AUDIO-VISUAL DISCREPANCY AND THE INFLUENCE ON VERTICAL SOUND SOURCE LOCALIZATION

Stephan Werner\(^1\), Judith Liebetrau\(^1\), \(^2\), and Thomas Sporer\(^2\)

\(^1\) Ilmenau University of Technology, Ilmenau, Germany
\(^2\) Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany

Email: \{\*stephan.werner, \*judith.liebetrau\}@tu-ilmenau.de
\*thomas.sporer@idmt.fraunhofer.de

ABSTRACT

Human audio perception is influenced by vision and vice versa. The effect and thresholds of perceptual fusion, for example the ventriloquism-effect, are well investigated for natural listening conditions in the horizontal plane. Modern reproduction approaches for realistic spatial audio, e.g. binaural reproduction, promise more realistic sound reproduction, though, including proper perception of direction, distance, and elevation. This raises the question if the thresholds of perceptual fusion in audio reproduction systems that consider elevation are the same as in natural listening conditions. To estimate the influence of audio-visual discrepancy on vertical sound source localization via binaural headphones, two experiments were conducted. Results show an effect of similar magnitude for the vertical and horizontal plane.

Index Terms— Psychoacoustics, acoustic testing, binaural auralization, localization, ventriloquist-effect

1. INTRODUCTION

Undoubtedly, audio perception is profoundly influenced by vision and vice versa. The widely known McGurk-effect [1] demonstrates that visual information is able to severely impair the perception of the sound of individual syllables. Another example is the ventriloquism-effect [2, 3]. Here, the perception of the sound source is influenced by a visual cue in such a way that it is localized off from its origin. If the local discrepancy is large enough, both stimuli will be perceived as two discrete sources. When the discrepancy gets smaller, the audio stimulus will be attracted by the visual cue, until at a given point perceptual fusion will be reached: Both stimuli will be perceived as a single one.

These effects and the threshold of perceptual fusion are well investigated for natural listening conditions. Newer sound reproduction systems intend to create real 3-D stereo sound scenes, giving the listener a feeling of presence or “being there” [4]. One example for such a reproduction system is binaural reproduction using headphones. Although binaural synthesis works well in principle, there are some challenges and unexplored issues with the playback of binaural recordings. In particular the question arises if the perceived discrepancy of the senses is the same as in natural listening conditions.

Two experiments were carried out to estimate the influence of audio-visual discrepancy on vertical sound source localization via binaural headphones. Experiment I investigates if participants experience perceptual fusion of the positions of competing stimuli. Psychometric functions are established. In experiment II, the participants had to indicate the location of a sound in presence of a competing stimulus. A dislocation of perception was measured with this method.

2. PREVIOUS RESEARCH

Several studies have been conducted to investigate the effect of ventriloquism, with different experimental designs and procedures. Bertelson and Radeau [3] found deviations in sound localization of approx. 4° for 7° difference between audio and visual stimuli, 6.3° for 15°, and 8.2° for 25° between the audio and visual stimuli using loudspeakers and flashlights as sources. The sources were placed in the horizontal plane and their location was rated via hand pointing. Seeber and Fastl [2] used a pointing method to investigate the audio-visual discrepancy in real and virtual environments. For real environments, the mean shifting in localization were 4.3°, 1.9°, and 4.2° for horizontal viewing directions of -40°, 0°, and +40°. The median plane was not investigated. Similar results were found in experiments with binaural synthesis via headphones for individualized binaural simulation (individual HRTFs) and smaller shifting for non-individual HRTFs. Bohlander [5] obtained deviations of 1.5° to 5.9° for 45° discrepancy between median plane and real environment. Alais and Burr [6] carried out experiments to measure psychometric functions and points of subjective equality for the ventriloquist effect in azimuth depending on stimuli discrepancy and diameter of the light point. They detected a strong influence of the diameter of the light point. For small sizes the perceived
direction varied, as expected, directly with the visual stimulus. Although the above-mentioned studies investigated audio-visual displacement thoroughly, the results were only obtained, and therefore valid, for horizontal displacement.

