Ventriloquism effect on distance auditory cues

N. Cote[a], V. Koehl[b] and M. Paquier[b]

[a]Institute of Electronics, Microelectronics and Nanotechnology, UMR CNRS 8520, ISEN department, 41 boulevard Vauban, 59046 Lille, France
[b]Université de Bretagne Occidentale, 6, avenue Victor Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France
nicolas.cote@isen.fr

Even though virtual reality applications are nowadays multimodal, developers focus their efforts on the visual rendering system. Complex visual rendering systems using stereoscopic techniques are employed in order to place visual objects in a three dimensions environment. Similar systems such as binaural rendering through headphones can be used for the auditory modality. However, several studies have reported a visual attractive effect in case of auditory-visual object which reduces the benefit of complex auditory rendering systems. A cognitive process combines both acoustic and visual cues and gives a higher influence to the visual modality. The resulting multimodal object is thus placed at the position of the visual cue. However, this effect has been less studied in the distance dimension. This study investigates the effect of separate visual and acoustic distance cues. For this purpose a binaural rendering is employed for acoustic cues and combined to a stereoscopic display for visual cues. The results show an asymmetrical ventriloquism effect in the distance dimension: the relative position of the sound source in comparison to the visual object has an influence on the perceived position of the auditory-visual object. A description and a possible explanation of this asymmetrical ventriloquism effect is detailed in this study.

1 Introduction

In cross-modal integration of non-coincident visual and acoustic cues, one modality can alter the perceptual process related to the other modalities. Such alterations are relatively frequent and a well-known example is the McGurk effect [12]: lip-movements can prevent the correct recognition of the phoneme heard. For instance, the phoneme /ba/ is perceived as /da/ when simultaneously the lip-movement /ga/ is seen. This effect shows that an interaction exist between hearing and vision during the perceptual and cognitive processes of information.

In the specific case of localization of multimodal objects, the "ventriloquism effect" is a famous cross-modal integration process where a visual object bias the perceived position of a sound source [15]. For instance, the image of a person speaking on a TV screen can attract the perceived position of speech sounds (coming from the TV loudspeaker) towards the lips of the person [2]. According to Ernst and Banks [7], the ventriloquism effect is related to the accuracy of the two senses vision and audition in terms of localization. Since vision is spatially more accurate than audition, it has a higher influence in the localisation of audiovisual objects. However, the audition can dominate vision for temporal cross-modal integration, see for instance Shams et al. [14]. The authors reported that vision can be biased by auditory information.

Even though ventriloquism effect has been well defined in azimuth and elevation, few papers were published on the cross-modal integration of auditory and visual distance estimates. The human auditory and visual systems are able to extract information about distance from many visual and acoustic cues. Zahorik et al. [17] produced an exhaustive list of the acoustic distance cues. This includes intensity, direct to reverberant energy ratio, spectrum coloration and cues related to binaural hearing (interaural time difference and interaural phase difference). Similar review has been produced by Cutting and Vishton [5] for visual distance cues. This includes relative size, linear perspective, aerial perspective, accommodation and cues related to binocular vision (convergence and binocular disparities). The available information on localization is largely enhanced by either subject or object motion.

Gardner [8] was the first to observe a dominance of vision over audition in distance. For this purpose, the author carried out an experiment with spatially non-coincident visual and acoustic cues. The subjects and sound sources (line of five loudspeakers) were placed in an anechoic chamber. The subjects perceived the sound source as coming from the nearest visual object. The effect observed was called "proximity-image effect" by the author. Then, Mershon et al. [13] repeated the experiment in both an anechoic and a reverberant room. The subjects located the sound source at either the nearer or farther visual object. These results show that the proximity-image effect is thus an example of a more general "visual-capture effect". More recently, Agganis et al. [1] used a virtual environment to study multimodal localisation in distance. The authors observed that vision biases the location of sound sources and this bias is larger in distance than in azimuth. It is worth noting that Zahorik [16] replicated Gardner's [8] experiment and observed no visual-capture effect. On the contrary, visual cues improved distance judgement accuracy. Overall, these results show that vision has an influence on sound source localisation but visual-capture effect is not a general effect.

