It sounds real when you see it. Realistic sound source simulation in multimodal virtual environments

Ágoston Török, Daniel Mestre, Ferenc Honbolygó, Pierre Mallet, Jean-Marie Pergandi, Valéria Csépe

It sounds real when you see it. Realistic sound source simulation in multimodal virtual environments

Ágoston Török[1,2,3], Daniel Mestre[4], Ferenc Honbolygó[1,3], Pierre Mallet[4], Jean-Marie Pergandi[4], Valéria Csépe[1,3]

Abstract Designing multimodal virtual environments promises revolutionary advances in interacting with computers in the near future. In this paper, we report the results of an experimental investigation on the possible use of surround-sound systems to support visualization, taking advantage of increased knowledge about how spatial perception and attention work in the human brain. We designed two auditory-visual cross-modal experiments, where noise bursts and light-blobs were presented synchronously, but with spatial offsets. We presented sounds in two ways: using free field sounds and using a stereo speaker set. Participants were asked to localize the direction of sound sources. In the first experiment visual stimuli were displaced vertically relative to the sounds, in the second experiment we used horizontal offsets. We found that, in both experiments, sounds were mislocalized in the direction of the visual stimuli in each condition (ventriloquism effect), but this effect was stronger when visual stimuli were displaced vertically, as compared to horizontally. Moreover we found that the ventriloquism effect is strongest for centrally presented sounds. The analyses revealed a variation between different sound presentation modes. We explain our results from the viewpoint of multimodal interface design. These findings draw attention to the importance of cognitive features of multimodal perception in the design of virtual environment setups and may help to open new ways to more realistic surround based multimodal virtual reality simulations.

Keywords Surround system · Ventriloquist illusion · Multisensory integration · Multilevel-modeling · Spatial audio

✉ Ágoston Török
torok.agoston@ttk.mta.hu

[1] Brain Imaging Centre, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2, 1117 Budapest, Hungary

[2] Doctoral School of Psychology, Eötvös Loránd University, Budapest, Hungary

[3] Department of Cognitive Psychology, Faculty of Pedagogy and Psychology, Eötvös Loránd University, Budapest, Hungary

[4] ISM UMR 7287, Aix-Marseille Université, CNRS, 13288 Marseille cedex 09, France

1 Introduction

For a long time, virtual reality designers and researchers have been using auditory stimulation to support visualization [1,2]. Even a single loudspeaker is sufficient to change the quality of experience, albeit this type of setup has at least one important limitation: it does not provide any spatial information, other than its own position. Spatial audio requires therefore a more complex approach, but the question is raised what is the minimal complexity that could provide the feeling of audio space (e.g. how many speakers are enough for the perception of a realistic sound environment). In the current study, instead of the traditional approach where researchers experimented with various sound generation methods (e.g. [3–5]), we investigated whether, in multimodal environments, a horizontal surround speaker setup was capable of effectively creating the illusion of two dimensional audio environment when perceived in the presence of visual objects. This illusion is called ventriloquism in the cognitive neuroscience literature [6]. The underlying mechanism is that the brain relies mostly on the visual loca-

tion of an object for its localization. Therefore even large distances between the auditory and visual stimulus locations (up to 20°) (e.g. [7]) are barely perceived by the user, even when s/he is warned of the possibility of discrepancies [8].

Concerning sound, the human brain uses binaural and monaural cues to localize sound sources [9]. Binaural cues are based on the fact that our ears are placed on both sides of the head. Therefore they receive auditory information from the same sound source at slightly different times (interaural time difference, ITD) and at different levels (interaural level difference, ILD). The duplex theory of hearing [9] states that sound localization is based on ITDs for low frequency sounds (under 1500 Hz), where phase differences are big enough to be perceived. For higher intensities, the shadowing effect of the skull serves as the basis for sound localization, attenuating the sound while it spreads through it (ILD). These two cues allow good localization in the horizontal plane (azimuth).

In the vertical plane however, sound localization is more difficult. Sounds on the medial plane cause no ITDs, because they are at the same angle and distance from both ears [9]. Vertical sound localization relies on the characteristics of the pinna, in that its shape and structure modifies the sound's spectrum as it reaches the inner ear. Nevertheless evidence shows that sound source localization in the vertical plane is poorer than that in the horizontal plane [9,10].

Currently, the most precise way of providing simulated spatial sounds is to use binaural recordings. If sounds from different spatial locations are recorded by microphones inserted into the ear canals of listeners, based on the recordings one can estimate a person's individual head related transfer function (HRTF, [3,5,11]). This HRTF can then be used to generate localized spatial sounds.

Although this is the most adequate way, it takes time and sophisticated equipment, and in most cases researchers and engineers use dummy heads (e.g. KEMAR, head acoustics) to simulate an individual's head. This approach has certain limitations; the ears of these dummy head models are made to be exactly identical, contrary to the human ears, where slight but significant differences certainly provide important cues for localization [9]. Setting up an accurate and individually tailored simulated binaural sound environment takes a lot of time and the process has to be repeated whenever a new person would like to experience it.

