AN EXPERIMENTAL EVALUATION OF THE INFLUENCE OF AUDITORY CUES ON PERCEIVED VISUAL ORDERS IN DEPTH

Delphine Devallez (1), Davide Rocchesso (2), Federico Fontana (1)

(1) University of Verona, Department of Computer Science
Strada Le Grazie 15, 37134 Verona, Italy
(2) IUAV, Department of Art and Industrial Design
Dorsoduro 2206, 30123 Venezia, Italy
devallez@sci.univr.it

ABSTRACT

We present an experiment investigating the influence of auditory cues on visual perceived orders in depth. Visual stimuli consisted in a layered 2D drawing of two squares respectively blue and red using semi-transparency. Auditory signals of the two words “red” and “blue” were presented simultaneously to the images. Subjects were required to determine which square appeared in front of the other in these cross-modal conditions. The coefficient of transparency as well as the audio level difference between the two speech signals “red” and “blue” were systematically varied. No significant influence of auditory cues on perceived order in depth was found, except when the visual information was totally ambiguous: in this case, the perceived order showed limited dependence on the acoustic information.

[Keywords: Sensory integration, Auditory-visual interaction, Depth perception]

1. INTRODUCTION

1.1. Depth rendering in human-computer interfaces

Techniques such as occlusion and perspective on 2D visual displays are largely used nowadays to render layered content and give a sense of depth. In 1993, Bier et al [1] introduced semi-transparent widgets appearing in front of an application and providing the user with tools for operating directly on the application beneath. In first-person engagement in video-games, as in boxing with the Nintendo Wii, transparency is also used to represent the alter ego of the player, and to distinguish him from the contender. In another context, the recent development of mobile TV opens the way for new audio-visual rendering techniques especially because of the limited size of the screen and reduced budgets. Sasse and Knoche [2] have indeed demonstrated that the requirements for audio and video quality depend on the context of use. For mobile TV, factors on the perceived quality include the shot types, the audio quality and the legibility of text if present. Watching a football match on a mobile phone is very illustrative: people expect to be able to recognize the players and see the ball, which is not as obvious as on a normal TV screen. For this kind of applications, we believe that simple techniques such as transparency could be used to give a sense of depth. In the auditory domain, similar techniques such as the manipulation of the direct-to-reverberant energy ratio would as well contribute to a consistent 3D rendering of audio-visual contents and improve the quality and efficiency of multisensory products, as highlighted by Spence and Zampini [3]. Tools for practical applications of auditory depth have already been proposed. Schmandt [4] proposed a tool called *acoustic zooming*, similar to the visual ability of focusing on a specific area of a display and applied to an auditory browsing environment of audio files. In a very similar manner, Fernström and McNamara [5] included a function called *aura* which restricted the user’s spatial range of hearing in a virtual soundscape in order to make the browsing task more efficient. However recent auditory interfaces have rather taken benefit of research on auditory directional perception to increasingly provide users with spatialized auditory displays, with applications ranging from scientific simulations for research purposes to entertainment and infotainment. In contrast with directional localization, relatively little attention has been given to auditory depth. To increase the degree of realism of the overall display as well as provide more information to the user, it seems indeed natural to render the depth dimension, both visually and auditorily.

It is therefore fundamental to understand how people locate images and sounds in the depth dimension as well as address interactions between audition and vision.

1.2. Previous studies in auditory-visual interactions

Past studies have demonstrated the improvement of some specific tasks by adding auditory stimuli to the visual ones (see [6], [7],[8] for reviews). These include improvement of target detection, decreased reaction times and localization improvement. In particular, cross-modal benefits are significant when spatial information in one sense is compromised or ambiguous.