In the study presented here, new tests to investigate the influence of audio-visual discrepancy on vertical sound source localization via binaural headphones were designed.

3. BINAURAL SYSTEM

The binaural system includes the recordings of individual binaural room impulse responses (BRIRs) for the used room and sound source positions and the auralization via headphones. The system was customized for each participant to avoid within-cone and out-of-cone confusion errors [7, 8]. The headphones were equalized using individual headphone transfer functions (HPTFs). In-ear microphones were used to measure individual BRIRs and individual HPTFs next to the eardrum of each test person. The measurements of the HPTFs were averaged over 5 recordings, repositioning the headphones for each recording. The inverse of a HPTF was calculated by a least-square method. A band-pass filter was applied between 80 Hz and 18 kHz. The measurements of the BRIRs were averaged over 3 recordings. Stax Lambda Pro headphones were used for playback.

4. EXPERIMENT I

The first experiment was intended to investigate participants’ experience of perceptual fusion of the positions of competing stimuli while listening to virtual sound reproductions over headphones. A test method was designed to investigate localization in virtual acoustics. In the first experiment, participants were provided with different test stimuli and had to report whether they perceive the audio stimulus below, in-plane, or above the visual stimulus.

4.1. Experimental design

The apparatus contains sound and visual source positions arranged on a segment of a circle with the test person in its center (see Fig. 1). The binaural auralization of the virtual loudspeakers via headphones is synthesized by a MATLAB audio player. White LEDs also arranged on the circle segment, with 5 mm diameter and approx. 15 cd luminous intensity, were used as visual sources. They were controlled by a MATLAB driven Arduino-Mega platform. During a test session, ambient light was dimmed to a minimum to keep visual distractions as low as possible.

4.1.1. Source positions

The combination of 4 sound source positions and 20 visual source positions were investigated. Table 1 shows the sound sources’ positions and their names.

A Geithain Mo-2 loudspeaker was used to measure the BRIRs for each of the four positions in a standardized listening lab (EBU Tech. 3276 [12] / ITU-R BS.1116-1 [11]). The distance from the loudspeaker to the listening point was 2.2 m. The height of the source positions was 1.26 m (i.e., the approximate ear position of a sitting person) for zero degree elevation. The recording positions of the BRIRs were identical to the listening position in the test. Custom-built in-ear microphones were used for measurements next to the eardrum [8].

Table 1: Azimuth and elevation of virtual sound source positions, used in experiment I

<table>
<thead>
<tr>
<th>Name</th>
<th>H0V0</th>
<th>H30V0</th>
<th>H0V25</th>
<th>H30V25</th>
</tr>
</thead>
<tbody>
<tr>
<td>azimuth</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>elevation</td>
<td>0°</td>
<td>+20°</td>
<td>+25°</td>
<td>+25°</td>
</tr>
</tbody>
</table>

Ten vertical positions at azimuths 0° and +30° were used for the visual sources. They covered a range from -10° to +35° elevation with 5° steps on a segment of a circle. Figure 1 shows the configuration of the experiments for the zero degree azimuth position. The black dots on the segment of a circle indicate the sound source positions. The grey dots indicate the visual source positions.

4.1.2. Conditions under test

All combinations of vertical audio and visual positions were used on each horizontal position. A six second long saxophone part without reverb and a white noise burst series (five bursts each with 30 ms duration and 3 ms cosine fade in/out and 70 ms silence between single bursts) were used as sound signals. The visual and audio stimuli were presented simultaneously. The order of the stimuli was randomized for each test person.

Figure 1: Positions of the audio and visual sources for experiment I and II; sound sources for playback via headphones are marked as black dots at 0° and +25° (exp. I) and at 0° and +20° (exp. II); visual (LED) positions marked as grey dots cover -10° to +35° with 5° intervals (exp. I) and white dots from -10° to +30° with 2.5° intervals (exp. II). Note that the sources were arranged on a segment of circle in experiment I, while experiment II had the sources arranged on a tangent plane.
4.1.3. Test panel
Two female and three male persons with normal hearing, aged between 24 and 33, participated in the listening tests. The participants were well experienced with listening tests. Prior to the test, a training session was done before the actual test started, familiarizing all listeners with the conditions and items under test. Participants additionally received a verbal and written introduction including definitions of the terms localization and externalization (following [9, 10]). Each participant had to listen to a selection of test stimuli consisting of stimuli with coinciding and diverging audio and visual source positions. Each training item had to be rated in order to become familiar with the testing procedure and to build an internal reference. Participants then had to judge the localization-differences between audio and visual stimuli for different deviations.