So far we have mentioned two ways for adding spatial sounds to a visual scene: while the one-speaker solution is cost-friendly and easy to use, it does not provide spatial information about the scene. Binaural recordings are rather expensive, difficult to setup and use but enable excellent spatial resolution. We can visualize these approaches on a two dimensional graph, where one dimension denotes simplicity of installation and use and the other the resolution of spatial information. In this scale the one speaker solution has high value on the easiness and low score on the spatial resolution dimension, while the binaural solution has the opposite values in both dimensions (see Fig. 1).

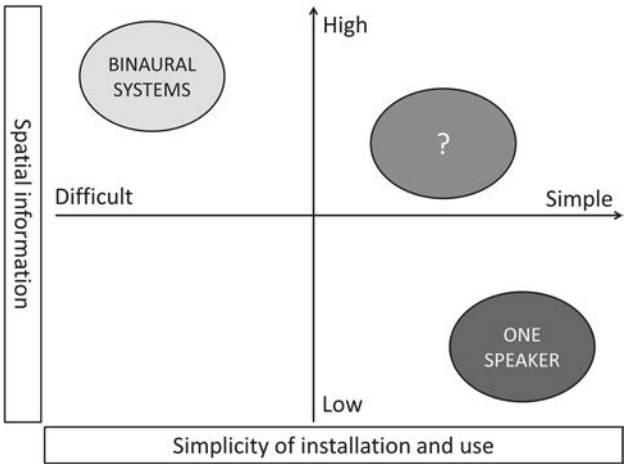

Fig. 1 Visualization of the two approaches in a two dimensional coordinate system, where one dimension is spatial information and the other is simplicity of installation and use. Binaural systems are difficult to set up properly, but provide a high degree of spatial information (localization of sound source). A single speaker is very easy to install, but does not provide spatial information other than its own position

The question arises whether there are any solutions having high values on both dimensions. If we take only audio, it is unlikely that there is a better possibility than an expensive installation, using a high number of speakers in all possible spatial directions (e.g. [4]). However, from the innovative point of view, price is also an important parameter of technical solutions. We suppose that there is a golden way to achieve the desired performance on a reasonable cost. In order to define this, however, one must take into account how human perception works in multimodal situations.

The human sensory systems are different in their respective performance and limits. Hearing is very good at detecting transient changes in our environment [12], but spatial localization based solely on audition is poorer than spatial localization based on vision. Vision, in contrast, is superior regarding spatial resolution, but often fails to detect transient changes in the environment [13]. Not only the senses themselves are different, but the brain is also adapted to their relative reliability in certain domains [14,15]. In multimodal situations the cognitive system combines information from different modalities in a weighted manner [16]. A well-known effect, called ventriloquism, demonstrates the relative importance and integration of auditory and visual information for spatial localization [10–17]. Ventriloquism refers to the phenomena where vision "captures" the apparent location of a concurrent sound. Sounds can be ventriloquized easily, without particular effort, that is to say without consciousness [18] or attentive access [19]. This effect is observed for offsets between the visual and auditory locations of stimuli up to 20° [7]. This could lead to the conclusion that the actual

spatial information of a sound has low weight in multimodal spatial localization and localization is mostly defined by the position of the visual stimulus component.

However sounds have an exquisite role in directing attention [20,21], so one of the greatest drawbacks of not having auditory spatial information is the weakness in directing the user's attention to certain places of the environment. The question could be raised whether this attention driving effect would be served adequately if sounds just approximated the exact location of the visual object by containing only horizontal information. As it was previously shown, the ventriloquism effect is even stronger in the vertical direction [10,22,23]. Taking this into account, one may assume that visual stimuli catch the vertical location of sounds in virtual environments with audio–visual properties. If this is true, a surround setup could be a golden mean between binaural and one speaker solutions, especially because surround speaker systems are easy to install and broadly available in the consumer market. Moreover, this kind of audio stimulation is readily available in most VR labs. Thus it would be important to see how realistic these audio–visual environments are for the human perception, and not from an objective point of view. Our current study aimed to serve this purpose.

Our hypotheses were the following:

1. Sounds can be ventriloquized in the vertical plane, therefore it is not necessary to provide vertical auditory spatial cues, and a sound system of good horizontal resolution is enough to provide a realistic audio–visual environment.
2. Sounds can be ventriloquized in the horizontal plane, thus a small mismatch or slight scarcity of sound simulation (e.g. because of asymmetric room reverberation characteristics) does not lead to measurable changes in perception.
3. In multimodal situations, sound source localization in environments using surround systems is as good as in environments using free field speakers.
4. The ventriloquist effect differs in the horizontal and vertical plane when using surround speakers.

