In Darkness Visual-Auditory Fusion Variations in Space Depend on Relation between Unimodal Egocentric Reference Frames

David Hartnagel (dhartnagel@imassa.fr)
Département de Sciences cognitives, Institut de médecine aérospatiale du Service de santé des armées
BP 73 Brétigny-sur-Orge, 91223 France
Université Paris 8

Alain Bichot (abichot@imassa.fr)
Département de Sciences cognitives, Institut de médecine aérospatiale du Service de santé des armées
BP 73 Brétigny-sur-Orge, 91223 France

Corinne Roumes (croumes@imassa.fr)
Département de Sciences cognitives, Institut de médecine aérospatiale du Service de santé des armées
BP 73 Brétigny-sur-Orge, 91223 France

Abstract

The present experiment investigates the frame of reference involved in Visual-Auditory (VA) fusion over space. This multisensory phenomenon refers to the perception of unity (Welch & Warren, 1986) resulting from visual and auditory stimuli despite their potential spatial disparity. The extent of this illusion depends on the eccentricity in azimuth of the bimodal stimulus (Godfroy, Roumes, & Dauchy, 2003). A recent study, performed in a luminous environment, has shown that variation of VA fusion is gaze dependent (Roumes, Hartnagel, & Godfroy, 2004). The present experiment questions the contribution of ego- or allocentric visual cues by repeating the experiment in total darkness. Auditory and visual stimuli were displayed in synchrony sustaining various spatial disparities. Subjects had to judge about their unity ("fusion" or "non fusion"). Results showed that VA fusion in darkness remains gaze-dependent despite the lack of any allocentric cues and reinforced the assumption that the reference frame of the bimodal space is neither head-centered nor eye-centered but results from an integration phenomenon.

Keywords: Perception, Multisensory, Visual-Auditory fusion, Space, Psychophysics.

Introduction

Perception of the world is basically multimodal; the environment is a unified perception of the various unimodal inputs (Gibson, 1966). The problem is that each sensory system (mostly studied per se) has a peculiar frame of reference for allocating a position in space. What are the rules of space perception then? Which reference frame does the brain use to localize multimodal stimuli? Whatever the modality Paillard showed that the frame of reference varies according to the context of the task or its aim (Paillard, 1987). In a multimodal perception context perception is not only driven by the modality that would seem to be the most adapted to the context and the aim. Most studies showed that vision is the most accurate modality for space perception and object location (Gibson, 1966), and that audition performs better in the temporal dimension (Blauert, 1983). However, it has been shown that the association of visual and auditory cues improves accuracy in time (Perrot, Saberi, Brown, & Strybel, 1990) and in space (Godfroy & Roumes, 2004). Thus, the contribution of the least efficient sensory modality is not negligible. To combine visual and auditory cues in space, the brain can tolerate a certain amount of spatial disparity between both unimodal parts of the stimulus. This phenomenon called "perceptual fusion" is usually investigated through the ventriloquism effect (Jack & Thurlow, 1973): the perception of the spatial location of a sound is biased in the direction of a visual stimulus (visual capture). The bias applied by a sensory modality on the other one is driven by the task. By focusing on the localization property of the stimulus, the observer allocates a higher weight to the visual cue. At the opposite, the perception of unity (i.e. the merging of the visual and the auditory cues into a unified feeling) allows to study VA fusion without giving preference to any modality.

Visual-Auditory fusion has been recently investigated by Godfroy, Roumes and Dauchy (2003). They showed that VA fusion capability varied with the eccentricity of the bimodal stimulus in the participant's perceptive field. They determined fusion areas for a set of stationary sounds issuing from loudspeakers spread over the anterior perceptive field by estimating the spatial extent over which a luminous spot could be displayed and still be perceived as fused with the sound over 50% of the trials. The smallest fusion areas laid in the median sagittal plane, and VA fusion areas were found to be symmetrical in relation to that plane. These variations over space followed closely the spatial resolution of the auditory system (Blauert, 1983). Audition was assumed to play a major role on VA fusion magnitude. Primary unimodal egocentric reference frames are drastically different for the visual and for auditory systems. As vision information is initially coded at retina level, and thus depends on the eye position, the vision reference frame is considered as eye-centered. For audition, spatial information estimation depends mainly on inter-aural time differences (ITD), inter-aural level differences (ILD) for azimuth, and on spectral cues for

