Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Rummukainen, Olli S; Schlecht, Sebastian J; Robotham, T; Plinge, Axel; Habets, Emanuël A P

Perceptual Study of Near-Field Binaural Audio Rendering in Six-Degrees-of-Freedom Virtual Reality

Published in:
2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)

DOI:
10.1109/VR.2019.8798177

Published: 01/01/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Rummukainen, O. S., Schlecht, S. J., Robotham, T., Plinge, A., & Habets, E. A. P. (2019). Perceptual Study of Near-Field Binaural Audio Rendering in Six-Degrees-of-Freedom Virtual Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* IEEE. https://doi.org/10.1109/VR.2019.8798177

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Perceptual Study of Near-Field Binaural Audio Rendering in Six-Degrees-of-Freedom Virtual Reality

Olli S. Rummukainen* Sebastian J. Schlecht Thomas Robotham Axel Plinge
Emanuël A. P. Habets

International Audio Laboratories Erlangen[1], Germany

ABSTRACT

Auditory localization cues in the near-field (< 1.0 m) are significantly different than in the far-field. The near-field region is within an arm's length of the listener allowing to integrate proprioceptive cues to determine the location of an object in space. This perceptual study compares three non-individualized methods to apply head-related transfer functions (HRTFs) in six-degrees-of-freedom near-field audio rendering, namely, far-field measured HRTFs, multi-distance measured HRTFs, and spherical-model-based HRTFs with near-field extrapolation. To set our findings in context, we provide a real-world hand-held audio source for comparison along with a distance-invariant condition. Two modes of interaction are compared in an audio-visual virtual reality: one allowing the participant to move the audio object dynamically and the other with a stationary audio object but a freely moving listener.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual Reality; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies;

1 INTRODUCTION

The space in close proximity to our body, the peripersonal space, has a special significance for our nervous system. Within this region objects can be grasped and manipulated without moving toward them. Avoiding potentially harmful objects is the most critical when they are close to our body. It may therefore be expected that the brain represents objects in the peripersonal space differently than objects at a distance. Body-part-centered reference frames have been identified for the visual peripersonal space, where receptive fields of multimodal neurons are activated only when a visual stimulus is located close to a specific body part [10]. Such integrated systems also exist for the auditory peripersonal space, where specific multimodal neurons respond to auditory-tactile stimuli within the proximity of the head, but not when the auditory stimulus is further away [8]. The study at hand presents a perceptual comparison of methods for binaurally rendering this near-field effect in a virtual reality environment.

To identify near-the-head sound sources, the auditory system uses a set of binaural cues that differ from the far-field cues. In the far-field (> 1.0 m), the head-related transfer functions (HRTFs) are practically independent of distance [3]. Within 1.0 m distance, as a point source moves lateral to the head, the interaural level difference (ILD) at low frequencies increases. In contrast for lateral sources in the far-field, the ILDs at low frequencies are very small [4, 5]. The low frequency ILD is most prominent when the source is within 0.5 m radius, reaching ILDs of 23 dB at 0.25 m distance [2]. Similarly, the interaural time difference (ITD) is affected by the source distance in the near-field for lateral sources, although the effect is less pronounced than for the ILDs. Measured from a dummy-head, the ITD increases on average by 100 μs moving from 10.0 m distance to 0.125 m [4] for a lateral sound source. Most of the increase occurs within the final 0.25 m from the head. For elevated sources, the HRTFs are fairly independent of distance, although some distance dependencies were recently found for sources close to the interaural axis [19]. For a precise description of the effects summarized here, we refer the reader to [4]. Apart from the binaural cues, more salient distance cues arise monaurally from the sound intensity and the direct-to-reverberant ratio [23]. However, the applicability of these cues depends heavily on our familiarity with the sound source and the acoustic environment.

Faithfully recreating the binaural localization cues may be essential for an immersive virtual audio experience. Currently, a set of HRTFs is a requirement for high quality binaural audio reproduction. They are typically measured, or modeled, only in the far-field, which results in a lack of near-field localization cues. The near-field cues may be re-introduced by modeling the near-field effect via a so-called distance variation function (DVF) [11, 21], which is shown to improve distance localization in the near-field compared to intensity cues only. The HRTFs may also be measured at multiple distances, which preserves the distance-dependent localization cues, but is a time consuming and laborious process [13].

In six-degrees-of-freedom (6 DoF) virtual reality, interaction in the near-field may take different forms. We may either approach a stationary sound object and inspect it without direct interaction, or, once within our peripersonal space, choose to actively move it. Sensory-motor coupling has been argued to form a basis for human cognition, where motor actions support sensory information processing [7]. Following this idea, holding a sound source, real or virtual, in our hand, the resulting auditory perception may be more accurately localized due to the proprioceptive cues compared to a case without direct interaction. On the contrary, spatial auditory resolution in the far-field during whole-body translation is impaired when compared to stationary listening [18]. Auditory reachability is a related cognitive concept, which has been shown to yield accurate distance estimates in the peripersonal space, further highlighting the cognitive importance of this region [16].