In his study, Hairston et al [6] examined the benefit of acoustical cues under conditions of myopia by presenting light-emitting diodes to the subjects with or without a broadband noise burst coming from the same location. While directional localization accuracy was equivalent for visual and multisensory targets under normal vision, the myopia condition showed a substantial improvement with the addition of auditory cues. In other words, when the visual sense gives ambiguous information, auditory cues have been shown to resolve the ambiguity. Sekuler et al [9] conducted an experiment on the perception of motion of two disks. Without any other cue, the visual stimulus may result into two different interpretations: either the disks stream through, or they bounce off each other. However, since collision often produce sounds characteristic of the impact, the absence of sound rather leads to the perception of streaming through. The perception of the scene was changed with the addition of a brief click at or near the point of coincidence, and promoted the perception of bouncing. Besides showing the effect of sound on visual motion, the authors also reported that the auditory stimulus did not need to be in perfect synchrony with the visual one but could be presented up to 150 ms before or after the visual coincidence point.

Another similar experiment was conducted by Ecker and Heller in [10] on the perception of motion. This time, the ambiguous visual stimulus consisted of a rolling ball that could either roll back in depth on the floor of a box, or jump in the frontal plane. Moreover, other ball's paths of different types and curvature in between were also presented to the subjects. The moving ball was either shown alone, accompanied with the sound of a ball rolling, or the sound of a ball hitting the ground. Similarly to the results of Sekuler [9], but this time with sounds other than transients, it was found that sound influenced the perception of the ball's trajectory, depending on the type of sound.

Frassinetti [11] also reported an improvement of visual tasks under auditory-visual conditions. Her experiment showed that the perceptual sensitivity for luminance detection of a green LED masked with four red LEDs was facilitated when an auditory stimulus (white noise burst) was presented at the same location and simultaneously to the visual stimulus.

Reaction time may also be speeded up by the presence of cues in different sensory modalities. Laurienti et al. [12] studied in particular the effect of semantically congruent auditory-visual stimuli on response time, using circles of red or blue color and the words "red" and "blue". Either unimodal or congruent bimodal stimuli (i.e. a red circle with the word "red") were presented to the subjects. A significant decrease of the response time was found under the auditory-visual conditions in comparison with unimodal auditory or visual conditions.

These aforementioned studies clearly showed an audio-visual integration of information. Handel [13] explains the human integration of multimodal cues based on the *unity assumption*. The latter considers temporal and spatial aspects of the auditory and visual inputs: if they are temporally synchronous and appear to come from the same spatial location, then they may refer to a single object. In the event that information from the two modalities is too conflicting, humans may decide that auditory and visual information come from two distinct objects. At the present time the processes governing the combination and integration of multiple sources of information are being quite well understood, it is however still unclear what determines the limits for interactions between signals from different senses [7, 8]. The interesting result for the present experiment is that if the spatial and temporal rules of multisensory integration are followed, auditory cues may help to resolve ambiguous visual information, especially for localization tasks. The experiment presented in this paper explores this paradigm in terms of perceived orders in depth.

2. RENDERING AUDITORY AND VISUAL DEPTH

2.1. Visual depth

A variety of techniques may be applied to render visual depth on a two-dimensional display. Among them, occlusion (also called interposition or overlapping) may be defined as follows: when an object is occluding part of another one, the latter is perceived as being further away. Occlusion, which is the easiest visual depth cue to implement, has been largely used in 3D computer graphics but presents the disadvantage of completely hiding the objects located in the background of another object. Therefore Zhai et al [14] proposed the use of *partial-occlusion*, which enables to see through the object that overlaps other objects. This cue is produced by *semi-transparency*, which means that the semitransparent surface can still be seen and does not block the view of any object that it occludes. To create the impression that one surface S_1 is in front of another surface S_2 by using semitransparency, the intensity I of the overlapping area is rendered by blending the color intensity of one surface, I_1 with the color intensity of the second surface, I_2, [15], according to:

$$I = \alpha I_1 + (1 - \alpha) I_2 \tag{1}$$

where α is the coefficient of transparency, lying between 0 and 1. If α = 1, the surface S_1 is opaque, therefore it appears in front of the surface S_2, and if α = 0, the surface S_1 is transparent, therefore the surface S_2 appears in front of S_1. As α varies from 0 to 1, the perceived surface in front will consequently change, from S_1 to S_2. In other words, Masin [16] suggested that the probability of perceiving a transparent surface in front increases as the color differences inside this surface decrease. Furthermore α = 0.5 refers to the point of equal probability: the probability of seeing S_1 in front equals the probability of seeing S_2 in front. Anderson [17] also called this phenomenon *bistable transparency*.

