Human Movement Science 56 (2017) 71–87

Contents lists available at ScienceDirect

Human Movement Science

journal homepage: www.elsevier.com/locate/humov

Full Length Article

Audiovisual integration increases the intentional step synchronization of side-by-side walkers

Dominic Noy[a,*], Sandra Mouta[b,c], Joao Lamas[c], Daniel Basso[d], Carlos Silva[b,c,e], Jorge A. Santos[a,c,f]

[a] *Department of Basic Psychology, School of Psychology, University of Minho, Portugal*
[b] *INESC TEC, Portugal*
[c] *Center of Computer Graphics, Portugal*
[d] *Faculdade de Engenharia, Universidade do Porto, Portugal*
[e] *Department of Informatics, University of Minho, Portugal*
[f] *Centro Algoritmi, School of Engineering, University of Minho, Portugal*

ARTICLE INFO

Keywords:
Interpersonal synchronization
Audiovisual integration
Optimal integration
Side-by-side walking
Point-light walkers

ABSTRACT

When people walk side-by-side, they often synchronize their steps. To achieve this, individuals might cross-modally match audiovisual signals from the movements of the partner and kinesthetic, cutaneous, visual and auditory signals from their own movements. Because signals from different sensory systems are processed with noise and asynchronously, the challenge of the CNS is to derive the best estimate based on this conflicting information. This is currently thought to be done by a mechanism operating as a Maximum Likelihood Estimator (MLE). The present work investigated whether audiovisual signals from the partner are integrated according to MLE in order to synchronize steps during walking. Three experiments were conducted in which the sensory cues from a walking partner were virtually simulated. In Experiment 1 seven participants were instructed to synchronize with human-sized Point Light Walkers and/or footstep sounds. Results revealed highest synchronization performance with auditory and audiovisual cues. This was quantified by the time to achieve synchronization and by synchronization variability. However, this auditory dominance effect might have been due to artifacts of the setup. Therefore, in Experiment 2 human-sized virtual mannequins were implemented. Also, audiovisual stimuli were rendered in real-time and thus were synchronous and co-localized. All four participants synchronized best with audiovisual cues. For three of the four participants results point toward their optimal integration consistent with the MLE model. Experiment 3 yielded performance decrements for all three participants when the cues were incongruent. Overall, these findings suggest that individuals might optimally integrate audiovisual cues to synchronize steps during side-by-side walking.

1. Introduction

Nonverbal coordination is fundamental for interactions between individuals such as in team sports, factory work, and also in simple everyday tasks. It requires a highly precise, mutual, and accurate spatiotemporal displacement of the body.

To understand interpersonal coordination, several studies investigated the ability to synchronize repetitive movements (Repp,

[*] Corresponding author.
E-mail address: id4256@alunos.uminho.pt (D. Noy).

http://dx.doi.org/10.1016/j.humov.2017.10.007
Received 25 June 2017; Received in revised form 9 October 2017; Accepted 15 October 2017

2005; Schmidt & Richardson, 2008). Synchrony can be defined as a bounded temporal relationship (Mörtl et al., 2012) between two oscillating entities (Pikovsky, Rosenblum, & Kurths, 2001). Synchronization differs from other natural every-day coordination tasks because the possible configurations of the effector systems are spatiotemporally constrained.

The dynamical system approach explains the phenomenon by assuming the presence of attractors that emerge because the entire inter- and intrapersonal system is governed by general laws (Coey, Varlet, Schmidt, & Richardson, 2011; Demos, Chaffin, Begosh, Daniels, & Marsh, 2012; Haken, Kelso, & Bunz, 1985; Issartel, Marin, & Cadopi, 2007; Oullier, De Guzman, Jantzen, Lagarde, & Kelso, 2008; Richardson, Marsh, Isenhower, Goodman, & Schmidt, 2007; Richardson, Marsh, & Schmidt, 2005; Strogatz, 2003; Schmidt & Richardson, 2008; Schmidt & O'Brien, 1997). However, while already Christian Huygens (in 1665) referred to the physical facts of the system in order to explain clock pendulum synchronization as an exchange of mechanical energy, explanations about the specific perceptual mechanisms underlying interpersonal synchronization are still scarce (see Colling & Williamson, 2014).

Recently, movement synchronization during side-by-side walking became a scope of inquiry (Nessler, De Leone, & Gilliland, 2009; Nessler & Gilliland, 2009; Nessler & Gilliland, 2010; van Ulzen, Lamoth, Daffertshofer, Semin, & Beek, 2008; Zivotofsky & Hausdorff, 2007; Zivotofsky, Gruendlinger, & Hausdorff, 2012). For achieving movement synchronization during side-by-side walking, individuals must perceive the spatiotemporal properties of both their own and the partners' movements. If not holding hands, estimations about the current gait cycle phase of the walking partner can be retrieved from auditory and visual cues. Detailed visual cues would be provided should the individual focus on the movements of the partner. It is assumed that this does not happen constantly owing to navigation and self-motion control demands (Warren, Kay, & Yilmaz, 1996). Therefore, peripheral visual cues are more likely to be used for perceiving the motion of the partner. Also, salient auditory cues are provided by footstep sounds produced by a short-lasting large upward force on the foot at heel strike (Pastore, Flint, Gaston, & Solomon, 2008).

In order to identify where an individual is within his/her gait cycle, it may be sufficient to watch the continuous displacement of the feet. Yet, during walking, the gaze is usually directed to future foot contact locations (Lappe, Bremmer, & Van den Berg, 1999). Nonetheless, the visual system provides cues for perceiving self-motion. From this, the current gait cycle position can be retrieved (Campos & Bulthoff, 2012). The most relevant visual cue for the perception of self-motion is optic flow. However, because the movements of different body parts (eye movements and head movement) are superimposed on global body displacement, optic flow cues have to be combined with other signals to robustly disambiguate the flow patterns (Lappe et al., 1999).

Cue integration seems to disambiguate information provided by single modalities (Cullen, 2012; de Winkel, Weesie, Werkhoven, & Groen, 2010). It was suggested that neural ensembles within the vestibular nucleus code self-motion by integrating the input from multiple afferent information (vestibular, visual, proprioceptive, somatosensory) and efference copies of the motor commands (Cullen, 2012; Fitzpatrick, Wardman, & Taylor, 1999). Similarly, besides visual self-motion inputs, the integration of proprioceptive and vestibular signals disambiguates information about the dynamics and kinematics of the body segments, and the integration of proprioceptive and cutaneous signals disambiguates information about the position of segments, relative to each other and relative to the surface (seeKaya, 2014).

Consequently, gait synchronization can be understood as cross-modally matching audiovisual signals from the partner with kinesthetic, cutaneous, visual, and auditory signals from their own movements. However, there is noise within each sensory system (Ernst & Bülthoff, 2004). Another caveat might be that signals from different sensory systems arrive asynchronously at higher processing levels due to different propagation, transduction, transmission, and processing times (Repp, 2005; Vroomen & Keetels, 2010). Furthermore, environmental cues might be ambiguous and more or less regular and accessible (see e.g., Cullen, 2012; Hartmann, 1983; Kopčo & Shinn-Cunningham, 2011; Kolarik, Moore, Zahorik, Cirstea, & Pardhan, 2016).

It is currently thought that the CNS derives the best output based on noisy and conflicting information by a mechanism operating as a Maximum Likelihood Estimator (MLE) (see Bayesian Optimal Integration Theory Alais & Burr, 2004; Ernst & Bülthoff, 2004; Ernst, 2006). According to MLE, sensory signals are optimally integrated by the assignment of weights. These are determined by how reliably each signal can represent the cue in question.

While there are many studies providing evidence for optimal integration mechanism underlying perceptual judgments and sensorimotor synchronization with simple rhythmic stimuli (Elliott, Wing, & Welchman, 2010; Wing, Doumas, & Welchman, 2010; Wright & Elliott, 2014; Sejdić, Fu, Pak, Fairley, & Chau, 2012), the contribution of different sensory systems in synchronization of side-by-side walkers is scarce and ambiguous (Nessler & Gilliland, 2009; Zivotofsky et al., 2012).

In former studies single modalities like visual, auditory, and haptic were masked, but it is not clear how effective the masking technique was or if other spatiotemporal cues were available. For instances: in Nessler and Gilliland's study (2009) vibrations from the surface floor might have been sensed and ear plugs did not prevent sounds produced by the walking partner entirely. Also, treadmill walking reduces optic flow and provides additional pacing cues (Zivotofsky et al., 2012).

Behavioral studies revealed a strong enhancement when complex, moving, and biological stimuli were presented through multiple modalities (Brooks et al., 2007; Wuerger, Meyer, Hofbauer, Zetzsche, & Schill, 2010; Arrighi, Marini, & Burr, 2009; Thomas & Shiffrar, 2013; Saygin, Driver, & de Sa, 2008; Bidet-Caulet, Voisin, Bertrand, & Fonlupt, 2005). This might be due to an improved disambiguation when using cues from multiple modalities to represent the object in question. Furthermore, neuroimaging studies showed that there are overlapping areas for the processing of biological stimuli and multimodal signals (seeThomas & Shiffrar, 2010; Grossman, Blake, & Kim, 2004; Barraclough, Xiao, Baker, Oram, & Perrett, 2005; Beauchamp, Lee, Haxby, & Martin, 2002; Peuskens, Vanrie, Verfaillie, & Orban, 2005; Saygin et al., 2008). Therefore, it is quite surprising that, to our knowledge, there are no attempts to test for the effects of multimodal integration when synchronizing with complex biological stimuli in a task such as side-by-side walking.

Of course, the control of all relevant variables might be quite difficult when merging multimodal and biological stimuli in a complex perceptual-motor task. In addition, the manipulation of sensory channels may cause unwanted side effects on relevant processes during walking. For example: manipulating the vestibular system, the proprioceptive system, or the somatosensory system, affects functions involved in maintaining stability and inter-limb coordination (Cullen, 2012; Cullen, 2016; Prochazka & Ellaway, 2012; Prochazka, Gritsenko, & Yakovenko, 2002; Ghez & Krakauer, 2000; Kaya, 2014). Also, the use of side-blinders constrains optic flow and therefore the perception of self-motion (Campos & Bulthoff, 2012; Lappe et al., 1999).

