Localization Performance in the Absence of Visual Cues for Binaural Renderings generated with a Virtual Artificial Head

Mina Fallahi[1], Martin Hansen[1], Steven van de Par[2,4], Simon Doclo[2,4], Dirk Püschel[3] and Matthias Blau[1,4]

[1] *Jade Hochschule Oldenburg, Institut für Hörtechnik und Audiologie*

[2] *Carl von Ossietzky Universität Oldenburg, Dept. für medizinische Physik und Akustik*

[3] *Akustik Technologie Göttingen*, [4] *Exzellenzcluster Hearing4All*

Introduction

A realistic spatial impression of a captured acoustical scene can be achieved by including Head-Related Transfer Functions (HRTFs) into the recordings. The Virtual Artificial Head (VAH) comprises a microphone-array based filter-and-sum beamformer that can be used to synthesize the directivity patterns of individual HRTFs by applying individually calculated spectral weights to the microphone signals. It has been demonstrated that the VAH can be used as an alternative to conventional artificial heads to create convincing virtual acoustic scenes [1]. Two key features of this technology are that (a) the same recording can be individualized *post-hoc* for different listeners, and (b) head tracking can be applied during listening by applying the spectral weights corresponding to the current head orientation of the listener.

In a recent study of the authors [2], the VAH was evaluated as perceptually convincing with respect to different perceptual attributes in a head-tracked binaural scenario and in direct comparison to the real sound sources in a reverberant room. Since in this setting the listeners could see the sound sources, it was not clear, to which extent the visual information about the sound source could have promoted the perception, especially in the reverberant environment, where the localization performance is less accurate. In the current study, a new experiment was performed to evaluate the localization performance with VAH signals in an anechoic room and in the absence of any visual cues. As an important part of such a localization task, a method had to be created to gather the responses, since different response techniques can lead to different localization performances [3]. Here, subjects had to map their responses onto a Graphical User Interface (GUI) while listening to virtual sources. A similar listening test was performed in which subjects listened to hidden real sources using the same GUI. The results indicated that also in the absence of visual cues, the localization performance with virtual sources generated with the VAH can be comparable to that of the real sound sources.

Spectral weights for the Virtual Artificial Head (VAH)

The directivity pattern $H(f, \theta_k)$ of the VAH as a filter-and-sum beamformer is defined as

$$H(f, \theta_k) = \mathbf{w}^H(f)\mathbf{d}(f, \theta_k), \tag{1}$$

with f denoting the frequency and θ_k the direction. The $N \times 1$ steering vector $\mathbf{d}$ is defined as the measured free-field acoustical transfer function between source at θ_k and the N microphones of the VAH and the $N \times 1$ vector $\mathbf{w}(f)$ contains the complex-valued spectral weights for the N microphones. To synthesize the desired directivity pattern $D(f, \theta_k)$ of the left or right HRTFs at $k = 1, 2, ..., P$ discrete directions, a narrow-band least-squares cost function defined as

$$J_{LS}(\mathbf{w}(f)) = \sum_{k=1}^{P} |H(f, \theta_k) - D(f, \theta_k)|^2, \tag{2}$$

was minimized. In addition, constraints were imposed on the resulting Spectral Distortion (SD) at each direction θ_k, $k = 1, 2, ..., P$ such that for all k