2.2. Auditory depth

As the distance between the sound source and the listener changes, some properties of the sound will change as well. For stationary listener and sound source, the most obvious and easy-to-implement piece of information is the intensity cue: when the sound source moves away from the listener, its intensity decreases. Other cues provide information about sound source distance, such as the direct-to-reverberant energy ratio, however in the case of speech signals Zahorik [18] showed that the intensity cue is weighted more than the direct-to-reverberant energy ratio. Gardner [19] studied speech signals in anechoic conditions and also showed the relatively good ability to estimate distance of such familiar sound sources with the intensity cue.

2.3. Implications for the design of the experiment

The experiment presented in this paper was designed to investigate the influence of auditory stimuli on the perceived order in depth of two surfaces. The unity assumption plays a major role in auditory-visual displays because it is a necessary condition for multimodal interactions. If the unity is too weak, the interaction between the two senses will be strongly reduced. With this in mind, it was decided to use colors as stimuli for the experiment: visual colors of two squares, e.g. red and blue, were associated with the recorded spoken words "red" and "blue" and the rendered visual orders in depth were consistent with the rendered audio orders in depth, i.e. if the red square appeared in front on the visual display, the audio signal "red" would be louder than (or equal to) the audio signal "blue", and vice-versa. Visual and auditory depths were rendered by manipulating respectively the coefficient of transparency of the overlapping surface and the intensity difference between the two audio signals.

3. EXPERIMENT

3.1. Method

3.1.1. Participants

Sixteen Italian volunteers (6 women and 10 men) participated in the experiment. Their ages ranged from 20 to 44 years and all of them had at least basic knowledge of the English language. All reported to have normal or corrected-to-normal vision and normal hearing. All studied or worked at the university of Verona, Italy. None of them worked in the field of crossmodal interactions and they were all naive as for the purpose of the experiment.

3.1.2. Stimuli and Apparatus

Visual stimuli. Each stimulus appeared in the middle of an Apple MacBook Pro 15-inch Widescreen Display (1440*900 pixels). The viewing distance was about 70 cm from the display. The patterns in figure 1 illustrate the stimulus shape. The stimulus consisted of two overlapping 7.8 cm * 7.8 cm squares in the middle of a permanent white rectangular background corresponding to the display area. One square was red ($C_1 = (1, 0, 0)$ in the RGB color space and the second one was blue ($C_2 = (0, 0, 1)$). To simulate transparency, the color C_3 of the overlapping area was a linear combination of the red and blue colors, such that

$$C_3 = C_1 * (1 - \alpha) + C_2 * \alpha \qquad (2)$$

where α was the coefficient of transparency and took nine values from 0.3 to 0.7 with a 0.05 increment. The overlapping squares appeared for 1 s, then the subsequent stimulus appeared 3 s after the subject answered by pressing a key. Theoretically, the point of bistable transparency arises at $\alpha = 0.5$ (which was verified during a preliminary visual experiment), while α values smaller than 0.5 make the red square appear in front of the blue one, and α values greater than 0.5 make the blue square appear in front of the red one. During the aforementioned preliminary visual experiment, the whole range of α values were explored from 0 to 1, and people were asked to determine under visual conditions only which square appeared to be in front of the other. It was found that no confusion arose for α values smaller than 0.3 or higher than 0.7.

