Cognition 153 (2016) 33–42

Contents lists available at ScienceDirect

Cognition

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

Original Article

Action-outcome learning and prediction shape the window of simultaneity of audiovisual outcomes

Andrea Desantis *, Patrick Haggard

Institute of Cognitive Neuroscience, University College London, London, UK

ARTICLE INFO

Article history:
Received 26 August 2015
Revised 4 March 2016
Accepted 8 March 2016
Available online 27 April 2016

Keywords:
Agency
Voluntary action
Action-outcome learning
Action-outcome prediction
Time perception
Audiovisual binding

ABSTRACT

To form a coherent representation of the objects around us, the brain must group the different sensory features composing these objects. Here, we investigated whether actions contribute in this grouping process. In particular, we assessed whether action-outcome learning and prediction contribute to audiovisual temporal binding. Participants were presented with two audiovisual pairs: one pair was triggered by a left action, and the other by a right action. In a later test phase, the audio and visual components of these pairs were presented at different onset times. Participants judged whether they were simultaneous or not. To assess the role of action-outcome prediction on audiovisual simultaneity, each action triggered either the same audiovisual pair as in the learning phase ('predicted' pair), or the pair that had previously been associated with the other action ('unpredicted' pair). We found the time window within which auditory and visual events appeared simultaneous increased for predicted compared to unpredicted pairs. However, no change in audiovisual simultaneity was observed when audiovisual pairs followed visual cues, rather than voluntary actions. This suggests that only action-outcome learning promotes temporal grouping of audio and visual effects. In a second experiment we observed that changes in audiovisual simultaneity do not only depend on our ability to predict *what* outcomes our actions generate, but also on learning the delay between the action and the multisensory outcome. When participants learned that the delay between action and audiovisual pair was variable, the window of audiovisual simultaneity for predicted pairs increased, relative to a fixed action-outcome pair delay. This suggests that participants learn action-based predictions of audiovisual outcome, and adapt their temporal perception of outcome events based on such predictions.

1. Introduction

Our environment comprises complex objects characterized by auditory, visual, and other sensory features that are processed by partially independent brain areas. The brain must be able to appropriately group information deriving from different senses in order to identify these objects and generate a unified perceptual experience of our surroundings.

Multisensory grouping partly depends on the co-occurrence in time of different sensations (Meredith, Nemitz, & Stein, 1987): multisensory interactions are stronger when two or more modalities are perceived as occurring simultaneously (Alais, Newell, & Mamassian, 2010; Stein & Meredith, 1990). Previous research showed that multisensory temporal simultaneity is important for guiding our actions by enabling fast and accurate responses (Colonius and Diederich, 2004; Colonius and Arndt, 2001; Frens, Van Opstal, & van der Willigen, 1995; Stein & Meredith, 1990).

A more radical view of grouping reverses the causal relation between perception and action, suggesting that action drives the perceptual processes that produce multisensory grouping (cf. Lewkowicz & Ghazanfar, 2009; James, 1890; Petrini, Russell, & Pollick, 2009; Piaget, 1963). Although, the specific role of action in multisensory processing remains under-researched, there is evidence that actions may shape such processes. Past research showed that action processes strongly mediate the perception unimodal stimuli. Indeed, sensory events caused and predicted by one's own actions are attenuated compared to stimuli that are externally-generated and predicted by sensory cues (cf, sensory attenuation, Blakemore, Wolpert, & Frith, 2000; Cardoso-Leite, Mamassian, Schütz-Bosbach, & Waszak, 2010; Hughes, Desantis, & Waszak, 2013). Research on sensory attenuation has also shown that this effect depends largely on action processes involved in the

* Corresponding author.
E-mail address: aerdna.desantis@gmail.com (A. Desantis).

http://dx.doi.org/10.1016/j.cognition.2016.03.009
0010-0277/

preparation of actions and in the neural prediction of the specific sensory consequences that our actions produce[1] (Stenner, Bauer, Heinze, Haggard, & Dolan, 2014; Blakemore et al., 2000; Cardoso-Leite et al., 2010; Wolpert, 1997). Taken together, these studies suggest that action processes mediate the transformation of physical stimulation into perceptual experience (Wilson & Knoblich, 2005). In line with these notion, the present study investigated whether the processes involved in action planning and action-outcome prediction also shape perceptual multimodal grouping. Supporting evidence for this hypothesis comes from a recent study showing that active exploration of audiovisual objects enhances memory and subsequent recognition of these objects compared to passive observation (Butler, James, & James, 2011).

Previous studies investigating the influence of action on perceptual experience, generally focussed on action outcomes confined to a single sensory modality. In real life, however, most actions produce multisensory effects. For example, speaking produces auditory, kinaesthetic and tactile inputs. Consequently, when preparing/executing an action the motor system might predict several sensory outcomes (i.e., auditory, tactile sensations) to occur *together* as a *common* outcome of our motor command. Importantly, with the term prediction we refer to the prediction of the content/identity of an action-outcome. In other words, we hypothesized that when we predict that our actions generate a specific combination of a sound and a visual input, we might tend to group these inputs into a simultaneous multisensory percept. In line with this notion, the *unity assumption* of multisensory perception (Jackson, 1953; Vatakis & Spence, 2007; Welch, 1999) states that sensory events that "go together" are experienced as simultaneous, even if they are slightly asynchronous (Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2007). Here, we test the hypothesis that action-outcome learning and prediction (i.e., the prediction of the content/identity of an action-outcome) facilitates the process of audiovisual binding. That is, action-outcome learning and prediction may enlarge a hypothetical "temporal window" within which the different multisensory components of an action outcome are perceived to be simultaneous. In this paper, we will refer to this concept as the Window of Audiovisual Simultaneity (WAS). This grouping process may be crucial, suggesting that the action system contributes to the unity and coherence of our perceptual experience (i.e., active exploration would help us create a unified and coherent representation of the external world, e.g., Piaget, 1963). Secondly, it might be essential to develop a healthy sense of agency. Indeed, when we generate outcomes composed of different features, the brain must be able to selectively bind the sensory components of these outcomes and not others, to prevent erroneous self-attributions.

To assess the role of action-outcome learning and prediction on multisensory binding, we conducted three experiments (one of them is reported in supplementary material) using a *mismatch paradigm,* in which the match/mismatch between predicted and actual action outcomes was varied (for similar methods see Baess, Widmann, Roye, Schröger, & Jacobsen, 2009; Cardoso-Leite et al., 2010; Desantis, Mamassian, Lisi, & Waszak, 2014). Participants were presented with two audiovisual pairs. They learned that one pair followed a left hand action, and the other pair followed a right hand action. Audio and visual inputs were presented simultaneously, but the interval between the action and the audiovisual pair was jittered. In a later test phase, each action could trigger either the same audiovisual pair as in the initial learning phase (the 'predicted' pair), or the pair associated with the other action ('unpredicted' pair). The latter case created a mismatch between predicted and actual action outcome. Importantly, the association of audio and visual components within each pair remained unbroken throughout the experiment: match/mismatch occurred between action and outcome, and never within the components of the outcome itself. In the test phase, the interval between the audio and visual components of each pair varied, and participants judged whether they were presented simultaneously or not. We hypothesised that learning a specific action-outcome relation would temporally bind the audio and visual components within the outcome pair (Fig. 1).