A few studies have evaluated the effect of near-field measured HRTFs on the accuracy of distance perception. In these studies, it was found that distance estimation based on binaural cues only utilizing loudness-normalized stimuli is not feasible [2]. Adding 3 DoF head tracking does not increase the importance of binaural cues in distance judgments or improve the overall accuracy [14]. Further studies with a modeled near-field effect found that an approaching relative motion is perceived more readily than a receding one [20].

In this study, near-field audio rendering is investigated in 6 DoF virtual reality. The focus is on general preference of near-field audio rendering given self-movement cues and a corresponding visual envi-

*e-mail: olli.rummukainen@iis.fraunhofer.de

[1]A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

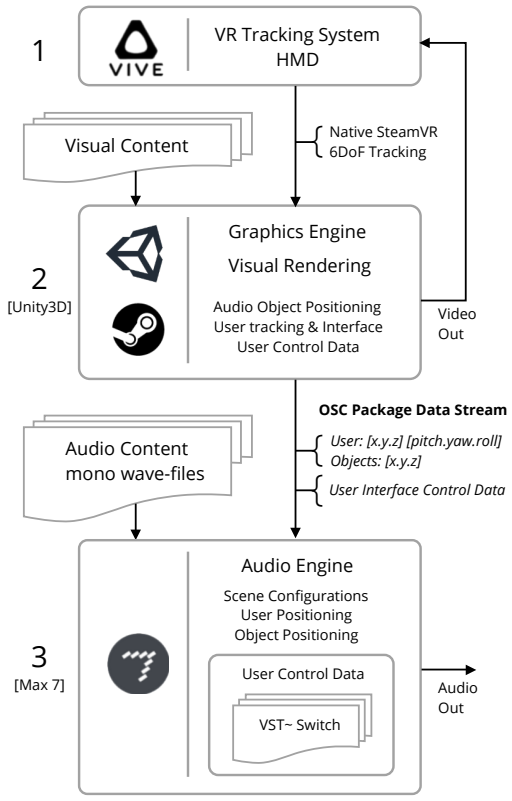


Figure 1: Overview of the real-time evaluation platform.

ronment, in contrast to previous unimodal auditory distance perception studies. We examine the influence of modeled near-field effect, far-field or multi-distance measured HRTFs, and intensity-invariant rendering on the desirability of the audio stimulus. We hypothesize that the real-world measured multi-distance HRTFs will yield the highest preference scores, whereas modeled, far-field measured, and intensity-invariant conditions will be scored lower. Furthermore, we assume holding the audio object in hand and dynamically moving it around the 6 DoF near-field will result in different preference weighting compared to 6 DoF full body motion with a static audio object. Finally, we hypothesize that a real-world loudspeaker will be the most desired audio condition instead of the virtual reproduction.

2 Method

2.1 Virtual reality environment

An important aspect of critical audio evaluation is the ability to switch between systems under test without relying on participants' auditory memory of conditions. This real-time comparison of conditions introduces an added complication in 6 DoF where the participant may ultimately dictate the audio content by their movements. While the fundamental audio content within a virtual scene is the same, participants' varying position and orientation means that audio in one participant's experience will be different to another.

A platform for real-time evaluation of binaural renderers has been developed allowing participants to switch between conditions, with no interruption to audio-visual sensory input. The basic structure is presented in this section; for a thorough walk-through, please see [15]. The platform may be broken up into three components: 1) VR device, 2) Graphical rendering engine, and 3) Audio rendering engine, as depicted in Figure 1. For Component 1, the HTC VIVE

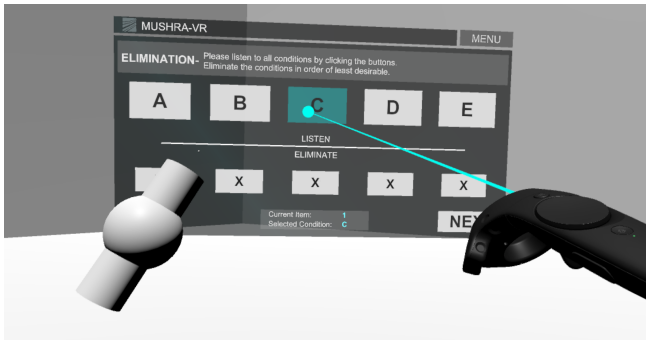


Figure 2: Virtual environment and the user interface. The sound source is modeled by the cylinder with a sphere denoting the audio object location.

Pro[2] head-mounted display (HMD) is used for positional tracking, visual presentation, and control interface. The tracking accuracy and latency are found suitable for reproducible scientific research [12]. For Component 2, game engine Unity3D is used for the graphical rendering engine, along with hosting positional information for all audio objects and participant's position and orientation using the SteamVR asset. All relevant positional and rotational data is then sent (via an Open Sound Control (OSC) data package at a 10 ms interval) to Max 7. In Component 3, binaural renderers are hosted in Max 7 and are fed, in parallel, the positional and rotational information received from Unity3D. All audio content is loaded into Max on a scene by scene basis and triggered to play, when the respective scene is loaded inside Unity3D.