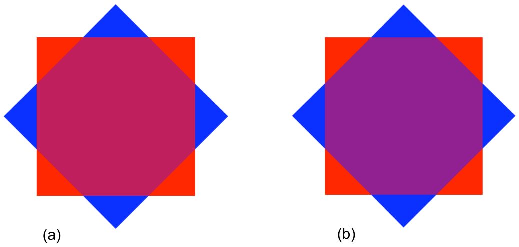

Figure 1: *Visual stimulus used for the experiment. (a) The red square appears in front of the blue square (α = 0.3). (b) Bistable transparency (α = 0.5).*

Auditory stimuli. Each visual stimulus was paired with an auditory stimulus consisting of the words "red" and "blue" presented simultaneously. These two sounds were recorded separately in a quiet room using a Marantz portable audio recorder PMD660 set at the same sound level for both sound signals. The speaker was the author herself and the two words were recorded in stereo with the built-in microphone in the uncompressed wav format, at 44.1 kHz sampling frequency. The two words "red" and "blue" were then time-aligned and sound files were normalized and shortened to 1 s. The two resulting auditory signals are shown in figure 2. In order to create an effect of auditory distance between the two signals, the sound level of each signal was manipulated digitally, while keeping the total sound level constant. The sound level difference ΔL was either -12, -6, -2, 0, 2, 6 or 12 dB, where negative values indicate that the sound level of "red" is greater than the sound level of "blue", and positive values indicate that the sound level of "blue" is greater than the sound level of "red". During the experiment auditory stimuli were presented over a pair of Beyerdynamic DT 770 headphones.

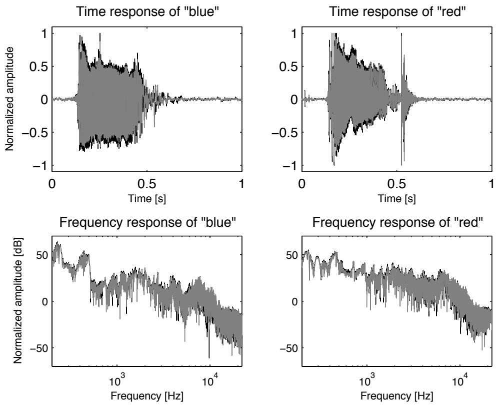

Figure 2: *Time and frequency responses of the words "blue" and "red".*

3.1.3. Design

The whole experiment consisted of three sessions per test subject, separated with breaks of about ten minutes. The visual and auditory stimuli were synchronized and had both a duration of 1 s. Besides they were congruent, i.e. both of them theoretically led to the same square being in front: negative ΔL values were combined with α values smaller or equal to 0.5, and positive ΔL values were combined with α values greater or equal to 0.5. The case where there was no audio cue, i.e. ΔL = 0, was combined with all the values of α. Therefore the test included 39 different combinations of visual and auditory stimuli, ordered randomly for each session and each subject. Furthermore, between two consecutive trials the blue and red squares were exchanged in order to avoid bias from a specific visual configuration. As a result each pair of α and ΔL was rendered twice in a different visual configuration during each session, giving a total of 78 visual-auditory stimuli per session.

3.1.4. Procedure

Before the experiment, a written instruction was given to each subject. Participants sat at a viewing distance of about 70 cm from the computer screen and wore headphones which played back the auditory stimuli. Possible auditory-visual interactions were not suggested to the subjects. For each stimulus people were asked to determine which square appeared to be in front of the other and press the key of the corresponding color. To answer, the “V” and “N” keys of the MacBook Pro keyboard were covered respectively with red and blue tags. No time limit to answer was specified, however the written instruction suggested to the participants not to think too much about their answer and rather follow their first impression. In addition to subjects’ answers, their response time were also recorded.

3.1.5. Results

Some subjects reported to have pressed the wrong key at least once during the experiment. The collected data can be represented for each subject by the percentage of answers for the blue square appearing in front, for each combination of auditory and visual stimuli. In that way a percentage smaller than 50% indicates that the answer “red” was given more often than the answer “blue”, and a percentage greater than 50% indicates that the subject answered more often “blue” than “red”. To assess the multisensory gain of combining redundant multisensory information, results were analyzed for each value of ΔL and were described by a psychometric function representing the percentage of answers “blue” as a function of the α value. The expected outcome is a psychometric function having a S shape: theoretically, answers for “blue” should increase as a function of α, from 0% for α = 0.3, to 100% for α = 0.7, while values of α close to 0.5 should lead to about 50% answers for “blue”. The boxplots of figure 3 summarize the distributions of answers for each value of α.

