

# Audio, visual, and audio-visual egocentric distance perception in virtual environments.

Marc Rébillat, Xavier Boutillon, Etienne Corteel, Brian F,g Katz

▶ **To cite this version:**

Marc Rébillat, Xavier Boutillon, Etienne Corteel, Brian F,g Katz. Audio, visual, and audio-visual egocentric distance perception in virtual environments.. Forum Acusticum, Jun 2011, Aalborg, Denmark. pp.482. hal-00619317

**HAL Id: hal-00619317**
**https://hal.science/hal-00619317**

Submitted on 6 Sep 2011

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Forum Acusticum 2011 - Aalborg - Denmark

# Audio, visual, and audio-visual egocentric distance perception in virtual environments

Marc Rébillat[1,2], Xavier Boutillon[2], Étienne Corteel[3], Brian F.G. Katz[1]
(1) LIMSI-CNRS, Université Paris Sud, France
(2) LMS, École Polytechnique, France
(3) *sonic emotion labs*, France

**Summary**
Previous studies have shown that in *real* environments, distances are *visually* correctly estimated. In *visual* (V) *virtual* environments (VEs), distances are systematically underestimated. In *audio* (A) *real* and *virtual* environments, near distances (< 2m) are overestimated whereas far distances (>2 m) are underestimated. However, little is known regarding combined A and V interactions on the egocentric distance perception in VEs. In this paper we present a study of A, V, and AV egocentric distance perception in VEs. AV rendering is provided via the SMART-I$^2$ platform using tracked passive visual stereoscopy and acoustical wave field synthesis (WFS). Distances are estimated using triangulated blind walking under A, V, and AV conditions. Distance compressions similar to those found in previous studies are observed under each rendering condition. The *audio* and *visual* modalities appears to be of similar precision for distance estimations in virtual environments. This casts doubts on the commonly accepted *visual capture* theory in distance perception.

**PACS no. 43.66.Qp, 42.66.Si**

## 1. Introduction

Virtual reality aims at providing users with a virtual world where they would behave and learn as if they were in the real world [1]. Correct distance perception in virtual environments is thus crucial for many virtual reality applications.

Because distance perception is a cognitive task, measurement protocols are needed to estimate the distance perceived by a participant. Existing measurements protocols for absolute egocentric distance can be divided in 3 main classes [2, 3]: verbal estimations, perceptually directed actions, and imagined actions. In *verbal estimation* protocols, participants assess perceived distance in terms of familiar units, such as meters. In *perceptually directed action* protocols, an object is presented to the participant who then has to perform an action, such as blind-walking, without perceiving the object any more. In *imagined action* protocols, the action is imagined instead of being performed and response times or other measures are used to infer the results of the action. The advantage of perceptually directed actions over other protocols is that the observer's perception of distance can be directly inferred from the action [2].

Using these protocols, numerous recent studies have focused on *visual* distance perception in *virtual* environments involving large screens [2–6]. These studies found that *visual* distances are systematically underestimated. In *audio virtual* environments [7–13] near distances (< 2m) are overestimated whereas far distances (>2 m) are underestimated. Little is known regarding combined *audio* and *visual* interactions on egocentric distance perception in the *real* world [14–17] and even less in *virtual* environments [18].

In this paper a study of *audio* (A), *visual* (V), and *audio-visual* (AV) egocentric distance perception in *virtual* environments is presented. AV rendering is provided via the SMART-I$^2$ platform[1] [19, 20] using tracked passive visual stereoscopy and acoustic wave field synthesis (WFS) [21]. Distances are estimated using a perceptually directed action (indirect blind walking [2, 22–25]) under A, V, and AV conditions.

The question to be discussed in the study is: *What kind of cross-modal interaction in audio-visual stimuli exists for distance perception in virtual environments?*

(c) European Acoustics Association

[1] SMART-I$^2$: Spatial Multi-user Audio-visual Real-Time Interactive Interface

## 2. Method

### 2.1. Experimental design

The goal of the current experiment was to study distance perception under three different rendering conditions (audio, visual, audio-visual). For that purpose, 5 distances in the *action space*, *i.e.* the space where "*we move quickly, talk, and if needed can throw something to a compatriot or at an animal*" [26]. were tested: 1.5 m, 2 m, 2.7 m, 3.5 m, 5 m.