4.2. Experimental procedure
Two listening test sessions were performed. The first session investigated the assumed influence of a visual foothold on sound localization. The second test verified the sound localization acuity in elevation without a visual foothold. The first test session was divided into three parts. The first part contained the training of the participants to establish perceptional localization and externalization. The second and third part consisted of three repetitions of the test stimuli respectively, separated by a break of ca. 5 minutes. The total amount of stimuli was 256 (3 repetitions x 2 sounds x 4 audio positions x 10 visual positions = 240 plus 16 training stimuli) per test person. A whole session took approx. 60 minutes.

The participants had to answer the following question: Do you perceive the audio stimulus below, in-plane, or above the visual stimulus? All participants were instructed to keep the head straight and forward during listening and rating, and to listen to the whole stimulus before rating. Their answer was filled in on a datasheet by the supervisor. Eye movements were explicitly allowed to increase the fixation and enable better localization of the two stimuli. Repeated listening to the stimulus pairs was possible when requested by the subjects.

The task of the second session was to rate the perceived sound localization without a visual stimulus (i.e., an illuminated LED) on a vertical scale ranging from -10° to +35° with 2.5° interval indicators (see Fig. 1). Participants named the perceived direction. The same test supervisors as in the first session performed the test.

4.3. Results
The ratings of the test persons for localization are presented as normalized frequency of their occurrence. Due to indistinguishable differences between the used sound signals saxophone and noise bursts, both signals’ frequencies were combined for further analysis. The results of the first session (training) show that all participants rated the stimuli with zero degree deviation between audio and visual stimulus correctly.

![Figure 2: Localization results as normalized frequency of the ratings for the acoustical positions H0V0 and H30V0 and both sound signals (saxophone and noise); the deviation between the audio and visual stimulus is shown on the x-axis; negative values indicate that the audio stimulus is positioned below the visual stimulus; the horizontal line indicates 50% of the ratings; embedded boxplots show a discrepancy of sound localization ratings without visual support, the negative deviation indicates that the source direction is perceived above the presented direction.](image)

Figure 2 shows the normalized frequencies of the ratings from all participants for the audio positions H0V0 and H30V0 as a function of audio-visual discrepancy. The occurrences for the answers "below", "in-plane", and "above" are shown. The embedded boxplots show the results of the second test session, indicating medians and 25% and 75% quantiles. The medians are located at zero degree deviation for both synthesized source positions, but the distributions are skewed to the left, indicating that participants localized the sound higher than the presented sound direction. Half of the ratings are within 2.5° deviation.

Figure 3 shows the normalized frequencies of the ratings from all test persons for the acoustical positions H0V25 and H30V25 as a function of audio-visual discrepancy. The embedded boxplot again show the ratings from the second listening test session. The ratings for “in-plane” for upper vertical sound source positions shown in Figure 2 is spread more than for the zero degree vertical positions shown in
As expected, lower localization acuity is observed for upper lateral positions. This observation is independent from visual feedback. Higher audio-visual discrepancies are more tolerable for upper lateral and upper frontal positions. The measured localization acuity in the median plane without visual foothold but with binaural presentation via headphones is comparable with real source listening [2, 3]. However, the angular resolution of the test was too coarse to identify how the localization discrepancy between visual and audio stimulus compares to human localization acuity. Furthermore, the vertical positions >30° were difficult to see for some test persons with glasses as head movements were forbidden and the borders of their glasses distorted the image.

5. EXPERIMENT II

The second experiment attempts to verify and refine the findings of experiment I with a slightly different method design. A new method was chosen for the indication of the localized sound source positions. In [2] a method utilizing a laser pointer to indicate localized direction is presented. As stated there, the so-called Propricetion Decoupled Pointer (Pro De Po) method shows less localization error and variance than most alternative localization methods, especially at lateral angles. Due to the promising results shown in [2] an adaption of this method was chosen for the indication of sound source localization for experiment II.