Taking these challenges into account, this is the first attempt to understand the sensory integration in movement synchronization of side-by-side walkers. In three experiments, the perceptual cues (visual and auditory) from a walking partner were virtually simulated. Given that auditory and visual cues inform about the time of the upcoming heel strike of a walking partner, it seems likely that the CNS integrates signals of both modalities to obtain the best estimate of the temporal onsets. We hypothesized that this should then improve synchronization. In Experiment 1, human real-sized Point Light Walkers (PLWs) obtained from the motion capture of walking individuals were used. Standard PLWs contain the spatiotemporal components of human motion. Their implementation allows a controlled manipulation of these components. Previously, it was shown that PLWs are an adequate mean for the study of intermodal perceptual processes as for instances recognition (see e.g., Thomas & Shiffrar, 2013), velocity, and simultaneity judgments (Mendonça, Santos, & López-Moliner, 2011; Silva et al., 2013). In this experiment, participants walked next to PLWs and were instructed to synchronize. Auditory, visual, or audiovisual cues about the PLW were provided. In Experiment 2, the virtual environment was improved by using a human-sized virtual mannequin as visual stimulus and stimuli positions were rendered in real-time. Thus, the spatiotemporal congruency of audiovisual cues was increased. In Experiment 3 the audiovisual asynchrony was manipulated and its effect on motor synchronization performance was analyzed.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Eight participants (7 naïve, 2 female, 6 male, age: $M = 28$, $SD = 3$, all right dominant hand) without gait disabilities took part in the experiment. All individuals gave informed consent for their participation.

2.1.2. Stimuli & experimental design

The experiment was programmed in Python (Python, 2016) using OpenGL for graphics presentation and OpenAL for audio playback.

2.1.2.1. Visual stimulus. Body kinematics during walking with different velocities (0.7–1.5 m/s) of 6 male and 8 female were previously captured with a Vicon motion capture system at 240 Hz. From eight gait cycles of these models, PLWs—with 13 dots generated in 2D coordinates and rendered as black dots—served as the visual stimulus. The 13 dots signalized the spatiotemporal positions during walking of the head, shoulders, elbows, hands, hips, knees, and ankles. The PLWs were presented by 3 DLP projectors Christie Mirage with a resolution per channel of 1400 × 1050 at 60 Hz. The three images were retro-projected side-by-side with blending areas between images resulting in a 7.20 m(H) × 2.10 m(V) flat screen in a dark room. The PLW dots were black (4 cd/m2) and the background was light gray (70 cd/m2). The PLWs maintained its original sizes and were projected in the sagittal plane walking from one side of the screen to the other one.

2.1.2.2. Auditory stimulus. The auditory stimulus was a footstep recorded from an individual with average stature in Portugal (1.62 m), including male and female population (Arezes, Barroso, Cordeiro, Costa, & Miguel, 2006) walking on a wooden floor barefoot with a velocity of 1.3 m/s. From these records, two footsteps were auralized by a MATLAB routine with head-related transfer functions (HRTF). These had approximately the acoustic properties of the sounds that reach the ear when produced by an individual walking next to the participant at 0.5 m. (Left foot: azimuth 90 degrees, elevation −72 degrees, relative to the right ear; right foot: azimuth 90 degrees, elevation −62 degrees). The intensity was matched to the recorded sound intensity at an average ear height (1.53 m). This was for the closer left foot 63 (dBA Leq) and for the right foot slightly lower. The footsteps were presented through wireless headphones (Sennheiser RS 120 II).

2.1.2.3. Audiovisual stimulus. The audiovisual stimulus was the PLW presented with the sound produced by the heel strikes of the PLW. In order to assure synchrony between visual and auditory stimulus, the time delay between a sound stimulus and a visual flash was measured (Lamas et al., 2015). A delay of 15 ms, with a precision of 3 ms, was applied to the auditory signal for achieving the correct temporal alignment taking into account audiovisual signal propagation times.

2.1.2.4. Conditions. The availability of sensory information and the start phase of the stimulus were manipulated to create the following conditions: for the sensory information, (1) the PLW was displayed temporally aligned with the presentation of footstep sounds (audiovisual condition – AV), (2) only the PLW was displayed (visual condition – V), or (3) only the footstep sounds were presented (auditory condition – A); for the start phase, the PLW/footsteps started (1) in midstance, (2) in midswing, or (3) with a heel strike of the left foot. These three (Sensory Information) x three (Start Phase) levels were combined constituting nine conditions. Each condition was repeated three times in three blocks making up 81 trials for each participant. In each block, the 27 trials were pseudo-randomly presented.

2.1.3. Procedure

The experiment was conducted at the Laboratory of Visualization and Perception of the University of Minho and Center of Computer Graphics. Prior to the experiment, the participants walked on a short walkway—a 13.50 m × 0.92 m wooden floor—in order to determine the comfortable walking velocity. Instructions were given to "walk comfortably but not too slow; walk as if you were walking with a friend". The participants walked as many times as needed until the velocity of three subsequent walks did not deviate more than 5% from the mean velocity of the three walks.

Then, one PLW was chosen according to gender, hip height (max. difference = 1.9 cm), and the comfortable velocity (max. difference = 0.1 m/s) to match relevant gait characteristics. The participants wore shorts and walked barefoot on the walkway, which was located next to the screen on which the PLW was displayed. Participants started walking 2.4 m before the screen began and stopped 3.2 m after the screen end. Two reflective markers were attached to the malleolus of the ankle of the participant and four markers were attached to the head. Marker positions were captured at 240 Hz by a Vicon motion capture system with 6 near-infrared cameras (MX F20 of 2 megapixels) and defined in a xyz-Euclidian frame. The participant walked alongside the projected stimuli for 7.20 m. Ten to 12 steps were registered. Participants were instructed to "Walk without interruption and do not reduce velocity until the walkway end. When the PLW is displayed, synchronize steps and maintain position at the side of the PLW. When not displayed, synchronize with the auditory footsteps. When PLW and footsteps are presented, synchronize with both and maintain the smallest distance as possible to the PLW". No instructions were given about gaze direction.

2.2. Analysis

All analyses and statistical inference were conducted with R Studio version 0.98. For each walk, 10 heel strikes were identified by the vertical displacement of the ankle of the participant (see Fig. A.3). The difference between the onsets of the heel strike of the participant was subtracted from the onsets of the heel strike of the PLW to compute the temporal asynchrony. A within-subject design (Steps [10] × Sensory [3] × Start [3] × Rep [3] × Block [3]) was used. Therefore, a maximum of 810 asynchrony records were obtained for each participant. Two participants were excluded because they were not able to correctly perform the synchronization task. Participants adapted quickly to the cadence and velocity of the stimulus so that step frequency was approximately matched since the start. This happened in 91% (D), 88% (E), 71% (B), 83% (C), 91% (J), and 89% (S) of all trials. For this reason, frequency matching was less relevant.

Synchronization was assessed by considering the time-dependent behavior of the asynchronies within each trial and the variability of the asynchronies between trials. However, it is important to consider that synchronization could be a transitory phenomenon (see van Ulzen et al., 2008) so that short walking samples might fail to capture it. Fig. A.4 shows the asynchronies as a function of step number for each individual and sensory condition. The plot illustrates that most asynchrony series converge logarithmically to a particular value. At steps ⩾6 the slopes of these converging curves approach zero. Fig. A.5 shows the asynchrony variability. Similarly, the asynchrony variability seems to stabilize logarithmically at steps ⩾6. The observations of Figs. A.4 and A.5 suggest that the participants followed the instructions and minimized relative asynchronies, and that once being minimized there were no later transitions to less synchronized states. Note that the above-described pattern is far more consistent in Experiment 2 compared to Experiment 1. This probably owes to artifacts of the setup in the first experiment, as discussed in Section 2.4.

The stabilization of asynchrony magnitude and variability indicates that participants did not attempt to further minimize the asynchrony. Assuming that participants were motivated to do so, this pattern suggests that at later steps (a) the asynchronies were perceived as synchronous, or (b) that the perception–action system was unable to further reduce the asynchronies. Nevertheless, it is plausible to assume that the participants attempted to get the asynchronies close to the converging point. This point was therefore interpreted as minimal achievable asynchrony (see alsoSemjen, Vorberg, & Schulze, 1998). Trials were considered outliers and excluded when the asynchronies at the last step did exceed 3 s.d. from this converging point. Asynchronies close to the converging point were framed synchronous.

Thus, whether an individual could maximize synchrony depended on the computation of the asynchrony and the selection and execution of adequate motor commands to reduce it. We expected that this strategy should be time dependent and that synchrony should become maximal when a certain temporal threshold is reached. Considering that the underlying processes operate under noisy conditions (sensory noise, motor noise, etc.), synchrony can be formalized as a random variable. Thus, the probability of observing more synchronized steps should increase with time. Then, the probability of synchronized steps can be represented by the proportion of synchronized steps.

Note that depending on the context, asynchronies can be represented as (a) the phase difference of two points in their cycles expressed in degrees or (b) its temporal separation expressed in milliseconds. While the former is normalized by the cycle interval, the interpretation of the latter is more intuitive. Both measures were used to analyze the results depending on the particular question on hand.

Firstly, to obtain the proportions of synchronized steps, asynchronies were (1) transformed into Discrete Relative Phase (DRP):

$$\mathrm{DRP_j} = \frac{\mathrm{t_{PLW,j}} - \mathrm{t_{participant,j}}}{\mathrm{t_{PLW,(j+1)}} - \mathrm{t_{PLW,j}}} 360,$$

where *t* is the onset of the *j* heel strike. The mean DRP at Step 10 (i.e., the last step that was considered) served as the best estimate of the converging point of each sensory condition. We considered DRPs as synchronous when falling into a 20 degrees interval around this point (see Nessler & Gilliland, 2009). Trials that were synchronized since start were excluded. After data exclusion, there were for each individual at least 80 asynchronies in each sensory condition (8 for each of 10 steps). Note that in most conditions it were registered many more asynchronies (up to 27). Secondly, proportions of synchronized steps were obtained by dividing the number of DRPs within the 20 degrees interval around the converging point by the total number of DRPs. A proportion was calculated for 5-time intervals ranging from 0s to 5s.

2.3. Results

The aim of this experiment was to verify if synchronization improves over time and whether this process differs depending on the available sensory cues. In order to model the improvement, we fitted cumulative normal distribution functions (cumulative Gaussians) to the proportions of steps considered as synchronous. We assumed that individuals attempted to minimize the asynchrony as fast as possible. This moment was quantified by the point at which the cumulative Gaussian reaches 75%. This "threshold" ($P[X \geqslant T] = 0.75$) represented the time at which 75% of the steps were synchronized (T). Here, T was interpreted as the time that is required to achieve synchronization.

In addition, once a participant reached the synchronization threshold, the performance could be further assessed through asynchrony variability. Assuming that audiovisual integration should improve the precision of estimates of the heel strike onsets, integration processes might be manifested by lower variability. Thus, we captured two different aspects of the synchronization process. The first was quantified by T and was interpreted as the time that is required to synchronize. The second quantified by the s.d. of asynchronies reflected synchronization precision.

