LOCALIZATION PERFORMANCE FOR BINAURAL SIGNALS GENERATED WITH A VIRTUAL ARTIFICIAL HEAD IN THE ABSENCE OF VISUAL CUES

Mina Fallahi[1] **Martin Hansen**[1] **Steven van de Par**[2,4]
Simon Doclo[2,4] **Dirk Püschel**[3] **Matthias Blau**[1,4]
[1] Institut für Hörtechnik und Audiologie, Jade Hochschule Oldenburg, Germany
[2] Dept. Medical Physics and Acoustics, University of Oldenburg, Germany
[3] Akustik Technologie, Göttingen, Germany
[4] Cluster of Excellence Hearing4all, Germany
mina.fallahi@jade-hs.de, matthias.blau@jade-hs.de

ABSTRACT

A Virtual Artificial Head (VAH) is an alternative to conventional artificial heads for creating binaural renderings of spatial sound fields. In contrast to conventional artificial heads, the scene captured with the VAH can be presented in an individualized head-tracked manner, by applying individually calculated spectral weights for different head orientations of the listener to the microphone signals. In this study, the localization performance when listening to virtual sound sources generated with the VAH was assessed in a listening test. In contrast to a previous study of the authors, reporting perceptually convincing localization ratings with the VAH [1], in this study, the localization test was performed without supplying the listener with any optical information about the sound source. The results indicate that the VAH performs well with respect to source azimuth and distance also without visual cues. Regarding possible optimization strategies, it appears to be beneficial to include only horizontal source directions in the calculation of spectral weights for the scenario considered here (anechoic environment and sources within $\pm 20^\circ$ from the horizontal plane).

1. INTRODUCTION

As an alternative to conventional artificial heads, a microphone array based filter-and-sum beamformer, referred to as Virtual Artificial Head (VAH), can be used to synthesize the directivity patterns of individual Head Related Transfer Functions (HRTFs), or their equivalence in the time domain, referred to as Head Related Impulse Responses (HRIRs) [2]. One advantage of the VAH is that the captured scene can be individualized post-hoc, by including the HRTFs of individual listeners. This is done by applying individually calculated spectral weights to the microphone signals. Another advantage of the VAH is that the spectral weights can be applied for different head orientations of the listeners, such that an individualized head-tracked binaural rendering is possible, without the need to rotate the VAH during the recording.

The VAH was recently evaluated as perceptually convincing for speech stimuli in a head-tracked scenario with respect to different perceptual attributes, including the source position [1]. The evaluation was based on direct comparison of the binaural signals generated with the VAH, presented over headphones to the signals played back from real sound sources in the same room. In this settings, listeners were able to see the sound sources. Besides the visual cues, also the use of head-tracking as done in [1] can promote the localization accuracy as well as the externalization [3]- [4], therefore, it was not clear to which extent the visual information of the sound source might have improved the ratings. Especially with respect to sound source localization, it is known that the presence of visual cues can influence the perception of the sound source position [5]- [6]. In order to assess the localization accuracy with VAH signals, this study performed a new localization experiment. The applied methods for generating the VAH binaural signals were almost the same as in [1], however, in the new study subjects were asked to localize the purely virtual sound sources while listening to head-tracked binaural signals generated with the VAH, without being supplied with any optical information about the sources. Subjects gave their responses by mapping the perceived source position onto a Graphical User Interface (GUI). Since the technique used for asking the perceived source position is also decisive for the localization accuracy [6]- [7], the localization test was also performed with hidden real sound sources in a separate session. The results of the test with hidden real sound sources served as a reference for evaluating the localization performance with the VAH binaural signals. This approach is considered appropriate here since the focus of the study was to investigate the VAH performance rather than to study the localization ability of the subjects.

After reviewing the chosen implementations for the VAH, the measurement setup for localization tests with real and virtual sound sources is described, followed by the discussion of the perceptual results.

2. VIRTUAL ARTIFICIAL HEAD (VAH) - THEORY AND CHOSEN PARAMETERS

2.1 Calculation of spectral weights for the VAH

The VAH aims for synthesizing the desired directivity pattern $D(f, \Theta_k)$ of the left or right HRTFs, with f denoting the frequency and $\Theta_k = (\theta_k, \phi_k)$, $k = 1, 2, ..., P$, the discrete directions with azimuth θ_k and elevation ϕ_k included in the directivity pattern. Considering the $N \times 1$ steering vector $\mathbf{d}(f, \Theta_k)$, defined as the free-field acoustic transfer functions between the source at direction Θ_k and the N microphones in the array, the resulting directivity pattern $H(f, \Theta_k)$ of the VAH is defined as

$$H(f, \Theta_k) = \mathbf{w}^H(f)\mathbf{d}(f, \Theta_k). \quad (1)$$

The complex-valued vector $\mathbf{w}(f)$ contains the spectral weights for the N microphones. These spectral weights were calculated by minimizing a narrow-band least-squares cost function, defined as

