Head-Tracker Based Auralization Systems: Additional Consideration of Vertical Head Movements

Philip Mackensen (1), Markus Fruhmann (1), Mathias Thanner (1), Günther Theile (1), Ulrich Horbach (2) and Attila Karamustafaoglu (2)
(1) Institut für Rundfunktechnik, München, Germany
(2) Studer Professional Audio AG, Regensdorf, Switzerland

Presented at the 108th Convention 2000 February 19-22 Paris, France

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd St., New York, New York 10165-2520, USA.

All rights reserved. Reproduction of this preprint, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.
Head-Tracker Based Auralization Systems: Additional Consideration of Vertical Head Movements

Philip Mackensen, Markus Fruhmann, Mathias Thanner, Günther Theile
Institut für Rundfunktechnik
Floriansmühlstr. 60, D-80939 München, Germany

Ulrich Horbach, Attila Karamustafaoglu
STUDER Professional Audio AG
Althardstr. 30, CH-8105 Regensdorf, Switzerland

Abstract
The headphone-based Binaural Room Scanning (BRS) auralization system relies on real-measured binaural room impulse response data and head-tracking, thereby offering high-quality listening for example in a virtual surround sound control room. Due to further developments of the BRS system it was suggested to be important that the head-tracker does support not only rotational movements of the head but also head movements in the vertical plane. Results from corresponding localization experiments are presented and discussed.

Introduction
For the acoustical display of surround sound material one can use either loudspeakers in a 5.1 arrangement or headphones. Though loudspeaker usage might be preferable it demands an acoustically treated (well-damped) room allowing for enough space to place the speakers correctly. This is a demand which often cannot be satisfied by small control rooms, e.g. OB vans. On the other hand the main disadvantage of using headphones is the in-head-localization of the acoustical events. Even by listening to recordings made with a stationary dummy-head the in-head-localization remains for certain directions (median plane) and is sometimes accompanied by front-back-reversals.

The headphone-based Binaural Room Scanning (BRS) procedure combines the advantages of both display techniques, i.e. listening to surround sound material by means of virtual loudspeakers using headphones without the unwanted artifacts formerly stated.
Auralization Methods

Model-based Auralization

Model-based auralization systems generate the aural room synthetically. Principally this synthesis consists of four steps. In the first step the sound field at the hot spot is recreated taking into account the acoustically characteristic properties of the room. Hence, for each sound source the direct sound, all the relevant reflections as well as the room-modes at the (virtual) listening spot have to be calculated. To model the properties of the human outer ear (pinna) is the second step. Here the binaural impulse response has to be determined by means of a database of stored HRTFs. The third step consists of the convolution of the source signals with the binaural impulse responses simulating the acoustical path between sound source and entrance to ear-channel (concha). Because for every direction of sound an individually corresponding HRTF has to be used, this step is rather process-power consuming. Depending on the number of reflections and sound sources taken into account, the geometry of the room, sound emission characteristics of the sources, etc. this processing-power can grow rapidly. Finally, in the forth step the calculated binaural signals are displayed through headphones.

To allow for head-tracking is inevitable for creating a realistic virtual environment, as has been shown in previous papers [1-4]. This implies, however, to repeat the steps aforementioned at least 60 times per second, thus for each orientation of the listener’s head a new set of binaural impulse responses has to be calculated [5,6]. As a consequence, it seems to be impossible by means of a model-based auralization system to render a real existing 5.1 studio control room with such an accuracy that absolutely no differences between the original and the virtual room can be heard – the required processing-power is not achievable at the moment. All simplifications, e.g. shorter impulse responses or less reflections, would diminish the impression to be actually in a real, acoustic environment.

Binaural Room Scanning

Already Thurlow et al. showed the relevance of head-movements in localization [7]. Recent experiments carried out at the Institut für Rundfunktechnik have confirmed this [1,4]. By using a dummy-head and allowing for head-movements, i.e. incorporating a head-tracking system, a localization comparable to normal hearing is possible. Artifacts like inversion of directions or in-head-localization often occurring in the case of the fixed dummy head completely vanish.
The main idea of the BRS auralization procedure is to capture the room impulse response *binaurally* by means of a rotatable dummy-head for every relevant orientation of the dummy-head. The diffuse-field [8] equalized dummy-head itself is placed in the sweet spot of a 3/2 surround sound loudspeaker setup. The MLS-based measurement is carried out for the horizontal plane with step-width of 6 degrees between -42° and 42°. Finally, for every loudspeaker a complete set of binaural (room) impulse response is stored in a database (fig. 1, [1,9]).