As a consequence of this process, the audio and visual components of the predicted outcome should more readily be experienced as simultaneous, even when slightly asynchronous, compared to the unpredicted outcome. To clarify whether this multisensory perceptual binding was indeed driven by *action*-outcome prediction, we compared a condition in which the participants voluntarily triggered the outcome through their own action, and a condition where the participants made no actions, but the outcomes were predicted by visual cues, with the same latency and probability relations as the action condition. We expected to observe no change of the WAS between predicted and unpredicted pairs when these were associated to visual cues and not actions.

In Experiment 2, we investigated whether the relation between action-outcome learning and temporal binding within the audiovisual outcome might itself be temporally tuned. That is, when participants learn the relation between an action and a multisensory outcome, they may also learn the time window within which the audio and visual components of the outcome should be bound together. Specifically, we hypothesised that a reliable temporal delay between an action and a predicted outcome should lead to a narrower temporal window for binding the predicted components of the outcome, relative to a variable delay. Evidence in support of this hypothesis comes from studies on sensory attenuation. Notably, research demonstrated that sensory attenuation of predicted action-outcome (by predicted action-outcome we mean the prediction of *what* outcome an action generates) occurs specifically around the time at which participants' expect the predicted outcome to occur (Bays, Wolpert, & Flanagan, 2005). Moreover, previous studies showed that prior experiences can recalibrate the window of audiovisual grouping (see Fujisaki, Shimojo, Kashino, & Nishida, 2004; Roseboom & Arnold, 2011; Spence & Squire, 2003; Vroomen, Keetels, de Gelder, & Bertelson, 2004), but it remains unclear whether action-outcome learning can induce such changes. Strategic tuning of the WAS could play an important role in parsing sensory events into those that are self-caused, and those that are externally-generated. For example, incorrectly setting too wide an action-based window for multisensory binding might lead to erroneous self-attribution of multimodal events, in a manner reminiscent of delusions of control.

2. Experiment 1

2.1. Materials and methods

2.1.1. Participants

Sixteen volunteers (12 women, average age = 21.28 years, *SD* = 3.78 years) were tested for an allowance of £ 7.5/h or course credit. Participants completed the experiment in two sessions on separate days (see supplementary material for inclusion criteria). All had normal or corrected-to-normal vision and hearing and were naïve as to the hypothesis under investigation. They all gave written informed consent. The experiments were conducted with ethical committee permission.

[1] In the present manuscript with the term *action-outcome prediction* or *prediction* we refer to the ability to predict *what* outcome an action generates. In other words, we refer to the prediction of the content or identity of an action-outcome. We use, instead, the term *temporal expectation* or *expectation* to refer to the ability to anticipate the *time* onset of action-outcomes.

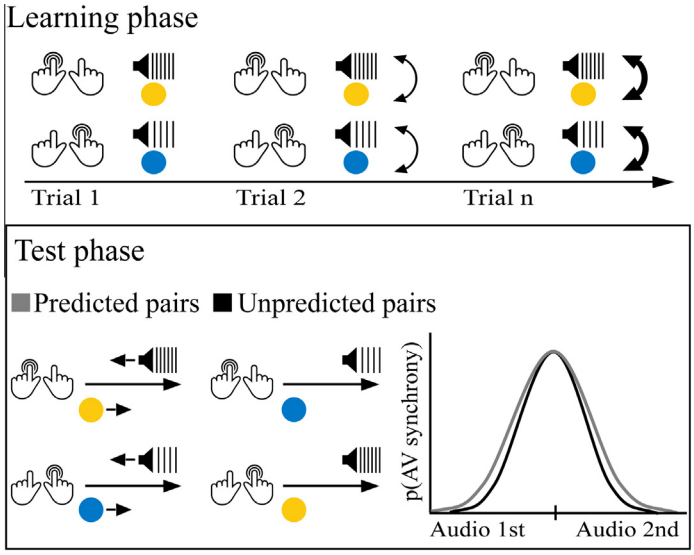

Fig. 1. Illustration of the expected results. (Top panel) Participants learned that specific actions generate specific audiovisual outcome pairs. We hypothesised that learning the association between action and outcome would drive the binding between the components of the outcome itself, represented by the curved double-headed arrows. Notably, action-outcome learning would lead the action system to consider the audio and visual components of the pair as *common outcomes* of a specific action. As a consequence, in agreement with the *unity assumption* participants would show more tolerance to audiovisual asynchronies when presented with predicted audiovisual pairs in the test phase compared to unpredicted pairs. In other words, their window of audiovisual synchrony will increase when their actions trigger predicted pairs (gray line) compared to unpredicted pairs (black line). Hand icon made by Yannick from www.flaticon.com.

2.1.2. Materials

Stimulus presentation and data acquisition were conducted using the psychophysics Toolbox (Brainard, 1997; Pelli, 1997) for Matlab 8.2.0 running on a PC computer connected to a 15-in. 60 Hz LCD monitor. Audio stimuli were presented via headphones (Sennheiser HD201).

2.1.3. Stimuli and procedure

Stimuli consisted of 4 pure-tones of 2.2 kHz, 1.6 kHz, 1 kHz, and 0.4 kHz of frequency and 4 colored Gaussian patches: magenta, green, yellow, and cyan. Tones were presented at 74dB SPL. Gaussian patches were 0.38° wide and were presented with a luminance level of 52.5 cd/m^2 in a dark gray background (14 cd/m^2) from a viewing distance of ~60 cm.

Audio and visual stimuli were combined to create four audiovisual pairs, two presented in the action condition and the other two in the sensory condition. For half of the participants, audiovisual pairs in the action condition were created by combining 2.2 kHz and 1 kHz pure tones with the magenta and green patches. In contrast, audiovisual pairs in the sensory-cue condition were created by combining 1.6 kHz and 0.4 kHz pure tones with yellow and cyan patches. For the remaining participants, audiovisual pairs in the action condition were composed of 1.6/0.4 kHz pure tones with yellow/cyan patches, and those in the sensory-cue condition were composed of 2.2/1 kHz pure tones with magenta/green patches.

The experiment consisted of a total of 64 blocks (32 action and 32 sensory blocks). Each block consisted of a learning phase followed by short test phase. The order of presentation of the action blocks and sensory blocks was counterbalanced across subjects. Participants completed the experiment in two sessions each consisting of 16 action and 16 sensory blocks. Each session lasted ~75 min.

2.1.4. Learning phases

Participants were led to learn action-stimulus and cue-stimulus associations (Fig. 2). In the *action* learning phases, in each trial they were presented at the center of the screen with a rhomb (cues size, 0.65° of width and 0.06° of thickness). This stimulus was used as fixation. They were informed to decide on each trial which button they wanted to press (left or right index finger key-press). They could execute their key-press at a time of their own choosing, but at intervals of at least 500 ms, with random and approximately equiprobable choice between the two alternatives. To help participants to perform a roughly equal number of left and right actions, feedback of the proportion of right and left key-presses was provided every 10 trials.

At key-press onset the fixation disappeared. According to the mapping to which participants were assigned, each left and right key-press generated a specific audiovisual pair. For instance, for one group of participants, the left action triggered a 2.2 kHz tone – yellow flash pair, and the right action triggered a 1 kHz tone – cyan patch pair. Action-audiovisual pair mappings were counterbalanced across subjects.

