Stephan Werner, Judith Liebetrau, Thomas Sporer:

Vertical Sound Source Localization Influenced by Visual Stimuli

Original published in:
Signal Processing Research, 2 (June 2013), 2, p 29-38.
ISSN Online: 2327-171X ; ISSN Print: 2327-1701
URL: http://www.seipub.org/spr/paperInfo.aspx?ID=2701
(Visited: 2013-07-12)

Vertical Sound Source Localization Influenced by Visual Stimuli

Stephan Werner*[1], Judith Liebetrau[2], Thomas Sporer[3]

Electronic Media Technology Lab, Ilmenau University of Technology, Ilmenau, Germany

Fraunhofer Institute for Digital Media Technology , Ilmenau, Germany

*[1]stephan.werner@tu-ilmenau.de; [2]judith.liebetrau@tu-ilmenau.de; [3]thomas.sporer@idmt.fraunhofer.de

Abstract

It is well-known that the perception of the position of audio and video stimuli is not independent. In general, video dominates the position if the position offset between audio and video is small. Most previous work focused on natural listening conditions and position offsets between audio and video in the horizontal plane. There is little research concerning offsets in vertical direction and artificial, auralized sound environments. Among different approaches to auralization of spatial audio, the binaural reproduction is especially very interestingas it offers proper perception of direction, distance, and elevation of sound sources at moderate cost.

This article addresses the question whether the thresholds of perceptual fusion of audio and video stimuli are the same in binaural reproduction systems and in natural listening conditions. To estimate the influence of audio-visual discrepancy on vertical sound source localization, two experiments have been designed. The test methods were optimized to improve usability and minimize rating errors. Both experiments resulted in psychometric functions of intersensory bias for competing audio and visual stimuli. For binaural reproduction, the obtained results showed an effect of similar magnitude for both the vertical and horizontal plane which is in good agreement with the results obtained from other studies in natural environments.

Keywords

Psychoacoustics; Acoustic Testing; Binaural Auralization; Localization; Ventriloquist-Effect

Introduction

It is established that audio perception is profoundly influenced by vision and vice versa. The widely known McGurk-effect (McGurk and MacDonald, 1976) demonstrates that visual information is able to severely impair the perception of the sound of individual syllables: Depending on the movement of the lips of a talking head the syllable perceived by a listener changes from /ba-ba/ (audio only) to /da-da/ (audio with video). Another example is the ventriloquism-effect (Seeber and Fastl, 2004), (Bertelson and Radeau, 1981). A puppet player creates the illusion that the puppet is talking. 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.

Many studies have investigated these effects and the thresholds for perceptual fusion in natural listening conditions. The target for technical systems for virtual reality is to create the illusion of being in a different audio-visual environment. Total immersion can only be achieved if audio is reproduced with 3-D audio systems (Heeter, 1992). An example for such an audio 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. Among these are the issues of personalization of head related transfer functions (HRTFs), the effects and compensation of head movements and the influence of the reproduction room. A particular question rarely addressed by other studies is whether the perceived discrepancy of visual and auditory stimuli in binaural reproduction is the same as in natural listening conditions.

With the advent of 3D Audio Systems (IOSONO, Dolby Atmos, Auro3D, etc.) audiovisual content with elevation has become available. Traditionally cinema positioned visual sound sources in the center channel only, but now proper positioning of audio has become possible. Studies from Ode et al. (2011) and others indicate that this gives an improvement of perceived AV-quality. It is for seen that 3D content will also reproduced on mobile devices using binaural

reproduction (International Organisation for Standardisation, 2012). Such systems might only use a limited number of BRIRs stored, interpolation of BRIRs might cause unwanted computational load and therefore it is necessary to find compromises including larger audio-visual discrepancy.

Two experiments were carried out to estimate the influence of audio-visual discrepancy on vertical sound source localization via binaural headphones. Experiment I investigates whether 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: The dislocation of perception was measured with this method.

Previous Research

Several studies have been conducted in the past to investigate the effect of ventriloquism in the horizontal plane, with different experimental designs and procedures. Bertelson and Radeau (1981) 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 (2004) 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 (1984) obtained deviations of 1.5° to 5.9° for 45° discrepancy between median plane and real environment. Alais and Burr (2004) 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 are valid, for horizontal displacement.