5.1. Experimental design

The second experiment has a similar design than experiment I. The main differences are the arrangement of sound and visual sources on a tangent plane instead of a spherical cap (see Figure 1), an increase of the number of visual sources, and using a pointer method similar to the Pro De Po method.

5.1.1. Source positions

Four sound and 34 visual source positions were used. The sound sources are displayed in Table 3.

<table>
<thead>
<tr>
<th>Name</th>
<th>H0V0</th>
<th>H20V0</th>
<th>H0V20</th>
<th>H20V20</th>
</tr>
</thead>
<tbody>
<tr>
<td>azimuth</td>
<td>0°</td>
<td>+20°</td>
<td>0°</td>
<td>+20°</td>
</tr>
<tr>
<td>elevation</td>
<td>0°</td>
<td>0°</td>
<td>+20°</td>
<td>+20°</td>
</tr>
</tbody>
</table>

Table 3: Azimuth and elevation of virtual sound source positions, used in experiment II

5.1.2. Conditions under test

Four Genelec 8030BPM loudspeakers were used to measure the BRIRs in a standardized listening lab (see experiment I). Svartek SV-25S in-ear microphones were used for measurements. The distance from the loudspeaker at H0V0 to the listening point was 2.2 m. The height of the source position was 1.26 m (approximate ear position of a sitting person) for zero degree elevation. Seventeen vertical
positions at azimuths 0° and +20° were used for the visual sources (LEDs). They covered a range from -10° to +30° with 2.5° steps. A black sound-transparent curtain was placed directly in front of the LEDs. The size of the light dots was 10 mm in diameter (ca. 0.26°) on the front side of the curtain. All combinations of acoustical and visual vertical directions were used for both horizontal directions. A dry acoustics male speech part with 4 seconds duration and the white noise burst trains from experiment I were used as sound signals. The visual and audio stimuli were presented at different times, the audio stimulus being delayed 150 ms to the visual stimulus. A lower fusion of both stimuli compared to a simultaneous occurrence was expected [3].

5.1.3. Test panel

Two female and four male persons with normal hearing, aged between 21 and 30, participated in the listening test. The participants were experienced with listening tests. Consistent with the first experiment, all participants had to complete a training session to become familiar with the selection of conditions under test, the test procedure, the input device, and to build an internal reference for the judgment. The selection consisted of test stimuli with both coinciding and diverging audio and visual source positions, and audio sources alone.

5.2. Experimental procedure

This experiment consisted of only one listening test session to investigate the assumed influence of a visual foothold on sound localization and to verify the sound localization acuity in elevation without a visual foothold. The test session was divided into three parts, the first being the training. The second and third part included two repetitions of the test stimuli of all combinations of visual and audio positions in randomized order. Furthermore, the audio positions without visual feedback were presented twice. A break of approx. five minutes was taken between parts to avoid listener fatigue. The number of stimuli was 320 per test person (2 repetitions x 2 sounds x 4 audio positions x 17 visual positions = 272 plus 2 repetitions x 2 sounds x 4 audio positions = 16 plus 32 training stimuli). One session took approx. 60 minutes.

Participants rated the sound event by pointing with a laser pointer in their left or right hand on a black curtain at the perceived incidence angle. The Curtain was placed direct in front of the LEDs. A webcam, controlled over a network connection, recorded the rating by taking a screenshot after participants pushed a button to trigger the camera. All participants were instructed to keep the head straight and forward during listening and rating, and to listen to the whole stimulus before rating. Eye movement was allowed. Repeated listening to stimuli was possible, if required.

5.3. Results

For the analysis a grid was projected with a video projector on the curtain and a screenshot with the webcam was taken. The projected grid was geometrically warped to fit the correct horizontal and vertical angles from a circle segment with its center at the listening position. The angular resolution of the grid was 1°. The grid was recorded once and it was not visible during experiment. The laser point from the test person was detected within the screenshot of each rating and compared to its position on the grid.