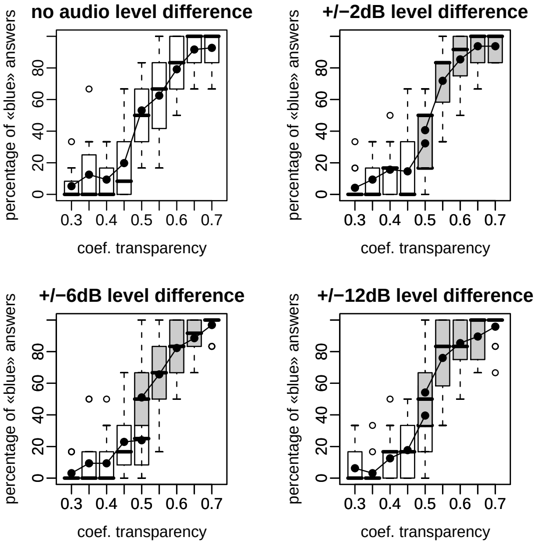

Figure 3: *Boxplots of the percentage of answers “blue” as a function of the coefficient of transparency. White boxplots represent results for negative or null audio level differences, and grey boxplots represent results for positive audio level differences. Means are connected by a line. Outliers are marked by ’o’.*

The four boxplots of figure 3 show a S shape as expected. Besides, for all boxplots answers are more spreaded around α = 0.5, also suggesting uncertainty in this region. However, comparing the case where there is no audio level difference, i.e. no audio cue, with cases where there are audio level differences, does not reveal any significant differences in people’s answers: if the audio cues would help them in answering correctly, answers for α slightly smaller than 0.5 should be less spreaded and closer to 0% and answers for α slightly greater than 0.5 should be closer to 100%.
To analyze in more details this phenomenon, paired t-tests have been applied to the mean of the answers for each α value. Although the case ΔL = ±2 dB might give relatively weak auditory cues, results did not suggest any significant difference between the case where there was no audio cue and cases with substantial level differences between the “red” and “blue” audio signals.
Another way to investigate the impact of auditory cues is to compare results at α = 0.5 when ΔL is positive or negative. Looking at figure 3 suggests that the most significant differences lie for ΔL equals to ±6 and ±12 dB. This is also illustrated in figure 4 which shows the distributions of answers for the various values of ΔL at α = 0.5. A paired t-test between the two distributions ΔL = -6/12 dB and ΔL = +6/12 dB shows a significant mean difference of 20.83% with t = 3.43 and p = 0.0017. This result suggests that at the exact point of bistable transparency (α = 0.5), people tend to base their judgment on the auditory information when the level difference between the two audio signals is high enough.

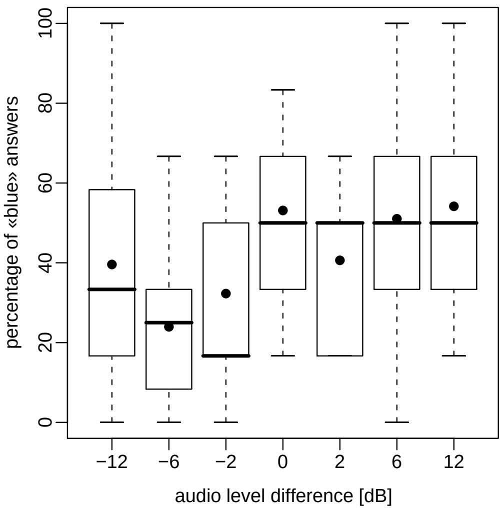

Figure 4: *Boxplots of the percentage of answers “blue” at α = 0.5 for the various values of auditory level differences. Solid circles represent the means of the distributions.*