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

Binaural System

For generating test stimuli, binaural recordings of individual binaural room impulse responses (BRIRs) for the used room and sound source positions and the auralization via headphones were prepared. The binaural system was customized for each participant to avoid within-cone and out-of-cone confusion errors (Kunze, Liebetrau, and Korn, 2012), (Møller, Sørensen, Jensen, and Hammershøi, 1996), (Werner and Siegel, 2011) and to increase the simulation´s similarity compared with the real loudspeakers (Begault and Wenzel, 2001). A listening lab with defined room acoustics and an adequate source receiver distance were chosen to include reverberation. Reverberation encourages the perception of externalization of an auditory illusion (Werner and Siegel, 2011), (Lindau and Brinkmann, 2010) and the impression of distance (Laws, 1973), (Shinn-Cunnigham, 2000). The receiver-source distance was chosen to be in the far-field of the loudspeaker and the receiver (head) in the effect that no variation of binaural cues depending on the distance is present (Kapralos, Jenkin, and Milios, 2003). 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 subject. The microphones are not removed between the BRIR and HPTF measurements. The measurements of the HPTFs were averaged over five recordings, repositioning the headphones for each recording. The inverse of a HPTF was calculated by a least-square method with minimum phase inversion (Schärer and Lindau, 2009). A band-pass filter was applied between 80 Hz and 18 kHz. The measurements of the BRIRs were averaged over three recordings. Stax Lambda Pro headphones were used for playback. The inherent insufficiences of the binaural synthesis are minimized by customize the system (Begault and Wenzel, 2001).

Experiment I

The intention of the first experiment was to investigate how participants experience perceptual fusion of the positions of competing visual 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.

Experimental Design

The apparatus contains sound and visual source positions arranged on a segment of a circle with the test participant 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 (Arduino, 2013). The LED arrays were visible during test. Ambient light was dimmed to a minimum to keep visual distractions as low as possible.

1) Source Positions

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

TABLE 1 AZIMUTH AND ELEVATION OF VIRTUAL SOUND SOURCE POSITIONS, USED IN EXPERIMENT I

Name	H0V0	H30V0	H0V25	H30V25
azimuth	0°	+30°	0°	+30°
elevation	0°	0°	+25°	+25°

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 / ITU-R BS.1116-1). 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 (Møller, Sørensen, Jensen, and Hammershøi, 1996).

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. Fig. 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.

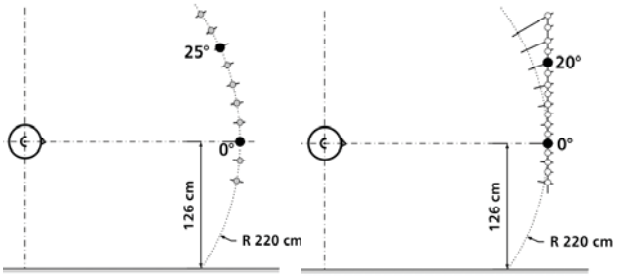

FIG. 1 POSITIONS OF THE AUDIO AND VISUAL SOURCES FOR EXPERIMENT I AND II; SOUND SOURCES FOR PLAYBACK VIA HEADPHONES ARE MARKED AS BLACK DOTS AT 0° AND +25° (EXP. I LEFT FIGURE) AND AT 0° AND +20° (EXP. II RIGHT FIGURE); VISUAL (LED) POSITIONS MARKED AS GREY DOTS COVER -10° TO +35° WITH 5° INTERVALS (EXP. I LEFT FIGURE) AND WHITE DOTS FROM -10° TO +30° WITH 2.5° INTERVALS (EXP. II RIGHT FIGURE). 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.

2) Test Conditions

All combinations of vertical audio and visual positions were used on each horizontal position. Two different types of audio content where used: An anechoic recording of saxophone (duration 6s) and a series of white noise burst (five bursts each with 30 ms duration and 3 ms cosine fade in/out and 70 ms silence between single bursts). The saxophone item was chosen because it has a spectral and tonal characteristic like human speech (Nykänen and Johannson, 2003), (Teal, 1963), but without the unwanted influence to distance perception caused by articulation or familiarization (Blauert, 2001). Both visual and audio stimuli were presented simultaneously. The order of the stimuli was randomized for each subject.

