Research Report

ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Psychological Science
22(7) 872–877
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611413293
http://pss.sagepub.com
SAGE

Twice Upon a Time: Multiple Concurrent Temporal Recalibrations of Audiovisual Speech

Warrick Roseboom and Derek H. Arnold
The University of Queensland

Abstract

Audiovisual timing perception can recalibrate following prolonged exposure to asynchronous auditory and visual inputs. It has been suggested that this might contribute to achieving perceptual synchrony for auditory and visual signals despite differences in physical and neural signal times for sight and sound. However, given that people can be concurrently exposed to multiple audiovisual stimuli with variable neural signal times, a mechanism that recalibrates all audiovisual timing percepts to a single timing relationship could be dysfunctional. In the experiments reported here, we showed that audiovisual temporal recalibration can be specific for particular audiovisual pairings. Participants were shown alternating movies of male and female actors containing positive and negative temporal asynchronies between the auditory and visual streams. We found that audiovisual synchrony estimates for each actor were shifted toward the preceding audiovisual timing relationship for that actor and that such temporal recalibrations occurred in positive and negative directions concurrently. Our results show that humans can form multiple concurrent estimates of appropriate timing for audiovisual synchrony.

Keywords

audiovisual simultaneity, audiovisual recalibration, subjective synchrony, audiovisual timing, temporal recalibration, audiovisual grouping, cross-modal grouping

Received 9/28/10; Revision accepted 3/16/11

Most physical events generate signals that can be encoded in multiple sensory modalities, such as audition and vision. Because of the vast difference between the speeds of light and sound, visual and auditory signals are subject to distance-dependent changes in the relative times at which they reach people's sensory receptors (see Alais & Carlile, 2005; Arnold, Johnston, & Nishida, 2005; King, 2005; Spence & Squire, 2003; Sugita & Suzuki, 2003). On the basis of their intensity, sensory signals are also subject to changes in the rates at which they propagate through the central nervous system (Burr & Corsale, 2001; Lennie, 1981). However, at least for proximate stimuli, humans are seldom aware of any asynchrony between an events' multisensory components. One reason for this might be that the brain has some capacity to correct for extrinsically and intrinsically generated differences in sensory coding times. One plausible mechanism for this function is audiovisual temporal recalibration. Following prolonged exposure (adaptation) to streams of asynchronous auditory and visual events, timing perception can recalibrate, such that the timing of the adapted asynchronous signals seems more synchronous than it did previously (Di Luca, Machulla, & Ernst, 2009; Fujisaki, Shimojo, Kashino, & Nishida, 2004; Hanson, Heron, & Whitaker, 2008; Harrar & Harris, 2008; Heron, Whitaker, McGraw, & Horoshenkov, 2007; Keetels & Vroomen, 2007; Miyazaki, Yamamoto, Uchida, & Kitazawa, 2006; Navarra et al., 2005; Navarra, Hartcher-O'Brien, Piazza, & Spence, 2009; Vatakis, Navarra, Soto-Faraco, & Spence, 2007, 2008; Vroomen, Keetels, de Gelder, & Bertleson, 2004). This subjective recalibration of audiovisual synchrony is often accompanied by increases in the range of timing differences across which auditory and visual signals seem synchronous (Hanson, Heron, & Whitaker, 2008; Navarra et al., 2005; Vatakis et al., 2007, 2008).

Although recalibration of audiovisual timing perception could be beneficial, humans exist in a cluttered environment and can concurrently encounter many events that generate multisensory signals. Given that the signals generated by these events may be subject to different delays due to variable viewing distances and signal intensities, it might not be beneficial

Corresponding Author:
Warrick Roseboom, University of Queensland, School of Psychology, McElwain Building, St. Lucia, Brisbane, Queensland 4072, Australia
E-mail: w.roseboom@psy.uq.edu.au

to have just one timing estimate for synchronous audiovisual inputs. In the experiment reported here, we investigated whether humans can simultaneously adapt to different timing relationships for different audiovisual combinations.