Further investigations on the influence of the audio signals on the answers have been performed. The psychometric functions where the average percentage of “blue” answers is plotted against α have shapes close to ogives, it is thus expected that they should become linear when the average percentages are expressed as z scores [20]. This transformation enables to quantify the slopes of the resulting linear functions and compare them. Figure 5 illustrates the z scores calculated from p values for ΔL = 0 dB and ΔL = ±12 dB. The shapes of the z scores are not linear over the whole range of α values, however they approach a linear behavior if the range of α values is restrained to [0.4 ; 0.6]. Linear regressions performed on the distributions for ΔL = 0 [slope = 10.07 with $s.d.$ = 1.11, $F(1,3)$ = 82.49, p = 0.003] and for ΔL = ±12 dB [slope = 10.48 with $s.d.$ = 1.31, $F(1,3)$ = 63.65, p = 0.004] do not show a significant difference between the two slopes (angle $\simeq 0.22°$). Therefore this latter result does not corroborate the influence of the audio cues found at α = 0.5 by comparing positive and negative audio level differences: it seems that while people may use auditory cues at α = 0.5, they do not use them for values of α slightly smaller or greater than 0.5. This may be partially explained by the rather good reliability of visual cues for these values, in particular the average percentage of answers "blue" when ΔL = 0 at α = 0.45 is 19.8%, thus significantly lower than the average percentage at α = 0.5 (a paired t-test gives a mean difference of 33.3% with t = 5.48 and p = 6.37e-05). However the mean difference between distributions at α = 0.5 and α = 0.55 is not significant, suggesting that the interval of visual confusion is not symmetric around 0.5.

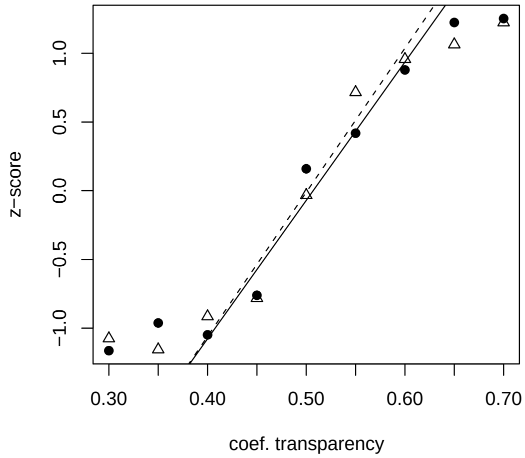

Figure 5: *Z scores of average percentage of answers "blue" as a function of transparency coefficient. Solid circles: results for ΔL = 0. Triangles: results for ΔL = ±12 dB. Solid line: linear regression of the means for ΔL = 0 and $\alpha \in [0.4; 0.6]$. Dashed line: linear regression of the means for ΔL = ±12 dB and $\alpha \in$* [0.4; 0.6].

Participants were all Italians and the auditory stimuli were the English words "red" and "blue", therefore the language aspect was also investigated. Since the Italian word for "blue" is "blu" pronounced identically, the language factor might be significant only if results differ according to the color, i.e. if there is an increase of answers "blue" for positive ΔL values in respect to answers "blue" for ΔL = 0, and no or less difference between answers "red" for negative ΔL values in respect to answers "red" for ΔL = 0. No such trend is found when one compared left portions of the four boxplots in figure 3 with their right portions, consequently the language factor is disregarded.

Finally, analysis of subjects' response time do not show any significant difference between the case $\Delta L = 0$ and $\Delta L \neq 0$. Results are shown in figure 6 and simply suggest that some people take more time to answer when the coefficient of transparency is close to 0.5 because it is visually more difficult to determine which square is in front of the other.

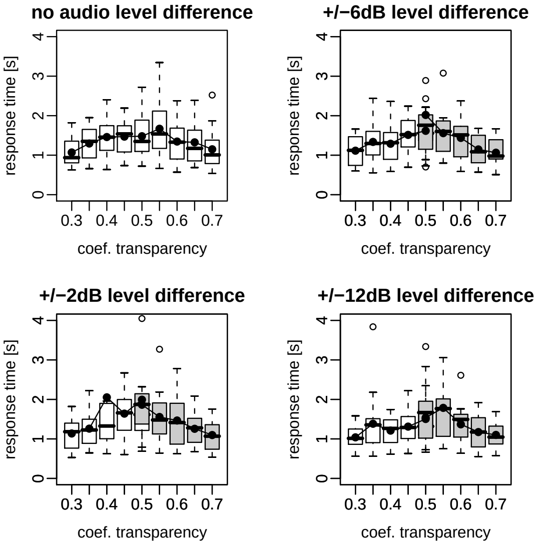

Figure 6: *Boxplots of the response time values as a function of the coefficient of transparency. White boxplots represent results for negative or null audio level differences, and grey boxplots represent results for positive audio level differences. Means are connected by a line. Outliers are marked by 'o'.*