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, 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 (Merimaa and Hess, 2004), (Hartmann and Wittenberg, 1996)). 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.

Experimental Procedure

The Experiment I consisted of one listening test sessions. Test investigated the assumed influence of a visual cue on sound localization for frontal, lateral and elevated directions of the stimuli. The test session was divided into three parts. The first part contained the training of the participants to establish perceptional localization and externalization. The training stimuli included the four directions, two sound signals, and congruence respectively divergence between the audio and visual stumulus. The second and third part consisted of three repetitions of the test stimuli respectively, separated by a break of ca. five minutes. The total amount of stimuli was 256 (3repetitions x 2sounds x 4audio positions x 10visual positions = 240 plus 16 training stimuli) per subject. 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. To avoid any movements or distraction by operating a computer interface all feedback of the subjects was done verbally only. Their answers were filled in 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.

Results

The ratings of the subjects for localization are presented as normalized frequency (percentage) of their occurrence. The differences in the results for the two items Saxophone and Noise Burst proved to be sufficiently small. Therefore the results of both items have been 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.

Fig. 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 in the figure. A horizontal line indicates the 50% point of the ratings.

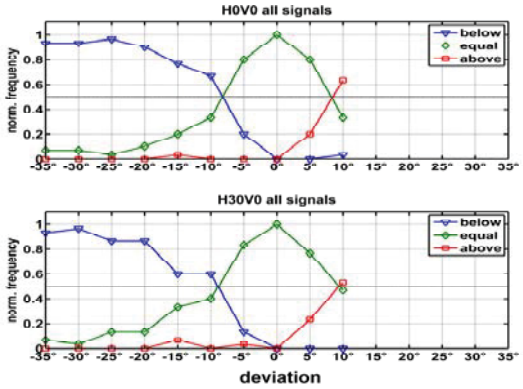

FIG. 2 LOCALIZATION RESULTS AS NORMALIZED FREQUENCY OF THE RATINGS FOR THE ACOUSTICAL POSITIONS H0V0 AND H30V0 AND BOTH SOUND SIGNALS (SAXOPHONE AND NOISE); THE DEVIATION BETWEEN THE AUDIO AND VISUAL STIMULUS IS SHOWN ON THE X-AXIS; NEGATIVE VALUES INDICATE THAT THE AUDIO STIMULUS IS POSITIONED BELOW THE VISUAL STIMULUS; THE HORIZONTAL LINE INDICATES 50% OF THE RATINGS.

Fig. 3 shows the normalized frequencies of the ratings from all subjects for the acoustical positions H0V25 and H30V25 as a function of audio-visual discrepancy. The ratings for “in-plane” for upper vertical sound source positions shown in Fig. 2 is spread more than for the zero degree vertical positions shown in Fig. 3. This leads to the conclusion that participants tolerate a larger deviation between the visual and audio source position for audio sources at higher elevation.

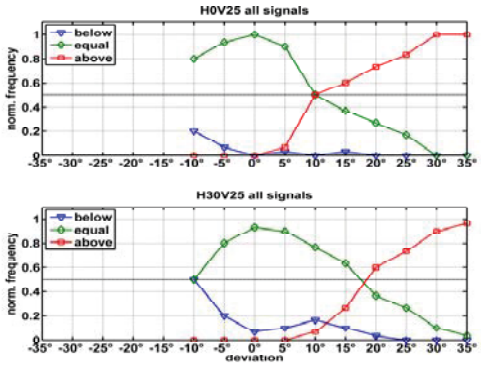

FIG. 3 LOCALIZATION RESULTS AS NORMALIZED FREQUENCY OF THE RATINGS FOR THE AUDIO POSITIONS H0V25 AND H30V25 AND BOTH SOUND SIGNALS (SAXOPHONE AND NOISE); THE DEVIATION BETWEEN THE AUDIO AND VISUAL STIMULUS IS SHOWN ON THE X-AXIS; POSITIVE VALUES INDICATE THAT THE AUDIO STIMULUS IS POSITIONED ABOVE THE VISUAL STIMULUS; THE HORIZONTAL LINE INDICATES 50% OF THE RATINGS.