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

subject to carefully chosen constraints. One set of constraints was imposed to the resulting Spectral Distortion (SD) at each direction Θ_k, $k = 1, 2, .., P$, by setting an upper and a lower limit, L_{Up} and L_{Low}, such that for all k

$$L_{\mathrm{Low}} \leq \mathrm{SD}(f, \Theta_k) = 10 \lg \frac{|\mathbf{w}^{\mathrm{H}}(f)\mathbf{d}(f, \Theta_k)|^2}{|D(f, \Theta_k)|^2} \mathrm{dB} \leq \mathrm{L_{Up}}. \quad (3)$$

An additional constraint was applied to the *mean* White Noise Gain ($\mathrm{WNG_m}$), defined as the ratio between mean output power of the microphone array over all P directions and the output power for spatially uncorrelated white noise [8], i.e.

$$\mathrm{WNG_m} = 10 \lg(\frac{1}{P} \sum_{k=1}^{P} \frac{|\mathbf{w}^H(f)\mathbf{d}(f, \Theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}) \mathrm{dB} \geq \beta. \quad (4)$$

The constraint on $\mathrm{WNG_m}$ was set to guarantee the robustness of the VAH against microphone self-noise or deviations in microphone characteristics and positions.

The Interior-Point algorithm was used to solve this constrained optimization problem and the solutions proposed in [8] were considered as the initial values for this iterative algorithm.

2.2 VAH implementation and constraint parameters

In this study, two different microphone arrays, referred to as VAH 1 and VAH 2, as shown in Fig. 1, were used to synthesize the directivity patterns of individual HRTFs. VAH 1 was a planar 2-D microphone array of 24 microphones with an extension of 20 cm $\times$ 20 cm [2]. VAH 2 was a 3-D microphone array, consisting of 31 microphones with the 11 cm (W) $\times$ 11 cm (L) $\times$ 6 cm (H) extensions. For both VAHs, the microphones were distributed according to a Golomb ruler, such that all inter-microphone distances were as different as possible.

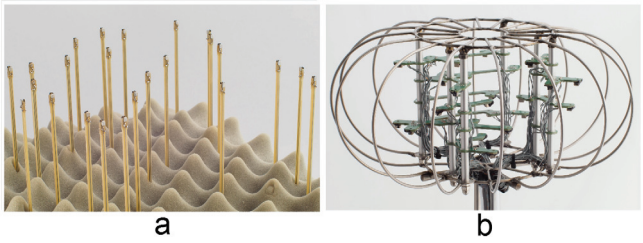

Figure 1. VAHs used in this study. (a) VAH 1: planar microphone array with 24 microphones. (b) VAH 2: 3-D microphone array with 31 microphones.

The ability of the two VAHs to meet the constraints as defined in Eqs. 3 and 4 depends on the values chosen for parameters L_{Up}, L_{Low} and β. In this study, the two constraint parameters L_{Low} and L_{Up} were chosen as -1.5 dB and 0.5 dB, respectively, as in [1]. This would lead to a maximum deviation of 2 dB in the resulting Interaural Level Differences at all P directions. For the minimum resulting $\mathrm{WNG_m}$, β was chosen as 0 dB, leaning on the results in [1]. Although the chosen value for β was evaluated only for VAH 1 in [1], the same value was also chosen for VAH 2 in the present study.

For a given set of parameters L_{Low}, L_{Up} and β, the synthesis accuracy with respect to the resulting Spectral Distortion and $\mathrm{WNG_m}$ depends on the number P of the discrete directions included in the calculation of the spectral weights. The synthesis is most accurate at the P discrete directions and deteriorates at other source directions. If P is increased, the constraint set for the resulting Spectral Distortion in Eq. 3, must be satisfied for a higher number of the synthesis directions. As a consequence, for some frequencies the resulting Spectral Distortion might not be kept in the desired range between $-$ 1.5 dB and 0.5 dB. On the other hand, a higher P also means that the accuracy would be distributed over all included P directions, meaning that more directions would benefit. An increased number P can also impact the satisfaction of the constraint in Eq. 4 such that the resulting $\mathrm{WNG_m}$ can be less than the desired minimum value β = 0 dB, leading to a reduced robustness. In this study, two cases for P were considered. P was either equal to 72 directions, all equally spaced in the horizontal plane, or P was equal to 3$\times$72=216, including source directions from the horizontal plane as well as the two elevations $\pm$15°. At each elevation, the sources were distributed equidistantly with 5° resolution. Spectral weights calculated for P = 72 horizontal directions are labeled with **El0** whereas the ones calculated for P = 216 directions (Elevations 0° and $\pm$15°) are labeled with **El0$\pm$15** in the remaining text.