The term ventriloquism refers to a "complete" visual-capture effect, i.e. the sound source is perceived at the position of the visual object. In addition, visual-capture effect is used for all types of dominance of vision over audition. For instance, this effect can occur for moving sound sources, see Zhou et al. [18].

This study aims at investigating the ventriloquism effect on distance and the corresponding human integration process of the visual and auditory modality. This study was motivated by the new possibilities of virtual reality applications and corresponding signal processing systems. The experiment presented in section 2 uses a virtual environment where subjects were asked to localize virtual visual and/or auditory objects in distance. The experimenters paid attention to the acoustic and visual displays characteristics: real-time processing, stereoscopy, large field of view, binaural rendering. Consequently, subjects' judgements obtained with this specific equipments are considered as similar to judgements obtained in a real environment [4, 9].

2 Experiment

The following section summarizes the test conditions and procedure used for the experiment carried out for this study. A complete description of the equipment and procedure can be found in Côté et al. [3].

An acoustic display and a visual one have been used to present a virtual environment to the subjects. Virtual environment enables to control all characteristics of the experiment and especially the position of the visual

object and the sound source. The same visual object and sound source were processed through different conditions according to four experimental variables: presentation modality (auditory-only, visual-only and bimodal), target distance (2, 3, 5, 10, 20 m), room effects (reverberation time of T_{60} = 370 ms or T_{60} = 860 ms) and amount of visual information (reduced or full visual cues). The resulting 48 test conditions tested in the experiment can be grouped in four blocks:

1. modality auditory alone (Block 1: conditions 1–10),
2. modality visual alone (Block 2: conditions 11–20),
3. bimodal, spatially coincident cues (Block 3: conditions 21–40),
4. bimodal, spatially non-coincident cues (Block 4: conditions 41–48).

The conditions were assessed in blocks. In each block, all combinations of target distance and reverberation and/or amount of visual information were presented in random order, with four repetitions per condition. The combination of modalities in distance perception were assessed by using eight conditions (Block 4) with spatially non-coincident acoustic and visual cues. These are defined by the target distances of the visual object, $\rho_{tar,V}$, and the sound source, $\rho_{tar,A}$, see table 1. The offset between the two target distances is defined as:

$$\Delta\rho = \rho_{tar,A} - \rho_{tar,V} \quad (1)$$

For these eight conditions the stimuli were placed in the visual environment with full visual cues and the room effect corresponding to the short reverberation time T_{60} = 370 ms has been used.

Table 1: Description of the eight bimodal conditions with spatially non-coincident acoustic and visual cues.

N	$\rho_{tar,V}$ (m)	$\rho_{tar,A}$ (m)
41	2	1
42	5	20
43	5	1
44	5	10
45	10	20
46	10	3
47	10	15
48	20	5

The visual target consisted in a virtual blue loudspeaker placed in a virtual room corresponding to the extension of the real test room through the visual display. The visual display was a 2.4 × 1.8 m^2 stereoscopic screen. A stereoscopic rendering technique has been used to provide binocular cues in addition to the monocular information provided by the visual environment. The sound source corresponds to a speech signal composed of two French sentences. The sound stimuli is processed by a binaural rendering system and reproduced through headphones. For this purpose, the sound source has been convolved with Binaural Room Impulse Responses (BRIRs) at the different target distances. The BRIR accounted for the Head-Related Impulse Responses (HRIRs) as well as the room effect.

In the present study the subjects were asked to report the egocentric distance of static visual object and/or sound source on a measurement scale. Throughout the experiment, the subjects were positioned on a chair placed at 2 m in front of the middle of the screen resulting in a 62° horizontal field of view and enabled subjects to view the near virtual ground surface. This large view on the virtual world serves as a reference frame to localise the target object (loudspeaker). After presentation of the sound and/or visual stimuli, subjects were asked to report their egocentric distance judgements by using a keypad. The experiment consisted of two sessions:

1. the auditory- (Block 1) and visual-only (Block 2) conditions. Half of the subjects started with the auditory block and half with the visual one.
2. the auditory-visual conditions (Block 3 and 4).