Table 2 lists the estimated deviation angles of ratings “in-plane” from all participants at the 50% point of normalized frequency. An increase of perceived deviation between audio and visual stimulus is visible for the elevated positions H0V25 and H30V25.

TABLE 2 ESTIMATED DEVIATIONS IN DEGREE FOR THE 50% POINT OF THE FREQUENCIES FOR RATING "EQUAL", (*: NO RELIABLE ESTIMATE AVAILABLE).

	H0V0	H30V0	H0V25	H30V25
50% point	+8°/-8°	+9°/-9°	+10°/*	+17°/-10°

A McNemar´s test was performed to estimate the significance of differences between the frequencies for ratings "in-plane" and "not in-plane" across the audio conditions. The rating "not in-plane" is thereby defined as the sum of ratings for "above" and "below". Significant differences (p<.05, N=500, DF=1) can be found between conditions H0V0 and H0V25, H0V0 and H30V25, and H30V0 and H30V25 (see table 3).

TABLE 3 CHI-VALUES AND PHI-VALUES (IN BRACKETS) FOR ANALYSIS OF DIFFERENCES (MCNEMAR´S TEST) BETWEEN THE RATINGS "EQUAL" AND "NOT EQUAL" FOR ALL ACOUSTICAL CONDITIONS; SIGNIFICANT VALUES ARE BOLD TYPE (P<.05, N=500, DF=1).

	H0V0	H30V0	H0V25	H30V25
H0V0	-	1.13(0.04)	**8.64(0.12)**	**14.73(0.16)**
H30V0		-	3.53(0.08)	**7.78(0.11)**
H0V25			-	0.81(0.04)

The reliabilities of the ratings over all subjects are shown in Fig. 4 for the 0° elevation direction and in Fig. 5 for the +25° elevation direction. The reliability is 100% for 0° vertical deviation for all test signals, except for the condition H30V25. A decrease of reliability is visible for increasing deviations. The visual and acoustical directions are not clearly separable by the subjects. The reliability is close to 100% if the vertical deviation increases further because the visual and acoustical directions are distinct separable. The reliability is used as an indicator for the influence of a visual cue on the localization of an acoustical event.

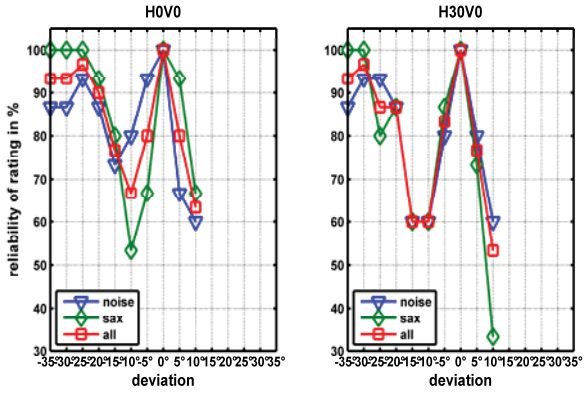

FIG. 4 RELIABILITY OF RATINGS OF ALL TEST PARTICIPANTS FOR THE TWO ACOUSTICAL POSITIONS H0V0 AND H30V0 AND TEST SIGNALS (SAXOPHONE, NOISE, AND BOTH SIGNALS TOGETHER); THE DEVIATION IN DEGREE BETWEEN THE AUDIO STIMULUS AND THE VISUAL STIMULUS IS SHOWN ON THE X-AXIS; A POSITIVE DEVIATION INDICATES THAT THE AUDIO STIMULUS IS ABOVE THE VISUAL STIMULUS.