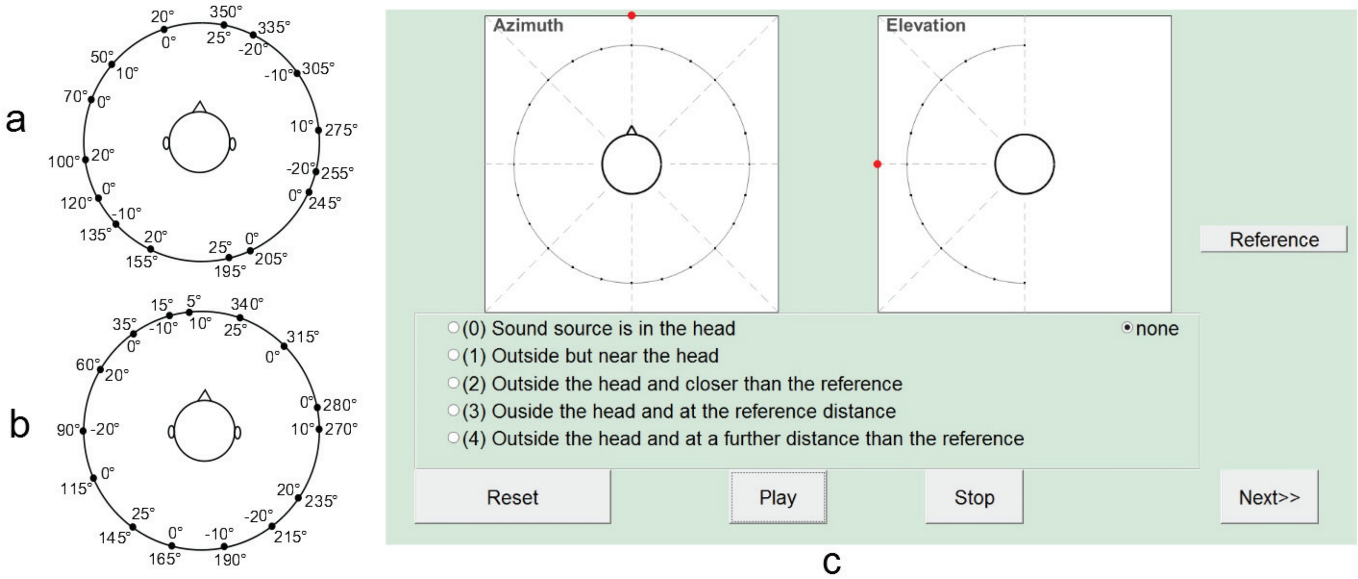
$$-1.5 \text{ dB} \leq \text{SD}(f, \theta_k) = 10 \lg \frac{\left|\mathbf{w}^{\text{H}}(f)\mathbf{d}(f, \theta_k)\right|^2}{|D(f, \theta_k)|^2} \text{dB} \leq 0.5 \text{ dB}. \tag{3}$$

The upper and lower limits chosen for the constraints on SD in Eq. 3 lead to a maximum of 2 dB deviation in the resulting Interaural Level Differences. Still, another constraint was imposed onto the minimum value of the resulting *mean* White Noise Gain (WNG$_\text{m}$) [4], in order to guarantee the robustness of VAH against microphone self-noise or deviations in the microphone characteristics, i.e.

$$\text{WNG}_\text{m} = 10 \lg\left(\frac{1}{P} \sum_{k=1}^{P} \frac{|\mathbf{w}^H(f)\mathbf{d}(f, \theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}\right) \text{dB} \geq 0 \text{ dB}. \tag{4}$$

The Interior-Point algorithm was used to solve this constrained optimization problem with the solutions proposed in [4] as initial values.

For the localization tests in this study, the VAH was a planar microphone array, 20 cm × 20 cm, with 24 microphones [1] and P= 72 horizontal directions were considered for the calculation of the spectral weights. The spectral weights were calculated for head orientations to the azimuth angles −90° to +90° in 5° steps and elevations −15° to +15° in 7.5° steps. For a given head orientation θ_h, $h \in 1, 2, ..., P$, this was done by taking the $D(f, \theta_k)$, $k = 1, 2, ..., P$, and the shifted steering vectors $\mathbf{d}(f, \theta_{k'})$ with $k' = h, h + 1, ..., P, 1, 2, ..., h - 1$ into Eq. 1 to 4.



○(0) Sound source is in the head

○(1) Outside but near the head

○(2) Outside the head and closer than the reference

○(3) Ouside the head and at the reference distance

○(4) Outside the head and at a further distance than the reference

●none

Figure 1: Tested source positions when localizing with (a) real sound sources (TestReal) and (b) virtual sources (TestVR). Numbers outside the circle indicate the azimuth angle and the ones inside the circle indicate the elevation angle of the sound source. (c): Graphical User Interface (GUI) for acquiring the responses on the source azimuth, elevation and distance. By clicking the 'Reference' button, subjects could listen to the reference source required for assessing the distance. The 'Reset' button was used to reset the head tracker during TestVR.

Methods

Two listening tests were performed in this study, consisting of localizing either virtual sound sources, referred to as TestVR, or localizing hidden real sound sources, referred to as TestReal. Both tests as well as the measurements were performed in the anechoic chamber of Institut für Hörtechnik und Audiologie at Jade University of Applied Sciences in Oldenburg. A total of ten normal-hearing subjects with individually measured HRTFs and Headphone Transfer Functions (HPTFs) took part in the test. The test signal was a dry recorded speech utterance of 15 s duration, spoken by a female speaker. For each test, 15 different source positions, as shown in Fig. 1a-b were considered.

