Universidade do Minho
Escola de Psicologia

Carlos César Loureiro Silva

Perceiving Audiovisual Synchrony as a Function of Stimulus Distance

Outubro de 2011

Carlos César Loureiro Silva **Perceiving Audiovisual Synchrony as a Function of Stimulus Distance**

UMinho|2011

Universidade do Minho
Escola de Psicologia

Carlos César Loureiro Silva

Perceiving Audiovisual Synchrony as a Function of Stimulus Distance

Dissertação de Mestrado
Mestrado Integrado em Psicologia
Área de Especialização em Psicologia Experimental

Trabalho realizado sob a orientação do
Professor Doutor Jorge Manuel Ferreira Almeida Santos

Outubro de 2011

Integrated Master in Psychology of University of Minho

Specialty of Experimental Psychology

Perceiving Audiovisual Synchrony as a Function of Stimulus Distance

Carlos Silva

Jorge A. Santos

Audiovisual perception is still an intriguing phenomenon, especially when we think about the physical and neuronal differences underlying the perception of sound and light. Physically, there is a delay of ~3ms/m between the emission of a sound and its arrival to the observer, whereas speed of light makes its delay negligible. On the other hand, we know that acoustic transduction is a very fast process (~1ms) while photo-transduction is quite slow (~50 ms). Nevertheless, audio and visual stimuli that are temporally mismatched can be perceived as a coherent audiovisual stimulus, although a sound delay is often required to achieve a better perception. A Point of Subjective Simultaneity (PSS) that requires a sound delay might point both to a perceptual mechanism that compensates for physical differences or to one that compensates for the transduction differences, in the perception of audiovisual synchrony. In this study we analyze the PSS as a function of stimulus distance to understand if individuals take into account sound velocity or if they compensate for differences in transduction time when judging synchrony. Using Point Light Walkers (PLW) as visual stimuli and sound of steps as audio stimuli, we developed presentations in a virtual reality environment with several temporal alignments between sound and image (-285ms to +300ms of audio asynchrony in steps of 30 ms) at different distances from the observer (10, 15, 20, 25, 30, 25 meters) in conditions which differ in the number of depth cues. The results show a relation between PSS and stimulation distance that is congruent with differences in velocity of propagation between sound and light (Experiment 1). Therefore, it appears that perception of synchrony across several distances is made possible by the existence of a compensatory mechanism for the slower velocity of sound, relative to light. Moreover, the number and quality of depth cues appears to be of great importance in the triggering of such a compensatory mechanism (Experiment 2).

Mestrado Integrado em Psicologia da Universidade do Minho

Área de Especialização de Psicologia Experimental

Percepção de Sincronia Audiovisual em Função da Distância do Estímulo

Carlos Silva

Jorge A. Santos

A percepção audiovisual é um fenómeno curioso, especialmente quando consideramos as diferenças físicas e neuronais subjacentes à percepção do som e da luz. Fisicamente, há um atraso de cerca de 3 ms/m entre a emissão de um som e a sua chegada ao observador, enquanto a velocidade da luz torna o seu atraso negligenciável. Por outro lado, sabemos que a transdução de um estímulo sonoro é um processo muito rápido (~1 ms) enquanto que a foto-transdução é um processo relativamente lento (~50 ms). Apesar destas diferenças, sabemos que estímulos auditivos e visuais temporalmente desalinhados podem ser percebidos como um estímulo audiovisual coerente. No entanto, para que tal aconteça, um atraso do som em relação à imagem é frequentemente necessário. Um Ponto de Simultaneidade Subjectiva (PSS) que requer um atraso do som pode ser um indício da existência tanto de um mecanismo perceptual que compensa as diferenças físicas, como de um mecanismo perceptual que compense as diferenças neuronais, na percepção de sincronia audiovisual. Neste estudo analisamos o PSS em função da distância de estimulação para perceber se temos em conta a velocidade do som ou se compensamos as diferenças ao nível dos processos de transdução, quando estamos a julgar a sincronia entre um estímulo auditivo e um visual. Usando Point Light Walkers (PLW) como estímulo visual e som de passos como estímulo sonoro desenvolvemos apresentações em ambiente de realidade virtual, com diferentes alinhamentos entre som e imagem (de -285ms a +300ms, em passos de 30 ms, de assincronia do audio) e a várias distâncias do observador (10, 15, 20, 25, 30, 25 metros), em condições que variavam segundo o número de pistas de profundidade apresentadas. Os dados mostram que há uma relação positiva entre PSS e distância de estimulação congruente com as diferenças entre som e luz, ao nível da velocidade de propagação (Experiência 1). Desta forma, parece-nos que a percepção de sincronia audiovisual ao longo de várias distâncias é possível através da existência de um mecanismo de compensação para a velocidade do som, mais lenta em relação à da luz. O número e qualidade das pistas de profundidade parecem também ter uma grande importância na activação deste mecanismo de compensação (Experiência 2).

Index

Figures

Tables

Graphs

1. INTRODUCTION

1.1 – Perceiving Audiovisual Synchrony

When we think about our daily life in normal situations of stimulation, we come together in a conclusion: we perceive a great deal of our world in an audiovisual fashion. Without forgetting other types of perceptual input, we have the clear sense that a large part of the environmental events, despite being composed by visual and auditory attributes that can be perceived separately, are best understood when perceived together. We expect to hear something when a book falls off our hands onto the ground, as we can guess that something was thrown to the ground when we hear a sound identifiable as a ground impact. If suddenly some physical events, like a strong clap of our hands, were to be produced without being accompanied by sound, the world would turn seemingly odd for us. Therefore, it is easy to agree on the benefits provided by the multisensory nature of the world in our relation with the physical environment that surrounds us. Gibson (1966) himself said that perception, despite being called "visual" or "auditory" is, in its essence, multisensory (*as cit. in* Vroomen & Keetels, 2010).

Nevertheless, the phenomenon of audiovisual perception, and, in particular, audiovisual perception of synchrony, is still quite an intriguing one, mainly because of the time dimension. Although time is essential for the perception of synchrony (see, for example, the phenomenon of *unity assumption*, section **1.2 –** "Theoretical Background"), we do not have a specific sensorial channel that deals with information about time on an absolute scale. No energy carries the duration information of a stimulus and no sensorial organ can record the exact time at which a stimulus occurs in order to compare it with the recording of another stimulus of a different modality on an absolute scale. In the perception of synchrony we deal with relative timing, i.e., time differences in the perceived occurrence of one stimulus when compared with the perceived occurrence of another distinguishable stimulus. In doing so, we face intricate problems, especially when we consider the physical and neural differences underlying the perception of sound and light (Harris, Harrar, Jaekl & Kopinska, 2009; Vroomen & Keetels, 2010).

When a natural audiovisual event occurs, the visual and the auditory signals are synchronic and thus emitted at the same time. This is mainly because there is a causal relation between a physical event and the propagation of signals that can be transduced by visual and auditory sensorial organs. However, considering the differences of propagation time for light and sound (sound takes about 3 milliseconds (ms) to travel 1 meter (m); light travels approximately 299 792 m in 1 ms), it is interesting to observe that, in our daily life, the sources of audiovisual stimulation are still perceived as synchronic. If physically we can expect a difference between the arrival time of sound and image of about 3ms/m, it is not obvious that an audiovisual stimulus should be perceived as such, at least in a range of distances that are big enough to create a considerable difference between the arrival of image and that of sound to the observer.

Furthermore, a similar problem arises when we consider the neural differences. Although physically light is faster than sound, the same is not true when considering the transduction process of both signals: vision requires a transduction process considerably slow when compared to that of audition. The acoustic transduction is a direct and fast mechanic process, taking ≈ 1 ms or less (Corey & Hudspeth, 1979; King & Palmer, 1985), while the visual transduction (i.e. phototransduction) is a relatively slow photochemical process that follows several cascading neurochemical stages, taking ≈ 50 ms (Lennie, 1981; Maunsell & Gibson, 1992). Therefore, although physically light reaches the individual faster than sound, for short distances the audio signal will become perceptually available several tens of ms before the image. Again, these temporal differences turn an apparent evident phenomenon into a huge interrogation.

These physical and neural differences led some philosophers to came up with the idea of a *horizon of simultaneity* (see, e.g., Pöppel, 1988; Dennett, 1992), which was the distance of stimulation where these differences will cancel each other out – somewhere around 15 m. However, psychophysical experimentation has shown that nothing special seems to occur at this distance. We still perceive as synchronic an audiovisual event taking place closer or further this range of stimulation. Hence, the question remains: How can we perceive an audiovisual event as such, when the audio and the visual streams will physically arrive at different times to the observer? Similarly, how can we perceive synchrony in an audiovisual event when the fact is that an auditory stimulus will be transduced several tens of milliseconds before a co-ocorrent visual stimulus?

1.2 – Theoretical Background

Historically, the study of multisensory perception has always been concerned with this fundamental question: how do the sense organs cooperate in order to form a coherent depiction of the world? In the realm of synchrony perception, one widely accepted hypothesis is referred to as the *unity assumption* and uses the concept of "amodal properties": properties that are not captured by a specific sensorial organ, such as *temporal coincidence*, *spatial coincidence*, *motion vector coincidence, and causal determination*. This hypothesis states that the more information on amodal properties different signals share, the more likely it is that we perceive them as originating from a common source or object and, consequently, as synchronic (see, e. g., Vroomen & Keetels, 2010; Welch, 1999). In fact, temporal coincidence has been considered one of the most important amodal properties in the phenomenon of unity assumption. In other words, we are more likely to perceive a multisensory stimulus as such if the information from different sensorial channels reaches the brain at around the same time; otherwise, when large time differences occur, we should perceive separate sensorial events. Although this appears perfectly reasonable, this is not as straightforward as it seems, as we can be seen by the example of audiovisual perception. Considering the physical and neural differences underlying the perception of sound and light, if we only perceived synchrony in situations where there was a

temporal coincidence in the processing of these two signals, we would only perceive as synchronic events around the horizon of simultaneity, which is clearly this is not the case.

In fact, several studies have shown that audiovisual integration does not need a straight temporal alignment between the visual and the auditory stimulus. We still perceive as "in synchrony" visual and auditory stimuli that are not actually being received or even emitted at the same time (e.g., Alais & Carlile, 2005; Arrighi, Alais & Burr, 2006; Dixon & Spitz, 1980; Kopinska & Harris, 2004; Sugita & Suzuki, 2003; Van Wassenhove, Grant & Poeppel, 2007). In a seminal study that triggered this recently renewed interest in the investigation of multisensory synchrony, Dixon and Spitz (1980) had their participants watch continuous videos of an audiovisual speech or of an audiovisual object action event (consisting in a hammer repeatedly hitting a peg). While watching the video, the audio and the visual streams were gradually desynchronized at a constant rate of 51ms/s until a maximum asynchrony of 500ms. The participants were instructed to report the exact moment when they noticed for the first time that there was an audiovisual asynchrony. On average, the auditory stream had to lag 258 ms or, otherwise, lead 131ms so that the participants could detect any audiovisual asynchrony in the speech condition. In the object action condition, the auditory stream had to lag 188ms or lead 75 ms so that the participants could detect any asynchrony.

The above cited studies have shown that temporally mismatched stimuli can be perceived as synchronized and, consequently, perceived as one multisensory stimulus, but only if we keep the onset difference between sound and image within certain limits. These limits have been termed as a *window of temporal integration* (WTI) (Vroomenet & Keetels, 2010; Harris et al., 2009). In multisensory perception this phenomenon can be defined as the range of temporal differences on the onset of two or more stimuli of different modalities, where these are best perceived as a unitary multisensory stimulus.

There has been a scientific effort to define the WTI across different domains of audiovisual perception either using **complex motion stimulus** – as in speech recognition (Van Wassenhove et al. 2007), music (Vatakis & Spence, 2006), biological movement (Arrighi et al., 2006; Mendonça, Santos, & Lopez-Moliner, 2011) and object action (Dixon et al., 1980; Vatakis et al., 2006), where causal relations, expectancies and previous experiences play an important role; or **simple stationary stimulus** like flash-click studies – where there are no naturally identifiable causal relations between the two streams (Fujisaki, Shimojo, Kshino, & Nishida, 2004; Sugita et al., 2003). The size of this temporal window (see **1.4** – "Assessing the Perception of Synchrony" for details on the measure of the WTI) appears to vary as a function of the phenomenon in study. Wassenhove and collaborators found a window with a range of approximately 200 ms, from -30 ms (by convention, a negative value means an "audio lead" and a positive value an "audio lag") to +170 ms, where participants could appropriately bind sound and image to create the correct speech recognition. Arrighi et al. found a temporal window of approximately 200ms and 100ms (with no tolerance for "audio lead") depending on the velocity of the visual stimulus for the biological movement of playing drums (1 and 4 drum cycles/s, correspondingly). In a study where the participants had to judge the synchrony between a

flashing LED (light emitting diode) and a clicking sound, Lewald & Guski (2004) found different temporal windows according to the distance of the source of stimulation: 24ms for distances of 1 meter and 68ms for distances of 50 meters.

The existence of a temporal window where non-synchronic stimuli can be judged as "in synchrony" and, therefore, be taken as a unique multimodal stimulus, is quite important due to evolutionary reasons. Furthermore, this range appears to be context-dependent, thus proving to be a good solution for the perception of synchrony in different situations of stimulation, despite the physical and neural differences in the perception of light and sound. Therefore, all of the first explanatory accounts for the audiovisual perception of synchrony make use of WTIs in order to explain how we can perceive synchrony despite the physical and neural differences between sound and light. As Vroomen and Keetels (2010) pointed out, this is the most straightforward reason why, despite these differences, information from different sensorial modalities are perceived as being synchronic: **because the brain is prepared to judge as "in synchrony" two stimulations streams given a certain amount of temporal disparity.**

The scientific exploration of this phenomenon has provided us with surprising findings. A large number of studies in audiovisual temporal alignment have frequently found that we perceive stimuli from different modalities as being maximally in synchrony if the visual stimulus arrives at the observer slightly before the auditory stimulus (e. g., Alais et al., 2005; Arrighi et al., 2006; Keetels & Vroomen, 2005; Kopinska & Harris, 2004; Sugita & Suzuki, 2003). This surprising finding has been termed the *vision-first bias* (Harris et al., 2010; Vroomen & Keetels, 2010). In a work that boosted a lot of scientific discussion on the vision-first bias, Sugita and Suzuki used a *temporal-order judgment* (TOJ) task (see **1.4** – "Assessing the Perception of Synchrony") to see what was the temporal relation between the emission of a sound (consisting in a burst of white noise) and a brief light flash that provided the best sensation of audiovisual synchrony. Central to this experiment is that several distances of visual stimulation were used. Moreover, the sound was always transmitted by headphones but had to be compared with flashes of light transmitted by LEDs located at distances of 1, 5, 10, 20, 30, 40 and 50 meters. What Sugita and Suzuki reported was that the *stimulus onset asynchrony* (SOA) that provides the best perception of synchrony is always a positive one (again, a negative value means that sound leads in relation with to image and a positive value means that sound is lagging with respect to image) and, most importantly, when distance increases bigger positive SOAs are needed in order to maximize the perception of synchrony. The SOA that provides the best sensation of synchrony has been termed the *point of subjective simultaneity* (PSS) (see section **1.4** – "Assessing the Perception of Synchrony" to understand the different forms of measuring the PSS) and what the results presented by Sugita and Suzuki show is that the PSS is positively correlated with distance. This correlation can be described by an increment of about 3ms in the PSS for each one-meter increment in stimulation distance. In fact, the results that they found were roughly consistent with the velocity of sound (at least up to 20 – 30 meters of visual stimulation distance) and can be quite well predicted by a linear model

based on this physical rule. Thus, what Sugita and Suzuki suggested is that the brain probably takes sound propagation into account when judging synchrony. Therefore, they concluded we rely on information about distance of stimulation in order to *compensate* for the differences in propagation velocity between sound and light. This compensation works by judging the temporal gap that physically exists as the temporal relation that provides the best sensation of synchrony. This is why PSS gradually increases with distance of stimulation.

Other studies have also pointed to the existence of a cognitive mechanism of compensation for differences in propagation velocity between sound and light guiding the judgment of audiovisual synchrony (e.g.,Engel & Dougherty, 1971; Kopinska & Harris, 2004, Alais & Carlile, 2005). In an experiment by Kopinska and Harris, a visual stimuli consisting of a 4cm bright disc displayed on a black background was paired with a tone burst of 50ms in different temporal relations (SOAs ranging from – 200ms to 200ms in steps of 25ms) and at different distances of stimulation (1, 4, 8, 16, 24 and 32m). An important difference from this work to that of Sugita and Suzuki (2003) is that sound was presented through loudspeakers (not headphones) and at the same location of the visual stimulus. By co-locating the visual and the auditory stimulus, Kopinska and Harris found no changes on the PSS with the increment of stimulation distance. So, regardless of distance of stimulation, the PSS was always around the SOA of 0ms. However, keeping in mind that in this case there was physical propagation of sound (because stimuli were co-located), the temporal disparity of the two streams at the time of arrival to the observer judged as the most synchronic was similar to that from Sugita and Suzuki's study (see **figure 1**).