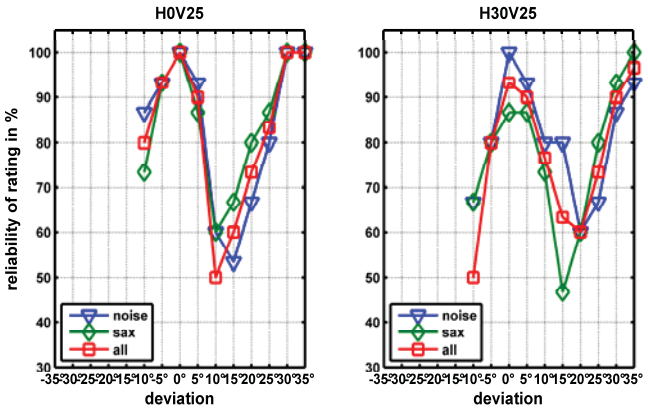

FIG. 5 RELIABILITY OF RATINGS OF ALL TEST PARTICIPANTS FOR THE TWO ACOUSTICAL POSITIONS H0V25 AND H30V25 AND TEST SIGNALS (SAXOPHONE, NOISE, AND BOTH SIGNALS TOGETHER); THE DEVIATION IN DEGREE BETWEEN THE AUDIO STIMULUS AND THE VISUAL STIMULUS IS SHOWN ON THE X-AXIS; A POSITIVE DEVIATION INDICATES THAT THE AUDIO STIMULUS IS ABOVE THE VISUAL STIMULUS.

As expected, audio-visual discrepancies in direction are more tolerable for upper lateral and upper frontal positions compared to lateral positions with 0° elevation. The estimated deviations cover a range from 8° for non-elevated positions to 17° for lateral and elevated positions. The presented results are affected by the localization accuracy without visual cues. 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 accuracy. The measured localization in the median plane with binaural presentation via headphones is comparable with real source listening (Seeber and Fastl, 2004), (Bertelson and Radeau, 1981). Furthermore, the vertical positions >30° were difficult to see for some subjects with glasses as head movements were forbidden and the borders of their glasses distorted the image.

Experiment II

The second experiment attempts to verify and refine the findings of experiment I with a slightly different test design. A new method was chosen for the indication of the localized sound source positions. Seeber and Fastl used a laser pointer to indicate localized direction (Seeber and Fastl, 2004). They proved that the so-called Proprioception 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 Seeber and Fastl (2004) an adaption of this method was chosen for the indication of sound source localization for experiment II.

Experimental Design

The principal setup of the second experiment is similar to the setup used in 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 the usage of a pointer method similar to the Pro De Po method. While acoustic and visual stimuli were presented simultaneously in experiment I the stimuli were presented with an offset in experiment II. The visual sources (LEDs), the pointing device, and the recording of the ratings were controlled by MATLAB and a MATLAB driven Arduino-Mega platform (Arduino, 2013).

1) Source Positions

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

TABLE 3 AZIMUTH AND ELEVATION OF VIRTUAL SOUND SOURCE POSITIONS, USED IN EXPERIMENT II.

Name	H0V0	H20V0	H0V20	H20V20
azimuth	0°	+20°	0°	+20°
elevation	0°	0°	+20°	+20°

2) Test Conditions

Four Genelec 8030BPM loudspeakers were used to measure the BRIRs in a standardized listening lab (see Experiment I). Svantek SV-25S in-ear microphones were used for BRIR and HPTF 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 as 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 (approx. 0.26°) on the front side of the curtain. All combinations of acoustical and visual vertical directions were used for both horizontal directions. Two audio stimuli were used in experiment II: An anechoic recording of male speech (duration 4 s) and the white noise burst sequence already used in experiment I. The visual and audio stimuli were presented at different times, the audio stimulus being delayed 150 ms to the visual stimulus caused by technical limitations of stimuli presentation and the recording of the rating with an IP camera. Due to this time difference less fusion of both stimuli compared to a simultaneous occurrence was expected (Bertelson and Radeau, 1981).

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 used for training consisted of test stimuli with both coinciding and diverging audio and visual source positions, and of test stimuli with audio sources only.

Experimental Procedure

This experiment consisted of one listening test session to investigate the assumed influence of a visual cue on sound localization and to verify the sound localization accuracy in elevation without a visual cue. 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 subject (2repetitions x 2sounds x 4audio positions x 17visual positions = 272 plus 2repetitions x 2sounds x 4audio 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 directly 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.

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 subject was detected within the screenshot of each rating and compared to its position on the grid.

Fig. 6 shows the grid with an exemplary rating marked as a cross at +9° vertical and +1° horizontal direction.