A vertical loudspeaker arc with 1.2 m radius hanging from a turntable mounted in the ceiling was used to present the target sources. The center of the arc in the middle of the room at 1.24 m height was chosen as the Listener Position. Loudspeakers were mounted into the arc at the different target source elevations in Fig. 1a-b. The turn table was rotated to bring the loudspeaker arc to the azimuthal target source positions. This rotation was both invisible and inaudible for the listeners.

The Graphical User Interface (GUI) shown in Fig. 1c was used to gather the responses, including the perceived azimuth, elevation and distance. For azimuth, the GUI showed the head, seen from above, with a circle around it and the subjects could click anywhere on this circle to give their responses. Similarly, for elevation, the head was shown from the side with a semicircle for elevations ranging from $-90°$ to $+90°$. The reference point of azimuth and elevation = 0°, corresponding to the frontal head orientation, was marked on the GUI as well as in the room in front of the subjects.

To gather information about the perceived source distance, for both tests, subjects were supplied with a reference source in the room positioned at azimuth and elevation = 0°. Subjects could not see the reference source. They had to give the perceived target source distance compared to this reference source using a scale from 0 to 4, corresponding to their perception (0) in head, (1) outside but near the head, (2) outside the head and closer than the reference, (3) outside the head and at the reference distance, or (4) outside the head and at a further distance than the reference. The reference source (the same loudspeaker model as the target sources) was adjusted to have the same level as the target sources, but it was positioned about 50 cm further than the target sources to the Listener Position. For TestVR, subjects were asked to take off the headphones while listening to the reference source. The "Reset" button shown in the GUI was used during TestVR to reset the head tracker and was removed from the GUI for TestReal.

During TestReal, subjects sat with their interaural center at the Listener Position. In order to eliminate any visual cues, subjects were seated inside an acoustically transparent curtain (Fig. 2) and the room was darkened. The only faint light source was the monitor display in front of the subjects, which they used to guide the test and to give their responses. The loudspeaker arc was rotated to the azimuthal target positions shown in Fig. 1a and the test signal was played back from the loudspeaker channel corresponding to the target elevation. Subjects were allowed to turn their head in the range of ±90° horizontally and ±15° vertically. Each target source was presented once, but subjects could listen to the presentation as long as desired. The order of presentation was randomized.

For TestVR, the VAH was positioned at the Listener Position and the room Impulse Responses (IRs) were

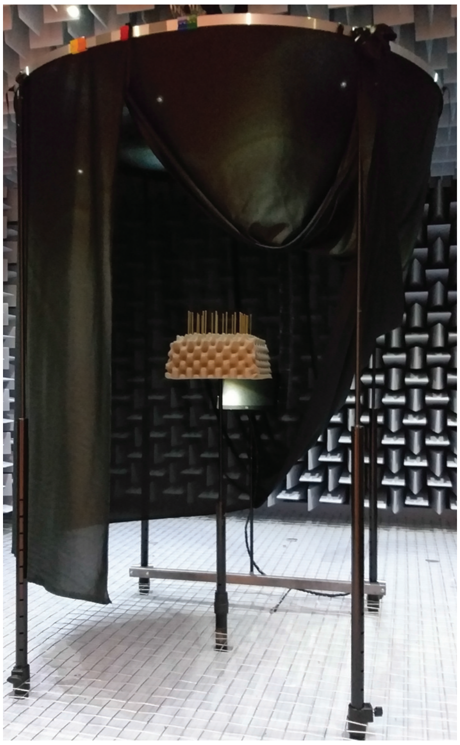


Figure 2: The VAH inside the acoustically transparent curtain at Listener Position. During the both listening tests subjects were seated at the Listener Position inside this curtain.

measured for the microphones of the VAH with respect to the target source positions in Fig. 1b. In order to keep the conditions comparable to TestReal, the VAH was positioned inside the acoustically transparent curtain (see Fig. 2). In order to create the Binaural Room Impulse Responses (BRIRs), the individually calculated complex-valued spectral weights for different head orientations were transformed to impulse responses and convolved with the measured IRs and the individual inverse HPTFs. During signal presentation, the test signal was dynamically convolved with the BRIRs for the current head orientation. Subjects sat at the Listener Position inside the acoustically transparent curtain in the darkened room. They wore the headphones with a custom made tracker mounted on the top of it to listen to head-tracked presentations of the virtual target sources. Each of the 15 target position was presented once. It should be noted that in addition to the BRIR set derived for the VAH as described above, four other BRIR sets were also evaluated using the same procedure. However, since the focus of the current contribution is on the test method, we discuss the results for the described variant only. The comparison between different VAH variants and a head-tracked version of a commercial dummy head is discussed in [5].