Audio and visual inputs were presented simultaneously for ~16 ms. Audiovisual pairs were presented at one of six possible time intervals 250, 283, 316, 333, 383 or 416 ms, after action execution. Action – audiovisual pair intervals were randomly selected in each trial.

In the *sensory-cue* learning phase, audiovisual pairs were not caused by any actions, but rather followed one of two visual cues: an empty circle and an empty square (cues size, 0.65° of width and 0.06° of thickness). The two visual cues were presented in random order and equally often. The offset time of the cues was randomly selected from a Gaussian distribution calculated from the mean and standard deviation of the participants' action time latencies in the action blocks. However, if the participant started the experiment with the *sensory* condition, cue offset time distribution was defined from the mean action time and standard deviation of the previous participant. As for the *action* learning phases, audiovisual pairs were presented at one of 6 possible time intervals (250, 283, 316, 333, 383 or 416 ms) after the offset of the visual cues. We timed the onset of the audiovisual pairs relative to the offset of the visual cue to match the action and the sensory cue condition. Indeed, in the action blocks the fixation disappeared at action onset, and then the audiovisual pair was presented after one of the intervals described above.

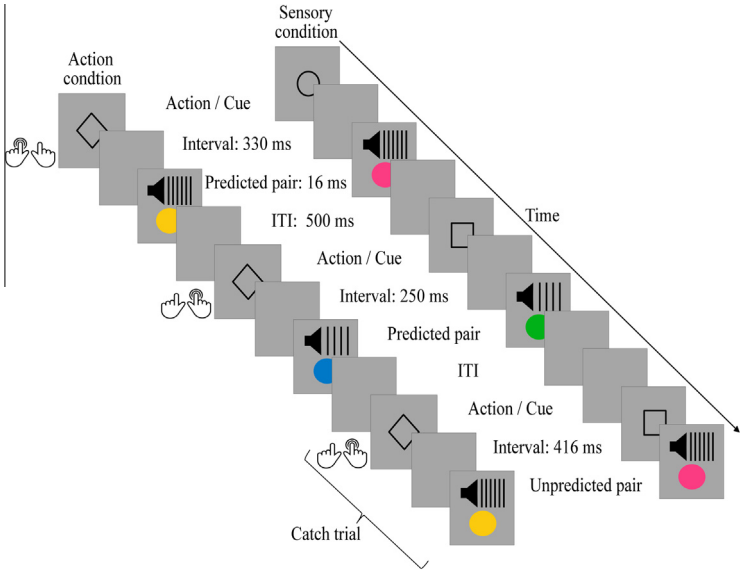

Fig. 2. Illustration of the learning phase. In the learning phase of the *Action condition* left and right actions were associated with specific audiovisual pairs. Audiovisual pairs were presented after a random delay of 250, 283, 316, 333, 383 or 416 ms relative to action onset. In the learning phase of the *sensory condition*, audiovisual pairs were associated with specific visual cues (circle or square). In both conditions, 20% of all trials were catch trials in which 'unpredicted' pairs were presented, i.e., the pairs that were associated with the other action/cue. Hand icon made by Yannick from www.flaticon.com.

Each cue was followed by an audiovisual pair. For instance, for a group of participants the circle-cue was followed by a 1.6 kHz tone – magenta patch pair and the square-cue by a 0.4 kHz tone – green patch pair. Cue-audiovisual pair mappings were counterbalanced across subjects.

To ensure that participants were paying attention to the audiovisual pairs presented, and that they learned the appropriate mapping between actions (or sensory cues) and audiovisual pairs, 20% of all trials were catch trials. In these trials participants were presented with an unpredicted audiovisual pair, i.e., the audiovisual pair that was associated with the other action/cue. Participants had to report these events by pressing both keys as fast as possible. A response time exceeding 1 s was considered as a *miss*. Learning phases consisted of 40 trials in the first block, and 20 trials thereafter.

2.1.5. Test phases

After each learning phase, participants completed 10 test trials, to assess how action/cue – audiovisual pair associations influenced the perception of audiovisual simultaneity. Unlike in the learning phase, the test phase involved asynchronous audio and visual components. The asynchrony between audio and visual components was varied randomly to be −266, −133, −86, −66, −33, 33, 66, 86, 133, or 266 ms (negative values indicate that audio preceded vision). At the end of each trial, participants indicated by pressing one of two foot pedals whether the audio and visual stimuli were simultaneous (left pedal) or not (right pedal).

In both the action and the sensory conditions, participants completed 'predicted' and 'unpredicted' trials in which the associations from previous learning phases between left/right action (*action* condition) or circle/square cue (*sensory-cue* condition), and the subsequent audiovisual pair was respected or violated, respectively (see Fig. 3). For instance, if in the *action* learning phase the 2.2 kHz tone – yellow flash pair was associated with the left keypress, then the same action triggered the same pair on half the trials in the test phase ('predicted' pair trials), while in the remaining trials it triggered the audiovisual pair previously learned to be associated with the right hand action, i.e., 1 kHz-cyan ('unpredicted' pair trials). Importantly, participants were informed that the identity of the stimuli presented was irrelevant for the simultaneity judgment task. However, they were asked to pay attention to stimulus identity, as they were required to indicate which audio or visual stimulus was presented in 20% of the trials at random. This was done in order to make sure that participants paid equal attention to predicted and unpredicted sounds and patches. In particular, in those trials, participants were presented with one of two following questions: "Which flash did you see?" or "Which sound did you hear?". The two possible answers (e.g., high sound, low sound) were presented to the left and right of the center of the screen. Participants responded by selecting one answer by pressing the left or the right pedal.

In total the 10 audiovisual SOAs (−266, −133, −86, −66, −33, +33, +66, +86, +133, +266 ms) were presented 12, 14, 16, 18, 20, 20, 18, 16, 14, and 12 times respectively, giving a total of 160 × 2 Cue (action and sensory conditions) × 2 Audiovisual Pair (predicted and unpredicted), namely 640 trials.

2.2. Data analysis

The proportion of "sound and flash simultaneous" responses for each audiovisual SOAs was calculated separately for each participant and condition: Action (present, absent) × Audiovisual pair (predicted, unpredicted). Psychometric functions were fitted using a (Gaussian) nonlinear regression model (details about fitting methods are reported in supplementary material, along with mean r^2 for each fit. Mean values for each condition, participant and experiment, as well as raw data are available from Open Science Framework http://osf.io/2kab9). Based on each individual function, we calculated the point of subjective simultaneity (PSS) for audiovisual asynchrony, corresponding to the peak of the Gaussian distribution with respect to the axis of abscissae. This gives an estimate of the temporal offset between sound and visual stimulus required for them to be perceived as simultaneous. However, our hypotheses focussed on the tolerance to audiovisual asynchrony rather than on the PSS: stronger temporal binding between audio and visual stimuli would result in an increase tolerance to audiovisual asynchrony, meaning that temporal asynchronies between visual and audio stimuli would go unperceived (cf. Jackson, 1953;