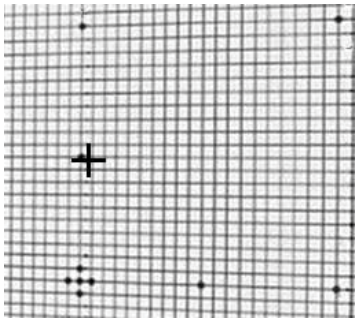

FIG. 6 SCREENSHOT OF THE PROJECTION OF THE WARPED GRID ON THE CURTAIN IN FRONT OF THE SUBJECT; AN EXEMPLARY RATING IS SHOWN AS A BLACK CROSS AT +9° VERTICAL AND +1° HORIZONTAL POSITION; 0° POSITION IS MARKED AS 5 POINTS IN LOWER LEFT PART OF THE FIGURE (CROPPED AND INVERTED PICTURE FOR BETTER VISUAL 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 cue, 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. The selection of the boarders are motivated by the results and the 50% point from experiment I. Significant results of one-sided sign test for the hypothesis of zero degree bias are given as asterisks in Fig. 7 and Fig. 8.

Fig. 7 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.

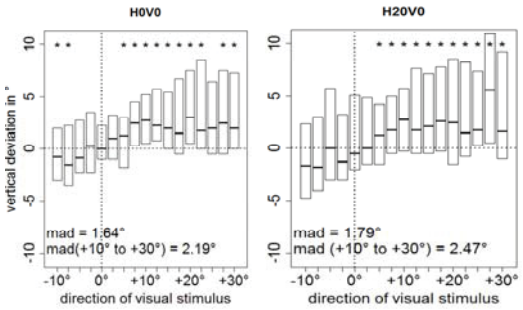

FIG. 7 VERTICAL DEVIATION IN DEGREE FOR THE CONDITION H0V0 (LEFT) AND H20V0 (RIGHT) RELATED TO THE DIRECTION OF THE VISUAL STIMULUS; MAD = MEAN ABSOLUTE DEVIATION; * P<.05 BY ONE-SIDED SIGN TEST.

Fig. 8 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.

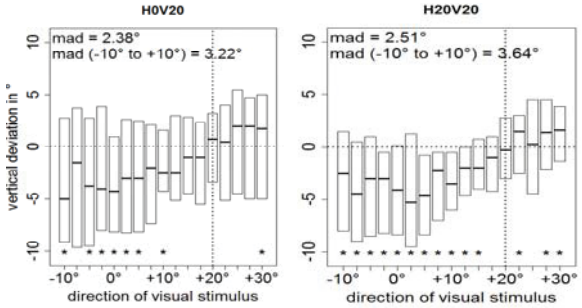

FIG. 8 VERTICAL DEVIATION IN DEGREE FOR THE CONDITION H0V20 (LEFT) AND H20V20 (RIGHT) RELATED TO THE DIRECTION OF THE VISUAL STIMULUS; MAD=MEAN ABSOLUTE DEVIATION; * P<.05 BY ONE-SIDED SIGN TEST.

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]. Fig. 9 shows the intersensory bias for the four conditions.

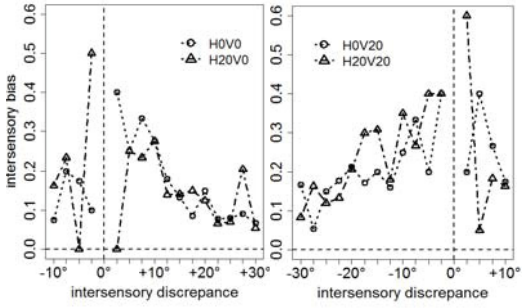

FIG. 9 INTERSENSORY BIAS FOR THE CONDITIONS H0V0 AND H20V0 (LEFT) AND H0V20 AND H20V20 (RIGHT).

The observed bias is consistent with literature (Seeber and Fastl, 2004), (Bertelson and Radeau, 1981) 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.