Qualitative assessment of the fitted Gaussians indicated an advantage of AV and A over V (Fig. 1a left). There might also have been a small advantage of AV over A for 5 out of 6 participants. The curves could be fully described by two parameters: (A) the slope of the cumulative Gaussian (= the SD of the Gaussian distribution) indicates the form. It is frequently used in psychophysical studies to quantify perceptual sensitivity (see e.g., Mendonça et al., 2011). Here, such variable is not relevant. It reflects how quickly individuals switched from non-synchronized to synchronized states while we were interested in how quickly individuals synchronized overall. (B) Any quartile of the curve ($P[X \geqslant T] = q$) reflects the location on the abscissa at which q% of synchronized steps were reached. We calculated ($P[X \geqslant T] = 0.75$) and interpreted it as synchronization threshold. T suggests that 5 of 6 individuals were faster synchronized in A than in V (Fig. 1b left). Three of 6 individuals were faster in AV than in A. Two individuals synchronized similarly in A and AV and for 1 individual performance in A was superior. Thus, considering the proportions obtained from pooled DRPs, there was a slight advantage in AV compared to A (Fig. 1b left). Overall, there were large variations within and among the individuals.

Next, s.d. of asynchronies was calculated from all steps occurring after time T (Fig. 1c left).

To test the MLE Model, s.d. of asynchronies in AV was predicted by

$$\hat{\sigma}_{MLE} = \sqrt{\frac{\hat{\sigma}_A^2 \hat{\sigma}_V^2}{\hat{\sigma}_A^2 + \hat{\sigma}_V^2}},$$

where $\hat{\sigma}_{MLE}$ is estimated by the s.d. of asynchronies in the auditory and the visual condition. S.d. of asynchronies obtained from the observations did not show a consistent pattern. The MLE predictions (M in Fig. 1c left) failed for all but 1 participant as indicated by 95% confidence intervals.

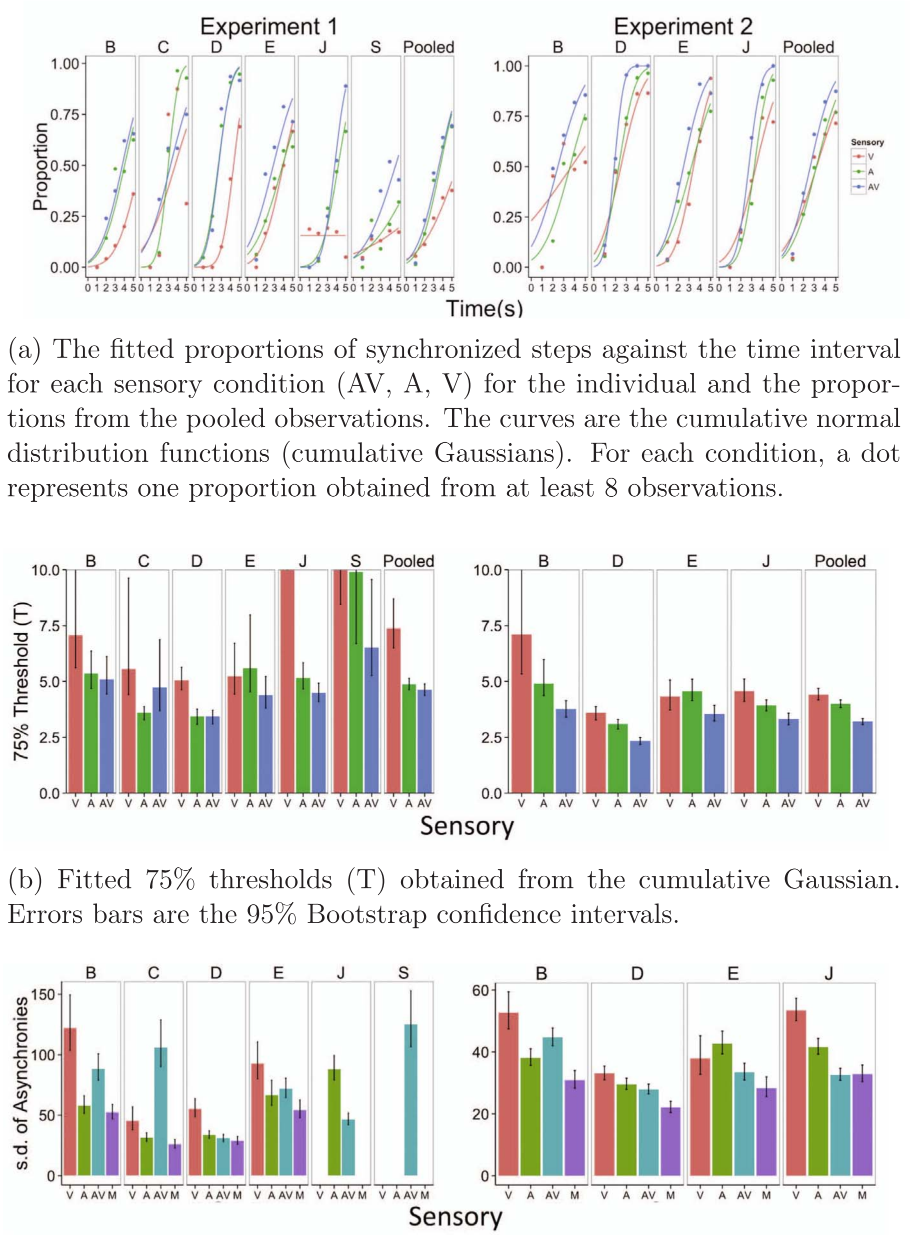
2.4. Discussion

In this first experiment the focus was on performance differences during side-by-side walking when the available cues from the walking partners were auditory, visual, or audiovisual. We treated synchronization as a random variable and assumed that the probability to be synchronized increases with time. Cumulative Gaussians were used in order to describe this synchronization effect. From the models, we estimated the time required to synchronize. As the second indicator of synchronization, we measured s.d. of asynchronies directly from the observations. We expected that audiovisual cues lead both to faster synchronization and to reduce variability.

Considering the time to synchronization, results revealed that participants were minimally faster with audiovisual cues compared to auditory cues. The absence of a clearer bimodal advantage contradicts the prediction of the MLE theory (see Bayesian Optimal Integration Hypothesis, Alais & Burr, 2004; Ernst & Bülthoff, 2004; Hove, Iversen, Zhang, & Repp, 2013). These outcomes might be interpreted as a result of auditory dominance caused by a superior temporal processing. It would result in a higher reliability of the estimation of temporal cue onsets (see Modality Appropriateness Hypothesis, Welch, DutionHurt, & Warren, 1986).

According to MLE, the variability in estimating the onsets by multiple sensory cues is always lower than when using individual cues. Thus, even highly unreliable cues should positively contribute to the final estimate. Yet, in this first experiment there was one participant who was slower with audiovisual cues than with auditory cues alone. This fact suggests that audiovisual cues can be distracting or more demanding. Such assumption is strongly supported by the synchronization variability (s.d. of asynchronies) showing extensive deviations from the MLE predictions for all but two participants (D & J).

However, artifacts of the auditory and visual cues in this experiment might have promoted the above-mentioned biases. PLWs preserve biomechanical and spatiotemporal properties of a walking person. They are appropriate for the study of, for example, recognition, velocity, and simultaneity judgments in focal vision (Mendonça et al., 2011; Silva et al., 2013). Yet, a crucial ability for PLW recognition is the extraction of structure from motion (Troje, 2008). Such ability was shown to be affected by eccentricity.



(a) The fitted proportions of synchronized steps against the time interval for each sensory condition (AV, A, V) for the individual and the proportions from the pooled observations. The curves are the cumulative normal distribution functions (cumulative Gaussians). For each condition, a dot represents one proportion obtained from at least 8 observations.

(b) Fitted 75% thresholds (T) obtained from the cumulative Gaussian. Errors bars are the 95% Bootstrap confidence intervals.

(c) S.d. of asynchronies obtained from the empirical observations. S.d. was calculated for asynchronies (in ms) of all steps occurring after time (T). Note that some conditions are lacking because in these cases T was estimated as being out of time range. M are the predictions of the MLE model. Error bars represent the 95% confidence intervals (CI).

Fig. 1. Synchronization performance with auditory, visual, & audiovisual cues for Experiment 1 (left) and Experiment 2 (right).

Eccentricity is clearly related to the decreasing capability to resolve stimulus details when the distance from the fovea increases (Gurnsey, Poirier, Bluett, & Leibov, 2006).

Consistently, studies revealed perceptual deficits of PLWs when the eccentricity of PLWs was increased (Ikeda, Blake, & Watanabe, 2005), at least when the PLW was not magnified appropriately (Gurnsey, Roddy, Ouhnana, & Troje, 2008). Note that we did not instruct gaze direction in order to maintain the paradigm as natural as possible. Nonetheless, in a control analysis we did not find any relation between head rotation and synchronization performance.

Moreover, in this first experiment the image did not change perspective and the sound source location was fixed to the lateral right side of the participant. This implies that (a) the visual perspective and distant cues did not change as a function of the relative position of the participant and (b) sound distance cues did not change at all. In addition, when for instances the PLW walked faster than the participant, the auditory delay should increase more than the visual delay due to the slower propagation of sound. Such real-time alignment of spatiotemporal signals did not happen. Incongruent signals can lead to the perception of asynchronies (Silva et al., 2013). This, in turn, should impair the integration of audiovisual signals (Spence, 2011) and therefore compromises the precision and accuracy of synchronization performance.

In sum, PLWs have been the standard stimuli in a wide range of perceptual experiments but might not be suitable for side-by-side walking studies. Also, audiovisual incongruencies might have impaired the synchronization performance. In Experiment 2, audiovisual cues were spatiotemporally congruent and changed in real-time as a function of participant behavior. In addition, further visual input was provided by replacing the PLW with a virtual mannequin stimulus.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Four individuals (all male, $M = 29$, $SD = 2.3$) from Experiment 1 participated in Experiment 2. The experiments were separated by 7 to 9 month. Pilot testing with both setups with one "control" participant (see Fig. A.6) and analyses of sequence effects (Fig. A.7) did not indicate any significant synchronization improvement through previous training.

3.1.2. Material and stimulus

The experiment was programmed in Python (Python, 2016) and Blender's logic bricks. BlenderVR 2.73 (Katz, Flinto, Tourain, Poirier-Quinot, & Bourdot, 2015) was used to coordinate and distribute the execution of the virtual environment.