Results and discussion

Azimuth: Response vs. target azimuths of ten subjects in TestReal and TestVR are shown in the top row of Fig. 3a. Responses shown with a '×' were suspected as front-back reversals. The lower row of Fig. 3a shows the absolute error between target and response azimuths. Reversals were excluded from the error calculation and are given as percentage of all presentations. The horizontal line shows the average absolute azimuth error over all target azimuths, which was 7.9° in TestReal. 1.3% of the responses were identified as reversals. These results demonstrate the localization performance when listening to real sound sources as well as the ability of the subjects to map their responses onto the GUI. The virtual sources generated with the VAH in TestVR were localized with 7.6° average azimuth error and 1.3% reversals, which is in perfect agreement with the results of TestReal.

Elevation: Response vs. target elevations of ten subjects and the signed elevation error (subtracting target from response elevation) are shown in the upper and lower of Fig. 3b, respectively. For TestReal, responses to target elevations between $-20°$ and $+25°$ extended from $-54°$ to $70°$. Subjects tended to overestimate the positive elevations and to underestimate negative elevations, which can also be observed in the signed error. This might have been caused by the difficulty of mapping the elevation responses onto the GUI. For TestVR, the positive signed error for negative and zero elevations indicates that subjects perceived these elevations at higher elevations. In contrast, positive elevations were perceived often at lower frequencies. In general, elevation perception with the VAH signals was less accurate compared to azimuth perception. This could be due to the fact that for the VAH variant discusses here, only horizontal source directions were considered in the calculation of spectral weights.

Distance: Fig. 4 shows the given source distance of real or virtual sources as scatter diagram vs. target azimuths. The area of each circle shows how many subjects chose each distance percept. Although the reference source was 50 cm further than the target sources, in TestReal, the majority of the subjects chose the score (3) which means that they perceived the target sources at the same distance as the reference source. As expected, none of the subjects perceived real sources in or near the head. In TestVR, 2.6% of the total given responses for the virtual source distance were in or near the head whereas the majority of responses were similar to the responses in TestReal.

As the results show, VAH signals led to similar localization performance as real sound sources with respect to source azimuth and distance. Elevation perception with the VAH under test was less accurate. The elevation results could be different for other VAH topologies and including elevated source directions in the calculation of spectral weights. The GUI may have introduced some difficulty to map the elevation responses, however, it could be used properly for gathering azimuth and distance responses. The results show that also in the absence of visual cues it is possible to have a similar localization performance with the VAH signals as with real sources.

Conclusion

The localization performance when listening to virtual sources generated with a Virtual Artificial Head (VAH)