Conclusions

Two experiments have been conducted to evaluate psychometric functions and intersensory bias of competing audio and visual stimuli. The ventriloquism effect for vertical positions was investigated for frontal and lateral azimuth directions. An individualized binaural auralization via headphones was used to increase the simulation´s similarity compared with real loudspeaker listening. The results from experiment I indicate that for upper and upper lateral directions an increase of audio-visual discrepancy is possible without disturbing perceptual fusion. The deviations are approx. 8° for non-elevated positions and approx. 17° for lateral elevated positions. The results are affected by the localization accuracy without visual cues which leeds to experiment II. From the results of experiment II it can be seen that the observed mean deviation of a maximum of 3.6° for an intersensory discrepancy from -10° to -30° at an audio position with 20° azimuth and 20° elevation (H20V20) is smaller than deviations reported in former experiments in the horizontal plane (see e.g. (Seeber and Fastl, 2004), (Bertelson and Radeau, 1981)). This observation might be caused of less fusion between the audio and visual stimuli due to the asynchronous onset of 150 ms between audio and visual stimulus. Another explanation is that the reduced resolution for localization of elevated sound sources might lead to a smaller influence of audio-visual discrepancy. 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.

ACKNOWLEDGMENT

This study was planned and conducted in a workshop at Ilmenau University of Technology. The authors would like to thank 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.

REFERENCES

Alais, D. and Burr, D., "The Ventriloquist Effect Results from near-Optimal Bimodal Integration", Current Biology, 14, 257-262, 2004.

Arduino, Arduino Mega 2560, http://www.arduino.cc, last call on 21st January 2013.

Begault, D. R. and Wenzel, E.M. "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), 904-916, 2001.

Bertelson, P. and Radeau, M., "Cross-Modal bias and Perceptual Fusion with Auditory-Visual Spatial Discordance", Perception and Psychophysics, 29 (6), 578-584, 1981.

Blauert,J.: "Spatial Hearing - The Psychophysics of Human Sound Localization". Revised Edition, Cambridge, London: MIT Press, 2001.

Bohlander, R., "Eye Position and Visual Attention Influence Perceived Auditory Direction", Percep. Mot. Skills, 59:483.510, 1984.

EBU Doc. Tech 3276-1998 (Second Edition): Listening Conditions for the Assessment of Sound Program Material: Monophonic and Two-Channel Stereophonic) and EBU Doc. Tech 3276-1999: Supplement 1: Multichannel Sound.

Hartmann, W. M. and Wittenberg, A., "On the externalization of sound images". J. Acoust. Soc. Am., 99(6), 3678-3688, 1996.

Heeter, C. "Being There: The Subjective Experience of Presence", Presence: Teleoperators and Virtual Environments, MIT Press, 1992.

International Organisation for Standardisation, Coding of Moving Pictures and Audio, ISO/IEC JTC1/SC29/WG11/w13194, Draft Call for Proposals for 3D Audio, Shanghai, China, 2012.

Kapralos, B., Jenkin M.R.M. and Milios E. "Auditory Perception and Spatial (3D) Auditory Systems". Technical Report, University York, Canada, 2003.

Kopčo, N., and Shinn-Cunningham, B. "Auditory Localization in Rooms: Acoustic Analysis and Behavior".

Proceedings of the 32nd EAA International Acoustics Conference of the European Acoustics, 2002.

Laws, P., "Entfernungshören und das Problem der Im-Kopf-Lokalisiertheit von Hörereignissen [Auditory Distance Perception and the Problem of "In-Head Localization" of Sound Images], Acustica, 29, 243-259 (NASA Technical Translation TT-20833), 1973.

Lindau, A., and Brinkmann, F.: "Perceptual Evaluation of Individual Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings". 3rd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, 137-142, 2010.

McGurk, H. and MacDonald, J. "Hearing Lips and Seeing Voices". In: Nature, Vol. 264, 746-748, 1976.

Merimaa, J. and Hess, W., "Training of Listeners for Evaluation of Spatial Attributes of Sound". Proc. of the 117th AES Conv., Preprint 6237, San Francisco, 2004.

Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D., "Binaural Technique: Do We Need Individual Recordings?". J. Audio Eng. Soc, 44(6), 451-469, 1996.

Nykänen, A. and Johannson, Ö. "Development of a Language for Specifying Saxophone Timbre". In: Proceedings if the Stockholm Music Acoustics Conference (SMAC). Stockholm, Schweden: SMAC, 2003.

Ode, S., Sawaya, I., Ando, A., Hamasaki, K., and Ozawa, K. "Vertical Loudspeaker Arrangement for Reproducing Spatially Uniform Sound", Audio Engineering Society Convention 131, Paper 8512, USA, 2011.