As an important advantage of the VAH, one can rotate the VAH virtually such that the resulting spectral weights correspond to a new head orientation of the listener. For a given head orientation $\Theta_h = (\theta_h, \phi_h)$, spectral weights can be calculated by taking the $D(f, \Theta_k)$, $k = 1, 2, ..., P$, and the shifted steering vectors $\mathbf{d}(f, \Theta_s)$ with $\Theta_s = (\theta_k + \theta_h, \phi_k + \phi_h)$ into Eqs. 1 to 4. In this study, for both VAH 1

and VAH 2, the spectral weights labeled as El0 and El0$\pm$15 were calculated for 37$\times$5=185 head orientations to the 37 azimuth angles θ_h of $-90°$ to $+90°$ in 5° steps and the 5 elevations ϕ_h of $-15°$ to $+15°$ in 7.5° steps.

3. METHODS

The localization experiment in this study consisted of two listening tests. The first one, referred to as **TestVR** was performed to assess the localization performance with binaural signals generated with the VAHs and played back over headphones, without supplying the subjects with any visual cues on the source position. During TestVR, the virtual sound source was presented at different positions dynamically (i.e. with head tracking) over headphones. Subjects sat in a darkened anechoic room with very limited optical information about their surroundings and had to give the perceived position of the virtual source (azimuth, elevation, and distance) using a Graphical User Interface (GUI). In order to verify the appropriateness of the used response technique, the second localization test was performed with (hidden) real sound sources in the same darkened room, which is referred to as **TestReal**. Both listening tests as well as the measurements required for preparing the binaural signals with the VAHs were performed in the anechoic chamber of the Institut für Hörtechnik und Audiologie at Jade University of Applied Sciences in Oldenburg (3.1m $\times$ 3.4m $\times$ 2m, $f_{\text{cutoff}} = 200$ Hz). For each test, a different set of 15 source positions was considered as target positions. The 15 azimuthal source directions for each test were chosen randomly at multiples of 5° such that the whole range of lateral angles between 0° and 355° could be represented. Six different elevations (0°, $\pm$10°, $\pm$20° and 25°) were assigned to the 15 target source positions such that a balance between the number of positive, negative and zero elevations in the front-back hemispheres could be maintained (see Fig. 2).

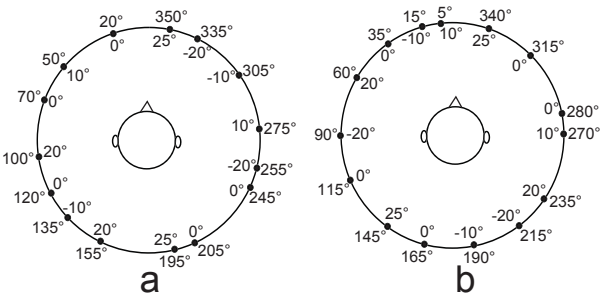

Figure 2. Target source positions when localizing with (a) real sound sources (TestReal) and (b) virtual sources (TestVR). Numbers outside the circle indicate the azimuth angle and the ones inside the circle indicate the elevation angle of the target source.

3.1 Target source positions in the room

A loudspeaker arc of 1.2 m radius was used to position the loudspeaker source that subjects needed to localize. The loudspeaker arc hung vertically from a turn table installed in the ceiling of the anechoic chamber. The turn table and the arc were mounted such that the center of the arc was at a height of 1.24 m and in the middle of the room with respect to X and Y axes. This position was defined as the Listener Position. Ten loudspeakers (SPEEDLINK XILU SL-8900-GY) were mounted in the arc at elevations between $-20°$ and 25° with 5° space between them, out of which the four loudspeakers at elevations $\pm$5° and $\pm$15° were not used in this study. Using the turn table, the loudspeaker arc could be rotated to any azimuthal source position. This setup was used to represent the real target sound sources for the TestReal as well as to generate the binaural signals with the VAHs for the TestVR (see section 3.5)

3.2 Subjects and test signal

A total of ten normal-hearing subjects with individually measured HRTFs and Headphone Transfer Functions (HPTFs) took part in TestReal and TestVR. Five subjects performed TestReal first whereas the other five started with the TestVR. For each subject, there was at least one week time between the two tests.

The test signal was a dry recorded speech utterance of 15 s duration, spoken by a female speaker. For TestReal, the test signal was played back from the real sound sources. For TestVR, the test signal was filtered with the synthesized HRIRs, as described in section 3.5, and was presented over headphones.

3.3 Graphical User Interface (GUI)

The localization task consisted of providing information about the perceived azimuth, elevation and distance of the real or virtual target sound sources. The GUI shown in Fig. 3 was used to conduct the tests and to gather the responses. This GUI was presented over the monitor display of a tablet computer, which was positioned in front of the subjects. For collecting the azimuth responses, the GUI showed the head seen from the top, with a circle around it. To enter the perceived source azimuth, subjects had to click on any point on this circle. An equivalent depiction of the head seen from the side was presented for giving the responses on the perceived source elevation. The reference position of azimuth and elevation = 0° in the darkened room, corresponding to the frontal head orientation, was marked with a white point at the top of the monitor display in the room as well as with a colored point on the GUI.

