Universidade do Minho
Escola de Psicologia

Ana Catarina Macedo Mendonça

Audiovisual Perception of Biological Motion

Junho de 2011

Audiovisual Perception of Biological Motion
Ana Catarina Macedo Mendonça
UMinho | 2011

Universidade do Minho
Escola de Psicologia

Ana Catarina Macedo Mendonça

Audiovisual Perception of Biological Motion

Tese de Doutoramento em Psicologia
Especialidade de Psicologia Experimental
e Ciências Cognitivas

Trabalho efectuado sob a orientação do
Professor Doutor Jorge Almeida Santos
e do
Professor Doutor Miguel Castelo-Branco

Junho de 2011

Às Bulcão e aos Mendonça

This work was fully funded by the Portuguese Foundation for Science and Technology by the PhD scholarship SFRH/BD/36345/2007, the re-equipment project REEQ/821/PSI/2005, and the Biomotion project PTDC/SAU-BEB/68455/2006.

This work was also funded by the Portugal-Spain Actions PT2009-0186 from the Spanish Government and E-134/10 from the Portuguese Conselho de Reitores das Universidades Portuguesas.

ACKNOWLEDGMENTS

This time family comes first. Madrecita, Padrecito, Mary, and Janico, you couldn't have done more.

A special thanks to Prof. Jorge A. Santos.

This thesis is the result of a joint effort from several team colleagues and research collaborations. With no specific order or relevance, I thank:

The group involved in the development of the visual stimuli, the point-light walkers. I specially thank Bruno Aragão, but also Rogério Pereira, and Liliana Magalhães from University of Minho, for all the work in the implementation of the Vicon® system and in the biological motion acquisitions. I thank Prof. Miguel Correia and Eduardo Soares, from the Faculty of Engineering of the University of Porto, for the effort in the development of the LabVIEW routines for generating the point-light walkers.

Prof. Guilherme Campos, Prof. Paulo Dias and Prof. José Vieira, the research collaborators from the Institute of Electronics and Telematics Engineering (IEETA) at the University of Aveiro, who disposed of their knowledge and algorithms and cooperated so closely in implementing the auralization system. Might this collaboration last for many years to come. I also thank João Pedro Ferreira, from University of Minho, for all the involvement in the audio work.

Prof. Joan López-Moliner, from the Institute of Brain, Cognition and Behavior at University of Barcelona for his warm welcome in Barcelona, for promptly disposing of his knowledge and time, and for his tireless collaboration in the audiovisual experiments.

Rogério Pereira for the development of the custom-built latency analyzer, an invaluable equipment to assess the delays between the visual and audio signals.

The team at the Laboratory of Visualization and Perception. Thank you Lili, Bruno, Carlos, Marta, Sandra, Roger, for making it a special place to work, for the selfless help, the availability, the good moments.

The Computer Graphics Center (CCG), for the investment and collaboration in the Laboratory of Visualization and Perception.

My friends, who never let me be without experimental subjects, who believed and worried with me. Special thanks to Nundi, Bruno, Adriana, Paulo and Jaime. I also thank some students from the discipline *Laboratório de Percepção*, for going beyond the scope of their work.

Last but not least, I thank other researchers who helped at several crucial moments of this project, namely Prof. Miguel Castelo-Branco, Dr. François Tonneau, Prof. Elisabete Freitas, Prof. Caetano Monteiro, Prof. Armando Machado, Prof. Sergio Nascimento.

AUDIOVISUAL PERCEPTION OF BIOLOGICAL MOTION

Abstract

This thesis comprises three experimental sections: a) Locating Auditory Sources with Non-Individualized HRTF-Based Auralizations; b) The Effect of Auditory Cues on Biased Biological Motion; and c) The Multisensory Integration of Biological Motion.

In the first experimental section, the accuracy and adaptation to auralized sounds was tested. Auralization is a powerful tool to increase the realism and sense of immersion in Virtual Reality environments. The Head Related Transfer Function (HRTF) filters commonly used for auralization are non-individualized, as obtaining individualized HRTFs poses very serious practical difficulties. It is therefore important to understand to what extent this hinders sound perception. In this section, we addressed this issue from a learning perspective. In a set of experiments, we observed that mere exposure to virtual sounds processed with generic HRTF did not improve the subjects' performance in sound source localization, but short training periods involving active learning and feedback led to significantly better results. We proposed that using auralization with non-individualized HRTF should always be preceded by a learning period.

In the second experimental section we addressed the effect of auditory cues on bistable biological motion representations. Perceiving humans in motion is a frequent and crucial task. However, visual stimuli alone are might be poorly informative and often result in a face forward bias. In this section, we intended to explore if related and meaningful sounds would reduce those visual biases. Participants were presented with visual, auditory or audiovisual walkers, which could be moving forward or away from the perceiver. The task was to discriminate walking motion direction. Overall, results in the audiovisual condition were significantly better than those of the visual condition, with more correct estimates and a lower bias, but similar to the auditory results. We

concluded that step sounds are a relevant cue, able to diminish the perceptual error and break the face forward illusion.

In the last experimental section, we explored the multisensory integration of cues of walking speed. After testing for audiovisual asynchronies (visual signals led auditory ones by 30 ms in simultaneity temporal windows of 76.4 ms), in the main experiment, visual, auditory and bimodal stimuli were compared to a standard audiovisual walker in a velocity discrimination task. Results in variance reduction conformed to optimal integration of congruent bimodal stimuli across all subjects. Interestingly, the perceptual judgements were still close to optimal for stimuli at the smallest level of incongruence. Comparison of slopes allowed us to estimate an integration window of about 60 ms, which is smaller than that reported in audiovisual speech.

We conclude that the audiovisual interactions of biological motion stimuli allow accuracy improvement and uncertainty reduction. These multisensory integration processes might be predicted by optimal mechanisms.

PERCEPÇÃO AUDIOVISUAL DE MOVIMENTO BIOLÓGICO

Resumo

A presente tese contempla três secções experimentais: a) Localização de fontes sonoras com auralizações baseadas em HRTFs não individualizados; b) O efeito de pistas auditivas em representações enviesadas de movimento biológico; e c) A integração multimodal de movimento biológico.

Na primeira secção experimental, testou-se a precisão e adaptação a sons auralizados. A auralização é uma ferramenta poderosa para melhorar o realismo e a sensação de imersividade em ambientes de realidade virtual. Os filtros de HRTF (*Head Related Transfer Functions*), frequentemente utilizados no processo de auralização, não são individualizados, uma vez que obter HRTF individualizados coloca sérias dificuldades práticas e elevados custos. Como tal, é importante compreender de que forma é que estes filtros genéricos afectam a percepção dos sons auralizados. Nesta secção, abordámos este tema numa perspectiva de aprendizagem. Num conjunto de experiências, constatámos que a mera exposição aos sons virtuais processados com HRTF genéricos não melhorava a performance dos sujeitos experimentais na localização de fontes sonoras. Contudo, curtos períodos de treino envolvendo aprendizagem activa e feedback conduziram a resultados significativamente melhores. Como conclusão, propomos que todo o uso de sons auralizados com HRTF não individualizados seja precedido de um período de aprendizagem.

Na segunda secção experimental, abordámos o efeito de pistas auditivas em representações bi-estáveis de movimento humano. Perceber humanos em movimento é uma tarefa frequente e crucial. Contudo, a visão pode ser pouco informativa e frequentemente leva a um viés frontal. Nesta secção, pretendemos analisar se sons relacionados e congruentes poderiam reduzir este viés. Apresentaram-se estímulos visuais, auditivos e audiovisuais de movimento biológico, que podiam mover-se em aproximação ou afastamento

do observador. A tarefa consistiu em identificar a direcção do movimento do estímulo. Os resultados da condição audiovisual foram significativamente melhores que os da condição visual, com mais estimativas correctas e menos viés, mas foram semelhantes aos resultados da condição auditiva. Concluímos que os sons de passos são uma pista relevante, capaz de diminuir o erro perceptivo e eliminar o viés frontal.

Na última secção experimental, abordou-se a integração multimodal de pistas de movimento biológico. Depois de se testar as assincronias percebidas (o estímulo visual deve ser apresentado antes do estímulo auditivo em 30 ms), na experiência principal, compararam-se estímulos visuais, auditivos e bimodais com um estímulo standard, numa tarefa de discriminação de velocidade. Os resultados de redução de variância dos estímulos audiovisuais congruentes foram bem previstos pelo modelo de integração óptima. Surpreendentemente, os julgamentos perceptivos dos estímulos ao nível mais baixo de incongruência também estiveram muito próximos das previsões óptimas. A comparação de curvas de ajustamento permitiu-nos estimar uma janela de integração multimodal de cerca de 60 ms, que é mais pequena que aquela reportada para a integração multimodal de estímulos verbais.

Conclui-se que as interacções audiovisuais de estímulos de movimento biológico permitem o aumento de precisão e a redução de incerteza. Estes processos de integração multissensorial podem ser explicados por mecanismos óptimos.

TABLE OF CONTENTS

ABREVIATIONS, SYMBOLS, AND UNITS

º Degree

σ Standard Devitation

2-IFC Two interval forced choice

3D 3 dimention

CAVE Cave Automatic Virtual Environment

Cd/m^2 Candela per squared meter

CI Confidence interval

dB Decibel

dB SPL Decibel in standard Pascal level, with standard 20 μPascal

deg Degree

deg/s Degree per second

hRIR Head related impulse response

HRTF Heard related transfer functions

Hz Hertz

ILD Interaural Level Difference

ITD Interaural Time Difference

JND Just noticeable difference

LVP Laboratory of Visualization and Perception

m Meter

m/s Meter per second

min Minute (time)

ms Millisecond

MLE Maximum likelihood estimation

n.s. Not significant

p Probability

PET Positron emission tomography

PLW Point-light walker

PSE Point of subjective equality

PSS Point of subjective simultaneity

s Second (time)

SC Superior colliculus

SD Standard deviation

SOA Stimulus onset asynchrony

STS Superior temporal sulcus

STSp Posterior superior temporal sulcus

VR Virtual Reality

GLOSSARY

Anechoic. Atribute of a sound or space where only direct sound is present. Sound reflections are eliminated or mostly limited.

Audiometric screening. An auditory test, to assess auditory sensitivity trough the determination of absolute thresholds at selected frequencies, for the left and right ears.

Auralization. Recreation of a spatial sound through digital audio processing techniques. Accounts for the presence of the listeners' body in space.

Azimuth. The position of an auditory source, relative to the listener's head, in the horizontal plane. Its perceptual discrimination relies mostly on binaural cues, such as ITD and ILD. Expressed in degree units.

Binaural cue. An auditory cue provided by differences in sound signals reaching the left and right ear. ITD and ILD are binaural cues.

Biological motion. The complex pattern that an animal exhibits while moving. In Psychophysics, biological motion refers to a specific type of human motion display, the point-light walker (PLW).

Bistability. Property of a stimulus that allows several percepts at the same time. When a percept is formed it becomes stable, making perceiver unaware of the other possible percept.

Bootstrap. A statistical method for assigning measures of accuracy to sample estimates. Allows accessing the properties of an estimator by constructing a number of resamples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset.

CAVE. Cave Automatic Virtual Environment. An immersive virtual reality environment where 3 or more projectors might be directed at several

surfaces, usually disposed as walls of a room, providing the surrounding of the visually defined reality in a coherent continuum.

Cronophotography. Photography produced by sequential exposure periods of the same frame or by several sequential frames.

Elevation. The position of an auditory source, relative to the listener's head, in the vertical plane. Its perceptual discrimination depends mostly on monoaural cues. Expressed in degree units.

HRTF. A head-related transfer function (HRTF) is a function that characterizes how an ear receives a sound from a point in space. A pair of HRTFs, for the two ears, is used to synthesize binaural sounds and produce the sensation that an acoustic object originates from a particular point in space.

Immersive. An immersive environment implies the feeling of presence within a virtual space, as if it was the external physical space.

Interaural Level Difference. The difference in signal intensity level, or wave amplitude at several frequencies, between the sound reaching the left and right inner ears. Expressed in dB.

Interaural Time Difference. The difference in signal arrival time between the left and right inner ears. Expressed in time units (s).

Inverted PLW. Upside-down rotated (180º) point-light walker.

Maximum likelihood estimation (MLE). A psychophysical model. In multisensory integration, it assumes that two signals in a compound stimulus interact at a sensory stage, maximizing the information and reducing sensory-related noise, thus increasing detection and reducing uncertainty.

Monoaural cue. An auditory cue provided by the signal characteristics, not by the difference in signals reaching the ears (see binaural cue). A sound's intensity and spectral composition are monoaural cues.

Multisensory. Involves two or more sensory modalities.

Pink noise. A noise where each octave carries an equal amount of noise power. There is a range of audible frequencies with intensity inversely proportional to frequency.

Point-light walker. A biological motion stimulus created through the display of only a few moving points of the human body, the point-lights. Most frequently, these point-lights follow the movement of the principal joints of the body.

PSE. A value in a continuum at which a given event is subjectively equal to a test event.

PSS. A point in a time continuum when an event is subjectively simultaneous to another event.

Pseudo-random. Similar to random. Here, pseudo-random is used to refer to stimuli presented randomly, but in a way that assures an equal number of trials per condition.

Probability summation. A psychophysical model. Commonly used in the modeling of detection thresholds. It assumes that two signals in a compound stimulus do not interact at a sensory stage, but that a binary choice is made at a subsequent decision stage.

Sensory dominance. Term used when, in multisensory processes, one sensory modality interacts with another in a winner-take-all fashion.

Spectral cues. In acoustics, spectral cues are related to the frequency spectrum of a sound and the power-frequency relation.

Step-cycle. A complete walking cycle in the visual stimulus, defined in biomechanics as the sum of movements made during locomotion, from the time a limb leaves the ground until it leaves the ground on the next occasion.

Translational component. The attribute in visual motion related to the stimulus displacement with a given direction and orientation in space. In biological motion, when absent, the walker moves as if it was on a treadmill. When present, it moves at a given, biomechanically defined, translational velocity, changing its position in space accordingly.

FIGURES

FOREWORD

The present doctoral thesis is based upon an FCT-approved PhD project under the theme of the audiovisual perception of biological motion. It intended to analyze the multisensory interactions in the processing of human motion stimuli.

The work here presented is the result of a 4-year work (2008-2011), mostly developed at the Laboratory of Visualization and Perception (LVP) in Guimarães. During that period great changes occurred at LVP. A new building received the laboratory, the visualization, virtualization and motion capture systems were implemented and a treadmill was integrated in what became a CAVE-like virtual enviroment room. Most importantly to the present work, the LVP soon abandoned its silent days with the introduction of the audio system. This introduction was crucial for the development of this doctoral project, versed in the audiovisual perception of biological motion.

With the implementation of the audio system several challenges were posed, new external cooperations arised, and some parallel work, intended mainly at calibrating and validating the auditory system, was carried out. For this reason, the first of 3 experimental sections in this thesis concerns only auditory experiments (*3.1. Locating auditory sources with non-individualized HRTF-based auralizations*). The text presented in that section is an unedited version of a paper published in the Journal of the Audio Engineering Society, in the Proceedings of the 129th (2010), under the title *On the improvement of auditory accuracy with non-individualized HRTF-based sounds*. It is also currently under review for publication in the same journal as a regular article. That work is co-authored by Prof. Jorge Santos, Prof. Guilherme Campos, Prof. Paulo Dias, Prof. José Vieira, and Eng. João Ferreira.

The second experimental section of the present work (*3.2. The effect of auditory cues on Biased Biological Motion Orientations*) is a refinement of previous work from my master thesis (Mendonça, 2007) and of a previous

paper (Mendonça & Santos, 2010) on the effect of auditory cues over the perceived direction of biological motion. It is a short experiment where biased displays of human motion are combined with auditory cues. The text presented in that chapter is based on an unpublished report co-authored by Prof. Miguel Castelo-Branco and Prof. Jorge Santos.

The third and final experimental section of this thesis (*3.3. The multisensory integration of biological motion*) describes experiments on optimal integration mechanisms with bimodal biological motion stimuli. It is an unedited version of the paper *The Benefit of Multisensory Integration with Biological Motion Signals*, published in Experimental Brain Research (2011). It is co-authored by Prof. Jorge Santos and by Prof. Joan López-Moliner.

I. INTRODUCTION

1.1. CONCEPTUAL FRAMEWORK

1.1.1. Biological Motion

The concept of biological motion refers to the complex pattern that any animal exhibits while moving. It is frequently studied in biomechanics, as well as in psychophysics. From the psychophysics point of view, biological motion is both a highly complex stimulus and one of great relevance in everyday life. Nowadays this term most frequently refers to a specific type of human motion display.

Biological motion as an object of empirical study took its first steps in the mid XIXth century. The first controlled recordings of motion capture can be traced back to 1878, when the photographer Muybridge intended to analyze in detail the motion pattern of galloping horses. For that purpose, Muybridge used an array of photographic equipments disposed throughout a hippodrome and achieved the first systematic biological motion decomposition. But it was the physiologist Marey (1884), a Muybridge's contemporary, who first studied the human locomotion. In order to decompose the biological motion, he invented the chronophotography, a photography produced by sequential exposures or snapshots which allowed him to obtain still or frozen images that respected the time intervals between positions. To simplify the images, eliminate the body forms and focus only on the motion cues, Marey marked the subjects with white bands in their members. This allowed the tracing of specific areas of the human body at any point in time (see figure 1).

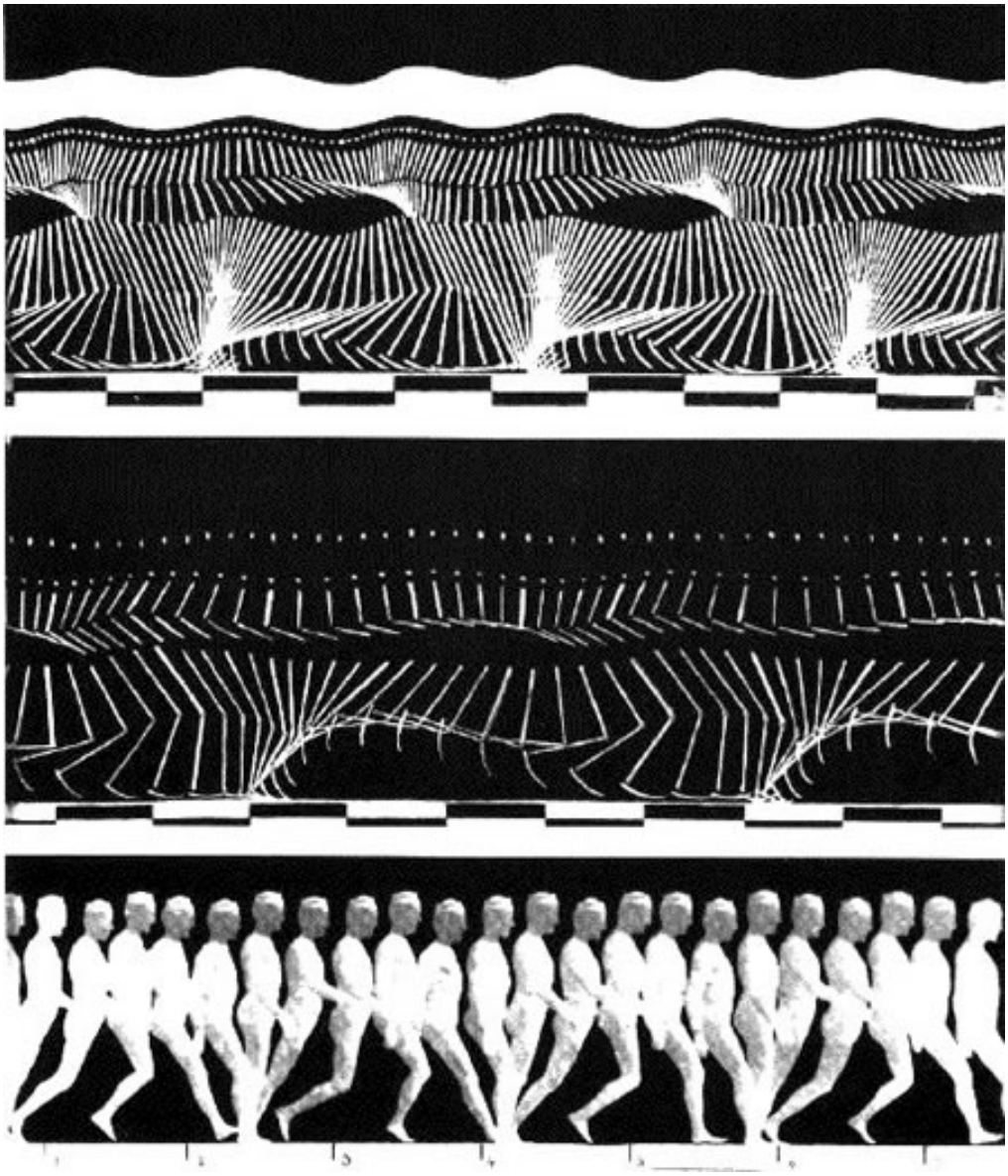

Figure 1: Marey, cronophotograph from *The Human Body in Action*, Scientific American, 1914.