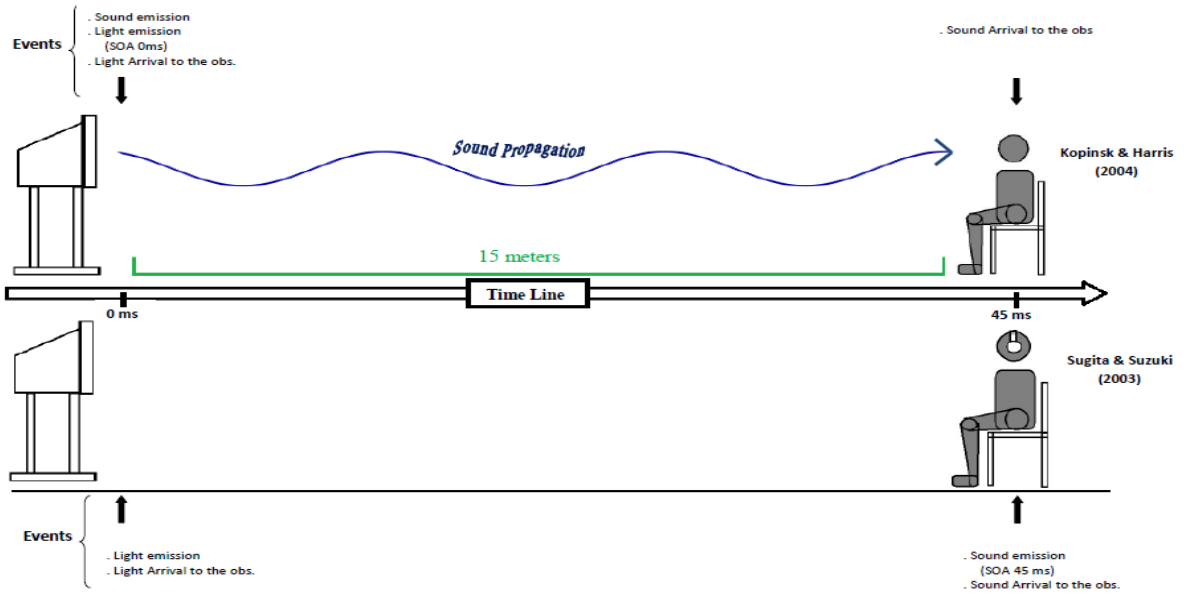


Figure 1. Depiction of the experimental differences between Sugita and Suzuki (2003) and Kopinska and Harris (2004), using as an example the stimulation distance of 15 m. For this distance the SOA perceived as the one giving the best sensation of perception is the SOA 0 ms in Kopinska and Harris (2004) and the SOA 45 ms in Sugita and Suzuki (2003), but note that the time relation between the arrival to the observer of the two stimuli is the same in both experiments: a 45 ms delay of sound.

Thus, this result, like the one from Kopinska and Harris (2004), could also be taken as evidence for the existence of a cognitive mechanism of compensation for the differences in

propagation velocity between sound and light. Both studies reach the same conclusion: a late arrival of sound to the observer provides a better sensation of synchrony than one simultaneous with light. Most importantly, this delay is distance dependent. Therefore, although the brain is prepared to judge as "in synchrony" certain amounts of temporal disparity between two stimulation streams, it also seems to have some level of discrimination to what is "more in synchrony" within the limits of WTI.

These experimental outcomes have been taken into consideration as a perceptual mechanism that helps us deal with the physical differences between the perception of sound and light (e.g. Engel et al., 1971; Sugita & Sugita, 2003; Kopinska & Harris, 2004;; Harris et al., 2009;): **we "resynchronize" the signals of an audiovisual event by shifting our PSS in the direction of the expected audio lag.**

However, while WTIs have been widely employed in explanations that try to account for the perception of audiovisual synchrony, some researchers have been reluctant in accepting a mechanism that compensates mainly for sound-transmission delays. For some researchers, a mechanism like this would be a remarkable computational feat, because it requires an "implicit knowledge" of some physical rules, namely the knowledge of sound propagation velocity (Lewald & Guski, 2004). Therefore, some argue that it is hard to accept that we perform the necessary calculations accurately when detailed information is required about both distance of stimulation and speed of sound. It has been argued that there are simpler mechanisms that can account for the perception of audiovisual synchrony and so, in accordance with the law of parsimony, we should adopt those as an explanation (Arnold, Johnston, & Nishida, 2005). Moreover, other studies have failed to demonstrate compensation for differences in propagation velocity (Arnold et al., 2005; Lewald & Guski, 2004; Stone et al., 2001). Lewald and Guski tried to replicate the findings of Sugita and Suzuki (2003) in a less artificial stimulus situation where they used the same kind of stimuli (sound bursts and LED flashes) but co-located (as in Kopinska and Harris, 2004) and with the experiment being performed in open field. Here, they found no compensation for distance. In fact, the PSS shifted according to the variation in sound arrival time, but in the direction of a sound lead and not of a sound lag as in Sugita and Suzuki (2003). So, in this case, participants had the best perception of synchrony when the auditory and visual signals were synchronic in their arrival at the observer's sensorial receptors. As Lewald and Guski pointed out, "this conclusion is in diametral opposition to the study of Sugita and Suzuki (2003)" (p. 121), and this discrepancy might be due to problematic procedures used in the former study. According to Lewald and Guski, there are two main problems in Sugita and Suzuki's study and both are related with the stimuli themselves:

1 – The sound stimuli were *not co-located* with the visual stimuli and consequently, there was no auditory distance information in the experiment;

2 – In the visual stimuli, because the perceived light intensity decreases with distance, the authors increased its luminance to compensate for this attenuation. However, by doing this they kept the perceived stimuli's luminance constant thus providing an incongruent cue with the distance increasing information. Moreover, parameters as size and contrast, extremely relevant in the

perception of distance, were not kept constant but could have been affected by this luminance manipulation.

According to Lewald and Guski (2004), these two stimulus manipulations made the design of the experiment inconsistent with everyday life. Thus, relying on the results of Lewald and Guski, if we keep natural all the physical changes in the stimuli caused by the increase of distance, we should not get evidence of compensation for the sound's slower time of propagation in the perception of synchrony. These authors refute the hypothesis that there might be an "implicit estimation" of sound-arrival time guiding the perception of audiovisual synchrony and, instead, advocate that we only use temporal windows of integration to deal with crossmodal temporal disparities.

Arnold et al. (2005) reached similar results to Lewald and Guski (2004) and also highlight the same problems in the work of Sugita and Suzuki (2003). However, both these investigations also present potential problems in the simulation of distance:

1 - The conduction of the experiment in open-field, as in the case of Lewald and Guski (2004), fails to provide the optimal conditions for auditory distance information. One of the most powerful depth cues in auditory perception is the ratio of the energies of direct and reflected sound (Bronkhorst and Houtgast, 1999). In open-field sound reflects only once (in the ground) and, as we can see by the work of Bronkhorst and Houtgast, we need around three to nine reflections to accurately perceive sound distance. Also, the most important distance cue in a situation of open-field – loudness – is frequently erroneously perceived as the level of the sound itself, which can cause misjudgments of stimulation distance;

2 – Arnold et al. (2005) manipulated the angular size and velocity (i. e. retinal size and velocity) of their visual stimuli to ensure that the size and velocity of the stimuli appeared constant while distance increased. This is pretty much the same problem pointed in the critique of Lewald and Guski (2004) to the work of Sugita and Suzuki (2003).

The work of Kopinska and Harris (2004) cited before seems to be free of all these problems: the sound is co-located and presented in a large corridor, so there are many sound reflections and, consequently, strong cues of auditory distance. The visual stimuli were not manipulated to keep some features constant across distances; and yet, in the end, the result clearly supports the existence of a mechanism of compensation for stimulation distance in the perception of audiovisual synchrony. Still, some authors have criticized their experimental design. Vroomen and Keetels (2010) pointed to the fact that in the study by Kopinska and Harris the distance of presentation was not randomized on a trial-by-trial basis, but instead was blocked by session. In an experimental situation where distance is blocked by session, we are being exposed for a great deal of time – 1 hour in the case of Kopinska and Harris – to quite specific temporal disparities and, mostly because one of the stimuli is constant, this became a prone situation for the observation of a *temporal recalibration* phenomenon. This could lead to a reduction of the effect of distance in the PSS and thus no shift in the PSS is observed, misleading one to conclude that distance is being compensated for (in studies where both stimulus are co-located).

Temporal recalibration has generally been taken into account as another possibility for explaining how the brain deals with crossmodal temporal differences in the perception of synchrony (Fujisaki et al., 2004; Navarra, Soto-Faraco, & Spence, 2007; Vroomen & Keetels, 2010; Harris et al., 2010). The term *recalibration* became well-known in the psychophysical scientific literature due to the work of Helmotz (1867) that showed a remarkable ease in the adaptation of the visuo-motor system to shifts on the visual field induced by wedge prisms. As noted by Vroomen and Keetels (2010): "Recalibration is thought to be driven by a tendency of the brain to *minimize discrepancies among the senses* about objects or events that normally belong together" (p. 878). So, by itself, this definition seems to suit up well to an explanation for the perception of audiovisual synchrony. But how does this temporal recalibration really works in the audiovisual perception? It is well documented and commonly accepted that the least reliable source of stimulation is adjusted towards the most reliable one (e. g. Ernst & Bulthoff, 2004; Di Luca, Machulla, Ernst, 2009). According to this, what may be happening in the audiovisual perception of synchrony is that one modality of the audiovisual stimulus (usually the sound) is being adjusted towards the other, because the visual one gives us more reliable information about localization or time of occurrence. Researchers have hypothesised that this adjustment can be made in one of three ways (Vroomen at al., 2010): (a) by adjusting the criterion of synchrony in order to be more willing to judge as "in synchrony" a certain temporal disparity to which we are constantly exposed to; (b) by adjusting the sensory threshold of one of the two modalities, delaying it or accelerating it in order to compensate for total temporal differences between the arrival to receptors and processing of the two modalities of stimulation; (c) by widening the WTI until the temporal disparity that we are being exposed to falls within the threshold for synchrony judgment.

Empirical support for these adjustments came from studies using the "exposure-test paradigm", where participants are exposed to a constant audiovisual stimulus with a certain temporal disparity between signals and then perform a typical task to evaluate the PSS (a TOJ or a *simultaneity judgment* (SJ) task; see section **1.4** – "Assessing the Perception of Synchrony"). Fujisaki et al. (2004) found that after an exposure phase where the participants were repeatedly presented with a tone and a flash separated by a fixed time lag, the PSS shifted in the direction of the exposed audiovisual lag. On average, the PSS without exposure phase (baseline) was -10 ms, but it changed to -32 ms after exposure to a -235 ms delay (audio lead) and to +27 ms after exposure to a +235 ms delay (audio lag). So, they found a total effect of recalibration of 59 ms. For Fujisaki and collaborators, this evidence of temporal recalibration has a clear adaptive value, because it might contribute, for example, for the human brain to compensate for the processing delays as in the case of a slower visual processing. Indeed, some studies have reported a decrease in reaction time to visual stimuli after exposure to an asynchronous audiovisual stimulus (Di Luca et al., 2007; Welch, 1999). Therefore, besides being another perceptual mechanism playing an important role in minimizing temporal discrepancies in an audiovisual scene, temporal recalibration may also be the mechanism responsible for compensating for the transduction time differences in the perception of sound and light. In fact, there is a persistent lack

of evidence for compensation of transduction time differences in studies manipulating the distance of stimulation. This could be due to a long history of exposure to this "veridical" neural lag and a consequent adaptation or temporal recalibration (Fujisaki et al. 2004). Because the transduction time difference between modalities is quite stable, a long process of temporal recalibration troughout lifetime might have canceled this difference.

On the whole, it can be that temporal recalibration acts as a mechanism that **corrects for timing differences by adapting the intersensory asynchrony via (a) adjustment of the criterion, (b) adjustment of the sensory threshold, (c) widening of the WTI.**

We still do not know precisely how this mechanism is related with the mechanism of compensation for propagation velocity differences, but one strong hypothesis is that temporal recalibration acts in situations where we should assume unity between two sensorial inputs despite the existence of a constant temporal disparity between them. In these cases, what probably happens is that temporal recalibration corrects for that difference, making way to the proper functioning of other mechanisms, such as compensation for distance.

As Fujisaki et al. (2004) pointed out, we should be especially careful in studies manipulating spatial localization, because recalibration could occur if we expose the participant for several minutes to a temporally unaligned audiovisual stimuli, where certain features (such as localization) are constants. This could lead us to erroneously conclude that the results are due to a mechanism of compensation other than temporal recalibration. However, even if temporal recalibration occur in the study of Kopinska and Harris (2004), the effect of temporal recalibration reported in several studies is not sufficient to account for the temporal relation changes in the audiovisual stimulus across distances. In Kopinska and Harris, from the nearest to the farthest condition, the sound arrives to the observer with a difference of 96ms. Some of the values for a total of recalibration found in the literature are: 59 ms for a stationary event and 48 ms for a motion stimulus – Fujisaki et al (2004); 27 ms in a SJ task and 18 ms in a TOJ task – Vroomen and Keetels (2009); 26 ms – Di Luca et al. (2007). Clearly there is something more going on than just temporal recalibration in the perception of audiovisual synchrony.

Another phenomenon that could help dealing with crossmodal temporal discrepancies is a long known phenomenon in psycophysics named the *ventriloquist illusion* (see, Bertelson, 1999; for a review). In *spatial ventriloquism* one can perceive a sound as coming from a visual stimulus when in fact they are not co-localized. We can think of it as the sound "being pulled" in the direction of the image in terms of perceived localization. But what some researchers regard as contributing for the perception of audiovisual synchrony is the *temporal ventriloquism* (Vroomen & Keetels, 2010; Harris et al., 2009). In this case, it is the visual stimulus that is "being pulled" to the auditory one. Thus, the illusion happens on the temporal dimension and not on the spatial one. This effect can be demonstrated using a quite famous perceptive effect: the flash-lag effect (FLE) (see, for example, Baldo & Caticha, 2005; Vroomen & Gelder, 2004). In the FLE, the stimulus is composed by a moving dot (although it could be almost any sort of visual stimulus) and a flashing one that appears only once.

At the time of its appearance the flashing dot is perceived as lagging behind the moving one, even though they are exactly at the same position when presented. Using this effect, Vroomen and Gelder (2004) collected evidence for the temporal ventriloquism illusion by introducing a clicking sound slightly before or after the flashing dot (they used intervals of 0, 33, 66 and 100 ms). What they found was that the presentation of the sound had the effect of anticipating or delaying in ~5% the perception of the flashing dot's occurrence time, whether the sound was presented before or after the flash, respectively. So, a sound presented 100 ms before the flashing dot would make it appear to be ~ 5 ms earlier. A sound presented 100 ms after the flashing dot would make it appear to be ~5 ms later. Similar studies reached the same conclusion (see, Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Scheier, Nijhawan, & Shimojo, 1999; Vroomen and Keetels, 2009): vision appears to be flexible in the time dimension and, consequently, a sound event presented slightly after or before (provided that the lag between them is below ~200 ms) can cause an illusion of temporal ventriloquism. This temporal ventriloquism can also be taken as a mechanism that contributes for the audiovisual perception of synchrony (Vroomen & Keetels, 2010; Harris et al., 2010) by **shifting the perceived visual onset time towards audition.**

Despite the importance of being aware of the contribution of this phenomenon for the audiovisual perception of synchrony, the quantitative changes in temporal perception that we find in these kind of studies are not sufficient to explain major changes in the PSS as in the study by Sugita and Suzuki (2003) or a constancy in the synchrony perception with an increase in 96 ms on the difference of arrival between the two streams to the observer, as in the study by Kopinska and Harris (2004). So, again, we stress that something more, which apparently is affect by distance manipulation, is involved in the perception of synchrony.

Considering the questions formulated in the end of **1.1** – "Perceiving Audiovisual Synchrony": How can we perceive an audiovisual event as such, if the audio and the visual signals physically arrive at different times to the observer? And how can we perceive synchrony in an audiovisual event when the fact is that an auditory stimulus will be transduced several tens of milliseconds before a co-ocorrent visual stimulus? We can now draw at least an outline of what should be the answer. Through the phenomenon of WTIs we show that temporally mismatched stimuli can still be perceived as synchronous. Furthermore, we might have three mechanisms to explain how inter-modality lags can be dealt within a certain context:

. Compensation for stimulation distance;

. Temporal recalibration;

. Temporal ventriloquism.

We still do not know how some of these mechanisms interact or, moreover, if some of them are really involved in the audiovisual perception of synchrony, as in the case of an active compensation for distance. However, if we want to give a satisfactory theoretical account of how

synchrony constancy is maintained, we need to have experimental evidence for each one of these three mechanisms, and especially for a mechanism that compensates for the differences in propagation velocity between sound and light. A mechanism like this is central to any attempt to explain the audiovisual perception of synchrony, because it is the only one that can account for large temporal disparities as the ones occurring when an audiovisual stimulus is being transmitted at large distances. Consider an audiovisual source located at 35 m from the observer: in this case there will be a delay of ≈105 ms between the arrival of the visual and the auditory stimulus to the observer due to physical differences. As we could see, none of the changes in the PSS due to temporal recalibration (and this mechanism requires a long exposure to the stimuli) or temporal ventriloquism could account for such a delay. Furthermore, most of the reported windows of temporal recalibration, if conceived as being static and with their center around the point of real synchrony, are not big enough to accommodate such values. So, if we want to give a satisfactory explanation on how synchrony perception occurs in a wide range of situations, we have to clarify and look for evidences of a mechanism that compensates for differences in propagation velocity in the perception of audiovisual synchrony.

1.3– Goals and Hipotheses of the Study

With this study we intended to contribute for the intense scientific discussion on the existence of a mechanism that compensates for the physical differences in the perception of audiovisual synchrony, by clarifying the relation between distance and the perception of synchrony. Also, we want to take in consideration some of the central critiques indicated to other studies that provided evidence for the existence of such a compensatory mechanism.