A virtual mannequin (see Fig. A.8) was created to be used as the visual stimulus. The spatiotemporal coordinates from the PLWs were used to determine the joint positions of the mannequin. The joints were connected by skin-colored cylinders, with relative sizes approximately proportional to the morphological dimensions. Because in Experiment 1 the joint position and the head were represented by small black dots and the virtual mannequin was built upon these dots, the virtual mannequin was larger. Increasing its size is one mean by which to magnify retinal stimulation so that visual discrimination performance becomes equal across the entire visual field (Gurnsey et al., 2008). Thus, the larger virtual mannequin increased the sensitivity for perceiving visual cues. Moreover, like in a real-world scenario, body segments that were closer to the participant occluded segments that were farther away. This provided additional depth cues.

The perspective for the projection of the mannequin and the sound source location were computed in real-time based on the relative position and the head rotation of the participant. To do so, the head coordinates of the participant were tracked by a Vicon motion capture system using Nexus 2.0 (Vic, 2016). In order to synthesize sound properties, an auralization process using non-individualized HRTFs was used from Oliveira et al., 2013 that included a simplified geometrical model of the experimental environment (e.g., reflections, distance, & latencies). To prevent delays during online auralization, sound samples for 450 different positions—5 distances relative to participant (−100 cm, −50 cm, 0 cm, +50 cm, +100 cm)∗90 head orientations of participants—were previously created and the appropriate ones played during the trial. In sum, both visual and auditory signals provided additional distance cues and an improved spatiotemporal congruency was achieved.

However, as expected in an immersive virtual environment, end-to-end system delays did occur from the motion capture to the update of stimulus presentation. The latencies for changing the perspective of the mannequin were of 93 ms (4*SD*) and of 50 ms (1*SD*) for the sound of footsteps. Therefore, to preserve the congruency in the audiovisual condition, a further delay of 43 ms was applied to the sound signals in A and AV condition.

3.1.3. Design and procedure

Before the experiment participants were trained in order to acquaint to location cues of the auralized sound. Samples were presented at azimuth (30 degrees, 90 degrees, & 150 degrees) and elevation −72 degrees, relative to the position of the right ear.

The experimental design was the same as Experiment 1, i.e. three different start phases of the stimulus and three sensory conditions, i.e. footstep sounds (A), virtual mannequin (V), or combined (AV). Each condition was presented three times in three blocks repeated in two sessions. This constituted 162 trials. The presentation was pseudo-randomized within a block. We expected an improvement in V due to the richer visual cues. Performance in A should increase owing to spatiotemporal correspondence. Finally, AV should lead to higher performance than A or V alone due to the integration of both signals, as predicted by MLE.

3.2. Analysis

A within-subject design (Steps [10] x Sensory [3] x Start [3] x Rep [3] x Block [3] x Sessions [2]) was used. It provided 1620 records of asynchronies for each participant. The exclusion criteria were the same as in Experiment 1. The stepping frequency was matched since the start for 100% (D), 96% (B), 54% (E), and 90% (C) of the trials.

3.3. Results

As previously, in order to capture performance differences in synchronization, we (a) obtained the 75% threshold (T) from the cumulative Gaussian functions and (b) calculated the standard deviation (s.d.) of asynchronies.

Considering the T, 3 of 4 participants were faster in A than in V (Fig. 1a right & b right). All participants were faster in AV than in V and A (for pooled observations: V $[T = 4.4s] >$ A $[T = 4s] >$ AV $[T = 3.2s]$; A-AV: Bootstrap: $p < .001$). The pooled observations illustrate the improvements in synchronization compared to Experiment 1 (Fig. 1 left) for all sensory conditions. This is indicated by the non-overlapping 95% Bootstrap confidence intervals of the means (Experiment 1: V $[T = 7.38] >$ A $[T = 4.87] >$ AV $[T = 4.62]$; Experiment 1 - Experiment 2, Bootstrap: $p < .001$).

For Experiment 2, the s.d. of asynchronies in AV was lower than in A and V and in A it was lower than in V for 3 of 4 participants (Fig. 1c right). M represents the prediction of the MLE. It correctly predicted s.d. reduction in AV for 2 (J & E) of 4 individuals. In addition, it pointed toward the correct direction for another individual (D), but here confidence intervals of AV and M did not overlap. All sensory conditions showed reduced s.d. of asynchronies compared to Experiment 1.

3.4. Discussion

In Experiment 2, a virtual mannequin substituted the PLW and both visual and auditory stimuli locations and perspective were updated in real-time depending on the head coordinates of the participant. First, the modifications of the stimuli in Experiment 2 increased synchronization performance compared to Experiment 1. All four individuals improved in synchronization velocity and variability. Second, in Experiment 2 all participants synchronized faster with audiovisual cues. Although the audiovisual advantage is consistent with the optimal integration theory, estimates of the time required to achieve synchronization (T) by an MLE model is not meaningful here.

According to MLE, the effects of cue integration should be manifested in an optimal reduction of variability of the sensory representation (Ernst & Bülthoff, 2004). This, in turn, should lead to more precise timing (Elliott et al., 2010). Consistently, three of four participants synchronized more precisely when audiovisual cues were provided. On the other hand, for only two of four individuals the MLE estimates matched the asynchrony variability of the audiovisual condition.

These inconsistencies might results from methodological shortcomings. In order to maintain sensory inputs as natural as possible, we investigated over-ground rather than treadmill walking. Because we conducted the experiments in a virtual environment, the walking distance was constrained to 7.2 m (Experiment 1 and 2). It could be that measures of variability were affected by the reduced number of steps. However, Fig. A.5 shows that the asynchrony variability consistently stabilized at minimal values after 5 steps. This suggests that 10 steps might be sufficient to maximize synchrony.

Since the PLW/mannequin was different across participants, another possible limitation is that some participants were trying to match signals that were more variable than that of other participants. Yet, the models, from which the stimuli were generated, were able to walk with an extremely constant pace (this was actually a model selection criteria). The step interval variances of all employed models were of 1 ms (D & E), 3.7 ms (B), 3.4 ms (C), 1.6 ms (J) and, 2 ms (S). In addition, we did not find any relation between these variabilities and the synchronization results. Although this does not rule out that the variability of other body segments may have produced some noise, it indicates that stimulus variability might have been a less relevant noise factor.

A more plausible explanation for the inconsistent results is that the MLE model implemented here does not allow the best fit. Asynchrony variability was computed from the steps. It specified the weights for each modality in the MLE model ($\hat{\sigma}_A$ & $\hat{\sigma}_V$). This model predicted then the variability of asynchronies with audiovisual cues ($\hat{\sigma}_{MLE}$). Thus, predictions were based on the asynchrony variability when the participants were trying to synchronize with visual cues and auditory cues alone.

However, for estimating the temporal onsets of their own heel strikes, audiovisual cues might be insufficient. Other relevant cues are provided by the vestibular system, the proprioceptive system, and the somatosensory system. While each system might provide ambiguous spatiotemporal cues, their combination should allow much less ambiguous estimates of the heel strike onsets. Thus, an adequate MLE model should include parameters of the reliability of estimates with each and all of these cues within a crossmodal framework. In short, the MLE model used here did not account for all the perceptual variables involved in the estimation process.

In addition, the parameter estimates of the MLE were obtained from the variability of the observed asynchronies. Asynchronies were computed from the temporal differences between the motor responses (i.e., stepping pattern). Different sub-processes within the perception–action loop could cause variability of a motor response. Variability may be inherent to the encoding of the events. This is modality dependent. But it also can be caused by the time-keeping of temporal intervals or the motor response implementation. The two latter processes are less dependent on the modality than the former process. Therefore, specifying parameters of the MLE model by the variability that is only caused by perceptual processes might lead to an overestimation of variance reduction by a MLE model of perception–action loops (Elliott et al., 2010).

Nevertheless, variability reduction was overestimated only for two participants (B & D). In addition, for one participant (B), variability was lower for auditory cues when compared to audiovisual cues. The fact that an unimodal cue condition revealed lower variability indicates that also in Experiment 2 the combination of bimodal cues may have caused some distraction or additional load. The spatiotemporal congruency of information provided by cues on each individual modality cue and/or combined was increased in Experiment 2 compared to Experiment 1. Yet, the visual perspective and the position of footsteps sound were updated with a delay of $\sim 90ms$. Participants frequently rotated the head to the right side up to 70 degrees relative to the walking direction. This mostly happened in conditions in which the visual cues were available. For the visual condition, this implies that the perspective was updated with a delay. This might have produced some marginal noise. However, for the audiovisual condition, a fast head rotation with delayed updates might have promoted incongruence between auditory and visual cues. During head rotation, sound cues indicated the source as lateral to the ear instead of being congruent with the visual stimulus by appearing slightly in front of the right ear. The rotation lasted only a few tenths of seconds, but it might have been sufficient to create additional noise during the heel strikes.

In conclusion, the present experiment clearly demonstrates that higher synchronization precision is achieved by the combined presentation of congruent audiovisual cues compared to auditory or visual cues alone. The MLE model suggests that this occurs because cues are integrated. Our results confirmed partly these assumptions but there were several sources of noise that prevent more robust conclusions.

The integration of signals is of advantage when they are coming from the same event. Signals are not integrated when the cues indicate a temporal separation between the events (Berniker & Kording, 2011) in order to prevent the erroneous integration of cues from different sources (Elliott, Wing, & Welchman, 2014). The maximal temporal separation at which signals are integrated is called window of temporal integration (WTI) (Vroomen & Keetels, 2010). To examine whether the benefits of audiovisual cues were promoted by their integration, we conducted a third experiment. In Experiment 3 the visual and the auditory stimuli were presented with different levels of temporal onsets. Conditions in which the temporal asynchronies between auditory and visual signals were small should reveal lower synchronization variability compared to conditions in which the temporal asynchronies were large.

4. Experiment 3

4.1. Methods

4.1.1. Participants

Three of the four individuals of Experiment 1 and 2 participated in Experiment 3 (all male, age: $M = 29$, $SD = 2$). The Experiment was conducted three months after Experiment 2 and pilot tests showed no noticeable training effects.

4.1.2. Material, stimuli, & design

The available auditory and visual cues were the same as in Experiment 2 but the footstep sounds (A) and the virtual mannequin (V) were presented throughout all trials (AV). Here, A or V were temporally phase shifted. That is, the heel strikes of the virtual mannequin and the footstep sounds were displayed with disparate temporal onsets. Phase Shifts were (a) in V or in A, (b) positive or negative, and the amount of phase shift ranged from −250 ms to +250 ms in 50 ms intervals (i.e., −250, −200, −150, −100, −50, 0, +50, +100, +150, +200, +250), constituting 23 conditions. Negative values signified that the shifted stimulus was presented earlier than the non-shifted stimulus; for positive values it was the opposite (see Fig. A.9). Each condition was repeated 10 times making up 230 trials presented in a pseudo-random order.