The methods for biological motion studies of nowadays are still quite similar to those proposed by Marey, seeking the extraction of motion patterns through the tracking of specific areas of the human limbs in time.

In 1973, Johansson published the first contemporary experimental data on biological motion perception. In his experiments, Johansson found that his simplified human motion stimulus, from now on referred to as point-light walker (PLW), produced a compelling and vivid impression of a person

walking. He contrasted only the major joints of the human body in motion and obtained 100 percent of correct immediate identifications of the stimulus. But when the stimulus was presented without motion this effect was disrupted, and only a senseless group of white dots was perceived.

From the pioneering work of Johansson emerged an experimental paradigm for the study of biological motion perception. It was found that from the depiction of these overly simplified PLW, subjects are able to access a large amount of information. They recognize actions such as running, jumping, eating and dancing (Dittrich, 1993; Norman, Payton, Long, & Hawkes, 2004), the actor's gender (Troje, 2002; Pollick, Lestov, Ryu, & Cho, 2002; Pollick, Kay, Heim, & Stringer, 2005), their identity (Cutting & Kozlowski, 1977; Troje, Westhoff, & Lavrov, 2005; Richardson & Johnston, 2005; Loula, Prasad, Harber, & Shiffrar, 2005), the subject's own walking pattern (Beardsworth & Buckner, 1981) and even emotions (Dittrich, Troscianko, Lea, & Morgan, 1996; Pollick, Paterson, Bruderlin & Sanford, 2001; Pollick et al., 2002; Atkinson, Dittrich, Gemmel, & Young, 2004).

The promptness to recognize biological motion might be related to underlying neural structures. Some initial studies in this field found that neurons in the superior temporal sulcus (STS) of primates respond selectively to faces and human movements (Perrett, Rolls & Caan, 1982, Perrett et al., 1985). The STS area is a point of convergence of the visual ventral and dorsal streams, with structure and motion processing functions, respectively. It is also connected to the amygdala and the orbitofrontal cortex, regions associated with the processing of stimuli with social and emotional relevance (Puce & Perrett, 2003). In a positron emission tomography (PET) study it was found that the posterior STS (STSp) responds only to congruent PLW (Bonda, Petrides, Oery, & Evans, 1996) and that the activation is stronger for upright than for inverted PLW (Grossman, Blake & Kim, 2004). Static human figures and complex rigid motion patterns produce barely, if any, activation in the STSp area (Beauchamp, Lee, Hoxby, & Martin, 2002; Peuskens, Vanrie, Verfaillie, & Orban, 2005). Cells in the STSp are direction-selective and activate mostly with frontal whole body motion, despite de

existence of some cells that respond to back-oriented motion (Puce & Perrett, 2003). This selectivity might be an underlying reason for the frontal bias found in the perception of biological motion, with will be discussed in greater detail in chapter 4.

Other brain areas respond selectively to the sight of human motion. It is the case of the F5 area in the premotor cortex. The particular aspect of these neurons is that they are equally activated when the primate sees and performs an action (Murata et al., 1997). For this reason, they are called mirror neurons, neurons for the observation, motor learning and preparing for action. These neurons also respond to motor activity-related sounds (Kohler et al, 2002).

The processing of human actions is different from the processing of rigid motion. We are more sensitive to human actions than to other motion patterns (Neri, Morrone, & Burr, 1998; Blake & Shiffrar, 2006). The sensitivity to biological motion is rapidly increased with the number of point-lights, and this increase is much faster than in the detection of rigid motion (Neri et al., 1998). Also, this information is integrated over periods of time much larger than with rigid motion. An accurate motion detection is found even in highly degraded conditions, like with diminished frame rates, altered point-light trajectories, or when biological motion is masked with random dot background noise, which does not occur in the detection of other motion patterns (Blake & Shiffrar, 2006). Even with some degree of scrambled dot motion and at artificially slow motion velocities there is an impression of a human in motion (Beintema, Olesiak, & Wezel, 2006).

Biological motion perception might be an intrinsic capability of the visual system. Infants aged 4 to 6 months exhibit preference for biological motion patterns (Fox & McDanie, 1982) and even 2-day-old newborn babies prefer biological motion displays rather than other non-biologial or inverted point-light displays (Simion, Regolin, & Bulf, 2007). Other species have been found to be sensitive to biological motion as well (e.g. Blake, 1993). These findings suggest that this visual expertise is part of an evolutionary ancient and non

species-specific system predisposing animals to preferentially attend other animals.

Human expertise in perceiving other humans in motion is not, however, limitless. Shiffrar and Pinto (2002) pointed some critical conditions for the processing of biologically defined motion patterns. First, PLW orientation is critical. Rotated and inverted PLW seem to be processed like any other motion pattern. They loose the biological benefit, while keeping the local components of the PLW. Second, time matters. If the temporal characteristics of the human action are not compatible with human motion dynamics, then these stimuli will be interpreted as rigid motion stimuli. Finally, the premotor brain areas are associated with the processing of biological motion, but they only process movements perceived as possible or executable. This indicates that not all perceived actions are equally treated by the visual system.

In sum, biological motion, and namely human motion, constitutes a particular class of stimulus. Despite its complexity it is easily detected and recognized. There are brain structures involved in the processing of biological motion that differ from those involved in the processing of other motion patterns. There is also a body of behavioral data to support the distinction of biological motion perceptual processes.

The study of biological motion perception has, however, remained mostly visual and associated with small computer display experiments. Also, the scope of these studies has been mostly limited to analyzing detection and recognition effects. The particularity of these stimuli is only starting to be duelly addressed in other, more naturalistic, interactive and multimodal perspectives. The scope of this work falls within this last case.

1.1.2. Multisensory Perception

The study of how information provided by the different senses is combined is an expanding and well-established area of interest. From its early days the way such different inputs yield a unified sense of the world has puzzled researchers.

There are several illustrative examples of perceptual effects which reveal that humans use information from different sensory modalities to reach a unified percept. One such broadly acknowledged example is the ventriloquist effect. In the common situation, a ventriloquist talks keeping his lips still while a puppet moves its mouth. This provides a compelling interpretation that it is indeed the puppet talking, as if the sound was actually produced by it. This effect, in which the sound's position is realigned with the visual stimulus' position, has been well demonstrated in multimodal research (e.g. Bermant & Welch, 1976; Bertelson & Radeau, 1981; Pick, Warren, & Hay, 1969). Another striking example of the effect of visual cues over signals form another modality is the rubber hand illusion. Here, subjects observed a dummy hand being touched while they felt simultaneous touches in their own hand, which they couldn't see. In this case, subjects reported feeling the rubber hand as their own (Botvinick & Cohen, 1998). This was the first reported case of an out-of-body experience produced by multimodal cue interactions.

But other data exist in which it is not the visual cue to determine the final sensory experience. In the double flash illusion, a variable number of light flashes were presented while an equally variable number of beeps were heard. Here, participants frequently failed to count the visual flashes, reporting a number that matched the auditory stimulus (Watkins, Shams, Tanakos, Haynes, & Rees, 2006). Another example is an effect frequently addressed as temporal ventriloquism, where a sound presented in close temporal proximity to a visual stimulus can alter the perceived temporal dimensions of the visual stimulus (Gebhard & Mowbray, 1959; Welsh, DuttonHurt, & Warren, 1986; Fendrich & Corballis, 2001; Vroomen & de Gelder, 2004). Also, in a psychomotor task, subjects tapped their finger at the same pace as an auditory beep, ignoring visual flashes, but failed to tap

according to the visual flash when rhythmic beeps were concurrently presented (Repp & Pennel, 2002; Repp, 2003).

In the examples yet described, one sensory modality revealed strongly influential over the other, an effect traditionally addressed as capture or sensory dominance. However, this is not always the case. One of the most puzzling effects reported in multisensory research is the McGurk-MacDonald effect (McGurk & MacDonald, 1976). Here, there was not a determinant influence of either sensory cue, but rather a combination that produced an entirely new percept. Subjects viewed a video depicting a man saying /ga/ and heard a male voice saying /ba/. When both stimuli were presented combined, most subjects reported perceiving a /da/.

It is not clear when and where these multisensory processes do occur in the brain. A variety of brain regions have been pointed out as serving multimodal processing functions, namely the frontal cortex, the intraparietal sulcus, the insula-claustrum and the superior temporal sulcus (see Calvert, 2001, for a review). It is believed that the activity of the superior colliculus (SC) cells is central in multimodal processing (e.g. Meredith & Stein, 1983; Meredith, Nemitz, & Stein, 1987; Meredith & Stein, 1996). These neurons respond to visual, auditory and tactile information. The SC (or optic tectum) is located in the midbrain and contains several layers of alternating white and grey matter, of which the superficial contain topographic maps of the visual field, and deeper layers contain overlapping spatial maps of the visual, auditory and somatosensory modalities (Affifi & Bergman, 2005). These neurons follow 3 rules, which have been associated with the principles of multisensory integration. The spatial rule determines that co-localization of the signals is required for activation – and therefore for integration to occur. The temporal rule states that stimulus co-occurrence in time is equally necessary for SC activity. Finally, the inverse effectiveness rule states that multisensory integration is more likely or stronger when the constituent unisensory stimuli evoke relatively weak responses.

In sum, there is evidence that multisensory processes occur in the brain with several interaction effects. There is not a fixed result when several sensory modalities diverge. Crucially, most often in everyday life the different sensory information about one same event is congruent. In the work reported

here we analyzed both a situation where visual and auditory information, although congruent, led to incompatible percepts (*2.2. The effect of auditory cues on Biased Biological Motion Orientations*), and we approached the integration benefit of congruent audiovisual information (*2.3. The multisensory integration of biological motion*) in biological motion perception.

1.1.3. Multisensory Perception of Biological Motion

In section 1.1.1 we approached biological motion perception from a unimodal perspective. Indeed, most of the current research on biological motion processing is devoted to understanding the human expertise in extracting information from simplified visual motion displays, the PLW. Human motion signals, however, are not exclusively visual.

The first evidence that audiovisual locomotor patterns might be processed in a different way from other crossmodal events came from a brain imaging study, which revealed that auditory footsteps also activated the STSp area (Bidet-Caulet, Voisin, Bertrand, & Fonlupt, 2005). In the STS, 23 percent of the neurons that respond to the sight of an action are significantly modulated by the sound of that action in bimodal conditions (Barraclough, Xiao, Baker, Oram, & Perrett, 2005). In the neurons whose visual response is increased by the addition of sound, the response is dependent upon stimulus congruency.

Some recent motion detection studies brought further support to the assumption that there were specific crossmodal interactions in the processing of biological motion. Sounds moving in the same direction as the PLW increased its detectability, whereas sound moving in an opposite direction inhibited the detection. No facilitatory/inhibitory effects were neither found with stationary sounds nor with inverted PLW (Brooks et al., 2007). With foot tapping point-lights, the detection was facilitated with sounds, but only when in synchrony with the visual stimuli (Arrighi, Marini, & Burr, 2009). Thomas and Shiffrar (2010) also found a facilitatory effect in the detection of biological

motion with step sounds, but not with pure tones. Again, this effect did not occur with inverted PLW.

The crossmodal temporal matching estimates are consistently better with upright than with inverted or scrambled walkers (Saygin, Driver, & de Sa, 2008). Sounds also affect the perceived gender of the PLW (Van der Zwan et al., 2009).

It should be noted that, despite being traditionally used as control stimuli in biological motion research, inverted PLW are not to our understanding uncontroversial in audiovisual experiments, as the causal effects between action and action-related sound are not preserved. The same critical perspective might be used to control stimuli which are stationary, out of phase, or not semantically congruent, as the pure tones. A more robust argument would emerge if jumping balls, cycloid motion or other complex mechanical actions were employed as control.

In sum, all data related to the crossmodal interactions in the processing of biological motion signals is recent. Sound and image appear to be processed in common brain areas. This might account for a possible increased detectability of human motion signals in audiovisual conditions. Some results also suggest that there might be some perceptive interactions, but these interactions still remain largely unexplored, and the data yet available are not beyond debate.

The work reported here will precisely attempt to focus the effect of crossmodal interactions in the perception of human motion signals. For that matter, some previous methodological aspects need to be considered regarding the nature of both the visual and the auditory stimuli.

1.2. METHODOLOGICAL CONSIDERATIONS

1.2.1. Visual Stimuli

In this section we will address the technical nature of the biological motion stimuli, the PLW.

Johansson (1973) used, as stimuli for his experiments, brief films of bright points marking the main joints of the human body in motion. For that purpose, he developed two different methods. In the first method, he attached flashlight bulbs to an assistant dressed in a tight dark suit and filmed her walking along a linear track in a dark room. This methodology revealed some inconveniences, namely the clumsiness of the wire connections and limitations of motion direction changes. The second method developed by Johansson took more advantage of the video systems. He used retroflective patches, rather than lamps over the joints and, in some cases, the patches were used as ribbons around the joints, to allow curvilinear tracks. He then adjusted the video contrast to the maximum, allowing only the retroflective points to show against a totally dark background.

Since Johansson's experiments, there has been little evolution in the PLW as a stimulus, and most of the observed changes were due to the technological advances that occurred since then. For example, Cutting (1978) developed a computer animation system based on artificial synthesis. This system had the benefit of producing PLW without depending on human filming, as the stimuli were simulated from an algorithm. On the other hand, while simplifying the process, this technique also posed some problems, namely the necessity of creating a new algorithm for every new action and some naturalness loss of the artificial synthesis (Dekeyser, Verfaillie, & Vanrie, 2002).

More recently, a variety of motion capture technologies emerged. Some of these technologies use magnetic sensors, and others infrared cameras to track reflective markers. These systems allow the recollection of 3D coordinates of the human motion, which are afterwards computationally animated (Blake & Shiffrar, 2006). These technologies allow a greater freedom in generating stimuli at any orientation or position. They are however still very expensive, demanding in terms of infrastructures, and time consuming.

Several research teams that made the effort to acquire new biological motion stimuli with these technologies chose to produce wide databases, and some are available to the scientific community. It is the case of Vanrie and Verfaillie's database (2004). These stimuli were precisely the ones used in the experiment of chapter 4. As seen in figure 1, they were treated to have shoulders, arms, hips, knees and feet always aligned in height. They also had some smoothing in the walking pattern to have all steps always with the same spatial coordinates and duration.

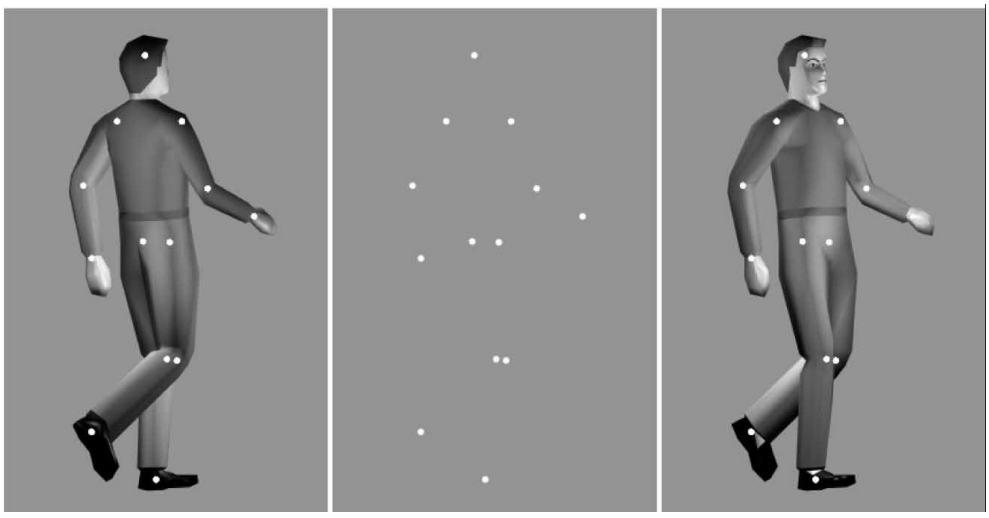

Figure 2: Example of a PLW from the Vanrie and Verfaillie database (2004).

The equalization of the stride to keep it constant was crucial, as it allowed for an easy solution in generating the auditory stimuli, which just had to respect the same stepping frequency of the visual stimulus.

Most recently, in the Laboratory of Visualization and Perception a Vicon® motion capture system was acquired. This system allowed the team to produce their own biological motion stimuli database. There were systematic motion acquisitions with athletes walking at different speeds, controlled by an external velocimeter. Afterwards, with an application developed specially for this effect, the PLW stimuli were created. An example of the output generated from the velocimeter is presented in figure 2.

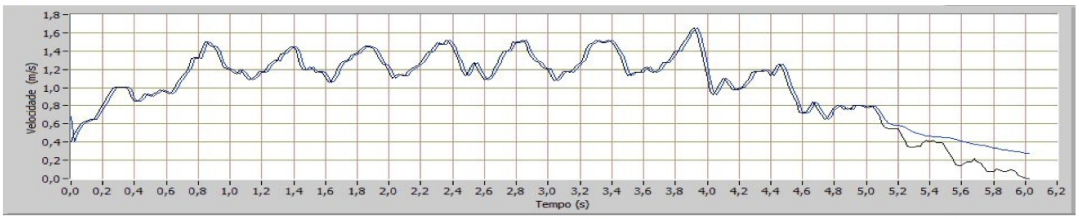

Figure 3: Example of a stride velocity output for a period of 6s.

As is well exemplified in figure 2, the velocity of human stride is not homogeneous. The notches in the graph correspond to the moments when a foot touched the ground. Steps are neither of equal duration nor size. The stimuli used in the experiments reported in chapter 5 were precisely from this database. These stimuli were not equalized nor smoothed, with the benefit of obtaining more natural biological motion patterns. However, to assure that the stimuli in that experiment were the most regular, there was a rigorous selection of motion sections at various velocities. The final stimuli corresponded to a walker taking 3 steps similar in length and duration and velocity was kept fixed. More detailed information is provided in chapter 5.

In sum, biological motion stimuli are complex and demanding, from a methodological point of view. There are many solutions for the generation of these stimuli, but there is not a more widely preferred option, as all have some benefits and some limitations. Stimulus specificity still remains dependent on each laboratory's techniques and methodological options.

1.2.2. Auditory Stimuli

In this section we will focus the nature of our auditory stimuli and the complex methodological and psychophysical issues they posed.

Auditory stimuli are variable in nature. Depending on the degree of technology involved, they can vary from a simple tone to a complex sound that simulates a multitude of aspects in an acoustic environment. Traditionally, the sounds used in audiovisual experiments have been simplified noises, displayed through loudspeakers. They varied either in rhythm or in localization. This last characteristic has traditionally been manipulated only in the azimuth dimension, changing signal intensity between two loudspeakers at the left and right side of the listener.

As was stated in the *Foreword*, when we started considering the development of audiovisual experiments, a wide virtual reality laboratory was being developed. From the start, we were interested in a solution which could be integrated with the visualization system, for the simulation of complex interactive environments. With the cooperation of Prof. Guilherme Campos, Prof. Paulo Dias and Prof. José Vieira from the University of Aveiro, we started implementing and testing auralization algorythms.

Auralization consists in the recreation of a spatial sound. The aim is to accurately simulate acoustic environments and provide vivid and compelling auditory experiences. These virtual sounds are a groundbreaking tool for the study of audition, auditory space, neural plasticity, and the study of audio-motor processes. Research with auralized sounds has also had practical applications like understanding cochlear implant listening adaptations, and audio engineering developments. Other application examples range from flight control systems to tools for helping the visually impaired. It also has a strong potential application in immersive environments and in the entertainment industry.

Acoustic simulation needs to take into account the influence not only of the room itself (wall reflections and attenuation effects) but also of the listener's physical presence in it. In fact, the interaction of sound waves with the listener's body – particularly torso, head, pinnae (outer ears) and ear canals – has extremely important effects in sound perception, notably interaural time and level differences (ITD and ILD, respectively), the main cues for source localization (Colburn, 1973; Stern & Colburn, 1978; Stern & Corlburn, 1985). Such effects can be mathematically described by the binaural impulse response for the corresponding source position, known as *Head Related Impulse Response (HRIR)*, or, more commonly, by its Fourier transform, the *Head Related Transfer Function* (HRTF).