A total of 15 volunteers between 24 and 45 years old participated in the experiment (12 men, 3 women). All participants reported normal vision (eventually corrected) and normal hearing. Each participant had to estimate the 5 distances 4 times under each rendering condition. They performed three sessions of 20 iterations each after a training phase of 2 iterations under each rendering condition. Participants had pauses between sessions and the whole experiment lasted approximately one hour. To limit the influence of learning effects, order of the sessions were balanced between the 6 possible permutations of the 3 conditions.

The chosen experimental design was therefore a repeated-measures design with two factors: the real distance $d_r$ (5 levels, fixed factor) and the rendering condition (3 levels, fixed factor). Tab. 2.1 summarises the chosen experimental design.

| *Independent variables* | | |
|---|---|---|
| Participant | 15 | Random variable |
| Rendering condition | 3 | A, V, AV |
| Rendered distance $d_r$ (m) | 5 | 1.5, 2, 2.7, 3.5, 5 |
| Repetition | 4 | 1, 2, 3, 4 |

| *Dependent variables* |
|---|
| Perceived distance $d_p$ (m) |

Table I. Independent and dependent variables used in the experimental protocol

### 2.2. Experimental setup

Experiments have been conducted in the audio-visual environment produced by the SMART-I$^2$ platform [19, 20]. In this system, front-projection screens and loudspeakers are integrated together to form large flat multi-channel loudspeakers also called *Large Multi-Actuator Panels* (LaMAPs). The rendering screens consist of two LaMAPs (2 m $\times$ 2.6 m with each supporting 12 loudpeakers) forming a corner (see Fig. 1).

Visual rendering is produced using tracked passive stereoscopy at 80 frames per second (in the GPU) and with a resolution of 1280 $\times$ 960 pixels on each screen. At the starting position (the black «×» in Figs. 2 and 3), the horizontal field of view is approximately 150 $^\circ$ and the vertical field of view approximately 70 $^\circ$. Graphical resolution [25, 27] and field of view [28] have been shown not to have an influence on *visual* distance perception.

Spatial audio rendering is realised using wave field synthesis [21]. This technology allows one to recreate physically the sound field corresponding to a virtual source at any given position in the horizontal plane, without the need for tracking. The inter-loudspeaker distance of 21 cm leads to an aliasing frequency $f_{al} \simeq 2$ kHz, above which the sound field is not spatially correctly reconstructed [10].

Furthermore, fine temporal and spatial calibration has been performed to ensure that the audio and visual renderings are coherent.

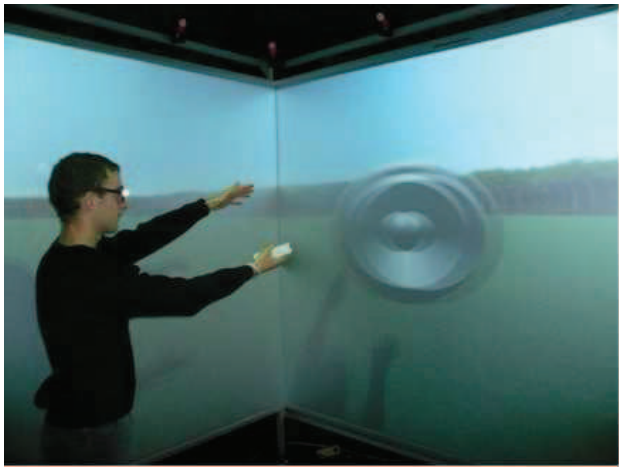


Figure 1. A participant and an audio-visual object in the virtual world. Visual rendering is front-projected on the front faces of the two LaMAPs which form a corner.