Normally, interfaces such as a mouse and/or a keyboard may be used to control such evaluation tests. However, as participants will not be positioned at a static location, the test control interface is implemented inside the VR environment itself, allowing full freedom of movement while not being forced to return to a specific location to interact with the experiment interface.

The interface is designed such that it can be instantiated anywhere in the VR scene. By pressing a button on a hand-held controller, a semi-transparent panel appears at eye level in the participants' field of view. The panel is presented in Figure 2. Pressing the button again hides the panel, allowing the user to fully explore the environment. When instantiated, a virtual laser pointer could be used to target buttons having traditional button-states (highlighted, pressed, inactive) for visual feedback of interaction.

2.2 Stimuli

The audio-visual virtual reality scene was composed of a large empty cube-shaped room created in the Unity3D game engine. The sound source was visually modeled as a cylinder having the same dimensions as its real-world counterpart: an UltimateEars Boom 2 loudspeaker[3]. The location of the virtual audio object was defined at the center of the real-world loudspeaker. The real loudspeaker was tracked by a 6 DoF tracker mounted on top of the loudspeaker, allowing to create an accurate virtual representation of it at the same spatial location. In the middle of the virtual cylinder, there was a sphere denoting the location of the sound object. The virtual audio object was played back via open-design AKG K1000 headphones mounted rigidly on the HMD. The open-design headphones allowed mostly unobstructed sound propagation from the real loudspeaker even when wearing the headphones. An RME UCX audio interface was used and all audio was processed at 48 kHz in 128 sample blocks.

[2]https://www.vive.com/eu/product/vive-pro/
[3]https://www.ultimateears.com/en-us/wireless-speakers/boom-2.html

Figure 3: Cortex Mk I dummy-head during calibration and the ceiling-suspended loudspeaker with a tracker mounted on top. The dummy-head is wearing an HTC Vive Pro HMD and the AKG K1000 open-design headphones.

Sound sources were selected to be an anechoic cello recording and pink noise. On the one hand, the cello sample is a natural signal that contains tonal structure and temporal variations that may mask rendering impairments in real application scenarios. On the other hand, these features may result in different rendering artefacts than in the pink noise case.

Both source signals were high-pass filtered with a cutoff frequency of 100 Hz. Their loudness was equalized at 0.5 m distance from the audio object with a diffuse field calibrated Cortex Mk I dummy-head to be 55 dB(A) sound pressure level for all the renderers and the real loudspeaker. The sound object was placed in front of the dummy-head, as shown in Figure 3. The inverse square law was applied by the audio renderers to realize distance gain with a minimum distance of 0.1 m from the center of the head. The experiment was run in a rectangular room with following dimensions: length 4.70 m, width 3.50 m, and height 2.80 m. The octave band reverberation time (RT_{60}) was measured to be 0.47 s at 1 kHz. The room acoustics affected only the real loudspeaker; the virtual room was modeled as anechoic. Thus, the direct-to-reverberant ratio cue was only available for the loudspeaker condition. Additionally, the real loudspeaker had a height of 0.18 m and a diameter of 0.067 m, which was not modeled by the binaural renderers. Therefore, close to the head, the real loudspeaker would become less point-like, whereas the virtual sound sources remain as point sources.

Two different binaural audio rendering principles were employed in the experiment: a structural-model-based parametric renderer and an HRTF-convolution-based renderer. For the model-based renderer, the ITD and ILD values were computed from a spherical head model [1, 6]. The processing consisted of a time-varying delay line and second-order shelving filters. No individualization was performed and the radius of the head model was set to 87 mm. No pinna model and therefore no elevation adjustment were included, but the near-field effect was modeled with the DVF method proposed in [21]. The HRTF-convolution renderer relied on a set of measured HRTFs, where it would select the pair that is closest to the sound object location. Both renderers were capable of realizing real-time rendering with 6 DoF tracking. There was no room acoustics modeling included.

Table 1: Renderers under test

Renderer	**Features**
Real loudspeaker	Ultimate Ears Boom 2
Modeled	Spherical head-model and near-field modeling
Far-field	HRTFs measured at 1.5 m
Multi-distance	HRTFs measured from 0.5 m to 1.5 m
Distance-invariant	Far-field HRTFs and no distance-based gain