Method

Procedure

Six participants,[1] all of whom reported normal or corrected-to-normal vision and hearing, completed six 1-hr experimental sessions (648 baseline trials and 648 test trials). Each session began with a baseline phase, in which participants viewed video clips of a male or female actor saying "ba" (see Supplemental Method in the Supplemental Material available online for a detailed description of the experiment). For each clip, participants were required to indicate whether the audio and video elements had occurred simultaneously. The offset of the audio relative to the video stream in these clips was manipulated (−400 ms, −300 ms, −200 ms, −100 ms, 0 ms, 100 ms, 200 ms, 300 ms, and 400 ms; negative offsets indicate that the audio stream preceded the video stream). Each baseline phase consisted of 108 trials: For each of nine audiovisual timing offsets, there were six trials with the male stimulus identity and six with the female stimulus identity; each identity was presented on the left side of the monitor in three trials and on the right side in the other three trials.

The baseline phase was followed by the adaptation phase, in which participants viewed 100 video clips, 50 presentations each of the male and female actors saying "ba." In this phase, the side of the screen on which each stimulus was presented remained constant (e.g., the female face always appeared on the left for half of the participants and always on the right for the other half of the participants). There were two adaptation conditions: In three of the six experimental sessions, clips of one actor were presented with the audio following the video by 300 ms, and clips of the other actor were presented with the audio preceding the video by 300 ms. In the other three sessions, these relationships were reversed. Participants were not required to make judgments in the adaptation phase; they were instructed only to remain fixated on a central cross while attending to the stimuli.

Participants then completed a test phase (Fig. 1). Each test presentation was preceded by a 7,200-ms adaptation top-up period, during which four clips with the same audiovisual relationships as those shown in the adaptation phase were presented for 1,800 ms each, with the male and female faces appearing on alternating sides of the screen. Participants were cued that this pretest adaptation top-up period was complete, and a test presentation was about to occur, by the central cross changing from red to white and a brief tone sounding. The test phase consisted of a single 1,700-ms clip of either the male or female actor. On congruent trials, the clip was presented on the same side of the screen as it had been during the pretest adaptation; on incongruent trials, the clip appeared on the opposite side of the screen as it had during the pretest adaptation (see Videos S1 and S2 in the Supplemental Material). As in the baseline phase, participants were instructed to make simultaneity judgments regarding the audio and video.