In contrast to the calculated, artificially synthesized room impulse responses of model-based auralization systems the BRS system uses *measured* impulse responses, containing all (linear) acoustical properties of the room and the sources (loudspeakers).

The impulse response and the sound field at the listening position don’t have to be calculated elaborately for the generation of the binaural signals, but a simple table-lookup procedure (dependent of the orientation of the listener’s head) establishes the binaural room impulse response used for the convolution with the input signals, which corresponds to the signals feeding the virtual loudspeakers. For each input signal the respective set of stored binaural impulse responses is used. Possibly an interpolation between two adjacent room impulse responses takes place beforehand (Figure 1, [1]). The result is an *acoustical clone* of the original, real existing listening room. Figure 2 depicts the complete functional outline of the BRS processor.

**First Practical Experiences**

During the 106th AES Convention in Munich a demonstration showed how realistic a data-based auralization by the BRS processor is. The listeners could compare a real 3/2 loudspeaker listening situation with the virtual situation, which was scanned in the same room previously. In the first run a voice was heard successively in all loudspeakers. In the second round the listeners had to put on headphones, and this time the voice was heard via the virtual loudspeakers (the real loudspeakers muted). The arising hearing event was such that the locations of virtual and real loudspeakers coincided. Moreover, some listeners even claimed to listen to hear the real loudspeakers in the second run until they put off the headphones.

Likewise positive was the feedback about the BRS processor experienced on a nine day lasting demonstration at the BBC where sound engineers and professionals had the chance to test the BRS processor. Each person had one hour to listen and to compare the listening situations, i.e. real room vs. virtual, acoustically cloned room, using their own listening material. None of the about 40 listeners experienced an in-head-localization or other unwanted artifacts. Much more, the acceptance to actually work with such a
system was very high - especially when working under acoustical unfavorable conditions [14].

Its first real application was a live broadcast of a classical concert at Westminster Cathedral. The unfavorable acoustic environment within an OB truck could be disregarded using the BRS processor.

Modification of the BRS Processor

As has been shown in previous listening tests the shape of the dummy-head and of the pinnae play a minor role in horizontal localization if head-tracking is enabled [1]. However, a few listeners reported some elevations, especially of the center speaker in the median plane. It is assumed that these unwanted elevations originate either from the difference in HRTFs between the listener and the dummy-head or from preventing further dynamic cues, e.g. pivoting [10-12].

The first attempt failed to cancel these elevations simply by an individual equalization of the binaural signals to compensate the deviations in the high frequency region, and so to eliminate the elevation effect [1]. Instead of diminishing the elevation the equalization process often evoked unnatural changes in timbre.

But instead of using individualized HRTFs for the scanning procedure, i.e. the exact replica of head and torso of the respective listener, it was suggested to implement a further degree of freedom of movement to provide more dynamic cues. The assumption that small, involuntary head movements are very essential for the up/down localization as well is supported by listening experiments from Wightman and Kistler [13]. Here localization of a static virtual source (no head movements) was impossible, whereas restricting to monaural hearing the localization was merely moderately degraded.

Head-Tracker Dependent Part of a Room Impulse Response

So far, the layout of the BRS processor allows for head rotations only. Long room impulse responses up to 80 ms could be stored and used for convolution to ensure for a high fidelity of the auralized room.

As shown by Thurlow either head rotations alone or in combination with pivoting are the most common used head movements in localization [7]. So it is suitable to allow for pivoting as a second head movement.

To consider pivoting brings about to repeat the scanning process for different vertical orientations of the dummy head. As a result a far greater amount
of room impulse responses for one room has to be stored and utilized dependent of the actual orientation of the listener's head. Still using impulse responses of 80ms length would increase the processing power needed beyond the available limit. Accordingly a simplification has to be made regarding the length of the impulse response which is dynamically, i.e. dependent on the head's orientation, convolved.

Only during a specific time of the whole impulse response ($t = 0 \text{ ms} \ldots t_{\text{dyn}}$) the orientation of the listener's head is considered and the respective room impulse response is chosen (Figure 3). After that ($t = t_{\text{dyn}} \ldots 85 \text{ ms}$) simply the impulse response of the 0 degree orientation is taken using a window of 1.3 ms length (i.e. 64 samples at 48 kHz sampling frequency) for the crossfading.