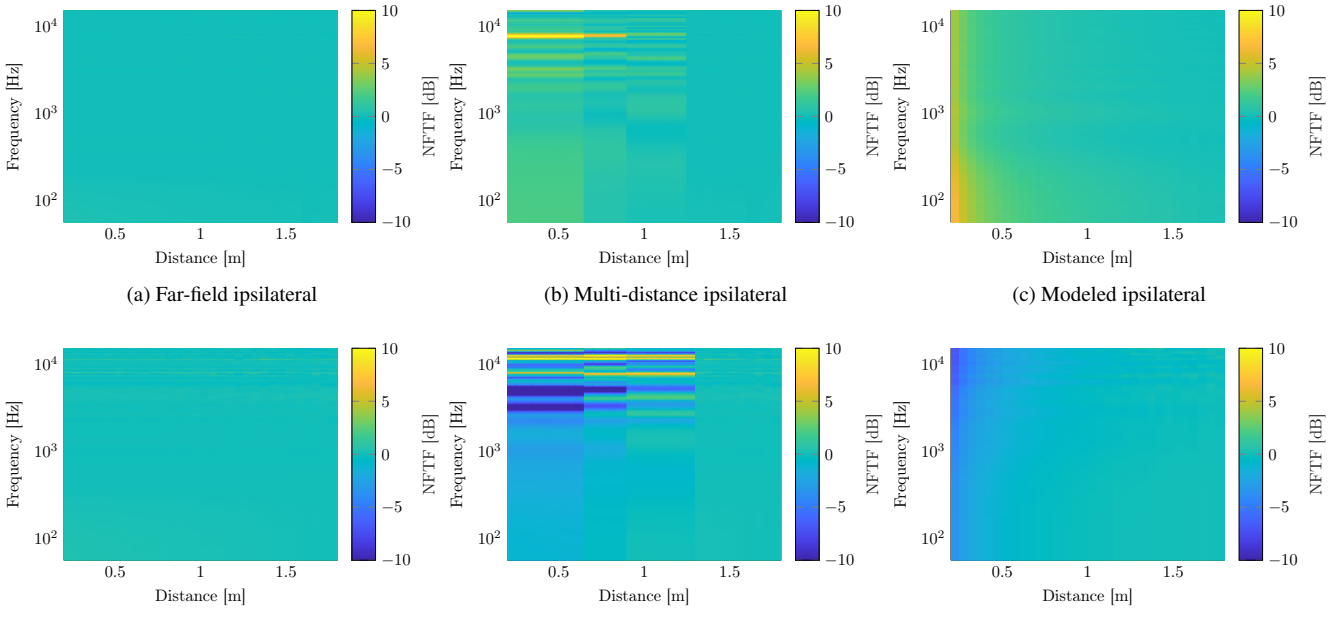
Additional conditions were created by adding three different versions of the convolution renderer with different HRTF sets and a gain setting. There was an HRTF set measured only in the far-field at 1.50 m distance, a multi-distance set with measurements at 0.50 m, 0.75 m, 1.00 m, and 1.50 m, and finally a version with the far-field HRTFs without distance gain adjustment. The HRTF sets were measured from a Neumann KU100 dummy head with a 2702-node Lebedev grid. The HRTF sets are publicly available from [13]. Together with the real loudspeaker there were five renderers in the test, summarized in Table 1.

Figure 4 displays simulated distance variation functions in the near-field for the different renderers when compared to the far-field HRTF normalized magnitude spectrum. The resulting spectra show only the distance-variation-dependent binaural effects and are called the near-field transfer functions (NFTF). The responses are simulated for both ears with a sound source located on the interaural axis. As expected, the far-field renderer does not produce any additional near-field cues (Panels 4a & 4d). The multi-distance renderer displays four prominent regions with increasingly stronger near-field cues as the distance decreases (Panels 4b & 4e). The model-based renderer displays a smoother but prominent effect of the near-field (Panels 4c & 4f). Notably, the modeled near-field effect modifies the response all the way up to the ear, whereas the multi-distance HRTF response does not anymore change within 0.5 m from the head.

Two scene types were employed to enable the study of source-listener interaction: *static* and *dynamic*. In the *static* scene, the real loudspeaker, and its virtual counterpart, was suspended from the ceiling at 1.5 m height from the floor. In this scene the participants were not allowed to directly manipulate the sound source; they were instructed to move around it themselves. In the *dynamic* scene the loudspeaker was detached from the ceiling and the participants held it in their hand whilst seated. Here, they were able to directly manipulate the sound source's location in space and also get accurate proprioceptive location information thanks to the hand-based interaction. Figure 5 shows a person wearing the HMD in the *dynamic* scene with the loudspeaker in one hand and the controller in the other. In both scenes the virtual loudspeaker was visually present resulting in further multi-modal location cues. Altogether, the participants completed four scene combinations: two different scene types and two source signals.

2.3 Participants

There were a total of 21 participants (5 female, 16 male). Their average age was 32.0 years (SD = 9.0). All of the participants provided a written informed consent to take part in the study. Nineteen of the participants reported having some experience of using VR equipment. All of the participants reported having normal auditory function.



(a) Far-field ipsilateral (b) Multi-distance ipsilateral (c) Modeled ipsilateral

(d) Far-field contralateral (e) Multi-distance contralateral (f) Modeled contralateral

Figure 4: Near-field transfer functions (NFTFs) for a source on the interaural axis from the three HRTF sets: far-field, multi-distance, and modeled. The NFTFs are obtained by normalizing the magnitude spectrum of the distance-dependent HRTFs with the corresponding far-field HRTFs [21].