### 2.3. Audio-visual stimuli

The *visual environment* consists of an open, grassy field with a forest in the background (see Fig. 1). The associated *audio environment* consists of the sound of wind in the trees accompanied by some birds songs in the distance (overall level of 36 dBA). Environmental sound level has been adjusted to be slightly above the background noise produced by the video-projectors (34 dBA).

The chosen *visual target* object is a footless 3D loudspeaker, approximately spherical, with a diameter of $\simeq$ 30 cm (see Fig. 1). Its foot has been removed to avoid window violation when the object is displayed in front of the screen. The floating loudspeaker is positioned at a height of 1.6 m and shadows are displayed.

The associated *audio target* object is a 4 kHz low-pass filtered white noise with a 15 Hz amplitude modulation. White noise has been chosen in order to have a wide spectral content and to allow subjects to rely on numerous audio localisation cues. The 4 kHz low-pass filtering has been performed to limit energy

above the aliasing frequency $f_{al} \simeq 2$ kHz and thus to limit possible incoherent spatial cues. In practice, it has been shown in [29] that even if not *exact*, spatial cues above $f_{al}$ were generally coherent with spatial cues below $f_{al}$ when using MAPs. This white noise has been modulated in amplitude to produce attack transients that are also useful in sound localisation. No room-effect (*i.e.* ground reflections) has been included. The sound level of the audio object corresponds to a monopole emitting 78 $dB_{SPL}$ when placed at 1 m and is thus well above the environmental sound level at each of the tested distances.

*Audio* and *visual* objects are always displayed coherently, *i.e.* at the same spatial position. In addition their *visual* size and *audio* level decrease naturally with distance. Participants are allowed to move within the rendering area. They can thus rely on all the cues naturally available in the corresponding real environment for distance estimations. These available audio-visual cues are summarised in Tab. II.

| Available cue | Modality | Class |
|---|---|---|
| Object size/level | A, V | Relative |
| Motion parallax | A, V | Absolute |
| Binocular/binaural cues | A, V | Absolute |
| Height in the visual field | V | Absolute |

Table II. Audio-visual cues available for distance estimation in the proposed experimental setup.

The audio-visual background environment was active in all the rendering conditions. In the *audio* condition, the spatialised sound corresponding to the virtual object was played with no image of the virtual object. The only visual image consisted of the open, grassy field with a forest in the background. In the *visual* condition, the 3D image of the virtual object was displayed with no corresponding sound. The only audio signal consisted of the sound of wind in the trees accompanied by some bird songs. In the *audio-visual* condition, the spatialised sound of the virtual object was rendered with its corresponding 3D image and the audio-visual environment.

### 2.4. Experimental task

Distance estimation is performed here in two phases: an *exploration* phase (see Fig. 2) and a *triangulation* phase (see Fig. 3). Participants start each iteration at the *start position*, indicated by a black «×» in Figs. 2 and 3. Virtual objects are displayed on the right screen of the SMART-I$^2$ platform and their real distance $d_r$ is relative to the start position (see Fig. 2).

During the *exploration* phase, subjects can move in the *exploration* area which is a rectangle of $1 \times 0.8$ m$^2$. During this phase, participants are instructed to move in the *exploration* area to acquire "a good mental representation of the object and of its environment". This *exploration* phase is depicted in Fig. 2.

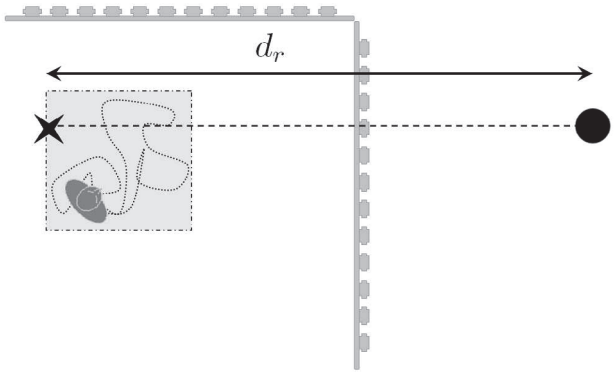



Figure 2. *Exploration phase. Start position*: black «×». *Exploration area*: grey rectangle. *Virtual object*: black disk placed at a distance $d_r$ from the *start position*. The dotted line indicates a hypothetical exploration trajectory performed by the subject.