For doing so, we used a highly ecological stimulus that represents a frequent situation of audiovisual events in everyday life. This audiovisual stimulus is composed by a Point-light Walker (PLW) of 13 dots walking on a front-parallel plane to the observer and by the sound of its steps. The PLW stimulus became popular due to the findings of Johansson (1973). What this author discovered was that, even when all the information about biological movement is reduced to a set of bright moving spots placed on the major human joints, the observer still has a vivid and apparently effortless perception of a moving person. However, when this set of bright spots is static, it loses all its representativeness of a human body. We use this type of stimulus because it allows us to isolate the information about human motion patterns (relevant to our situation of synchrony judgment), separating it from pictorial and formal patterns of a human body (irrelevant to our situation of synchrony of judgment) and thus creating a stimulus that is both highly representative of the real human movement and easy to manipulate (see **figure 2**).

By using this type of audiovisual stimulus we account for a recurrent critique: studies defending the evidence of compensation for distance of stimulation only use simple and stationary depictions rarely present in everyday life and frequently without a causal relation underlying the visual and the auditory signals (Lewald & Guski, 2004; Vroomen & Keetels, 2010, Harris et al, 2009). When

more complex stimuli are used, like those with movement and those involving causal relations, it is difficult to observe evidence of compensation for distance of stimulation (Arnold et al, 2005; Harris et al, 2009). Furthermore, by using motion stimuli causally related, we also took in consideration recent findings showing that biological motion stimuli are preferred when compared with constant velocity motion stimuli, in the study of audiovisual synchrony (Arrighi et al., 2006). In a study where the participants had established a baseline PSS to an audiovisual stimulus consisting in footages of a professional drummer playing a conga drum, Arrighi and collaborators found that the PSS was not different when the visual stimulus was a computerized abstraction of the drummer but with the same movement (biological movement) presented in the footages. However, when the same artificial visual stimulus had an artificial movement (constant velocity movement) the PSS was significantly different, even when the frequency was the same as in both conditions.

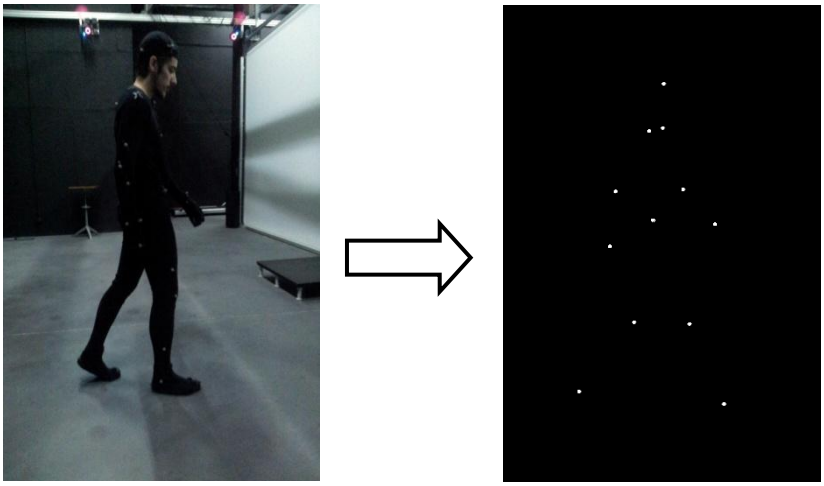


Figure 2, On the left is a frame of a participant in a session of biological motion capture. The markers that we see placed over the suit in some body joints are designed to reflect signals transmitted by the 6 cameras. In this way we can record in real time the position (along 3 axes) of each marker in order to design an animated representation of the human body movement (image on the right).

Another advantage of these stimuli is that they allow us to easily eliminate attributes that can be viewed as redundant and also to have a high control of parameters that are critical for the perception of distance – such as angular size, angular velocity, elevation, intensity, contrast, and contextual cues. Thus, in order to assess the existence of a mechanism that compensates for the differences in propagation velocity, we presented an audiovisual stimulus in a virtual reality (VR) environment that emulates the real physical world in such a way that we could simulate several distances by manipulating some depth cues (see **2.1** and **3.1** – "Method"). By doing this we were able to check how PSS changes when simulated distance increases. In experiment 2, we tried to assess the importance of some depth cues in the functioning of a possible mechanism that compensates for differences in propagation velocity. We did so by manipulating the number of depth cues presented in the visual stimulus (see **3.1** – Method).

In short, we will be able to provide support to the argument of compensation for differences in propagation velocity as one of the mechanisms that makes synchrony perception possible if:

1- We find a positive relation (similar to the rule of physical propagation of sound) between the shift of the PSS (towards an "audio lag") and the simulated distance of the visual stimulus;
2- This relation is dependent on the number and quality of the depth cues.

However, before going into experimental details, we want to provide some insight on how to measure synchrony.

1.4 – Assessing the Perception of Synchrony

The scientific literature has provided us with two methods to assess the perception of synchrony: *Simultaneity Judgements* (SJ) and *Temporal-Order Judgment* (TOJ) tasks. These methods have been used for a long time and in a wide range of studies involving all kinds of sensory stimulation.

In an SJ task the participant is confronted with a forced-choice decision between the options "synchronous" or "not synchronous" when two stimuli are presented in a certain temporal relation. Typically, these outputs are reported as a frequency distribution of the "synchronic" answer plotted as a function of the SOA between the two stimuli. This distribution usually follows a bell-shaped Gaussian curve (see **figure 3**) where the peak of the distribution is taken as the PSS and so the correspondent SOA is interpreted as the temporal relation between the stimuli that elicits the best perception of synchrony. The standard deviation (σ +/- 34%) is often used as an approximation to the WTI, because it can be interpreted as the range of SOAs at which the observer judges the two stimuli streams more often as being synchronic. Nevertheless, there is still some variability in the way in which WTI is calculated in studies using SJ task. For most of the authors the WTI is equivalent to the *just noticeable difference* (JND), calculated as the average interval (between audio-first SOAs and visual-first SOAs) at which the participant responds with 75% "synchronous" answer (Vroomen & Keetels, 2010; Lewald & Guski, 2004 for an analogue procedure but in a TOJ task).

In a TOJ task the participant is confronted with a forced-choice decision between the options of "auditory stimulus-first" and "visual stimulus-first". Typically, when we plot the frequency of "visual stimulus-first" as a function of the SOA we get a Cumulative Gaussian function curve (see **figure 3**), which means that in virtually none of the trials with highly negative SOAs thus the observer perceive the visual stimulus first; inversely, in virtually all the trials with highly positive SOAs the observer perceives the visual stimulus as appearing first. Uncertainty lies somewhere closer to the center of the SOAs spectrum and maximum uncertainty is where the SOA corresponds to the 50% crossover point (i. e. the point at which the proportion of "auditory stimulus-first" equals the proportion of "visual stimulus-first"). From this we infer the PSS in a reasoning of the type: you cannot tell which stimulus appears first because you perceive them as simultaneous. The WTI is also matched with a 75% JND in this type of task and its value can easily be defined by taking half of the temporal interval that goes from the SOA that elicited 25% of synchronous answers and the SOA that elicited 75% of synchronous answers.

Although these two tasks look just like different means of achieving the same results, we have to be very cautious about this interpretation. In a study that explores the problem of which task to use in order to measure synchrony, Eijk, Kohlrausch, Juola, and Van de Par (2008), looked at PSS and WTI results of several studies in audiovisual perception and pointed some interesting differences between studies using each task. What seems to be the most striking difference is that when we look for studies that present negative PSSs as an overall experimental outcome (i.e. The PSS is located at a SOA where sound is appearing first than image, contradicting the vision-first bias and the physical rules under natural conditions of stimulation) the TOJ task was found to be the most frequently used. Moreover, studies using an SJ task seem to yield similar results between them, even when this task involves more than two categories as an a SJ2 task (where the observer can answer "audio first", "synchronous", or "video first") and independently of the complexity of the stimulus. On the other hand, TOJ tasks yield more inconsistent results and, more importantly, seem to yield different results according to the stimulus complexity.

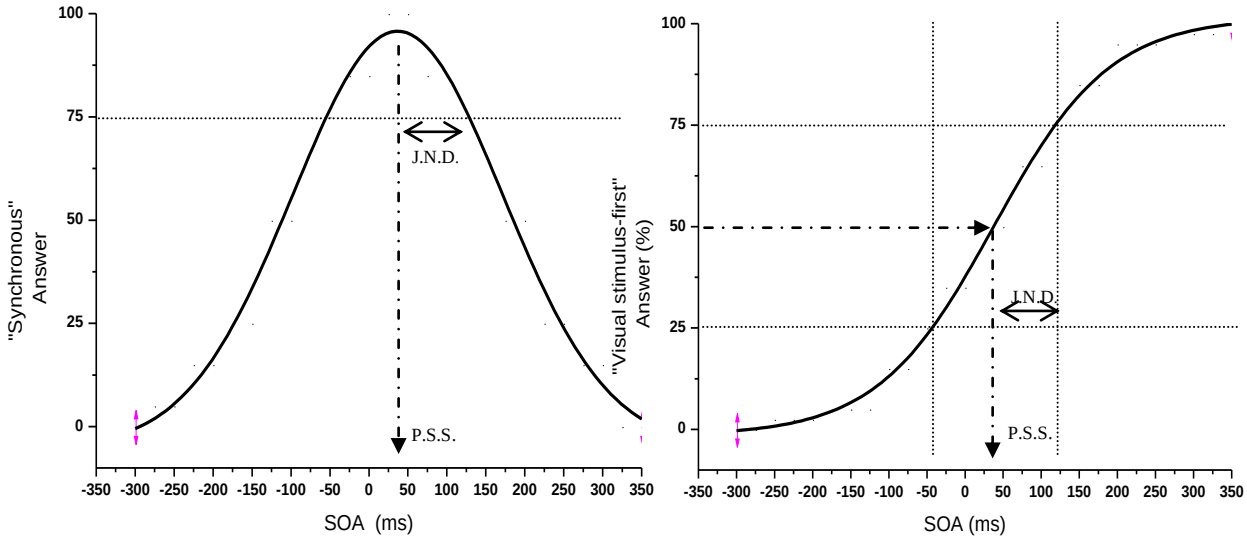


Figure 3, graphical representation of the PSS and the JND as measured an a SJ task (left) and in a TOJ task (right).

Hence, several researchers argued that TOJ and SJ might measure different things. Alan (1975) stated that SJ tasks are emphasizing the judgment of "synchrony" versus "successiveness", while TOJ tasks are emphasizing the judgment of "order", which in turn requires the perception of successiveness. Considering this, and as Eijk et al. (2008) point out, the PSS derived of a TOJ task is more vulnerable to the individual strategy because, at least within the limits of the WTI, we should not perceive successiveness and therefore we could not perform a reliable order judgment. In fact, Eijk and collaborators concluded that the PSS derived from a TOJ task (but not from an SJ task) varies greatly according to the participant and it can occur anywhere in the range of the WTI. Also, as pointed by Harris et al. (2009), the decisions in the TOJ task are "based on the remembered temporal sequence and [are] therefore vulnerable to postperceptual biases." (pg. 237) Conversely, the SJ task yields a more direct measure of perceived synchrony (Zampini, Guest, Shore, & Spence, 2005).

Finally, and considering all this information, we should prefer the SJ task when the goal is the measurement of the PSS in audiovisual perception; therefore, this was our choice in the experiments presented bellow.

2. EXPERIMENT 1: Searching for evidences of compensation for differences in propagation velocity

2.1 – Method

2.1.1 – *Participants*

Six participants, aged 20-28 years old, all underwent visual and auditory standard screening tests and had normal hearing and normal, or corrected to normal, vision. All of the participants were university students and all gave informed consent to participate in the study.

2.1.2 – *Stimuli and Materials*

The experimental tasks were performed in a darkened room located at the Centro de Computação Gráfica (CCG) in the University of Minho, especially designed to create an immersive sensation along with the projection of a VR environment.

For the projection of a VR environment we used a cluster of 3 PCs with a NVDIA® Quadro FX 4500 graphic boards, that works with a custom software of projection running on top of OpenGL and using VR/Juggler as a "virtual platform" (an interface of communication between the hardware of a VR system – for example, a head-mounted display or a projection based system – and the software used in the design of the virtual world). Each of the PCs forming the cluster was connected to one image channel using 3chip DLP projectors Christie Mirage S+4K with a resolution of 1400x1050 pixels and a frame rate of 60Hz per channel. Only one projector was used for the projection of the visual stimulus and the area of projection was the central area with 2.80 m high by 2.10 m wide of a PowerWall projection surface of 3 m x 9 m. The auditory stimulus was projected through a set of Etymotics ER-4B in-ear phones processed by a computer with a Realtec Intel 8280 IBA sound card. Due to hardware processing time, all the latencies in the visual and auditory channels were measured (using a custom-built latency analyzer – consisting in an Arm7 microprocessor coupled with light and sound sensors) and corrected by adjusting the SOAs to the system latency.

2.1.2.1 - Visual Stimuli

The visual stimuli were PLWs composed of 13 white dots (see **fig. 2**), generated in the Laboratory of Visualization and Perception (University of Minho) using a Vicon® motion capture system with 6 cameras MX F20 and a set of custom LabVIEW implemented routines.

All stimuli corresponded to the correct motion coordinates of a 1.87 meters high, 17 year-old male, walking in a front-parallel plane to the observer, at a velocity of 1.5 m/s and with an average

interval between steps of 530 ms. The duration of each stimulus was 1.8s and so the participant was presented with three visual steps during this time (at 412 ms, 912 ms and 1458 ms; see **fig. 5**).

In order to simulate six distances of stimulation (10, 15, 20, 25, 30 and 35 meters), we changed the visual angular size and angular velocity of the stimuli according to the expected changes in the real physical world (see **fig. 4**).

Figure 4. Geometrical representation of how the angular size of an object decreases with distance. The same geometrical relation is the cause of angular velocity decrease with distance.

We these changes by applying the following geometrical rules:

(a) For the angular sizes, $\alpha = [\text{Tan}^{-1} (s/d)] \times 2$; where α is the angular size in a certain distance of stimulation *d* in cm for a PLW, in which *s* equals half its size.

(b) For the angular velocities we had to first calculate the complete angular path $\Theta = [\text{Tan}^{-1} (p/d)] \times 2$; where Θ is the angular path for a certain distance of stimulation *d* for a PLW, in which *p* equals half of the total distance walked (135 cm). We get the angular velocity by dividing Θ by the time of presentation (1.8 seconds in this case).

Thus, in order to simulate a distance of stimulation of 10 meters we had to project a PLW that in reality has 1.74 meters and walks 270 cm (a walking speed of 1.5 m/s during 1.8 seconds), with an angular size of about 10º and an angular velocity of ≈8.5 deg/s. Therefore, as the surface of projection was 4 meters from the observer, in order to get these angular values and thus get a virtual distance of 10 meters, the projected PLW had to have a metric value of 70 cm high and a walking path of about 104 cm. In the same rationale, to simulate a PLW at 35 meters from the observer we had to generate a PLW with an angular size of about 2.8º and an angular velocity of about 2.4 deg/s. Therefore, we had to project a PLW with the metric values of 20 cm high and a walking path of about 15 cm.

The PLW was composed by white dots (54 cd/m^2) that moved against a black background (0.4 cd/m^2). In order to create a good immersive sensation there was no illumination in the room except from the one coming from the projection itself.

Using this type of visual stimuli, we made available only two important pictorial depth cues: *familiar size* and *elevation*. Familiar size depth cue is of great importance in the judgment of distance and works by combining the previous knowledge that we have about the size of a certain object with the decreasing, with distance, of the visual angle from that object's projection in the retina. Elevation depth cue, also known as "relative height" or "height in the plane", works by locating vertically higher, in the visual field, objects that are farther away. To clearly understand this depth cue you just

have to notice how the horizon line is seen as vertically higher than the foreground. Furthermore, the visual stimuli also present, inherently, two dynamic depth cues: the *amplitude of the step* (wider steps represent a closer presentation) and the *angular velocity* (a smaller angular velocity translates into a further distance of presentation).

With a PLW of this kind we get all these depth cues working just by decreasing the angular size and velocity, as well as by decreasing the angular size of the dots composing it, and by gradually increasing its elevation according to distance of stimulation.

2.1.2.2 - Auditory Stimuli

The auditory stimuli were step sounds from the database of controlled recordings from the College of Charlston (Marcell, Borella, Greene, Kerr, & Rogers, 2000). They correspond to the sound of a male walking over a wooden floor and taking three steps at exactly the same velocity as the visual stimuli. So, there was, as in the visual stimuli, an interval between steps of 530ms (see **fig. 5**). These sounds were auralized as free-field (with no distance information from reverberations) by a MATLAB routine with head related transfer functions (HRTFs) from the MIT database (http://sound.media.mit.edu/resources/KEMAR.html). With this auralization process we matched the angular velocities of the auditory stimuli to those of the visual stimuli. Thus, in all the audiovisual stimuli the sound moves in the same direction and with the same velocity of the visual stimulus (movement in the front-parallel plane to observer). Nonetheless, this MATLAB routine did not allow us to manipulate the localization of the auditory stimuli in a front-perpendicular plane to observer. In other words, although we provided accurate auditory information about direction and velocity, we did not provide important information about distance of stimulation.

2.1.2.3 – Visual and Auditory Stimuli Relation

In order to present several audiovisual stimuli, we put together the visual and the auditory stimulus in 19 different time relations or, in other words, 19 different SOAs, with specific spectrums of SOAs for even and odd distances. The SOAs took the following values:

. For a PLW at a distance of 10, 20, or 30 meters: -240 ms; -210 ms; -180 ms; -150 ms; -120 ms; -90 ms; -60 ms; -30 ms; 0 ms; 30 ms; 60 ms; 90 ms; 120 ms; 150 ms; 180 ms; 210 ms; 240 ms; 270 ms; 300 ms;

. For a PLW at a distance of 15, 25, or 35 meters: -225 ms; -105 ms; -165 ms; -135ms; -105 ms; -75 ms; -45 ms; -15 ms; 0ms; 15 ms; 45 ms; 75 ms; 105 ms; 135 ms; 165 ms; 195 ms; 225 ms; 255 ms; 285 ms.