2.4 Procedure

The participants were instructed to eliminate the audio renderers in the order of least desirability [17, 22]. It was made clear that there is no audio reference, but their self-motion and the visual presentations should be understood as creating the reference for expected auditory stimulation. The real loudspeaker was explicitly not defined as a reference, since it would have potentially biased the listeners' attention to factors that may not be relevant for the overall experience, and the maximum achievable quality would have been defined by it. Instead, quality degradation was expected to result from mismatch between the expectations and the perceived auditory stimulus. The participants were instructed to pay attention especially on spatial impression, i.e., how well the auditory stimulus is co-located with the visual and haptic sensation. The interface (Figure 2) allowed the participants to switch between the five renderers via buttons (A-E) as many times as required. They eliminated iteratively the least desirable renderer via an eliminate button. Since all renderers continuously received tracking data they remained updated on the current position and orientation of the listener with respect to the audio object making the switch between renderers seamless.

The participants were first familiarized with the VR system and interface in a special scene before beginning with the actual experiment. They were instructed on how to operate the controller to bring up and hide the control panel. They were also able to go through a dummy elimination test without audio to get used to the interface buttons. The four experiment scenes were evaluated after the familiarization scene. The order of static and dynamic scenes was counterbalanced within the participants, but the cello sample was always evaluated before the pink noise sample, since the noise sample was considered as more critical for evaluating the renderers. This was done not to bias the participants to listen for artefacts that may be only salient with the noise signal but not with the natural cello signal. Average time to complete evaluation of the four scenes after the setup and familiarization was 13.4 min (SD = 5.5 min).

3 Results

3.1 Preference modeling

In this analysis the rankings are assumed to result from a series of independent paired comparisons between the conditions. Given J objects (indexed by j and k, $j < k$), a Bradley-Terry (BT) model defines one paired comparison j,k where the probability to prefer sample j over sample k ($j \succ k$), or vice-versa, is formulated as follows [9]:

$$p(j \succ k|\pi_j,\pi_k) = \frac{\pi_j}{\pi_j+\pi_k} \quad \text{or} \quad p(k \succ j|\pi_j,\pi_k) = \frac{\pi_k}{\pi_j+\pi_k}, \quad (1)$$

where the π are the locations of samples on a preference scale limited between $0 \leq \pi \leq 1$. Equation 1 may be further formulated as:

$$p(y_{j,k}) = \left(\frac{\sqrt{\pi_j}}{\sqrt{\pi_k}}\right)^{y_{j,k}}, \quad (2)$$

where $y_{j,k}$ is a response to a comparison j,k and takes the value of 1, if object j is preferred to k and value of -1 in the opposite case. The probability of a pattern of paired comparisons, or a ranking, with a response of $\mathbf{y} = (y_{1,2}, y_{1,3}, ..., y_{j,k}, ..., y_{J-1,J})$ is finally given by [9]:

$$p(\mathbf{y}) = \prod_{j<k}\left(\frac{\sqrt{\pi_j}}{\sqrt{\pi_k}}\right)^{y_{j,k}}, \quad (3)$$

which is the product of all pairwise probabilities defined by a given ranking. The preference scale locations are estimated from the ranking data with the R-package *prefmod* [9].

Figure 6 displays the estimated preference scores separately for each scene. The resulting preference scale is a ratio scale, meaning there is a real zero and distances between breaks on the scale are equal. The preference scores sum up to one in each scene. In the *static* scene with the cello sample a large difference in preference may be observed from the least preferred conditions *modeled* and *distance-invariant* to the most preferred *multi-distance* condition,

Figure 5: Participant during the *dynamic* hand-held interaction scene. Interaction with the experiment is done via the Vive controller and a user interface within VR.

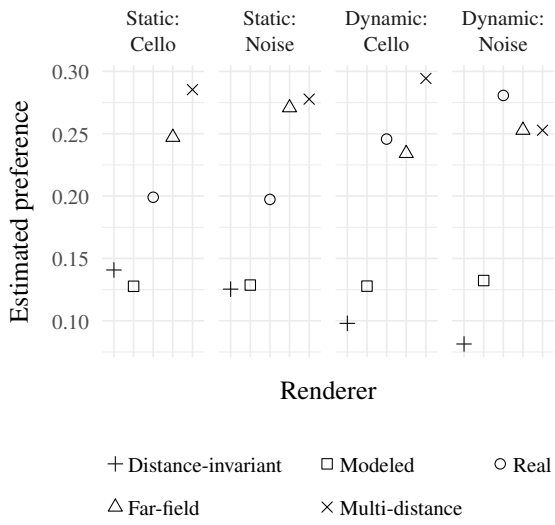


Figure 6: Preference modeling parameter estimates.

which is almost twice more likely to be preferred than the *modeled*. The *real* condition is rated almost exactly in the middle between the least and most preferred binaural renderers. The *far-field* condition gets closest to the *multi-distance*.

The *static* scene with pink noise sample shows similar overall ranking of the conditions, but the scale is more compressed from the lowest and highest ratings than with the cello sample. There are no longer preference differences between the *modeled* and the *distance-invariant*, or the *far-field* and the *multi-distance*.