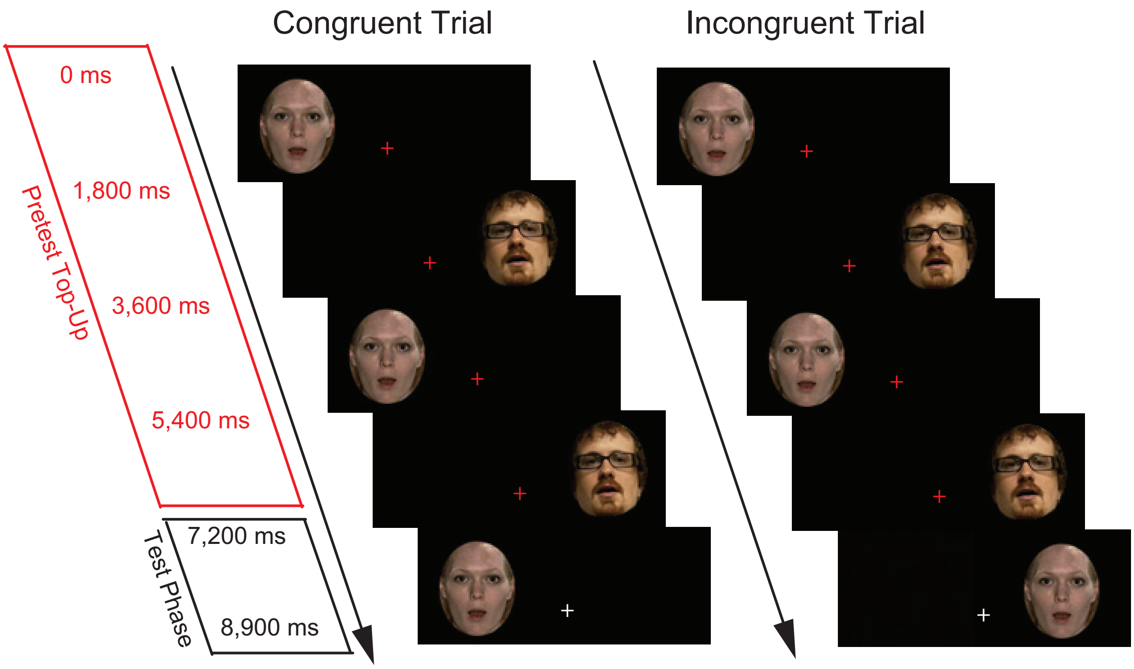

Fig. 1. Examples of the test-trial sequence. Each test trial began with a pretest adaptation top-up, in which four video clips of male and female actors saying "ba" appeared alternately to the left and right of a red central fixation cross, for 1,800 ms each. The fixation cross then turned white (at 7,200 ms), and a test stimulus (a single presentation of the male or female actor saying "ba") appeared for 1,700 ms. On congruent trials, this clip was presented on the same side of the screen as the clip featuring the same actor had been during the pretest adaptation. On incongruent trials, it appeared on the opposite side of the screen.

Data analysis

Truncated Gaussian functions were fit to distributions of reported synchrony between auditory and visual elements of stimuli as a function of the physical offset of these stimuli (mean quality of function fit for 6 participants: $\eta^2 = 0.924$, $SD = 0.022$); peaks were taken as estimates of the point of subjective synchrony (PSS; see Fujisaki et al., 2004). The full-width half-maximum (FWHM) of the distribution fits were taken as estimates of the range of audiovisual timing differences across which audio and video signals seemed synchronous. Functions were fit to data from both baseline and test phases (see Fig. 2). Differences between baseline and test PSS and FWHM estimates were calculated for each participant, for each of the two adaptation conditions. These estimates were averaged to quantify the effects of adapting to audio streams that either preceded or followed video streams.

Results

Data discussed in this section refer to congruent trials (e.g., when the female stimulus was presented on the right during both adaptation and test). We found that, overall, following adaptation to an audio stream that preceded a video stream, there was a shift in PSS between baseline and test of −32 ms, single-sample $t(5) = 14.21$, $p < .001$. Conversely, following adaptation to an audio stream that followed a video stream, there was a PSS shift between baseline and test of +36 ms, single-sample $t(5) = 3.15$, $p = .026$ (Fig. 3a). In both cases, the change in PSS was toward the adapted asynchrony. There was no difference in the absolute magnitude of temporal recalibration induced by adapting to audio that preceded or followed video, paired-sample $t(5) = 0.22$, $p = .832$. In isolation, each of these recalibrations is consistent with previous findings concerning temporal recalibration following exposure to an