The time $t_{\text{dyn}}$ depends on the scanned room. As a useful upper boundary of control rooms to be auralized with the BRS processor the Bavaria Filmstudio Munich has been chosen because of its long reverberation time (worst case). In a listening test $t_{\text{dyn}}$ for a front channel (center) and for a surround channel (left surround) has been determined. The listener was offered female speech (EBU SQAM-CD, track 49) using two auralized control rooms: the “original” room, where head movements were considered for the whole impulse response length and a “shortened” room. In the latter only up to $t_{\text{dyn}}$ the orientation of the listener's head was accounted for. Several “shortened” rooms were used for comparison with the time $t_{\text{dyn}}$ varied between 7 ms and 43 ms at a step width of 4 ms. By using rotational head movements the subject had to state if it can hear a difference between them. Twelve subjects participated in this test.

Figure 4 and Figure 5 show the mean value (full line) and the 95% confidential interval (dotted line) for $t_{\text{dyn}}$ of the center or the surround channel respectively. Here the value 1 denotes a 100% detection of a difference between the two rooms, whereas 0 means no difference at all was perceived. Principally the same applies to Figure 6 but here no discrimination between front and surround channel was made.

As can be seen, up to $t_{\text{dyn}} = 15 \text{ ms}$ the differences are obvious, above 15 ms difficult to observe. The differences to be perceived at first are different size of the auralized room or a reduced distance of the sound source. With increasing $t_{\text{dyn}}$ an emphasis of the low frequencies or apparent displacement of the sources occur. Finally, if $t_{\text{dyn}}$ is further increased, even these characteristic features disappear making a distinction impossible. For both, the mean values of center channel and surround channel $t_{\text{dyn}}$ is found to be about 23 ms. To ensure that no difference can be heard between using the “full” length of the impulse response and the “reduced dynamic length” it is recommended to choose $t_{\text{dyn}}$ not less than 30 ms.
Though this investigation of \( t_{\text{dyn}} \) was made for a certain room, and these results cannot be transferred to different room with certainty, it is very improbable that artifacts arise using similar or even smaller rooms. However, the auralization of bigger rooms is already limited by the dimension of the memory. And even using a 100% dynamic calculation (\( t_{\text{dyn}} = 85 \, \text{ms} \)) an exact reproduction of the original room is impossible.

### Listening Experiments on Elevation

**Impact of Head Movements on Localization in the Median Plane**

To actually scan a room with different vertical orientations of the dummy head it was necessary to modify the mounting unit with the step motor. An additional joint was installed enabling to pivot the dummy head in defined vertical displacement angles in the median plane (Figure 7). The “new” BRS procedure was such that for eight different inclinations of the dummy head (\(-20^\circ \ldots 15^\circ\) using a stepwidth of \(5^\circ\)) the respective “horizontal” plane was scanned (\(-42^\circ \ldots 42^\circ\), every \(6^\circ\)) by means of the MLS technique. The inclination angles were chosen because it was assumed that the potential users of the BRS processor are mainly professionals working at mixing consoles. In that case the listener rather looks at the desktop (console) keeping the head inclined downwards.

The room used for the experiment had a volume of about \(65 \, \text{m}^3\), and it was equipped with wall-to-wall carpet, curtains and wooden inventory. Its reverberation time was small enough for a loudspeaker setup to be auralized using the BRS processor.

For the investigation of the localization performance in the median plane a different loudspeaker arrangement than the standard 3/2 setup (according to the ITU-R 775 recommendation) had to be taken. Nine loudspeaker positions were chosen which are listed in the following table [15]. Additionally two further loudspeakers were placed outside the median plane to check if there is an effect on both, vertical and horizontal localization.

In the experiment 21 listeners participated. Male speech (track 50 of the EBU SQAM-CD) was presented through one of the eleven virtual loudspeakers (using the headphone-based BRS system), and the subject’s task was to indicate its apparent horizontal and vertical position on a sheet. Full circles (with the head as the center) denoted the horizontal plane or the median plane, respectively. After completing a test run, a total of 20 source position were displayed subsequently, each twice for cross-checking, both with and without vertical head movements (pivoting).
### Experimental Results (Median Plane)

The results of the experiments allowing for horizontal head-tracking only are displayed in Figure 8. It comprises all elevations for every subject together in one plot. Except for an elevation angle of 90° the dispersion seems at first glance to be fairly big (about 60°). Only the vertical position is either perceived correctly or in the head (denoted by 0°). All in all, the vertical localization does not seem to be very good.

Astonishingly, incorporating a *vertical* head-tracking (additionally to the horizontal one) does not improve the localization performance drastically (especially, if one compares what impact horizontal head-tracking has on the localization performance in horizontal plane [1]). Accordingly, the dispersion in Figure 9 is not smaller than in Figure 8. Even the opposite is the case for the elevation angle of 90°.

