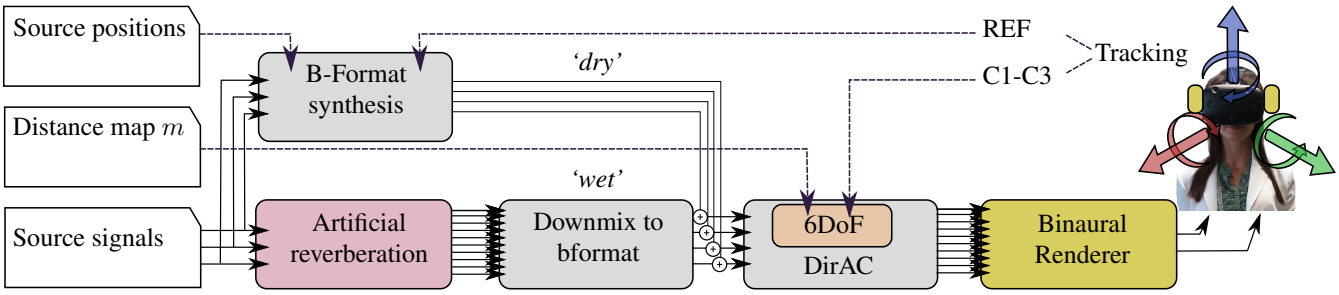

Fig. 4: The signal paths for reference rendering and DirAC. In the reference case, the tracking data is used to change the positioning and rotation of the object-based synthesis (top left). In the other conditions C1-C3, the tracking data is applied in the DirAC domain (right).

gain is set to that of sources in a distance of 2 m from the listener. This condition is used as an anchor.

C2 The proposed method for 6DoF reproduction without distance information. The listener position is used to change the direction vector. All sources are located on a sphere outside of the walking area. The radius of the sphere was fixed to 2 m and the distance-dependent gain is applied.

C3 The proposed method of 6DoF reproduction with distance information. The listener position and orientation is accounted for. The distance information is used to compute the correct DOA at the listener position, and the distance-dependent gain is applied.

So the reference is not a perfect synthetic rendering, but the best possible recording position. This way the limitations of the FOA signal, namely the large width of the sources, is still audible. Nevertheless the spatial adaptation to the listeners movement is perfect. Similarly, the anchor is not a clearly bad low pass filtered version but just the 3DoF version. So it is spatially deficient in the regard that there is no distance attenuation or changing of DOAs depending on the tracked listeners' position but otherwise sounds plausible.

3.2. Technical realization

The experiment was realized using a HTC VIVE for tracking and reproduction, Unity 3D Engine for Graphics, and Max MSP for audio. The platform is described in full detail in [24]. The user wears the VIVE HMD and DT 770 Pro headphones driven by an RME Babyface, both connected to a PC running Unity and Max. A unity script encapsulates the tracking data and sends it as an open sound control (OSC) message to the Max patch. The switching of the active rendering as well as the MUSHRA rating and scene switching were realized by dedicated interaction scripts sending OSC messages to Max such as 'condition 2 selected', 'active condition rated 45', etc. Note that the Max patch took care of the randomization of conditions internally, so that neither the user nor the Unity scripts would know which is which.

In order to investigate the ability of the proposed method to reproduce the sound reproduced as if recorded at a single location in 6DoF, a dedicated rendering pipeline was constructed as shown in **Fig. 4**. A collection of virtual studio technology (VST) plugins was integrated using Max MSP 7. The key principle was to keep the same processing chain for all test conditions, so that the timbre is as similar as possible.

An FOA signal is generated from each source with distance attenuation with a dedicated VST in the Max patch. In case of the reference condition, the virtual microphone was placed at the listeners tracked position. In all other conditions, it was the fixed recording position in the center of the walking areas.

In scene 1, artificial reverberation is added to the source signal in a time-invariant manner by a dedicated VST (Fraunhofer IIS Reverb). Early reflections from the boundaries of the shoebox-shaped room are added with accurate delay, direction, and attenuation. Late reverberation is generated with a spatial feedback delay network (FDN) which distributes the multichannel output to the virtual loudspeaker setup [28]. The frequency-dependent reverberation time T_{60} was between 90 to 150 ms with a mean of 110 ms. A tonal correction filter with a lowpass characteristic was applied subsequently.

This is rendered to a 8+4+4 virtual speaker setup. Eight speakers are uniformly distributed along the medial plane at 0, 45, 90 degrees etc. An additional four are placed both at the top and bottom in a cross formation at $\pm 45°$ elevation. The reverberated signal is then converted to B-format by multiplying each of the virtual speaker signals with the B-format pattern of their DOA in an Ambisonics encoder VST. The reverberant B-format signal is added to the direct signal.