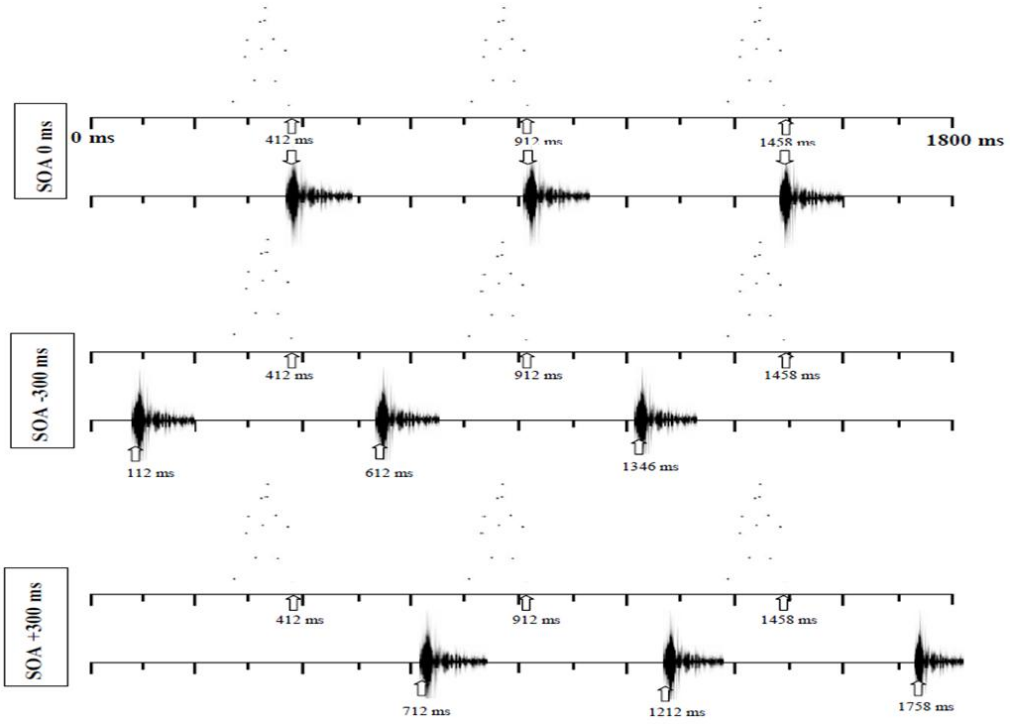


Figure 5, examples of audiovisual stimuli where we can see when steps occur. The first step occurred 412 ms after the start of the trial, the second step at the 912th ms and the third step occurred at the 1458th ms. Also, we can see the temporal relations between the occurrence of the visual step and the occurrence of the auditory step, in the SOAs 0 ms, -300 ms, and +300 ms.

Remember that negative values indicate that sound is being presented first and positive values mean sound being presented after the visual stimulus (see **fig.5**). The reasons why we had a different spectrum of SOAs for even and odds distances were essentially two: Firstly, we wanted to always present the theoretical values that, according to the compensation for sound velocity hypothesis, would provide the best sensation of audiovisual synchrony (for e.g., 30 ms for stimulation at 10 meters from the observer, 45 ms at 15 meters, and so on). Secondly, we wanted to keep the experimental sessions within a reasonable duration, in order to avoid situations of fatigue. If we intended to have just one spectrum of SOAs that accounted for all the theoretical values we would need a spectrum of 38 SOAs – if we wanted to keep a constant time difference between them – and we would have to match them with all the distances of stimulation. This would make sessions extremely and unnecessarily long. Therefore, we had two groups of 19 SOAs presented, each one presented at 3 distances, comprising a total of 114 different audiovisual stimuli.

2.1.3 – Procedure

The experimental sessions were blocked by distance of stimulation so there were 6 experimental sessions for each participant completed in a pseudo-random order (see **table 1**). In each session 19 different audiovisual stimuli were presented, corresponding to the different SOAs at that specific distance of stimulation. All stimuli were randomly presented with 40 repetitions each and with an inter-stimulus interval of 1.6 seconds. Each session took about 43 minutes to complete and a break,

no longer than 3 minutes, was always taken in the middle of the session. It took about 4 hours for each participant to complete the experiment.

Participant	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
1	20	25	30	35	10	15
2	15	20	25	30	35	10
3	10	15	20	25	30	35
4	35	10	15	20	25	30
5	30	35	10	15	20	25
6	25	30	35	10	15	20

Table 1. Presentation order of each stimulation distance (in meters) for each participant.

Before each experimental session the participants were shown 10 repetitions of an audiovisual stimulus in which the sound appeared with a 300 ms lead, and, another 10 repetitions of an audiovisual stimulus in which the sound appeared with a 330 ms lag. This preliminary session was taken in order to check if participants were able to perceive any kind of asynchrony. Note that none of the SOAs used in this preliminary session were then used in the experimental session.

At the beginning of the experimental session, the following instructions were given: "You will participate in an audiovisual perception study in which you will be presented with several audiovisual scenes of a PLW walking at a certain distance. I want you to pay close attention to the audiovisual scene, because you will have to judge its audiovisual synchrony during the intervals between scenes. Thus, after each scene, if you think that the auditory and the visual streams were synchronized click the right button; otherwise, if you think that the auditory and the visual streams were not synchronized click the left button" (SJ task).

The participant was seated in a chair 4 m from the screen and in line with the center of the projection area. In each scene the participants were presented with a PLW walking from left to right and taking three steps at a velocity of 1.5 m/s, while listening trough in-ear phones to three steps with the same angular velocity of the visual stimuli and in a given temporal relation with the visual stimulus. Thus, participants had three moments where they could judge the audiovisual synchrony of the stimulus. After the presentation of each audiovisual stimulus and during the inter-stimulus interval, the participant had to answer in a two key mouse according to the instructions.

2.2 – Results

Among the participants of this experiment, three had some background knowledge about the thematic of the study and the remaining three were naive to the purpose of the experiment. Nevertheless, according to a t-test for independent samples there is no significant difference between these groups with regard to values of PSS ($t(33) = -0.81$, n. s.) and WTI ($t(33) = -0.95$, n. s.) for the different distances. Given this result, we chose to run the same individual analyses for all the participants and also include all of them in global analyses.

Figure 6 shows the results for each of the six participants. Each graphic plots the PSS (in ms) as a function of stimulation distance (in meters). Five out of six individual results conformed well to a linear function (participants 1, 2, 3, 4, and 5 see **table 2**). Therefore, these results can be directly compared with the two models of compensation also presented in the graphics. A linear function was fitted to the experimental data (blue triangles) and it is possible to compare this linear function with two theoretical models – one that explains the PSS as a result of a compensation mechanism for differences in propagation velocity (black linear function) and another that explains the PSS as a result of a compensation mechanism for transduction time differences (red linear function). A model representing a compensation mechanism for differences in propagation velocity can be expressed by the function $y = 3x$ (with y representing the PSS in ms and x representing the distance of stimulation in meters), because we know that sound takes about 3 ms to travel 1 m, while the propagation velocity of light is perceptually unnoticed (see **1.1** – "Perceiving Audiovisual Synchrony"). On the other hand, a model representing a mechanism of compensation for differences in transduction time can be expressed by the function $y = 50$, because this is approximately the temporal difference between the transduction of sound and light (with transduction of light being slower) and remains constant regardless of the stimulation distance.

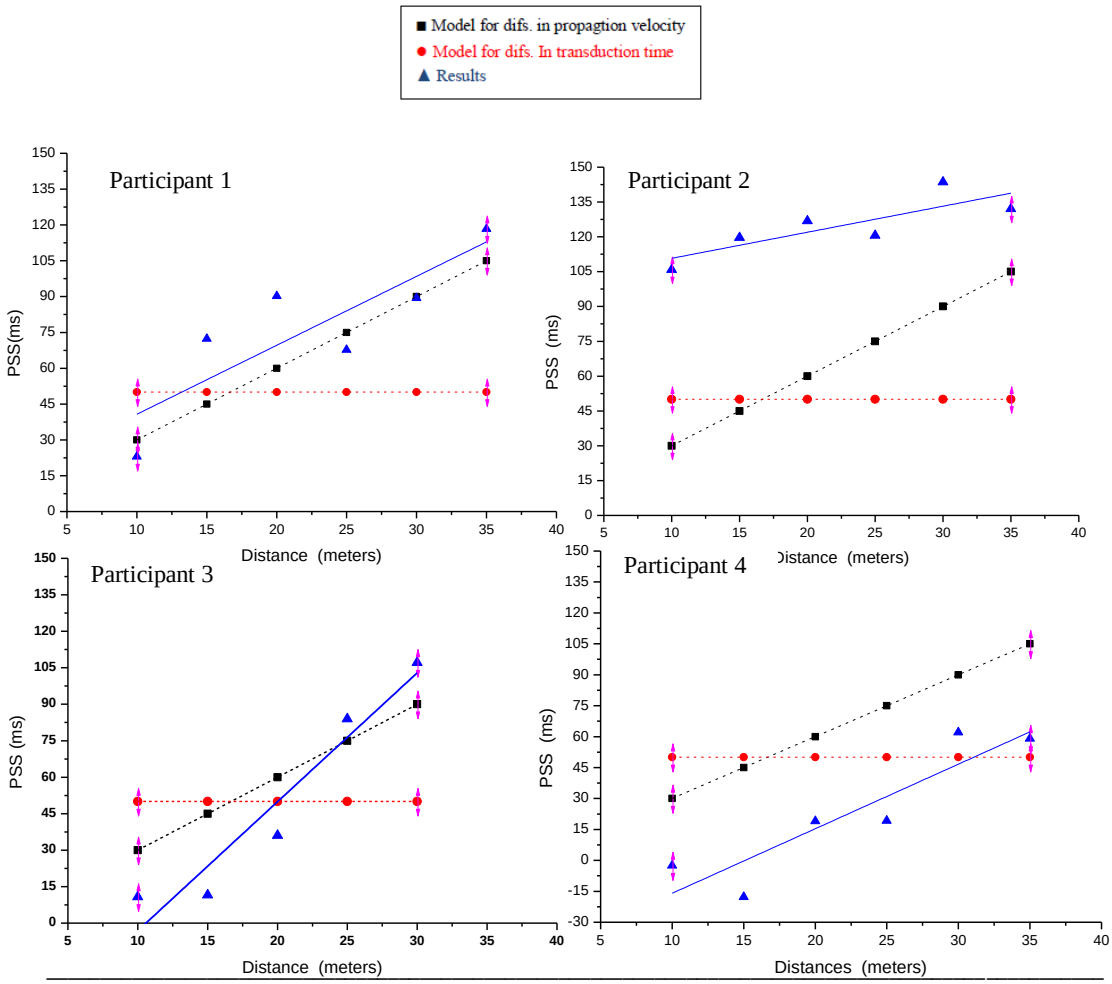


Figure 6. Graphs of the PSS plotted as a function of distance for each of the participants. Black squares correspond to the theoretical values predicted by a mechanism that compensates for differences in propagation time. Red dots correspond to the theoretical values as predicted by a mechanism that compensates for differences in transduction times. Blue triangles are the PSS found for each participant in each distance of stimulation. A linear function was fitted to each group of data.

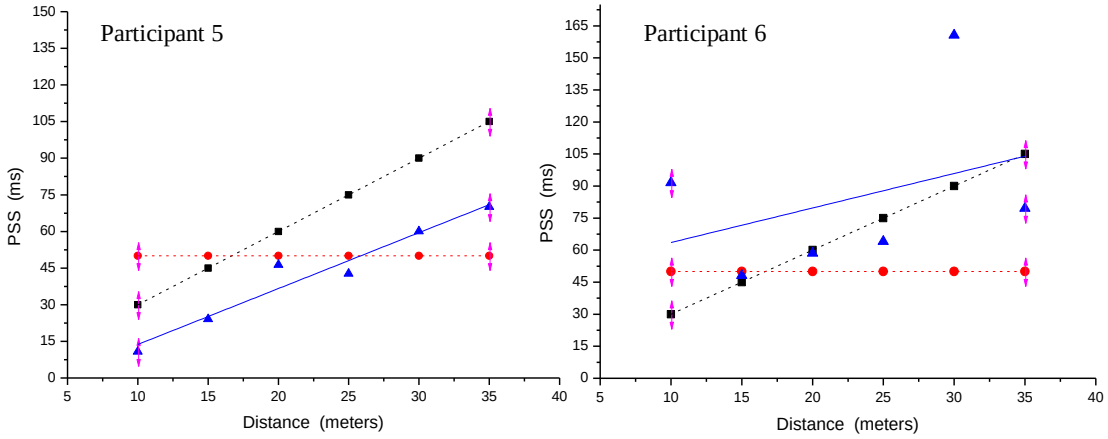


Figure 6 (cont.). Graphs of the PSS plotted as a function of distance for each of the participants. Black squares correspond to the theoretical values predicted by a mechanism that compensates for differences in propagation time. Red dots correspond to the theoretical values as predicted by a mechanism that compensates for differences in transduction times. Blue triangles are the PSS found for each participant in each distance of stimulation. A linear function was fitted to each group of data.

Part.	PSS 10m (WTI)	PSS 15m (WTI)	PSS 20m (WTI)	PSS 25m (WTI)	PSS 30m (WTI)	PSS 35m (WTI)	Linear Fit
1	23 (98)	72 (89)	90 (88.5)	67,7199 (97.5)	89 (105.5)	118 (108)	y = 2.88x + 11.88 (adj. R-Square = 0.65)
2	106 (51)	120 (93)	127 (66)	121 (77)	144 (70)	132 (59)	y = 1.12x + 99.52 (adj. R-Square = 0.59)
3	11 (130)	12 (110.5)	36 (143)	84 (120)	107 (125)	**M**	y = 5.3x – 56.12 (adj. R-Square = 0.89)
4	-3 (58.5)	-18 (51.5)	19 (60)	19 (58.5)	62 (62.5)	59 (63.25)	y = 3.13x – 47.28 (adj. R-Square = 0.78)
5	11 (213)	24 (190)	46 (184)	43 (136)	60 (191)	70 (242)	y = 2.29x – 9.17 (adj R-Square = 0.93)
6	92 (70.5)	48 (48)	59 (63.5)	64 (66)	161 (50)	80 (45)	y = 1.62x + 47.42 (adj. R-Square = 0.08)

Table 2. Value of the PSS and the WTI (both in ms) for each participant in the several distances of stimulation. In the last column are the equations and the values of adjustment for each of the linear functions fitted to the individual data.

All the participants in which data conformed well to a linear function had a positive slope close from that in the model representative of a compensation mechanism for the differences in propagation time (2,88 for participant 1; 1,12 for participant 2; 5.3 for participant 3; and 3,13 for participant 4, and 2,29 for participant 5). Indeed, a one sample t-test indicates that there is no significant difference between the mean slope for the linear functions fitted to the data and the slope of 3 for the model that explains the PSS as a result of a compensation mechanism for the differences in propagation velocities ($t(5) = -0.46$, n.s.). Moreover, there is a significant difference between the data slopes and the zero slope for the model of compensation for differences in transduction time ($t(5) = 4.5$, $p < 0.01$), with the mean of the data slopes being significantly higher than the zero-slope.

When we look at the values of PSS of all the participants we see a mean difference of 96,2 ms between the lower and the highest PSS found, reflected in a mean difference of 24 m. Theoretically, these values should be: a difference of 75 ms between the lowest and the highest PSS reflecting a difference of 25 m in the distance of stimulation, for the model that explains the PSS as a result of a compensation mechanism for sound velocity; no difference in the PSS, regardless of the increment of

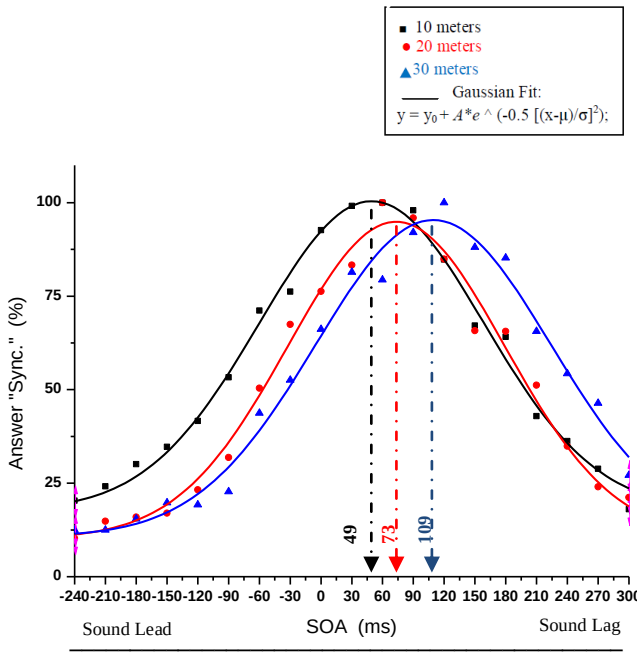
distance of stimulation, for the model that explains the PSS as a result of a mechanism that compensates for transduction time differences. Thus, these results indicate that PSS is increasing with distance and, indeed, correlation tests show that there is a positive correlation between the PSS value and the distance of stimulation ($r = 0.49$, $p < 0.01$), with a higher PSS being associated to higher distances.

Nevertheless, there is a lot of individual variability concerning the intersection point of the linear function fitted to the data, with this parameter being negative for some participants (3, 4 and 5) and positive for others (1, 2 and 6).