4.1.3. Procedure

As in Experiment 1 and 2, the participant started walking from 2.4 m before the screen. Then, when passing a threshold of 0.3 m before screen start, the AV stimulus was presented spatiotemporally synchronous for 3.3 m. In this way, the participant had approximately five steps to get synchronized. Figs. A.4 and A.5 illustrate that this should be sufficient. When the participant passed the 3.6 m threshold, a phase shift was applied to V or A according to a predefined value. As a control, one synchronous condition (0 ms phase shift) was included. In order to remove artifacts, the stimuli were occluded immediately before phase shift. That is, when the virtual trajectory of the stimulus passed 3.3 m, the footstep sound disappeared for one step and the virtual mannequin disappeared behind a green square of 0.6 m × 2.1 m. During occlusion, the phase shift was applied and then the virtual mannequin re-appeared from behind the square and the footsteps sounds were presented again for 3.3 m. Participants were instructed to "synchronize steps with the mannequin and the footsteps sounds and maintain the smallest distance to the stimuli. If footstep sounds and mannequin are asynchronous, synchronize with which you feel more comfortable at that moment".

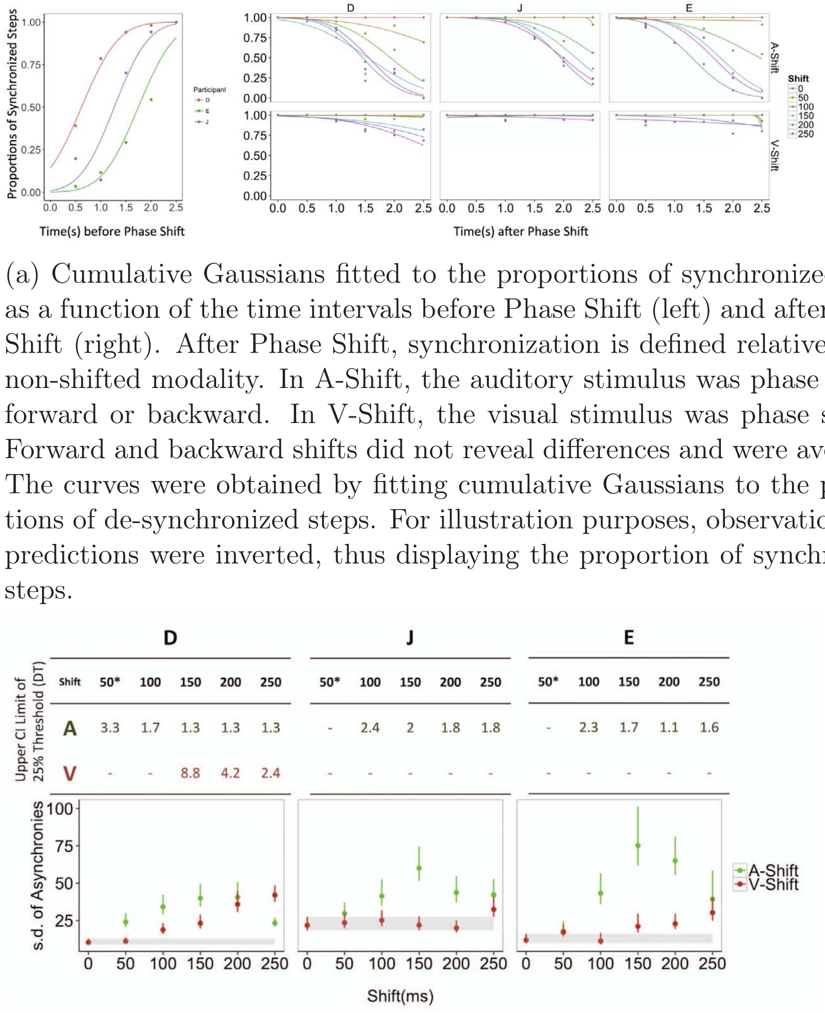
4.2. Analysis

The walkway was 0.6 m larger compared to the previous experiments. As in Experiment 1 and Experiment 2, in Experiment 3 were analyzed 10 steps (~5 s); 5 steps (~2.5 s) before phase shift and 5 steps (~2.5 s) after phase shift.

We determined the effects of phase shift of auditory (A) and visual (V) cues on synchronization performance by comparing it with the performance before phase shift. For this reason, it first had to be determined whether synchronization was achieved before phase shift. The converging point was estimated by calculating the average of the DRP at the last step before phase shift. Each trial in which the DRP at the last step deviated less than ± 20 degrees from the converging point was considered a "synchronization succeeded" trial and was included in the further analysis. This was 100% (D), 97% (J), and 89% (E) of all trials.

4.3. Results & discussion

Experiment 3 was conducted in order to examine whether temporal coincidence of audiovisual cues improves synchronization through their integration. Temporally incongruent cues should lead to increased synchronization variability because their temporal separation prevents cue integration (Elliott et al., 2014; Ernst & Bülthoff, 2004). Note that the methodological shortcomings discussed in Experiment 2 might also account for some variability in Experiment 3 since the same stimuli and a similar setup were used in both



(a) Cumulative Gaussians fitted to the proportions of synchronized steps as a function of the time intervals before Phase Shift (left) and after Phase Shift (right). After Phase Shift, synchronization is defined relative to the non-shifted modality. In A-Shift, the auditory stimulus was phase shifted forward or backward. In V-Shift, the visual stimulus was phase shifted. Forward and backward shifts did not reveal differences and were averaged. The curves were obtained by fitting cumulative Gaussians to the proportions of de-synchronized steps. For illustration purposes, observations and predictions were inverted, thus displaying the proportion of synchronized steps.

	D					J					E				
Shift	50*	100	150	200	250	50*	100	150	200	250	50*	100	150	200	250
A	3.3	1.7	1.3	1.3	1.3	-	2.4	2	1.8	1.8	-	2.3	1.7	1.1	1.6
V	-	-	8.8	4.2	2.4	-	-	-	-	-	-	-	-	-	-

(b) Top: Upper confidence interval of the 25% De-synchronization Threshold (DT) as a function of Phase Shift. DT is the estimate of time(s) at which 25% of steps are not synchronized. The estimation was made for a time interval of maximal 20s. Dashes (-) indicate when DT exceeded this domain and hence did not occur. Asterisks (*) indicate when the 95% CI of A-Shifts and V-Shifts did overlap. Bottom: S.d. of asynchronies obtained from the observations at the last step for A-Shifts (green) and V-Shifts (red). The grey-shaded region is the 95% CI of the 0-Shift condition.

Fig. 2. Effects of phase shifts on synchronization performance.

experiments. Phase shifts were implemented to create the temporal asynchrony. Thus, we first estimated the time to achieve synchronization before phase shift from the proportions of synchronized steps. The 75% synchronization threshold was T = 1.01 s (D), T = 1.63 s (J), and T = 2.13 s (E) (see Fig. 2a left). The synchronization thresholds were smaller (faster synchronization) than in Experiment 2 mainly because of the implemented cut-off criterion through which we excluded all non-synchronized trials at the last step before phase shift. The criterion was such high because we were here interested in the effects of the phase shift and not on how much synchronization could be achieved overall. After phase shift, synchronization could be maintained or impaired. Impairment was determined against the non-shifted modality. The proportion of steps that were no longer synchronized quantified its probability of occurrence. We expected that it increases with the size and the time elapsed after phase shift. Cumulative Gaussians were fitted to the proportions of the asynchronous steps. The time at which 25% steps were not synchronized was estimated from the curves and termed 25% De-synchronization Threshold (DT).

Overall, the cumulative Gaussians fit well the proportions of A-Shifts but not of V-Shifts (Fig. 2a right). A-Shifts mean that the auditory cue was shifted relative to the visual cue. For V-Shifts it was the opposite. Smaller A-shifts compared to V-Shifts led to DT. In addition, A-Shifts led to faster DT when being larger than 50 ms. Thus, when cues were temporally separated, individuals' heel-strike onsets seem to shift into the direction of the cues indicated by the auditory modality. Again, this suggests a stronger influence of auditory cues for the control of synchronization.

In cases of DT, it is difficult to study cue integration because it cannot be clarified whether a new converging point was reached and stabilized. Nevertheless, it is unlikely that the converging point had been reached within the available 2.5s after phase shift considering that individuals required at least 1.8 s (D) to 3.6 s (E) with congruent audiovisual cues before phase shift.

Next, we considered the upper limits of the 95% Bootstrap confidence interval of DT (Fig. 2b top). At most V-Shifts and smaller A-Shifts the confidence limits exceeded the domain of the applied model, which was 20 s. When DT was larger than 20s, we labeled it as infinite. An infinite DT indicates that synchronization was maintained with the non-shifted modality. Here, integration effects could be determined.

As previously, synchronization performance was quantified by the s.d. of asynchronies obtained from the observations. Results of Experiment 2 had revealed that s.d. with congruent AV was lower than with A. Here, we observed that also with small shifts variability equaled AV. That is, below 100 ms (D), 250 ms (J), and 150 ms (E) shifts, s.d. was within the 95% confidence interval of the 0-phase shift condition (i.e., the s.d. is within the gray-shaded area in Fig. 2b bottom). Note that the s.d. is not directly comparable between Experiment 2 and 3 because in Experiment 3 all trials were excluded in which synchrony was not reached before phase shift. Nevertheless, these results indicate that cues became integrated because participants maintained low levels of variability (Ernst & Bülthoff, 2004). When synchronization was maintained with the non-shifted modality, as indicated by an infinite DT, and the asynchronies between the cues became larger, s.d. increased. This suggests that cues with larger shifts functioned as distractors.

Finally, at A-Shifts between 200 ms and 250 ms, s.d. decreased again. This may owe to a wear-off of the distractor effects. For V-Shifts, s.d. increased remarkably slower and did not decrease. These patterns are similar to findings of previous studies about finger-metronome synchronization using target-distractor paradigms (Bertelson & Aschersleben, 2003; Repp & Keller, 2004; Kato & Konishi, 2006; Hove et al., 2013). The asymmetric distractor effect could be attributed to a superior ability of the auditory system to extract temporal structure from isochronous stimulus sequences (Grahn, Henry, & McAuley, 2011; Su, 2014). Then, a temporal shift in a widely regular step sound leads to the perception of disruption of a sequence. This elicits stronger error correction processes than when the disruption is not being perceived (Repp, 2005). Temporal shifts of the visual mannequin may not have been perceived that fast.