The common process for obtaining HRTFs is trough the systematic recording of the impulse responses of a sound inside the human ear. Most frequently, there are head-and-torso manikins having anatomical features similar to a human with rubber ears, placed in a room where sound is presented at a dense array of positions in space. The rubber ears are shaped according to an average model, thus providing a measurement that corresponds to a general anatomical configuration, but not to a specific person (e.g. Algazi, Duda, & Thompson, 2001). These are called non-individualized HRTFs. It is possible to generate individualized HRTFs by creating ear models from the listener and performing new acquisitions for each new user.

The HRTF filters are used in the auralization of sounds by applying them to the auditory stimuli with the desired virtual source position. It is believed that through this technology it is possible to accurately locate a sound source in space and provide an immersive virtual experience. Despite the great technological breakthroughs in this field, little pychophysical validation had been carried out when we first started implementing it.

Several questions remained about the perceptual efficiency of auralized sounds. There were several concerns that these sounds might not be accurately externalized, seeming to come from inside the listener's head rather than from an external position. Also, no reliable empirical information

existed about how non-individualized sounds were heard, if listeners could perceive auditory space from them or if there was any adaptation process. If limitations were found at this level, then the auditory cues in our audiovisual experiments would fail to provide the desired information. As such, a number of experiments were developed to validate these algorithms and to propose an experimental protocol for the use of non-individualized HRTF sounds. These experiments will be described in detail in section 2.1.

1.2.3. Combining Visual and Auditory Cues

A last methodological consideration regards the implementation of the combined visual and auditory stimuli. There is vast evidence that for two different modality signals to be perceived as one event (the unity assumption) all stimuli must be co-localized and temporally aligned (e.g. Meredith et al., 1987; Meredith & Stein, 1996; Arrighi et al., 2009). While the space dimension was assured through visual calibrations and audio localization experiments, the audio-visual matching in time required yet another technical development.

To assure that our system actually displayed the visual and audio signals at the same time, we had our software developers index the sound triggers at the moment that the first frame was displayed by the computer cluster. More importantly, we implemented a function that allowed us to manipulate the audio-visual delays at the level of the millisecond, a time interval which is thought to be below the differential threshold in the perception of these two signals combined (for a reviw see Van Eijk, Kohlrausch, Juola, & Van de Par, 2008).

To accurately control the temporal alignment of the sound onset and the displayed frames for each stimulus, an additional care was taken. An external clock was implemented to assess the timing differences between the image display and the earphone output. This custom-built latency analyzer

(Arm7 microprocessor coupled with light and sound sensors) allowed further extensive tests with all stimuli, where average delays and variations were determined. This allowed us to fine tune the timing of the audio signals.

Naturally, assuring physical synchronicity does not provide evidence that indeed stimuli were perceived as being displayed simultaneously. On this matter, a small experiment was conducted, further detailed in section 2.3.

1.3. MOTIVATION AND EXPERIMENTAL GOALS

This thesis comprehends 3 experimental sections, corresponding to 3 major research objectives. The global scope of the work here described was the study of the multisensory processes in the perception of biological motion. The specific goal was to understand if congruent auditory sounds would decrease discrimination errors and bias (experiment *The effect of auditory cues on Biased Biological Motion Orientations,* section 2.2), and would reduce uncertainty (experiment *The multisensory integration of biological motion*, section 2.3). An additional research objective emerged from the implementation of auralized sounds, as described in topic 1.2.2. The auditory experiments (section 2.1, *Locating auditory sources with non-individualized HRTF-based auralizations*) were intended to validate the spatial propertied of the sound stimuli and to develop a protocol for the preparation of experimental participants. The auditory work was indeed only conducted after conclusion of the section 2.2 experiments, which didn't use the auralization algorithms above described. However, having in mind a more natural sequence of research themes, they are presented first. Therefore, we have, by presentation order, the following research objectives:

1. *Locating auditory sources with non-individualized HRTF-based auralizations*

Here, we intended to understand if the non-individualized auralized sounds were accurately perceived in space. Not only were we interested in the localization accuracy, we also had the objective of determining if there was an adaptation process. As the stimuli were generated with algorithms corresponding to an average pinnae, different from the listeners' own anatomy, we expected that some adaptation or learning took place. Due to the novelty of the work reported in this section, eventually these experiments became a separate area of interest, with both psychophysical relevance and applied value by their own.

There were 3 experiments. In the first experiment the goal was to assess if listeners adapted spontaneously to the auralized stimuli while testing azimuth localization accuracy.

In the second experiment we intended to test if the adaptation process could be favored by a training period. We expected to reduce localization error in azimuth through an active learning process. In the third experiment the same training was tested in an elevation discrimination task. From these experiments we expected to determine a quick procedure that would allow us to prepare subjects for the use of the non-individualized auralized sounds.

2. *The effect of auditory cues on Biased Biological Motion Orientations*

When we fist started studying the multisensory processes in biological motion perception little was know about auditory-visual interactions with this specific type of stimuli. Our fist approach was to design a simple experiment where we tested if sounds could influence the perceived motion characteristics of a walker. As it is assumed that multisensory processes might reduce errors or biases by the presence of redundant information, we presented ambiguous and biased displays of PLW and matched them with congruent unbiased auditory cues.

It was expected that, with additional auditory congruent information, biased misjudgments would drop and performance would be enhanced. Results were analyzed in terms of the bias levels in the visual-only, auditory-only and in the audiovisual conditions.

3. *The multisensory integration of biological motion*

The final experiments intended to determine the benefit of multisensory integration in biological motion discrimination.

First, and having in mind that perceived simultaneity is required for signal integration (e.g. Arrighi, Marine, & Burr, 2009), a preliminary experiment was conducted. We measured simultaneity judgments at several audiovisual walking asynchronies and obtained a point of subjective simultaneity.

The main experiment consisted of a velocity discrimination task. We hypothesized that congruent audiovisual stimuli would lower the discrimination thresholds. The discrimination accuracy was compared to optimal predictions and analyzed at different velocity congruency levels.

II. EXPERIMENTS

2.1. LOCATING AUDITORY SOURCES WITH NON-INDIVIDUALIZED HRTF-BASED AURALIZATIONS

2.1.1. Introduction

In this section we will address the auditory accuracy, adaptation and learning processes involved in the use of auralized sounds.

As was described in section 2.2.2, the auralization process needs to take into account the interaction between sound waves and the listeners' body. The shaping provided by someone's anatomy over the audio signals contains the key elements for good auditory source localization (Plenge, 1974; Wightman, Kistler, & Perkins, 1978; Blauert, 1983; Searle, Braida, Cuddy, & Pavis, 1976). The filters used in audio engineering for the simulation of this shaping are commonly known as HRTFs (Head Related Transfer Functions).

Since they depend on anatomic features such as the size and shape of head and ears, HRTFs vary considerably from person to person. Moreover, even for the same person they will vary with age and reveal no symmetry between left and right ear responses. Given this variability, spatial audio simulations should use individualized HRTFs (Wenzel, Arruda, Kistler, & Wightman, 1993). However, these would be extremely difficult to obtain in practice for wide or representaive samples of listeners; HRTF recordings are effortful, expensive and very time consuming, requiring anechoic rooms, arrays of speakers (or accurate speaker positioning systems), miniature microphones, and specialized software and technicians. Due to these practical difficulties, most systems make use of generic (non-individualized) HRTFs, measured on manikins or head-and-torso systems equipped with artificial pinnae designed to approximate as best as possible an 'average' human subject.

It has been suggested that satisfactory auralization can be obtained using generic HRTFs (Loonis, Klatsky, & Golledge, 1999). Wenzel and colleagues (1993) compared the localization accuracy when listening to external free-field acoustic sources and to virtual sounds filtered by non-individualized HRTFs. Several front-back and up-down confusions were found, but there was overall similarity between the results obtained in the two test situations. A similar result was found in the auralization of speech signals (Valjamae, Larson, Vastfjal, & Kleiner, 2004). Most listeners can obtain useful azimuth information from speech filtered with non-individualized HRTFs.

On the other hand, there are indications that individualized HRTF-based systems do differ from generic ones. There is a significant increase in the feeling of presence when virtual sounds are processed with individualized binaural filters instead of generic HRTFs. Differences in the intensity of the auditory virtual experience are also reported (Begault & Wenzel, 1993). Interestingly, some authors have suggested that the perception of spatial sounds with non-individualized HRTFs might change over time. Begault and Wenzel (1993) observed several individual differences, which suggested that some listeners are able to adapt more easily to the spectral cues of the non-individualized HRTFs than others. Asano, Suzuki and Stone (1990) also claimed that reversal errors (such as the front-back and up-down confusions) decrease as subjects adapt to the unfamiliar cues in static anechoic stimuli.

In this context, a research question emerged: can humans learn to accurately localize sound sources processed with HRTF sets different from their own? There is evidence that the mature auditory brain is not immutable, but instead holds the capacity for reorganization as a consequence of sensory pattern changes or behavioral training (Gilbert, 1998). Shinn-Cunningham, Durlach and Held (1998) trained listeners with “supernormal” cues, which resulted from the spectral intensification of the peak frequencies. With repeated testing, during a single session, subjects adapted to the altered relationship between auditory cues and spatial position. Hofman (1998) addressed the consequences of manipulating spectral cues over large periods of time, adapting moulds to the outer ears of the subjects. Elevation cues, which

depend mostly on monoaural cues (for a review see Middlebrooks & Green, 1991) were initially disrupted, but azimuth discrimination was preserved. These elevation errors were greatly reduced after several weeks, suggesting that subjects learned to associate the new patterns with positions in space.

The broad intention of this study was to assess the baseline accuracy and the learning processes in the use of non-individualized HRFTs. It was expected that such study would bring new insights on the human perception of auditory space, with general psychophysical interest as well as applied applied potentials, both to our work and to other professionals in the audio field.

Specifically, we intended to validate our auralization algorithms and assure that users of such generically spatialized sounds become able to fully extract the relevant information from their listening experiences, and in as short time as possible. The experiments were intended to understand under which conditions subjects will be readily prepared, namely by tackling the questions: *Do listeners adapt spontaneously without feedback?* (2.1.2); and *Can we improve the adaptation process?* (2.1.3 and 2.1.4).

2.1.2. Azimuth Accuracy Without Feedback

This experiment intended to assess the localization accuracy of inexperienced subjects as they became gradually more familiarized with the non-individualized HRTF processed sounds. We tested their ability to discriminate sounds at fixed elevation and variable azimuth in 10 consecutive experimental sessions (blocks), without feedback. We analyzed the evolution of the subjects' performance throughout each block.

2.1.2.1. Method

i) Participants

Four naïve and inexperienced young adults participated in the experiment. They all had normal hearing, verified by standard audiometric screening at 500, 750, 1000, 1500 and 2000 Hz. All auditory thresholds were below 10 dB SPL and none had significant interaural sensitivity differences.

ii) Stimuli and Apparatus

The stimuli consisted of pink noise sounds.

The sounds were auralized (see section 1.2.2) at 8 different azimuths: 0º (front), 180º (back), 90º (left and right), (45º left and right), and 135º (left and right). They had constant elevation (0º) and distance (1m). For this purpose, the original (anechoic) sound was convolved with the HRTF pair corresponding to the desired source position. The resulting pair of signals – for the left and the right ear – was then reproduced through earphones.

The HRTFs set were recorded using a KEMAR dummy head microphone at the Massachusetts Institute of Technology (Gardner & Martin, 2010). Sounds were reproduced with a Realtec Intel 8280 IBA sound card, and presented through a set of Etymotics ER-4B MicroPro in-ear earphones.

iii) Procedure

All sounds were presented pseudo-randomly for 3 seconds, with 1 second interstimulus interval. There were 10 blocks of 10 stimulus repetitions each. Participants were told to indicate the perceived sound source location for each stimulus.

The answers were recorded by selecting, on a touch screen, one of the eight possible stimulus positions.

2.1.2.2. Results

The average correct responses in azimuth localization was 65%, well above chance as random responses would have resulted in averages of 12.5% The left and right 90º sounds were the most accurately located, with a correct response rate of 78%. Similarly to what had been found in previous studies (Wenzel et al., 1993), there were several front-back confusions that account for the lower accuracy at 0º (62% correct answers), 180º (43%), left/right 45º (60%) and left/right 135º (69%).

Analyzing the average participant's performance along time (Figure 4), we see that the overall accuracy remained constant. There were individual differences between participants. Listener 1 was less accurate (50.4% correct answers), listeners 2 and 3 performed near average (61.9% and 71.1%, respectively) and listener 4 had the best azimuth localization performance (85.1%). However, none of the participants revealed a tendency to improve their performance.

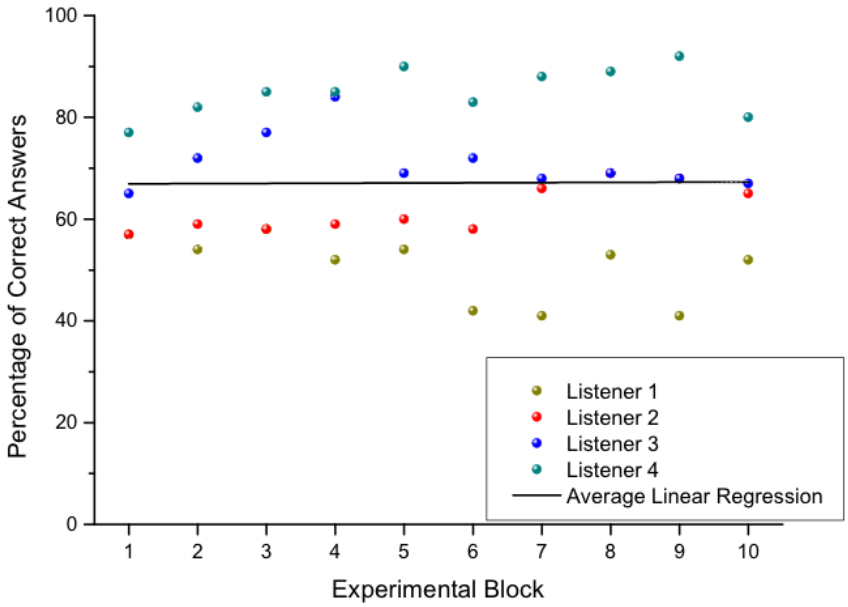

Figure 4. Percentage of correct answers by experimental block and linear regression.

The linear regression results revealed a slope coefficient close to zero (0.04), meaning almost no change in the percentage of correct responses. The correlation values confirmed that the experimental block number does not account for the listeners' accuracy (r^2=0.00), and the relation between them is insignificant (p=0.958).

Our results reveal that naïve participants are able to discriminate sounds at several azimuths well above chance. This high accuracy was, however, below our expectations. Previous studies on sound localization with real external sound sources (Middlebrooks & Green, 1991) had revealed that the azimuth stimuli separated by 45º, as ours were, should not be confused, as localization errors were always below 20º. We therefore would expect good discrimination of all sounds and some ceiling performances. We hypothesized that ceiling effects were not observed due to some front-back confusions. However, these confusions could not explain the low results in the lateral stimuli. Therefore, we also hypothesized that no ceiling effects were observed due to the low adaptation to the non-individualized HRTFs.

On the other hand, throughout the exposure blocks, localization accuracy did not increase, leading to the conclusion that simple exposure is not enough for significant localization improvement in short periods of time.

In view of these conclusions, a second experiment was developed where, for the same amount of time, listeners were trained to discriminate sound source locations.

2.1.3. Learning Azimuths

In this experiment we tested the participants' accuracy in localizing sounds at several azimuths before and after a short training program. For this program we selected only a small number of sounds and trained them through active learning and response feedback.

2.1.3.1. Method

i) Participants

Four young adults participated. None of them had any previous experience with virtual sounds. They all had normal hearing, tested with a standard audiometric screening, as described in 3.1.2.

ii) Stimuli and Apparatus

As in the first experiment, all stimuli consisted of pink noise sounds, auralized with the same algorithms and software.

All stimuli varied in azimuth, with elevation (0º) and distance (1 m) fixed. Azimuths ranged from the front of the subjects head to their right ear, spaced at 6º intervals (from 6º left to 96 º right). Only these azimuths were used, aiming to assure that other effects such as front-back biases and individual lateral accuracy asymmetries did not emerge, as they were not within the interest of our study. All sounds had a 3 second duration, with an interval of 1 second between each stimulus.

iii) Procedure

The experiment started with a pre-test. In the pre-test, all sounds were presented pseudo-randomly with 4 repetitions each. Participants had to indicate, on a continuum displayed on a touch screen (figure 5A, blue area), the point in space where they estimated the sound source to be.

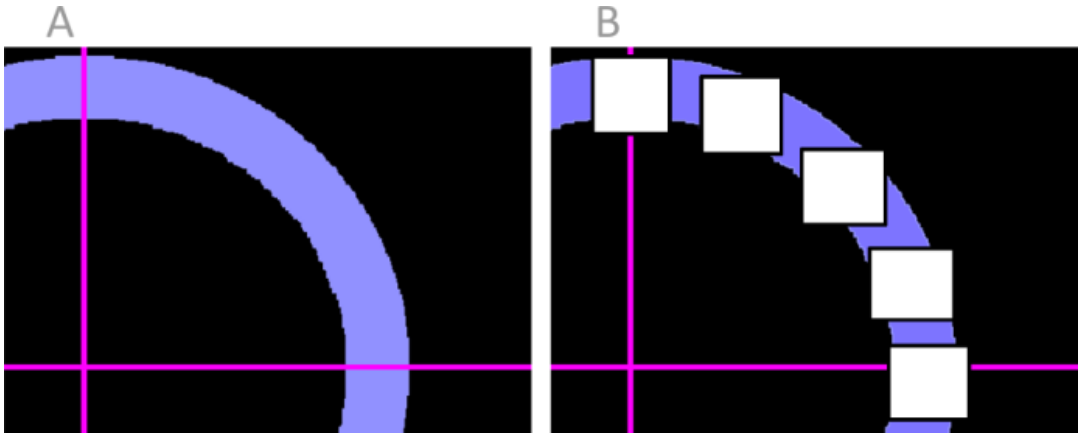

Figure 5. Touch screen in the pre-test and post-test (A). Touch screen in the training program (B).

After the pre-test, participants engaged in a training period. The trained sounds corresponded to the frontal (0º), lateral (90º) and two intermediate azimuths (21º and 45º) (see white areas in figure 5B).

The training conformed to the following steps:

- Active Learning: Participants were presented with a sound player where they could hear the training sounds at their will. To select the sounds, there were several buttons on the screen, arranged according to the spatial the corresponding source position. The participants were informed that they had 5 minutes to practice and that afterwards they would be tested.

- Passive Feedback: After the 5 minutes of active learning, participants heard the training sounds and had to point their location on a touch screen (figure 5B). After each trial, they were told the correct answer. The passive feedback period continued until participants could answer correctly in 80 percent of the trials (5 consecutive repetitions of all stimuli with at least 20 correct answers).

When the training period ended, participants performed a post-test, an experiment equal to the pre-test for comparison purposes.

2.1.3.2. Results

i) Pre-Test

Results from the pre-test and post-test sessions are displayed in Figure 6. Orange and purple bars display the average response distance (in degrees) to the given stimulus position. This gives us an estimate of the size of the error in locating the position of the sound source. Gray bars display the mean theoretical error (in degrees) that would be obtained if participants responded randomly.

