Spatial grouping resolves ambiguity to drive temporal recalibration

Kielan Yarrow [1*], Warrick Roseboom [2] & Derek H Arnold [2]

1. Department of Psychology,
City University London

2. School of Psychology,
The University of Queensland

Running head: Spatial grouping drives temporal recalibration

* Author for correspondence:

Kielan Yarrow,
Social Science Building,
City University,
Northampton Square,
London EC1V 0HB

Tel: +44 (0)20 7040 8530
Fax: +44 (0)20 7040 8580
Email: kielan.yarrow.1@city.ac.uk

Abstract

Cross-modal temporal recalibration describes a shift in the point of subjective simultaneity (PSS) between two events following repeated exposure to asynchronous cross-modal inputs – the adaptors. Previous research suggests that audio-visual recalibration is insensitive to the spatial relationship between the adaptors. Here we show that audio-visual recalibration can be driven by cross-modal spatial grouping. Twelve participants adapted to alternating trains of lights and tones. Spatial position was manipulated, with alternating sequences of a light then a tone, or a tone then a light, presented on either side of fixation (e.g. left tone - left light - right tone - right light etc.). As the events were evenly spaced in time, in the absence of spatial-based grouping it would be unclear if tones were leading or lagging lights. However, any grouping of spatially co-localised cross-modal events would result in an unambiguous sense of temporal order. We found that adapting to these stimuli caused the PSS between subsequent lights and tones to shift *toward* the temporal relationship implied by spatial-based grouping. These data therefore show that temporal recalibration is facilitated by spatial grouping.

Introduction

Imagine you have gone to the cinema, and are irritated to discover that the soundtrack is temporally misaligned with the images: It has a noticeable lead. However, you are surprised to find that the asynchrony becomes more bearable as time passes. Finally, upon leaving the cinema you thank the usher, and are shocked to find that his facial movements seem strangely detached from his reply: His voice seems to lag behind. You have adapted to the cross-modal temporal misalignment in the cinema and are now out of synch with the real world.

Could this actually happen? Persistent exposure to temporally offset sights and sounds can indeed bring about a temporal realignment of vision and audition (Di Luca, Machulla & Ernst, 2009; Fujisaki, Shimojo, Kashino & Nishida, 2004; Hanson, Heron & Whitaker, 2008; Harrar & Harris, 2008; Heron, Roach, Whitaker & Hanson, 2010; Navarra, Hartcher-O'Brien, Piazza & Spence, 2009; Vroomen, Keetels, de Gelder & Bertelson, 2004; Vroomen & Keetels, 2010). However, the effect tends to be small (~25 ms) and as such may not be readily apparent in daily conversation. The original reports used beeps and flashes (Fujisaki et al., 2004; Vroomen et al., 2004), perhaps suggesting recalibration at a firework display rather than at a bad movie. Subsequent studies have found recalibration in situations closer to the scenario described by using video and soundtrack stimuli (Navarra et al., 2005; Vatakis, Navarra, Soto-Faraco, & Spence, 2007; Vatakis, Navarra, Soto-Faraco, & Spence, 2008), but again, the effect was modest (~15 ms). Nonetheless, recalibration effects have strong implications for our understanding of temporal perception. They suggest that our sense of multisensory timing is more flexible

than straightforward accounts based on hardwired differences in neural processing times might imply (e.g. Paillard, 1949; Schroeder & Foxe, 2004).

Here, we will focus on one rather perplexing feature of the phenomenon as described to date: The spatial congruence of the adapting events does not seem to matter. Many other multisensory interactions show a strong dependence on spatial coincidence – such that it is common to speak of a "spatial rule" in multisensory binding (Holmes & Spence, 2005; Stein & Stanford, 2008). In contrast to this, Fujisaki et al. (2004) found that the magnitude of temporal recalibration was almost identical when the sound was presented over headphones compared to when it was presented from a hidden speaker positioned directly below the visual stimulus. Similarly, Keetels & Vroomen (2007) combined an LED flash directly in front of their participants with a sound burst presented from either the same location or from a position directly to the left or right. Recalibration did not differ statistically between these two kinds of adaptor.

To explain these negative findings, it may help to consider the deliberately sparse adaptation conditions in a typical recalibration experiment. Each bisensory pair of adapting events is repeated many times in a consistent relationship. Critically, these pairs can easily be grouped on the basis of temporal proximity, because the interval between each presentation of a bisensory pair greatly exceeds the offset between the paired events. Hence the experiment is set up to generate strong temporal proximity-based grouping. Thus it may not be surprising that additional cues pertinent to the binding of each bisensory pair have little power to further affect grouping, and thus the degree of temporal recalibration.