In order to test these hypotheses, we designed two experiments in a Cave Automatic Virtual Environment (CAVE) setup, installed in the Mediterranean Virtual Reality Center (http://www.crvm.eu). Participants were asked to locate sound sources occurring with or without simple visual stimuli (Gaussian blobs). The paradigm was based on that reported in the study of [24]. Sound sources were either free field speakers (free field condition) or their simulated copies delivered through a stereo speaker set (surround condition). We used left, middle and right sound directions to test whether there was any difference in the ventriloquism effect, depending on from where the participants heard the sound. Visual stimuli were placed on the vertical plane in the first experiment and on the horizontal plane in the second experiment.

2 Methods

2.1 Participants

All participants were recruited as volunteers from Aix-Marseille University, Marseilles, France. 6 participants (1 female, mean age 32.4 years, min 25 years, max 48 years) took part in the (1st) and 5 (1 female, mean age 27 years, min 21 years, max 41 years) in the (2nd) experiment. Participants had normal hearing and had normal or corrected to normal vision. Because of population variability in the ability for stereoscopic perception, participants were tested for stereo vision using stereoscopic random dot figures (Randot Stereotests, Stereo Optical Co.). Stereo vision was adjusted for each participant based on their interocular distances. Every participant gave written informed consent prior to the experiment. Each participant took part in one experiment. They did not receive any compensation for the experiments. The study involved exclusively non-invasive perceptual measurements, and was approved by the Institute of Movement Science Laboratory Review Board. The experiment was conducted in accordance with the Declaration of Helsinki.

2.2 Apparatus

The experiment took place in a dimly lit hall designed for the virtual reality equipment. The walls were painted black and the hall had no windows. The CAVE had a set of four screens; a 3×4 m frontal display, two 3×4 m lateral displays and a 3×3 m fiber optic screen on the ground. Participants sat in a comfortable chair at a 1.2 m distance from the frontal screen with their eye-level at about 1.15 m from the ground level of the CAVE. For the experiments we defined our setup so that all visual stimuli were on the frontal screen, this way we avoided any bias caused by brightness transitions on the edges of the screens. The frontal screen's resolution was 1400×1050 pixels. Visual stimuli were light-blobs (visual angle 7.6°) with a Gaussian envelope. Blobs were presented for 16.67 ms (one frame). The baseline luminance of the screen was 0.006 $\mathrm{cd/m^2}$ and the luminance of the visual stimuli was 0.35 $\mathrm{cd/m^2}$. Participants wore passive stereo-glasses (Infitec) and the projectors used static stereo image rendering. It should be noted that, in the case of such short exposures and the kind of visual stimuli used, no actual conscious depth perception occurs. At the beginning of each trial, a fixation cross appeared on the screen in the center, at 1.1 m height.

The acoustic stimuli of 16.67 ms duration were broadband noises high pass filtered at 250 Hz. Sounds were delivered via

seven identical speakers of a 7.1 surround system (Creative Inspire p7800), sound pressure level at the participants position was 65 dB SPL. Speakers were placed on a 2.99 m radius circle with its center at 0.76 m from the ground level of the CAVE in the participants' position. The seven speakers were placed 10.5° from each other, with speakers 2 (−21°), 4 (0°) and 6 (+21°) as free field speakers, and speakers 1 and 7 for the surround condition to simulate speaker position 2, 4 and 6. Speakers 3 and 5 were not used in the current study. We used panning (inter-speaker sound level differences) to create the stereo sounds projected at the position of the free field speakers, sound levels were matched between the free field and surround conditions. The participants used a Flystick (ART Flystick 2) to respond. The flystick's position was logged by infrared cameras (ART), this way participants could easy and naturally locate sounds. The 3D orientation of the flystick was used as an indication of the perceived sound direction.

2.3 Procedure

Each trial began with a fixation cross, participants were asked to move the cursor with the flystick to the fixation cross, to ensure that at the start of each trial their hand was in the same position, and that they fixated the central cross. The fixation cross disappeared after 1000 ms and the test stimulus occurred with a 50 ms onset delay. Each trial consisted of either single auditory or audio–visual stimulus presented for 16.67 ms followed by 420 ms blank screen and then an "x" sign as participants used the flystick's cursor as a response tool by the participants. According to the instruction given, they had to respond as accurately as possible by moving the cursor to the location of the auditory stimuli. A new trial started after they responded. The participants never had to respond to the visual stimuli, but were asked to always keep their eyes open till the end of the experiment.

In Experiment 1, visual stimuli had a vertical offset relative to sounds. There were three sound positions (left, middle, right), two sound types (free field, surround) and six visual positions (no visual, −21°, −10.5°, 0° (same position as the sound), 10.5°, 21° relative to the sounds on the vertical axis).

In Experiment 2 a similar design was used, only the six visual positions were spread out horizontally, not vertically.