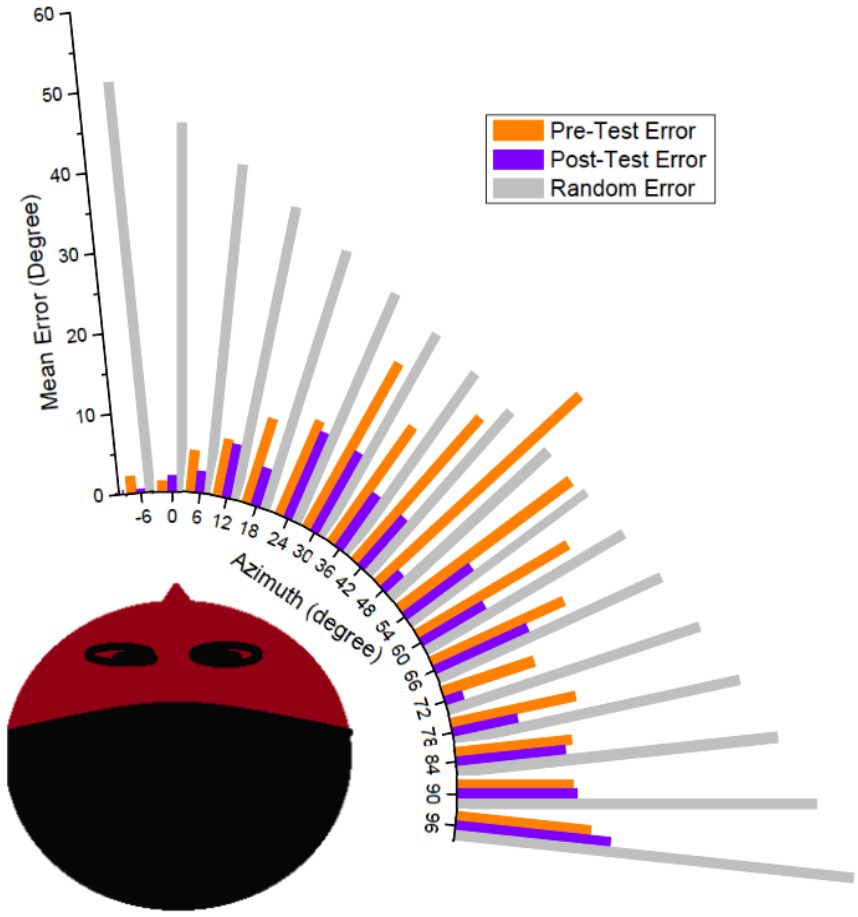

Figure 6. Average response error (degree) in the Pre-Test and Post-Test sessions, and theoretical error level if listeners responded randomly.

Analyzing the pre-test results (Figure 6, orange bars), we observe that azimuth discrimination is easier for frontal stimuli: the average error is below 5

degrees. The absence of rear stimuli, which prevented any front-back confusions, may help explain these results. As in the previous experiment, listeners were fairly precise in identifying lateral source positions. Sounds were most difficult to locate at intermediate azimuths (between 40º and 60º). For these positions, pre-test localization accuracy was close to random (from 22º to 36º of average errors), revealing an overall inability of the subjects to discriminate such sound positions.

On average, in the pre-test, participants had a localization error of 15.67º.

ii) Training Period

The training sessions were very successful for all participants. All took less than 30 minutes and, in average, they lasted 22 minutes.

Learning curves are displayed in figure 7, where individual azimuth discrimination accuracy is plotted as a function of the time elapsed since the start of the training period.

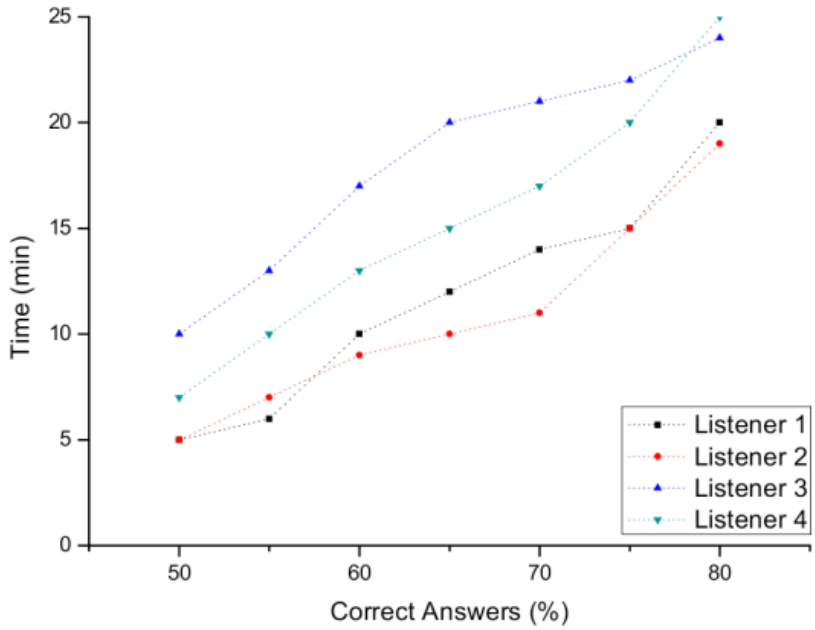

Figure 7. Individual accuracy evolutions in the azimuth localization training sessions.

All participants reached the 80% criterion. Despite the differences in learning velocity, a smooth progression was observed for all of them.

iii) Post-Test

The post-test results (Figure 3, purple bars) revealed a large error reduction of 7.23º on average, from 15.67º in the pre-test to 8.44º in the post-test. Despite individual differences, all participants revealed the similar learning effects. This difference was statistically significant in a paired samples T-test ($t_{(3)}$=7.23, p≤0.005). The error reduction was most expressive in the intermediate azimuths, where the average error decreased 20 degrees. Analyzing the trained azimuths (0º, 21º, 45º, 66º, 90º), we observe that performance enhancement was substantial not only for these stimuli, but also for others, not trained. As an example, the best error reduction was obtained with the 48º azimuth, a non-trained stimulus. In contrast, the 90º azimuth, a trained one, revealed similar results in both sessions. These findings allow us to conclude that the trained discrimination abilities for some stimuli positions are generalized to other, non-trained, auditory positions.

2.1.4. Learning Elevations

Here, an elevation discrimination task was carried out using the same methodology as in the previous experiment. Elevation is known to be perceived less accurately than azimuth or distance, probably because it depends mostly upon monoaural information (Middlebrooks & Green, 1991). This experiment was designed to investigate whether or not the learning effect found with azimuths could be attributed to an improved interpretation of the binaural information contained in the HRTFs.

2.1.4.1. Method

i) Participants

Four inexperienced subjects took part in the experiment after undergoing auditory testing with the same standard screening as previously described.

ii) Stimuli and Apparatus

As in the previous experiments, all stimuli consisted of pink noise sounds, auralized with the same algorithms and software. Here, the stimuli varied in elevation, but not in azimuth (0º) or distance (1m). They ranged from the front of the listeners' head (0º in elevation) to the top (90º in elevation) in 10º intervals. Stimuli did not go beyond 90º, as the HRTF database was limited to these elevations. Participants were aware that no back stimuli were present, but no instruction was given regarding stimuli below 0º.

All sounds had 3-second duration, with 1 second interstimulus intervals.

iii) Procedure

This experiment followed the same procedure as the previous one.

In the training period, the sounds were positioned at elevations of 0º, 50º and 90º. Figure 8 shows the touch screen used in the pre-test and post-test

sessions (A), as well as the touch screen with the 3 defined elevations, which were trained (B).

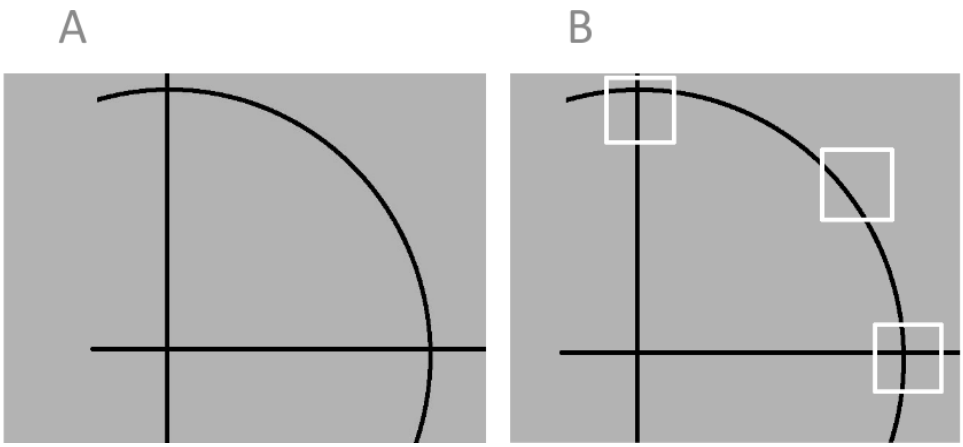

Figure 8. Touch screen in the pre-test and post-test (A). Touch screen in the training period (B).

2.1.3.2. Results

i) Pre-Test

Figure 9 presents the average error (in degrees) between the subjects' answers and the stimulus elevations in the pre and post-test sessions. It also shows the theoretical errors that would be obtained if subjects responded at chance.

In the pre-test session, the average error was 40.8º. The subjects were unable to localize all sounds ranging from 0º to 50º; the worst results were in the frontal (0º) stimuli (55º average error). Overall, participants were less accurate in estimating a sound position in elevation than in azimuth.

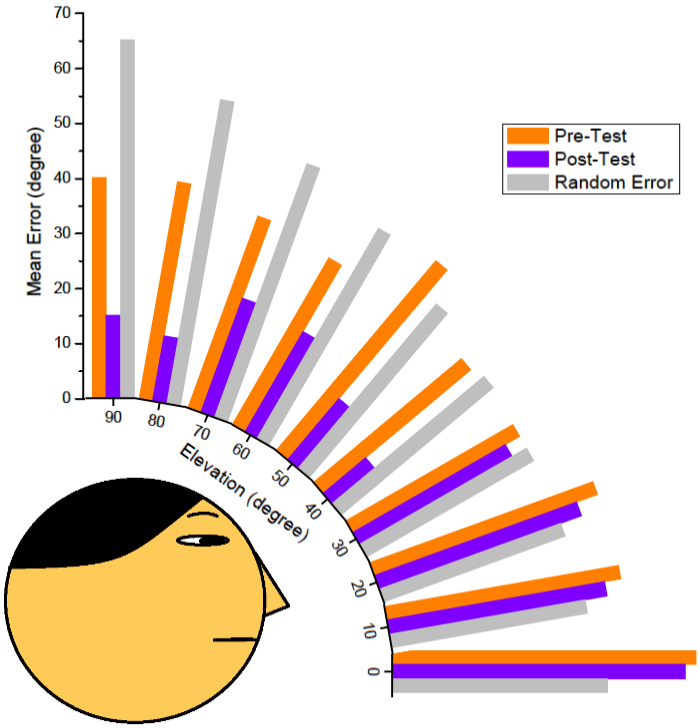

Figure 9. Average response error in the Pre-Test and Post-Test sessions, and theoretical response errors if listeners responded randomly.

ii) Training Period

Training sessions were faster than those of experiment 1, as there were only 3 trained elevations. On average, they took 17 minutes (Figure 10).

Only one subject (listener 3) did not evolve as expected. After 10 minutes testing, this subject was still making excessive mistakes, and was allowed a second active learning phase (5 minutes), after which the 80 percent accuracy was rapidly achieved.

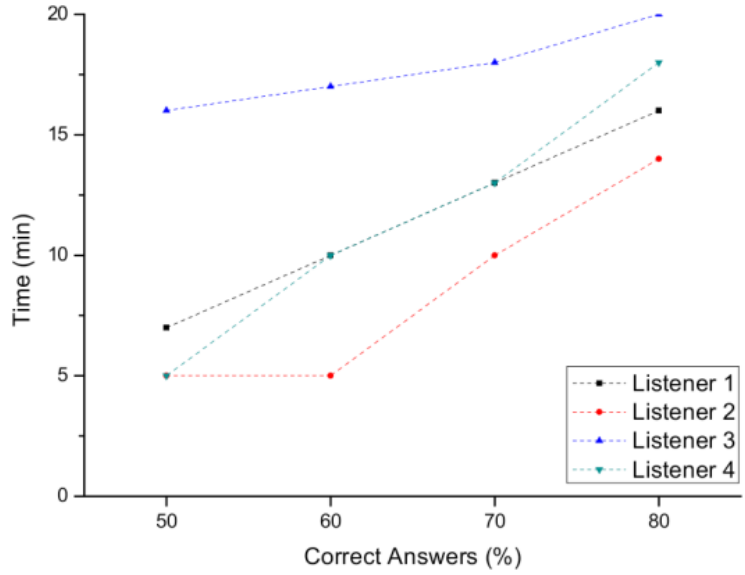

Figure 10. Individual accuracy evolutions in the elevation training sessions.

iii) Post-Test

The post-test results were better than those of the pre-test for all subjects. This difference was significant in a paired samples T-test ($t_{(3)}$=14.23, p≤0.001) The average error decreased 14.75 degrees, to a mean of 26.5º. Like in the previous experiment, after training errors were reduced in half. The training effect was most expressive for the upper stimuli, namely at 80º, 40º and 50º elevations. Among these stimuli, the only trained one was at 50º. On the other hand, sounds at 0º elevation, a trained stimulus, revealed the smallest accuracy increase in the post-test session.

Similarly to what was found in the previous experiment, training was highly effective and generalized well to other, untrained, stimuli.

2.1.5. Discussion

Auralization is of great interest to practitioners in psychophysics, the audio engineering, audio industry, in immersive environments, entertainment, as well as in a variety of other applications. Research efforts in this field have led to sophisticated simulations, which include the effect of the individual anatomical shaping upon the sound waves that reach the listeners' ears. But, as such shaping varies considerably among different people, the perceptual caveats of using non-individualized approaches must be investigated. In this section, we were specifically interested in better understanding the evolution in perceptual accuracy as a subject familiarizes with non-individualized HRTFs. We intended to understand if listeners adapt spontaneously without feedback in a reasonably short time and/or if we could somehow accelerate the adaptation process.

In the first experiment - *Azimuth Accuracy Without Feedback*, we addressed the listeners' adaptation process to static non-individualized azimuth sounds without feedback. Throughout 10 short experimental consecutive sessions, we measured the percentage of correct answers in position discrimination. Results revealed an azimuth discriminability well above chance, but below expectations. Throughout time, there was an overall absence of performance improvement in all subjects. We concluded that simple exposure is not enough for significant accuracy evolution to be achieved in short periods of time. Such exposure learning had been claimed in previous works (Valjmae et al., 2004; Asano et al., 1990), in an attempt to explain individual differences in accuracy results. Our results did not reveal those effects. Adaptation without training was, however, demonstrated before (Hofman et al., 1998), but over wide periods of time (weeks) and with spatial feedback, as participants of those experiments carried the moulds inside their ears in their daily lives during the whole period.

Pursuing the intent of preparing untrained listeners to take full advantage of non-individualized HRTFs, we designed a second experiment – *Learning*

Azimuths, where subjects were trained in sample sounds with a short program combining active learning and feedback. In a pre-test, participants revealed good discrimination abilities for frontal stimuli, but performed very poorly in the intermediate (40º to 60º) azimuths. After the training sessions, in a post-test, all azimuths were identified above chance, with results significantly better than the pre-test ones. More importantly, the training benefit was observed not only in the trained sample azimuths, but was generalized to other stimulus positions. One might argue that an overall learning of the new HRTF-based cues took place, and was then applied to the other untrained stimuli.

We could speculate that the learning effect found in the second experiment might be explained by a fast recalibration to new ITD and ILD values (which could be found in other non-auralized audio systems, like stereo or surround), rather than an adaptation to the new binaural and spectral cues altogether. In a final experiment – *Learning Elevations*, we tested the same training program, with stimuli varying in elevation and with fixed azimuth. Elevation alone is known to be poorly discriminated, when compared to azimuth, mostly because it depends upon monoaural cues, such as the spectral shaping of the pinnae and inner ear. Results in the pre-test of this experiment revealed poor source discrimination ability at almost all elevations, particularly from 0º to 50º. Indeed, with unfamiliar HRTF filters, auralized sounds carried little elevation information for the untrained subjects. A large difference was found in the post-test, where some discriminability arose. Again, the performance benefit was generalized across stimuli and was not restricted to the trained elevations. This finding further supports the assumption that indeed the new HRTF-shaped frequencies were learned.

Both in experiments *Learning Azimuths* and *Learning Elevations*, the learning sessions had the approximate duration of 20 minutes. Longer sessions might have led to better performance improvements. We stress, however, that in preparing listeners for auralized interfaces time should not be the criterion. In our sessions, each participant revealed a different profile and learned at a different velocity. Fixing a goal (such as 80% accuracy) will allow a way of assuring all listeners reach an acceptable adaptation.

We conclude that in binaural auralization using generic HRTF it is possible to significantly improve the auditory performance of an untrained listener in a short period of time. However, natural adaptation to static stimuli is unlikely to occur in a timely manner. Without any training, several source positions are poorly discriminated. In view of this, we argue that virtual sounds processed through non-individualised HRTFs should only be used after learning sessions. We propose that these sessions might involve a small sound sampling, active learning and feedback.

Future studies in this field should focus on the endurance of the learned capabilities over time, generalization limits, and the training effects over the final auditory virtual experience.

2.2. The effect of auditory cues on Biased Biological Motion Orientations

2.2.1. Introduction

Accurately perceiving humans in motion is an important ability. In most daily activities, from driving to walking on a sidewalk, playing sports or watching a movie, we rely on this ability to perceive the actions others are performing from their motion patterns. As was addressed in section 1.2.1, there is vast consensus that we are indeed experts in perceiving biological motion. Nonetheless, this expertise is not without limits.

Like in many other perceptual tasks, under impoverished conditions, biological motion stimuli might not provide enough information for an accurate representation to be formed. One such example is the frontal bias in visual biological motion discrimination. When lighting conditions are low or excessive, artificial or highly contrasted (as when the human body is backlit), direction, form, and texture cues, such as facial and clothing features are frequently unavailable, and the human in motion becomes bistable, allowing both a frontal and a back-oriented interpretation (see figure 11b for an illustration; see also figure 2). In such situations, a face forward bias emerges, and the human in motion is more frequently perceived facing the viewer, rather than backwards. If all the perspective, size and distance cues were removed or kept constant, the frontal interpretation would be preferred in about eighty percent of the instances (Vanrie, Dekeyser, & Verfaillie, 2004).

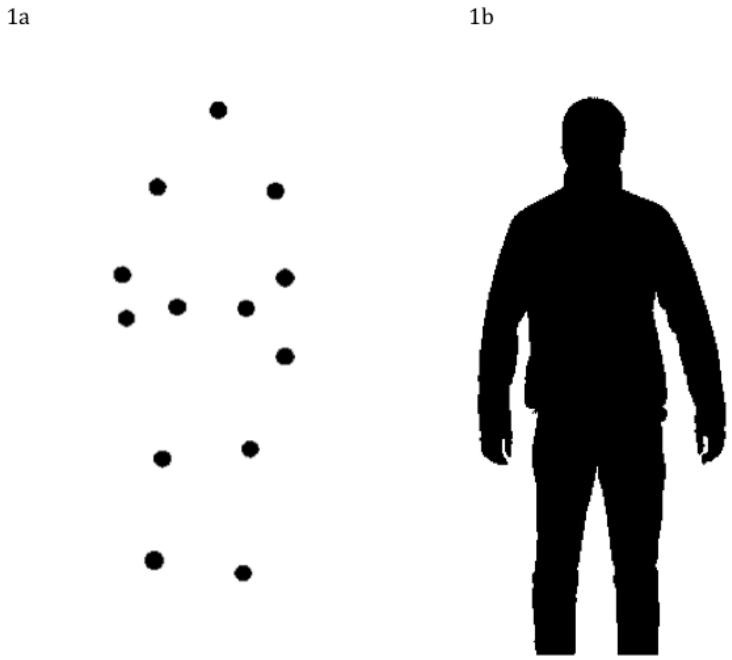

Figure 11. Ambiguous human body displays. Left figure (1a) presents the biased visual stimulus; Right figure (1b) shows a backlit man who might be perceived in a frontal or back orientation.

Frontal biases are not exclusive to human motion. They arise in the processing of looming sounds (for a review, see Neuthoff, 2001) and looming visual concentric gratings (Parker & Alais, 2007). An evolutionary perspective might account for these biases. One such explanation effort is provided by the error management theory (Haselton & Nettle, 2006). This theory predicts that, under uncertainty, natural selection favors the least costly error, which with biological motion would be assuming that the perceived human could collide or interact with the perceiver. Although adaptive in general, this bias might however lead to misjudgments, errors and inadequate actions in several practical situations.

The scope of the present work is precisely to approach this visual bias and to understand if, given additional congruent environmental information,

namely step sounds, this bias could be diminished, thereby providing more accurate estimations of human motion direction.

Research on multimodal perception has shown that combining cues from different senses improves perceptual performance (i.e., lowers detection or discrimination thresholds and variance) and is also likely to reduce reaction times (Driver & Spence, 1994; Driver & Spence, 2004; Vroomen & de Gelder, 2000). Only a few multimodal studies have focused on moving stimuli. Such studies revealed some specific effects.