The two sessions were separated in time by at least 36 hours.

A total of 24 subjects participated in the experiment (results from 2 subjects were rejected due to incoherence in their judgements). Participants were naive with respect to the purpose of the experiment. They had normal or corrected to normal vision and reported no hearing impairments.

3 Results

First, a test of normal distribution has been applied to subjects' judgements for each condition. The analysis denotes an asymmetry towards the higher distance values and a large spread of the subjects' judgements around the mean egocentric perceived distance. Since the subjects' judgements do not follow a normal distribution, non-parametric statistical measures were employed to analyse the perceived distance values. The 95% confidence intervals are calculated by using a Bootstrap technique [6]. This technique is employed to approximate a parameter (here the variance around the condition mean) by constructing a number of resamples of the distance values (here, up to 2000 samples). An additional statistical test is employed to detect significant effects: the *Friedman* (paired samples) test.

This paper focuses on the bimodal conditions only (Block 3 and 4). Three types of statistical tests were performed: all positions presented in figure 1 are compared by pairs. Position 2 corresponds to the eight conditions with spatially non-coincident cues, see table 1. Position 1 and 3 corresponds to conditions with spatially coincident cues: in Position 1, the object is placed at $\rho_{tar,A}$ and in Position 3, the object is placed at $\rho_{tar,V}$. Therefore, in Positions 1 and 3 (conditions in Block 3), the visual object and the sound source are at the same target distance whereas in Position 2 (conditions in Block 4) they are separated by $\Delta\rho$.

The first test (Position 1 *vs* Position 3) is a preliminary test which considers objects with spatially coincident cues only. This first statistical test tries to answer the following question: does a shift in target distance of both modalities together has an influence on the subjects' judgements? Then, two others tests were performed: Position 2 *vs* Position 1 and Position 2 *vs* Position 3. They detect whether a shift in visual (respect. auditory) target distance $\rho_{tar,V}$ (respect. $\rho_{tar,A}$) alone has an influence on subjects' judgements. These two tests try

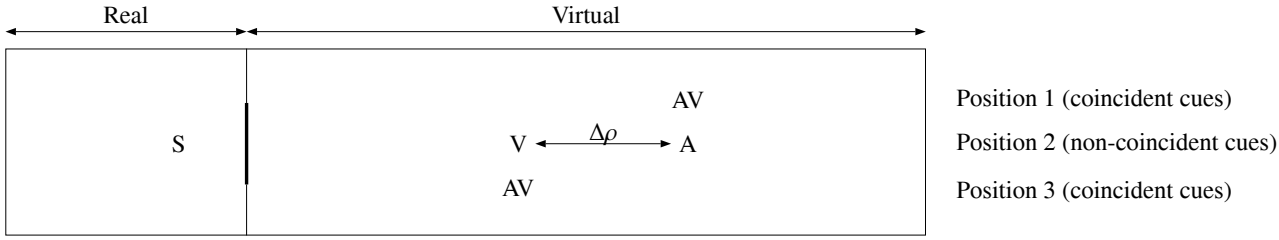

Figure 1: Dimension of the test room and of its virtual extension (top view). The labels correspond to the positions of the subject (S), the visual object (V, $\rho_{tar,V}$) and the sound source (A, $\rho_{tar,A}$).

to answer a second question: does the shift in target distance of one modality alone influence the perception process?

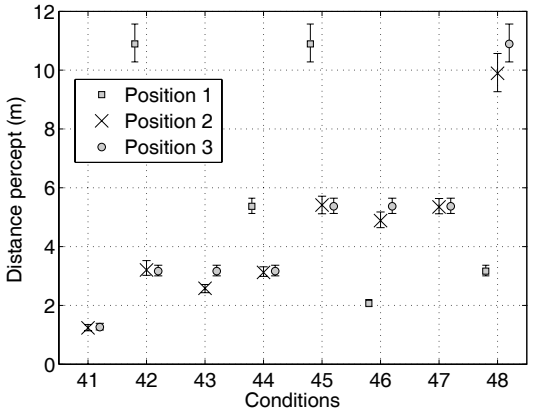

Figure 2: Distance percepts for the eight spatially non-coincident conditions described in table 1.