Figure 2 illustrates the possible stimulus presentations for the left sound position in the two experiments. In both experiments each trial was repeated 20 times, making a total of 750 presentations. The experiments lasted 1 h with one or, if participants needed, two 5 min long breaks.

2.4 Data analysis

We first inspected the data for outliers. We rejected every response where RT was less than 300 ms or more than 4000 ms. Due to significant time uncertainty (variable delay)

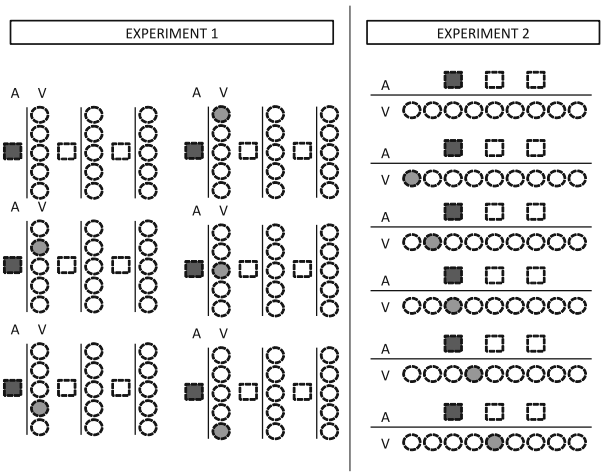

Fig. 2 Possible stimulus presentation sets for a left sound. *Dark squares* indicate the sound positions and *grey circles* mark the place of the synchronously presented visual stimuli. In the figure, audio and visual stimuli positions are presented with a *separator line* for illustrative purposes

caused by the computer cluster system, we used response times just for filtering. After the removal of outliers, on average 92 % of each participant's data in Experiment 1 and 82 % in Experiment 2 remained and were entered in the analyses. Errors as projection of response bias in the direction of the distractor stimulus were also measured. Multimodal errors (perceived shift) were calculated for each condition relative to its respective average perceived unimodal response direction. We used this approach because some participants tended to mislocalize sounds on the vertical plane, thus analyzing relative bias to veridical sound positions would distort our results [25]. Since our data were collected from a relatively small sample (six and five participants) and sphericity was violated, our data structure is not well suited for standard ANOVA analyses [26]. The averaging of responses across conditions also caused considerable data loss, so we decided to run multilevel modeling, to be able to deal with possible differences between individuals and with the whole range of responses. We used sound type (real, surround); sound direction (left, center, right); visual stimulus direction (1–5) as fixed factors and participants' ID as random effect in the models, we used Restricted Maximum Likelihood (REML) estimation in SPSS.

3 Results

3.1 Results of Experiment 1

We built five models to explore the underlying effects in our study. First we built an empty model where we entered only the Participants' ID as a random factor. It was also the baseline to evaluate further models, so we compared

the information criteria to the initial model's [Schwarz's Bayesian Criterion (BIC) = 26,079.46]. Wald Z test for the $\tau 00$ was not significant, indicating that there was no systematic difference between subjects. The residual variance ($\sigma 2$) was high enough to reach significance (Wald Z = 41.49; p < 0.001), meaning that it was justified to enter individual level predictors in the model. Therefore we entered sound type, sound direction, and visual stimulus direction in the model as fixed factors. We handled visual stimulus direction as scalar variable, sound type and sound direction as dummy variables. We used this approach because prior inspection of raw data indicated a possible difference between the center and the two eccentric positions. Comparison of the model estimating linear fit and nominal (dummy) estimates favored the handling of Sound direction as nominal variable (BIC: 22,348.23 < 22,376.90).

First we only took the main effects into consideration. We ran the model with variance components specifications because the covariance was too small to estimate [27]. This model had far better goodness of fit values (BIC = 22,348.23) and revealed a significant effect for Visual stimulus direction. Linear fitting to the data showed a 5.18° difference in pointing response bias from level to level of Visual stimulus direction (see Fig. 3). The model's intercept was the estimated bias for the surround speaker set simulating a sound in the right position presented with a light blob in the downmost position. We did not find significant variations of Visual stimulus direction factor between subjects. Based on the estimations and their standard errors it can be said that localizations were clearly affected by the visual distractors, and localization bias showed linear connection with visual stimulus distance.

In order to take into account interactions that may modulate the results, we entered first all two ways and then the possible three way interaction in our model. These effects brought further improvement (BIC_{2way} 22,241.07, BIC_{3way} 22,240.24; see supplementary Table 1) to our model, revealing a significant interaction between sound direction and sound type (F (2, 3414.00) = 26.25; p < 0.001) and between Sound direction and Visual stimulus direction (F (8, 3414.00) = 5.90; p < 0.001). An overall lower localization of sounds for the real left (p < 0.01) and for the real central sound sources (p < 0.001) were behind the sound direction and sound type interaction. The interaction of visual stimulus direction and sound direction was caused by a significantly (p < 0.001) stronger visual capture in the central presentations (see Fig. 3). Because the two way interaction of Sound type and Visual stimulus direction and the three way interaction were not significant (p > 0.7) we left them out from our final model. This model had better information criteria than the one containing only main effects (BIC 22,239.02). Summary of the final model can be seen in Table 1.