When both visual and auditory motion signals are present, detection performance improves (Alais & Burr, 2004a; Wuerger, Hofbauer, & Meyer, 2003). Also, vision has an impact on the auditory motion aftereffects, but sound does not influence the visual aftereffects (Kitagawa & Ichihara, 2002; Vroomen & de Gelder, 2003). Concerning the motion direction estimates, visual cues change the perceived direction of the auditory signal, but the opposite effect does not occur (Kitajima & Yamashita, 1999; Matteef, Hohnsbein, & Noak, 1985; Sanabria, Soto-Faraco, & Spence, 2007; Soto-Faraco, Lyons, Gazzaniga, Spence, & Kingstone, 2002; Soto-Faraco, Spence, & Kingstone, 2004).

There are, however, some exceptional situations in which sound affects the perceived visual motion pattern, as when combining suprathreshold auditory motion with subthreshold visual motion (Meyer & Wuerger, 2001), and in the bouncing ball effect (Sekuler, Sekuler, & Lau, 1997). The bouncing effect if found when two objects move toward each other. In the colliding moment, two possible visual interpretations arise: either the objects move past each other (streaming) or they collide and bounce back (bouncing). Such bistable motion pattern is most often seen as streaming, however, when added a sound in the colliding moment, the perceived motion will bias toward bouncing. The disambiguating effect of the auditory sound persists even when the visual motion sequence is less ambivalent and more consistent with streaming (Groove and Sakurai, 2009).

In sum, although vision is predominantly the most influential cue in estimating motion direction, audition may, under bistable circumstances, be determinant. Whether both these effects predict how we perceive the direction of human motion is however yet to be established.

In this section we intended to test if the face forward bias of the walkers could be reduced by congruent disambiguating auditory cues. It was expected that, with additional auditory congruent information, the biased misjudgments would drop and the orientation discrimination accuracy would be enhanced.

To test if auditory steps sorted any effect over the visual face forward illusion, we used a visual representation of a human in motion, a point-light figure (see figure 1a), and coupled it with footstep sounds. Both the visual and the auditory stimuli could be moving towards or away from the subjects. The task was to point the direction of the stimuli across visual, auditory, and audiovisual experimental conditions.

2.2.2. Method

2.2.2.1. Participants

Ten participants performed the experimental task. They were all naïve with respect to the experimental purposes and had never seen the point-light stimuli before. They were also unaware of the biasing effects of the stimuli. All subjects' visual and hearing abilities were tested before the experiments. All had normal or corrected-to-normal vision; their listening thresholds were below 5 dB SPL for 500 Hz, 750Hz, 1000 Hz and 1500 Hz frequencies for both ears.

2.2.2.2. Stimuli and Apparatus

The visual biological motion stimuli were point-light walkers made of thirteen dots. The white dots subtended 0.1 deg of visual angle at the starting

point (luminance of 79 cd/m^2) and moved against a black background (0.07 cd/m^2) at 30 frames per second. From head to toe, the point-light walkers subtended 4.8 deg of visual angle for observers seated at 3 m from the screen. The visual stimuli were projected on a 2.78 m x 2.09 m screen. The walkers had the common translational component removed, as if walking on a treadmill, and completed one step-cycle per second. They were animated in 3D coordinates, simulating the veridical behavior of the point-light spheres in perspective for the defined visual angle. As such, the size of the spheres changed, expanding and contracting with the approaching and receding movements of the limbs. Also, the distance between the spheres would change according to the perspective of the observer. These stimuli could be presented with a front or with a back direction, but only the above referred perspective clues changed with each direction.

The experiment was programmed with a custom application based on Open GL running over Vr/Net Juggler software. For the projection we used a computer with a graphics card Nvidia Quadro FX 4500. The image was displayed by a 3 chip DLP projector Christie Mirage S+4K at 1400x1050 pixel resolution, and a refresh rate of 60 Hz.

The auditory stimuli consisted of step sounds animated in Open AL software, an application which models audio sources in a 3D space to be heard by a single listener who also has a determined position in space, accounting for distance, object mass, air and floor density/propagation, interaural time and level differences and reflections. Sounds were animated by placing the sources at the same distance as for the visual walkers in a virtual world with ground but no walls, at 1 m per second, one step cycle per second (matching the visual motion). There were two types of sound tracks, as the steps could be approaching or moving away. The auditory stimuli were produced by a Realtec Intel 8280 IBA sound card and presented by two large speakers placed at both the image limits.

Estimating sound distance and motion in depth depends on several acoustic properties, such as familiar sound level, reverberation, spectral shifts and changes in sound pressure at the listener's ears (Coleman, 1963; Blauert,

1983; Ashmead, Deford, Nortington, & Nortington, 1995). All these features were available in the auditory stimuli, although it is assumed pressure changes were the most influential cues in discriminating motion direction (Shaw, McGowan, & Turkey, 1991; Zou et al., 2007). The sounds arriving at the participants' ears were measured by a Bruel & Kjaer Head and Torso Simulator type 4128-C with Bruel & Kjaer Pulse Analyzer type 3560-C and Pulse CPB Analysis software. Auditory information at both ears was comprised between 50 Hz and 8 kHz, but it peaked at 630 Hz both for looming and receding sounds. Approaching sounds reached the ear at a maximum level of 26.4 dB SPL for the first step and 34.6 dB SPL for the last step. Receding sounds had a maximum level of 26.4 dB SPL for the first step and 23.2 dB SPL for the last step.

In the audiovisual condition, both auditory and visual stimuli were triggered in phase with a constant 4 ms stimulus onset of the visual output, a delay considered well within the simultaneity window (see Van Eijk, Kohlraush, Juola, & Van de Par, 2008). Signal delays were measured by the custom-built latency analyzer with an Arm7 microprocessor coupled with light and sound sensors (see section 2.2.3).

2.2.2.3. Procedure

The experiment consisted of a 2-forced choice direction discrimination task, using the method of constant stimuli. There were three experimental conditions: auditory, visual, and audiovisual. In the audiovisual condition both the point-light walker and the step sounds were presented in phase, and always moving in the same direction. Each presentation had the duration of 2 s, corresponding to two step-cycles or four steps. There were six different types of stimuli, as each experimental condition was presented in a front and in a back direction. Each stimulus type was repeated 100 times. The interval between each trial lasted for 1 s.

The participants were tested individually in a darkened room. They were first informed about the nature of the stimuli: moving points that could resemble

a walking human, step sounds, or both simultaneously. The instruction was to indicate as quickly as possible to where the stimuli were directed in each presentation. The participants answered on a touch screen, with two defined response areas, corresponding to the two possible directions of the stimulus (front or back).

2.2.3. Results

Correct answers were compared for the visual, auditory and audiovisual conditions and for the front and back directions of the stimuli. Differences between the percentages of correct answers for both stimuli directions were analyzed with paired samples t-tests, as all results followed normal distributions. The differences in variance between all three experimental conditions across the front and back directions of the stimuli were tested with a two-way repeated measures F ANOVA (3x2). The frontal bias was specifically addressed by analyzing the proportion of “front” answers; two-way F ANOVAs were also performed.

In the visual condition, the percentage of correct answers was of 67.2. We found a face forward bias, as the correct answers were higher when the stimulus was oriented to the participant (86 percent) rather than when it was oriented away from him (48 percent) (Figure 12). The difference between the correct answers for the front and back stimuli was significant ($t_{(9)}$=19.44, $p<0.001$). Additionally, the standard deviation for the percentage of correct answers was higher for the back direction of the point-light walker (δ=17) when compared the frontal direction (δ=10). With no exception, all the participants had worse results for the back stimuli.

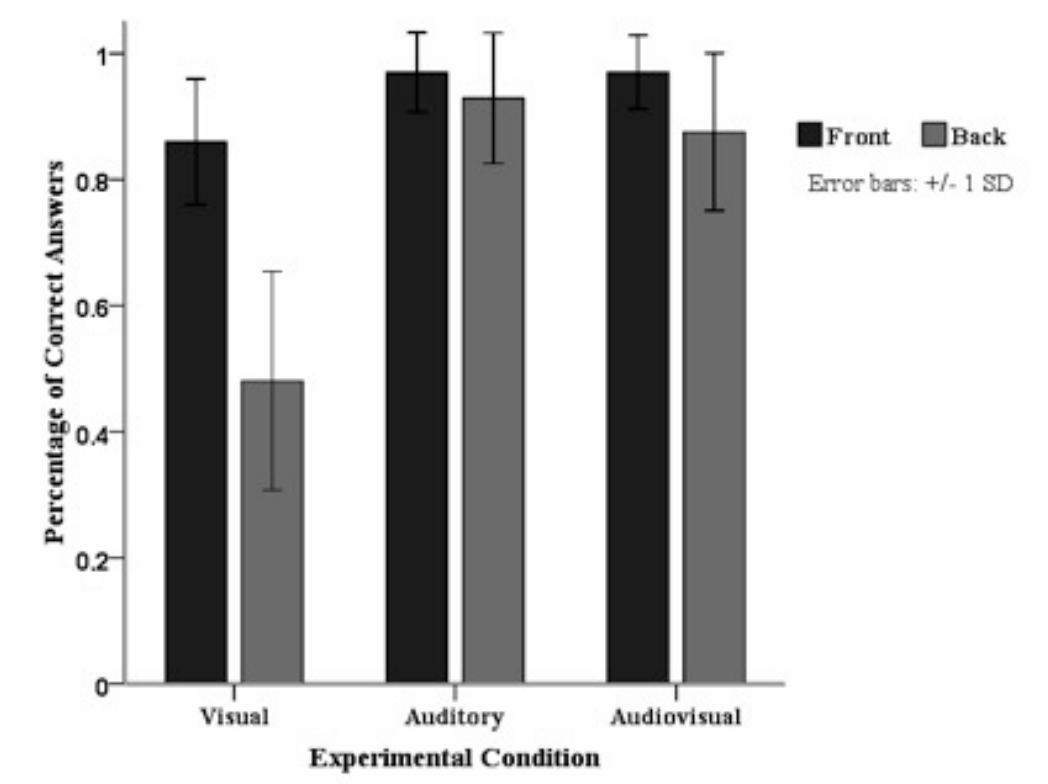

Figure 12. Percentage of correct answers for front and back stimuli in the Visual, Auditory and Audiovisual experimental conditions.

In the auditory condition, results showed that the footstep sounds had robust perceptual properties, as their direction could correctly be identified in 94.9 percent of the trials. Interestingly, in this condition results were also better when the stimulus was looming: 97 percent correct answers for front directions against 93 percent in back directions, and these differences were significant ($t_{(9)}$=4.94, p<0.001) The standard deviation was higher when the stimulus was receding (δ=13) compared to when it was looming (δ=6). These observations are consistent with a looming bias that generalizes across conditions.

The audiovisual results were also better than the visual (Figure 12). Participants answered correctly in 92.3 percent of the trials. There were again better results for the frontal stimuli directions (97.1 percent) than for the back ones (87.6 percent) and this difference was statistically significant ($t_{(9)}$=8,87, p<0.01). The standard deviation was also smaller in the front results (δ=06), when compared to the back ones (δ=14).

The two-way F ANOVA analysis confirmed the difference between the

results of the front and back stimuli ($F_{1,9}$=361, p<0.001) and revealed an interaction effect between experimental condition and stimulus direction ($F_{2,9}$=134, p<0.001). The differences between the visual, auditory and audiovisual results were statistically significant ($F_{2,9}$=385, p<0.001). However, post-hoc Sheffé test revealed that these differences were only significant between visual and auditory condition (p<0.001), and between visual and audiovisual condition (p<0.001). The difference between auditory and audiovisual condition was not significant (p=0.25, n.s.). Overall, these results demonstrate that sound was a determinant factor in the final audiovisual percepts.

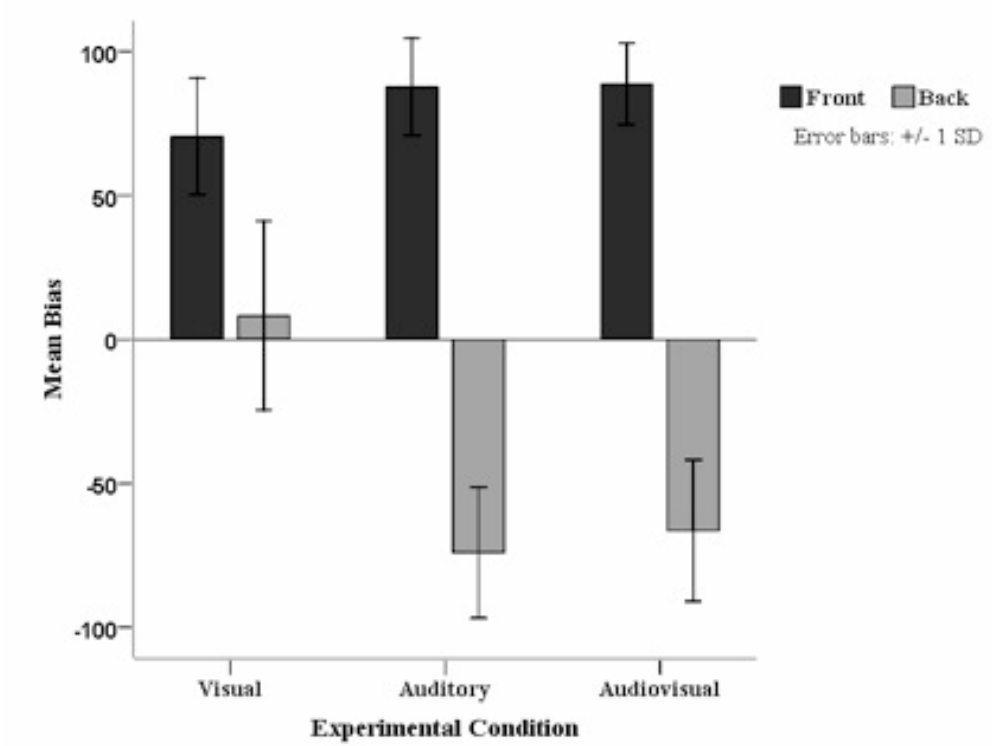

Figure 13: Frontal bias in all experimental conditions. The base line refers to an equal ratio of “front” and “back” answers. Positive values reflect more “front” than “back” answers, negative values reflect the opposite. A bias level of 100 would reflect 100 per cent of “front” answers.

By computing the frequency of front answers, we are able to analyze the perceptual bias. To obtain a direct measure of the frontal bias the results were all converted so that a 50 percent proportion of front answers would correspond to zero bias, 100 percent front answers would be the maximum bias and more back than front answers (less than 50 percent front answers) would be a negative bias or back bias (figure 13).

In the visual condition the mean bias level was 39.7, mostly because when the stimuli were presented facing away the participants failed to answer accordingly, still showing a slight tendency to answer front (8.3 bias). In the auditory condition, results show that participants effectively responded “front” when the stimulus was looming (a strong frontal bias level of 87.7) and responded “back” when the stimulus was receding (a negative bias level of -74). The mean results of the auditory condition show a bias of 6, meaning that effectively there were more “front” than “back” answers to the auditory stimulus, but few in number and inconsistent across subjects. The audiovisual bias was of 11.2 in average, showing that although not totally eliminated, the visual bias was significantly reduced. In this condition, not only for frontal but also for back stimuli, the results resembled those obtained in the auditory condition.

Comparing the bias levels across conditions and stimulus direction, we observed once again that there were differences between front and back stimuli ($F_{1,9}$=4650, p<0.001) and that there was a significant interaction between experimental condition and stimulus direction in bias results ($F_{2,9}$=426.7, p<0.001). Differences in bias results across conditions were also significant ($F_{2,9}$=120.5, p<0.001), but again Sheffé post-hoc test revealed that these differences were only significant between visual and auditory (p<0.001) and visual and audiovisual conditions (p<0.001). The bias results were not different between auditory and audiovisual conditions (p=0.059, n.s.). In sum, the frontal bias is almost absent in the audiovisual condition, as it is in the auditory condition. We conclude that human step sounds were a strong cue in disambiguating human motion direction.

2.2.4. Discussion

Perceiving biological motion, although a frequent and relevant task, might be highly biased concerning body direction estimates. The scope of this section was to understand if providing additional, non-visual cues could reduce this bias. More specifically, we intended to analyze the impact of auditory footstep sounds over the motion direction estimates.

We presented visual point-light walkers, auditory step sounds or both simultaneously, moving towards or away from the participants.

In the visual condition, we found an effect of the face forward bias. Participants responded more frequently “front” than “back” for both frontal and back-oriented stimuli. These results are by themselves surprising, as there were perspective cues available that could have helped discriminate the stimulus direction. Results confirmed that vision alone might be a limited source of motion direction information in biological motion. Conversely, auditory results yielded good, near-ceiling, performances. This was a new finding, as the perceptual accuracy in determining motion direction of naturalistic biological sounds had never been established before. Our results show that step sounds might be considered as a viable source of motion direction information. In the audiovisual condition, good discrimination levels were found. The face forward bias was almost absent and did not differ from the bias obtained in the auditory condition.

Research with rigid motion had shown that sound tends to be perceptually realigned in order to match the visual motion, but the opposite does not occur (Soto-Faraco et al., 2002). Sound did, however, disambiguate visual motion direction in highly undefined presentations (e.g. Sekuler et al., 1997). In those works, the streaming bias was largely reduced and the alternative motion pattern, the bouncing, arose.

Analogously, our visual stimuli were ambiguous and biased. This level of ambiguity might be a key to explaining why sound exerted such influence in the

audiovisual percepts. In the results we report here, sound did not realign with the visually perceived motion, and it actually seemed to dominate as the main source of directional information.

These results seem to point to a global benefit of multisensory cue integration. It becomes arguable that under undefined displays humans choose the least costly interpretations (such as a forward moving walker), but additional sensory cues take over such interpretations, allowing for the most accurate estimate to strive.

Following these findings, a new experiment was developed, to specifically address the benefit of multisensory integration with congruent audiovisual cues.

2.3. THE MULTISENSORY INTEGRATION OF BIOLOGICAL MOTION

2.3.1. Introduction

Ever since multisensory processes started to be studied, either in psychophysics, sensory processing, computer sciences, or other related areas of interest, there has been a search for the modeling of the integrative processes and their effects.

Several models have been proposed to account for signal interactions in multimodal cue integration. Classical hypotheses like modality precision, directed attention and modality appropriateness (for a review, see Welsh & Warren, 1980) assumed that, given a discrepancy between two sensory modalities, one modality would bias the other, becoming dominant. These models accounted for several effects, such as those addressed in 2.1.2. For example, in the ventriloquist effect and in the rubber hand illusion, subjects relied on the visual cue to determine the position of an object or limb. The modality appropriateness model explained this sensory bias assuming that vision was the most appropriate sensory modality to convey spatial attributes. This model thus predicted that, in spatial tasks, vision would always become the dominant cue. Conversely, in the double flash illusion and in temporal ventriloquist effect, subjects failed to report visual number, duration or rhythm, as the auditory cues biased the estimates. Here, the modality appropriateness model predicted that audition was the most accurate sensory modality to access temporal-related events.

Most recently alternative models gained relevance, in which signal integration was conceptualized as a probabilistic process (e.g. Massaro & Friedman, 1990; Ernst & Bülthoff, 2004). The Maximum Likelihood Estimation theory (MLE) has earned growing attention by explaining the traditional crossmodal biases as an imbalance in relative reliability of each sensory cue. Here, sensory dominance is explained not by a winner-take-all mechanism, as

previously hypothesized, but by an imbalance in relative cue weighting. For an example, in the ventriloquist effect, vision provided more reliable spatial information than audition, which might be expressed in a higher weight of the visual spatial estimate, compared to the weight attributed to the auditory estimate. Alais and Burr (2004b) demonstrated this hypothesis in an experiment where the visual cue was degraded. Here, while more noise was added to the visual stimulus, the auditory stimulus gradually gained relevance in the spatial estimates.

According to this model, multisensory integration allows an accuracy maximization by the combination of redundant cues. It is assumed that sensory estimates are naturally noisy and vary in precision. The combination of congruent signals decreases uncertainty, and therefore variability (Clarke & Yuille, 1990). The integrated percept is optimal, as it has the maximum reliability, and thus the least possible variability.