elevation (Blauert, 1983). These cues vary with head position; so, the auditory reference frame is considered as head-centered. In Godfroy et al.'s experiment (2003), the participant performed the VA fusion task looking straight ahead; so both reference frames (eye and head) were superimposed. Even if the fusion areas followed the auditory resolution, it could not be inferred that VA fusion was based on a head centered auditory reference frame.

Does the gaze direction affect VA fusion in space? This question was the purpose of an intermediate experiment (Roumes, Hartnagel, & Godfroy, 2004).

Participants sat in the middle of a hemi-cylindrical screen on which a green background (80°H x 60°V) was displayed. They kept their head and body 10° horizontally rotated relative to the apparatus axis of symmetry by means of a bite board (either rightward or leftward). A red fixation cross was displayed: either straight ahead (i.e. in alignment with the 10° rotation) or 20° laterally shifted (i.e. in alignment with the alternate head orientation, 10° over the axis of symmetry of the screen). In the first case, the visual and the auditory frame of reference were aligned; in the second case, they were dissociated by 20°. Gaze orientation was monitored by an eye-tracker system. Then, a spot of green light was presented in synchrony with a pink noise delivered by one of the 15 loudspeakers located behind the screen. Participants had to judge about the unity of the bimodal stimulus (fusion or non fusion). Results showed that the VA fusion areas vary with gaze position (i.e. gaze shift alters fusion areas in azimuth) and that the reference frame of VA fusion is neither head-centered (auditory) nor eye-centered (visual) but appears to be a dynamic integration of both sensors (ears and eyes).

However, VA fusion areas were not entirely symmetrical relatively to the median sagittal plane when the participant fixated straight ahead (i.e. in the aligned auditory and visual reference frames condition) as in Godfroy et al.'s experiment (2003). The orientation of the subject's head and the gaze were both laterally shifted relative to the axis of symmetry of the experimental apparatus. The edges of the projected luminous background may have provided the observer with an allocentric visual reference frame. Investigating the "Roelofs effect" Bridgeman, Peer and Anand (1997) showed that peripheral cues biased visual localization in a direction opposite to the shift of the visual frame. It was suggested that a shifted frame biased the apparent midline (Dassonville, Bridgeman, Kaur Bala, Thiem, & Sampanes, 2004). So, allocentric visual cues (i.e. the edges of the visual display) may partially account for previous results on VA fusion (Godfroy et al. 2003; Roumes et al. 2004). The present experiment aims to determine if VA fusion only depends on the relationship between unimodal egocentric reference frames and does not depend on the peripheral allocentric frames. We propose that VA fusion space in darkness (without any allocentric visual cues) is neither head-centered nor eye-centered but comes from an integrative phenomenon, based on both sensory modalities. In order to investigate this hypothesis, subjects performed a VA fusion task with unimodal reference frames aligned or dissociated. This task was run in total darkness and in a uniform uninformative low noise background to avoid any effect of allocentric information.

Methods

Subjects

Seven volunteers participated in this study, 4 women and 3 men, aged from 25 to 45. They all had normal or corrected to normal vision and no auditory defect.

Apparatus

Stimuli Control The subject was located at the axis of symmetry of an acoustically-transparent, hemi-cylindrical screen, 120 cm in radius and 145 cm in height. The subject's head was maintained by a custom bite-board with the eyes at mid height of the screen.

The head and body was rotated 10° leftward of the axis of symmetry of the screen to increase the space of investigation when the fixation spot was presented 20° to the right. No alternate rightward orientation shift was tested because no laterality effect had been found in the reference luminous experiment (Roumes et al. 2004). The orientation of the gaze was monitored with an ASL 504 (50 Hz) eye-tracker placed 45 cm in front of the subject at a level lower than the investigated field of view to prevent from any visual masking. To avoid any effect of allocentric cues, the experimental room was in total darkness, and noise level was reduced as much as possible (< 39 dB_A).