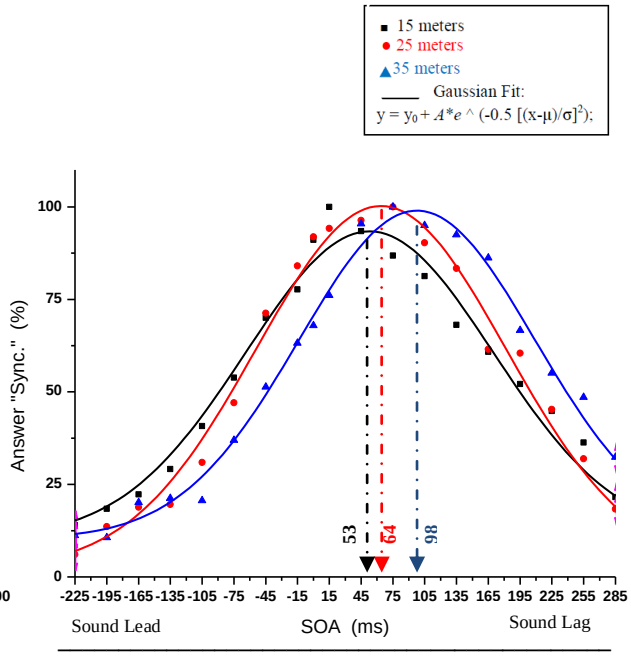
WTIs found in the individual data and measured according to the exposed in section **1.4** and figure **2** seem to have no relation with distance ($r = -0.001$, n. s.).

Considering these individual data, and despite the fact that participants' 6 data regression resulted in a linear model with a low explanatory power, we chose to run a PSS and WTI analysis in a pool data of the six participants. We chose to include participant 6 in this analysis, in spite of the simple regression results, because this participant also showed an increasing tendency in the PSS for several consecutive distances. For participant 6 PSS increases from distance 15 m to distance 20 m, from 20m to 25 m and from 25 m to 30 m. Nevertheless, a decrease in the PSS from distance 10 m to distance 15 m and from 30 m to 35 m is sufficient to explain the poor fit of a linear model to this participant's data. Thus, because all the participants showed a similar tendency in several consecutive distances, we decided to pool all the data in order to get a global indicator of how PSS changes according to distance of stimulation. We chose to run the same analysis in a data pool, instead of opting for more traditional global indicators once we realized that quite often the SOA that elicited the highest percentage of "synchronous" answer would not necessarily be the PSS. Due to the nature of the Gaussian regression used in this analysis, the center of the curve (interpreted as the PSS) was often pulled away from the SOA with the highest percentage of "synchrony" answer because these percentages were higher in one of the two extremes of the distribution. If we had chosen another global indicator like, for example, an overall mean of the PSS for each distance, we would be increasing this Gaussian regression's peculiarity and, potentially, getting values that do not reflect what really happened.

Graphs 1 and 2 show a fitting of a Gaussian function to the pooled data for distances of 10m, 20m, 30 m, and distances of 15 m, 25 m, 35m, respectively. The function fitted to the data is of the type $y = y_0 + A*e \wedge (-0.5\ [(x-\mu)/\sigma]^2)$, where A is the curve amplitude (difference between the highest and the lowest value of the distribution in the y-axis), y_0 is the lowest value of the distribution in the y-axis, σ^2 is the variance and (μ, y_0+A) is the peak of the distribution's coordinates. All the data grouped by distance conformed well to a Gaussian function (see **table 3**) and, as can be seen, graphs 1 and 2 clearly show that the Gaussian curve center is progressively moving towards a higher sound delay as the distance of stimulation increases.



Graph 1. Proportion of "synchronized" answers as a function of the SOA for a data pool of distances 10, 20 and 30 meters. A fit of a Gaussian function was performed in order to get the PSS and WTI values for each distance of stimulation



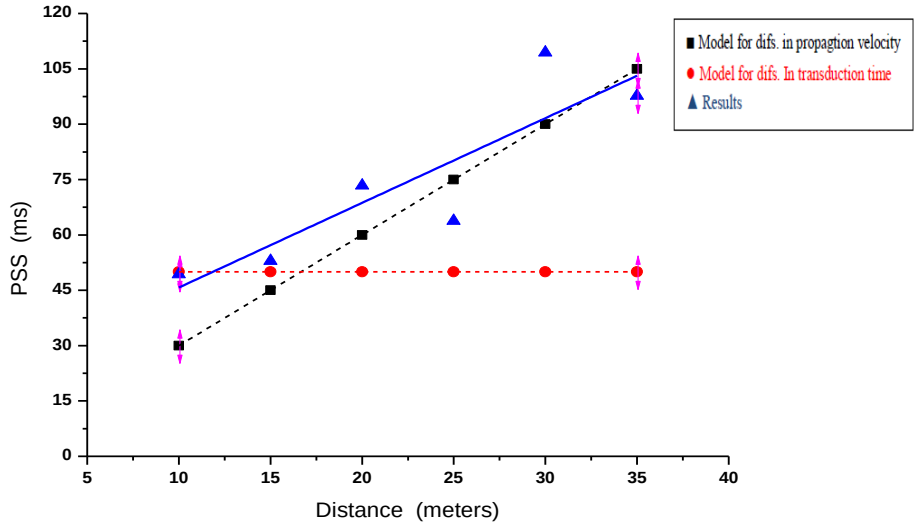
Graph 2. Proportion of "synchronized" answers as a function of the SOA for a data pool of distances 15, 25 and 35 meters. A fit of a Gaussian function was performed in order to get the PSS and WTI values for each distance of stimulation

Distance	Gaussian Function	R^2	ANOVA Results
10 m	$y = 17.6 + 82.7e \wedge (-0.5[(x-49.3)/109.3]^2)$	.987	$F(4, 15) = 1798.2, p < 0.001$
15 m	$y = 10.4 + 83e \wedge (-0.5[(x-52.9)/116.6]^2)$	.966	$F(4, 15) = 691.7, p < 0.001$
20 m	$y = 10.2 + 84.7e \wedge (-0.5[(x-73.4)/105.8]^2)$	.980	$F(4, 15) = 875.3, p < 0.001$
25 m	$y = 2.1 + 98.1e \wedge (-0.5[(x-63.8)/117.8]^2)$	.979	$F(4, 15) = 878.2, p < 0.001$
30 m	$y = 10.6 + 84.7e \wedge (-0.5[(x-109.4)/114.6]^2)$	.979	$F(4, 15) = 908.2, p < 0.001$
35 m	$y = 10.3 + 88.7e \wedge (-0.5[(x-97.7)/111.2]^2)$	.986	$F(4, 15) = 1420.1, p < 0.001$

Table 3. Equation and adjustment value of each of the Gaussian function fitted, by distance, to the pooled data. In the last column is the F-ratio and the value of its significance, for each regression.

Graph 3 plots the PSS from the pooled data as a function of the distance of stimulation. Again, a linear function ($y = 2.29x + 22.7$) was fitted to the data with a good adjustment ($r^2 = .72$; $F(1, 4) = 6.69$, $p < 0.05$). Comparing this linear function with the two models also presented in the graph, we can clearly say that data is closer to a model representative of a compensation mechanism for the differences in propagation velocity than to a model representative of a compensation mechanism for the differences in transduction time. Thus, these results indicate that PSS is increasing with distance. Nevertheless, the correlation between PSS from the pooled data and distance is only marginally significant ($r = 0.79$, $p < 0.1$). Moreover, in the linear function fitted to the data, its slope and especially its interception with y-axis are somewhat different from the same parameters in the model representative of a compensation for differences in propagation velocity.

Interestingly, when we look at the values of PSS of the pooled data we see a difference of 60 ms between the lowest and the highest PSS found, reflected in a difference of 20 m. If we wanted to represent these differences with a linear function, this would be a linear function with a slope of 3.



Graph 3. PSS plotted as a function of distance for a data pool. Black squares correspond to the theoretical values predicted by a mechanism that compensates for differences in propagation time. Red dots correspond to the theoretical values as predicted by a mechanism that compensates for differences in transduction times. Blue triangles are the PSS found for each participant in each distance of stimulation. A fit of a linear function was performed in each group of data.

Distance	PSS	WTI
10 m	49	70.5
15 m	53	82.5
20 m	73	77
25 m	64	91
30 m	109	95
35 m	83	88

Table 4. Values of the PSS and WTI (both in ms) for the pooled data in the several distances of stimulation.

Finally, when we pooled the data, WTI seems to have a positive correlation with distance of stimulation ($r = 0.81$, $p < 0.05$). Results showed that wider WTIs are associated with higher distances of stimulation (see **Table 4**).

2.3 - Discussion

Results from experiment 1 point towards the existence of a compensation mechanism for differences in propagation velocity. We can see, both from the individual and the pooled data, that frequently PSS increases from one distance of stimulation to the following. Thus, quite probably, we are taking into account sound velocity when judging audiovisual synchrony. The results did not show any evidence of compensation for differences in transduction time, which is in accordance with the lack of evidence for this type of compensation in scientific literature, as exposed in **1.2**. It is possible that, because these temporal differences are constant in most of the stimulation contexts, we

recalibrate throughout our lifetime the perception of synchrony to account for this crossmodal difference. Thus, we should have developed a constancy mechanism for audiovisual processing time differences similar to those involved in the active touch perception (Harris et al., 2010).

Despite these conclusions, the slopes found in the linear function fitted to the pooled data indicate that PSS is changing, according to distance, at a slower rate than the speed of sound. In the pooled data analysis we found a rate of change in the PSS of 2.29 ms/m. Given that sound travels at a speed of about 3ms/m, this change in the PSS is approximately 1 ms/m slower. Another important outcome is the high variability in the y-axis interception point as demonstrated in the individual data. Assuming that we compensate for differences in propagation velocity, these outcomes could clearly mean one (or both) things:

1 – Participants are misjudging the distance of the visual stimuli by: (a) judging them as closer than they actually are, as we can see by the cases in which the y-axis interception is negative and the PSS is generally lower than predicted by the model of compensation for sound velocity (participants 3, 4 and 5); (b) judging them as further away than they actually are, as in cases where the y-axis interception is positive and the PSS is generally higher than predicted by the model of compensation for sound velocity (participants 1 and 2);

2 – Compensation for differences in propagation velocity is just one of the mechanism at work in the perception of audiovisual synchrony at different distances of stimulation.

Although we cannot easily either refute or defend this second justification, we can discuss, at some length, the first justification for the experimental outcomes.

Former studies supporting the existence of a compensation mechanism for differences in propagation velocity were undertaken under natural conditions of stimulation (Engel et al. 1971; Sugita & Suzuki, 2003; Kopinska & Harris, 2004; Alais & Carlile, 2005) and, therefore, provided all the depth cues that one can get in everyday life (in the visual stimuli: relative size, familiar size, perspective, elevation, lighting, aerial perspective, curvilinear perspective, accommodation, occlusions, texture gradients, shadings, blurring, stereopsis and convergence; in the audio stimuli: intensity attenuation with distance, reflected sounds and inter-aural differences of intensity and frequency). In the present study we were able to control the number and quality of visual cues and present only two pictorial depth cues, familiar size and elevation, and two dynamic depth cues, amplitude of the step and angular velocity. These were the only visual depth cues presented because, despite the fact that they are probably the most powerful ones, they allow for an easy manipulation and allow us to maintain a virtual environment without many elements and, consequently, keep the diminished light environment, important for the generation of an immersive sensation in this experiment. It is clear to us that other visual depth cues, namely shading and lighting, were inevitably presented. Even so, due to the oversimplified nature of the stimuli, these cues would never be as informative as they could be in real world visual scenarios. Nevertheless, despite the presentation of only two pictorial and two dynamic visual depth cues with enough quality to guide a distance

judgment, we were able to find evidence for compensation for distance of stimulation. However, what can be happening is that due to the presentation of fewer depth cues than in a natural situation, participants are misjudging distance in the ways presented above in point 1. This could be the reason for the high variability in the y-axis and in the slopes of the linear functions adjusted to the data PSSs in graphs of figure 6 and graph 3.

Another important detail that could be contributing for the misjudgment of distances is the lack of information about distance in the auditory stimuli. We implemented HRTFs models that can simulate the inter-aural differences in terms of frequency and intensity, but we were not capable of simulating intensity attenuation with distance nor reflected sounds. These limitations could lead to incongruent information about the localization of the audiovisual stimulus because, while the visual stimulus projections can be simulated as farther or nearer from one condition to another, the auditory stimulus never changes its information about distance (i. e., distance is kept constant throughout different conditions). Consequently, this can result in misjudgments of stimulation distance, because we are providing incongruent information about distance when we increase it in the visual stimulation but not in the auditory one. In sum, if distance is misjudged in all the ways described above, it is expectable that we observe variations in the PSS in relation to what was expected according to the model of compensation for propagation velocity.

Regarding point 2 exposed before, it is possible that due to the procedure used in this experiment we enhanced the existence of some temporal recalibration. By presenting visual distances of stimulation blocked by session, we submitted the participant for a great deal of time to the same visual stimulus. As exposed in section **1.2**, this is a situation prone to temporal recalibration: we have a source of stimulation that is constant and reliable and this could lead the participant to temporally recalibrate in the ways presented in section **1.2**. The general outcome of this recalibration is a tendency to gradually minimize the temporal differences between the two stimuli and thus approach the PSS to the 0 ms SOA. Overall, if temporal recalibration occurs, it can explain the lower slope in the linear regression fitted to the data, when compared with the slope of the model of compensation for differences in propagation velocity.

One way to clarify both issues of distance and temporal recalibration and, at the same time, to explore the role of depth cues in synchrony perception is to design an experiment with two conditions which differ in the number and quality of depth cues, and presenting all the distances of stimulation within the same session in a randomized fashion. By doing this we could grasp the role of certain depth cues in the triggering of a mechanism that compensates for distance of stimulation while avoiding the phenomenon of temporal recalibration. We hypothesize that the more depth cues presented, the more similar the results will be in respect to the model that explains the perception of synchrony by a compensatory mechanism for the differences of propagation velocity (i.e. a linear function fitted to the PSS, plotted against distance, will have a slope closer to 3 than the ones found in experiment 1; moreover, the y-axis interception point should be closer to zero). On the other hand,

having less depth cues then the ones presented in this experiment will lead to results increasingly unrelated with distance. To test these hypotheses we conducted a second experiment.

3. EXPERIMENT 2: Assessing the role of depth cues

3.1 – Method

3.1.1 – *Participants*

Four participants aged 22 -28 took part in this experiment. None of them had participated in the first experiment. All underwent visual and auditory standard screening tests and had normal hearing and normal, or corrected to normal, vision. All of the participants were university students and all gave informed consent to participate in the present study.

3.1.2 – *Stimuli and Material*

All the material and facilities used in experiment 1 were used as well in experiment 2 (see section **2.1.2**).

3.1.2.1 –Visual Stimuli

In experiment 2 we had two types of visual stimuli:

a) One, consisting of a PLW similar to the one in experiment 1, but at a velocity of 1.1 m/s and with three different durations (1.08 ms; 1.12 ms; 1.17 ms). This PLW took only one step (at the 542nd ms in the minimum duration; at the 563rd ms in the medium duration; and at the 583rd ms at the maximum duration). We reduced the moments of synchrony judgment from three to one step because of two reasons: Firstly, several participants in experiment 1 reported that they had made their synchrony judgment based only on the first step and the other two steps were either confusing the judgment or completely ignored. Secondly, this second experiment had two experimental conditions (see **3.1.3** – "Procedure"), which meant that total duration would be twice that of experiment 1, had we kept the same stimulus duration. By presenting just the first step we provide sufficient moments to have a clear synchrony judgment, while keeping a tolerable experimental duration. Also, this PLW had three different durations because we wanted to randomize the time between the beginning of the stimuli and the step, so that a constant time could not be used as a marker for the occurrence of the visual step, turning it into a cue for the existence, or not, of synchrony.

Notwithstanding, the fundamental difference between this stimuli and the one from experiment 1 is that, apart from the depth cues already presented in that experiment, we also added a perspective depth cue consisting in the graphic simulation of several rectangles with the same physical dimensions, but located at different distances (see **fig. 7**). This perspective depth cue was design to give the impression of a room with 9.7 meters wide, 4.5 meters high (dimensions close to the room were the experiment took place), and 35 meters long. The floor and wall lines were virtually located at

10, 15, 20, 25, 30, and 35 meters from the observer. Thus, the PLW corresponding to each one of these distances was presented as walking right on top of the correspondent floor ground line.

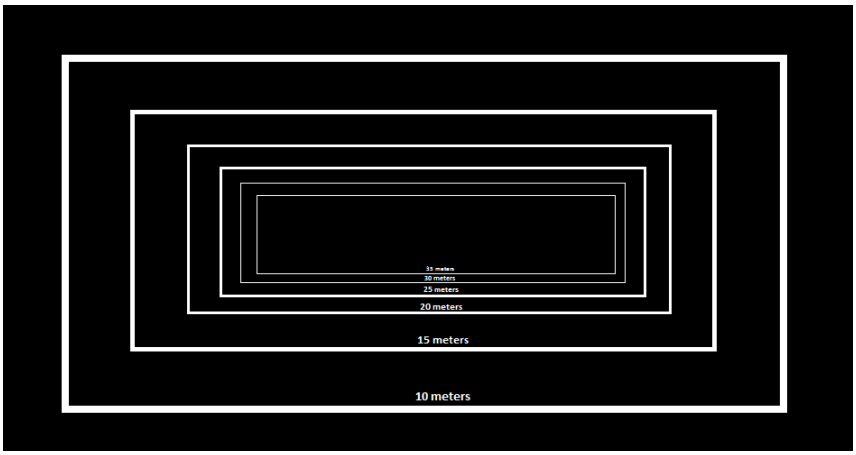


Figure 7. Perspective depth cue.