Over the last years, a few optimal mechanisms were found in multimodal cue integration, namely in the discrimination of visuo-tactile size (Ernst and Banks 2002), audiovisual position (Alais & Burr, 2004b; Battaglia et al., 2003), event number (Andersen, Tiippana, & Sams, 2005; Shams, Ma, & Beierholm, 2005) and arrival time (Wuerger, Meyer, Hofbauer, Zetzsche, & Schill, 2010). In velocity discrimination, however, integration effects are not as robust. Vision affects the perceived velocity of sounds in a weighted average fashion (López-Moliner & Soto-Faraco, 2007), but there is only a weak tendency for optimal integration of the velocity estimates, which defaults to probability summation when unimodal weights diverge (Bentvelzen, Leung, & Alais, 2009).

Here, we were interested in assessing the integration mechanisms of audiovisual biological motion signals at different walking speeds. In walking motion, both velocity and rhythmic cues are present. The translational global motion of the walker provides velocity information, while the frequency of steps in the auditory stimuli and local limb motion in the visual stimuli bring rhythmic information. Such phenomena are interdependent, as when the human walker accelerates or decelerates, so does the frequency of the local periodic events. Therefore, for biological motion, velocity and rhythmic information are coupled

and could not be isolated without violating biomechanical constraints and the corresponding perceptual coherence. In rigid motion, as opposed to biological motion, these cues are independent, but it is still a mater of debate how subjects recover speed from them. In this study, we did not separate both sources of information, as we intended to obtain plausible biomechanical and perceptual walking representations. Therefore, while specifically addressing walking speed, this study comprised simultaneously rhythm and velocity discriminations of human motion.

Having in mind that auditory delays are frequently required for auditory-visual subjective simultaneity to be reached (e.g. Alais & Carlile, 2005; Arrighi, Alais, & Burr, 2006; Eijk et al. 2008; Vatakis and Spence 2006a, 2006b; Di Luca, Machulla, & Ernst, 2009), a preliminary experiment was conducted. We measured simultaneity judgments at several audiovisual walking asynchronies and obtained a point of subjective simultaneity.

The main experiment consisted of a velocity discrimination task. Visual, auditory, and audiovisual walkers, which either were congruent or incongruent in velocity, were compared to a standard audiovisual stimulus walking at an average rate. We hypothesized that congruent audiovisual stimuli would lower the discrimination thresholds. The discrimination accuracy was compared to optimal predictions and analyzed at different velocity congruency levels. Results were discussed in light of the MLE model and the temporal windows of integration.

2.3.2. Method

2.3.2.1. Participants

Two subjects took part in the preliminary experiment (*Measuring Perceived Asynchronies*) and seven subjects took part in the main experiment.

In the preliminary experiment, one participant was an author (CA) and the other was an untrained naïve subject (SM), one left and the other right handed, one trained and the other untrained in the auralized sounds. In the main experiment, one participant was an author (CA), 3 were trained but naïve to the purpose of the study (BO, LA, RO) and 3 were untrained naïve subjects (CO, JO, RA), 3 left handed and 4 right handed. Both the author and the trained participants were trained in the auralized sounds. All underwent visual and auditory standard screening and had normal or corrected-to-normal sensory acuity.

2.3.2.2. Stimuli and Apparatus

The visual stimuli were point-light walkers of 13 white dots, generated in the Laboratory of Visualization and Perception (see section 2.2.1) by a Vicon motion capture system with 6 cameras MX F20 at 240 Hz and a set of custom LabVIEW implemented routines. All stimuli corresponded to the correct motion coordinates of a 17 year-old male, 1.87 m high, walking at five translational velocities (1.1, 1.3, 1.5, 1.7 and 1.9 m/s) with constant time intervals between steps (590, 560, 530, 500, and 470 ms, respectively, with a maximum variability of ±4 ms). When virtualized as walking at 10.73 m from the observer, the visual angular velocities corresponded to 6, 7, 8, 9, and 10 deg/s. The point-light stimuli were presented through a computer with a graphics card Nvidia Quadro FX 4500, and a 3 chip DLP projector Christie Mirage S+4K, with 1400x1050 pixel resolution at 60 Hz. The image was projected onto a 2.78x2.09 m screen area in a dark room. On the screen, the white dots (54 cd/m^2) moved against a black background (0.4 cd/m^2) and each dot subtended 0.9 deg of the visual angle.

The auditory stimuli were step sounds from the database of controlled recordings from the College of Charlston (Marcell, Borella, Greene, Kerr, & Rogers, 2000). They corresponded to the sound of a male walking over a concrete floor. These sounds were auralized by a MATLAB routine with generic head-related transfer functions (see section 2.2.2) from the MIT database (http://sound.media.mit.edu/resources/KEMAR.html), as free-field, at the same

five velocities and step frequencies as the visual stimuli, at a distance of 10.73 m in space. The auditory stimuli were presented with a Realtec Intel 8280 IBA sound card through a set of in-ear earphones Etymotics ER-4B. The sounds reached the listeners at a maximum level of 35 dB SPL, peaked at 260 Hz, as measured with a Brüel & Kjaer Head and Torso Simulator type 4128-C with Brüel & Kjaer Pulse Analyzer type 3560-C and Pulse CPB Analysis software.

The audiovisual stimuli were programmed and presented through a custom application running on top of Vr Juggler code. The temporal alignment of the sound onset and the displayed frames for each stimulus was checked and adjusted prior to the experiment using a custom-built latency analyzer (see section 2.2.3). The sensors were targeted at the precise frame of the visual footstep and at the earphones' paired output. Due to limitations of the visualization system, the SOA accuracy was of ±1 ms and the precision across trials was within a maximum range of ±6 ms.

2.3.2.3. Procedure

i) Measuring Perceived Asynchronies

We first wanted to know the magnitude of the perceived asynchrony between the visual and auditory cues to motion and the width of the integration window to help us interpret the data from the main experiment. Audiovisual congruent stimuli moving at 6, 7, 8, 9 and 10 deg/s, with step frequencies of 1.69, 1.78, 1.89, 2.0, and 2.13 Hz respectively, were presented with a duration of 1.6 s. Stimuli delays consisted of the time differences between the frame when the visual foot touched the ground and the moment when the auditory step sound reached the subject's ear. Each SOA was randomly varied across all stimulus velocities, from -120 ms (auditory lead) to 120 ms (visual lead) in 30 ms intervals. All stimuli were pseudo-randomly presented, 20 trials per asynchrony level for each velocity (5x20), with an interstimulus interval of 0.8 s. Participants had to judge if the visual and the auditory signals were synchronized or not.

ii) Main Experiment

In each trial of the main experiment there was a standard stimulus and a test stimulus. The standard stimulus was always an audiovisual congruent walker, moving at 8 deg/s, with a step frequency of 1.89 Hz and the duration of 1.6 s. The test stimuli could either be unimodal (visual or auditory), or bimodal. There were 5 possible stimulus velocities of the visual and auditory stimuli: 6, 7, 8, 9 or 10 deg/s, with the respective 5 step frequencies of 1.69, 1.78, 1.89, 2.0, and 2.13 Hz.

The bimodal stimuli corresponded to all visual and auditory velocities factorially paired (5x5), which were therefore congruent or incongruent. Incongruence varied from 1 deg/s of velocity difference to 4 deg/s. There were three possible test stimulus durations: 1.5, 1.6 and 1.7 s. This duration variation was too small to introduce differences in the number of visual and auditory steps, but it was large enough to assure that the starting and ending points as well as the first and last frame of the point-light stimuli were not informative. As such, for example, slower stimuli did sometimes travel longer distances than faster stimuli.

Both visual and auditory signals moved from left to right and crossed hemifields at half the stimulus duration. There were always 3 steps, and stimuli were arranged in such way that the second step occurred precisely when crossing the visual mid-line. Thus, in the bimodal incongruent stimuli, the second step would always be perceptually aligned and the mismatch was equally distributed by both the first and the third step (see figure 14). Also, and as a result from the preliminary experiment, where we found that a delay was required for simultaneity to be reached (see figure 15), all sounds in this experiment lagged an additional amount of 30 ms.

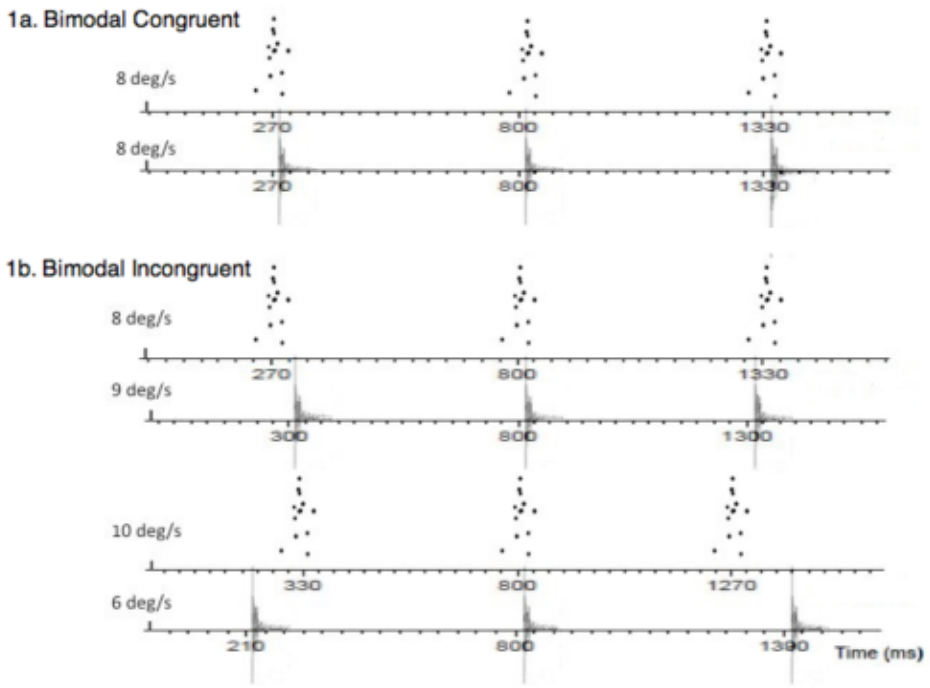

Figure 14. Examples of bimodal stimuli. 1a. Bimodal Congruent. Both visual (upper line) and auditory (lower line) signals moved at 8 deg/s (standard stimulus). The first step occurred 270 ms after the trial start, the second step occurred at the 800th ms (half of the 1600 ms long stimulus) and the third step occurred at the 1330th ms. There was an actual delay of 30 ms between the moment when the visual foot reached the ground and the moment when the sound reached the subject's inner ears. 1b. Bimodal incongruent. The upper example is a stimulus with 8 deg/s of visual velocity 9 deg/s auditory velocity. The second step co-occurred after 800 ms, but the first and last steps were mismatched; the sound arrived 30 ms later than the image in the first step and 30 ms earlier in the third step (0 ms later in the first step and 60 ms later in the last one, considering the 30 ms audio lag). The bottom example is the most incongruent bimodal stimulus. Here, the maximum time mismatch was 120 ms (150 ms in the last step, with the audio lag).

There were 20 trials per stimulus, distributed in 4 experimental blocks. Each block presented all stimulus types (visual, auditory, and bimodal) pseudo-randomly. In a 2-IFC task, participants were instructed to respond in which interval (standard or test) was the fastest walker.

2.3.3. Results

2.3.3.1. Perceived Asynchronies

In the preliminary experiment several auditory-visual SOAs were tested in a simultaneity judgment task. The relative frequency of "synchronized" estimates as a function of SOA was well fitted by a Gaussian curve (figure 15).

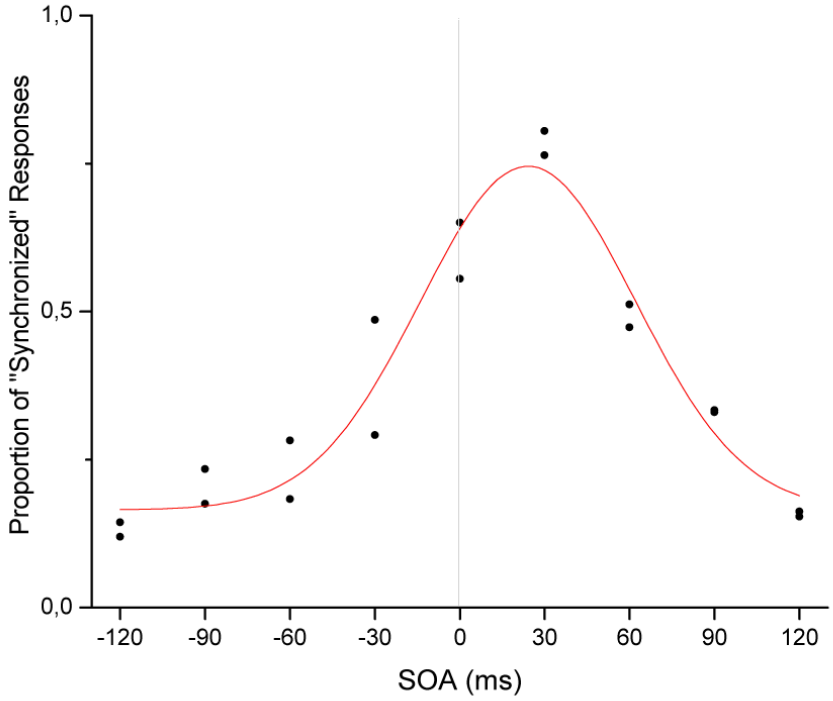

Fig. 15. Proportion of the pooled synchronized responses as a function of each auditory-visual delay, from -120 ms (auditory first) to 120 ms (visual first). Each dot corresponds to the responses of each of the two participants (100 trials/dot). Gaussian fit with mean=25 ms and σ=38.2 ms.

The overall results show that as the absolute SOA increased the average simultaneity judgments decreased. However, the distribution revealed a positive bias, as the highest proportion of synchronized answers occurred not for the stimuli with the 0 ms SOA (relative frequency of 0.6), but for the stimuli with 30

ms SOA, vision had to be presented first (a 0.78 relative frequency). The best-fit Gaussian, adjusted with MLE fitting procedure, was thus shifted to the right, with a point of subjective simultaneity (PSS) of 25 ms (95% CI of 5.27) and a discrimination threshold (1 σ) of 38.2 ms.

These results support the broad understanding that for subjective simultaneity to be reached visual lead is required (e.g. Arrighi et al. 2006).

Results in this experiment were taken into account in the subsequent experimental design. Since our stimuli were displayed with several realistic depth cues such as perspective, relative change of the walkers' size and dot dimension, as well as sound attenuation effects, we chose to keep a delay of the auditory signal for all stimuli, consistent with our distribution peak and signal propagation differences. Therefore, in the main experiment, congruent bimodal walkers (proximal stimuli) were always mismatched by 30 ms and incongruent walkers followed the same rule, with varying delays.

2.3.3.2. Main Experiment

In the main experiment, visual, auditory and audiovisual walkers at different speeds and congruency levels were compared to a standard bimodal walker at 8 deg/s. As we used a speed discrimination task instead of a detection one (as it has often been the case) in which the standard was always bimodal some caveats need to be taken into account in the analysis. We compared auditory, visual and audiovisual psychometric curves, which have been obtained against the same (bimodal) reference. When uncertainty is high, potential biases might appear in the unimodal conditions in the form of choosing the bimodal standard more often. If subjects did so, we would be able to detect it by observing fluctuations in the mean (PSE) of the curve. Nevertheless, to further clarify the sensory modality impact of the standard stimulus, an experiment not reported here was run with a similar methodology as in the main experiment, where visual, auditory and audiovisual standards were used with 3 all-naïve participants. The results from that experiment revealed no differences in the discrimination slopes. Once these cautions were taken, MLE predictions

in variance reduction were tested for all subjects individually and for the pooled data results.

As there were no differences between results for each stimulus duration, nor starting/ending points, all data were pooled together for analysis purposes. All results in the main experiment were presented with respect to stimulus velocity (not stimulus step frequency). This was a matter of choice, mainly due to the fact that participants were indeed instructed to make a velocity judgment.

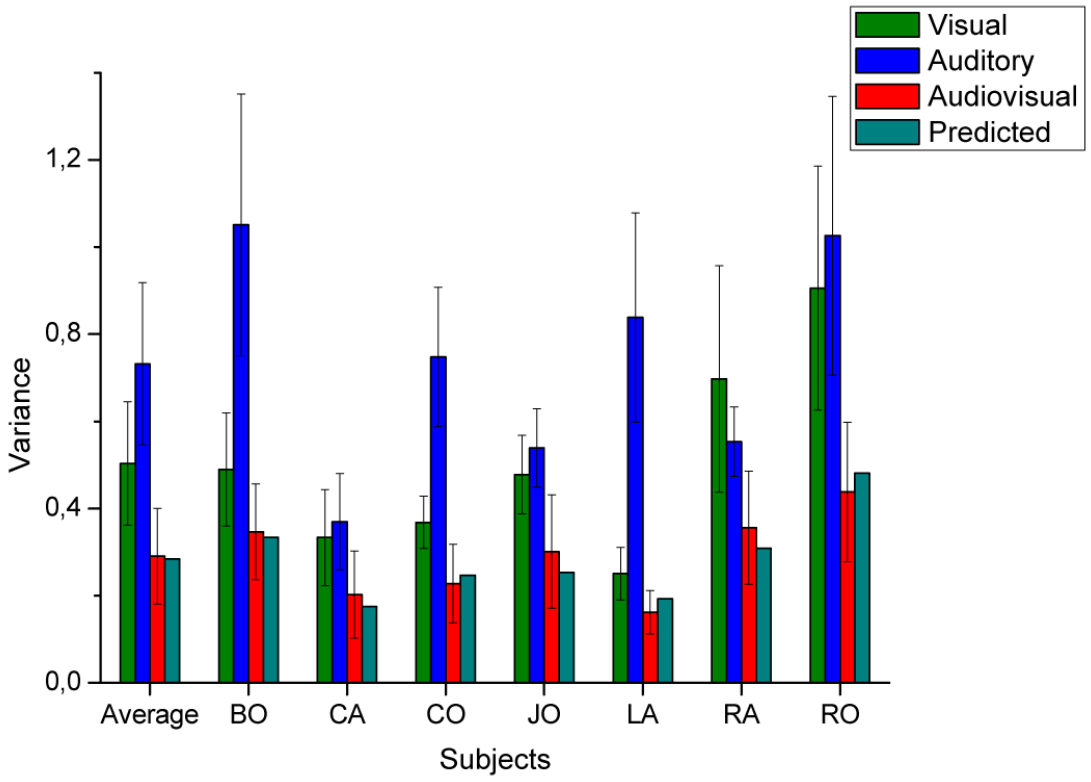

Figure 16. Velocity discrimination variances in the visual, auditory and audiovisual congruent stimuli and the optimal integration predictions for each subject. Variance values were obtained by the squared standard deviations of the cumulative Gaussians fitted for each participant. From those values, the predicted optimal integration variances were calculated as $\sigma^2_{AV}=\sigma^2_A\sigma^2_V/\sigma^2_A+\sigma^2_V$. Error bars display 95 percent confidence intervals obtained from each Gaussian's standard error.

We first obtained the proportion of faster responses against the velocity for the different conditions and fitted cumulative Gaussians to these data

distributions. This was done for individual subjects and pooled data. The parameters of the fit, mean and standard deviation (SD), defined the point of subjective equality (PSE) and the discrimination threshold (JND) respectively. The individual and group variances for each condition were obtained from the estimated SD. The group and individual variability in terms of the variance is presented in figure 16.

According to MLE, the benefit of multiple sensory cue integration lays in the reduced variance and therefore increased precision of the perceptual judgments. Overall, unimodal precision varied greatly across subjects. Variance was lower in the visual than in the auditory discriminations of 3 participants, and 4 participants had overlapping unimodal variances. Nonetheless, optimal integration predictions were within the 95% confidence interval of the audiovisual congruent condition for all subjects. The discrimination data pooled over subjects are displayed in figure 17 together with the best cumulative Gaussian fits.

Congruent audiovisual stimuli yielded the steepest slopes (smallest deviation, σ=0.6), better than those of visual (σ=0.8) and auditory (σ=0.92) stimuli alone, and significantly different from best unimodal (bootstrap p=0.043). Interestingly, bimodal stimuli with a velocity incongruence of only 1 deg/s also revealed a steep function (σ=0.78) with a slope which did not differ from bimodal congruent nor best unimodal (bootstrap p=0.554 and p=0.32 respectively). The integration disruption was observed in the stimuli with a velocity incongruence of 2 deg/s (σ=1.08), and bimodal stimuli with larger velocity differences did not fit Gaussians. Importantly, PSE values remained close to standard in all conditions.