To provide information about the perceived source distance, subjects had to be supplied with a reference point. Due to the lack of optical information in the experiment design, the reference distance was presented with a reference sound source (the same loudspeaker type as mounted

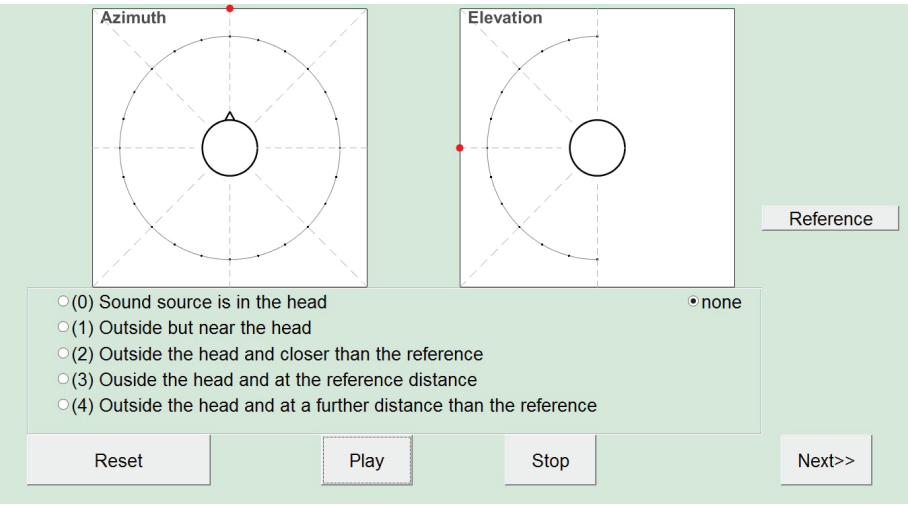

Figure 3. Graphical User Interface for acquiring the responses on the source azimuth, elevation and distance. Upon clicking the Reference button, subjects could switch to the signal coming from the reference source in room, to give their percept of the source distance. The "Reset" button was used to reset the head tracker during TestVR.

in the arc). This reference source was located at a fixed position in the room, nearly at azimuth and elevation = 0° and at 1.7 m distance to the Listener Position. By clicking the "Reference" or "Play" buttons on the GUI, subjects could switch the signal to come from the reference source or from the (real or virtual) target source. Subjects had to judge the perceived distance compared to the 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 was first positioned at the same distance as the target sources and the reference source and the target sources were then adjusted to have the same level (55 dBSPL) at the Listener Position. Then, the reference source was displaced back by 50 cm.

Subjects had no information on the exact position of the reference source and were instructed not to consider this source as a reference for azimuth and elevation, but only for the perceived distance. In addition, for TestVR, subjects were instructed to take off the headphones while listening to the reference source.

The "Reset" button shown in GUI in Fig. 3 was used during TestVR to reset the head tracker. For TestReal, this button was omitted from the GUI.

3.4 Localization with real target sound sources (TestReal)

During TestReal, subjects sat in the room, with their interaural center at the Listener Position. To eliminate any visual information on the source positions, subjects were seated inside an acoustically transparent curtain (see Fig. 4 a,b) and the room was darkened. The only light source was the monitor in front of the subjects, which they used to conduct the test and to give their answers using the GUI. The loudspeaker arc was rotated to one of the 15 azimuthal target positions around the subject (Fig. 2a) and the test signal was presented from the loudspeaker channel corresponding to the source elevation. Subjects were informed that they could rotate their heads in the allowable range of ±90° horizontally and ±15° vertically, and were asked, not the leave this range even if they perceive the sound sources behind them. Each of the 15 target source positions was presented once in a randomized order. Prior to that, five of them were chosen randomly to be presented at the beginning of the TestReal as familiarization trials without feedback. The responses given to these five target sources were discarded from the evaluations.

3.5 Localization with virtual target sound sources (TestVR)

To generate the binaural signals for the TestVR, VAH1 and VAH2 were positioned at the Listener Position in the anechoic room. Impulse Responses (IRs) were measured for each of the 24 microphones of VAH 1 and 31 microphones of VAH 2, and for each of the target source positions shown in Fig. 2b. In order to keep the environmental conditions comparable to TestReal, the IR measurements with the VAHs was performed with the VAHs positioned inside the same acoustically transparent curtain as in TestReal and in the presence of the monitor display (see Fig. 4 c). The corresponding individually calculated spectral weights for 185 head orientations and the two constraint cases El0 and El0±15, as introduced in section 2.2, were applied to the measured IRs with VAH 1 and VAH 2. This resulted in a total of four sets of individually synthesized Binaural Room Impulse Responses (BRIRs) with VAH 1 and VAH 2 for each of the 15 target source positions.