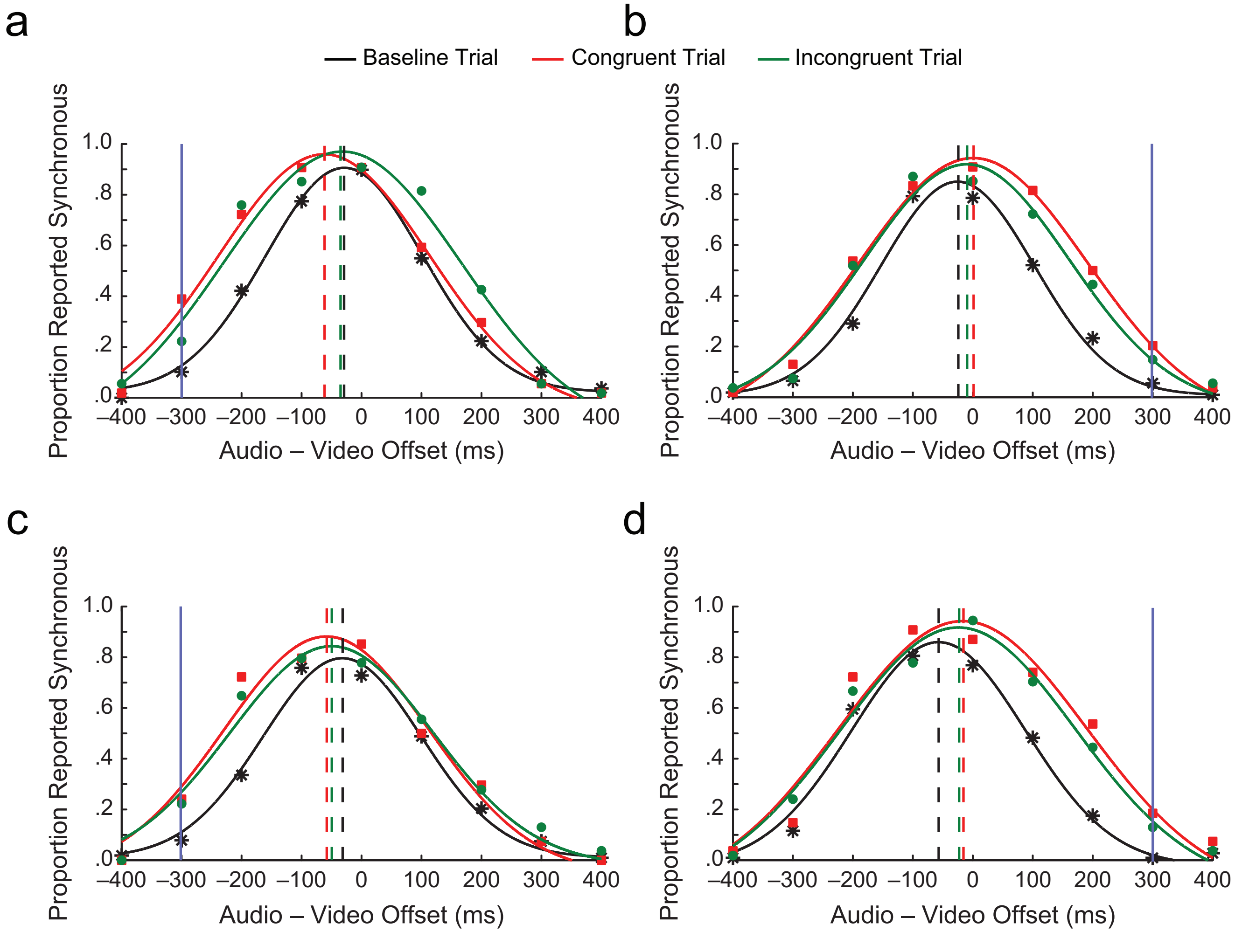

Fig. 2. Proportion of trials on which the audio and video streams were reported to be synchronous as a function of the physical timing offset of the audio stream relative to the video stream (negative numbers indicate audio preceding video). The graphs show Gaussian curves fitted to the averaged data for the baseline and test trials, and the dotted vertical lines highlight the peak of each curve. Graphs on the left show results for test trials following adaptation to stimuli in which the audio stream preceded the video stream when the adapting and test stimuli were (a) female and (c) male. Graphs on the right show results for test trials following adaptation to stimuli in which the audio stream followed the video stream when the adapting and test stimuli were (b) male and (d) female. The vertical blue lines highlight the relative offsets of the audio and video streams in the adaptation phase.