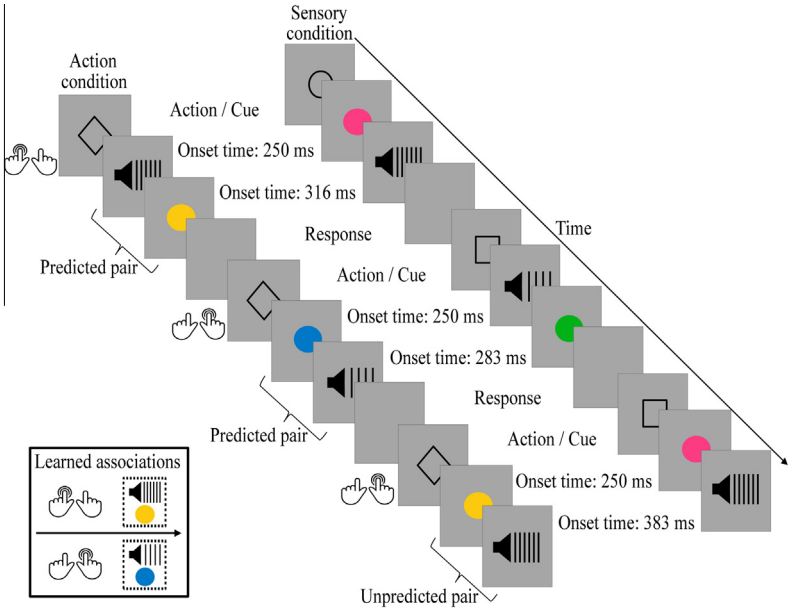

Fig. 3. Illustration of the test phase. In both the sensory and the action test phases, the onset time of the audio and visual components of the audiovisual pairs was varied. Participants completed an audiovisual simultaneity judgment task. In both action and sensory blocks, they were presented with 'predicted' and 'unpredicted' pairs. For instance, in 'unpredicted' trials the action/cue was followed by the sound and flash associated with the other action/cue. This created a mismatch between the pair presented and the pair predicted on the basis of the previously-learned association. Importantly, the association of audio and visual components *within* each pair remained unbroken throughout the experiment. Only the association between action and audiovisual pair was either respected ('predicted' pair trials) or violated ('unpredicted' pair trials) in the test phase. Hand icon made by Yannick from www.flaticon.com.

Vatakis & Spence, 2007; Vatakis et al., 2008; Welch, 1999). In other words, when two modalities are bound, participants would show higher tolerance to audiovisual asynchrony, and thus larger WAS value. We estimated participants' tolerance to audiovisual asynchrony using the standard deviation (SD) of the psychometric function. Higher SD values indicate higher tolerance to audiovisual asynchrony, corresponding to a loss of information about the relative timing of sound and flash. Significance value was set at $p < .05$ for all statistical tests.

2.3. Results

A repeated measure ANOVA on PSS values with Action (present, absent) and Audiovisual pair (predicted, unpredicted) as factors showed and no interaction $F(1,15) = .360$, $p = .557$, $\eta_p^2 = 0.023$, no main effect of Action $F(1,15) = .677$, $p = .424$, $\eta_p^2 = 0.043$, and no main effect of Audiovisual pair $F(1,15) = .319$, $p = .581$, $\eta_p^2 = 0.021$.

The SD of the Gaussian fit provides a direct estimate of the Window of Audiovisual Simultaneity (WAS) showed a significant interaction $F(1,15) = 11.971$, $p = .003$, $\eta_p^2 = 0.444$. No main effects of Action and Audiovisual pair were observed, $F(1,15) = .131$, $p = .722$, $\eta_p^2 = 0.009$, and $F(1,15) = 1.346$, $p = .264$, $\eta_p^2 = 0.083$, respectively. We then explored the interaction using simple effects. We performed two paired two-tailed *t*-tests to assess differences between 'predicted' and 'unpredicted' pairs in both the action and sensory conditions. The analyses showed that SD values were higher for 'predicted' pairs compared to 'unpredicted' pairs when these followed an action $t(15) = 2.890$, $p = .011$, $d = 0.564$. Participants were less sensitive to audiovisual asynchronies for predicted compared to unpredicted pairs. In other words, the window of tolerance to audiovisual asynchrony was higher for the predicted ($M = 103$ ms, $SD = 36$ ms) compared to unpredicted pairs ($M = 88$ ms, $SD = 21$ ms). No difference was observed between predicted and unpredicted audiovisual pairs when preceded by visual cues $t(15) = 1.645$, $p = 0.121$, $d = 0.228$. Thus, participants' tolerance to audiovisual asynchrony for predicted pairs increased only when these pairs were generated by an action (Fig. 4). This suggests that the audio and visual components of a predicted pair (i.e., the same pair that participants' action generated in the learning phase) were reported as more closely bound together in time than the components of an unpredicted pair (i.e., the pair that was associated with the other action).

We then assessed whether the absence of a difference between predicted and unpredicted trials in the sensory condition was due to the fact that participants did not learn cue-stimulus associations. We computed a repeated measure ANOVA on identification d' for both action (d': $M = 4.140$, $SD = 0.700$) and sensory condition (d': $M = 4.060$, $SD = 0.726$) in the learning phase. The analysis showed no significant effect of Action $F(1,15) = .319$, $p = .580$, $\eta_p^2 = 0.021$. Consequently, participants did learn both action and cue-audiovisual pair associations.

We also investigated whether there were differences in the allocation of attentional resources for the predicted and unpredicted events in the action compared to the sensory condition of the test phase. We conducted a repeated measures ANOVA on identification performances in the catch trials with Action (present, absent) and Audiovisual pair (predicted, unpredicted) as factors. The ANOVA showed no significant interaction $F(1,15) = .036$, $p = .852$, $\eta_p^2 = 0.002$, no main effect of Action $F(1,15) = 1.601$, $p = .225$, $\eta_p^2 = 0.096$ and Audiovisual pair $F(1,15) = 2.501$, $p = .135$, $\eta_p^2 = 0.143$. This suggests that participants' attention was equally focused to stimuli in all conditions. Proportion of correct identification for the four conditions show that participants identified correctly almost all stimuli: predicted pair (action): $M = 0.932$, $SD = 0.084$; unpredicted pair (action): $M = 0.949$, $SD = 0.053$; predicted pair (sensory): $M = 0.949$, $SD = 0.044$; unpredicted pair (sensory): $M = 0.962$, $SD = 0.036$.