Once a "a good mental representation" has been acquired, the perceived distance $d_p$ is estimated by means of triangulated blind walking [2, 22–25]. During the *triangulation* phase, depicted in Fig. 3, participants close their eyes, make a 40 ° right-turn and walk blindly for an imposed distance of $\simeq$ 2 m. A handrail guide has been included to help during blind-walking. Subjects stop at the end of the guide, turn in the direction where they thought the object is, and make a step forward in that direction. Participants are told that the perceived distance will be calculated from this step. They then indicate that they have completed a trial by pressing a wiimote button. After that they can open their eyes and go back to the *start position* for a new iteration. The experimental protocol is fully automated.

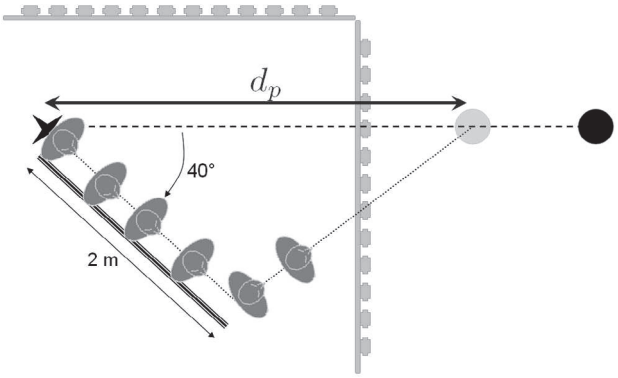



Figure 3. *Triangulation phase. Start position*: black «×». *Guide*: large continuous black lines. *Virtual object*: black disk placed at a distance $d_r$ from the *start position*. The dotted line indicates a hypothetical trajectory performed by the subject. *Perceived object*: grey disk placed at the estimated perceived distance $d_p$ from the *start position*.

## 3. Results

Because of large differences observed relative to other participants of the mean estimated distances, data from 2 subjects have been removed from the analysis. Results from the 13 remaining participants are shown in this section.

### 3.1. Data treatment

Position of the head of the participant (middle of the eyes) and corresponding times are recorded for each iteration during both the *exploration* and the *triangulation* phases. Examples of *exploration* and *triangulation* trajectories are shown in Fig. 4. Perceived distances are estimated from the *triangulation* trajectory as follows. A line $y = ax+b$ is fitted to the trajectory points during the forward step. Afterwards, the estimated perceived distance is given by $d_p = -b/a$.

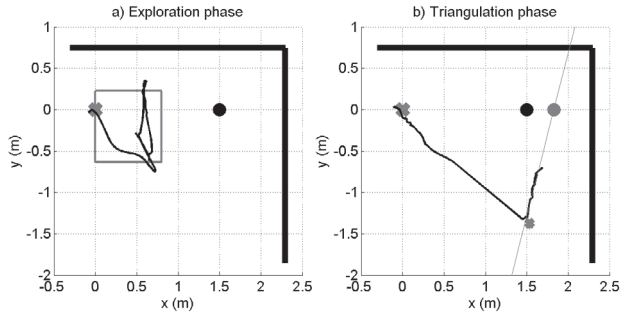



Figure 4. Example of *exploration* and *triangulation* trajectories (participant 2, iteration 15, *audio* condition, $d_r = 1.5$ m and $d_p = 1.83$ m). *Exploration area*: grey rectangle. *Beginning position*: big grey «×». *End of the guide*: small grey «×». *Virtual object*: black disk. *Perceived object*: grey disk. *Participant trajectories*: solid black lines.

### 3.2. Analysis of perceived distances

Results for the dependant variable $d_p$, *i.e.* the perceived distance, are given in Tab. III and shown in Fig. 5 for each condition. On average, participants performed the *exploration* task in 12.2 s $\pm$ 6.6 s and the *triangulation* task in 8.6 s $\pm$ 2.3 s. As expected, perceptual compressions are observed in the different rendering conditions.