Summarizing the results of Experiment 1, we found strong visual capture for all sound positions, especially for central sounds. We also found a slight difference in sound localization errors between Real and Surround conditions, although the effect of visual capture did not differ between sound gen-

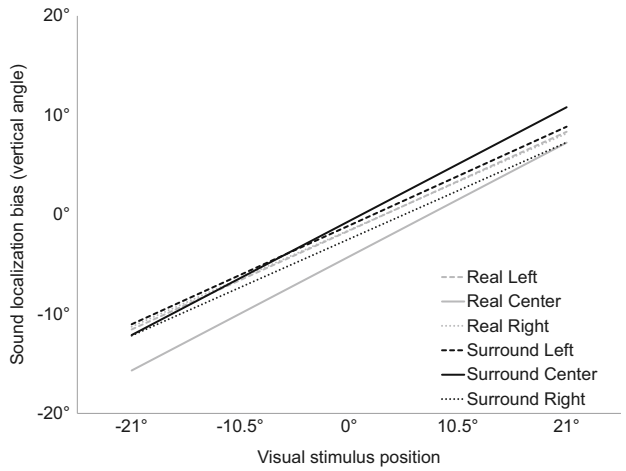

Fig. 3 Estimated attraction for each level of sound type and direction. The strong capture of visual stimuli is visible in the graph. Free-field sounds especially in the central position are localized lower

Table 1 Summary of the final model in Experiment 1

Fixed effects	Est	SE
Intercept (surround.right.1)	−17.07*	5.16
Sound type (real)	0.92	0.62
Sound type (surround)	0.00	0.00
Sound direction (left)	1.04	0.94
Sound direction (center)	−0.82	0.95
Sound direction (right)	0.00	0.00
Visual stim dir (1–5)	4.86*	1.36
Sound type (r) × sound direction (l)	−1.45**	0.49
Sound type (r) × sound direction (c)	−4.50***	0.50
Sound direction (l) × visual stim dir	0.11	0.17
Sound direction (c) × visual stim dir	0.87***	0.18
Random effects		
Residual	35.45***	0.86
Intercept (subject = ID)	156.18	99.55
Sound type (subject = ID)	0.77	0.57
Sound direction (subject = ID)	1.49*	0.75
Visual stim dir (subject = ID)	11.06	7.02
Fit statistics		
AIC	22,208.31	
BIC	22,239.02	

Est estimate, *SE* standard error, *AIC* akaike information criterion, *BIC* Schwarz's Bayesian information criterion; visual stim dir: 1, downmost; 5, upmost; *l* left, *c* center, *r* real

* p < 0.05

** p < 0.01

*** p < 0.001

eration types. Overall real sound sources especially in the center were localized lower, possibly due to an intrinsic variation of responses.

3.2 Results of Experiment 2

We ran a multilevel modeling on the data of Experiment 2 using the same factors as in Experiment 1. First we built an empty model as our baseline. We entered visual stimulus direction, sound direction and sound type in the second model both as fixed and as random effects. We built a model where sound direction was entered as an ordinal variable and one where it was a three level nominal variable. The later had better fit criteria (BIC 17,534.72 < 17,761.48) so we decided to use it as nominal. Model 2 had better fit parameters to the data, and revealed a tendentious effect of Visual stimulus direction (estimation 2.41, SE 1.083; see Fig. 4). Besides the fixed effect of visual stimulus direction, the random effect of Sound direction was significant ($p < 0.05$), indicating that the effect of sound direction differed between subjects.

We entered first the two way interactions in the third and then all interactions in the fourth model. The fourth model had better fit parameters than the third. The three way interaction was not significant, but two way interactions revealed that visual capture is weakest for the right sounds (estimate 1.53, SE 1.12), slightly stronger for left sounds (estimate 2.22, SE 1.46) and the strongest for sounds in the center (estimate 3.38, SE 1.46; supplementary Table 2).

Because neither the three way interaction nor the interaction between sound type and visual stimulus direction were significant we decided to leave them out of our final model. This model had the best fit parameters to the data. The effect of visual stimulus direction was not significant, but the interaction between visual stimulus direction and sound direction