Subsequently, the mixed signal is processed in the parametric DirAC domain with a dedicated VST we developed. In case of the reference condition, no changes are made based on the tracking data and it is set to the recording position ($\boldsymbol{l} = [0,0,0]^T$ and $\boldsymbol{o} = [0,0,0]^T$). The processing is only done to keep the timbre and delays identical. An added benefit is that the switch between conditions can be realized seamlessly. In the other conditions, the 6DoF processing based on the tracking is applied to varying degree. In C3, only rotation is applied. In C2 and C3 translation according to the listener position is applied as well. Only in C3 the distance information is used. In C2, all sources are assumed slightly outside the walking area. The signal is then converted into a channel signal for a 8+4+4 loudspeaker configuration using EFAP panning [27].

These signals are the convolved with generic HRTFs with another VST. Each of the 16 channels is convolved with a far-field HRTF corresponding to the spherical coordinates for both left and right ear. Then the signal is output from Max to the headphones.

Fig. 5: The indoor scene. The sound is coming from the person, the radio and the open window, each source marked with concentric circles. The microphone position is marked by a cross. The user can walk in the area marked by the dashed rectangle on the floor.

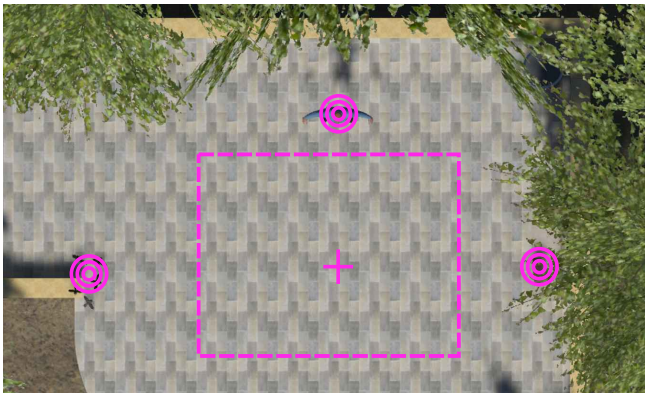

Fig. 6: The outdoor scene. The sound is coming from the person, the radio and the birds, each source marked with concentric circles. The microphone position is marked by a cross. The user can walk in the area marked by the dashed rectangle on the floor.

3.3. Scenes

For the experiments, a room with an walking area of about 3.5 × 4.5 m was used. The listening tests were conducted separately in two scenes. Both scenes were constructed realistically with visual representations of the sound objects. This was done to provide a baseline visual immersion for VR content.

The first scene was an indoor scene in a virtual living room. **Fig. 5** shows a top view. The cross shows the recording position. In about 0,5 m a virtual human speaker is placed, in 1 m distance a radio playing a string loop. In 2 m distance, just at the edge of the walking area, birds singing could be heard from a window.

The second scene was placed outdoors. **Fig. 6** show a top view. The same sound sources were played back, but this time outside the walking area. The human was placed in 2 m, the radio in 3 m and the birds, this time in view as birds on the floor, in 4 m distance.

4. Results

The listening tests were conducted on different days with a total of 25 subjects, some only participating on the first

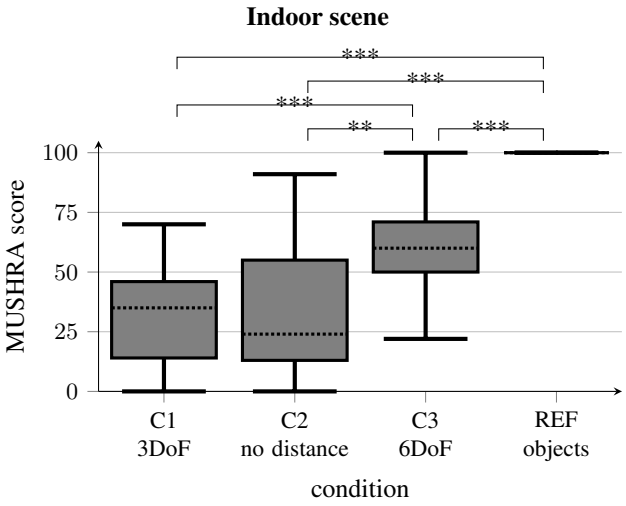

Fig. 7: MUSHRA ratings for the indoor scene (N=20).