2.4. Preliminary discussion

Participants showed more tolerance to audiovisual asynchrony when their actions generated a predicted audiovisual pair,

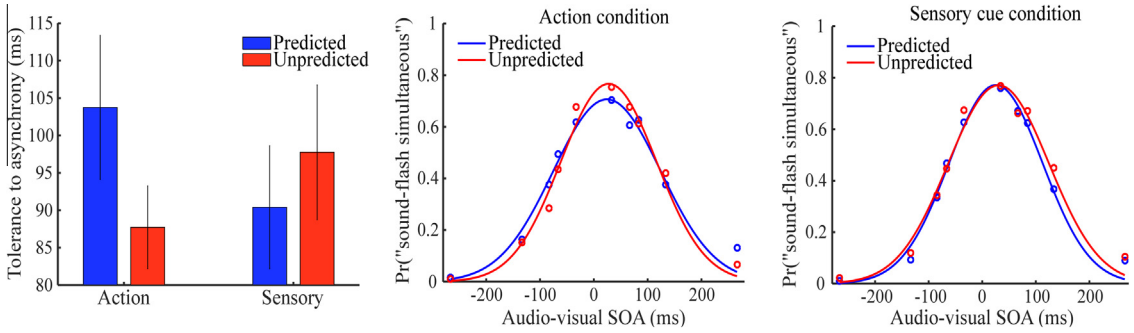

Fig. 4. (Left panel) Mean tolerance to asynchrony (SD) values for all conditions (averaged across all participants). High SD values indicate high tolerance to audiovisual asynchronies, i.e., a wide WAS. (Central panel and right panel) Proportion of "sound and flash simultaneous" responses for predicted and unpredicted effects in the action and sensory cue condition, respectively (averaged across all participants) as a function of the 10 audiovisual SOAs.

compared to when the pair violated the action-outcome associations acquired in the learning phase. Interestingly, the modulation of the WAS was doubly-specific. First, it was found only for outcome pairs that followed actions and not for the same pairs presented after a sensory cue. Second, the effect was found only when the action produced the outcome pair predicted from the association acquired in the learning phase compared to when the same action generated an unpredicted pair.

3. Experiment 2

Voluntary action control is characterized not only by the ability to predict *what* the consequences of our actions are, but also *when* these predicted consequences will occur (Bays et al., 2005). For instance, when I press the switch to turn on the light I predict the *light* to turn on *immediately* after I pressed the switch. In Experiment 2 we investigated whether the relation between action-outcome learning and temporal binding within the audiovisual outcome might itself be temporally tuned. That is, when participants learn the temporal relation between an action and a multisensory outcome, they may also learn the time window within which the component or the predicted audio and visual effects should be bound together.

Previous research demonstrated that sensory attenuation of predicted action-outcome is temporally tuned. Notably, sensory attenuation occurs specifically around the time at which participants' expect a predicted action-outcome to occur (Bays et al., 2005; see also Desantis, Roussel, & Waszak, 2014). For instance, when participants produced with a right action a tactile stimulation over the left index fingertip, attenuation of the predicted tactile sensation was temporally tuned on the time at which fingers would normally make contact (i.e., around 0 ms delay, Bays et al., 2005). Interestingly, this attenuation was not observed when the same tactile stimulation was temporally predictable but generated externally.

Similarly, we hypothesised that the WAS might be temporally tuned based on previously experienced action-audiovisual pair intervals. That is, learning that a given audiovisual pair follows an action with a fixed delay might allow strategic tuning of a narrow WAS, within which the audio and visual components would be grouped. Conversely, a variable action-outcome pair delay would require a wider WAS. In the context of action control and agency, tuning the window of simultaneity might be essential as grouping sensory inputs outside a plausible action-related time window might lead to binding the wrong sensory inputs. This might in turn cause erroneous self-attribution of multimodal events, in a manner reminiscent of delusions of control.

In Experiment 2 we investigated whether learning a fixed or a variable delay between action and outcome would result in a strategic adjustment of the WAS. Variable action-outcome intervals in the learning phase were expected to replicate Experiment 1, with a greater WAS for predicted compared to unpredicted audiovisual pairs. However, learning a fixed action-pair interval might narrow the WAS, since participants would have a precise temporal expectation of when the predicted audiovisual pair would occur. This would decrease their tolerance to asynchronies for predicted pairs presented at a fixed action-outcome delay, compared to predicted pairs presented at a variable interval in the learning phase. We did not have any hypotheses for the temporal tuning of the unpredicted pairs because this tuning has, to our knowledge, never been tested.

3.1. Materials and methods

3.1.1. Participants

Sixteen volunteers (9 women, average age = 22.56 years, *SD* = 2.80 years) participated in the experiment for an allowance of £ 7.5/h. All had normal or corrected-to-normal vision and hearing and were naïve as to the hypothesis under investigation. They all gave written informed consent.

3.1.2. Materials

See experiment 1

3.1.3. Stimuli and procedure

Stimuli consisted of 2 pure-tones of 2.2 kHz and 1 kHz of frequency and 2 colored Gaussian patches: yellow and cyan. Sound-patch associations were counterbalanced across participants. Participants completed 64 *action* blocks. Each block consisted of a learning phase followed by a short test phase. Participants completed the experiment in two sessions taking place in two different days. Each session lasted ~75 min.

3.1.4. Learning phases

As for Experiment 1 participants executed left or right index finger key-presses at a time of their own choosing. Each action was associated with one audiovisual pair. For instance, the left action triggered a 2.2 kHz tone – yellow patch pair and the right action triggered a 1 kHz tone – cyan patch pair. Action-audiovisual pair mappings were counterbalanced across subjects.

The audio and visual components of both pairs were presented simultaneously for a duration of ~16 ms. However, for one of the actions, audiovisual pairs were presented with a variable action-audiovisual pair interval of 250, 280, 310, 340, 380 or 410 ms. Instead, audiovisual pairs triggered by the other key-press were always presented after a fixed delay of 330 ms (i.e., the mean of the variable delays). For half of the participant the left hand was associated with a variable action-outcome pair interval and the

right with a constant action-outcome pair interval. For the other half the reversed association was used.

To ensure that participants were paying attention to the audiovisual pairs, 20% of all trials were catch trials. In those trials, participants were presented with either a louder sound or a brighter patch (13.3% louder/brighter than the standard stimuli). Participants were required to report the change in saliency by pressing both left and right key together. Thus, in a refinement of the method of experiment 1, the dimension defining a catch trial was now orthogonal to the action-outcome pair relation. This change aimed to avoid any interference between the task to perform in the learning phase and the task to perform in the test phase. Learning phases consisted of 40 trials in the first block, and 20 trials thereafter.

3.1.5. Test phase

The test phase was as for the action trials in Experiment 1. Participants' actions triggered the 'predicted' audiovisual pair (i.e., the pair that was associated with that action in the learning phase), in 50% of the trials, and the 'unpredicted' pairs on the remaining trials. Participants judged audiovisual simultaneity as before.

3.2. Results

A repeated measure ANOVA on PSS values with learned action-audiovisual pair interval (variable, fixed) and Audiovisual pair (predicted, unpredicted) as factors showed no interaction $F(1,15) = 2.070$, $p = .170$, $\eta_p^2 = 0.121$. Similarly, we observed no main effect of learned interval $F(1,15) = 1.518$, $p = .237$, $\eta_p^2 = 0.092$ and no main effect of $F(1,15) = .004$, $p = .989$, $\eta_p^2 < 0.000$.

ANOVA of the WAS showed no main effect of Learned interval or Audiovisual pair: $F(1,15) = .320$, $p = .580$, $\eta_p^2 = 0.021$, and $F(1,15) = 1.558$, $p = .231$, $\eta_p^2 = 0.094$, respectively, but a significant interaction $F(1,15) = 5.697$, $p = .031$, $\eta_p^2 = 0.275$. To explore the pattern of this interaction, we used simple effect testing. The WAS estimate was higher for predicted (M = 100 ms; SD = 26 ms) than unpredicted (M = 88 ms, SD = 24 ms) pairs following a variable action-pair interval in the learning phase: $t(15) = 2.779$, $p = .014$, $d = 0.463$ (Fig. 5). No difference was found for fixed action-pair interval (predicted: M = 93 ms, SD = 27 ms, unpredicted: M = 97 ms, SD = 27 ms: $t(15) = 0.853$, $p = 0.407$, $d = 0.148$). Finally, the difference between the WAS of predicted pairs for fixed action-pair intervals (100 ms) and for variable action-pair intervals (93 ms) did not quite reach the boundary of statistical significance: $t(15) = 2.044$, $p = 0.059$.