To clarify how strong the influence of additional, *vertical* head-tracking is, the median values of both experimental outcomes are plotted in Figure 10. Compared to "pure" horizontal head-tracking a small improvement in the region between -30° and 60° is visible. But still, loudspeakers at 0°, 30° and 60°, respectively, seem to be perceived unnaturally raised. Figure 11 and Figure 12 display this situation in more detail.

Here the elevation of the four frontal sources is apparent. The impact of added vertical head-tracking is bigger for the two speakers at -30° and 0° than on the rest. The interquartiles even do not confirm this. They seem to be nearly constant for all positions, so that one cannot conclude from the additional vertical head-tracking to a better localization performance.

---

**Table: Loudspeaker Positions**

<table>
<thead>
<tr>
<th>Loudspeaker</th>
<th>Elevation</th>
<th>Azimuth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-30°</td>
<td>0°</td>
</tr>
<tr>
<td>2</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>3</td>
<td>30°</td>
<td>0°</td>
</tr>
<tr>
<td>4</td>
<td>60°</td>
<td>0°</td>
</tr>
<tr>
<td>5</td>
<td>90°</td>
<td>0°</td>
</tr>
<tr>
<td>6</td>
<td>120°</td>
<td>0°</td>
</tr>
<tr>
<td>7</td>
<td>150°</td>
<td>0°</td>
</tr>
<tr>
<td>8</td>
<td>180°</td>
<td>0°</td>
</tr>
<tr>
<td>9</td>
<td>210°</td>
<td>0°</td>
</tr>
<tr>
<td><strong>outside the median plane</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>45°</td>
<td>45°</td>
</tr>
<tr>
<td>11</td>
<td>135°</td>
<td>135°</td>
</tr>
</tbody>
</table>
For the sound sources in the rear of the head the median values indicate a good localization performance. Only the source at 210° is perceived slightly elevated. Here, the pivoting does improve the localization a little bit.

**Experimental Results (Off-Median Plane)**

Principally, the same applies to the two loudspeakers outside the median plane at 45° and 135° azimuthal angle, respectively. Again, allowing for additional, vertical head-tracking does not improve the vertical localization (Figure 13, Figure 14). For the frontal loudspeaker the dispersion without the vertical head-tracking is even much smaller than in the other case.

However, regarding the azimuthal localization a slight improvement can be seen in the case of the frontal loudspeaker when allowing for vertical head-tracking (Figure 15, Figure 16). Also the interquartil is smaller. Thus, allowing for vertical head-tracking seems to have an influence on horizontal localization in the frontal region.

**Influence of Additional Vertical Head-Tracking on Timbre**

Finally, the listeners were asked whether they determine (unnatural ?) variations in timbre brought on by using a different HRTF than the listener’s HRTF. Therefore, some music was played using the virtual speaker at 30° elevation angle and the subjects were asked to turn their heads and listen to the timbre. The listener had to decide between the three statements:

| 1. no changes in timbre perceived |
| 2. changes in timbre seem to be natural |
| 3. strong, unnatural variations are present |

The experimental result was that only 18% of the listeners heard strong, unnatural variations in timbre, whereas 36% identified the timbral changes as normal, and 46% did not recognized changes in the timbre at all.

**Impact of Head Movements on localization Using a 3/2-Surround Loudspeaker Arrangement with Different Elevations**

The results of the previous listening tests suggest that additional, vertical head-tracking does improve the localization only slightly. However, the horizontal localization of the off-median frontal source seems to be improved significantly.

In the following experiment a more practical listening situation is investigated than merely using loudspeakers in the median plane at various elevation angles. Therefore, the standard surround sound loudspeaker setup was chosen, whereas the center speaker and only the right frontal and the
right surround speaker were chosen. For symmetrical reasons the left side speakers were left out. This situation was scanned for auralization in the first part of the experiment. An (visually) opaque curtain around the listener covered the real loudspeakers to prevent optical cues.

At the center speaker position (0° azimuthal angle) three loudspeakers with different elevation angles (-10°, 0°, 10°) were placed. The same applied to the frontal right position at 30° (azimuth). Only two loudspeakers were used in the rear position (0° and 10° elevation angle). All loudspeakers had the distance of 2 meters to the listener.

Again the listeners were offered the same male speech example as was used in the median-experiment (SQAM-CD, track 50). In the first run they had to wear headphones to determine the position (especially, the elevation angle) of the virtual loudspeakers auralized by the BRS processor. Each virtual loudspeaker was displayed two times with and one time without processing the vertical head-tracking information. As used in the previous “median-experiment”, circles represented the horizontal and vertical plane, respectively, to indicate the position of the hearing event. In the second part of the experiment the elevation of the real loudspeakers had to be identified. Here the listeners did not wear headphones.