It was interesting to have a comparison between the binaural signals generated with the VAHs and the binaural sig-

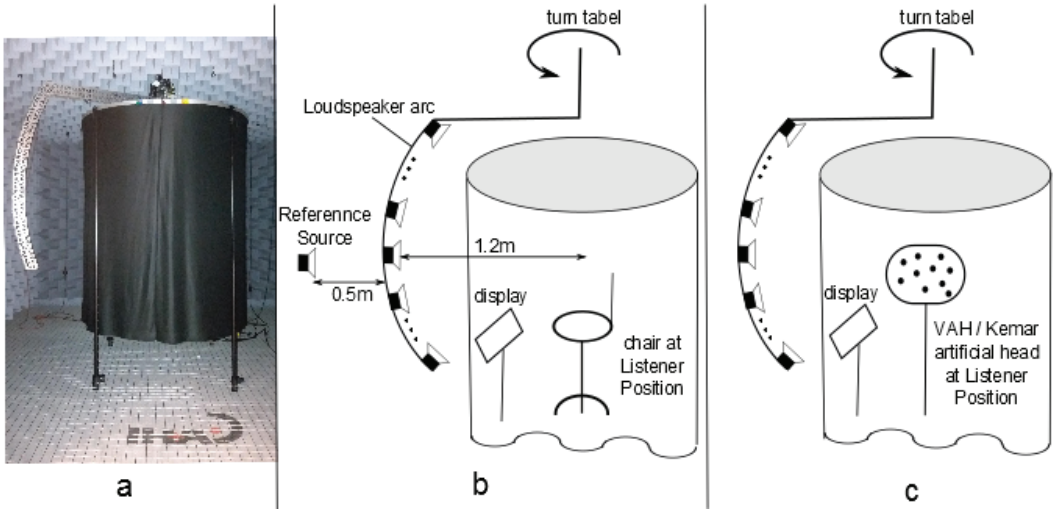

Figure 4. (a) Acoustically transparent curtain and the loudspeaker arc in the anechoic chamber of Institut für Hörtechnik und Audiologie at Jade University of Applied Sciences. (b) Experiment setup during TestReal and TestVR (Loudspeaker arc was used during TestReal to represent the target sources. During TestVR, virtual sources were presented via headphone). (c) Measurement setup for capturing the impulse responses with the VAHs or with the Kemar artificial head inside the curtain.

nals generated with a traditional artificial head, as it was also investigated in our previous study [1]. Therefore, the IR measurement for the 15 target source positions shown in Fig. 2b was also performed for the two ears of the Kemar artificial head (GRAS KEMAR type 45BB). However, binaural signals recorded with a traditional artificial head can be presented only for one fixed orientation of the artificial head. In order to nevertheless enable a dynamic binaural presentation with the signals generated with the Kemar artificial head, the IR measurement for each target position had to be repeated 37 times for the 37 orientations of the Kemar artificial head to the azimuthal orientations $-90°$ to $90°$ in $5°$ steps. This resulted in a total of $37\times15=555$ measurements. Note that for traditional artificial heads this is a nonrealistic use case with an enormous work load which cannot be applied to their standard application, e.g. for recordings in a concert hall. However, this nonrealistic use case of the Kemar artificial head was still used because otherwise the Kemar artificial head would clearly lose out against VAH 1 and VAH 2 during the localization task.

In order to generate the binaural signals for the TestVR, the test signal was filtered with the four synthesized BRIRs sets derived from VAH 1 and VAH 2, as well as from Kemar artificial head, and was subsequently filtered with individually measured inverse HPTFs. The binaural signals generated with the VAHs are referred to as **VAH1 El0**, **VAH1 El0±15**, **VAH2 El0** and **VAH2 El0±15**. The binaural signals generated from different head orientations of the Kemar artificial head, are referred to as **HTK** (Head-Tracked Kemar).

During TestVR, subjects sat with their interaural center at the Listener Position inside the acoustically transparent curtain, wearing the headphones (Sennheiser HD 800) with a custom made tracker mounted on top of it. They were instructed to reset the tracker before starting to listen to the virtual sound source by keeping their heads oriented to the marked reference position on the top of the display monitor and pressing the "Reset" button on the GUI. They were encouraged to make use of the possibility to rotate their heads within the allowable range ($\pm90°$ horizontally and $\pm15°$ vertically).

Each of the 15 target source positions was presented five times, i.e. once with each of the five BRIRs (either generated with the VAHs or with HTK). This resulted in a total of 75 virtual sources, which were presented in a randomized order. Prior to that and as in TestReal, five of them were chosen randomly to be presented at the beginning as familiarization and were discarded from the evaluations.

4. RESULTS

Azimuth: Response azimuths of ten subjects vs. target azimuths presented in TestReal and TestVR are shown in the upper row of Fig. 5. Each marker represents the response of each of the ten subjects. In the lower row of Fig. 5, the absolute error between target and response azimuth angles, averaged over the ten subjects, is shown. If a pair of target and response azimuths were at the two different sides of the interaural axis, a front-back reversal occurance was suspected and the response was therefore excluded from the calculation of absolute errors. These cases are marked with a '$\times$' in the upper row of Fig. 5. Target and response pairs within $\pm7.5°$ off the interaural axis were not checked for being reversals and were considered normally in the error calculation. The horizontal line in the lower row of Fig. 5 represents the average absolute localization error over all presented target azimuth angles and the number of cases suspected