The perceived distances obtained in the present repeated-measures design were analysed using a repeated-measures two-way analysis of variance. There were two within-subjects factors: rendered distance $d_r$ (5 levels, fixed factor) and rendering condition (3 levels, fixed factor). Rendered distance $d_r$ was significant at the 5 % level with $F(4, 40) = 89$ and $p < 10^{-4}$. Surprisingly, rendering condition was not significant at the 5 % level as $F(2, 20) = 1.98$ and $p = 16.5$ %. After the effects of individual differences have been removed, the percentage of variability associated with rendered distance is $R^2 = 89.93$ %. As a post-hoc test, series of Bonferroni corrected t-tests have been performed and all the rendered distance pairs have been found to be significantly different. The different distances $d_r$ are thus well recognized by the participants independently of the rendering condition.

| | **A** | **V** | **AV** |
|---|---|---|---|
| **1.5 m** | 1.92 ±0.42 | 1.60 ±0.35 | 1.65 ±0.40 |
| **2 m** | 2.14 ±0.53 | 2.03 ±0.37 | 2.01 ±0.29 |
| **2.7 m** | 2.44 ±0.57 | 2.30 ±0.37 | 2.27 ±0.35 |
| **3.5 m** | 2.60 ±0.53 | 2.42 ±0.41 | 2.59 ±0.69 |
| **5 m** | 2.96 ±0.83 | 2.79 ±0.76 | 2.91 ±0.78 |

Table III. Mean and standard deviation (in m) of the perceived distances $d_p$ as a function of rendered distance $d_r$ for each rendering condition (see Fig. 5).

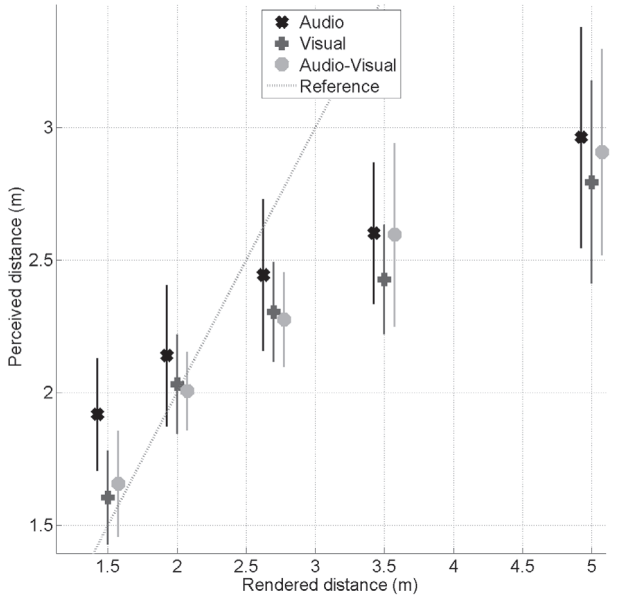



Figure 5. Perceived distances $d_p$ as a function of rendered distance $d_r$ for each rendering condition (see Tab. III). «×,○,+»: Mean under each modality. *Vertical lines*: Standard deviation under each modality. For the sake of readability, results corresponding to the different rendering conditions have been slightly shifted.

## 4. Discussion

### 4.1. Modality precision

Azimuth judgements based on the *visual* modality are much more accurate than azimuth judgements based on the *audio* modality. This is not the case for distance judgements. A Kruskal-Wallis test performed

on the standard deviations given in Tab. III with rendering condition as an independent factor gives $\chi^2(2, 12) = 2.66$ and $p = 26.5$ %. Standard deviations for the *audio*, *visual*, and *audio-visual* perceived distances are thus not significantly different. This is coherent with results from [23] where real-world distance are estimated using blind direct walking under *audio*-only and *visual*-only conditions. The *audio* modality is therefore as precise as the *visual* modality for the perception of distances in virtual environments when rendered distances are between 1.5 m and 5 m.