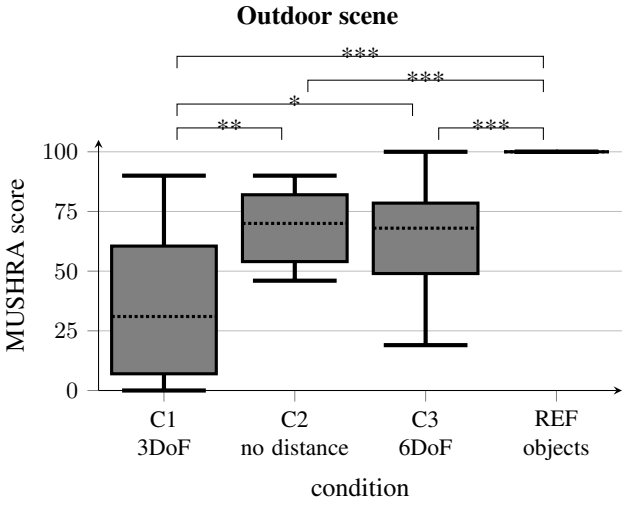

Fig. 8: MUSHRA ratings for the outdoor scene (N=18).

day. They were 24-40 years old, male and female. Their VR experience had a wide range from almost none to quite extensive. After excluding subjects that scored the reference with less then 90 points, there are 20 remaining out of 24 for scene one and 18 out of 23 for scene two. The scores for both scenes are shown in **Fig. 7** and **Fig. 8** as box plots. The dotted line represents the median score, the boxes the 1st to 3rd quartile, the whiskers are at ±1.5 inter-quartile range (IQR). Stars indicate significant differences according to a pairwise permutation test using one million permutations [29], * = $p < 0.1$, ** = $p < 0.01$, *** = $p < 0.001$.

4.1. Indoor Scene

As indicated in **Fig. 7**, all 20 subjects scored the reference with 100 points in the indoor scene. There was no significant difference between C1 and C2 ($p = 0.74$) that is both the 6DoF without distance and the 3DoF were rated worst with around 30 points. In both cases the sound came from the wrong side when walking behind sources, which may have dominated all other effects. The 6DoF scheme with distance information is rated better with around 60 points. The difference between using distance information (C3) or not

(C2) is significant ($p \leq 0.01$). The reference is clearly valued better than all other conditions ($p \leq 0.001$), as is 6DoF with distance (C3) vs 3DoF (C1) ($p \leq 0.001$).

4.2. Outdoor Scene

As can be seen in **Fig. 8**, all 18 subjects scored the reference with 100 points in the outdoor scene. In contrast to the indoor scene, there was no significant difference between the 6DoF rendering with (C3) and without (C2) distance information ($p = 0.70$). The 3DoF rendering (C1) was ranked lower but not as clearly. However it is rated significantly below both C3 ($p \leq 0.02$) and C2 ($p \leq 0.01$). Again, the reference is clearly valued higher than all other conditions ($p \leq 0.001$).

4.3. Discussion

The 3DoF anchor is harder to spot than the usual anchors, which can confuse subjects, especially those used to non-VR MUSHRA tests. This can explain the rather big variance of scores. In, e.g., [30] mono mix was used as anchor, which leads to a clearer distinction. This could have been used in addition. Still, the requirement of a reference condition and a lower anchor condition is tricky in VR. Alternative methods are emerging to avoid this problem [31].

We did not exclude results when subjects rated C1 and C2 similar, which happed often in the first scene as the DOA was audibly wrong for both. There is a strong and significant distinction between cases with correct and faulty DOA in the indoor scene. This distinction is shifted towards a lower bar of 3DoF in the second scene. As long as there was a perceptible loudness change and reasonable DOAs, the scene was accepted as good. Even though there was a misplacement by using 2 m instead of the true up to 4 m distance of the sources, there was no significant difference in the subject's rating of C2 and C3.

5. Conclusion

A novel method for reproducing spatial audio recordings in 6DoF was evaluated. The method employs distance information to reproduce sound recorded at a single position at different points in the space the listener can move in. This distance information is important in scenarios with close sources, which the listener can move around. A listening test was conducted in two separate scenarios. The first was an indoor scenario with sources reachable by the listener; the second was an outdoor scenario where the sound sources are unreachable. A MUSHRA test showed a clear significant preference for the use of distance information in the first scenario, but not in the second. In the second scenario the 6DoF reproduction was clearly preferred to the 3DoF with no distance attenuation. This indicates an enhanced realism by correct placement for close sources and by distance attenuation for sound source in close to medium distance.

References

[1] D. A. Bowman and R. P. McMahan, "Virtual reality: How much immersion is enough?," *IEEE Computer*, vol. 40, no. 7, pp. 36–43, 2007.

[2] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization.," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339, 1940.