In this paper we introduce a simple manipulation which allows us to show clearly that spatial coincidence does in fact influence cross-modal temporal recalibration: We remove temporal cues to grouping for our bisensory events, while providing spatial cues that generate an implied direction of temporal asynchrony. Spatial location is known to provide a powerful cue for the grouping of perceptual elements. In audition, for example, auditory stimuli that are usually interpreted as parts of a single stream can segregate into multiple streams when presented from different spatial locations (Bregman, 1990).

The adaptor sequence that we have used is depicted in Figure 1. Observers were exposed to a train of alternating flashes and beeps that were equally spaced in time, such that any flash could be interpreted as leading the subsequent beep or lagging the preceding one. However, the spatial position from which the flashes and beeps arose could be used as a cue to disambiguate this situation, implying a constant asynchrony between sequential pairs of events presented to one side of fixation. To anticipate our results: We found robust cross-modal temporal recalibration in a direction consistent with events having been interpreted according to a spatial rule.

<INSERT FIGURE 1 AROUND HERE>

Methods

Design

The repeated-measures design comprised two adaptation conditions: *Light lagging* and *light leading*. The interval between lights and tones (200 ms) was physically

identical in the two conditions. The labels therefore reflect the temporal relationship *implied* by the spatial arrangement of the stimuli. For the light lagging condition, the adaptation train contained multiple repetitions of the sequence left tone – left light – right tone – right light; for the light leading condition, it was left tone – right light – right tone – left light. Hence spatial grouping implied a consistent lag or lead where none really existed. The order in which participants completed the two conditions was counterbalanced.

After the presentation of an adapting sequence, participants were shown pairs of test events separated by 11 possible stimulus onset asynchronies (SOAs: -350, -250, -150, -100, -50, 0, 50 100, 150, 250, 350; negative numbers denote lights before tones). During a block of trials, participants were shown each of these 11 timing relationships on 10 occasions, all in a pseudorandom order. This yielded 110 trials per block. Each participant completed two blocks of trials for each condition, therefore 220 trials per condition.

Participants

Twelve naïve participants (8 male) with normal or corrected-to-normal vision took part in exchange for either money or course credits. All procedures were approved by the University of Queensland School of Psychology ethics committee.

Apparatus & Stimuli

A PC running Matlab (The MathWorks; U.S.A.) interfaced with a RX8 Multi I/O Processor (Tucker-Davis Technologies) produced stimuli at 100 kHz. The RX8 Multi I/O

Processor controlled two green light-emitting diodes (LEDs) mounted on two speakers, as well as a central yellow fixation LED. The fixation LED was located 57 cm in front of the observer, while the speakers and green LEDs were located 25cm to the left and right of fixation. The participant's head was supported by a chin rest, with eyes approximately 25 cm above the speakers. Peripheral LED flashes lasted 10 ms plus a 5 ms linear onset/offset ramp. Auditory stimuli were 10 ms 1000 Hz pure tones with a 5 ms linear onset/offset ramp.

Procedure

A block of trials began with 60 seconds of adaptation, while participants fixated the central LED. Peripheral stimuli were presented at a constant rate of 5 Hz alternating between beeps and flashes. A pattern of four stimuli, two on the left and two on the right, was presented repeatedly (see design). Following adaptation, a trial was signalled by the brief offset then onset of the central LED. Audiovisual pairs were then presented (with the auditory components beginning 500 ms after central LED onset) simultaneously on both sides of fixation (i.e. two synchronous lights with an SOA relative to two synchronous tones). Participants judged whether the test lights and tones had been synchronous or asynchronous. Two seconds later, a top up adaptation train was presented for 5 s (i.e. 6.25 repetitions of the four-stimulus pattern) before the next trial commenced. To ensure that adaptation was robust, a second full (i.e. 60 s) adaptation train was presented in the middle of each block (i.e. every 55 trials).

Data analysis

The proportion of times that each participant judged audiovisual pairs as synchronous was determined for each SOA in each condition. Data were fitted with a difference of cumulative Gaussians function, which is the model implied if observers categorise the difference in arrival time between the auditory and visual stimuli by saying "synchronous" if the difference falls between two criteria (see Schneider & Bavelier, 2003, appendix A.1 for a derivation)[1]. A maximum-likelihood fit was obtained using the Nelder-Mead simplex algorithm (Nelder & Mead, 1965; O'Neill, 1971) to estimate the point of subjective simultaneity (PSS). The model also yielded two additional parameters, reflecting the typical placement of criteria for simultaneity, and noise (in transmission latencies and/or the consistency with which criteria were maintained). Standard two-tailed parametric tests were used to assess differences in these parameters across conditions.