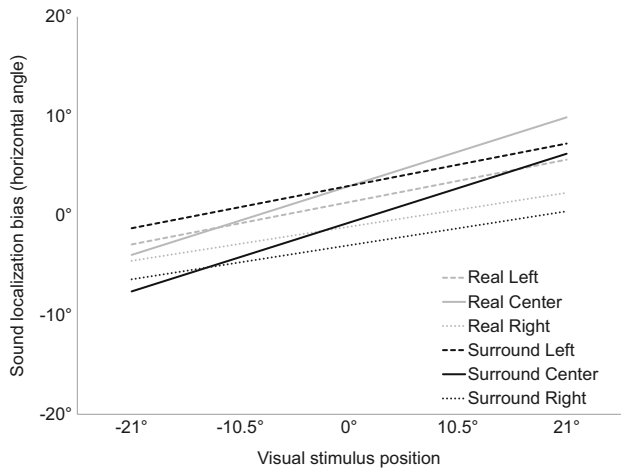

Fig. 4 Estimated attraction for each level of sound type and direction. Strong capture of visual stimuli is visible only for central sounds. Surround sounds in the center are localized more to the left

was significant again ($l_{eft} < 0.1$ (0.083); $c_{enter} < 0.001$). The significant interaction of visual stimulus direction and sound direction meant that the effect of visual capture was most salient for central sounds, regardless of sound type. Interaction between sound type and sound direction indicated that simulated and real sound sources were perceived somewhat differently. The random effect of sound direction was also significant, so sound localization was variable between subjects. Table 2 summarizes the parameter estimates of our final model.

Summarizing the results of Experiment 2, we found that visual capture was strongest in the center, and this effect was not different for real and surround sounds. As in Experiment 1, we found that participants localized real and surround sounds differently in bimodal situations in the horizontal plane, however no clear trend was apparent, rather a slight variance in the perception of sound sources.

4 Discussion

In the present experiments, we investigated the audio source localization ability of 11 subjects by measuring their per-

Table 2 Summary of the final model in Experiment 2

Fixed effects	Est	SE
Intercept (surround.right.1)	−8.12*	4.32
Sound type (real)	1.84+	0.82
Sound type (surround)	0.00	0.00
Sound direction (left)	4.71	3.30
Sound direction (center)	−2.98	3.31
Sound direction (right)	0.00	0.00
Visual stim dir (1–5)	1.71	1.10
Sound type (r) × sound direction (l)	−3.47***	0.68
Sound type (r) × sound direction (c)	1.83**	0.69
Sound direction (l) × visual stim dir	0.42+	0.24
Sound direction (c) × visual stim dir	1.75***	0.25
Random effects		
Residual	50.33***	1.41
Intercept (subject = ID)	64.81	52.78
Sound type (subject = ID)	1.11	0.93
Sound direction (subject = ID)	25.37*	12.82
Visual stim dir (subject = ID)	5.93	4.23
Fit statistics		
AIC	17,393.01	
BIC	17,422.23	

Est estimate, *SE* standard error, *AIC* akaike information criterion, *BIC* Schwarz's Bayesian information criterion; visual stim dir: 1, leftmost; 5, rightmost; *l* left, *c* center, *r* real
* $p < 0.05$
** $p < 0.01$
*** $p < 0.001$

formance in multimodal situations. We aimed to evaluate the usability of surround systems in supporting visualization and creating realistic perceptual situations. We first hypothesized that sounds can be ventriloquized on the vertical plane. In order to investigate this, in the first experiment we looked at how well participants localized sound sources when they occur with synchronous, but vertically displaced visual distractors. We found that the visual distractor positions greatly affected the subjects' localization judgments, but the effect was slightly stronger for centrally presented sounds. This effect was persistent both in free field speaker and in surround situations. Our second hypothesis was that horizontal ventriloquist illusion could compensate sound source simulation scarcity. Thus, in the second experiment we presented sounds and visual distractors with horizontal offsets. We found that the participants' judgment of sound source locations was affected by visual distractors. Furthermore, the effect of visual distractors was greater for sounds in the center.

The third hypothesis stated that sound source localization is not different between surround and free field speakers. In contrast, we observed slight differences between surround and free field speakers in both experiments. One likely explanation for the variance is that there was some difference in the speakers' characteristics. Alternatively, the asymmetry in the reverberation structure of the experimental hall could alter the reverberation properties of sounds. Because we used identical speakers and sound levels were measured for each speaker separately, it is more likely that the asymmetry of the experimental hall contributed to the differences in localization. This further highlights the importance of visual capture and multimodal stimulation to prevent perceptual changes caused by imperfection of sound source modeling and rendering.

Handling of visual stimulus positions as scalar variables led to a better fit than using them as dummy variables, and the effect of visual stimuli depended on the sound's direction. We found the strongest capture in both experiments for central sound sources. One possible explanation for this is that the shifts in sound source position judgments were strongly affected by the size of the visual stimulus' cortical representation. Visual stimuli closer to the focus (and so the fovea) have greater cortical representation [28]. ITDs and ILDs are more pronounced farther away from the medial plane [9]. Participants reported similar impressions in the debriefing after the experiments.