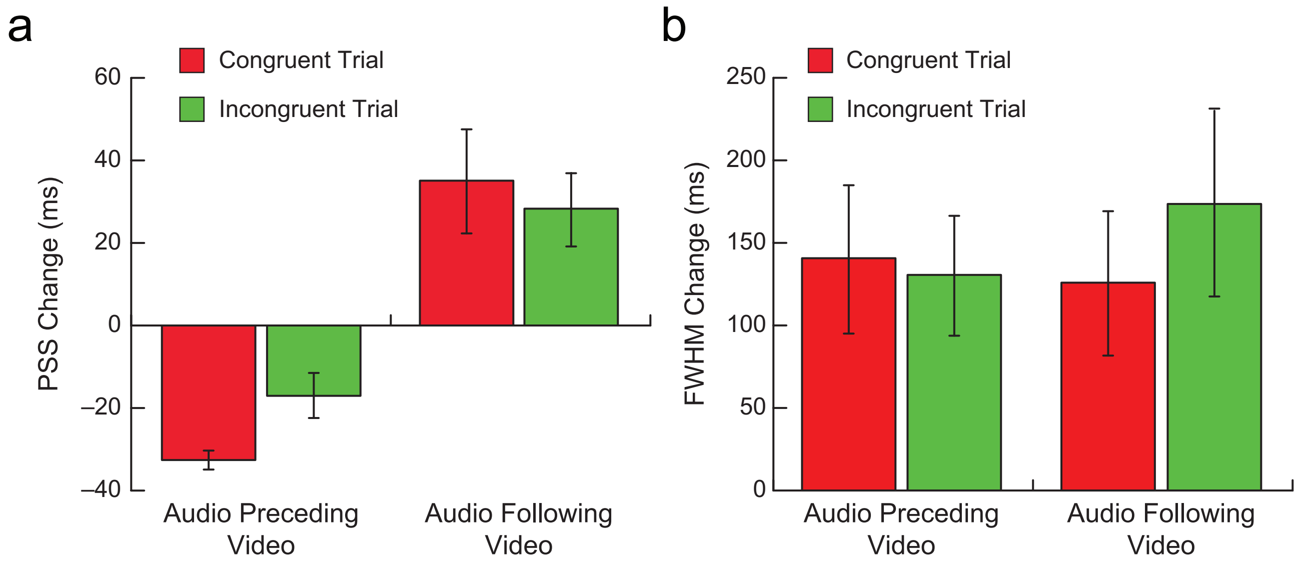

Fig. 3. Mean change in (a) point of subjective synchrony (PSS) and (b) full-width half-maximum (FWHM) between baseline trials and test trials. Results are shown separately for congruent and incongruent trials in which audio streams preceded video streams and in which audio streams followed video streams. Error bars show standard errors of the mean for the 6 participants.

asynchronous audiovisual relationship (e.g., Di Luca et al., 2009; Fujisaki et al., 2004; Hanson, Heron, & Whitaker, 2008; Harrar & Harris, 2008; Heron et al., 2007; Keetels & Vroomen, 2007; Navarra et al., 2005, 2009; Vatakis et al., 2007, 2008; Vroomen et al., 2004). However, in this experiment, we were able to induce concurrent positive and negative temporal recalibrations of subjective synchrony for the two different audiovisual pairs.

A repeated measures analysis of variance (ANOVA) confirmed this finding and revealed a significant interaction between the direction of the adapting relationship (audio preceding or following video) and experimental phase (baseline or test) in the raw PSS values, $F(1, 11) = 54.44$, $p < .001$. We also found a significant increase in the FWHM of response distributions in adaptation trials relative to baseline trials for both audio preceding video (140 ms), single-sample $t(5) = 3.1$, $p < .001$, and audio following video (126 ms), single-sample $t(5) = 3.52$, $p = .017$ (Fig. 3b).

Spatially or Contextually Based Recalibration?

We examined whether differential temporal recalibrations were tied to adapted spatial locations (left or right of fixation) or to adapted stimulus identities (male or female) by switching the locations of test stimuli such that they were incongruent with their positions during adaptation. In these circumstances, temporal recalibration effects were similar to when adaptation and test stimuli were in the same locations. Specifically, when the audio stream preceded the video stream during adaptation, there was a PSS shift between baseline and test of −17 ms, single-sample $t(5) = 2.72$, $p = .042$, and when the audio stream followed the video stream, there was a shift between baseline and test of +31 ms, single-sample $t(5) = 2.68$, $p = .044$ (Figs. 2 and 3).

Although these shifts might appear somewhat mitigated relative to when the stimuli were in the same positions during the adaptation and test phases, these reductions were not statistically significant—audio preceding video: paired-sample $t(5) = 0.91$, $p = .402$; audio following video: paired-sample $t(5) = 2.5$, $p = .055$. Analysis also revealed that audio and video signals seemed synchronous across greater audiovisual timing differences after adaptation, with FWHMs increased by 131 ms following exposure to an adaptor in which the audio preceded the video, single-sample $t(5) = 3.0$, $p = .03$, and by 174 ms following exposure to an adaptor in which the audio followed the video, single-sample $t(5) = 2.85$, $p = .036$. A repeated measures ANOVA, with direction of adapting relationship (audio preceding or following video) and stimulus presentation location (congruent or incongruent) as factors, revealed no effect of presentation location on FWHM changes, $F(1, 5) = 0.105$, $p = .759$.