Both the *dynamic* scenes, with cello and pink noise, show similar overall ordering of the conditions. The *real*, *far-field*, and *multi-distance* are closely together with the highest preference scores, while the *modeled* and *distance-invariant* remain significantly lower on the scale. The *multi-distance* achieves the highest rating with the cello sample, while the *real* is rated the highest with pink noise. The *distance-invariant* is rated clearly lower than the *modeled* in the *dynamic* scenes but not in the *static* scenes.

3.2 Tracking data

The participants' and the loudspeaker's location was recorded in 6 DoF during the experiment at an 0.1 s interval. This data may be used to shed more light on the obtained preference ratings, especially since the participants' movements were not guided in any way during the experiment. Figure 7 presents a derived preference modeling with an added participant-specific covariate calculated from the variation in the distance between the source and listener. This is a binary covariate obtained by normalizing the source-listener distance throughout a scene and finding the variance of this data. Finally, the participant pool is divided into two parts based on the amount of found variation by labeling the lower half of them with *small* and the upper half with *large*.

In Figure 7, through all the scenes, the left-most estimates are obtained from the participants who did relatively less distance-varying movements in favor of rotational movement either around the sound source (*static*) or by moving the source around their head (*dynamic*). Moreover, participants scoring low in the distance variation scale may have moved relatively less in general. This type of movement is of specific interest since distance estimation in the near-field has been so far the most studied perceptual aspect of near-field audio rendering. By classifying people according to the distance variation, we may isolate the sub-group who have been affected by the potentially more effective binaural distance cues in the near-field.

The most notable effects in Figure 7 happen on the *far-field* condition in *static: cello* scene and on the *multi-distance* condition in the *dynamic: cello* scene. The *far-field* condition increases in preference the more participants induced variations in distance by moving around the sound source themselves. In contrast, in the *dynamic: cello* scene the *multi-distance* condition shows reduced preference by the large distance variation sub-group. More consistent effects in the *dynamic* scenes are found for the *real* loudspeaker which increased in preference for the sub-group who showed relatively more distance variation. An opposite, but more subtle, effect is observed in the *static* scenes. Effects for the other conditions are less pronounced and not equally consistent.

Figure 8 shows the median time spent in each scene. The *dynamic* and *static* scenes were counterbalanced within participants to remove ordering effects. There is potentially a learning effect moving from the cello sample to noise, since they were always evaluated in this order. A significant reduction of time can be observed with the cello samples between the *dynamic* and *static* scene, where the median value for *static* scene is 220 s and for the *dynamic* 169 s. Similar effect is not observed between the noise samples, which were always evaluated after the cello scene.

A summary statistic of the median distance between the source and listener measured from the center of the head to the sound object is shown in Figure 9. The distance is approximately 0.20 m shorter in the *dynamic* scenes compared to the *static* scenes.

4 Discussion

The *far-field* and *multi-distance* conditions were consistently the most preferred in 6 DoF multimodal virtual reality. Looking at

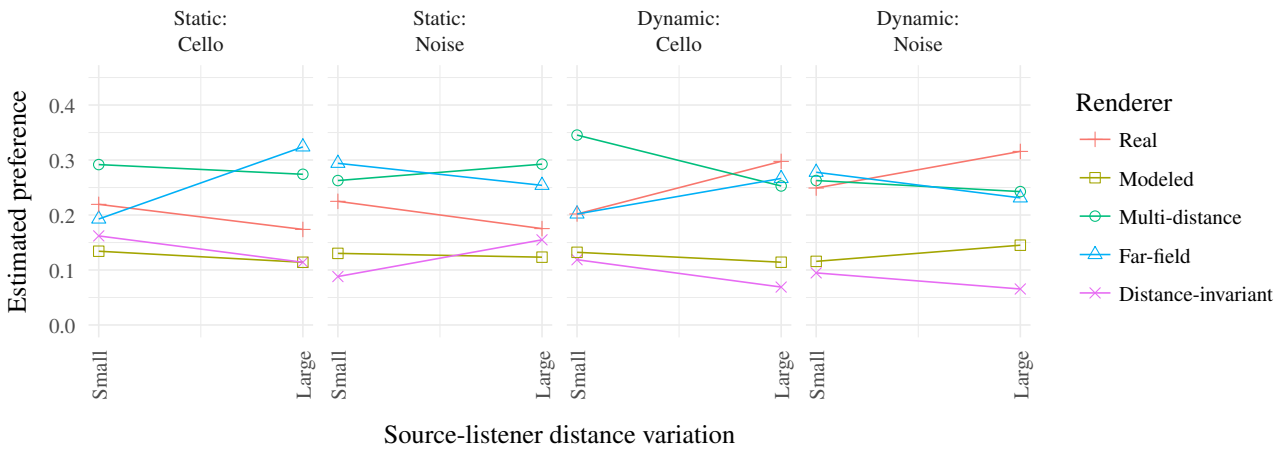


Figure 7: Preference estimates with a participant-specific covariate for variation in distance between source and listener.