### 4.2. Limits of *visual capture* theory

The *visual capture* theory, already demonstrated in [14–18] for audio-visual distance perception, suggests that the *visual* modality is the more reliable and dominates the *audio* modality for localisation tasks. In the hypothesis that the *audio* modality is as precise as the *visual* modality (see Sec. 4.1), doubt is cast on the validity of such a theory. However, the present results are not sufficient to conclude on this point. Additional studies involving differences between the audio and visual rendered distances (incoherent stimuli presentation) are needed to clearly investigate the limits of the *visual capture* theory.

### 4.3. Virtual is not real

Examining Tab. III and Fig. 5, it can be seen that the mean values in the *audio*, *visual*, and *audio-visual* conditions are not significantly different. This has been assessed by a repeated-measures 2-way ANOVA which found no significative differences between the different rendering conditions. We can form the hypothesis that the perceived distance is, in the virtual environment, the same in the *audio*, *visual*, and *audio-visual* rendering conditions (same compression curve in Fig. 5). We note that this hypothesis is not true in real environments [30, 31]. We can thus wonder for what reason visual and audio cues are treated the same way in virtual environments while being treated differently in real environments?

### 4.4. Influence of the physical rendering setup

A final discussion point in the results presented in Fig. 5 is the point at which rendered distance equals perceived compressed *audio*, *visual*, or *audio-visual* distances. This point is situated here around 2.3 m. This distance is the distance at which is effectively positioned the user from the right LaMAP, *i.e.* the flat array of loudspeakers used as a front projection screen (see Sec. 2.2). When a virtual object is rendered at a distance of 2.3 m, its image and sound position are thus *physically* and *perceptually* close to the position of the LaMAP. The physical rendering setup thus seems to act here as an *anchor* between the real and virtual worlds.

## Acknowledgements

The authors thank all the volunteers who took part in the experiment. Thomas Chartier and Philippe Cuvillier, currently students at the École Polytechnique (France) are also thanked for their help in the design and performance of preliminary tests. Finally, a special thanks is given to Matthieu Courgeon for the time he spent on visual rendering.

## References

[1] F. P. Brooks. What's real about virtual reality? *IEEE Computer Graphics And Applications*, 19(6):16–27, 1999.

[2] E. Klein, J. E. Swan, G. S. Schmidt, M. A. Livingston, and O. G. Staadt. Measurement protocols for medium-field distance perception in large-screen immersive displays. *IEEE Virtual Reality 2009, Proceedings*, pages 107–113, 2009.

[3] T. Y. Grechkin, T. D. Nguyen, J. M. Plumert, J. F. Cremer, and J. K. Kearney. How does presentation method and measurement protocol affect distance estimation in real and virtual environments? *ACM Transactions on Applied Perception*, 7(4):26, July 2010.

[4] C. Armbruster, M. Wolter, T. Kuhlen, W. Spijkers, and B. Fimm. Depth perception in virtual reality: Distance estimations in peri- and extrapersonal space. *Cyberpsychology & Behavior*, 11(1):9–15, February 2008.

[5] A. Naceri, R. Chellali, F. Dionnet, and S. Toma. Depth perception within virtual environments: a comparative study between wide screen stereoscopic displays and head mounted devices. *2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns*, pages 460–466, 2009.

[6] I. V. Alexandrova, P. T. Teneva, S. de la Rosa, U. Kloos, H. H. Bülthoff, and B. J. Mohler. Egocentric distance judgments in a large screen display immersive virtual environment. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, APGV '10, pages 57–60, New York, NY, USA, 2010. ACM.

[7] J. M. Loomis, R. L. Klatzky, and R. G. Golledge. *Mixed reality: Merging real and virtual worlds*, chapter Auditory distance perception in real, virtual, and mixed environments. 1999.

[8] H. Y. Kim, Y. Suzuki, S. Takane, and T. Sone. Control of auditory distance perception based on the auditory parallax model. *Applied Acoustics*, 62(3):245–270, March 2001.

[9] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 111(4):1832–1846, April 2002.

[10] E. Corteel. *Caractérisation et Extensions de la Wave Field Synthesis en conditions réelles*. PhD thesis, Université de Paris 6, 2004.