This pictorial depth cue adds several new depth cues not presented in experiment 1, namely: *perspective*, given by the convergence of parallel lines at infinity (in this case those lines are not draw, but one can have the sensation of parallel lines due to the alignment of the rectangle vertices); and *relative size*, because although the real size of the rectangles is always the same, they look smaller as they are farther away. At the same time, this depth cue also improves the quality of familiar size and elevation pictorial depth cues, because now the participant can compare the knowledge of the dimensions of a human figure at a certain distance with the knowledge of the relative dimension of the figure inside a room. Also, the elevation effect is more clearly represented in the lines of the ground. Moreover, more information about lighting is present due to the lower luminance of objects at a greater distance (here we get information from the PLW itself as in experiment 1 and from the lines that delimit the room). So, these visual stimuli provide more information about distance of stimulation than the ones presented in experiment 1. Additionally, the temporal discrimination could also improve, because the PLW is walking on top of the ground line, which gives us more accurate information about the moment when its foot touches the ground.

b) The second type of visual stimulus used in experiment 2 was conceived in order to eliminate as many depth cues as possible. In these stimuli we had no contextual distance cues, and so we had, as in experiment 1, only the PLW walking against a black background. In order to provide even less information about distance than in experiment 1 (that had all the information about familiar size and elevation of the PLW), we present only the feet with a random size of the dots and at a constant elevation (see **fig. 8**). By doing this, we eliminate all the bodily cues about familiar size and keep elevation constant. Thus the only distance cue presented is the amplitude of the step (with wider steps meaning a closer presentation). Nevertheless, amplitude of step is not a strong cue, because in order to give accurate information about distance, the several presentations would have to keep velocity, frequency, and amplitude constant. Although we do this in this experiment, this is not a natural situation in our everyday life, where these parameters are quite variable. All the temporal and

spatial parameters of the step movement were the same as in the first visual stimuli described in this experiment.

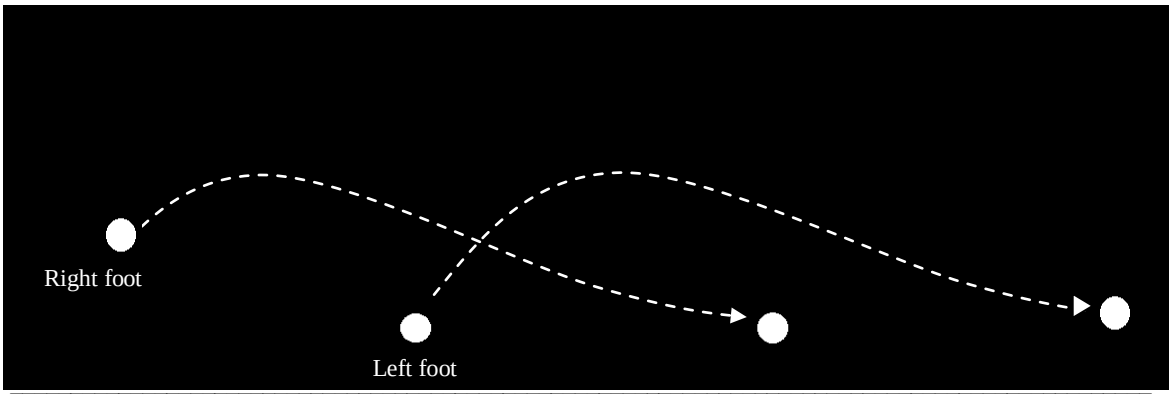


Figure 8. Movement description of the second visual stimuli. Only two dots are projected, each one corresponding to one foot. At the beginning of the trial, the left foot is in contact with the ground and the right foot is already in a movement phase of the step cycle (swing phase). During the trial time only the right foot will hit the ground and the left foot ends the trial in a swing phase.

3.1.2.2 – Auditory Stimuli

The auditory stimuli were one-step sounds from the database of controlled recordings from the College of Charlston (Marcell et al. 2000). This step sound corresponds to the sound of a male human walking over a wooden floor and taking one step. The sounds were auralized as free-field (with no distance information from reverberations) by a MATLAB routine with HRTFs from the MIT database (http://sound.media.mit.edu/resources/KEMAR.html). This auralization process allowed us to present the step sound in a central localization matched with the spatial position (in a front-parallel plane to the observer) where the visual foot touches the ground. Again, no information about distance of stimulation in the front-perpendicular plane to the observer was presented.

3.1.2.3 – Visual and Auditory Stimuli Relation

For this experiment we used the same SOAs as in experiment 1.

3.1.3 – *Procedure*

Unlike what happened in experiment 1, experimental sessions were not blocked by distance of stimulation. Thus, audiovisual stimuli at several distances were randomly presented in the same experimental session. In each session, 36 different audiovisual stimuli were presented (corresponding to the different SOAs for both even and odd distances) at 6 different distances, in a total of 114 audiovisual scenes. All scenes were randomly presented with 10 repetitions each. There was an inter-stimulus interval of 1.6 seconds between each scene, where a fixation cross marking the distance at which the next stimulus would appear. Each session took about 52 minutes to complete and an break, no longer than 3 minutes, was always taken in the middle of the session. It took 8 sessions (4 for each condition) to complete the experiment, in a total time of experiment of about 7 hours for each participant. Two of the participants completed the "full depth" condition first (the condition using the

visual stimuli presented in section **3.1.2.1 – (a)**) and the other two completed the "low depth" condition first (the condition using the visual stimuli presented in section **3.1.2.1 – (b)**).

Before each experimental session the participants were exposed to 10 repetitions of an audiovisual stimulus in which the sound appeared with a 300 ms lead, and to 10 repetitions of an audiovisual stimulus in which the sound appeared with a 330 ms lag. This preliminary session was devised in order to check if participants were able to perceive any kind of audiovisual asynchrony. Note that none of the SOAs used in this preliminary session were then used in the experimental session.

At the beginning of the experimental session the following introductions were given: "You will participate in an audiovisual perception study in which you will be presented with several audiovisual scenes of a PLW walking at a certain distance. I want you to pay close attention to the audiovisual scenes because you will have to judge its audiovisual synchrony during the intervals between scenes. Thus, after each scene, if you think that the auditory and the visual streams were synchronized click the right button; otherwise, if you think that the auditory and the visual streams were not synchronized click the left button" (SJ task).

The participant was seated in a chair 4 meters from the screen and aligned with the center of the projection area. In each scene participants were visually presented with a PLW walking from left to right and taking one step at a velocity of 1.1 m/s, while listening, trough in-ear phones, to one step with the same angular location of the visual stimuli and in a given temporal relation with the visual stimulus (see **fig. 3**). Thus, participants had one moment where they could judge the audiovisual synchrony of the stimulus. After the presentation of each audiovisual stimulus and during the inter-stimulus interval, the participant had to answer in a two key mouse according to the instructions.

3.2 – Results

Among the participants of this experiment, two had some background knowledge about the thematic of the study and the remaining two were naive to the purpose of the experiment. According to a t-test for individual samples there is no significant difference between these groups with regard to the PSS and WTI values in the "full depth condition" ((t (22) = -1.06, n. s.) for differences in the PSS; (t (34) = -0.85, n. s.) for differences in the WTI) and in the "low depth condition" with regard to the WTI (t ((34) = -0.73, n. s). There is, however, a significant difference in the "low depth condition" regarding the PSS (t (22) = 7.64, $p < .01$). Participants with background knowledge appear to have higher PSSs (in direction of an audio lag) when compared to participants with no experience in psychophysical studies. We ran the same individual analyses for all the participants and, despite the above cited differences and having in mind the porpoise of this experiment, we still chose to include all of them in a global analysis similar to the one in experiment 1.

Figure 9 shows the results for each of the four participants in graphs similar to those of figure 6, but this time plotting the results of the two experimental conditions: "full depth condition" (blue

dots) and "low depth condition" (orange triangles). All of the individual results conformed well to a linear function and, therefore, we can compare the results in the two conditions and also compare the results of each condition to the model of compensation for differences in propagation velocity.

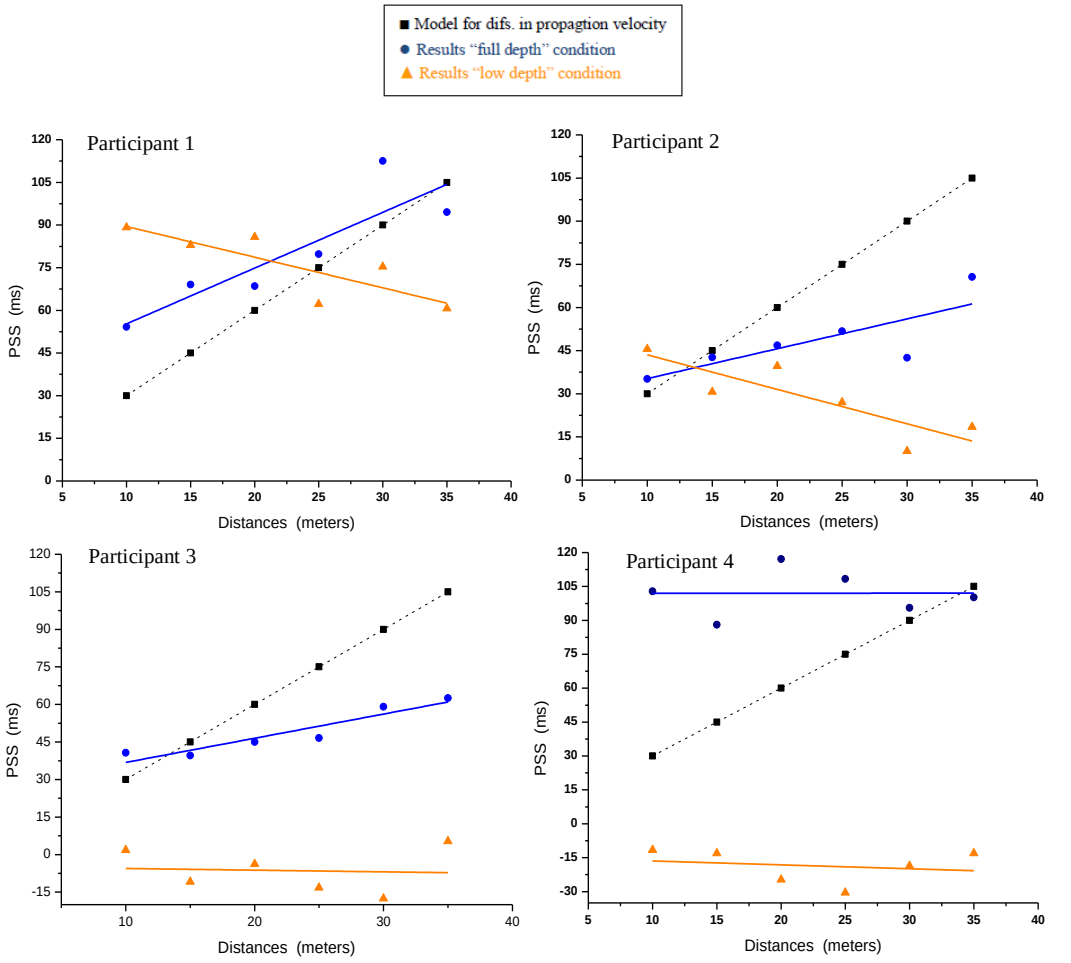


Figure 9. Graphs of the PSS plotted as a function of distance for each of the participants in the two conditions of stimulation. Black squares correspond to the theoretical values predicted by a mechanism that compensates for differences in propagation time. Blue dots are the PSS found for each participant in each distance of stimulation in the "full depth" condition, and orange triangles are the PSS found for each participant in each distance of stimulation in the "low depth" condition. A fit of a linear function was performed in each group of data.

In the "full depth condition" all the linear functions fitted to the data had a positive slope, with the exception of the almost-zero slope for participant's 4 data (see **table 5**). Nevertheless, a one sample t- tests show that there is a significant difference between the data's mean slope, which is lower than the prediction of a slope of 3 given by the model of a compensation mechanism for propagation velocity ($t(3) = -5$, $p < .05$). Still, the data's mean slope is marginally higher than zero ($t(3) = 2.5$, $p < .1$).

Part. / Condition		PSS 10m (WTI)	PSS 15m (WTI)	PSS 20m (WTI)	PSS 25m (WTI)	PSS 30m (WTI)	PSS 35m (WTI)	Linear Fit
1	"Full Depth Cond."	54 (98.5)	69 (113)	69 (106)	80 (96.5)	113 (78)	95 (39)	$y = 1.96x + 35.62$ (adj. $R^2 = 0.71$)

	"Low Depth Cond."	89 (108)	83 (117)	86 (98.5)	62 (88)	75 (86.5)	61 (79.5)	y = -1.07x + 100.3 (adj. R^2 = 0.60)
2	"Full Depth Cond."	35 (86)	43 (70)	47 (77)	52 (69.5)	43 (87)	71 (60.5)	y = 1.04x + 24.88 (adj. R^2 = 0.54)
	"Low Depth Cond."	46 (66)	31 (69)	40 (60.5)	27 (62)	10 (49.5)	18 (68)	y = -1.20x + 55.50 (adj. R^2 = 0.66)
3	"Full Depth Cond."	41 (107)	40 (109)	45 (104)	47 (116.5)	59 (90.5)	63 (116.5)	y = 0.97x + 27.16 (adj. R^2 = 0.85)
	"Low Depth Cond."	2 (116)	-11 (115)	-4 (113.5)	-13 (121)	-18 (109)	5 (117)	y = -0.07x – 4.88 (adj. R^2 = 0.42)
4	"Full Depth Cond."	103 (132.5)	88 (106)	117 (125)	108 (119)	96 (101)	100 (100.5)	y = 102 (adj. R^2 = 0.71)
	"Low Depth Cond."	-12 (91.5)	-13 (103)	-25 (91)	-30 (64.5)	-19 (101.5)	-13 (91)	y = -0.17 – 14.73 (adj. R^2 = 0.71)

Table 5. Individual values of the PSS and the WTI (both in ms) for the several distances of stimulation in the "full depth" condition and in the "low depth" condition. In the last column are the equations and the values of adjustment for each of the linear functions fitted to the individual data.

Moreover, when we look at the values of PSS of all the participants in the "full depth condition", we see a mean difference of 36.43 ms between the lowest and the highest PSS found, reflected in a mean increase of 20 m in the distance of stimulation. Theoretically, these values should be a difference of 75 ms between the lowest and the highest PSS reflecting a difference of 25 m in the distance of stimulation, for the model that explains the PSS as a result of a compensation mechanism for sound velocity. Thus, although we obtain smaller values relatively to the ones derived from this analysis presented in experiment 1, these results appear to indicate that PSS is increasing with distance. In fact, correlation tests show that there is a marginally positive correlation between the PSS value and the distance of stimulation ($r = 0.327$, $p < 0.1$), with higher PSSs being associated with higher distances. However, there is still a lot of variability concerning the intersection point of the linear functions fitted to the individual data.

WTIs found in the individual data concerning the "full depth" condition seem to have only a marginal significantly correlation with distance ($r = -0.38$, $p < 0.1$). However, this time the association is on the opposite direction from the one found in the pooled data of experiment 1: the length of WTIs is getting smaller while distance is increasing.

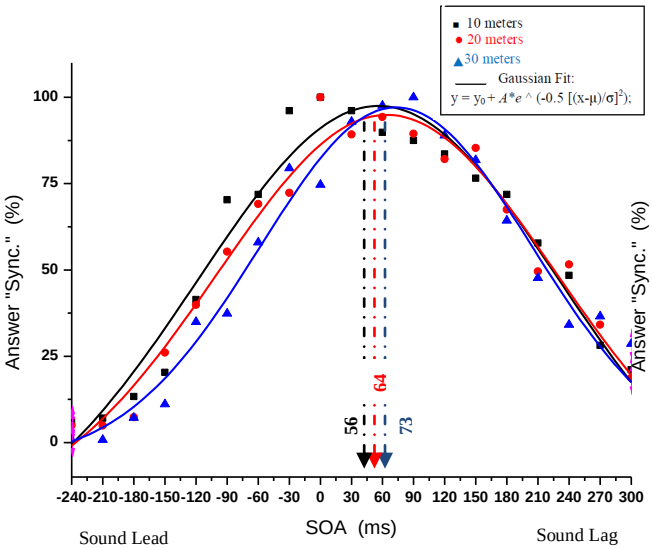
In the "low depth" condition there is more variability in the slopes of the linear functions fitted to the data; nevertheless, and more importantly, none of these slopes are positive. In fact, they are all negative and, in some cases, quite close to zero (see participants 3 and 4). A one sample t-test indicates that there is a significant difference between the mean slope for the linear functions fitted to the data in the "lower depth" condition and the theoretical slope for the model of compensation represented in the graphs ($t(3) = -12.3$, $p < 0.01$), with the slope of a model of compensation for differences in propagation velocity being significantly higher than the mean slope from the data linear regression. On the other hand, there is no difference between the mean slope for the linear functions

fitted to the data and a zero slope (t (3) = -2.1, n.s.). In the "low depth" condition, when we look at the values of PSS of all the participants, we see a mean difference of 25.5 ms between the lowest and the highest PSS found, reflected in a mean decrease of 20 m in the distance of stimulation. This outcome is quite different from the expected according to a model of compensation for differences in propagation velocity (because, for participants 1and 2, PSS is decreasing with distance, which is the opposite of what we have seen so far). Furthermore, correlation tests show that there is no correlation between the PSS value in the "low depth" condition and the distance of stimulation (r = -0.14, n.s.). Again, there is a lot of variability concerning the intersection point of the linear functions fitted to the data.