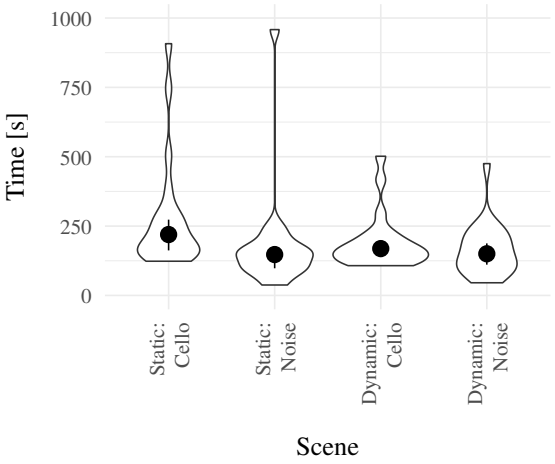


Figure 8: Median time spent in a scene with bootstrapped 95 % confidence intervals. The violin plots show the probability density of the data.

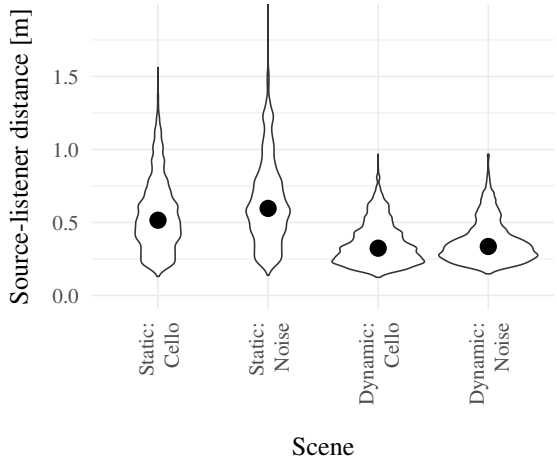


Figure 9: Median source-listener distance in a scene. The violin plots show the probability density of the data.

the overall preference modeling (Figure 6), there appears to be little benefit of including HRTFs measured at multiple distances for enhanced near-field effect. However, the nearest distance in the used HRTF set was 0.50 m, which is considered to be the threshold within which most of the near-field distance-dependent cues are the most prominent [2, 4]. Therefore, future studies must include HRTFs from even closer distances to get a more complete picture of their effect. Interesting directions for further research include high resolution near-field HRTFs from, *e.g.*, numerical simulations from participants' head scans or from otherwise individualized head models.

The *multi-distance* condition appears to be more preferred than the *far-field* condition in both the *static* and *dynamic* scenes with the *cello* source signal. This signal includes temporal variations and a continuous tonal structure which potentially facilitate the discrimination of the subtle near-field cues while masking the artefacts of HRTF switching in the distance dimension. Therefore, our first hypothesis on the highest preference for multi-distance HRTFs receives at least partial support.

The *real* loudspeaker condition received the highest score in only one of the tested scenes (*dynamic*: *noise*). It performed especially poorly in the *static* scenes, which rejects our hypothesis that the *real* loudspeaker would be preferred instead of the virtual renderings. There are a few factors potentially explaining the low preference of the loudspeaker condition. Firstly, it was the only condition with room acoustics included, since the experiment was done in a normal quiet listening room instead of an anechoic chamber. This resulted in early reflections and late reverberation that did not match the virtual space, and which were missing from the binaural conditions. Second, the timbre of the loudspeaker was perceivably different compared to the headphone stimuli, despite the high-pass filtering and loudness calibration. Third, the distance gain seemed to slightly differ between the real loudspeaker and the binaural conditions at close distances. It is likely that the far-field defined inverse square law does not apply in the near-field. Moreover, the loudspeaker is not a point-source that is usually assumed in loudness experiments.

Most participants commented informally after the experiment having identified the real loudspeaker, but still having preferred the binaural conditions since they were more point-like and appeared to

match what they were seeing in the virtual environment better. In future experiments with real-world sound sources compared with virtual sounds a better matching of the conditions is essential. This may be achieved for example by equalizing the source signals for different playback systems, having an anechoic chamber or a room acoustic model matching the real space, employing a smaller loudspeaker, and modeling the sound producing object more truthfully in the virtual environment (*e.g.* source extent).

Another surprising finding is the poor performance of the modeled near-field effect, which was found in previous research to aid in distance estimation [20, 21]. Here, the results may be more due to the underlying rendering method, which relied on HRTF modeling from a spherical head model. The lack of elevation cues in this model is likely such a critical shortcoming that the potentially helpful near-field modeling is masked by it. Therefore, drawing conclusions on the goodness of the DVF method is not feasible based on the results presented here. For future experiments with the DVF modeling the binaural rendering principle should be kept similar across all conditions.