In conclusion, both the asymmetric effects on DT thresholds and s.d. of asynchronies clearly suggest a stronger reliance on the auditory modality. When DT did not happen, cues might have been integrated serving as aid when being small and as distractors when being larger. Overall, the results of the Experiment 3 offer further support to the claim that the temporal onset of audiovisual signals is crucial and that their misalignment reduces synchronization performance.

5. General discussion

The main goal of this work was to investigate the relative contribution of audiovisual cues for synchronizing steps during side-by-side walking. In an attempt to bridge the gap between highly controlled sensorimotor synchronization paradigms and more ecologically valid studies about interpersonal coordination, three experiments were conducted using a virtually simulated walking partner.

Through Experiment 2 it was found that the presence of audiovisual cues increased synchronization performance compared to auditory or visual cues alone. In order to synchronize steps, an individual has to estimate the time of the upcoming heel strike onsets. This can be based on multiple cues. But here, we focused on the integration of heel strike sounds and visual cues of the motion of the virtual walking partner. An optimal way of estimation is to integrate signals from both modalities. This may lead to estimates with increased reliability (Ernst & Bülthoff, 2004). Increasing the reliability of the estimate should then lead to more stable synchronization. In Experiment 2, the MLE model predicted partly the synchronization variability for audiovisual cues. However, the task required that the participants cross-modally matched audiovisual cues from the partner with audiovisual, somatosensory, and kinesthetic cues from their own movements. Up to now, neither the MLE model nor any other currently available model seems able to capture the integration mechanisms in such a cross-modal sensorimotor matching paradigm. A possible experimental approach would be to manipulate all of these cues within a cross-modal framework (Elliott et al., 2010).

Multimodal signals should be integrated when the signalized unimodal event onsets fall into a WTI (Vroomen & Keetels, 2010). Hence, in Experiment 3, it was tested for bimodal integration by manipulating the asynchrony between the stimulus onsets. At small asynchronies, the variability was as small as in the synchronous condition. This is consistent with the MLE model because the integration of slightly asynchronous signals should still reduce the variability of the final estimate of the onsets (Ernst & Bülthoff, 2004). At larger asynchronies, the synchronization variability increased. This can be attributed to distractor effects of the temporally displaced stimuli. These arise when signals from supposedly separate (or independent) events are integrated to code the same event (e.g., Repp & Penel, 2004). When the asynchrony is very large, this does not happen. Then, asynchronous cross-model events are coded and perceived as separate (Keetels & Jean, 2012). Therefore, the point where distraction is maximal might indicate the size of the WTI. However, results of Experiment 3 weaken this assumption because the points of maximal distraction were much larger than expected from the WTIs determined by previous studies (see e.g., Mendonça et al., 2011; Vatakis & Spence, 2006). Moreover, a decreasing variability at larger shifts was only observed when the auditory stimulus was shifted but not when the visual stimulus was shifted.

Yet, there must be caution in drawing conclusions about the exact sensory integration mechanism. The size of the WTI differed largely between previous studies being much larger and different for complex biological stimuli (see e.g., Repp & Penel, 2002; Repp & Penel, 2004; Vatakis & Spence, 2006; Arrighi, Alais, & Burr, 2006). Furthermore, in the present experiments individuals synchronized with a non-adaptive stimulus. This is similar to sensorimotor synchronization studies in which movements had to be synchronized with metronome events (e.g., Repp, 2005), but it differs from interpersonal coordination studies where both individuals were mutually adaptive (e.g., Schmidt & Richardson, 2008). The underlying mechanism of both tasks might be different and for complex stimuli and tasks, cues that are incongruent might still be combined and be beneficial in other ways.

It must be highlighted that Experiment 2 and Experiment 3 used individuals that had also participated in Experiment 1. The pilot tests and sequence order effect analyses do not entirely rule out that participants might have improved their performance in subsequent experiments due to training. However, since this was a very novel and complex paradigm, we used the same individuals in order to have some comparison standard across the experiments. Because the sample sizes were very small, using different participants across the experiments would probably have hampered the interpretations to a greater extent. Fig. A.6 provides some hints for the absence of large training effects. Nevertheless, previous walking studies investigating spontaneous synchronization have shown that there is a large amount of variability among pairings (Zivotofsky et al., 2012). Although we assume that variability across pairings may be reduced for intentional synchronization, the small sample size limits the generalizability of our results.

It is also unclear if these findings can be generalized to spontaneous synchronization. Specifically, we do not know if the same mechanisms are shared by both the spontaneous and the intentional synchronization. Nevertheless, our results are generally consistent with those from Nessler and Gilliland, 2009, where spontaneous synchronization was greatest with audiovisual cues followed by auditory cues and then visual cues. Furthermore, several former studies on spontaneous synchronization reported an auditory dominance effect (see e.g., Repp, 2005), which was also found in our study. Besides one participant in Experiment 2, synchronization was consistently faster and less variable with auditory cues and the distraction effect was stronger with auditory shifts than with visual shifts. Therefore, the underlying sensory integration mechanisms in intentional and spontaneous synchronization might be comparable.

However, rather than assuming a superior temporal processing (see e.g., Repp, 2005; Van Wassenhove, Grant, & Poeppel, 2007), we here suggest that the auditory bias might be explained by an increased ease of matching unimodal stimuli. For asynchrony estimation, all available information can be used. We assume that during walking without obstacles people do usually not observe their feet. Then, kinesthetic and somatosensory signals from one's own body and sounds produced through ground contact signalize the heel strike onsets. It was shown that temporal estimates are much more accurate when evaluating stimuli with a sharper rise time of energy, which are here both heel strike sounds (Van der Burg, Cass, Olivers, Theeuwes, & Alais, 2009). Moreover, fusion limits at which two cues are perceived as one are much lower for multimodal stimuli (~4 Hz for audiovisual) (Fujisaki & Nishida, 2009) compared to unimodal stimuli (~25 Hz for visual) (Fujisaki & Nishida, 2005). In addition, in the present setup, the online generation of auditory stimulus position and visual perspective was delayed leading to spatiotemporal conflicts. This could have further promoted the reliance on single modalities. Consequently, it seems easier to match external auditory signals rather than cross-modally matching auditory, somatosensory, and kinesthetic signals with external visual signals. Support for this is provided by a study showing that tempo matching during stepping on place was facilitated through unimodal matching sounds with auditory feedback of the heel strikes compared to cross-modal matching sounds with haptic feedback (Maculewicz, Erkut, & Serafin, 2016).

Finally, although our study focused on the information processing mechanisms, the results could also be interpreted within the dynamical system framework. Specifically, the heel strikes could represent attractors accessible through visual and/or auditory cues. The attractor stability can be identified by the time to synchronization and its variability. Within this framework, our results suggest that (1) auditory cues represent more stable attractors than visual ones and (2) multimodal stimuli might increase the attractor stability of unimodal cues.

In short, we provided evidence that the audiovisual cues from the walking partner are integrated in order to intentionally synchronize steps during side-by-side walking. Although the model of optimal integration might explain such synchronization effects, further variables—particularly other sensory systems and the signals individuals receive from their movements—have to be considered in order to draw a more complete picture of integration mechanisms that are involved in the perception–action loop of movement synchronization.

Appendix A

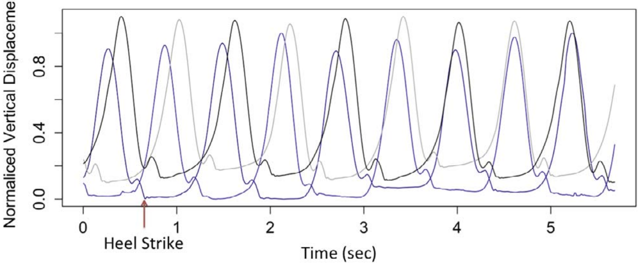


Fig. A.3. Normalized vertical displacement of the markers attached to the ankles as a function of time. Displayed are the position of left ankle (blue) and right ankle (dark blue) of the participant and the position of the left ankle (gray) and right ankle (black) of the PLW. For illustration, the time series of the PLW are here vertically shifted on the y-axis by +0.1. Heel strike onsets were identified by a peak finding algorithm detecting the second local minima (red arrow) in the vertical displacement of the ankle within a stride cycle. Due to signal noise, motion of the skin, and movement variability, this was not always possible and the positions had then to be estimated manually. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

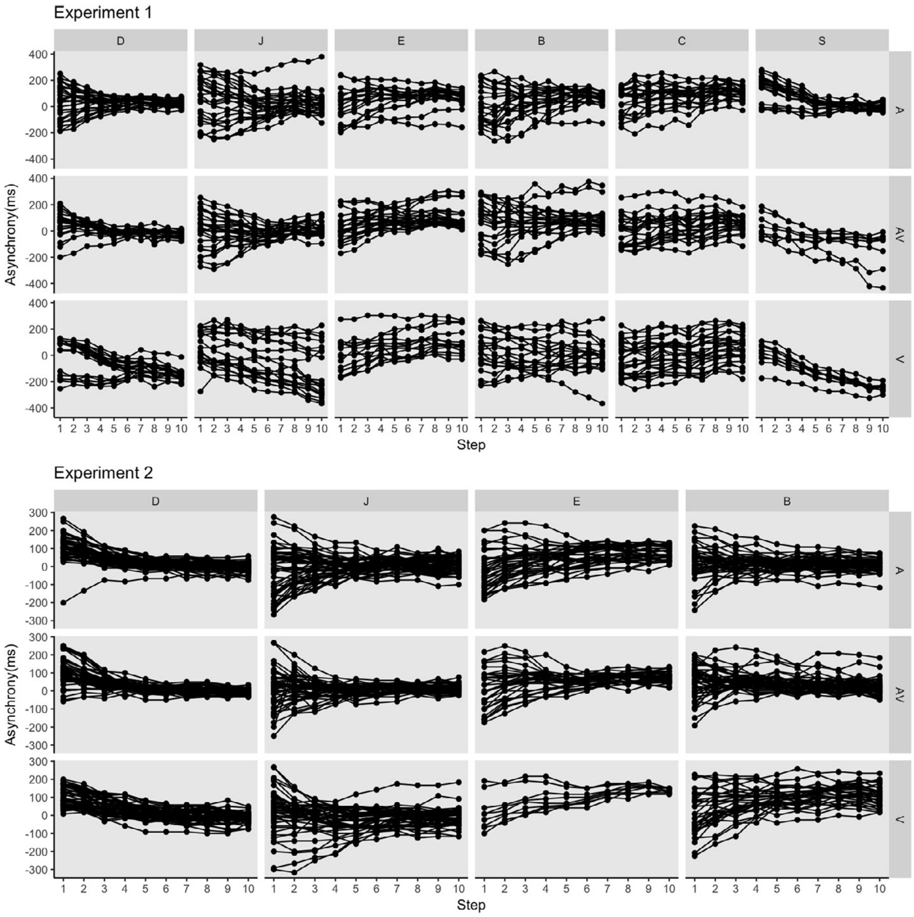


Fig. A.4. Asynchronies(ms) as a function of step number for visual (V), auditory (A), and audiovisual (AV) information of each participant for Experiment 1 and Experiment 2. Each line-segment represents one trial.