Alignment or dissociation between the visual and the auditory reference frames was controlled before the bimodal stimulus onset. The fixation spot was provided by one of two red laser beams (λ=535 nm, < 1 mW), placed behind the subject. The spot was displayed either straight ahead or 20° laterally shifted on the right. Stimuli could only be presented if the subject was looking at the red fixation spot with an angular error less than 1.66° for a mean duration of 500 ms (fixation time was randomly sampled in a 300-700 ms interval).

Such a feedback between the eye-tracker sampling and the experimental software was used to control initial position of the gaze and to guarantee the spatial configuration of the two references frames at the bimodal stimulus onset. The bite-board controlled the head position (i.e. the auditory reference frame) and the eye-tracker controlled the eye position (i.e. the visual reference frame)

The visual part of the bimodal stimulus was provided by a green laser beam (Melles-Griot, λ = 532 nm, 5 mW) attenuated with optical filters in order to reduce the luminance of the visual stimulus to the expected value. The position of the laser beam was adjusted through mirrors mounted on rotating motors driven by an electronic unit (Acutronic). This apparatus was placed over and behind the subject in order to avoid head masking. The visual stimulus onset/offset was controlled by a rotating rigid flag mounted on a brushless

motor driven by an electronic unit (Aerotech), to avoid sound effect of classical laser shutters.

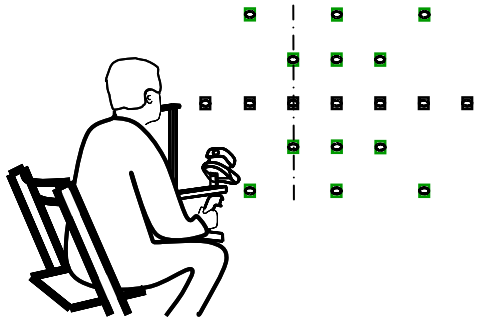

Figure 1: Experimental apparatus: Subject sat at the axis of symmetry of an acoustically transparent hemi-cylindrical screen. His/her head was maintained by a custom bite-board and the gaze was monitored by an eye-tracker. Behind the screen, 19 loudspeakers were oriented toward the subject's head.

The auditory part of the bimodal stimulus was delivered by one of the 19 loudspeakers (LS) located behind the screen, oriented toward the subject's head (Fig. 1). Seven loudspeakers were located at eye height 10° laterally spaced in azimuth from -30° to +30° relative to the middle of the apparatus (negative values for left location, positive values for right location). Six LS were placed ±10° above and below eye level, two at -10° azimuth, two at 0° and two at +10°. Six loudspeakers were placed ±20° in elevation, two at -20° azimuth, two at 0° and two at +20°.

Stimuli The multimodal stimulus consisted in a 49 dB_A broad-band pink noise presented for 500 ms in synchrony with a 1° spot of light, 3 $cd.m^{-2}$ in luminance.

Variables

Disparity For each of the 7 loudspeakers at eye level, 61 disparities between the spot and the sound source were tested (Fig. 2a). The spot could be displayed 0° to 20° horizontally apart from the center of the associated LS.
At 0° of elevation relative to the LS, horizontal disparity was tested with a 2.5° step in azimuth. At ±10° or ±20° elevations, the spot could be displayed with a 5° step in azimuth. Eight additional positions of the spot were also considered in vertical alignment with each LS in order to achieve a 5° step sampling of disparity from -30° to +30° in elevation.
For all other LS, only 9 disparities were tested (Fig. 2b). The spot could be displayed 0° to 20° apart from the center of the LS, with a 5° step in azimuth.
This sampling of disparities was selected in order to achieve a better resolution than previous experiments (Godfroy, Roumes, & Dauchy, 2003; Roumes, Hartnagel, & Godfroy, 2004) in defining the two dimensional shape of VA fusion areas.