4. CONCLUSIONS

A significant influence of auditory cues on visual perception of depth has been found only for the restricted case of theoretical bistable transparency (α = 0.5) and substantial sound level difference between the two speech signals. In this situation of weak visual cues, subjects tend to make their decision based on the level difference between the two auditory signals "red" and "blue". Except for this case, no influence of the auditory stimulus has been found. The most obvious explanation might be the lack of unity: auditory and visual stimuli are not perceived as coming from a unitary event, therefore judgments are most of the time made using only visual cues.

For signals coming from different senses to be integrated, the brain has to establish a correspondence between these signals and decide whether they come from the same object or event. This auditory-visual integration depends on the level of abstraction of the auditory and visual representations that are involved. In the specific case of auditory-visual speech perception, one may argue that speech would more easily fuse with the vision of the mouth generating the words, like in the case of the ventriloquism effect. According to recent studies [21, 22], two hypotheses could explain the integration of the senses in this case and are based on low-level processes: first there is a strong temporal correlation between the auditory and the visual signals (e.g. of the sound level with the degree of lip aperture), and secondly a coherence of movement (a spectro-temporal variation of the audio signal may be correlated with a movement of the lips). Thereby several mechanisms of auditory-visual integration may cooperate, at different levels of processing in the brain. For auditory-visual perception of speech, the integration process seems to be based on low-level processes which could explain the robust integration of audio and visual information. In our experiment on the contrary, a higher semantic level of processing in the brain is necessary to give a meaning to the words "red" and "blue ", which could explain why our experiment does not reveal any auditory-visual unity: the content of the auditory signals is not taken into account in the integration process and is therefore irrelevant for judging the visual order of the two squares.

Nevertheless, results of the present experiment conflict with previous investigations by Ecker and Heller [10]. Auditory stimuli used in the two experiments were of different nature: Ecker and Heller used recorded sounds of rolling and impacting whereas we used speech signals. However Ballas and Howard [23] suggested that speech and everyday sounds, including rolling and impacting, are similar in several aspects, in particular everyday sounds may be thought of as a form of language because they are integrated on the basis of cognitive processes similar to those used to perceive speech. Causes explaining the difference between results from Ecker et al and ours are not obvious. However differences in the design of the two experiments might give some clue. First, the instructions given to the subjects were different: while Ecker and Heller instructed their subjects to "make a judgment about a ball and the path it travels", therefore not specifying on which sense to base their judgment, we gave the participants the instruction to "determine which square appears in front" therefore implying a visual judgment. Besides their experiment dealt with dynamic auditory and visual information whereas ours used static stimuli. Therefore in addition to temporally and spatially coincident auditory-visual cues, dynamic information from both senses may reinforce the auditory-visual unity. In order to verify this assumption, a proposed follow-up of the present experiment is to introduce a dynamic factor by delaying one visual square and its corresponding auditory stimulus. However it is uncertain whether this dynamic cue may improve the auditory-visual integration: as it was mentioned earlier the auditory and visual signals refer to different levels of representation, and most of all the process of sensory integration is still largely unknown. In particular, if one distinguishes between structural (e.g. spatio-temporal correspondence) and cognitive (e.g. semantic congruency) factors, further research is needed to understand their respective contributions in the process of sensory integration, in addition these two types of factors may not be clearly separated [8].

Despite the possible limitation of audio-visual integration of depth cues that was demonstrated in this paper, we believe that techniques based on auditory-visual cues of depth, such as visual partial occlusion and auditory intensity, could be very valuable for applications such as mobile TV, in particular to exaggerate perspective effects and improve thereby the perceived quality of interaction and content fruition.