Figure 2 shows the perceived egocentric distances for the bimodal conditions with spatially non-coincident cues (Position 2) and coincident cues (Positions 1 and 3). Error bars correspond to the 95% confidence intervals calculated with the *Bootstrap* technique.

3.1 Position 1 *vs* Position 3

This first statistical test focuses on conditions with coincident modalities only: comparison between objects placed at the sound source target distance, $\rho_{tar,A}$, and objects placed at the visual object target distance, $\rho_{tar,V}$. For instance, the distance percepts for two bimodal objects placed at 5 and 20 m were compared (corresponding to the condition 42, $\rho_{tar,V} = 5$ m and $\rho_{tar,A} = 20$ m, see table 1). This test is used to verify the influence on the distance percepts of a shift in target distance equivalent to the offset Δ_ρ. For this purpose a *Friedman* statistical test for each pair of conditions was applied. However, the subjects did not judge the position of bimodal objects with coincident cues at target distances 1 and 15 m. Therefore only conditions 42, 44, 45, 46 and 48 were analysed. Results of the *Friedman* test indicated a significant difference in distance percepts for each pair of conditions ($p < 10^{-5}$), see the second column of table 2. Thus, a shift in target distance of both modalities together has an influence on the subjects' judgements.

3.2 Position 2 *vs* Position 1

The eight conditions with non-coincident cues (Position 2) have been compared to positions with spatially coincident cues placed at the auditory target distanc,e $\rho_{tar,A}$ (Position 1), see figure 1. For instance, the condition 42 was compared to an object with coincident cues placed at 20 m, see table 1. This statistical test quantifies the impact of the visual stimulus displacement. A *Friedman* test showed a significant difference for each pair of conditions, see the third column of table 2. The displacement of the visual stimulus has an influence on the perceived distance.

3.3 Position 2 *vs* Position 3

Last, the conditions have been compared to conditions with spatially coincident cues placed at the visual target distance, $\rho_{tar,V}$ (Position 3), see figure 1. This statistical test quantifies the impact of the sound stimulus displacement. The results of the eight *Friedman* tests are depicted in the fourth column of Table 2. Here, the results depend on the relative position of the sound source in comparison to the visual object. There is a significant difference for conditions 43, 46 and 48 only. The conditions with the sound source placed behind the visual object, i.e. $\rho_{tar,A} > \rho_{tar,V}$, provide similar perceived distance to conditions with coincident cues. In that case, the subjects perceived the bimodal object at the position of the visual object. In other terms, the position of the auditory object is displaced towards the position of the visual object. This phenomenon corresponds to the ventriloquism effect described in section 1. However, table 2 shows no ventriloquism effect in case the sound source is placed ahead the visual object (condition 41 excepted), i.e. $\rho_{tar,A} < \rho_{tar,V}$. These results show an influence of distance and arrangement of the modalities (i.e. visual object or sound source nearer) on the ventriloquism effect.

Table 2: Results of the three statistical tests

N	1 *vs* 3	2 *vs* 1	2 *vs* 3
	p	p	p
41	-	-	.564
42	$< 10^{-5}$*	$< 10^{-5}$*	.999
43	-	-	.001*
44	$< 10^{-5}$*	$< 10^{-5}$*	.796
45	$< 10^{-5}$*	$< 10^{-5}$*	.467
46	$< 10^{-5}$*	$< 10^{-5}$*	.002*
47	-	-	.492
48	$< 10^{-5}$*	$< 10^{-5}$*	.002*

4 Discussion

It is worth noting that cross-modal integration appears for coherent sources, i.e. when visual and acoustic information form a unified object in time and in position [10]. For large deviations between the two modalities, the central nervous system is able to separate the two objects that have produced the respective modalities. During the experiment, few subjects detected conflict between the visual and acoustic cues. However, they took into account both cues to determine the position of the bimodal virtual object.