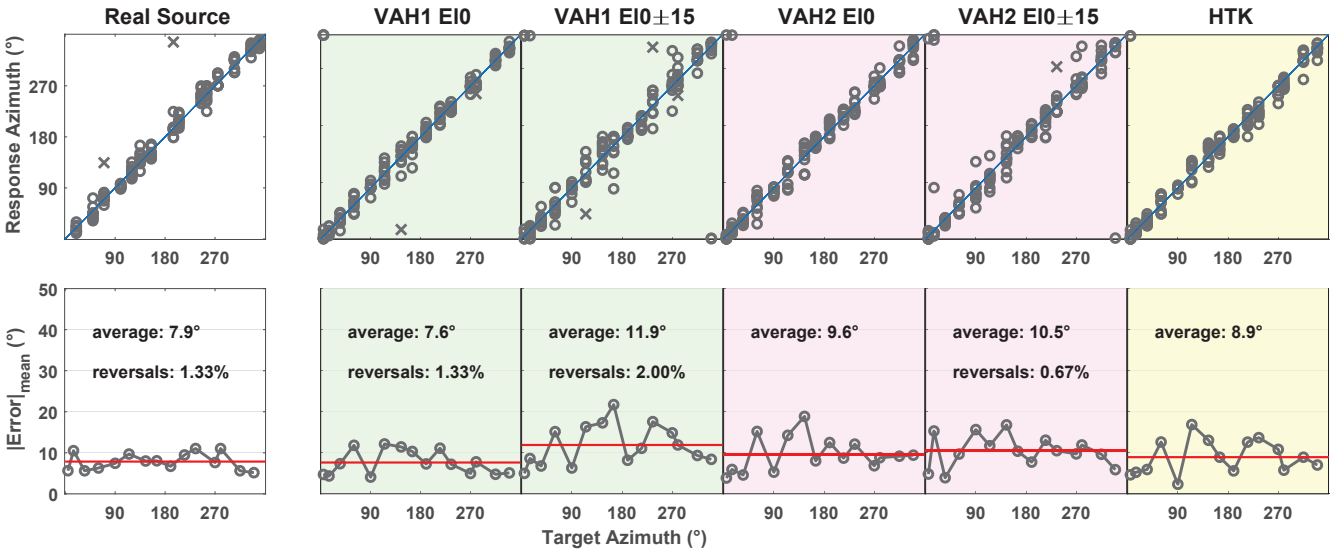

Figure 5. Top: Response azimuth (ordinate) vs. target azimuth (abscissa) when listening to real sound sources (TestReal) as well as to virtual sound sources (TestVR). Each circle represents the response of each of the ten subjects. Responses classified as reversals are marked with a '×'. Below: Absolute error between target and response source azimuth, averaged over ten subjects. The averaged absolute error over all target angles is shown with the horizontal line and listed as "average". The number in % gives the percentage of responses classified as front-back reversals (excluded from the error calculation).

as reversals is given as a percentage value.

For TestReal, the average absolute azimuth error over all target directions was 7.9° and 1.3% of the responses were classified as reversals. These values represent the ability of the subjects to localize real sound sources in the absence of optical cues, when mapping their responses onto the GUI.

For TestVR, the VAH1 El0 case showed a similar azimuth error (7.6°) and reversals rate (1.3%) compared to TestReal. For VAH2 El0 and HTK, the azimuth error was slightly higher, with no reversals. The largest azimuth error was observed for both VAH cases labeled with El0±15. Including more directions in the calculation of spectral weights for the VAH reduced the synthesis accuracy for these cases, as already discussed in section 2.2. This led to higher azimuth errors.

Elevation: Response vs. target elevations of ten subjects are shown in the upper row of Fig. 6. The difference between response and target angles was in general higher for elevation than for the azimuth angles. This is in accordance with the fact that the auditory spatial resolution is smaller in the vertical direction than in the horizontal direction [9].

For TestReal, subjects' responses to the presented target elevations of between −20° and 25° extended from −54° to 70°. In general, subjects tended to underestimate negative elevations and to overestimate positive elevations, which might have been caused by the difficulty of mapping the responses onto the GUI. This can also be seen in the signed elevation error (subtracting target elevations from response elevations), averaged over ten subjects, as shown in the lower row of Fig. 6.

For TestVR, the responses were different from TestReal. The positive signed error for the target elevations −20°, −10° and 0°, as shown in the lower row of Fig. 6, indicates that these target elevations were often perceived at a higher elevation. In contrast, the positive target elevations 10°, 20° and 25° were often perceived at a lower elevation, leading to negative signed errors. For binaural signals VAH2 El0±15 and HTK, the accuracy was comparable to the results of TestReal. For VAH2 El0±15, the smaller errors compared to the other VAH cases was due to the inclusion of the elevations ±15° in the calculation of the spectral weights such that the synthesis accuracy at non-horizontal directions was higher compared to VAH2 El0, VAH1 El0 and VAH1 0±15. In contrast to VAH 2, including the elevations ±15° was not advantageous for VAH 1 with respect to elevation errors. One explanation could be the 3-D topology of VAH 2 compared to the 2-D topology of VAH 1. With microphones distributed at different elevations, the variation of signals coming from different elevated directions can be captured better with VAH 2 compared to the planar microphone array of VAH 1. Another reason can be the higher number of microphones in VAH 2 compared to VAH 1, which supported the satisfaction of the constraints. In general, the elevation responses to the binaural signals in TestVR were less accurate than the responses in TestReal.