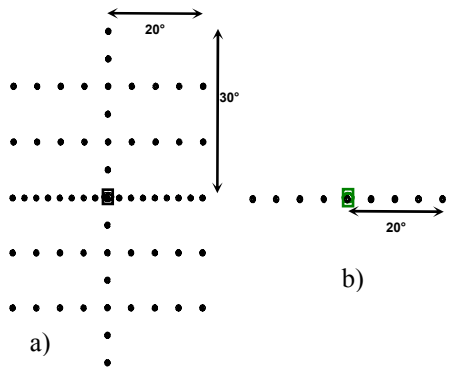

Figure 2: Disparity tested for a) the 7 loudspeakers at eye level, b) the 12 other loudspeakers. Squares represent the loudspeaker and dark dots represent all the potential positions of the spot of light.

Eccentricity Variation of fusion over the perceptive field was addressed through the location of the referred LS (i.e. 9 eccentricities from -30° to +30° in azimuth with a 10° step).

Dissociation/Alignment of the Frames of Reference When the red fixation spot appeared (straight ahead or 20° laterally shifted), the subject had to look at it to activate the multimodal stimulus presentation. According to the visual fixation, the visual and auditory reference frames could be either aligned (i.e. the subject looked at the fixation point straight ahead) or 20° laterally dissociated (i.e. the subject looked at the fixation point 20° to the right).

Task

The subject had to judge the perception of unity emerging from the bimodal stimulus using a joystick. When the spot and the pink noise were perceived as coming from a unique and common location in space, "fusion" response was selected by pulling the joystick. When the two unimodal stimuli were perceived as issuing from two distinct locations, the "non fusion" response was selected by pushing the joystick. The red fixation spot was randomly presented at either of its two locations. The bimodal stimulus was only presented if the subject looked at the point for the required duration; stimulus presentation was gaze-dependent (Fig. 3). One trial took about 3 s to be completed.

Analysis

All bimodal combinations were repeated 5 times for each subject. Previous experiment on VA fusion has tested the effect of subject orientation so that all bimodal combination had been tested 10 time, results were very close to results issuing from the 5 repetitions. So we consider that the limited number of replications do not affect the reliability of the data.
A rate of fusion was derived for each disparity tested. Then, fusion limit was estimated using a probit analysis, from the 50% "fusion" response rate.

For each of the 7 loudspeakers at eye height, 12 limits were estimated, 10 in azimuth (5 on the left and 5 on the right) and 2 in elevation. For the other 12 LS, only 2 limits in azimuth were calculated. These limits allowed defining the so-called "fusion area" for each location of a loudspeaker. Statistical analysis was performed on fusion limits for each subject (7) in the two experimental conditions tested (Reference frames aligned or dissociated). For 7 subjects, it represented 1316 limits in azimuth and 196 in elevation

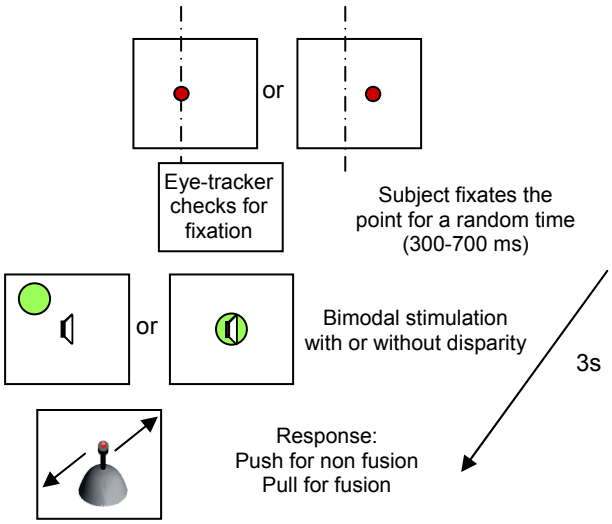

Figure 3: Test trial: Subject had to fixate a red point either straight ahead or 20° right shifted for a random time checked by the eye-tracker to trigger the bimodal stimulus; subject responded "fusion" or "non-fusion" using a joystick.