In their everyday life, humans perceive their surrounding environment by using several sources of information: vision, audition, touch, etc.. These senses receive correlated information which enables an almost perfect description of these objects (in terms of identification and localisation). Information are extracted from multiple cues that are combined to provide humans accurate percepts of the target position. Ventriloquism effect implies that subjects' judgements are based exclusively on vision, i.e. the most reliable modality for localisation. Results described in section 3 show a visual-capture effect of the sound source. However, ventriloquism effect seems to be asymmetrical and thus appears in case the sound source is placed behind the visual object. The position of the sound source has a small but significant "auditory-capture effect" on the perceived distance of bimodal objects. Thus, ventriloquism is not a global effect and humans combine information from multiple modalities in a statistical manner.

According to psychophysicists, the auditory distance estimate and visual one are linearly combined into a single cross-modal percept of the target distance. Weightings are applied to each distance estimate which takes into account the reliability of the corresponding estimate and its discrepancy with the other estimates: a smaller weight is associated to a less reliable cue in the integration process [7]. These simple models are coherent with neurophysiology studies. For instance, Ma et al. [11] showed that estimates integration is performed in the central nervous system by a simple linear combination of neural activity. An statistical analysis of the subjects' judgements should reveal the weights apply by human to both modalities, vision and audition.

Acknowledgement

The authors would like to thank all test subjects. This research has been funded by the Finistère General Council (29), France.

References

[1] B. T. Agganis, J. A. Muday, J. A. Schirillo, "Visual biasing of auditory localization in azimuth and depth", *Perceptual and Motor Skills* **111**(3), 872–892 (2010)

[2] P. Bertelson, G. Aschersleben, "Automatic visual bias of perceived auditory location", *Psychonomic Bulletin & Review* **5**(3), 482–489 (1998)

[3] N. Côté, V. Koehl, M. Paquier, F. Devillers, "Interaction between auditory and visual distance cues in virtual reality applications", *In Proc. of Forum Acusticum*, 1275–1280, Aalborg, DK (2011)

[4] S. H. Creem-Regehr, P. Willemsen, A. A. Gooch, 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)

[5] J. Cutting, P. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth", *In Perception of Space and Motion*, Chapter 3, 69–117 (1995)

[6] A. C. Davison, D. V. Hinkley, "Bootstrap Methods and Their Application", Cambridge Univ. Pr. (1997)

[7] M. O. Ernst, M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion", *Nature* **415**(6870), 429–33 (2002)

[8] M. B. Gardner, "Proximity image effect in sound localization", *J. Acoust. Soc. Am.* **43**(1), 163 (1968)

[9] V. Interrante, B. Ries, L. Anderson, "Distance perception in immersive virtual environments, revisited" *In Virtual Reality Conference*, 3–10 (2006)

[10] M. S. Landy, L. T. Maloney, E. B. Johnston, M. Young, "Measurement and Modeling of Depth Cue Combination: In Defense of Weak Fusion" *Vision Research* **35**(3), 389–412 (1995)

[11] W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, "Bayesian inference with probabilistic population codes". *Nature Neuroscience* **9**(11), 1432–8 Epub (2006)

[12] H. McGurk, J. MacDonald, "Hearing lips and seeing voices" *Nature* **264**(5588), 746–748 (1976)

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

[14] L. Shams, Y. Kamitani, S. Shimojo, "What you see is what you hear" *Nature* **408**(6814), 788 (2000)

[15] R. W. Thurlow, C. E. Jack, "Certain determinants of the "ventriloquism effect" " *Perceptual and Motor Skills* **36**(3c), 1171–1184 (1973)

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

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

[18] L. Zhou, J. Yan, Q. Liu, H. Li, C. Xie, Y. Wang, J. Campos, H. Sun, "Visual and auditory information specifying an impending collision of an approaching object" *In Proc. of the 12th Int. Conf. on Human-computer interaction: interaction platforms and techniques* (4551), 720–729, Beijing, CH, Springer-Verlag (2007)