We conducted a repeated measures ANOVA on identification d' for both variable (d': $M = 3.678$, $SD = 0.620$) and fixed learned interval (d': $M = 3.946$, $SD = 0.682$) in the learning phase to assess whether participants were equally paying attention to bimodal stimuli in both intervals exposure. The analysis showed no significant effect of learned interval $F(1,15) = 3.669$, $p = .075$, $\eta_p^2 = 0.196$.

Finally, we assessed whether participants attended equally to predicted and unpredicted pairs in the test phase. The analyses showed no significant interaction $F(1,15) = .103$, $p = .753$, $\eta_p^2 = 0.007$, no main effect of learned action-audiovisual pair interval $F(1,15) = 0.975$, $p = .339$, $\eta_p^2 = 0.061$ and no main effect of audiovisual pair $F(1,15) = 0.039$, $p = .846$, $\eta_p^2 = 0.002$. The proportion of correct responses for each condition were as follow: predicted pair variable interval: $M = 0.928$, $SD = 0.122$; unpredicted pair variable interval: $M = 0.921$, $SD = 0.087$; predicted pair fixed interval: $M = 0.932$, $SD = 0.096$; unpredicted pair fixed interval: $M = 0.941$, $SD = 0.084$.

4. General discussion

The current study investigated whether action-outcome learning and the prediction of sensory outcomes promote audiovisual temporal grouping. In Experiment 1 participants exhibited more tolerance to asynchrony of an audio and visual input that was predictable from their action, compared to audio and visual inputs that were not so predicted. In other words, the window of audiovisual binding was wider when participants' actions were followed by predicted compared to unpredicted pairs. No change in the window of audiovisual simultaneity was observed when audiovisual pairs followed visual cues, rather than voluntary actions. This suggests that the cognitive processes of selecting and executing an action preparatorily engage a process for multisensory binding of the specific predicted outcomes of the chosen action. Mere statistical predictability of the same outcomes, as in our visual cue condition, was not sufficient.

These results provide further evidence suggesting that learning and prediction based on actions has different influences on our experience of the world from learning and prediction based on sensory cues. Past research on unimodal perception showed that events that are predicted from our actions are attenuated compared to the same events but when they are predicted by sensory cues and not actions (Blakemore et al., 2000; Cardoso-Leite et al., 2010; Hughes et al., 2013). Our study suggests that processes involved in the preparation of action and in action-outcome prediction, not only influence unimodal perception but also mediate multimodal grouping. Notably, action-outcome learning would lead the system to predict a specific pair of audio and visual outcomes to occur *together* as a *common* consequence a specific action. Consequently, only the audio and visual stimuli participants'

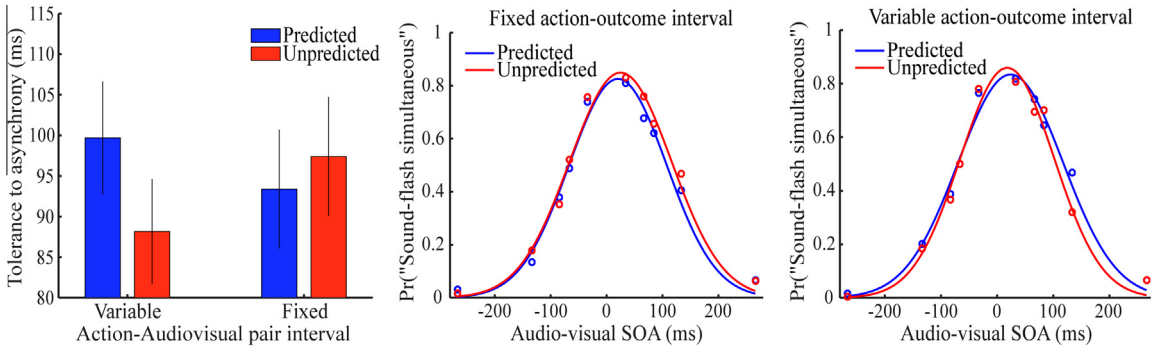

Fig. 5. (Left panel) Mean tolerance to asynchrony (SD) values for all conditions (averaged across all participants). High SD values indicate high tolerance to audiovisual asynchronies, i.e., a wide WAS. (Central panel and right panel) Proportion of "sound and flash simultaneous" responses for predicted and unpredicted effects in the variable and fixed action-outcome interval condition, respectively (averaged across all participants) as a function of the 10 audiovisual SOAs.

expected to generate from the specific chosen action were grouped into a simultaneous multisensory outcome. Evidence supporting this interpretation comes from a recent fMRI study showing that actions mediate audiovisual object learning. Butler et al. (2011) showed that active exploration of audiovisual objects consistently improved subsequent object recognition compared to passive observation. Moreover, they observed that the active exploration group exhibited strong activation of motor-related areas (Supplementary Motor Area, Cingulate Gyrus and Cerebellum) during the perception of previously learned audiovisual associations. These motor areas also showed strong functional connectivity with sensory areas (e.g., visual areas). Interestingly, Supplementary Motor Area, Cingulate Gyrus and the Cerebellum have often been associated to action preparation, action-outcome prediction and other processes linked to the comparison of predicted and actual action-outcomes (Blakemore, Wolpert, & Frith, 1999; Haggard & Whitford, 2004; Nachev, Kennard, & Husain, 2008). Finally, the active exploration group exhibited stronger activity of brain regions typically involved in audiovisual integration (e.g., STS). Similarly, these action processes might also explain the results we observed in our action conditions. Recent studies have shown that action preparation/execution leads to a modulation of the sensory areas representing the predicted outcomes of an action, event before stimulus onset (cf. pre-activation model, Desantis, Roussel et al., 2014; Kühn, Keizer, Rombouts, & Hommel, 2011; Roussel, Hughes, & Waszak, 2013; SanMiguel, Widmann, Bendixen, Trujillo-Barreto, & Schröger, 2013; Stenner, Bauer, Haggard, Heinze, & Dolan, 2014; Stenner, Bauer, Heinze et al., 2014; Waszak, Cardoso-Leite, & Hughes, 2012). One might speculate that action processes responsible for the prediction action-outcome synchronized the activity of the visual and auditory areas representing the predicted outcomes of an action. As a consequence, predicted audio and visual components of an outcome would be reported more often as simultaneous compared to the components of an unexpected pair.

Importantly, the greater WAS for predicted as opposed to unpredicted action outcomes was found only when the participants learned that the predicted outcome followed actions after a variable, rather than a fixed delay (see supplementary material for a study replicating these results). Thus, prior experience of action-outcome time intervals, tunes the "temporal window" within which the components of a learned audiovisual pair are perceived to be simultaneous. These results suggest a priority of predictions: prediction of *what* will happen can trigger strategic adjustments of perception, but only if *when* it will happen is uncertain. Temporal predictability overrides the effects of content predictability.