[3] K. I. McAnally and R. L. Martin, "Sound localization with head movement: implications for 3-D audio displays," *Frontiers in Neuroscience*, vol. 8, pp. 210, 2014.

[4] H. Hacıhabiboglu, E. D. Sena, Z. Cvetković, J. Johnston, and J. O. Smith III, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 36–54, May 2017.

[5] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order Ambisonics with distance information," in *Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality*, Redmond, U.S.A., Aug. 2018.

[6] M. Frank, F. Zotter, and A. Sontacchi, "Producing 3D audio in Ambisonics," in *Audio Eng. Soc. Conf.*, Mar. 2015.

[7] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer, 2017.

[8] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, July 2016.

[9] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of Ambisonic recordings," in *2nd International Conference on Spatial Audio*, Erlangen, Germany, 2014.

[10] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42, 2015.

[11] O. Thiergart, K. Kowalczyk, and E. A. P. Habets, "An acoustical zoom based on informed spatial filtering," in *International Workshop on Acoustic Signal Enhancement*, Sept. 2014, pp. 109–113.

[12] H. Khaddour, J. Schimmel, and F. Rund, "A novel combined system of direction estimation and sound zooming of multiple speakers," *Radioengineering*, vol. 24, no. 2, June 2015.

[13] O. Thiergart, G. D. Galdo, M. Taseska, and E. A. P. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

[14] C. Schörkhuber, P. Hack, M. Zaunschirm, F. Zotter, and A. Sontacchi, "Localization of multiple acoustic sources with a distributed array of unsynchronized first-order Ambisonics microphones," in *Congress of Alps-Adria Acoustics Assosiation*, Graz, Austria, Jan. 2014.

[15] J. G. Tylka and E. Y. Choueiri, "Soundfield navigation using an array of higher-order Ambisonics microphone," in *Audio Eng. Soc. Conf. on Audio for Virtual and Augmented Reality*, Los Angeles, California, U.S.A., 2016.

[16] E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, and T. Zernicki, "Toward six degrees of freedom audio recording and playback using multiple Ambisonics sound fields," in *Audio Engineering Society Convention*, York, U.K., Mar. 2019.

[17] J. Zamojski, P. Makaruk, L. Januszkiewicz, and T. Zernicki, "Recording, mixing and mastering of audio using a single microphone array and audio source separation algorithms," in *Audio Engineering Society Convention*, New York, U.S.A., Oct. 2017.

[18] A. Iljazovic, F. Leschka, B. Neugebauer, and J. Plogsties, "The influence of 2-d and 3-d video playback on the perceived quality of spatial audio rendering for headphones," in *Audio Engineering Society Convention*, Oct. 2012.

[19] S. Merchel and S. Groth, "Adaptively adjusting the stereophonic sweet spot to the listener's position," *Journal of the Audio Engineering Society*, vol. 58, no. 10, pp. 809–817, 2010.

[20] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head related transfer functions of human subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, 1995.

[21] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, The MIT Press, Oct. 1996.

[22] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3D Ambisonic based binaural sound reproduction system," in *Audio Eng. Soc. Conf.*, June 2003.

[23] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[24] T. Robotham, O. Rummukainen, J. Herre, and E. Habets, "Evaluation of binaural renderers in virtual reality environments: Platform and examples," in *Audio Engineering Society Convention*, New York, U.S.A., Oct. 2018.

[25] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[26] O. Thiergart, G. Del Galdo, F. Kuech, and M. Prus, "Three-dimensional sound field analysis with directional audio coding based on signal adaptive parameter estimators," in *Audio Engineering Society Convention on Spatial Audio: Sense the Sound of Space*, Oct. 2010.

[27] C. Borß, "A polygon-based panning method for 3D loudspeaker setups," in *Audio Engineering Society Convention*, Los Angeles, CA, U.S.A., Oct. 2014, pp. 343–352.

[28] S. J. Schlecht and E. A. P. Habets, "Sign-agnostic matrix design for spatial artificial reverberation with feedback delay networks," in *Audio Engineering Society International Conference on Spatial Reproduction*, Tokyo, Japan, 2018.

[29] P. Good, *Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer Series in Statistics. Springer, 2 edition, 2000.

[30] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets, "Online vs. offline multiple stimulus audio quality evaluation for virtual reality," in *Audio Engineering Society Convention*, New York, U.S.A., Oct. 2018.

[31] O. Rummukainen, T. Robotham, S. J. Schlecht, A. Plinge, J. Herre, and E. A. P. Habets, "Audio quality evaluation in virtual reality: Multiple stimulus ranking with behavior tracking," in *Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality*, Redmond, U.S.A., Aug. 2018.