Distance: The results for the given source distance of the real or virtual sources are shown as scatter diagrams vs. target source azimuth in Fig. 7. The area of each circle indicates how many subjects chose each distance percept.

As described in section 3.3, the real or virtual target sources were expected to be perceived closer than the reference source due to the 50 cm distance between them. For TestReal however, the majority of the subjects chose

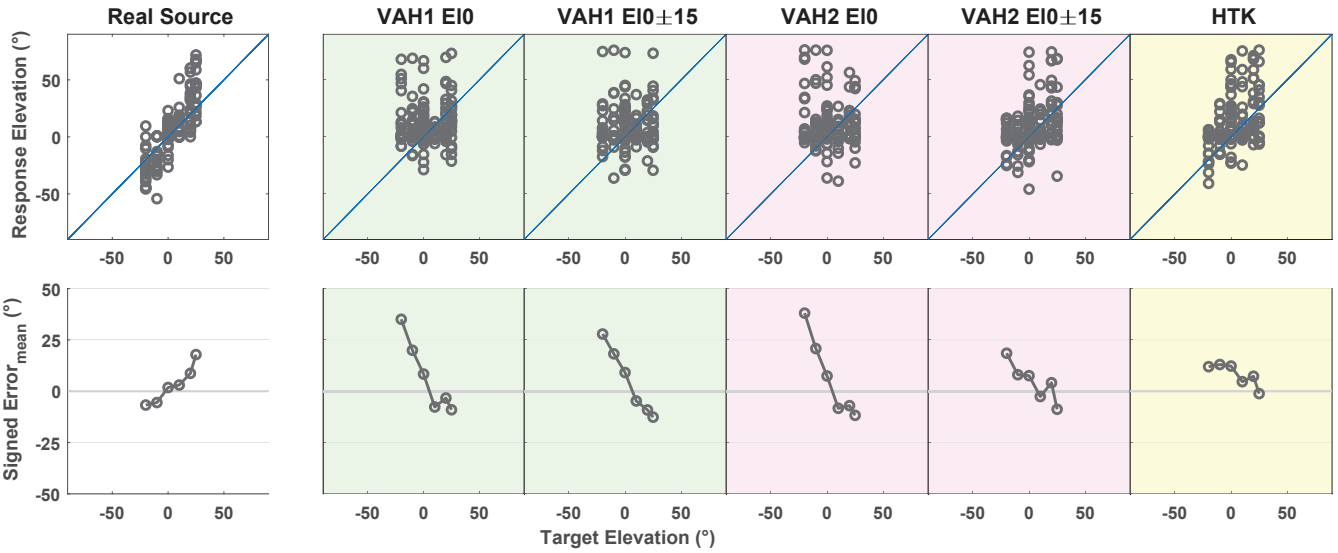

Figure 6. Top: Response elevation (ordinate) vs. target elevation (abscissa) when listening to real sound sources (TestReal) as well as to virtual sound sources (TestVR). Each point represents the response of each of the ten subjects. Below: Over ten subjects averaged signed elevation error (response − target)

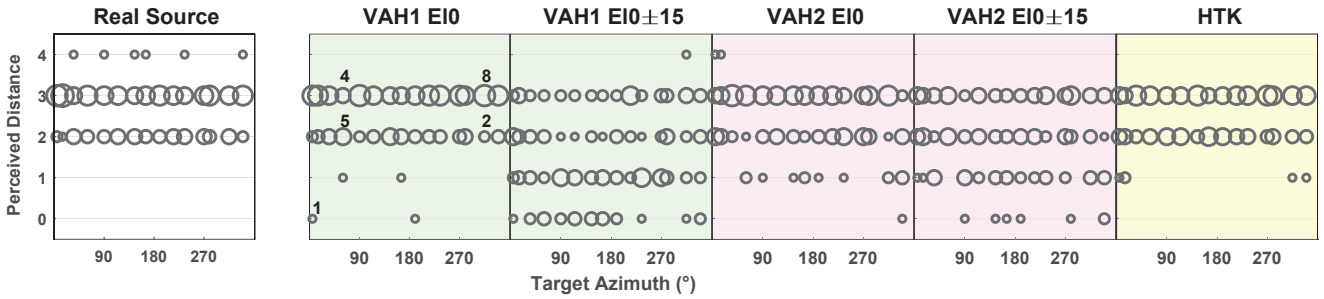

Figure 7. Perceived source distance (ordinate) vs. target azimuth (abscissa) when listening to real sound sources (TestReal) as well as to virtual sound sources (TestVR) on a scale between 0 and 4, corresponding to (0): in head, (1): outside the head but hear the head, (2): outside the head and closer than the reference, (3): outside the head and at the reference distance, (4): outside the head and at a further distance than the reference. Area of each circle with the numbers shown exemplarily for VAH 1 El0 indicate how many subjects chose each distance range.

the score (3), which means that the real target sources were perceived at the same distance as the reference source. It seems that the level difference between reference and target sources due to their different distances to Listener Position was not enough for some subjects to perceive the target source closer than the reference. None of the subjects perceived the real target sources near or in the head, which is an expected result.