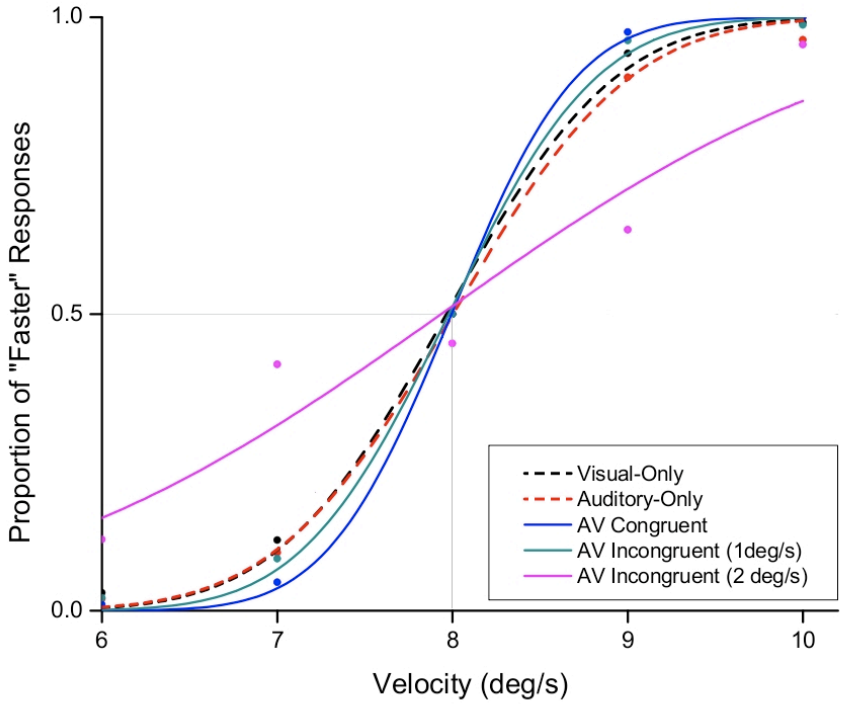

Figure 17. Best-fit cumulative Gaussians, with unconstrained parameters, obtained from the faster-than-standard judgments of all participants (420 trials per dot) for the unimodal (dashed lines) and audiovisual (full lines) stimuli.

The unexpected steepness of the slopes at the smallest level of incongruence led us to wonder about a window of audiovisual integration, which might stretch to such velocity mismatch. While little is known about windows of velocity integration, there is a growing body of information on the temporal windows of such processes (e.g. Van Wassenhove, Grant, & Poeppel, 2007). Taking a new look over figure 1, we might notice that indeed our incongruent stimuli do vary in temporal alignment. Summing up the 30 ms auditory delay we added to all bimodal stimuli (see preliminary experiment), we have that congruent stimuli were mismatched in 30 ms, 1 deg/s incongruent stimuli were maximally mismatched in 60 ms and 2 deg/s incongruent stimuli in 90 ms. In figure 5 the discrimination thresholds are displayed as a function of the temporal mismatch of the bimodal stimuli.

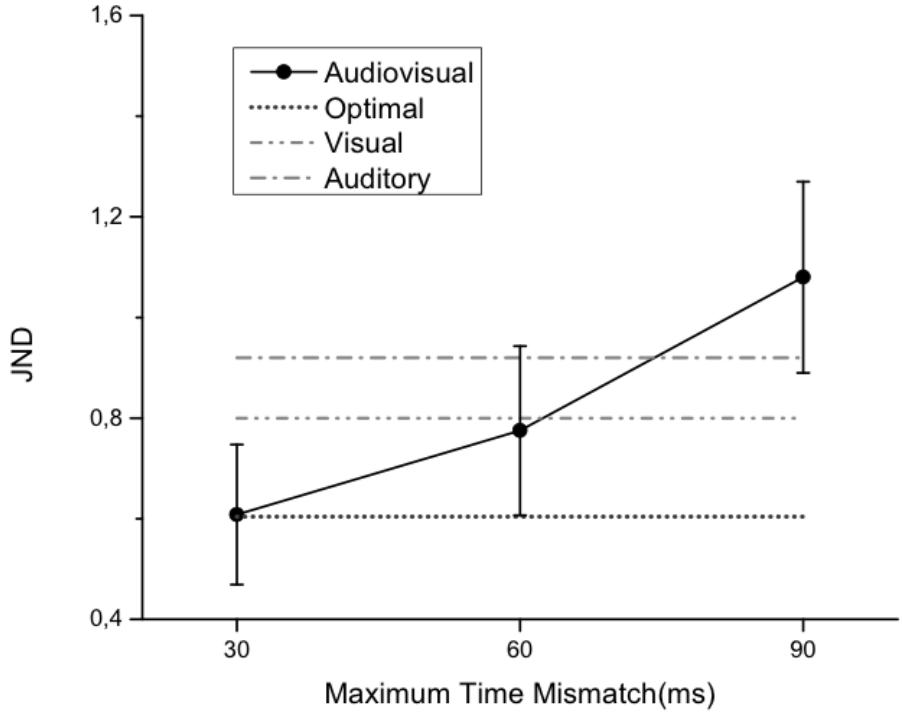

Figure 18. Just noticeable differences (JND) as a function of the temporal alignment. Error bars are confidence intervals calculated with the bootstrap technique in 1000 simulations (Efron & Tibshirani, 1993).

The JND of the congruent stimuli, which were mismatched in 30 ms, fell well within optimal integration prediction and, as said above, was significantly lower than the best unimodal. The sensitivity of stimuli with the 60 ms mismatch was better than the best unimodal (visual) and worse than audiovisual congruent, but did not differ from either. Stimuli mismatched in 90 ms had the highest discrimination thresholds. These results are consistent with the optimal mechanisms in the processing of biological motion. They further support the conceptualization of a window of integration within which multisensory biological signals might be fused.

2.3.4. Discussion

The experiments we reported in this section were devoted to understanding the multisensory integration of biological motion at several walking speeds. In a preliminary experiment, we assessed the audiovisual PSS of the point-light steps and the footstep sounds to accurately produce perceptually aligned and congruent bimodal stimuli. We found that stimuli with a 30 ms auditory lag were the most perceived as synchronized. This result is consistent with the assumption that visual lead is necessary for perceived simultaneity of audiovisual stimuli, and it is in line with several PSS values recently reported in literature (e.g. Arrighi et al., 2006; Vatakis et al., 2006a, 2006b; see Van Eijk et al., 2008 for a review). Also, this finding supports the assumption that the necessary asynchrony might follow physical rules (Alais & Carlile, 2005; Sugita & Suziki, 2003), since for our stimuli's virtual localization an actual 30 ms auditory delay was to be expected. The width of the Gaussian fit for the synchronized judgments might also be informative, as it revealed a JND of ±38.2 ms. Such value, which reflects a temporal window of perceived audiovisual simultaneity for human walkers, is consistent with the integration window found in the main experiment (stimuli lagging from -13.2 to 63.2 ms would be perceived as synchronous). This window is smaller than that reported for audiovisual speech, object action and music play (Van Wassenhove et al., 2007; Vatakis & Spence, 2006a, 2006b), but larger than the window for flashes-clicks, and other non-biological motion signals (e.g. Arrighi et al., 2006; Van Eijk, 2008). However, for a more accurate understanding of the simultaneity perception in audiovisual biological motion, a broader study is needed.

In the main experiment, we focused on the multisensory integration of biological motion at different walking speeds. There were visual, auditory and bimodal human walkers, congruent and incongruent, which were compared to a standard audiovisual walker in a velocity discrimination task. All participants discriminated more accurately the bimodal congruent than the unimodal stimuli, and their accuracy with such stimuli conformed to optimal integration mechanisms.

Optimal mechanisms had been found before with audiovisual motion in a time-to-arrival experiment (Wuerger et al., 2010). However, those findings were obtained in an experiment where participants had received a large amount of training, with feedback, before the experimental sessions. Furthermore, during those procedures, subjects were only exposed to congruent stimuli, and therefore they would be able to develop optimal strategies. Conversely, during our experiments, participants remained untrained, since most of the randomly displayed bimodal stimuli were incongruent (20 out of 25), and thus integration learning was impaired. Another study directly assessed the audiovisual integration of velocity signals and looked for optimal integration mechanisms (Bentvelzen et al., 2009). In that study, the authors had found only a weak tendency to follow MLE predictions, as just some of the subjects did discriminate velocity in an optimal fashion, whereas the others had performed worse than with the best unimodal stimuli. The authors had interpreted this as separate integration mechanisms derived from different unimodal weights: only those subjects who had similar visual and auditory accuracies would follow MLE. Here, in contrast, optimal integration was a robust effect across all subjects, despite different unimodal discrimination performances, previous training or experimental goal awareness.

The integration effects that we find are consistent with previous neuroimaging data. The posterior superior temporal sulcus (STSp), a brain region selectively activated for the sight of biological motion stimuli, is also hypothesized to be a structure where multisensory integration might occur (Calvert, 2001). Single cell-recordings showed significant interactions between visual and auditory processing of human actions in this area (Barraclough et al., 2005), and fMRI studies revealed that STSp responds to the auditory stimulation of human steps (Bidet-Caulet et al., 2005). Our findings might be read as further evidence for a possible neural specialization in the processing of human motion-related signals.

An unexpected outcome of our findings was the discrimination slope of the stimuli at the smallest level of incongruence. The JND of these stimuli was still lower than in the best unimodal and remained close to optimal. A window of

integration might explain this result. According to this hypothesis, optimal mechanisms are not exclusive to a given pair of audiovisual stimuli, but might occur along a small continuum of slightly mismatched stimuli, which still allows for unity assumption. Such windows of integration have long been studied, and carry an extensive body of experimental support in audiovisual speech (e.g. Yabe, Tervaniemi, Reinikainen, & Määtänen, 1997; Van Vassenhove et al., 2007). In our study, stimuli at the smallest level of incongruence were mismatched in 1 deg/s of angular velocity and were also temporally mismatched in -0/+60 ms. Therefore, we propose the existence of a temporal window of integration in biological motion signals, which might be as large as 60 ms wide. Interestingly, this value is within the -13.2/+63.2 ms synchrony interval we report in the preliminary experiment. On the other hand, this window is much smaller than the ±200 ms audiovisual integration window found in audiovisual speech (Van Vassenhove et al., 2007). A narrower window in speed perception would have a higher adaptive value as it would allow to ascertain small discrepancies in situations when judging the speed accurately is crucial.

In sum, firstly we argue that audiovisual congruent walkers benefit from cue integration in an optimal fashion. We cannot fully claim for MLE predictions, as we didn't address PSE weighting by directly manipulating cue reliability. Nevertheless, we should stress that we found an optimal benefit of multisensory integration in lowering discrimination thresholds. Secondly, we argue that these integration processes might occur within a well-defined temporal interval of stimulus alignment. Further research should approach the spatiotemporal constraints of biological motion multisensory cue fusion.

III. FINAL DISCUSSION

This thesis comprehended 3 experimental sections, corresponding to 3 major research objectives. The global scope of the work here described was the study of the multisensory processes in the perception of biological motion. The goal was to understand the benefit of redundant bimodal cue information in perceptual accuracy, if congruent auditory sounds would decrease discrimination errors and bias, and would reduce uncertainty. From the implementation of auralized sounds, we also intended to validate the spatial properties of the sound stimuli and to develop a protocol for the preparation of subjects for use these stimuli. There were therefore, by section order, 3 sets of experiments: Locating Auditory Sources with Non-Individualized HRTF-Based Auralizations; The Effect of Auditory Cues on Biased Biological Motion; and The Multisensory Integration of Biological Motion.

3.1. Locating Auditory Sources with Non-Individualized HRTF-Based Auralizations

Here we intended to test the auralized stimuli. As we intended to take full advantage of this technology in posterior experiments, our first concern was to obtain an empirical validation of the audio algorithms. Several questions needed to be addressed. Firstly, we intended to assure that listeners could recover spatial properties from these virtual sounds, thus validating them. Secondly, we were concerned with the localization accuracy, which should be precise enough that no alternative audio system would need to be used in the audiovisual experiments. Also, we were concerned with the lack of literature available on the listener's adaptation to these sounds and were interested in addressing this issue in a way that would provide us with both new findings and a practical solution.

In the first experiment - *Azimuth Accuracy Without Feedback*, we addressed the listeners' adaptation process to static non-individualized azimuth sounds without feedback.

We found that indeed subjects recovered spatial properties from the sounds, as the responses for front, back, lateral and intermediate positions were well above chance. Still, with such a small sampling of well-spaced sounds (sounds differed from one another in 45º), we expected higher accuracy levels. No ceiling effects were found. Also, throughout the experimental blocks, no performance improvement occurred. This was surprising for us, as subjects were expected do become gradually more familiarized with both the stimuli and the task itself. As they kept answering to the different azimuth stimuli, some stimulus regularities should have been found and some optimization should have occurred. We concluded that simple exposure is not enough for significant accuracy evolution to be achieved in short periods of time. As such, by the end of this experiment, two concerns remained: we still did not know if auralized sound localization was an ability which could evolve with training and if adaptation processes could take place; and we remained focused in assuring that a short protocol for the testing and preparation of subjects should be defined.

In a second experiment – *Learning Azimuths*, subjects were trained in sample sounds with a short program combining active learning and feedback. Here, we found that azimuth discrimination was best for the frontal and lateral stimuli, but that at the intermediate positions there was an auditory blur before training. The training program produced a significant performance improvement in all subjects. This finding demonstrated for the first time that auralized sound localization was indeed an ability prone to adaptation and learning processes. A crucial result was that the learning effect was not limited to the trained stimuli. The training benefit was also found in other, non-trained, azimuths, as if a generalization process took place. This encouraging finding led us to believe that indeed a global HRFT learning took place, as if subjects developed an understanding of the new anatomically based spatial map.

To better understand if this was the case, and if this learning was not limited to the adaptation to auralized the binaural cues (ITD and ILD), we developed a third experiment with elevation stimuli. In elevation, with azimuth kept constant, binaural cues are less relevant (e.g. Middlebrooks & Green, 1991). Here, only spectral cues, resulting mostly from the signal shadowing and shaping of the head and pinnae, are present.

In the final experiment - *Learning Elevations*, we tested the same training program, with stimuli varying in elevation and with fixed azimuth. Results were very similar to the previously found with azimuths. In the pre-test, elevation cues were barely discriminated, but there was a large improvement after training and this improvement was also found for untrained stimuli. This finding further supported the assumption that indeed new HRTF-shaped spectral cues were learned.

Several practical implications derived from this study. From our initial perspective, we obtained an empirical validation of the auralization system. This allowed us to extend the use of the virtual audio stimuli to other experiments. Also, we achieved a better understanding of the baseline accuracy in discriminating these sounds and developed a method for the preparation of subjects. Two key factors underlie our training program. First, knowing that the HRTF cues are learned and that this learning is extended beyond those trained sounds allows us to choose small sound samplings and thus reduce training duration. On the other hand, having as training criterion a standard accuracy level, such as 80 percent correct answers, allows us to assure that all subjects achieve a similar stage of acquaintance with the HRTFs. This is crucial in reducing the frequently found high intersubject variability in auditory accuracy and provides a homogenization of the experimental participants' abilities.

One matter of debate is the training technique. Choosing active learning and response feedback proved to be effective in short training periods, but other methods should be tested and compared. Among other solutions, the implementation of real-time audio should be considered. In real-time auralization, the auditory sources move as the listener moves his head. This would provide additional information, like continuous audio motion and

proprioceptive cues, which we believe would not only enhance the auditory experience but also largely reduce the adaptation time. However, several technical limitations pose some limits to this approach, and this technology is still under implementation in our facilities.

But other outcomes emerged from this study. From a psychophysical perspective, new data was obtained on the perceptual accuracy of sound localization. Also, we highlighted the plasticity and ability to learn spatial auditory cues from anatomical features different from ones' own body. These findings are of great interest to the fields of auditory rehabilitation, neurosciences in general, and the area of cochlear implants in particular.

Finally, our findings were most welcomed in the area of the audio engineering. Here, many technical advances have emerged in recent years, but the validation and understanding of how human listeners interact with these virtual sounds is still scarce. We proposed that indeed non-individualized auralized sounds provide accurate spatial cues, but that learning and adaptation processes should be considered. This finding provided new insights on why frequently the software developers felt great immersivity with their own sounds, but other users reported several intracranial auditory experiences (sound appearing to come from inside the head, not from a position in the external space), without accurate an spacialization of the sounds. Also, several technicians reported to finally have an understanding on why they systematically preferred sounds generated from their own HRTF database and failed to have satisfactory listening experiences with their colleagues' work. Above all, these findings highlighted that the validation and algorithm comparing studies should account for familiarity and listening adaptation and learning processes.

Several questions, however, remain unanswered. How and under what conditions intracranial experiences give way to immersive experiences should be carefully analyzed. Also, the evolution of the learned capabilities over time and generalization limits and the effect of training different sounds are matters of interest, which should be addressed in future studies.

3.2. The Effect of Auditory Cues on Biased Biological Motion

When we first started to address the auditory-visual interaction in the processing of biological motion, little was known about it. There was a wide body of knowledge on multisensory processing and there were several studies on biological motion perception, revealing that these biologically relevant stimuli had some specificity in their processing. Our primary goal in a first approach was to find if some interaction could be found.

In everyday life multimodal sensory information from a single event is available and redundant, and this redundancy is hypothesized to be the key to improving sensory accuracy. With this in mind, we conceived an experiment where the visual stimuli were poorly informative and prone to biased judgments, and analyzed the effect of contingent redundant auditory cues. We used PLWs in a frontal-parallel orientation, which are bistable, as they might be perceived as moving facing the viewer or facing away. These particular stimuli are highly biased as they most frequently are perceived in a frontal orientation. The scope of this study was to understand if providing additional auditory cues could reduce this bias. More specifically, we intended to analyze the impact of auditory footstep sounds over the motion direction estimates.

In the visual condition, we replicated the effect of the face forward bias. Participants responded more frequently “front” than “back” for both frontal and back-oriented stimuli. Conversely, auditory results yielded good, near-ceiling, performances. In the audiovisual condition, good discrimination levels were found. Perceptual accuracy was enhanced and the bias was largely reduced. The results in the audiovisual condition were significantly different from the reuslts in the visual condition. They were, however, not different from the auditory stimuli.

This dominance-like effect might be interpreted in several perspectives. In light of the modality appropriateness theory, one would argue that in this particular task the most appropriate sensory modality was audition, and hence an auditory dominance emerged. This strict interpretation seems to contradict

several findings revealing that in motion and spatial tasks it is the visual information that dominates the perceptual experience (e.g. Soto-Faraco et al., 2002). On the other hand, in a cue weighting perspective, it is arguable that due to its low reliability, vision was poorly weighted and that the auditory estimate gained relevance.

As this experiment was not conceived to test an integration model, and to avoid the unfashionable debates on sensory dominance effects, we chose to regard our findings in light of the disambiguating action of the auditory cues. Previous data had found that sound disambiguates visual motion direction in highly undefined presentations (e.g. Sekuler et al., 1997). Analogously, our visual stimuli were ambiguous and biased. This level of ambiguity might be a key to explaining why sound exerted such influence over the audiovisual percepts.

We argue that under undefined displays humans choose the least costly interpretations (such as a forward moving walker), but additional sensory cues take over such interpretations, allowing for the most accurate estimate to strive.

Several questions remained unanswered in this study. From our findings we were still unable to determine whether biological motion audiovisual processing differed from other motion-related multisensory processes. Also, despite finding an integration benefit in the discrimination accuracy, we were still unable to determine the quality of this integration and to approach a model to explain it. Some of these questions were addressed in the following experiments.

3.3. The Multisensory Integration of Biological Motion

In this study we intended to directly approach the multisensory integration mechanisms in biological motion. There were two experiments. A first, preliminary, experiment addressed the simultaneity judgments of auditory and visual stimuli. The main experiment focused the multisensory integration of biological motion at different walking speeds.

In the preliminary experiment we found that the audio lag of 30 ms produced the highest rates of synchronized estimates. Determining PSS values is a matter of current active debate. It has obvious potential applications to all audiovisual systems and it is useful to all practitioners and researchers who work with multisensory cues. Oddly, different values have been found under different experimental conditions and they are not the same for different stimuli (see Van Eijk et al., 2008 for a review). Also highly debated is if humans compensate or not for the slow propagation speed of sound and for the transduction latency in the brain processing of light. Our result is in line with several studies, which state that visual lead is required for the perceived audiovisual simultaneity to be reached. As both our visual and auditory stimuli were spacialized, and placed at a certain distance in space, our result is also in line with a possible processing account for the audio propagation speed.