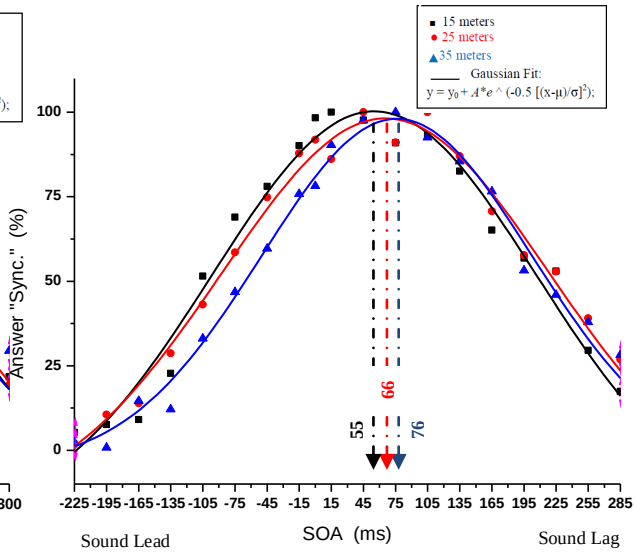
In the individual data concerning the "low depth" condition, no relation was found between the length of the WTI and the distance of stimulation (r = -0.19, n. s.).

Considering the success of the linear regressions for all the individual data, we chose to run a PSS and WTI analysis in a pooled data of the four participants in both conditions.

Graphs 4 and **5** show the fitting of a Gaussian function to the pooled data, for distances of 10 m, 20 m, 30 m (graph 4) and 15 m, 25 m, 35 m (graph 5) in the "full depth" condition. All the data, grouped by distance conformed well to the Gaussian function (see **table 6**) and, as in experiment 1, we can see that the center of the Gaussian curve is progressively moving towards a higher sound delay as the distance of stimulation increases. This increment appears generally lower than the correspondent one in experiment 1 (see **graphs 1** and **2**). There is a difference of only 20 ms between the lowest and highest PSS, while in the pooled data of experiment 1 this difference is of 60 ms. Nevertheless, there is no significant difference in the PSS values between the pooled results in experiment 1 and the pooled results in the "full depth condition" of this experiment (t (10) = 0.701, n. s.).

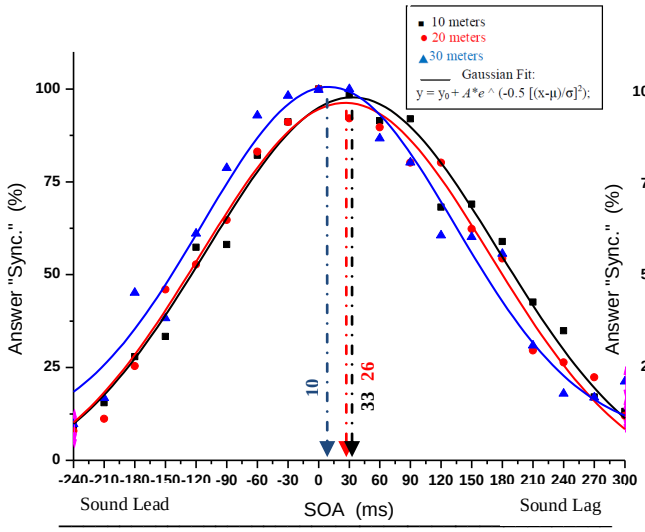


Graph 4. Proportion of "synchronized" answers as a function of the SOA for a data pool of distances 10, 20 and 30 meters in the "full depth" condition. A fit of a Gaussian function was performed in order to get the PSS and WTI values for each distance of stimulation.

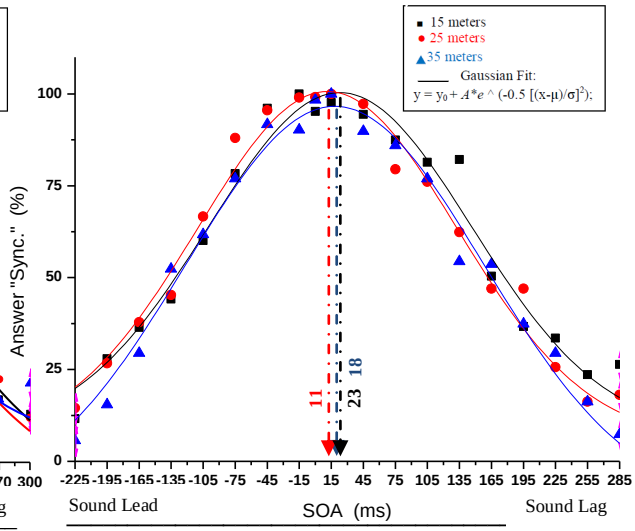


Graph 5. Proportion of "synchronized" answers as a function of the SOA for a data pool of distances 15, 25 and 35 meters in the "full depth" condition. A fit of a Gaussian function was performed in order to get the PSS and WTI values for each distance of stimulation.

Graphs 6 and 7 show the fitting of a Gaussian function to the pooled data for all the distance of stimulation, but in the "low depth" condition. All the data grouped by distance conformed well to the Gaussian function (see **table 6**). In general, we can see in both in graphs that the center of the Gaussian curve hardly moves from one distance to another (especially in graph 7) and when it does move, it does so in the direction of a lower sound delay: the opposite outcome of the same data analysis in experiment 1 and in the pooled data for the "full depth" condition in experiment 2. Also, for the first time within a PSS analysis with a data pool, we found several PSSs with a value quite close to zero, the value of real physical synchrony (see distances 25 m and 30 m). Furthermore, statistical tests show that the PSS values in the "low depth" condition are significantly different from the pooled PSS results in experiment 1 ($t(10) = 5.28$, $p < 0.01$) and from the PSS results in the pooled data for the "full depth" condition ($t(10) = 8.75$, $p < 0.01$).



Graph 6. Proportion of "synchronized" answers as a function of the SOA for a data pool of distances 10, 20 and 30 meters in the "low depth" condition. A fit of a Gaussian function was performed in order to get the PSS and WTI values for each distance of stimulation.

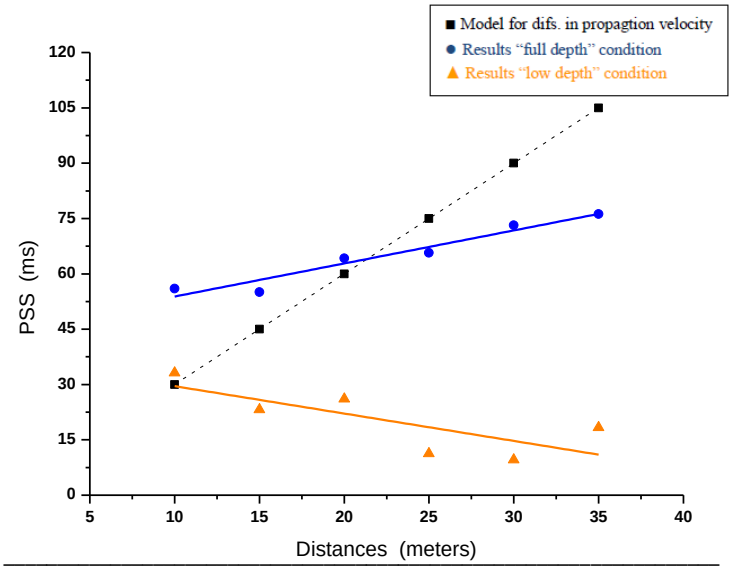


Graph 7. Proportion of "synchronized" answers as a function of the SOA for a data pool of distances 15, 25 and 35 meters in the "low depth" condition. A fit of a Gaussian function was performed in order to get the PSS and WTI values for each distance of stimulation.

Distance		Gaussian Function	R^2	ANOVA Results
10 m	"Full depth"	$y = -31 + 128.6e \wedge (-0.5[(x-56)/174.7]^2)$	.946	$F(4, 15) = 343.9$, $p < 0.001$
	"Low Depth"	$y = -7 + 104.8e \wedge (-0.5[(x-33.2)/142.9]^2)$	.973	$F(4, 15) = 727.9$, $p < 0.001$
15 m	"Full depth"	$y = -22.8 + 123.1e \wedge (-0.5[(x-55.1)/151.5]^2)$	.964	$F(4, 15) = 526.6$, $p < 0.001$
	"Low Depth"	$y = 7.6 + 92.7e \wedge (-0.5[(x-23.2)/123.2]^2)$	.959	$F(4, 15) = 569.7$, $p < 0.001$
20 m	"Full depth"	$y = -21.4 + 116.4e \wedge (-0.5[(x-64.2)/163.2]^2)$	.963	$F(4, 15) = 489.7$, $p < 0.001$
	"Low Depth"	$y = -7.7 + 104e \wedge (-0.5[(x-26.1)/141.8]^2)$	.974	$F(4, 15) = 731.6$, $p < 0.001$
25 m	"Full depth"	$y = -17.6 + 115.8e \wedge (-0.5[(x-65.7)/152.5]^2)$	.98	$F(4, 15) = 1061$, $p < 0.001$
	"Low Depth"	$y = 5.8 + 94.9e \wedge (-0.5[(x-11.3)/121.6]^2)$	.973	$F(4, 15) = 800.9$, $p < 0.001$
30 m	"Full depth"	$y = -5.8 + 102.9e \wedge (-0.5[(x-73.2)/131.5]^2)$	.964	$F(4, 15) = 435.4$, $p < 0.001$
	"Low Depth"	$y = 5.9 + 94.7e \wedge (-0.5[(x-9.6)/124.1]^2)$	.946	$F(4, 15) = 357.7$, $p < 0.001$
35 m	"Full depth"	$y = -4.8 + 102.9e \wedge (-0.5[(x-76.2)/126.1]^2)$	.982	$F(4, 15) = 977.7$, $p < 0.001$
	"Low Depth"	$y = -12.7 + 109.3e \wedge (-0.5[(x-18.3)/139.5]^2)$	.977	$F(4, 15) = 819.2$, $p < 0.001$

Table 6. Equations and adjustment values for each of the Gaussian function fitted, by distance, to the pooled data in both conditions of stimulation. In the last column are the F-ratios and the value of its significance, for each regression.

Graph 8 plots the PSS from the pooled data as a function of the distance of stimulation. A linear function was fitted to the PSSs obtained in the "full depth" condition (y = 0.9x + 45) with a good adjustment (r^2 = 0.93; F (1,4) = 63.84, p < 0.01). Similarly, a linear function was fitted to the PSSs obtained in the "low depth" condition (y = -0.7x + 37) with a roughly good adjustment (r^2 = 0.5, F (1,4) = 5.77, p < 0.1). In this graph we can easily compare these two linear functions between them and with the model of compensation for differences in propagation velocity.



Graph 8. PSS plotted as a function of distance for a data pool. Black squares correspond to the theoretical values predicted by a mechanism that compensates for differences in propagation velocity. Blue dots are the PSS found for each participant in the "full depth" condition and orange triangles are the PSS found for each participant in the "low depth" condition. A fit of a linear function was performed in each group of data.

Distance		PSS	WTI
10 m	"full depth"	56	108
	"low depth"	33.2	100
15 m	"full depth"	55	103
	"low depth"	23.2	98
20 m	"full depth"	64.2	100
	"low depth"	26.1	95.5
25 m	"full depth"	65.6	101.5
	"low depth"	11.3	91.5
30 m	"full depth"	73.2	91
	"low depth"	9.5	98.5
35 m	"full depth"	76.2	90
	"low depth"	18.3	93

Table 7. Values of the PSS and WTI (both in ms) for the pooled data by distance and stimulation condition.

One can easily see from the graph that the results from the two conditions present an opposite tendency. While the PSS from the "full depth" condition appears to be increasing with distance, the PSS from the "low depth" condition appears to be decreasing with distance. In fact correlation tests

show that the "full depth" condition PSSs are marginally correlated with distance ($r = 0.791$, $p < 0.1$) in the sense that higher PSSs are associated with higher distances. On the other hand, the PSSs in the "low depth" condition are also marginally correlated with distance ($r = -0.77$, $p < 0.1$), but in the opposite sense: higher PSSs are associated with lower distances.

Again, the slopes and the interceptions of the y-axis with the data linear regression functions appear to be fairly different from the same parameters in the compensation model represented in graph 8. Nevertheless, the only data linear function that resembles, or presents a similar tendency, with the one from the compensation model for differences in propagation velocity is the linear function fitted to the data from the "full depth" condition.

Regarding the WTIs in the pooled data, the tendency of a negative correlation remains between this measure and distance of stimulation, for the "full depth" condition ($r_{sp}= -0.81$, $P < 0.05$), while no correlation between these two variables is found in the "low depth" condition ($r_{sp}= -0.41$, n.s.).

3.3 - Discussion

Experiment 2 was conducted in order to clarify the results that, in experiment1, were pointing towards the existence of a compensation mechanism for sound propagation velocity. The hypothesis that we placed after experiment 1 was that in an experiment manipulating distance cues, we should get essentially two results: An increment of the PSS along with an increment in distance of stimulation closer (than in experiment 1) to the model of compensation for differences in propagation velocity, in conditions with more depth cues than the ones presented in experiment 1; results increasingly unrelated with distance and, therefore, increasingly different from the model of compensation for differences in propagation velocity, in conditions with less depth cues than the ones presented in experiment 1.

As we could see by the results in experiment 2, if we fitted a linear function $y = 3x$ to the data in any of the conditions, we would probably get a poor data adjustment. Nevertheless, if we consider as an approximation to the model of compensation for differences in propagation velocity a set of data that shows an increment in the PSS associated to an increment in distance of stimulation (independently of the magnitude of that increment) then, both in the individual data (perhaps with the exception of participant 4) and in the pooled data, the results of the "full depth" condition are clearly closer to this model than the results from the "low depth" condition. Indeed, it was only in the "full depth" condition that we found positive slopes in the linear functions fitted to the data. In fact, the negative slope found in some of the individual data and in the polled data (**graph 8**) cannot be easily explained by the known explanatory hypothesis for the perception of synchrony across different distances. None of the theoretical models presented before and taken as possible mechanisms of compensation for physical and neural differences across the two sensorial modalities predict a

negative correlation between the PSS and distance of stimulation. Moreover, there is a weaker relation between PSS and distance in the "low depth" condition than in the "full depth" condition.

Despite this unexpected outcome in the "low depth" condition (negative slopes) for some individual data and for the pooled data, we should also pay close attention to the results of participants 3 and 4 (see **fig. 9**). What one can see in these results is that PSS is practically unaffected by distance and remains quite close to the point of physical synchrony (SOA 0ms) which is exactly the predictable outcome for a situation where no compensation for differences in propagation velocity exists.

Overall, we think that we succeed in confirming part of the hypothesis raised in the discussion of experiment 1: evidence for the existence of a mechanism of compensation for differences in propagation velocity is stronger when more depth cues are presented. Independently of the slope found in both conditions of stimulation, the most important result withdrawn from this experiment should be that the presence and quality of depth cues appears to be important in the activation of a mechanism of compensation for distance.

Nevertheless, it is important to discuss the fact that the slope from the linear function fitted to the pooled data in the "full depth" condition (0.9) is quite different from the slope of 3 in the model of the mechanism of compensation for differences in propagation velocity. In the "full depth condition" we were expecting, at least, slopes generally higher than the ones found in the pooled data linear regression in experiment 1. More depth cues were available in this condition and, consequently, the results should be closer to the model of compensation for differences in propagation velocity. Furthermore, we were also expecting less variability in the y-axis interception and a general result in this parameter close to zero. This would mean that the increment in the number of depth cues was improving the judgment of stimulation distance. Contrary to this expectation we obtain a slope from the linear regression in the pooled data (0.9) lower than the one in experiment 1 (2.29) and the individual variability in the y-axis interception was maintained.

In our view, it is possible that this counterintuitive outcome is mainly due to three simple changes in the procedure from one experiment to another:

1 - After experiment 1, we tried to account for the critique and consequent alert of Vroomen and Keetels (2010) for the possible effects of blocking by session the presentation of the different distances in the work of Kopinska and Harris (2004). As exposed in section **1.2**, there is some evidence that exposure for a large amount of time to specific temporal disparities could lead to phenomena of temporal recalibration. Thus, in experiment 1, the lower slope in the data when compared to the model that compensates for differences in propagation velocity could be a consequence of some effect of temporal recalibration. Therefore, we wanted to avoid this in experiment 2, in order to assure that a possible effect of distance in the PSS was exclusively due to a mechanism of compensation for distance. This was ensured by presenting the PLW in each session at several distances in a randomized fashion. However the trade off using this procedure was a more visually demanding task. The fact that we presented a PLW every 3 seconds at a different location in

the screen and with a different angular size and velocity that had to be subject to a judgment, could be, for untrained participants, a highly demanding task. We know that uncomfortable stimulation could impair the results in some perceptual tasks and this may explain this unexpected outcome. In fact, some participants, essentially the less experienced ones in terms of psychophysical studies, reported eyestrain after some sessions in the "full depth" condition, but not in the "low depth" condition. Therefore, it is possible that a highly demanding task, in terms of visual stimulation, could be somewhat disruptive for the judgment of stimulation distance. Future works should take this into account. We could avoid this problem by presenting less distances of stimulation at each session or, by blocking the presentations by distance, choosing, for example, to run more sessions with less duration in order to avoid temporal recalibration.

2 - The implementation of the perspective depth cue may have lead to the emergence the perceptual phenomenon known as *Slant effect*. This effect, first reported by Gibson (1950), is well-defined by a tendency to misjudge visual distance in environments that try to simulate depth perception (as in screen projections), by judging the far stimuli as closer than they really are. Of course, the more immersive the virtual environment is the less manifestations of the Slant effect we should get. Thus, as we cannot guarantee a fully immersive VR environment it is possible that the slant effect is contributing to the misjudgment of the highest distances. The outcomes of this effect are clear, especially in the data of participants 2 and3 and in the pooled data.