The notion that prior learning of action-outcome interval might shape the window of audiovisual simultaneity is supported by both research on action control and audiovisual binding. For example, sensory attenuation of predicted outcomes occurs in a specific time window, which is centred on the time at which stimulations usually occurred in our past experience (Bays et al., 2005; see also Desantis, Roussel et al., 2014). In audiovisual integration, the brain is able to recalibrate its window of audiovisual binding based on past experience of a delay between audio and visual stimuli (Fujisaki et al., 2004; Spence & Squire, 2003; Vroomen et al., 2004). Finally, recent studies showed that having an accurate temporal expectation of an incoming auditory stimulus can strongly improve audiovisual judgments of simultaneity (Petrini et al., 2009; see also Cook, Van Valkenburg, & Badcock, 2011).

In voluntary action control, adjusting the WAS based on learned action-outcome relations may contribute to sensorimotor attribution. Computational theories of motor control highlight the need to separate events that are caused by one's own actions from other, external events (Wolpert, 1997). Integrating sensory inputs outside a plausible WAS might lead to self-attribution of external events. In fact, this frequently occurs: most explicit studies of attribution agree that people over-attribute events to their own agency (Van den Bos & Jeannerod, 2002; Franck et al., 2001; Daprati et al., 1997), and psychotic patients do so to an even greater extent (Daprati et al., 1997). When temporal cues to agency are absent, then an expected sensory event is more likely to fall within this window of binding than an unexpected sensory event. Thus, *plausible* sensory events might be incorrectly self-attributed. More broadly, we suggest that over-attribution in agency judgement could be a by-product of the tendency to group multisensory events prior to attribution of the entire group. Further experiments are required to investigate this hypothesis.

Finally, one disadvantage of using simultaneity judgment task is that the size of the window of audiovisual simultaneity can be influenced by a change of decision criterion. For instance, an observer who is more liberal in responding "simultaneous" will show a large window compared with a more conservative observer (see Vroomen & Keetels, 2010). Thus, in our case, participants might show a general bias to report as simultaneous events that are congruent with previously learned associations. However, any general bias to perceive associated events as simultaneous should have affected responses to all predicted pairs, including predicted pairs preceded by visual cues (Experiment 1), and pairs that were generated by actions with fixed time intervals in the learning phase (Experiment 2).

However, we cannot entirely rule out the possibility that our results are partly driven by a change in perceptual decisions that would affect specifically action-outcomes. In other words, action-outcome prediction might have not affected sensory processing itself, but the interpretation of the readout of sensory areas. Notably, the brain might prefer to *consider* that only the audio and visual components of predicted action-outcomes should be bound together. As a consequence, participants would be more tolerant toward audio-visual asynchrony when pairs occurred as predicted outcomes of an action. Further studies should investigate the influence of perceptual decision changes and sensitivity changes on the effect we observed.

Another possible criticism to our studies would suggest that temporal discrimination was impaired not because audio and visual components were bound together in time, but because predicted outcomes were attenuated. In other words, sensory attenuation of predicted outcomes would impair time perception. However, several arguments indicate that our effect is unlikely due to sensory attenuation. Several studies have shown that action-outcomes are attenuated when participants can predict their identity (Cardoso-Leite et al., 2010; Roussel et al., 2013). However, research on intentional binding showed that this kind of prediction does not affect time perception of unimodal outcomes (Desantis, Hughes, & Waszak, 2012; Haering & Kiesel, 2014). For instance, Haering and Kiesel (2014) did not observe any effects of outcome predictability on temporal sensitivity. Thus, even though predicted outcomes are attenuated, time perception is not impaired. This suggests that the effect we observed is due to the fact that action-outcome learning and prediction binds the audio and visual feature of the action-outcome together.

To conclude, our study shows that action-outcome learning and prediction can produce temporal grouping of visual and auditory stimuli. Notably, instrumental actions that generated specific multisensory outcomes promoted the grouping of the predicted audio and visual components of the outcome. A temporally-correlated sensory cue did not promote temporal grouping in the same way. Moreover, the window of audiovisual simultaneity seems to depend on our past experiences of *when* audiovisual outcomes are likely to occur: repeated exposure to a fixed action-outcome interval would decrease our tolerance to audiovisual asynchrony

compared to a variable action-outcome interval. Taken together our results suggest that actions might represent an important factor contributing to the coherence of our perceptual experience of the external world. We believe that this process might also be important for body-ownership. Indeed, several studies showed that body ownership depends at least partially on multisensory integration (Ehrsson, Holmes, & Passingham, 2005; Tsakiris & Haggard, 2003). Action might boost such integration, thus facilitating the creation of a coherent representation of one's own body (Ma & Hommel, 2015).

Acknowledgements

This research was supported by ERC Advanced Grant HUMVOL, awarded to PH. PH was additionally supported by an ESRC Professorial Fellowship. We are grateful to the UCL Speech Hearing and Phonetic Sciences and in particular to Steve Nevard for his help with dB SPL measurements. We are also grateful to Dr. Karolina Moutsopoulou and Dr. Nobuhiro Hagura for their comments at the initial stage of this project, to Ms. Ksenia Vinogradova for her participation in the data collection of a pilot version of the current experiments. Finally, we are grateful to Ms. Catherine Bird for her participation in the data collection of the present experiments.

Appendix A. Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2016.03.009.

References

Alais, D., Newell, F. N., & Mamassian, P. (2010). Multisensory processing in review: From physiology to behaviour. *Seeing and Perceiving, 23*(1), 3–38. http://dx.doi.org/10.1163/187847510X488603.

Baess, P., Widmann, A., Roye, A., Schröger, E., & Jacobsen, T. (2009). Attenuated human auditory middle latency response and evoked 40-Hz response to self-initiated sounds. *European Journal of Neuroscience, 29*(7), 1514–1521. http://dx.doi.org/10.1111/j.1460-9568.2009.06683.x.

Bays, P. M., Wolpert, D. M., & Flanagan, J. R. (2005). Perception of the Consequences of Self-Action Is Temporally Tuned and Event Driven. *Current Biology, 15*(12), 1125–1128. http://dx.doi.org/10.1016/j.cub.2005.05.023.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (1999). The cerebellum contributes to somatosensory cortical activity during self-produced tactile stimulation. *NeuroImage, 10*(4), 448–459. http://dx.doi.org/10.1006/nimg.1999.0478.

Blakemore, S.-J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself? *NeuroReport: For Rapid Communication of Neuroscience Research, 11*(11), R11–R16. http://dx.doi.org/10.1097/00001756-200008030-00002.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. http://dx.doi.org/10.1163/156856897X00357.

Butler, A. J., James, T. W., & James, K. H. (2011). Enhanced multisensory integration and motor reactivation after active motor learning of audiovisual associations. *Journal of Cognitive Neuroscience, 23*(11), 3515–3528. http://dx.doi.org/10.1162/jocn_a_00015.

Cardoso-Leite, P., Mamassian, P., Schütz-Bosbach, S., & Waszak, F. (2010). A new look at sensory attenuation. *Psychological Science, 21*(12), 1740–1745. http://dx.doi.org/10.1177/0956797610389187.