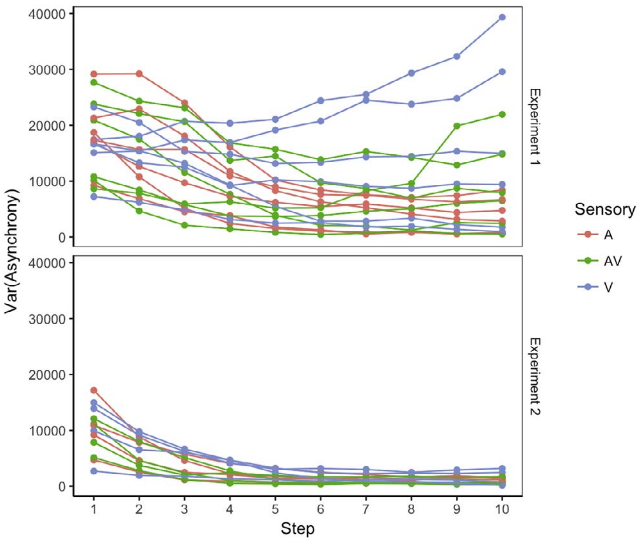


Fig. A.5. Asynchrony variance as a function of step number for Experiment 1 and Experiment 2. Each line-segment represents a participant. The sensory conditions are color-coded.

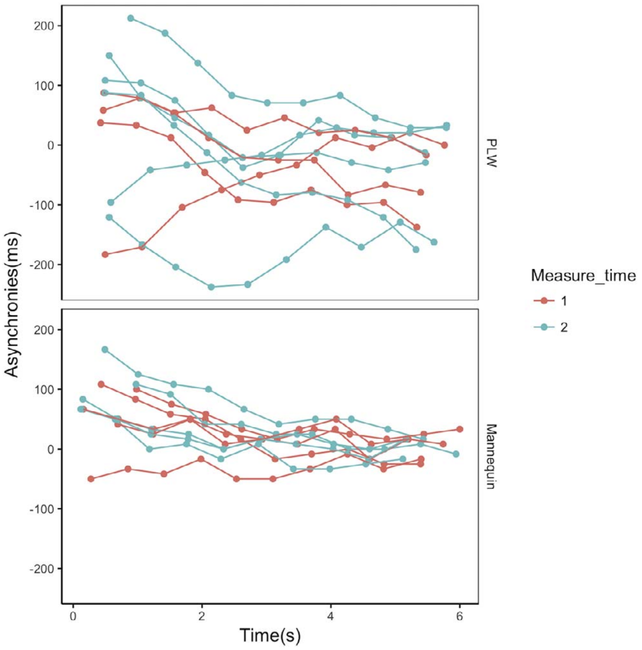


Fig. A.6. Asynchronies(ms) as a function of Time(s) separated by the time of measurement and stimulus (PLW and Mannequin). The asynchronies were obtained from the performance of a collaborate who synchronized 4–6 times in the AV sensory condition with the two types of stimuli (PLW and Mannequin) at two measurement times (Measure time 1 and Measure time 2). Measure time 2 was 7 month later than Measure time 1. The plot illustrates that the asynchrony series do not differ between the measurement times with the same stimulus.

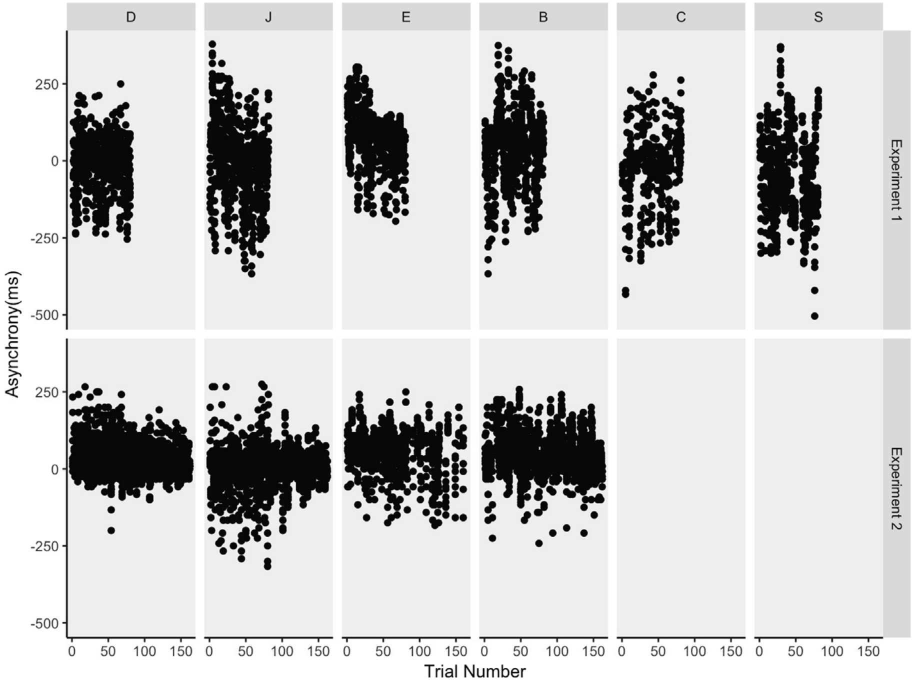


Fig. A.7. Asynchronies(ms) as a function of Trial, for each participant in Experiment 1 and Experiment 2. The plot illustrates that there is no consistent pattern of asynchrony change as the trial number increases. This suggests the absence of sequence effects.

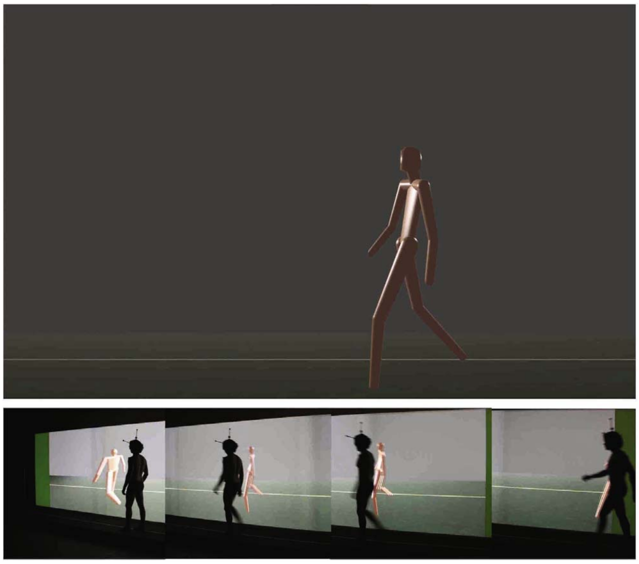


Fig. A.8. Stimulus and setup of Experiment 2. Top: Screenshot of the virtual mannequin. Bottom: Four snapshots of a participant walking next to the virtual mannequin projected on the screen. Note that the perspective changed as a function of position, which is well illustrated in the fourth snapshot (bottom left). The "antenna-like device" on the head of the participant delivered most reliable position and head rotation coordinates for the online render of image and sound.

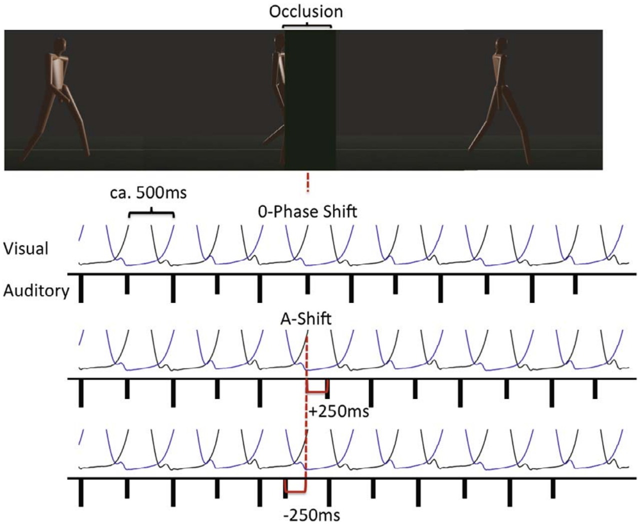


Fig. A.9. Stimulus manipulations of Experiment 3. Top: Presentation of the virtual mannequin (V) and the occlusion area (green). A Phase Shift (red) was applied during occlusion (green) of the stimuli. Bottom: Time series of the vertical ankle displacement (blue: left foot, black: right foot) (V) and footstep sounds representing the heel strike onsets (black stripes) (A). In red is exemplified the position of a positive (+) and negative (−) 250 ms Phase Shift of A (A-Shift). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

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

Arezes, P.M., Barroso, M.P., Cordeiro, P., Costa, L.G. d., & Miguel, A.S. (2006). Estudo antropométrico da população portuguesa. ISHST.

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

Arrighi, R., Marini, F., & Burr, D. (2009). Meaningful auditory information enhances perception of visual biological motion. *Journal of Vision, 9* 25–25.

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

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

Berniker, M., & Kording, K. (2011). Bayesian approaches to sensory integration for motor control. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*, 419–428.

Bertelson, P., & Aschersleben, G. (2003). Temporal ventriloquism: Crossmodal interaction on the time dimension: 1. evidence from auditory–visual temporal order judgment. *International Journal of Psychophysiology, 50*, 147–155.

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

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

Campos, J. l., & Bülthoff, H.H. (2012). Multimodal integration during self-motion in virtual reality. In M. Murray, & M. Wallace (Eds.), The Neural Bases of Multisensory Processes. Taylor and Francis.

Coey, C., Varlet, M., Schmidt, R. C., & Richardson, M. J. (2011). Effects of movement stability and congruency on the emergence of spontaneous interpersonal coordination. *Experimental Brain Research, 211*, 483–493.

Colling, L. J., & Williamson, K. (2014). Entrainment and motor emulation approaches to joint action: Alternatives or complementary approaches? *Frontiers in Human Neuroscience, 8*.

Cullen, K. E. (2012). The vestibular system: Multimodal integration and encoding of self-motion for motor control. *Trends in Neurosciences, 35*, 185–196.

Cullen, K. E. (2016). The neural encoding of self-generated and externally applied movement: implications for the perception of self-motion and spatial memory. In S. Bernard, C. Lopez, T. Brandt, P. Denise, & P. Smith (Eds.). *The Vestibular System in Cognitive and Memory Processes in Mammals*. Frontiers.