3 - Finally, and perhaps most importantly, the judgment of distance could again have been affected by the spatial incongruence between the auditory and the visual stimuli. Has exposed in the discussion of experiment 1 we were not capable of simulate important auditory depth cues as intensity attenuation with distance and reflected sounds. These limitations could lead to incongruent information about localization of the audiovisual stimulus. Notwithstanding, in experiment 1 the auditory depth cue of angular velocity was still informative due to the existence of three steps, each one in a different position in a fronto-parallel plane to the observer. However, when we reduced the number of steps in the auditory stimuli of experiment 2, we unintentionally reduced the information about distance provided by this depth cue. Thus, the level of incongruence between the auditory and the visual stimuli in respect to information about localization was even higher in experiment 2. While visual distance was changing from one scene to another, the auditory distance was kept constant throughout the session. It is possible that participants were using some strategy to deal with this incongruence and calculate audiovisual distance, like averaging the perceived presentation distance of both visual and auditory stimulus. If this was the case, the audiovisual stimulus may have been systematically judge as closer than what we intended.

In regard to WTIs, we had an ambiguous result concerning their possible relation with distance in experiment 1. Therefore, we were looking forward to the WTIs results in experiment 2 in order to clarify this question. In the "lower depth" condition we did not find any relation between the WTIs length and distance of stimulation, and this outcome was somehow expectable due to the limited

information regarding distance of stimulation. However, in the "full depth" condition we found a marginal correlation in the individual data and a significant correlation in the pooled data between WTIs length and distance. Only this time, and unlike the results in experiment 1, we found a negative correlation between these two variables. At a first glance, this experimental outcome appears to be of little help for our experimental purposes when we decided to measure the relation between distance and WTIs length. We did want to explore the relation between distance and WTIs length, because of the findings of Lewald and Guski (2004). In their work, cited above in **1.2**, besides finding a negative correlation between distance of stimulation and PSS value that lead them to conclude that there was no compensation for distance of stimulation, the authors also found a positive relation between the WTIs length and distance of stimulation. What both these experimental outcomes together lead them to conclude was that WTIs alone were dealing with the crossmodal temporal disparities resulting from changes in distance of stimulation, by increasing their length in higher distances of stimulation. So, according to Lewald and Guski, we did not need a moving WTI (as predicted by the models of compensation for distance) because we have the capability of "stretching it" enough to account for the crossmodal temporal disparities in high distances of stimulation. In other words, we become more tolerable to crossmodal temporal disparities in higher distances of stimulation, but we will always perceive as more synchronic stimuli that are indeed temporally synchronic when they arrive at the sensorial receptors. Despite finding evidence in contrary for the relation between PSS and distance of stimulation, we also found an opposite relation between the WTIs length and distance of stimulation. It seems that, both from our individual and pooled data in experiment 2, WTIs of smaller lengths are associated to higher distances of stimulation. Thus, this makes it even more imperative to find a complementary mechanism that accounts for the perception of audiovisual synchrony over high stimulation distances.

In sum, the data from experiment 2 lead us to conclude that there is a mechanism of compensation for distance of stimulation that works by shifting our PSS in the direction of the expected audio lag in a certain distance of stimulation. It is possible that this mechanism of compensation takes into account sound velocity in order to correct crossmodal temporal disparities. Nevertheless, based on our results, it is not easy to clearly assert that this mechanism of compensation for distance of stimulation is taking sound velocity as a model of compensation, essentially because the slopes of the data linear fits are generally lower than the theoretical slope for this model. However, we know that results in this type of experiments are usually quite variable between participants (Harris et al, 2010; Eijk et al, 2008), and that these kind of tasks are quite demanding for untrained participants. So, further studies should be made in order to clarify the magnitude of the positive correlation between PSS and stimulation distance. In our perspective, this positive relation between the PSS and distance of stimulation, along with the existence of WTIs, is what makes audiovisual perception of synchrony possible over large distances of stimulation.

4. GENERAL DISCUSSION

The two experiments presented found evidence for several perceptual phenomena generally related in scientific literature to the realm of audiovisual synchrony perception. Firstly, we did show, unequivocally, that unity assumption happens even when there is no temporal co-occurrence between two stimuli modalities. Thus, a precise temporal alignment between sound and image is not a requisite for the perception of audiovisual synchrony. Secondly, we show plenty of evidence for the vision-first bias. Participants were willingly inclined to judge as synchronic audiovisual presentations where sound was transmitted shortly after the image. Thirdly, and more importantly, this vision-first bias tends to be accentuated when distance of stimulation increases. In other words, the farther the distance of audiovisual stimulation, the higher the SOA that we judge as giving the best perception of synchrony is.

Our results corroborate the conclusions of Sugita and Suzuki (2003), Kopinska and Harris (2004), Alais and Carlile (2005) and all the other investigations that found evidence for a mechanism of compensation for differences in propagation velocity in the perception of synchrony. But the innovation that this work brings to this scientific discussion is that, for the first time, we found evidence for the existence of such a mechanism in an entire environment of VR and with highly ecological stimuli that represent a causal relation between the visual and the auditory stimulus. Despite all the scientific work developed is this area, some questions would remain unanswered without control over the number and quality of depth cues. Thus, undertaking this work in such a VR environment allowed us to manipulate the number and quality of depth cues and the results were of great relevance in two ways: Firstly, as seen in experiment 1, we found evidence that a combination of only two depth pictorial cues, such as familiar size and elevation, and two depth dynamic cues, step amplitude and angular velocity, are enough to trigger a mechanism of compensation for distance of stimulation. Secondly, as could be seen from experiment 2, evidence for the existence of this mechanism of compensation appears to be dependent on the number of depth cues presented in the audiovisual stimuli. Moreover, we obtain these results using motion stimuli that are closer to a natural situation of audiovisual stimulation than simple stationary stimuli, such as beeps and flashes, traditionally used in this kind of study. In fact, this is, to our knowledge, the first work that presents evidence for compensation for differences in propagation velocity using motion stimuli with causal relations between the two streams of stimulation. Note that the use of stimuli of this nature in a study manipulating distance was greatly facilitated by the use of a VR environment.

All in all, we think we can clearly frame our results within the notion of *simultaneity constancy* firstly introduced in the work of Kopinska and Harris (2004) and later developed by Harris et al. (2009). These authors argue that the ability to maintain the perception of synchrony over a range of different situations of stimulation (referred by them as simultaneity constancy) becomes possible trough one of two ways:

1 - *Deterministic models for simultaneity constancy*: Where signals could be resynchronized in the brain using deterministic models based on the knowledge of the factors responsible for the asynchrony. In these explanatory models, the stimuli have their delays corrected before reaching a decision center. This type of models implies changes in the timing of information as it passes through the nervous system and, consequently, this would reflect changes in the processing time of a particular sensorial modality. Furthermore, to correct these delays we should have implemented a sort of "library" containing all the sources of variations in the relative timing of two synchronic signals (with propagation velocity and transduction time certainly being the more important ones). Although these types of models appear to be good explanations for the problem of synchrony perception and, moreover, present good predictions concerning the PSS value in certain stimulation situations, they have a hard time finding neural evidences for the existence of a such resynchronization mechanism (since changes in reaction times should depend on distance of stimulation) and in dealing with the implication of having very situation-specific predictions (Harris et al., 2009). As Harris and collaborators pointed out, if we have to delay the processing time of light in order to match it with the time it that takes for a sound to propagate from a distant event, the added delay would affect all the subsequent perceptions involving the visual stimuli, unless we computed a new delay for every type and situation of stimulation. Clearly this would be quite complex and not so advantageous. There is another class of models that seems to have a good fitting to the data found in this type of experiments, while at the same time being less complex than these deterministic models.

2 – *Probability models for simultaneity constancy*: When desynchronized signals are judged as synchronized because they fall into a certain time window located at a certain temporal relation between the two stimuli. These models are probabilistic rather than deterministic and they follow several stages. The first stage is concerned with the appropriate binding of the stimuli and follows spatial, temporal and cause/effect criterion as exposed in the notion of unity assumption (**1.2** – "Theoretical Background"). In a second stage, and after the correct binding, the time differences between the modalities are acknowledged and this differences in two ways: (1) a primary one, consisting in the generation of an internal expectation of the time differences normally associated to those stimuli in that particular context (for example, a particular distance); and (2) a more operational one, where this difference is compared with the generated temporal expectancies in a certain context in order to make a decision based on a probabilistic model similar to the ones used in our data analysis (see **figure 10**). If the time disparity between signals falls within the expected time window, a decision of "synchrony" is made; otherwise, if this disparity falls outside the time window expected for that context of stimulation, a decision of "not synchrony" is made.

As Harris et al. (2010) pointed out, in these probabilistic models, "expectancies are continuously updated based on the context so that veridical and constant simultaneity can be perceived despite changes in the context" (p. 248). Using as an example the audiovisual perception of synchrony at a distance of 25 m, this model would work by comparing the delay with which sound and image were

captured with a probability function with a peak at a 75 ms delay of sound based in previous experience. Thus, SOAs closer to this peak would be increasingly evaluated as in synchrony. A change in the context of stimulation will, consequently, create a new probability function with a different peak in order to compare it with the new crossmodal disparity.

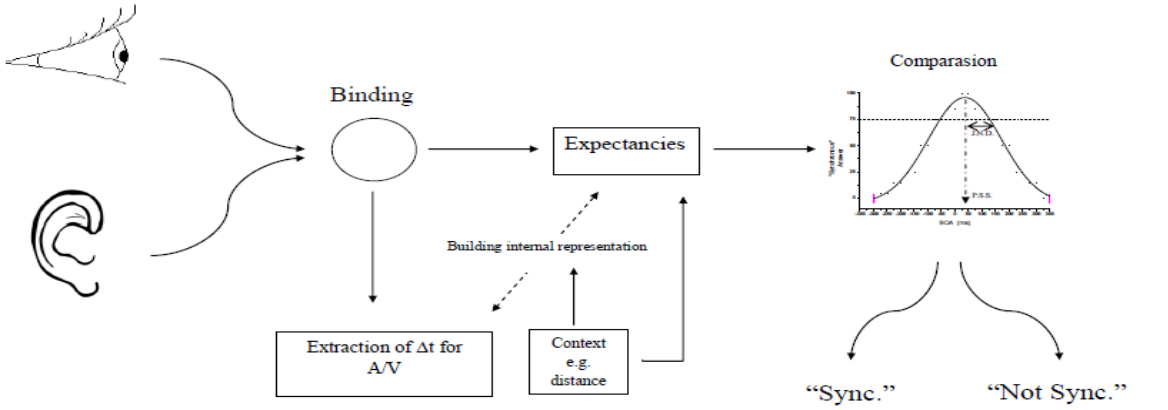


Figure 9. Probability model for simultaneity constancy as based in the model presented by Harris et al. (2009). The time difference between the two stimuli is being used to "feed" a certain temporal expectancy relative to that context of stimulation and to be used in the decision about that temporal disparity falling, or not, within the temporal window of a synchrony judgment.

This probabilistic model appears to be an explanatory scheme quite suitable to our data. The distances cues presented in our experiments (context) define a certain expectation relative to time disparities between stimuli, which are different from distance to distance. When an audiovisual stimulus is presented with a SOA closer to the temporal disparity expected in that given context, the probability of "synchrony" judgment increases.

Furthermore, we can explain PSSs under the predicted value through a model of compensation for differences in propagation velocity as a consequence of the generation of a defective expectancy caused by insufficient information relative to visual and auditory distance of stimulation. Also, this could be the reason why number and quality of depth cues were a determinant factor for the appearance of some evidence of compensation for stimulation distance in experiment 2.

In sum, with this work we think we have developed a procedure that allows us to give a clearer explanation about the role of depth cues in the triggering of a compensation mechanism. Future investigations should have an increasing number of conditions differentiated by the number and quality of depth cues presented. Something of the outmost importance is the implementation of more auditory depth cues (such as intensity decrement with distance and reflected sounds). This would make it possible to have, also with the auditory stimuli, several levels of proximity to natural conditions of stimulation (e.g., developing conditions with different numbers of sound reverberations – we know that more reverberations translate into more distance cues). Theoretically, with this kind of procedure, it would be possible to access the individual role of each depth cue, and also the effect of several combinations between them in the triggering of this compensation mechanism for distance of stimulation; in other words, if we want to frame this hypothesis in the probabilistic model of Harris et al. (2010), we could access the individual role of each depth cue, such as the effect of several

combinations between them, in the generation of the expectancy for the judgment of audiovisual synchrony. Experiments of this kind should be accompanied by previous psychophysical distance judgments of the audiovisual stimulus used in each session in order to be able to compare that judgment with the synchrony judgment and, thus, clarify even more the relation between perceived distance of stimulation and PSS.

5. CONCLUSION

Understanding the perceptual mechanisms that allow the perception of audiovisual synchrony in the real world is of fundamental importance, not only for the theoretical development in psychophysical investigation, but also, and equally important, for the development of VR environments. A satisfactory answer to the questions placed at the end of section **1.1** will be of a great scientific importance and, at the same time, allow a huge leap in the technological development of interactive VR environments and, consequently, in the development of its numerous applications, such as virtual games, 3D cinema, virtual tourism, virtual audio navigation tools and other systems integrating audiovisual stimulation. As Dias, Campos, Santos, Casaleiro, Seco & Sousa Santos, (2008) pointed out: "we are still very much in the 'silent era' of VR" (oral presentation's abstract). Great developments have been made in the generation of increasingly immersive visual interactive VR environments, but audiovisual interactive VR environments are still conceived as an enormous challenge, mostly due to ignorance about some audiovisual processes, such as the audiovisual perception of synchrony (Begault, 2000).

This work brings new evidence for the mechanisms that guide the perception of synchrony, and, at the same time, give us confidence that this could be a starting point in the development of a procedure that may allow a clear answer to the problem of audiovisual perception, thus providing a theoretical development in psychophysical investigation and a technological development in the area of interactive VR environments.

6. REFERENCES

Alais, D., & Carlile, S. (2005). Synchronizing to real events: Subjective audiovisual alignment scales with perceived auditory depth and speed of sound. *PNAS, 102*, 2244-2247.

Arnold, D. H., Johnston, A. & Nishida, S. (2005). Timing sight and sound. *Vision Research,* 45, 1275-1284.

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

Baldo, M. V. C., & Caticha, N. (2005). Computational neurobiology of the flash-lag effect. *Vision Research, 45*, 2620-2630.

Begault, D. R. (2000). *3-D sound for virtual reality and multimedia*. NASA Ames Research Center, Moffet Field California.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersllebn, T. Bachmann, & J. Müsseler (Eds.*), Cognitive contributions to the perception of spatial and temporal events* (pp. 347-362). Amsterdam: Elsevier.

Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature, 397,* 517-520

Corey, D. P., & Hudspeth, A. J., (1979). Response latency of vertebrate hair cells. *Biophysic Journal, 26,* 499-506.

Dekeyser, M., Verfaillie, K., & Vanrie J. (2002). Creating stimuli for the study of biological-motion perception. *Behavior Research Methods, Instruments, & Computers, 34 (3),* 375-382.

Dennett, D., & Kinsbourne, M. (1992). Time and the observer. *Behavioral and Brain Sciences, 15 (2),* 183-247.

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

Dias, P., Campos, G., Santos, V., Casaleiro, R., Seco, R., Sousa Santos, B. (2008). 3D reconstruction and auralization of the "painted dolmen" of antelas. In Proc. of the Electronic Imaging 2008 Conf., SPIE Vol. 6805, 6805OY, Three-Dimensional Image Capture and Applications 2008, San Jose, California, USA. 28-29 Jan. 2008.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719-721.

Engel, G. R., & Dougherty, W. G. (1971). Visual-auditory distance constancy. *Nature, 234,* 308.

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

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7,* 773-778.

Gibson, J. J. (1950). The perception of visual surfaces. *The American Journal of Psychology, 63,* 367-384.

Harris, L., Harrar, V., Philip, J., and Kopinska, A. (2009). *Space and time in perception and action*. Cambridge: Cambridge University Press.

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

Keetels, M., & Vroomen, J. (2005). The role of spatial disparity and hemifields in audio-visual temporal order judgments. *Experimental Brain Research, 167,* 635-640.

King, A. J., & Palmer, A. R., (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research, 60,* 492-500.

Kopinska, A., & Harris, L. R. (2004). Simultaneity constancy. *Perception, 33,* 1049-1060.

Lennie, P. (1981). The physiological basis of variations in visual latency. *Vision Research*, 21, 815-824.

Lewald, J., Guski, R. (2004). Auditory-visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. *Neuroscience Letters*, *357*, 119-122.

Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Psychology*, *31*, 1096-1106.

Maunsell, J. H. R., & Gibson, J. R., (1992). Visual response latencies in striate cortex of the macaque monkey. *Journal of Neurophysiology*, *68 (4)*, 1332-1344.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research*, *17*, 154-163.

Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Adaptation to audiotactile asynchrony. *Neuroscience Letters*, *413*, 72-76.

Pöppel, E. (1988). *Mindworks*. Boston: Harcourt Brace Jovanovich.

Scheier, C. R., Nijhawan, R., & Shimojo, S. (1999). Sound alters visual temporal resolution. *Investigative Ophhalmology & Visual Science*, *40*, 41-60.

Stone, J. V., Hunkin, N. M., Porril, J., Wood, R., Keeler, V., Beanland, M., Port, M., & Porter, N. M. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society*, *268*, 31-38.

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

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 and Psychophysics*, *70 (6)*, 955-968.

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

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*, 134-142.

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.

Vroomen, J., & Keetels, M. (2009). Sounds change four-dot masking. *Acta Psychologica*, *130*, 58-63.

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics*, *72 (4)*, 871-884.

Welch, R. B. (1999). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In G. Aschersllebn, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 371-387). Amsterdam: Elsevier.

Zampini, M., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics*, *67 (3)*, 531-544.