Recommendation ITU-R BS.1116-1 (10/1997) "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems". International Telecommunication Union, Radio communication Assembly, 1997.

Schärer, Z. and Lindau, A. "Evaluation of Equalisation Methods for Binaural Signals", Proc. of the 126th AES Conv., Preprint 7721, 2009.

Seeber, B. and Fastl, H., "On auditory-Visual Interaction in Real and Virtual Environments", In Proc. ICA 2004, 18th Int. Congress on Acoustics, Kyoto, Japan, volume III, Int. Commission on Acoustics, 2293–2296, 2004.

Shinn-Cunningham, B. "Distance cues for Virtual Auditory Space". Special Session on Virtual Auditory Space, Proceedings of the First IEEE Pacific-Rim Conference on Multimedia, 13-15 December 2000, Sydney, Australia, pp. 227-230, 2000.

Teal, L. "The Art of Saxophone Playing". Miami, USA: Summy-Birchard Music, 1963.

Werner, S. and Siegel, A., "Effects of Binaural Auralization via Headphones on the Perception of Acoustic Scenes". Proc. of the 3rd International Symposium on Auditory and Audiological Research ISAAR, 215-222, Denmark, 2011.

Stephan Werner was born in Merseburg, Germany in 1981. He finished high school (german: Gymnasium) in 1999. In 2000 he started studying Media Technology at Ilmenau University of Technology and received his Master of Science (Diplom Ingenieur) in 2007 with the topic "Separation of wanted signals from noise signals based on vesicle filtering in a neuronal auditory model".

In 2007 he was research assistant at Fraunhofer Institute for Digital Media Technology in Ilmenau. Since 2008 he is a research and teaching assistant at "Ilmenau University of Technology" at the Electronic Media Technology Lab in Ilmenau, Germany. Currently, he aspires to a PhD degree about "Effects in perception of auditory illusions". His main research interests are binaural synthesis related with room acoustics, context dependencies, and perceptional evaluation.

Judith Liebetrau studied "Media Technology" at "Ilmenau University of Technology" and received her Dipl.-Ing. Degree (Master of Science) in 2006. Having graduated from university, she started to work at the department of Acoustics at the Fraunhofer Institute for Digital Media Technology IDMT. Her main work focuses on research concerning video and sound quality assessment as well as audio-visual perception. She authored and co-authored papers on perceptual evaluation of sound quality and psychoacoustical effects. She has also participated in the Working Party 6C (WP 6C)–"Programme production and quality assessment" of the standardization body ITU-R and is Rapporteur for audio quality assessment as well as chairmain of the Rapporteur Group for the revision of ITU-R BS.1534-1.

In 2011 she started on a DFG (German Research Foundation) founded project at Ilmenau University of Technology. Here, she tries to answer the question what parts of music influence emotions and how it does. Automatic information retrieval and music classification considering emotional aspects is also be part of this project.

Thomas Sporer was born in 1964 and earned a M.Sc. in computer science (Diplom-Informatiker) from the Universität Erlangen-Nürnberg in 1988 and received his Ph.D. in electrical engineering in 1998. From 1988 to 1989 he worked at the Fraunhofer Institute for Integrated Circuits in Erlangen, Germany in the audio research group on

perceptual audio coding. Since then he returned to the university where he was working in the department of electrical engineering as a research and teaching assistant, continuing to support the development of mp3 and aac, but mainly focusing on perceptual measurement.

In June 1997 he returned to the Fraunhofer Institute in Erlangen. In 2000 he moved to a newly founded Fraunhofer group in Ilmenau/Thüringen, which became the Fraunhofer Institute for Digital Media Technology IDMT in 2004. He is currently head of "Perception and Ergonomics" at the department Acoustics and Deputy Director of the institute. His research topics include perceptual audio coding, subjective and objective assessment of audio quality, spatial audio, and techniques for the protection of multimedia data like scrambling and watermarking. Since 1999 he has taught at the Ilmenau University of Technology and being professor at the University of the Arts Berlin since 2010.

Prof. Dr.-Ing. Thomas Sporer has involved in the standardization efforts for perceptual audio measurement in ITU-R TG10/4 and EBU B/AIM. In addition, he is member of ITU-R WP6C, WP6B, SMPTE DC28, IEC TC100/TA11 and MPEG.