5. REFERENCES

[1] E.A. Bier, M.C. Stone, K. Pier, W. Buxton, and T.D. DeRose, "Toolglass and magic lenses: the see-through interface," in *Proc. of the 20th annual conference on Computer graphics and interactive techniques*, 1993.

[2] M.A. Sasse and H. Knoche, "Quality in context - an ecological approach to assessing qos for mobile tv," in *Proc. of the 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, Germany*, Sept. 2006, pp. 11–20.

[3] C. Spence and M. Zampini, "Auditory contributions to multisensory product perception," *Acta Acustica United with Acustica*, vol. 92, pp. 1009–1025, 2006.

[4] C. Schmandt, "Audio hallway: A virtual acoustic environment for browsing," in *Proc. of the 11th annual ACM symposium on User interface software and technology, San Francisco, California, United States*, April 1998.

[5] M. Fernström and C. McNamara, "After direct manipulation - direct sonification," *ACM Transactions on Applied Perception*, vol. 2(4), pp. 495–499, 2005.

[6] W.D. Hairston, P.J. Laurienti, G. Mishra, J.H. Burdette, and M.T. Wallace, "Multisensory enhancement of localization under conditions of induced myopia," *Experimental Brain Research*, vol. 152, pp. 404–408, 2003.

[7] M.O. Ernst and H.H. Bülthoff, "Merging the senses into a robust percept," *TRENDS in Cognitive Sciences*, vol. 8(4), pp. 162–169, 2004.

[8] C. Spence, "Auditory multisensory integration," *Acoust. Sci. & Tech.*, vol. 28(2), pp. 61–70, 2007.

[9] R. Sekuler, A.B. Sekuler, and R. Lau, "Sound alters visual motion perception," *Nature*, vol. 385, pp. 308, 1997.

[10] A.J. Ecker and L.M. Heller, "Auditory-visual interactions on the perception of a ball's path," *Perception*, vol. 34, pp. 59–75, 2005.

[11] F. Frassinetti, "Enhancement of visual perception by crossmodal visuo-auditory interaction," *Experimental Brain Research*, vol. 147(3), pp. 332–343, December 2002.

[12] P.J. Laurienti, R.A. Kraft, J.A. Maldjian, J.H. Burdette, and M.T. Wallace, "Semantic congruence is a critical factor in multisensory behavioral performance," *Exp. Brain Res.*, vol. 158, pp. 405–414, 2004.

[13] S. Handel, *Perceptual Coherence*, Oxford University Press, 2006.

[14] S. Zhai, W. Buxton, and P. Milgram, "The partial-occlusion effect: Utilizing semitransparency in 3D human-computer interaction," *ACM Transactions on Computer-Human Interaction*, vol. 3(3), pp. 254–284, 1996.

[15] J.D. Foley, A. Van Dam, S.K. Feiner, and J.F. Hughes, *Computer Graphics Principles and Practice*, Addison-Wesley, 1990.

[16] S.C. Masin, "The luminance conditions of transparency," *Perception*, vol. 26, pp. 39–50, 1997.

[17] B.L. Anderson, "A theory of illusory lightness and transparency in monocular and binocular images: The role of contour junctions," *Perception*, vol. 26, pp. 419–453, 1997.

[18] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. Acoust. Soc. Am.*, vol. 111(4), pp. 1832–1846, 2002.

[19] M.B. Gardner, "Proximity image effect in sound localization," *J. Acoust. Soc. Am.*, vol. 43, pp. 163, 1968.

[20] G.A. Gescheider, *Psychophysics: method, theory, and application*, Lawrence Erlbaum Associates, 1985.

[21] K.W. Grant and P.-F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *J. Acoust. Soc. Am.*, vol. 108(3), pp. 1197–1208, 2000.

[22] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, pp. B69–B78, 2004.

[23] J.A. Ballas and J.H. Howard, "Interpreting the language of environmental sounds," *Environment and Behavior*, vol. 19(1), pp. 91–114, 1987.