The two scene types, *static* and *dynamic*, yielded different preference ratings and durations. Judging by the duration only, the hand-based direct manipulation of the sound source appears to be more intuitive and to enable faster elimination decisions based on the scenes with the cello samples. The participants saw either of these scenes first in the test. The *real* loudspeaker was ranked in the *dynamic* scenes equally good as the best binaural conditions, which did not happen in the *static* scenes. Furthermore, the source was on average closer to the head in the *dynamic* scenes (Figure 9) reducing the effect of room acoustics. In these scenes the localization was further aided by proprioceptive cues.

5 Conclusions

This perceptual study examined the near-field binaural audio rendering with non-individualized HRTFs. The outcomes of this study suggest the use of far-field or multi-distance measured HRTFs for near-field rendering. The multi-distance HRTFs were preferred over the far-field HRTFs for a complex source signal with temporal variation (cello). More studies will be needed with the modeled near-field effect. The real loudspeaker was equally preferred with the best binaural renderings when dynamic source manipulation was allowed, but without direct interaction it was preferred less. The dynamic scenes were faster to complete and yielded differing preference rankings compared to the static scenes. Overall, this study provides the groundwork for further investigations into auditory perception in the near field in 6 DoF virtual environments.

Acknowledgments

The authors wish to thank Simon Schwär for implementing and configuring the HRTF-convolution based renderer used in this study.

References

[1] V. R. Algazi, C. Avendano, and R. O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *The Journal of the Audio Engineering Society*, 49(6):472–479, 2001.

[2] J. M. Arend, A. Neidhardt, and C. Pörschmann. Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set. In *29th Tonmeistertagung*, pp. 356–363. Cologne, Germany, 2016.

[3] D. S. Brungart. Near-field virtual audio displays. *Presence: Teleoperators and Virtual Environments*, 11(1):93–106, 2002.

[4] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999.

[5] D. S. Brungart and B. D. Simpson. Auditory localization of nearby sources in a virtual audio display. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 107–110. New Paltz, NY, USA, 2001.

[6] R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. *The Journal of the Acoustic Society of America*, 104(5):3048–3058, 1998.

[7] A. K. Engel, A. Maye, M. Kurthen, and P. König. Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, 17(5):202–209, 2013.

[8] A. Farnè and E. Làdavas. Auditory peripersonal space in humans. *Journal of Cognitive Neuroscience*, 14(7):1030–1043, 2002.

[9] R. Hatzinger and R. Dittrich. prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, 18(10):1–31, 2012.

[10] N. P. Holmes and C. Spence. The body schema and multisensory representation(s) of peripersonal space. *Cognitive Processing*, 5(2):94–105, 2004.

[11] A. Kan, C. Jin, and A. van Schaik. A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. *The Journal of the Acoustical Society of America*, 125(4):2233–2242, 2009.

[12] D. C. Niehorster, L. Li, and M. Lappe. The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research. *i-Perception*, 8(3):1–23, 2017.

[13] C. Pörschmann, J. M. Arend, and A. Neidhardt. A Spherical Near-Field HRTF Set for Auralization and Psychoacoustic Research. In *Audio Engineering Society 142nd Convention*, pp. 1–5. Berlin, Germany, 2017.

[14] C. Pörschmann, J. M. Arend, and P. Stade. Influence of head tracking on distance estimation of nearby sound sources. In *DAGA*, pp. 1065–1068. Kiel, Germany, 2017.

[15] T. Robotham, O. Rummukainen, and E. A. P. Habets. Evaluation of Binaural Renderers in Virtual Reality Environments: Platform and Examples. In *Audio Engineering Society 145th Convention*, pp. 1–5. New York, NY, USA, 2018.

[16] L. D. Rosenblum, A. P. Wuestefeld, and K. L. Anderson. Auditory Reachability: An Affordance Approach to the Perception of Sound Source Distance. *Ecological Psychology*, 8(1):1–24, 1996.

[17] O. S. 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 Conference on Audio for Virtual and Augmented Reality*, pp. 1–10. Redmond (WA), USA, 2018.

[18] O. S. Rummukainen, S. J. Schlecht, and E. A. P. Habets. Self-translation induced minimum audible angle. *The Journal of the Acoustical Society of America*, 144(4):EL340–EL345, 2018.

[19] S. Spagnol. On distance dependence of pinna spectral patterns in head-related transfer functions. *The Journal of the Acoustical Society of America*, 137(1):EL58–EL64, 2015.

[20] S. Spagnol, E. Tavazzi, and F. Avanzini. Relative Auditory Distance Discrimination With Virtual Nearby Sound Sources. In *8th Int. Conference on Digital Audio Effects (DAFx-15)*, pp. 1–6. Trondheim, Norway, 2015.

[21] S. Spagnol, E. Tavazzi, and F. Avanzini. Distance rendering and perception of nearby virtual sound sources with a near-field filter model. *Applied Acoustics*, 115:61–73, 2017.

[22] F. Wickelmaier, N. Umbach, K. Sering, and S. Choisel. Comparing three methods for sound quality evaluation with respect to speed and accuracy. In *Audio Engineering Society 126th Convention*, pp. 1–10. Munich, Germany, 2009.

[23] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United with Acustica*, 91(3):409–420, 2005.

View publication stats