[11] H. Wittek, S. Kerber, F. Rumsey, and G. Theile. Spatial perception in wave field synthesis rendered sound fields: Distance of real and virtual nearby sources. In *Audio Engineering Society Convention 116*, 5 2004.

[12] H. Wittek. *Perceptual differences between wavefield synthesis and stereophony*. PhD thesis, Department of Music and Sound Recording School of Arts, Communication and Humanities University of Surrey, 2007.

[13] G. Kearney, M. Gorzel, F. Boland, and H. Rice. Depth perception in interactive virtual acoustic environments using higher order ambisonic soundfields. In *Proccedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010.

[14] M. B. Gardner. Proximity image effect in sound localization. *Journal of the Acoustical Society of America*, 43(1):163, 1968.

[15] D. H. Mershon, D.H. Desaulniers, T. L. Amerson, and S. A. Kiefer. Visual capture in auditory distance perception: proximity image effect reconsidered. *Journal of Auditory Research*, 20(2):129–136., 1980.

[16] D. H. Mershon, D. H. Desaulniers, S. A. Kiefer, T. L. Amerson, and J. T. Mills. Perceived loudness and visually-determined auditory distance. *Perception*, 10(5):531–543, 1981.

[17] P. Zahorik. Estimating sound source distance with and without vision. *Optometry and Vision Science*, 78(5):270–275, May 2001.

[18] A. Bowen. *Visual Localization Accuracy Determines the Bias of Auditory Targets in Azimuth and Depth*. Master's thesis, Wake Forest University, 2010.

[19] M. Rébillat, E. Corteel, and B. F.G. Katz. SMART-I$^2$ "Spatial Multi-user Audio-visual Real-Time Interactive Interface". *125th Convention of the Audio Engineering Society*, 2008.

[20] M. Rébillat, B. F.G. Katz, and E. Corteel. SMART-I$^2$: "Spatial Multi-user Audio-visual Real-Time Interactive Interface", a broadcast application context. *Proceedings of the IEEE 3D-TV conference*, 2009.

[21] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

[22] S. S. Fukusima, J. M. Loomis, and J. A. DaSilva. Visual perception of egocentric distance as assessed by triangulation. *Journal of Experimental Psychology- Human Perception and Performance*, 23(1):86–100, February 1997.

[23] J. M. Loomis, R. L. Klatzky, J. W. Philbeck, and R. G. Golledge. Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics*, 60(6):966–980, 1998.

[24] J. M. Loomis and J. M. Knapp. *Virtual and Adaptive Environments: Applications, Implications, and Human Performance Issues*. Lawrence Erlbaum Associates, 2003.

[25] W. B. Thompson, P. Willemsen, A. A. Gooch, S. H. Creem-Regehr, J. M. Loomis, and A. C. Beall. Does the quality of the computer graphics matter when judging distances in visually immersive environments? *Presence-Teleoperators and Virtual Environments*, 13(5):560–571, 2004.

[26] J. E. Cutting. How the eye measures reality and virtual reality. *Behavior Research Methods Instruments & Computers*, 29(1):27–36, February 1997.

[27] B. R. Kunz, L. Wouters, D. Smith, W. B. Thompson, and S. H. Creem-Regehr. Revisiting the effect of quality of graphics on distance judgments in virtual environments: A comparison of verbal reports and blind walking. *Attention Perception & Psychophysics*, 71(6):1284–1293, August 2009.

[28] S. H. Creem-Regehr, P. Willemsen, A. A. Gooch, and W. B. Thompson. The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, 34(2):191–204, 2005.

[29] E. Corteel, K. V. NGuyen, O. Warusfel, T. Caulkins, and R. Pellegrini. Objective and subjective comparison of electrodynamic and map loudspeakers for wave field synthesis. *30th International Conference of the Audio Engineering Society*, 2007.

[30] W. M. Wiest and B. Bell. Stevens exponent for psychophysical scaling of perceived, remembered, and inferred distance. *Psychological Bulletin*, 98(3):457–470, 1985.

[31] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United With Acustica*, 91(3):409–420, May 2005.