Results

Raw data from all subjects are presented in Figure 4. The graphs represent the fusion rate as a function of disparity from the center of the LS, in azimuth (Fig. 4a) and in elevation (Fig. 4b). It represents the distribution of the fusion rate for all trials and all subjects for each of the 19 loudspeakers. As shown by the interpolation of all the raw data (Fig. 5) fusion rate in the 2 dimensions of space varied with the horizontal eccentricity of the LS. To analyze this variation, fusion limits were defined.

Individual fusion rates were considered to determine the fusion limits and then the means fusion areas (Fig. 6). Fusion areas defined for each loudspeakers (Fig. 6) match well with the interpolation of the raw data shown in Figure 5. When fusion limits in elevation could not be established (i.e. subjects still responded "fusion" for more than 50% of the trials for disparity up to 30°, fusion limit was arbitrarily set at 30°. For some subjects, at some levels of elevation, fusion limits in azimuth could not been defined due to the lack of responses "fusion". So, when statistical analysis compares fusion limits from the whole set of subject, degrees of freedom is under the foresee one.

The distribution of fusion areas changed depending on the relative positions of the unimodal reference frames. When the reference frames were aligned, the overall spatial distribution of areas was symmetrical in relation to the loudspeakers straight ahead. Whereas, when the reference frames were 20° laterally shifted, the spatial distribution was symmetrical in relation to the loudspeakers located in-between the auditory reference frame and the visual reference frame; that is, in relation to the LS located at 0° in azimuth (Fig. 6).

VA fusion limits in azimuth varied with the horizontal eccentricity of the LS ($F_{(6,1243)}$ = 21.783; p<.001). This variation along the perceptive field appeared laterally deviated in the direction of gaze shift in the reference frames dissociated condition. However, statistical analysis showed no significant difference between the 2 conditions ($F_{(1,1243)}$ = 0.973; NS).

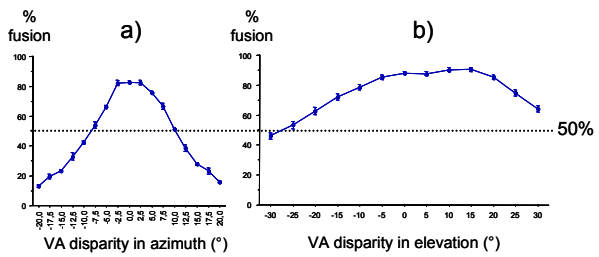

Figure 4: Probability fusion rate for all subjects and all LS a) as a function of azimuth disparity, b) as a function of elevation disparity. The 0° disparity corresponds to the center of the tested LS.

As the interaction between reference frame conditions and eccentricity was significant (∩ F(6,1243) = 4.93; p<.05) we analyzed the sign of the differences in the limits of VA fusion between the aligned and the dissociated conditions along the horizontal eccentricity. As shown in Figure 7, differences between the two conditions were significant for 5 paired eccentricities out of the 7 tested. Moreover, the sign of the differences changed as if the curve of the dissociated condition would have been shifted 10° to the right (toward gaze shift).

When reference frames were aligned, fusion areas ±20° and ±10° apart from straight ahead were symmetrical relative to the median sagittal plane (t_9 = 1,594; p = .1455 and t_{13} = 1.942; p = .0742, respectively in paired comparisons of limits in azimuth between LS ±20° and ±10° relative to straight ahead). Degrees of freedom of t-test change because the number of fusion limits estimated change with horizontal eccentricity (i.e. it depends on the number of LS tested).