Assessing the simultaneity threshold is also of great relevance to understanding the human tolerance for the multisensory delays. Our JND revealed a small window of ±38.2 ms. This simultaneity window might be a key factor to understand the windows of multisensory integration. To the issues of perceived asynchronies and simultaneity windows in the processing of biological motion stimuli, further, more extensive experiments are required, with the systematic manipulation of spatial cues and distances.

In the main experiment, we approached the multisensory integration of biological motion at different walking speeds. All participants discriminated more accurately the bimodal congruent than the unimodal stimuli, and their accuracy with such stimuli conformed to optimal integration mechanisms. These findings

were more robust than what was found in previous studies (Bentvelzen et al., 2009) on optimal mechanisms with speed discrimination in rigid motion.

Our data, for the first time, brings support to the argument that indeed biological motion multisensory processing might be qualitatively different from the processing of multisensory rigid motion stimuli. This interesting result should however be read with caution. A direct comparison between rigid motion and biological motion might wrongfully attribute the robustness of our findings to the high attunement to biologically related stimuli. There were however no direct comparisons with other realistic motion patterns where a semantic relation and causal effect between visual action and action-related sound were tested.

A final interesting debate from this study regards the hypothesis of a window of integration within which small stimulus discrepancies are allowed. We found, for the stimuli at the smallest level of incongruent, an unexpected discrimination slope, steeper that any of the unimodal slopes, and still close to optimal. To account for this intriguing result we proposed that there might be a window within which optimal mechanisms are found. This is a new hypothesis, as the MLE model predicts that only congruent redundant stimuli are combined in an optimal fashion. Whether such hypothesis had never been posed within these integrative studies, it had already been approached in multisensory language research (e.g. Yabe, Tervaniemi, Reinikainen, & Määtänen, 1997; Van Vassenhove et al., 2007). In that case, integrative phenomena like the McGurk-MacDonald effect had been used to study an asynchrony window where the visual and auditory syllables were merged into a new percept. Our proposed window of biological motion integration (-0/+60 ms) was much smaller than that reported for the integration of audiovisual speech (±200 ms), but well within the synchrony interval found in the preliminary experiment (-13.2/+63.2 ms).

Our findings for the first time brought new evidence on the benefit of multisensory integration with biological motion stimuli. There was a robust tendency to follow optimal mechanisms, which had not been observed with rigid motion. Even our stimuli at the smallest level of incongruence showed some integrative benefit. These results should be further compared to other

meaningful crossmodal stimuli to determine if they are indeed attributable to a biological motion processing specialization. Also, a new study with a manipulation of smaller incongruence levels would allow to more accurately define the window of biological motion integration.

IV.CONCLUDING REMARKS

The work presented in this thesis was framed within larger projects that are not limited to the work here presented and won't cease by the end of this doctoral program.

Within the Re-equipment and the Biomotion projects, the visualization system was implemented in LVP, new biological motion stimuli were created, and several psychophysical experiments were carried out. But further related work, regarding the brain processes involved in the multisensory integration of biological motion, is still currently being developed. In those experiments we expect to shed some light on the brain activity during the perception of congruent and incongruent, synchronous and asynchronous, walking stimuli.

The multisensory integration study, framed by the luso-spanish joint actions, is also currently continuing in new, more interactive work, where audiovisual human motion will be addressed in perception-and-action experiments, namely in walking synchronization experiments. Also, some related work is being developed on the audiovisual simultaneity windows.

Finally, the audio experiments opened several cooperation doors and many research interests. Framing these cooperations, a new funded project begun (Acousticave), where several psychophysical experiments will be carried out in close proximity with audio engineers. This scientific-technological link has proven very fruitful, in a loop where psychophysical advances boost engineering advances, which in turn bring new psychophysical benefit.

This thesis is also a stage mark for a personal path of professional growth. In these years, many scientific lessons were learned, of which intellectual humility stands out. Several technical competences were acquired and, above all, the ability to learn new competences will certainly remain. Finally, the social and cooperative dimension of the scientific process was a key factor in this path. Learning to communicate with technicians and colleagues from other knowledge backgrounds was an ongoing process throughout these years, and developing joint work in external collaborations also determined the researcher I came to be.

In future perspectives, all I expect is much science, many challenges, but few certainties.

REFERENCES

Affifi, A. K., & Bergman, R. A. (2005). *Functional neuroanatomy: Text and atlas*. UK: McGraw-Hill Professional.

Alais, D., & Burr, D. (2004a). No direction-specific bimodal facilitation for audiovisual motion detection. *Cognitive Brain Research, 19,* 185-194.

Alais, D, & Burr, D. (2004b). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257-262.

Alais, D., & Carlile, S. (2005). Synchronizing to real events: subjective audiovisual alignment scales with perceived auditory depth and speed of sound. *Proceedings of the National Academy Sciences, 102,* 2244-2247.

Algazi, V. R., Duda, R. O., & Thompson, D. M. (2001). The CIPIC database. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics:* New York.

Andersen, T. S., Tiippana, K., & Sams, M. (2005). Maximum likelihood integration of rapid flashes and beeps. *Neuroscience Letters, 380,* 155-160.

Arrighi, R., Alais, D., Burr, D. (2006). Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *Journal of Vision, 6,* 260-268.

Arrighi, R., Marini, F., & Burr, D. (2009). Meaningful auditory information enhances perception of visual biological motion. *Journal of Vision, 9(4),* 1-7.

Asano, F., Suzuki, Y., & Stone, T. (1990). Role of Spectral cues in median plane localization. *Journal of the Acoustical Society of America, 80,* 159-168.

Ashmead, D. H., DeFord, L., Nortington, D., & Nortington, A. (1995). Contribution of listner's approaching motion to auditory distance perception. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 239-256.

Atkinson, A. P., Dittrich, W. H., Gemmel, A. T., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception, 33,* 717-746.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience, 17*, 377-391.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America, 20,* 1391-1397.

Beardsworth, T., & Bucker, J. (1981). The ability to recognize oneself from a video recording of one's movement without seeing one's body. *Bulletin of the Psychonomic Society, 18,* 19-22.

Beauchamp, M., Lee, K., Haxby, J., & Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron, 34,* 149-159.

Begault, D. R., & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors*, *35(2),* 361-376.

Beintema, J., Olesiak, A., & Wezel, R. (2006). The influence of biological motion perception on structure-from-motion interpretations st different speeds. *Journal of Vision, 6,* 712-726.

Bentvelzen, A., Leung, J., & Alais, D. (2009). Discriminating audiovisual speed: Optimal integration defaults to probability summation when companent reliabilities diverge. *Perception, 28,* 966-987.

Bernant, R. I., & Welch, R. B. (1976). The effect of degree of visual-auditory stimulus separation and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills, 43,* 487-493.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics, 29,* 587-584.

Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlupt, P. (2005). Listening to a walking human activates the temporal biological motion area. *NeuroImage, 28,* 132-139.

Blake, R. (1993). Cats perceive biological motion. *Psychological Science, 4,* 54-57.

Blake, R., & Shiffrar, M. (2006). Perception of human motion. *Annual Review of Psychology, 58,* 12.1-12.27.

Blauert, J. (1983). *Spatial hearing* (pp 116-137, 175-177). Cambridge, MA: MIT Press.

Bonda, E., Petrides, M., Oery, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neurosciences, 16,* 3737-3744.

Botvinick, M., & Cohen, J. (1998). Rubber hands “feel” touch that eyes see. *Nature, 391,* 756.

Brooks, A., Zwan, R., Billard, A., Petreska, B., Clarke, S., & Blanke, O. (2007). Auditory motion affects biological motion processing. *Neuropsychologia*, *45*, 523-530.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex, 11,* 1110-1123.

Clarke, J. J., Yuille, A. L. (1990). *Data fusion for sensory information processing*. Boston, Kluwer Academic.

Colburn, H. S. (1973). Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination. *Journal of the Acoustical Society of America, 54,* 1458-1470.

Coleman, P. D. (1963). An analysis of cues to auditory depth perception in free space. *Psychological Bulletin, 66,* 302-315.

Cutting, J. (1978). A program to generate synthetic walkers as dynamic point-light displays. *Behavior Research Methods, Instruments, & Computers, 1,* 91-94.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9,* 353-356.

Dekeyser, M., Verfaillie, K., & Vanrie, J. (2002). Creating stimuli for the study of biological motion perception. *Behavior Research Methods, Instruments, & Computers, 34,* 375-382.

Di Luca, M., Machulla, T. K., Ernst, M. O. (2009). Recalibration of multisensory simultaneity: Cross-modal transfer coincides with change in perceptual latency. *Journal of Vision, 9,* 1-16.

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception, 22,* 15-22.

Dittrich, W., Troscianko, T., Lea, S., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception, 25,* 727-738.

Driver, J., & Spence, C. (1994). Spatial synergies between auditory and visual attention. In C. Uniltá & M. Moscovitch (Eds.), *Attention and*

performance: Conscious and nonconscious information processing (pp. 311-331). Cambridge, MA: MIT Press.

Driver, J., & Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In C. Spence, & J. Driver (Eds.), *Crossmodal space and crossmodal attention.* Oxford: Oxford University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, Chapman and Hall.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429-433.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Tends in Cognitive Science, 8,* 162-169.

Fendrich, R., & Corballis, P. M. (2001). The temporal cross-capture of audition and vision. *Perception & Psychophysics, 63,* 719-725.

Fox, R., & McDanie, C. (1982). The perception of biological motion by human infants. *Science, 218,* 486-487.

Gardner, B., & Martin, K. *HRTF Measurements of a KEMAR Dummy-Head Microphone*. url: http://sound.media.mit.edu/resources/KEMAR.html visited - June 2010.

Gebhard, J. W., & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology, 72,* 521-528.

Gilbert, C. D. (1998). Adult cortical dynamics. *Physiology Review, 78,* 467-485.

Groove, P. M., & Sakurai, K. (2009). Auditory induced bounce perception persists as the probability of motion reversal is reduced. *Perception, 38(7),* 951-965.

Grossman, E., Blake, R., & Kim, C. (2004). Learning to see biological motion: brain activity parallels behaviour. *Journal of Cognitive Neuroscience, 16,* 1667-1669.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 1,* 47-66.

Hofman, P. M., Van Ristwick, J. G. A., & Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience, 1,* 417-421.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14,* 201-211.

Kitagawa, N., & Ichihara, S. (2002). Hearing visual motion in depth. *Nature, 416,* 172-174.

Kitajima, N., & Yamashita, Y. (1999). Dynamic capture of sound motion by light stimuli moving in three-dimensional space. *Perceptual and Motor Skills*, 89, 1139-1158.

Kohler, E., Keysers, C., Umilta, M., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science, 297,* 846-848.

Loonis, J. .M., Klatzky, R. L., & and Golledge, R. G. (1999). Auditory distance perception in real, virtual and mixed environments. in Ohta, Y., & Tamura, H. (Eds.) *Mixed Reality: Merging Real and Virtual Worlds*. Tokio: Ohmsha.

López-Moliner, J., & Soto-Faraco, S. (2007). Vision affects how fast we hear sounds move*. Journal of Vision, 6,* 1-7.

Loula, F., Prasad, S., Harber, K., & Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology, 31,* 210-220.

Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology, 22*, 830-864.

Marey, E-J. (1884). Analyses cinématiques de la marche (chronophotograph). *Comptes Rendus Hebdomadaires des séances de l'Académie des Sciences, 98,* 1218-1225.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information given multiple sources of information. *Psychological Review, 97,* 229-252.

Mateeff, S., Hohnsbein, J., & Noack, T. (1985). Dynamic visual capture: Apparent auditory motion induced by dimensional space. *Perception, 14,* 721-727.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Mendonça, C. (2007). A bi-estabilidade nas representações de movimento humano a partir de estímulos visuais e auditivos. *(Master Thesis),* Braga: University of Minho.

Mendonça, C., & Santos, J. A. (2010). A bi-estabilidade nas representações de movimento humano a partir de estímulos visuais e auditivos. *Análise Psicológica, 2,* 321-331.

Mendonça, C., Santos, J. A., Campos, G., Dias, P., Vieira, J., & Ferreira, J. (2010). On the improvement of auditory accuracy with non-individualized HRTF-based sounds. *Journal of the Audio Engineering Society, Proceedings of the 129th AES Convention,* San Francisco.

Mendonça, C., Santos, J. A., & López-Moliner, J. (2011). The benefit of multisensory integration with biological motion signals. *Experimental Brain Research.* DOI: 10.1007/500221-011-2620-4.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus. I. Temporal factors. *Journal of Neuroscience, 7,* 3215-3229.

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science, 221,* 389-391.

Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology, 75,* 1843-1857.

Meyer, G. F., & Wuerger, S. M. (2001). Cross-modal integration of auditory and visual motion signals. *NeuroReport, 12,* 2557-2560.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. Sound localization by human listeners. *Annual Review of Psychology, 42,* 135-159.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., & Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology, 78,* 2226-2230.

Neri, P., Morrone, M., & Burr, D. (1998). Seeing biological motion. *Nature, 395,* 894-896.

Neuthoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecological Psychology, 13,* 87-110.

Norman, j. F., Payton, S. M., Long, J. R., & Hawes, L. M. (2004). Aging and the perception of biological motion. *Psychology and Aging, 19,* 219-225.

Parker, A., & Alais, D. (2007). A bias for looming stimuli to predominate in binocular rivalry. *Vision Research, 47,* 2661-2674.

Perrett, D., Rolls, E., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47,* 329-342.

Perrett, D., Smith, P., Mistlin, A., Chitty, A., Head, A., et al. (1985). Visual analysis of body movements by neurons in the temporal cortex of the macaque monkey: A preliminary report. *Behavioral Brain Research, 16,* 153-170.

Peuskens, H., Vanrie, J., Verfaillie, K., & Orban, G. (2005). Specificity of regions processing biological motion. *European Journal of Neuroscience, 21,* 2864-2875.

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics, 6,* 203-205.

Plenge, G. (1974). On the difference between localization and lateralization. *Journal of the Acoustical Society of America, 56,* 944-951.

Pollick, F., Kay, J., Heim, K., & Stringer, R. (2005). Gender recognition from point-light walkers. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1247-1265.

Pollick, F. E., Lestov, V., Ryu, J., & Cho, S. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research, 42,* 2345-2355.

Pollick, F. E., Patterson, H. M., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition, 82,* B51-B61.

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. *Philosophical Transactions of the Royal Society of London, 358,* 435-445.

Repp, B. (2003). Phase attraction in sensorymotor synchronization with auditory sequences: Effect of single and periodic distractors on synchronization accuracy. *Journal of Experimental Psychology, 29,* 290-309.

Repp, B., & Pennel, A. (2002). Auditory dominance in temporal processing: New evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology, 5,* 1085-1099.

Richardson, M., & Johnston, L. (2005). Person recognition from dynamic events: The kinematic specification of individual identity in walking style. *Journal of Nonverbal Behavior, 29,* 25-44.

Sanabria, D., Soto-Faraco, S., & Spence, C. (2007). Spatial attention and audiovisual interactions in apparent motion. *Journal of Experimental Psychology. Human Perception and Performance, 33,* 927-937.

Saygin, A. P., Driver, J., & de Sa, V. R. (2008). In the footsteps of biological motion multisensory perception: Judgments of audiovisual temporal relations are enhanced for upright walkers. *Psychological Science, 19,* 469-475.

Searle, C. L., Braida, L. D., Cuddy, D. R., & Pavis, M. F. (1976). Model for auditory localization. *Journal of the Acoustical Society of America*, *60,* 1164-1175.

Sekuler, R., Sekuler, A. B., Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308.

Shams, L., Ma, W. J., & Beierholm, M. (2005). Sound-induced flash illusion as an optimal percept. *NeuroReport, 16,* 1923-1927.

Shaw, B. K., McGowan, R. S., Turkey, M. T. (1991). An acoustic variable specifying time-to-contact. *Ecological Psychology*, *3,* 253-261.

Shiffrar, M., & Pinto, J. (2002). The visual analysis of bodily motion. In Prinz, W., & Hommel, B. (Eds.), Common Mechanisms in Perception and Action – Attention and Performance XIX (pp. 381-399). Oxford: Oxford University Press.

Shinn-Cunningham, B. G., Durlach, N. I., & Held, R. M. (1998). Adapting to supernormal auditory location cues. I. Bias and resolution. *Journal of the Acoustical Society of America, 103,* 3656-3666.

Simion, F., Regolin, L., & Bulf, H. (2007). A predisposition for biological motion in the new born baby. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 809-813.

Soto-Faraco, S., Lyons, J., Gazzaniga, M. S., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research, 14,* 139-146.

Soto-Faraco, S., Spence, C., & Kingstone, A. (2004). Cross-modal dynamic capture: Congruency effects in the perception of motion across sensory modalities. *Journal of Experimental Psychology, 30,* 330-345.

Stern, R. M., & Colburn, H. S. (1978). Theory of binaural interactions based on auditory-nerve data. IV. A model for subjective lateral position. *Journal of the Acoustical Society of America, 64,* 127-140.

Stern, R. M., & Colburn, H. S. (1985). Lateral-position and interaural discrimination. *Journal of the Acoustical Society of America, 77,* 753-755.

Sugita, Y., & Suzuki, Y. (2003). Implicit estimation of sound-arrival time. *Nature, 421,* 911.

Thomas, J. P., & Shiffrar, M. (2010). I can see you better if I can hear you coming: Action-consistent sounds facilitate the visual detection of human gait. *Journal of Vision, 10,* 1-11.

Troje, N. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision, 2,* 371-387.

Troje, N., Westhoff, C., & Lavrov, M. (2005). Person identification from biological motion: Effects of structural and kinematic cues. *Perception and Psychophysics, 67,* 667-675.

Valjamae, A., Larson, P., Vastfjall, D., & Kleiner, M. (2004). Auditory pressure, individualized Head-Related Transfer Function, and illusory ego-motion in virtual environments. *Proceedings of the Seventh Annual Workshop in Presence*, Spain.

Van der Zwan, R., MacHatch, C., Kozlowski, D., Troje, N. F., Blanke, O., & Brooks, A. (2009) Gender bending: Auditory cues affect visual judgments of gender in biological motion displays. *Experimental Brain Research, 198,* 373-382.

Van Eijk, R. L. J., Kohlrausch, A., Juola, J. F., & Van de Par, S. (2008). Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type. *Perception & Psychophysics, 70,* 955-968.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in audiovisual speech perception. *Neurophychologia, 45,* 598-607.

Vanrie, J., Dekeyser, M., & Verfaillie, K. (2004). Bistability and biasing effects in the perception of biological motion. *Perception, 33,* 547-560.

Vatakis, A., & Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 111,* 134-142.

Vatakis, A., & Spence, C. (2006b). Audiovisual perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters, 393,* 40-44.

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Crosssmodal effects of auditory organization on vision. *Journal of Experimental Psychology*: Human Perception ands Performance, 26, 1583-1590.

Vroomen, J., & de Gelder, B. (2003). Visual motion influences the contingent auditory motion aftereffect. *Psychological Science, 14,* 357-360.

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 513-518.

Watkins, S., Shams, L., Tanakos, S., Haynes, J., & Rees, G. (2006). Sound alters activity in human v1 in association with illusory visual perception. *Neuroimage, 31,* 1247-1256.

Welsh, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contribution of audition and vision to temporal rate perception. *Perception & Psychophysics, 39,* 294-300.

Welsh, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638-667.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized Head-Related Transfer Functions. *Journal of the Acoustical Society of America, 94,* 111-123.

Wightman, F. L., Kistler, D. J., & Perkins, M. E. (1978). A new approach to the study of human sound localization. in Yost, W., & Gourevich, G. (Eds), *Directional Hearing,* New York: Sringer-Verlag.

Wuerger, S., M., Hofbauer, M., & Meyer, G. F. (2003). The integration of auditory and visual motion signals at threshold. *Perception & Psychophysics, 65,* 1188-1196.

Wuerger, S., Meyer, G., Hofbauer, M., Zetzsche, C., & Schill, K. (2010). Motion extrapolation of auditory-visual targets. *Information Fusion. 11,* 45-50.

Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *Neuroreport, 8,* 1971-1974.

Zhou, L., Yan, J., Liu, Q., Li, H., Xie, C., Wang, Y., et al. (2007). Visual and auditory information specifying an impending collision of an approaching object. In J. Jacko (Ed.), *Human-Computer interaction, Part II* (pp.726-729). Berlin: Springer-Verlag Berlin Heidelberg.