Figure 4 shows the grid with an exemplary rating marked as a cross at +9° vertical and +1° horizontal direction.

Figure 4: Screenshot of the projection of the warped grid on the curtain in front of the subject; an exemplary rating is shown as a black cross at +9° vertical and +1° horizontal position; 0° position is marked as 5 points in lower left part of the Figure (cropped and inverted picture for better visual presentation).

The quantiles of the data from the localization test with presentation of visual stimuli (test trial) were normalized to the corresponding results from the localization test without visual stimuli (control trial). The influence of the visual foothold, i.e., the deviation was then calculated as the difference of the medians between the normalized test trials and the control trial for each audio position. A mean absolute deviation (mad) of the medians was calculated over all visual directions and over a range from +10° to +30° for V0 conditions and over a range from -10° to +10° for V20 conditions. Significant results of one-sided sign test for the hypothesis of zero degree bias are given as asterisks in Figure 5 and Figure 6.

Figure 5 shows the vertical deviation for condition H0V0 and H20V0 under the influence of visual stimuli. A significant vertical deviation is observed for visual stimulus directions of greater than or equal to +5° (except +25° for H0V0) and smaller than or equal to -7.5° for H0V0, and the mad increases for lateral positions.
Figure 5: Vertical deviation in degree for the condition H0V0 (left) and H20V0 (right) related to the direction of the visual stimulus; mad = mean absolute deviation; * p<.05 by one-sided sign test.

Figure 6 shows the vertical deviation for condition H0V20 and H20V20 under the influence of visual stimuli. Significant vertical deviations are observed for all visual stimulus directions smaller than or equal to +15° and greater than or equal to +22.5° (except +25°) for H20V20. The condition H0V20 shows the same trend, but with no significant (p<.05) results for some directions. The mad is increasing for upper lateral condition. A stronger increase is observed between the frontal and upper conditions.

Figure 6: Vertical deviation in degree for the condition H0V20 (left) and H20V20 (right) related to the direction of the visual stimulus; mad=mean absolute deviation; * p<.05 by one-sided sign test.

The observed bias is consistent with literature [2, 3] for intersensory discrepancies in azimuth for real sound sources and binaural synthesized sources. An imbalance can be observed between positive and negative discrepancies. This is not reported for experiments in azimuth. Furthermore, slightly higher bias is found for the V20 conditions.

6. CONCLUSION

Two experiments were conducted to evaluate psychometric functions and intersensory bias of competing audio and visual stimuli. The ventriloquism effect for vertical positions is investigated for frontal and lateral azimuth directions. An individualized binaural auralization via headphones is used. Experiment I indicates that an increase of audio-visual discrepancy is more tolerable for upper and upper lateral directions to maintain perceptual fusion. In experiment II, the observed mean deviation of a maximum of 3.6° for an intersensory discrepancy from -10° to -30° at H20V20 are smaller than deviations from former experiments in horizontal plane (see e.g. [2, 3]). This might be an effect of less fusion between the audio and visual stimuli caused by the asynchronous onset of audio and visual stimulus. Alternatively this could be an indication of less influence of audio-visual discrepancy for elevated sources caused by the lower resolution in localizing elevated sound sources. However, we can show that the measured ventriloquism effect for an individualized binaural synthesis via headphones has similar magnitudes for elevated source positions as it has in the horizontal plane for virtual and real environments.

An intersensory bias was calculated by dividing the median of the deviation and the intersensory discrepancy between the audio and visual stimuli. The bias was a direct bias with a minimum influence of adaptation effects [3]. Figure 7 shows the intersensory bias for the four conditions.

Figure 7: Intersensory bias for the conditions H0V0 and H20V0 (left) and H0V20 and H20V20 (right).
7. ACKNOWLEDGEMENTS

This study was planned and conducted in a workshop at Ilmenau University of Technology. The authors would like to thank the attendees M. Bauer, T. Brass, and H. Gräber for their support in planning and conducting the experiments, as well as S. Schneider for proofreading and the test participants for their interest in research and participation in this study.

8. REFERENCES