In comparison to real target sources, for the two binaural signals VAH1 El0±15 and VAH2 El0±15, the virtual sound sources were frequently perceived in and near the head (48% and 26.3% of the total given responses, respectively). The synthesis error, due to the higher number of included source directions, led to virtual sources with insufficient externalization for these VAH cases. For VAH1 El0, VAH2 El0 and HTK, 2.6%, 9.3% and 3.3% of the total given responses were in and near the head, respectively; however, the majority of the subjects perceived the distance of the virtual target sources in TestVR comparable to the real target sources in TestReal.

5. DISCUSSION

As the results show, azimuth and distance perception with the VAH signals were similar to real sound sources, whereas elevation perception with the VAHs was often inaccurate. In our previous study [1], in which the localization ratings were high, not only the sources could be seen, but the experiment task was also different. It consisted in rating the match between the position of the real and virtual sources in general, rather than giving responses separately for the absolute azimuth and elevation of the virtual sources. Nevertheless, the results in the present study showed that even in the absence of optical cues and within

a more challenging localization task, it is possible to generate virtual sound sources with the VAH, with comparable localization performances in azimuth and distance as with real sound sources. VAH cases, in which only horizontal source positions were included (VAH1 El0 and VAH2 El0), performed better than the cases were non-horizontal directions were considered.

According to the results, head-tracked presentations generated with measured non-individual HRTFs of Kemar artificial head (HTK) led to similar azimuth and distance responses as with VAH1 El0 and VAH2 El0. For elevation, HTK was better than most of the VAH cases, however, still comparable to the VAH2 El0±15. This confirms our previous findings [1], [10] that, for speech signals and with a head tracked signal presentation, individual HRTFs may not constitute an important advantage over generic (dummy head) HRTFs. It should be stated again, that the head-tracked presentation of the HTK signals in this study was an unusual use case of the traditional artificial heads. The rate of front-back reversals as well as in and near head localizations may be higher for signals generated with traditional artificial heads, if the impractical effort to represent such signals dynamically, as done in this study, is not taken. The results confirm the advantage of the Virtual Artificial Head as a suitable substitute for the conventional artificial heads.

6. SUMMARY AND CONCLUSION

The aim of this study was to assess the localization accuracy for virtual sound sources generated with the Virtual Artificial Head (VAH), in the absence of visual cues. Two different VAHs were used to capture free-field impulse responses for different source positions. Individually calculated spectral weights for 185 head orientations were applied to these impulse responses and then convolved with the test signal (speech) to represent the virtual sound sources in a head-tracked scenario. No visual information about the sources was supplied to the listeners. The responses to the azimuth, elevation and distance of the virtual sources were mapped onto a Graphical User Interface (GUI). The responses given to real hidden sources using the same GUI gave a reference against which the localization accuracy with virtual sources could be evaluated. The results showed that even in the absence of visual cues and in anechoic conditions, it is possible to have similar localization accuracy with the VAHs in the azimuth and distance as with real sources. For elevation, the VAH technology needs to be improved, either by modifying the array topology or by modifying the constraints defined for the calculation of the VAH spectral weights. Nevertheless, the ability of presenting head-tracked signals, as applied in this study, is a great advantage of the VAH technology which improves the localization accuracy. Further investigations should concern the performance of the VAH with other signals such as music or broadband noise and including other perceptual attributes such as spectral coloration in the evaluation.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] 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. *Proc. of the 23rd International Congress on Acoustics, Aachen, Germany*, 9−13 September, 2019.

[2] 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.

[3] Begault, D. R., Wenzel, E. M., Anderson, M.R. Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, 49(10), pp. 904-916, 2001.

[4] Brimijoin, W. O., Boyd, A. W., Akeroyd, M. A. The contribution of head movement to the externalization and internalization of sounds. *PLOS ONE*, 8(12), 2013.

[5] Kytö, M., Kusumoto, K., Oittinen, P. The ventriloquist effect in augmented reality. *IEEE International Symposium on Mixed and Augmented Reality*, pp. 49-53, 2015.

[6] Tabry, V., Zatorre, R. J., Voss, P. The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in Psychology*, Vol.4, Article 932, 2013.

[7] 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.

[8] 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.

[9] Blauert, J. *Spatial hearing: the psychophysics of human sound localization*, Revised ed. Cambridge, Massachusetts: MIT Press, 1997.

[10] Blau, M., Budnik, A., van de Par, S. Assessment of perceptual attributes of classroom acoustics. *Proc. of Institut of Acoustics*, Vol.40.Pt.3.2018.