The observed effects are comparable to those of [22,24, 29]. Although our methodology was based on these earlier studies, important differences exist. References [24–29] used only one movable near sound source (cca. 50 cm) in a soundproof chamber whereas in our experiments sound sources were much farther (3 m) away from the viewpoint in a reverberating hall. This difference is even more important since near and far sound sources are localized differently [30]. This could be also important when we consider why the effects were different for sound directions. Another important difference is that while [24] used a led array as visual stimuli, positioned at the distance of the sound sources, in our case Gaussian blobs were projected to the frontal screen at a distance of 1.5 m from the participant's viewpoint. Moreover the screen was not curved, but the blobs were stereo-projected to a virtual sphere at 2.99 m from the viewpoint. The last important difference was that in contrast to the earlier studies we allowed participants to respond freely both horizontally and vertically simply by moving their hands, this way we could avoid artifacts caused by unnatural response methods, such as choice from a button array, or button rotation.

Our fourth hypothesis claimed that vertical and horizontal ventriloquist effects are different in size. We decided not to compare the data of the experiments in one analysis, because the random effects in the models were different for the two experiments indicating sample variability, although indirect comparison is possible. The fact that MLM showed more consistent effect of visual stimuli for vertical arrangements indicates that visual capture is stronger in the vertical plane. Earlier, with different methodology, [10] reached very similar conclusions.

The present study has certain limitations. Based on the methodology we used, we cannot decide whether the sounds were really perceived close to the visual stimuli or the effect was caused by post perceptual response strategies. After the experiments, the participants reported that they felt sometimes that sounds and flashes were coming from elsewhere. The fact that responses fell between the visual and sound positions and were not centered at the place of the visual position might mean that the participants did try to locate the sounds and not simply chose the position of the visual stimuli. Our methodology was based on standard ventriloquism paradigms, which were also affected by this criticism [6]. However, there are other studies showing that the ventriloquist effect occurs in non-transparent (i.e. where the discrepancy is so little that it is not possible to differentiate consciously the audio and visual signal's location) paradigms as well, so it cannot be solely defined by response strategies. It is also important to note that the brain responses elicited by ventriloquized and non-ventriloquized sounds differ at early processing stages [31]. A preattentive brain response, the mismatch negativity observed in EEG studies, is sensitive to the ventriloquist effect [7,19], further supporting the suggestion that response strategies do not cause the effect by themselves. Although the above limitations must be kept in mind, it is highly unlikely that the effects found can be fully attributed to conscious, decision-related processes.

Our research fits well within the scope of cognitive infocommunications [32]. It shows that our brain does not just co-evolve with infocommunication devices but we can use

the specific features of our cognitive system in the design of better and/or cheaper infocommunication tools. Similar approaches, already present in the literature, demonstrated how perceptual based illusions can benefit multimodal user interfaces [33]. Furthermore, we think that our approach shows a channel where knowledge from explorative psychophysical research could flow into the field of applied cognitive ergonomics.

To know more about how multimodal integration works in virtual reality, further studies are needed, utilizing brain imaging and electrophysiological methods. The question of how the brain perceives virtual environments is already a major topic in neuroscience research [34,35]. However, studies involving recordings of brain activity in interactive conditions are mostly lacking, mainly due to signal processing limitations.

To sum-up, in the present experiments we found that (1) the ventriloquist effect works in virtual reality, (2) sounds can be ventriloquized both vertically and (3) horizontally, and (4) there is a slight deterioration in the sound source position judgments when using surround system compared to free field speakers. In conclusion, researchers and virtual reality designers should be able to use surround systems efficiently in the generation of realistic audio–visual scenes, because the human perceptual system is well adapted to the experienced mismatches in audio and visual positions. This is important also for the design of multimodal user interfaces [36] where the scope is to build an intuitive and natural UI that frees up the visual system [20], with the help of audio stimulation that directs and lowers the load on the attentional system.

Acknowledgments The research leading to these results has received funding from the European Community's Research Infrastructure Action—grant agreement VISIONAIR 262044—under the 7th Framework Programme (FP7/2007-2013). Ágoston Török was supported by a Young Researcher Fellowship from the Hungarian Academy of Sciences. Thanks to Dénes Tóth for his help during the statistical analyses. Thanks to Orsolya Kolozsvári for her help in the preparation of the manuscript.