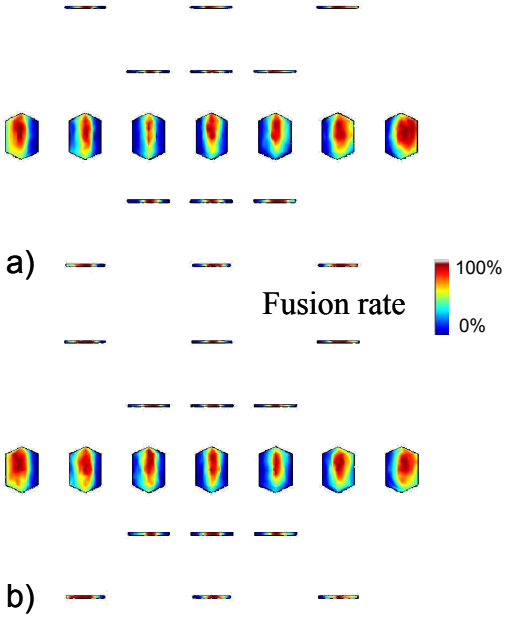

Figure 5: Interpolations from all responses "fusion" for each LS a) when the reference frame are aligned, b) when they are dissociated.

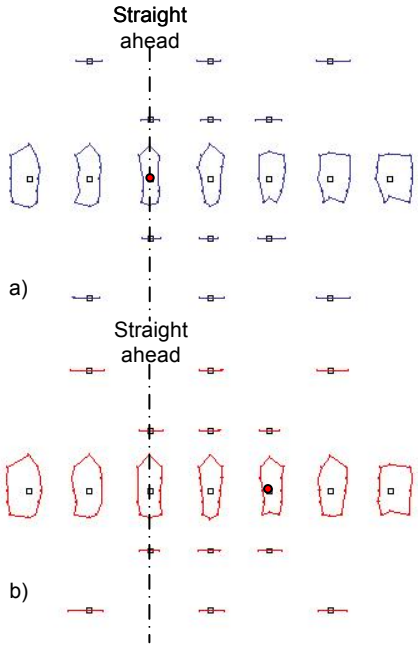

Figure 6: Fusion areas for all loudspeakers a) in the reference frames aligned condition, b) in the reference frames dissociated condition. The point of fixation, in each condition, is figured as the filled red circle. Space between the centers of fusion areas were voluntary increased to disambiguate data reporting.

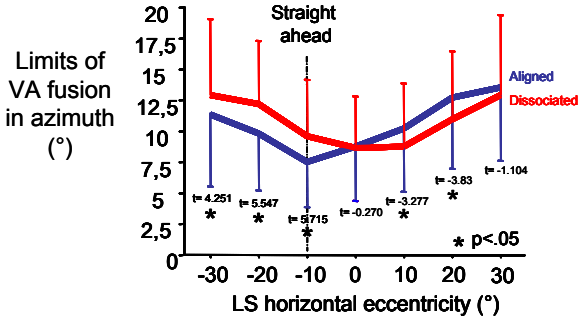

Figure 7: Limits of fusion in azimuth as a function of horizontal eccentricity of the LS. Stars refer to a significant difference between limits of fusion in the reference frames aligned condition and the reference frames dissociated one in a paired comparison.

Discussion

In daily life, humans move their eyes constantly so superimposition of head and eye reference frames is seldom. Most experiments use a fixation point straight ahead. Such is the case in basic VA fusion investigation performed by Godfroy, Roumes and Dauchy (2003). The present experiment shows that, even in a dark and silent room, visual-auditory fusion in the azimuth dimension is gaze dependent.

Aligned Reference Frames Condition For the whole set of loudspeakers, (i) the narrowest fusion areas in the azimuth dimension were in the median sagittal plane; (ii) fusion limits in azimuth increased with lateral eccentricity in the perceptive field and (iii) the fusion areas were symmetrical in relation to the sagittal median plane. From these three characteristics, the head-centered frame of reference hypothesis for VA fusion is emphasized. This result is consistent with a previous experiment where subjects had their gaze fixed straight ahead congruently with the allocentric visual cues (Godfroy, Roumes, & Dauchy, 2003). When compared to a previous experiment (Roumes, Hartnagel, & Godfroy, 2004) where background gave allocentric asymmetrical cues (relative to the subject median sagittal plane), present results showed fusion areas that are more symmetrical relative to the median sagittal plane. So, allocentric cues may alter VA fusion mechanism.