Colonius, H., & Arndt, P. (2001). A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & Psychophysics, 63*, 126–147.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: A time-window-of-grouping model. *Journal of Cognitive Neuroscience, 16*, 1000–1009.

Cook, L. A., Van Valkenburg, D. L., & Badcock, D. R. (2011). Predictability affects the perception of audiovisual synchrony in complex sequences. *Attention, Perception & Psychophysics, 73*(7), 2286–2297. http://dx.doi.org/10.3758/s13414-011-0185-8.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition, 65*(1), 71–86. http://dx.doi.org/10.1016/S0010-0277(97)00039-5.

Desantis, A., Hughes, G., & Waszak, F. (2012). Intentional binding is driven by the mere presence of an action and not by motor prediction. *PLoS ONE, 7*(1), e29557. http://dx.doi.org/10.1371/journal.pone.0029557.

Desantis, A., Mamassian, P., Lisi, M., & Waszak, F. (2014). The prediction of visual stimuli influences auditory loudness discrimination. *Experimental Brain Research, 1–8*. http://dx.doi.org/10.1007/s00221-014-4001-2.

Desantis, A., Roussel, C., & Waszak, F. (2014). The temporal dynamics of the perceptual consequences of action-effect prediction. *Cognition, 132*(3), 243–250. http://dx.doi.org/10.1016/j.cognition.2014.04.010.

Ehrsson, H. H., Holmes, N. P., & Passingham, R. E. (2005). Touching a rubber hand: Feeling of body ownership is associated with activity in multisensory brain areas. *The Journal of Neuroscience, 25*(45), 10564–10573. http://dx.doi.org/10.1523/JNEUROSCI.0800-05.2005.

Franck, N., Farrer, C., Georgieff, N., Marie-Cardine, M., Daléry, J., D'Amato, T., & Jeannerod, M. (2001). Defective recognition of one's own actions in patients with schizophrenia. *American Journal of Psychiatry, 158*(3), 454–459. http://dx.doi.org/10.1176/appi.ajp.158.3.454.

Frens, M. A., Van Opstal, A. J., & van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics, 57*, 802–816.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773–778. http://dx.doi.org/10.1038/nn1268.

Haering, C., & Kiesel, A. (2014). Intentional Binding is independent of the validity of the action effect's identity. *Acta Psychologica, 152*, 109–119. http://dx.doi.org/10.1016/j.actpsy.2014.07.015.

Haggard, P., & Whitford, B. (2004). Supplementary motor area provides an efferent signal for sensory suppression. *Cognitive Brain Research, 19*(1), 52–58. http://doi.org/16/j.cogbrainres.2003.10.018.

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin, 139*(1), 133–151. http://dx.doi.org/10.1037/a0028566.

Jackson, C. V. (1953). Visual factors in auditory localization. *The Quarterly Journal of Experimental Psychology, 5*, 52–65. http://dx.doi.org/10.1080/17470215308416626.

James, W. (1890). *The principle of psychology* (Vol. 2)New York, NY, US: Dover Publications.

Kühn, S., Keizer, A., Rombouts, S. A. R. B., & Hommel, B. (2011). The functional and neural mechanism of action preparation: Roles of EBA and FFA in voluntary action control. *Journal of Cognitive Neuroscience, 23*, 214–220.

Ma, K., & Hommel, B. (2015). The role of agency for perceived ownership in the virtual hand illusion. *Consciousness and Cognition, 36*, 277–288.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience, 7*(10), 3215–3229.

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews Neuroscience, 9*(11), 856–869. http://dx.doi.org/10.1038/nrn2478.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. http://dx.doi.org/10.1163/156856897X00366.

Petrini, K., Russell, M., & Pollick, F. (2009). When knowing can replace seeing in audiovisual integration of actions. *Cognition, 110*(3), 432–439. http://dx.doi.org/10.1016/j.cognition.2008.11.015.

Piaget, J. (1963). *The origins of intelligence in children*. Norton.

Roseboom, W., & Arnold, D. H. (2011). Twice upon a time: Multiple concurrent temporal recalibrations of audiovisual speech. *Psychological Science, 22*(7), 872–877. http://dx.doi.org/10.1177/0956797611413293.

Roussel, C., Hughes, G., & Waszak, F. (2013). A preactivation account of sensory attenuation. *Neuropsychologia, 51*(5), 922–929. http://dx.doi.org/10.1016/j.neuropsychologia.2013.02.005.

SanMiguel, I., Widmann, A., Bendixen, A., Trujillo-Barreto, N., & Schröger, E. (2013). Hearing silences: Human auditory processing relies on preactivation of sound-specific brain activity patterns. *The Journal of Neuroscience, 33*(20), 8633–8639. http://dx.doi.org/10.1523/JNEUROSCI.5821-12.2013.

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology, 13*(13), R519–R521. http://dx.doi.org/10.1016/S0960-9822(03)00445-7.

Stein, B. E., & Meredith, M. A. (1990). Multisensory integration. Neural and behavioral solutions for dealing with stimuli from different sensory modalities. *Annals of the New York Academy of Sciences, 608*, 51–65. discussion 65–70.

Stenner, M.-P., Bauer, M., Haggard, P., Heinze, H.-J., & Dolan, R. (2014). Enhanced alpha-oscillations in visual cortex during anticipation of self-generated visual stimulation. *Journal of Cognitive Neuroscience, 26*(11), 2540–2551. http://dx.doi.org/10.1162/jocn_a_00658.

Stenner, M.-P., Bauer, M., Heinze, H.-J., Haggard, P., & Dolan, R. J. (2014). Parallel processing streams for motor output and sensory prediction during action preparation. *Journal of Neurophysiology*. http://dx.doi.org/10.1152/jn.00616.2014.

Tsakiris, M., & Haggard, P. (2003). Awareness of somatic events associated with a voluntary action. *Experimental Brain Research, 149*(4), 439–446. http://dx.doi.org/10.1007/s00221-003-1386-8.

van den Bos, E., & Jeannerod, M. (2002). Sense of body and sense of action both contribute to self-recognition. *Cognition, 85*(2), 177–187. http://dx.doi.org/10.1016/S0010-0277(02)00100-2.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the 'unity effect' reveals that speech is special. *Journal of Vision, 8*(9). http://dx.doi.org/10.1167/8.9.14. 14.1–11.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the 'unity assumption' using audiovisual speech stimuli. *Perception & Psychophysics, 69*(5), 744–756.

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72*(4), 871–884. http://dx.doi.org/10.3758/APP.72.4.871.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research, 22*(1), 32–35.

Waszak, F., Cardoso-Leite, P., & Hughes, G. (2012). Action effect anticipation: Neurophysiological basis and functional consequences. *Neuroscience & Biobehavioral Reviews, 36*(2), 943–959. http://dx.doi.org/10.1016/j.neubiorev.2011.11.004.

Welch, R. B. (1999). Chapter 15 Meaning, attention, and the 'unity assumption' in the intersensory bias of spatial and temporal perceptions. In T. B. and J. M. Gisa Aschersleben (Ed.), *Advances in Psychology* (Vol. Volume 129, pp. 371–387). North-Holland.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin, 131*(3), 460–473. http://dx.doi.org/10.1037/0033-2909.131.3.460.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences, 1*(6), 209–216. http://doi.org/16/S1364-6613(97)01070-X.