Discussion

In the experiment reported here, we found that humans can simultaneously adapt to two audiovisual timing relationships, and that this adaptation results in concurrent positive and negative temporal recalibrations. Subjective estimates of audiovisual synchrony were shifted in the direction of the adapted asynchrony for each particular stimulus. It is interesting to note that the changes in subjective timing were, in all cases, associated with an increase in the range of audiovisual timing offsets across which auditory and visual signals were judged to be synchronous (the audiovisual simultaneity window). Overall, our data show that humans can maintain multiple concurrent estimates of appropriate timing for synchronous audiovisual inputs. Moreover, our data suggest that the processes by which these estimates are derived are primarily constrained by contextual information, in this case by speaker identity, as opposed to the spatial origins of visual input.

Our observation that audiovisual temporal recalibration is not constrained by spatial location is broadly consistent with the findings of previous studies (Fujisaki et al., 2004; Keetels & Vroomen, 2007; though see also Hanson, Roach, Heron, & McGraw, 2008; Yarrow, Roseboom, & Arnold, in press). Of note, however, is that the audio signals in our study were presented via headphones and, consequently, there was no difference in the location of the audio presentation when the visual stream was presented in a different location. It has previously been reported that temporal recalibration magnitude does not change when sounds are presented via headphones rather than by speakers that are colocalized with visual signals (Fujisaki et al., 2004; see also Di Luca et al., 2009). However, given the established importance of spatial correspondence in multisensory grouping (e.g., Driver & Spence, 1998; Slutsky & Recanzone, 2001; Stein & Meredith, 1993), we think the provision of stronger auditory location cues might reveal an important role for spatial correspondence in audiovisual temporal recalibration (see Hanson, Roach, et al., 2008; Yarrow et al., in press).

The difficulties faced when trying to distinguish synchrony from asynchrony are not restricted to audiovisual perception. In this context, the inherent dilemma is that due to the vastly different speeds of light and sound, auditory and visual signals reach sensory receptors at different relative times from different viewing distances (Alais & Carlile, 2005; Arnold et al., 2005; King, 2005; Spence & Squire, 2003; Sugita & Suzuki, 2003). However, as the rates at which sensory signals propagate through the central nervous system are generally correlated with signal intensity (Burr & Corsale, 2001; Lennie, 1981), it is possible to envisage situations in which signal intensities will differ across sensory modalities. For instance, a brightly lit high-contrast object may make only a soft sound on collision with another object, thus resulting in an intense visual but subtle auditory signal. Such situations could also arise within a single sensory modality. Neural analyses of visual color and motion, for instance, are relatively independent (Livingstone & Hubel, 1988). Thus, it is possible to have a stimulus that alters dramatically in color but only subtletey in direction of motion, and this could exacerbate the tendency for direction changes to seem delayed relative to the color changes (Arnold, 2005; Arnold & Clifford, 2002; Moutoussis & Zeki, 1997; Nishida & Johnston, 2002).

Given that the speeds at which sensory signals propagate through the central nervous system vary with the intensity of the input signals, humans may have to adopt dynamic strategies when judging timing relationships across a broad range of contexts. Thus, although we have demonstrated concurrent positive and negative temporal recalibrations using audiovisual speech, we believe that similar effects are possible for other combinations of sensory events, such as audio-tactile, visual-tactile, or visual color and visual motion. Such processes would ensure that, regardless of the combination of sensory signals, any possible differences in sensory signal timings could be resolved. Consistent with these views, the findings of previous studies have shown that people's sense of timing for multiple tactile stimuli is malleable (Miyazaki et al., 2006) and, more recently, that a form of temporal recalibration can occur for multiple visual events, specifically for color and direction-of-motion changes (Arnold & Yarrow, 2011).

We believe our results are important given the cluttered nature of the environment in which humans exist. In the real world, there is seldom, if ever, a single source of audiovisual information, but many concurrent sensory signals. Thus, the potential benefit of an adaptive mechanism that maintains multiple concurrent estimates of audiovisual synchrony perception is clear. Calibrating all audiovisual inputs to a single timing relationship could be dysfunctional, as distinct audiovisual stimuli will often be subject to divergent physical and neural transmission delays. But in the experiment reported here, we demonstrated that the human sensory system is capable of establishing simultaneous positive and negative temporal recalibrations for distinct audiovisual stimuli. Thus, audiovisual temporal recalibration does not necessarily generalize to all audiovisual inputs but can be specific for distinct pairings of audiovisual stimuli.