Dissociated Reference Frames Condition Fusion areas vary with the eccentricity of the loudspeakers in the perceptive field. This variation significantly differs from the aligned reference condition. The narrowest fusion areas in the azimuth dimension are those from the loudspeakers laying on the vertical axis, including the point of fixation. But there is no significant difference between those fusion areas and the fusion areas located on the vertical axis in-between straight ahead and the gaze position (i.e. the vertical axis including 0°

in azimuth in the apparatus). These latter fusion areas are the most intrinsically symmetrical ones and all other fusion areas are symmetrically organized on each side. These results are in line with the previous experiment (Roumes et al. 2004) where background gave allocentric asymmetrical cues. The current experiment, preventing any allocentric bias, emphasizes a dual contribution of vision and audition to define the locations in space where bimodal stimuli can still be perceived as one. So, the reference frame for fusion space can neither be considered as eye-centered nor head-centered but resulting from a relative contribution of these two egocentric reference frames.

2D Fusion Areas The 2D shape of VA fusion areas can be derived from the seven loudspeakers at eye level, for which both azimuth and elevation limits could be estimated. They were anisotropic: limits in elevation were always greater than those in azimuth. In their experiment, Godfroy, Roumes and Dauchy (2003) inferred the two dimensional shape of the VA fusion areas only from disparities between the unimodal stimuli in horizontal and vertical spatial alignment with the center of each loudspeaker. So, the fusion areas were figured as diamond-shapes centered on their respective loudspeaker. In the current experimental design, fusion areas could be more precisely defined from a larger combination of azimuth and elevation disparities. They are not diamond-shaped, the fusion limits in azimuth keeping rather constant whatever the elevation component in the disparity. Along the horizontal eccentricity VA fusion areas tend to increase in azimuth (Fig. 7). These variations of VA fusion areas over the perceptive field follow closely the spatial resolution of audition (Blauert, 1983; Perrot et al. 1990).

Conclusion

Even if VA fusion follows closely the properties of the auditory space it also depends on the relative position of both unimodal sensory captors. This latter effect is mainly due to egocentric cues as it remains effective in darkness without any visual allocentric biases. The frame of reference for VA fusion space is neither head-centered nor eye-centered but is the result of a multisensory integration.

Acknowledgement

Special thanks to Patrick Sandor for his helpful technical support and to Sylvain Hourlier for rewording. We also greatly thank all the volunteers who participated in this study.

References

Blauert, J. (1983). *Spatial Hearing. The psychophysics of human sound localization*. London: The MIT press.

Bridgeman, B., Peery, S., & Anand, S. (1997). Interaction of cognitive and sensorimotor maps of visual space. *Perception & Psychophysics, 59*(3), 456-469.

Dassonville, P., Bridgeman, B., Kaur Bala, J., Thiem, P., & Sampanes, A. (2004). The induced Roelofs effect: two visual systems or the shift of a single reference frame? *Vision Research, 44*(6), 603-611.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Westport, Connecticut: Greenwood Press.

Godfroy, M., & Roumes, C. (2004). Multisensory enhancement of localization with synergetic visual-auditory cues. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Chicago, IL: Lawrence Erlbaum Associates.

Godfroy, M., Roumes, C., & Dauchy, P. (2003). Spatial variations of visual-auditory fusion areas. *Perception, 32*(10), 1233-1245.

Jack, C. E., & Thurlow, W. R. (1973). Effect of degree of visual association and angle of displacement on the ventriloquism effect. *Perceptual and Motor Skills, 37*, 967-979.

Paillard, J. (1987). Comment le corps bâtit l'espace. *Science et vie, 158*, 36-42.

Perrot, D. R., Saberi, K., Brown, K., & Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. *Perception & Psychophysics, 48*(3), 214-226.

Roumes, C., Hartnagel, D., & Godfroy, M. (2004). Audio-visual fusion and frames of reference. *Proceedings of the European Congress of Visual Perception - Perception Supplement* (pp 142). Budapest, Hungary: Pion.

Welch, R. B., & Warren, D. H. (1986). Intersensory interactions. In K. Boff, L. Kaufman & J. Thomas (Eds.), *Handbook of perception and human performance. Sensory processes and perception* (Vol. 1). New York: Wiley Interscience.