**Experimental Results (Elevated 3/2 Loudspeaker Arrangement)**

While conducting the experiment it was obvious that the listeners had some problems to determine (hear) the elevation of a hidden sound source. Additionally, the correct marking on the sheet of a relatively small elevation (±10°) of a hearing event was difficult, too (see Figure 17 - Figure 19). Often the virtual loudspeakers were indicated as much higher elevated as they actually were perceived.

Using the real loudspeakers, however, it was much easier for the listeners to determine the elevation. They decided and marked the location faster than in the case of the virtual speakers, but still, some of the loudspeakers were perceived slightly elevated (Figure 20 - Figure 22).
Summary

The Binaural Room Scanning (BRS) method was shortly reviewed stressing on the difference between model-based and data-based auralization systems, e.g. the BRS processor made by Studer Professional Audio AG.

Because unwanted elevations occurred using only a head-tracking system for the horizontal head movements, it was suggested to allow additionally for vertical movements, too. Therefore a modification of the dummy head’s mounting system had to be made in order to enable pivoting. The scanning procedure itself was such that for different inclinations of the dummy head the sound sources had to be scanned.

In a first experiment the localization performance in the median plane was investigated. Here, the height (elevation) of eleven differently elevated loudspeakers in the median plane was to be determined by the subjects while both, enabling and disabling the vertical head-tracking.

The results showed that additional vertical head-tracking just improved the localization only a little bit (if ever). Mostly, the frontal sources seem to be elevated, and this elevation persists, even when using the additional vertical head-tracking.

However, the azimuth of a frontal off-median loudspeaker, tested as well, was perceived more accurate when allowing for pivoting. This result led to a further listening experiment.

In this second listening experiment a standard surround sound loudspeaker setup was taken. For symmetrical reasons only the center and right speakers were used. The frontal loudspeaker were presented with three different elevations (0° and ±10°), the surround speaker only using two (0° and 10°).

Again, the virtual loudspeakers were perceived elevated, but even the real loudspeakers were located with some small elevation. A better vertical localization was not caused by using an additional vertical head-tracking.

However, there are still a number of uncertainties to be clarified before we can state that the vertical dynamic cues are unimportant. It is suggested to perform further experiments on this subject, in particular on the importance of (individual) exactness of HRTF’s in the median plane.
References


Figures

Binaural Room Scanning (BRS)
The Complete System

Figure 1  Scanning real room and rendering of the virtual room

Binaural Room Scanning (BRS)
Functional outline of the BRS-Processors

Figure 2  Functional outline of the BRS processor
Figure 3  Room impulse response of scanned room. The part between 0 ms and $t_{\text{dyn}}$ is used in dependence of the actual position of the listeners head. For later times only the room impulse response of the 0 degree orientation is used.

Figure 4  Average and 95% confidential interval for the center channel. The value 1 denotes a 100% recognition of a difference, whereas 0 means no difference at all was perceived.
Figure 5  Average and 95% confidential interval for the left surround channel. The value 1 denotes a 100% recognition of a difference, whereas 0 means no difference at all was perceived.

Figure 6  Average and 95% confidential interval over all test results. The value 1 denotes a 100% recognition of a difference, whereas 0 means no difference at all was perceived.
Figure 7  Dummy head with added joint and torso

Figure 8  Distribution of hearing event elevations over all subjects without head-tracking the pivoting movement
Figure 9  Distribution of hearing event elevations over all subjects with head-tracking the pivoting movement.

Figure 10  Median values of hearing event elevations in the entire region $-30^\circ \ldots 210^\circ$. 
Figure 11 Median and interquartiles of hearing event elevations (-30°... 60°)

Figure 12 Median and interquartiles of hearing event elevations (90°... 120°)
Figure 13 Distribution of perceived elevation angles of elevated sound sources outside the median plane (left without vertical head-tracking, right with)

Figure 14 Elevation angle of elevated sound sources outside the median plane (median and interquartiles).
Figure 15 Distribution of perceived azimuthal angles of elevated sound sources outside the median plane (left without vertical head-tracking, right with).

Figure 16 Azimuthal angle of elevated sound sources outside the median plane (median and interquartiles).
Figure 17 Perceived elevation angle of the center speaker

Figure 18 Perceived elevation angle of the right speaker
Figure 19 Perceived elevation angle of the right surround speaker

Figure 20 Difference between virtual and real center speaker regarding the perceived elevation angle
Figure 21 Difference between virtual and real right speaker regarding the perceived elevation angle

Figure 22 Difference between virtual and real right surround speaker regarding the perceived elevation angle