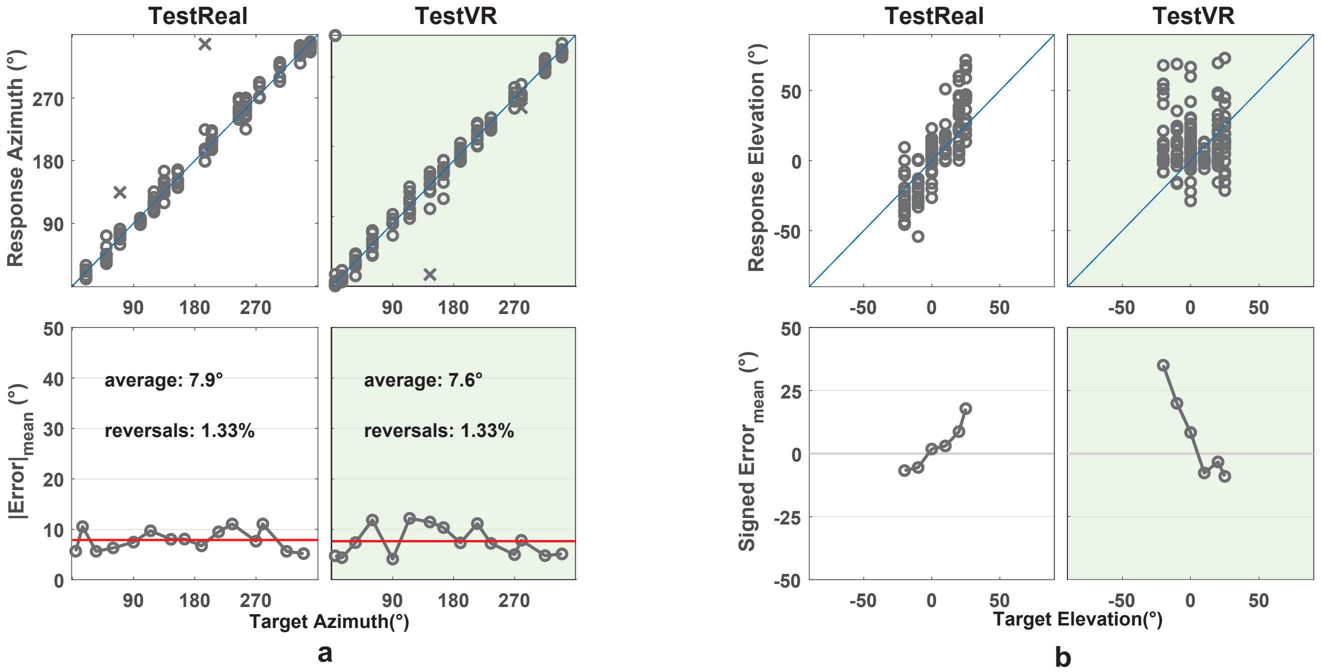


Figure 3: (a) Top: Response vs. target azimuths when listening to real sources (TestReal) or to virtual sources generated with the VAH (TestVR). Responses shown with a '×' were suspected to be front-back reversals. Below: Absolute azimuth error averaged over ten subjects. The error averaged over all target angles is shown with the horizontal line. (b) Top: Response vs. target elevations for TestReal and TestVR. Below: Signed elevation error (response − target) averaged over ten subjects.

was compared to localization with real sound sources, both in the absence of visual cues. Results indicated that also in the absence of visual cues, it is possible to have similar localization performance with VAH signals compared to real sound sources with respect to source azimuth and distance. The comparison of responses given to real and virtual sources using the same GUI offered a suitable method to evaluate the VAH performance with respect to source localization. This method can be used to further investigate the localization performance with the VAH signals in other environments and with other test signals.

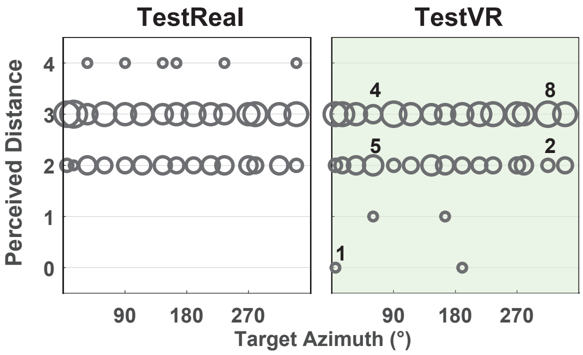


Figure 4: Perceived source distance vs. target source azimuth when listening to real sound sources (TestReal) as well as to virtual sound sources (TestVR) on a scale between 0 and 4 (refer to the text for more details). Area of each circle and the numbers shown indicate how many subjects chose each distance range.

Acknowledgement

This work was funded by Bundesministerium für Bildung und Forschung under grant no. 03FH021IX5.

References

[1] Rasumow, E., Blau, M., Doclo, S., van de Par, S., Hansen, M., Püschel, D., Mellert, V. Perceptual evaluation of individualized binaural reproduction using a virtual artificial head. *J. Audio Eng. Soc.*, 65(6), pp. 448-459, 2017.

[2] Fallahi, M., Hansen, M., Doclo, S., van de Par, S., Püschel, D., Blau, M. Individualized dynamic binaural auralization of classroom acoustics using a virtual artificial head. *ICA 2019, Aachen, Germany.*

[3] Iyer, N., Thompson, E. R., Simpson, B. D. Response techniques and auditory localization accuracy. *The 22nd International Conference on Auditory Display, Canberra, Australia*, 2-8 July, 2016.

[4] Rasumow, E., Hansen, M., van de Par, S., Püschel, D., Mellert, V., Doclo, S., Blau, M. Regularization approaches for synthesizing HRTF directivity patterns. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), pp. 215-225, 2016.

[5] Fallahi, M., Hansen, M., van de Par, S., Doclo, S., Püschel, D., Blau, M. Localization performance for binaural signals generated with a virtual artificial head in the absence of visual cues. Submitted to *Forum Acusticum 2020*