Acknowledgments

We thank Thomas Wallis and Helen Quinn for their assistance.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by an Australian Research Council Discovery Grant and an Australian Research Fellowship awarded to D. H. Arnold.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Note

1. Participants included 1 of the authors (W. Roseboom) and 5 volunteers who were naive to the purpose of the experiment. Volunteers were recruited from the University of Queensland undergraduate student population. The methods in this study were approved by the University of Queensland Human Ethics Board and conform to Declaration of Helsinki standards.

References

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

Arnold, D. H. (2005). Perceptual pairing of colour and motion. *Vision Research*, *45*, 3015–3026.

Arnold, D. H., & Clifford, C. W. (2002). Determinants of asynchronous processing in vision. *Proceedings of the Royal Society B: Biological Sciences*, *269*, 579–583.

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

Arnold, D. H., & Yarrow, K. (2011). Temporal recalibration of vision. *Proceedings of the Royal Society B: Biological Sciences*, *278*, 535–538.

Burr, D. C., & Corsale, B. (2001). Dependency of reaction times to motion onset on luminance and chromatic contrast. *Vision Research*, *41*, 1039–1048.

Di Luca, M., Machulla, T.-K., & Ernst, M. O. (2009). Recalibration of multisensory simultaneity: Cross-modal transfer coincides with a change in perceptual latency. *Journal of Vision*, *9*(12), Article 7. Retrieved from http://www.journalofvision.org/content/9/12/7

Driver, J., & Spence, C. (1998). Crossmodal attention. *Current Opinion in Neurobiology*, *8*, 245–253.

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

Hanson, J. V. M., Heron, J., & Whitaker, D. (2008). Recalibration of perceived time across sensory modalities. *Experimental Brain Research*, *185*, 347–352.

Hanson, J. V. M., Roach, N. W., Heron, J., & McGraw, P. V. (2008). Spatially specific distortions of perceived simultaneity following adaptation to audiovisual asynchrony. *Perception*, *37*(ECVP Abstract Suppl.), 27.

Harrar, V., & Harris, L. R. (2008). The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity. *Experimental Brain Research*, *186*, 517–524.

Heron, J., Whitaker, D., McGraw, P. V., & Horoshenkov, K. V. (2007). Adaptation minimizes distance-related audiovisual delays. *Journal of Vision*, *7*(13), Article 5. Retrieved from http://www.journalofvision.org/content/7/13/5.short

Keetels, M., & Vroomen, J. (2007). No effect of auditory-visual spatial disparity on temporal recalibration. *Experimental Brain Research*, *182*, 559–565.

King, A. J. (2005). Multisensory integration: Strategies for synchronization. *Current Biology*, *15*, R339–R341.

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

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, *240*, 740–749.

Miyazaki, M., Yamamoto, S., Uchida, S., & Kitazawa, S. (2006). Bayesian calibration of simultaneity in tactile temporal order judgement. *Nature Neuroscience*, *9*, 875–877.

Moutoussis, K., & Zeki, S. (1997). A direct demonstration of perceptual asynchrony in vision. *Proceedings of the Royal Society B: Biological Sciences*, *264*, 393–399.

Navarra, J., Hartcher-O'Brien, J., Piazza, E., & Spence, C. (2009). Adaptation to audiovisual synchrony modulates the speeded detection of sound. *Proceedings of the National Academy of Sciences, USA*, *106*, 9123–9124.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, *25*, 499–507.

Nishida, S., & Johnston, A. (2002). Marker correspondence, not processing latency, determines temporal binding of visual attributes. *Current Biology*, *12*, 359–368.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport*, *12*, 7–10.

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology*, *13*, R519–R521.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Sugita, Y., & Suzuki, Y. (2003). Audiovisual perception: Implicit estimation of sound-arrival time. *Nature*, *421*, 911. doi:10.1038/421911a

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Temporal recalibration during asynchronous audiovisual speech perception. *Experimental Brain Research*, *181*, 173–181.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: Temporal order versus simultaneity judgments. *Experimental Brain Research*, *185*, 521–529.

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

Yarrow, K., Roseboom, W., & Arnold, D. H. (in press). Spatial grouping resolves ambiguity to drive temporal recalibration. *Journal of Experimental Psychology: Human Perception and Performance*.