Demos, A. P., Chaffin, R., Begosh, K. T., Daniels, J. R., & Marsh, K. L. (2012). Rocking to the beat: Effects of music and partner's movements on spontaneous interpersonal coordination. *Journal of Experimental Psychology: General, 141*, 49.

de Winkel, K. N., Weesie, J., Werkhoven, P. J., & Groen, E. L. (2010). Integration of visual and inertial cues in perceived heading of self-motion. *Journal of Vision, 10* 1-1.

Elliott, M. T., Wing, A. M., & Welchman, A. E. (2010). Multisensory cues improve sensorimotor synchronisation. *European Journal of Neuroscience, 31*, 1828–1835.

Elliott, M. T., Wing, A. M., & Welchman, A. E. (2014). Moving in time: Bayesian causal inference explains movement coordination to auditory beats. *Proceedings of the Royal Society of London B: Biological Sciences, 281*, 20140751.

Ernst, M. O. (2006). A bayesian view on multimodal cue integration. *Human body perception from the inside out, 131*, 105–131.

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

Fitzpatrick, R. C., Wardman, D. L., & Taylor, J. L. (1999). Effects of galvanic vestibular stimulation during human walking. *The Journal of Physiology, 517*, 931–939.

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166*, 455–464.

Fujisaki, W., & Nishida, S. (2009). Audio–tactile superiority over visuo–tactile and audio–visual combinations in the temporal resolution of synchrony perception. *Experimental Brain Research, 198*, 245–259.

Ghez, C., & Krakauer, J. (2000). The organization of movement. *Principles of Neural Science, 656*, 668.

Grahn, J. A., Henry, M. J., & McAuley, J. D. (2011). Fmri investigation of cross-modal interactions in beat perception: Audition primes vision, but not vice versa. *Neuroimage, 54*, 1231–1243.

Grossman, E. D., Blake, R., & Kim, C.-Y. (2004). Learning to see biological motion: Brain activity parallels behavior. *Journal of Cognitive Neuroscience, 16*, 1669–1679.

Gurnsey, R., Poirier, F. J., Bluett, P., & Leibov, L. (2006). Identification of 3d shape from texture and motion across the visual field. *Journal of Vision, 6* 1-1.

Gurnsey, R., Roddy, G., Ouhnana, M., & Troje, N. F. (2008). Stimulus magnification equates identification and discrimination of biological motion across the visual field. *Vision Research, 48*, 2827–2834.

Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*, 347–356.

Hartmann, W. M. (1983). Localization of sound in rooms. *The Journal of the Acoustical Society of America, 74*, 1380–1391.

Hove, M. J., Iversen, J. R., Zhang, A., & Repp, B. H. (2013). Synchronization with competing visual and auditory rhythms: bouncing ball meets metronome. *Psychological Research, 77*, 388–398.

Ikeda, H., Blake, R., & Watanabe, K. (2005). Eccentric perception of biological motion is unscalably poor. *Vision Research, 45*, 1935–1943.

Issartel, J., Marin, L., & Cadopi, M. (2007). Unintended interpersonal co-ordination: Can we march to the beat of our own drum? *Neuroscience Letters, 411*, 174–179.

Kato, M., & Konishi, Y. (2006). Auditory dominance in the error correction process: A synchronized tapping study. *Brain Research, 1084*, 115–122.

Katz, B. F., Flinto, D. Q., Tourain, D., Poirier-Quinot, D., & Bourdot, P. (2015). Blendervr: Open-source framework for interactive and immersive vr. IEEE Press.

Kaya, D. (2014). Proprioception: The forgotten sixth sense. Proprioception and Gender. Foster City, USA: OMICS Group eBooks.

Keetels, M., & Jean, V. (2012). Perception of synchrony between the senses. In M. Murray, & M. Wallace (Eds.). *The Neural Bases of Multisensory Processes chapter 9* (pp. 147–177). London: CRC Press/Taylor and Francis.

Kolarik, A. J., Moore, B. C., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics, 78*, 373–395.

Kopčo, N., & Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources a. *The Journal of the Acoustical Society of America, 130*, 1530–1541.

Lamas, J., Silva, C.C., Silva, R., Mouta, S., Campos, J.C., & Santos, J.A. (2015). Measuring end-to-end delay in real-time auralisation systems. In: 10th European Congress and Exposition on Noise Control Engineering, Maastrich, Netherderlands (pp. 791–796).

Lappe, M., Bremmer, F., & Van den Berg, A. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences, 3*, 329–336.

Maculewicz, J., Erkut, C., & Serafin, S. (2016). An investigation on the impact of auditory and haptic feedback on rhythmic walking interactions. *International Journal of Human-Computer Studies, 85*, 40–46.

Mendonça, C., Santos, J. A., & López-Moliner, J. (2011). The benefit of multisensory integration with biological motion signals. *Experimental Brain Research, 213*, 185–192.

Mörtl, A., Lorenz, T., Vlaskamp, B. N., Gusrialdi, A., Schubö, A., & Hirche, S. (2012). Modeling inter-human movement coordination: synchronization governs joint task dynamics. *Biological Cybernetics, 106*, 241–259.

Nessler, J. A., De Leone, C. J., & Gilliland, S. (2009). Nonlinear time series analysis of knee and ankle kinematics during side by side treadmill walking. Chaos: An Interdisciplinary. *Journal of Nonlinear Science, 19*, 026104.

Nessler, J. A., & Gilliland, S. J. (2009). Interpersonal synchronization during side by side treadmill walking is influenced by leg length differential and altered sensory feedback. *Human Movement Science, 28*, 772–785.

Nessler, J. A., & Gilliland, S. J. (2010). Kinematic analysis of side-by-side stepping with intentional and unintentional synchronization. *Gait & posture, 31*, 527–529.

Oliveira, A., Campos, G., Dias, P., Murphy, D.T., Viera, J., Mendonça, C., & Santos, J. (2013). Real-time dynamic image-source implementation for auralisation. In Proceedings of the 16th International Conference on Digital Audio Effects: (pp. 368–372). York.

Oullier, O., De Guzman, G. C., Jantzen, K. J., Lagarde, J., & Kelso, J. S. (2008). Social coordination dynamics: Measuring human bonding. *Social Neuroscience, 3*, 178–192.

Pastore, R. E., Flint, J. D., Gaston, J. R., & Solomon, M. J. (2008). Auditory event perception: The source-perception loop for posture in human gait. *Attention, Perception, & Psychophysics, 70*, 13–29.

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

Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). A universal concept in nonlinear sciences. *Self, 2*, 3.

Prochazka, A., & Ellaway, P. (2012). Sensory systems in the control of movement. *Comprehensive Physiology*.

Prochazka, A., Gritsenko, V., & Yakovenko, S. (2002). Sensory control of locomotion: reexes versus higher-level control. *Sensorimotor Control of Movement and Posture* (pp. 357–367). Springer.

Pyt. (2016). Python 2.0.https://www.python.org.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychonomic Bulletin & Review, 12*, 969–992.

Repp, B. H., & Keller, P. E. (2004). Adaptation to tempo changes in sensorimotor synchronization: Effects of intention, attention, and awareness. *Quarterly Journal of Experimental Psychology Section A, 57*, 499–521.

Repp, B. H., & Penel, A. (2002). Auditory dominance in temporal processing: new evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 1085.

Repp, B. H., & Penel, A. (2004). Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychological Research, 68*, 252–270.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science, 26*, 867–891.

Richardson, M. J., Marsh, K. L., & Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 62.

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

Schmidt, R. C., & O'Brien, B. (1997). Evaluating the dynamics of unintended interpersonal coordination. *Ecological Psychology, 9*, 189–206.

Schmidt, R. C., & Richardson, M. J. (2008). Dynamics of interpersonal coordination. Springer.

Sejdić, E., Fu, Y., Pak, A., Fairley, J. A., & Chau, T. (2012). The effects of rhythmic sensory cues on the temporal dynamics of human gait. *PloS One, 7*, e43104.

Semjen, A., Vorberg, D., & Schulze, H.-H. (1998). Getting synchronized with the metronome: Comparisons between phase and period correction. *Psychological Research Psychologische Forschung, 61*, 44–55.

Silva, C. C., Mendonça, C., Mouta, S., Silva, R., Campos, J. C., & Santos, J. (2013). Depth cues and perceived audiovisual synchrony of biological motion. *PloS One, 8*, e80096.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971–995.

Strogatz, S.H. (2003). Sync. Hyperion Books.

Su, Y.-H. (2014). Visual enhancement of auditory beat perception across auditory interference levels. *Brain and Cognition, 90*, 19–31.

Thomas, J. P., & Shiffrar, M. (2010). I can see you better if i can hear you coming: Action-consistent sounds facilitate the visual detection of human gait. *Journal of Vision, 10* 14-14.

Thomas, J. P., & Shiffrar, M. (2013). Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony. *Journal of Vision, 13* 8-8.

Troje, N. F. (2008). Biological motion perception. *The senses: A comprehensive reference, 2*, 231–238.

Van der Burg, E., Cass, J., Olivers, C., Theeuwes, J., & Alais, D. (2009). Efficient visual search from nonspatial auditory cues requires more than temporal synchrony. Temporal multisensory processing and its effects on attention, (pp. 63–84).

van Ulzen, N. R., Lamoth, C. J., Daffertshofer, A., Semin, G. R., & Beek, P. J. (2008). Characteristics of instructed and uninstructed interpersonal coordination while walking side-by-side. *Neuroscience Letters, 432*, 88–93.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual 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.

Vic. (2016). Nexus 2.0.http://www.vicon.com/products/software/nexus.

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

Warren, W. H., Kay, B. A., & Yilmaz, E. H. (1996). Visual control of posture during walking: Functional specificity. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 818.

Welch, R. B., DutionHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics, 39*, 294–300.

Wing, A. M., Doumas, M., & Welchman, A. E. (2010). Combining multisensory temporal information for movement synchronisation. *Experimental Brain Research, 200*, 277–282.

Wright, R. L., & Elliott, M. T. (2014). Stepping to phase-perturbed metronome cues: multisensory advantage in movement synchrony but not correction. *Frontiers in Human Neuroscience, 8*.

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

Zivotofsky, A. Z., Gruendlinger, L., & Hausdorff, J. M. (2012). Modality-specific communication enabling gait synchronization during over-ground side-by-side walking. *Human Movement Science, 31*, 1268–1285.

Zivotofsky, A. Z., & Hausdorff, J. M. (2007). The sensory feedback mechanisms enabling couples to walk synchronously: An initial investigation. *Journal of Neuroengineering and Rehabilitation, 4*, 28.