References

1. Nacke LE, Grimshaw MN, Lindley CA (2010) More than a feeling: measurement of sonic user experience and psychophysiology in a first-person shooter game. Interact Comput 22(5):336–343
2. Zhou Z, Cheok AD, Yang X, Qiu Y (2004) An experimental study on the role of 3D sound in augmented reality environment. Interact Comput 16(6):1043–1068
3. Hu H, Zhou L, Ma H, Wu Z (2008) HRTF personalization based on artificial neural network in individual virtual auditory space. Appl Acoust 69(2):163–172
4. Seeber BU, Kerber S, Hafter ER (2010) A system to simulate and reproduce audio-visual environments for spatial hearing research. Hear Res 260(1–2):1–10
5. Wenzel EM, Arruda M, Kistler DJ, Wightman FL (1993) Localization using nonindividualized head-related transfer functions. J Acoust Soc Am 94(1):111–123
6. Vroomen J, Gelder B De (2004) Perceptual effects of cross-modal stimulation: ventriloquism and the freezing phenomenon. Handb Multisens Process 3(1):1–23
7. Colin C, Radeau M, Soquet A, Dachy B, Deltenre P (2002) Electrophysiology of spatial scene analysis: the mismatch negativity (MMN) is sensitive to the ventriloquism illusion. Clin Neurophysiol Off J Int Fed Clin Neurophysiol 113(4):507–518
8. Török Á, Kolozsvári O, Virágh T, Honbolygó F, Csépe V (2013) Effect of stimulus intensity on response time distribution in multisensory integration. J Multimodal User Interfaces 8(2):209–216. doi:10.1007/s12193-013-0135-y
9. Middlebrooks JC, Green DM (1991) Sound localization by human listeners. Annu Rev Psychol 42:135–159
10. Thurlow WR, Jack CE (1973) Certain determinants of the "ventriloquism effect". Percept Motor Skills 36(3):1171–1184
11. Raykar VC, Duraiswami R, Yegnanarayana B (2005) Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. J Acoust Soc Am 118(1):364–374
12. Shams L, Kamitani Y, Shimojo S (2000) Illusions. What you see is what you hear. Nature 408(6814):788
13. Henderson RM, Orbach HS (2006) Is there a mismatch negativity during change blindness? Neuroreport 17(10):1011–1015
14. Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14(3):257–262
15. Ohshiro T, Angelaki DE, DeAngelis GC (2011) A normalization model of multisensory integration. Nat Neurosci 14(6):775–782
16. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PloS One 2(9):e943
17. Howard IP, Templeton WB (1966) Human spatial orientation. Wiley, New York
18. Bertelson P, Aschersleben G (1998) Automatic visual bias of perceived auditory location. Psychon Bull Rev 5(3):482–489
19. Stekelenburg JJ, Vroomen J, de Gelder B (2004) Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. Neurosci Lett 357(3):163–166
20. Brewster S (1998) The design of sonically-enhanced widgets. Interact Comput 11(2):211–235
21. McDonald JJ, Ward LM (2000) Involuntary listening aids seeing: evidence from human electrophysiology. Psychol Sci 11(2):167–171
22. Hartnagel D, Bichot A, Roumes C (2007) Eye position affects audio-visual fusion in darkness. Perception 36(10):1487–1496
23. Werner S, Liebetrau J, Sporer T (2013) Vertical sound source localization influenced by visual stimuli. Signal Process Res 2(2)
24. Besson P, Richiardi J, Bourdin C, Bringoux L, Mestre DR, Vercher J-L (2010) Bayesian networks and information theory for audio-visual perception modeling. Biol Cybern 103(3):213–226
25. Wozny DR, Shams L (2011) Recalibration of auditory space following milliseconds of cross-modal discrepancy. J Neurosci Off J Soc Neurosci 31(12):4607–4612
26. Hoffman L, Rovine MJ (2007) Multilevel models for the experimental psychologist: foundations and illustrative examples. Behav Res Methods 39(1):101–117
27. Heck RH, Thomas SL, Tabata LN (2010) Multilevel and longitudinal modeling with IBM SPSS (Google eBook). Taylor Francis, New York
28. Rovamo J, Virsu V (1979) An estimation and application of the human cortical magnification factor. Exp Brain Res 37(3):495–510
29. Besson P, Bourdin C, Bringoux L (2011) A comprehensive model of audiovisual perception: both percept and temporal dynamics. PloS One 6(8):e23811
30. Moore DR, King AJ (1999) Auditory perception: the near and far of sound localization. Curr Biol 9(10):R361–R363

31. Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, Heinze H-J, Hillyard SA (2007) Neural basis of the ventriloquist illusion. Curr Biol CB 17(19):1697–1703
32. Baranyi P, Csapo A (2012) Definition and synergies of cognitive infocommunications. Acta Polytech Hung 9(1):67–83
33. Lee J-H, Spence C (2009) Feeling what you hear: task-irrelevant sounds modulate tactile perception delivered via a touch screen. J Multimodal User Interfaces 2(3–4):145–156
34. Haans A, IJsselsteijn WA (2012) Embodiment and telepresence: toward a comprehensive theoretical framework. Interact Comput 24(4):211–218
35. Kober SE, Kurzmann J, Neuper C (2012) Cortical correlate of spatial presence in 2D and 3D interactive virtual reality: an EEG study. Int J Psychophysiol Off J Int Organ Psychophysiol 83(3):365–374
36. Ghirardelli TG, Scharine AA (2009) Auditory–visual interactions. In: Letowski, El Michael B, Russo tomasz R (eds) Helmet-mounted displays: sensation, perception, and cognition issues. U.S. Army Aeromedical Research, Fort Rucker, pp 599–618