Results

Figure 2 part A shows raw data alongside the MLE fit for the combined data from all participants. Figure 2 part B shows equivalent data for one naïve participant, selected because their individual PSSs closely matched the sample mean values. Stimulus onset asynchronies, shown along the x axis, denote the time of the light relative to the beep in test trials (i.e. negative SOAs imply the light came first). In general, the fitting procedure captured the qualitative features of the data well. Model fits were assessed formally using the deviance statistic. If the model is a good one (and to the extent that asymptotic approximations hold) deviance should follow a chi-squared distribution and exceed 19.68

only 5% of the time (Wichmann & Hill, 2001). This value was exceeded in only 2/24 individual fits.

<INSERT FIGURE 2 AROUND HERE>

PSS estimates were calculated for each participant based on the best fit to their data, with a negative value indicating that, on average, the light had to be presented *before* the tone to be judged as simultaneous. The group mean PSSs are shown in Figure 2 part C. The PSS was slightly negative in the light lagging condition (-19.5 ms) and showed a more pronounced negative bias in the light leading condition (-56.0 ms), showing that the spatial cues in the two adaptation sequences differentially influenced participants' sense of audio-visual synchrony. Importantly, the PSS was shifted in the direction of the implied adapting asynchrony (true in 11/12 participants). This difference was confirmed with a paired-sample t-test ($t = 4.52$, $df = 11$, $p = 0.001$). Additional parameters derived from the model fits, shown in Table 1, did not differ reliably between conditions[2].

<INSERT TABLE 1 AROUND HERE>

Discussion

We presented participants with two kinds of adaptation trains consisting of lights and tones. The trains had identical (and ambiguous) temporal properties which would not

be expected to generate strong and consistent grouping into bimodal pairs. However, they differed in their spatial properties, such that the trains could be grouped into bimodal pairs coming from one side and then the other in alternation. Our design ensured that the consistent matching of audiovisual elements could be achieved easily: We used stimulus pairs repeating at 2.5 Hz, whereas synchrony judgements only break down at around 4 Hz (Fujisaki & Nishida, 2005). Hence this spatial grouping should have implied a consistent lag or lead between matched bimodal pairs.

Test stimuli presented at a range of SOAs were used to determine points of subjective simultaneity after adaptation. The mean PSS differed reliably between the two conditions, in line with their having shifted in the direction of the adapting asynchrony implied by the spatial arrangement of adaptors. Hence our spatial grouping cue appeared to resolve the temporal ambiguity regarding the pairing of bimodal events, and thus gave rise to a consistent interpretation which evidently drove an audio-visual temporal recalibration. The success of our spatial grouping cue is consistent with much previous research which suggests that spatial coincidence is important when grouping multisensory events (Holmes & Spence, 2005; Stein & Stanford, 2008).

We obtained a negative PSS in both of our adaptation conditions, which may seem surprising. However, the unadapted PSS for audiovisual stimuli is consistently found to occur when lights precede sounds (reviewed in van Eijk, Kohlrausch, Juola & van de Par, 2008) so a value in the range of -20 to -55 ms, as implied here, is reasonable. We did not take a baseline measure in our experiment because there was no need to do so. Our basic claim is that recalibration can, under the right circumstances, depend on spatial cues to stimulus grouping. Evidencing this claim requires only that we

demonstrate differences in the PSS when spatial grouping cues differ but other grouping cues remain constant. These differences in the PSS imply a different magnitude of recalibration in the two conditions, because there are no other reasonable mechanisms by which the PSSs might have come to differ. Of course, recalibration relative to baseline may have occurred either in both conditions or in just one of them, but this is irrelevant to the logic of our demonstration.

We would like to emphasise that our adapting trains were identical in all respects relevant to implied grouping *except* for the spatial cues that we deliberately inserted. It could be argued that our temporally-ambiguous adaptation trains actually contained temporal cues to grouping, because we used an asynchrony (+/-200 ms) that was objectively ambiguous, but may not have been subjectively ambiguous given the baseline bias outlined above. However, such a tendency would have encouraged lights to group with succeeding sounds *regardless of experimental condition*. Similarly, participants could perhaps segregate an ambiguous train of this kind by grouping the first pair of stimuli they received together and then repeating this grouping strategy for the remainder of the adaptation period. However, both of our trains began with a tone followed by a light, so the implied grouping would again be identical with respect to this cue.

Why did we obtain a spatial modulation of audio-visual temporal recalibration when previous attempts have failed to do so (Fujisaki et al., 2004; Keetels & Vroomen, 2007)? We suspect that those authors used adaptation trains with such strong temporal cues to grouping that spatial cues could do little to affect the perceived pairings. We suggest that when a single audio and a single visual input are presented close together in time, they are likely to group despite the spatial arrangement. Essentially, what we have

done here is to increase our experiment's sensitivity to detect spatial modulation by presenting a greater number of distinct audio and visual events. Recently it has been shown that this simple manipulation can have profound effects on measures of audio-visual simultaneity (Roseboom, Nishida & Arnold, 2009).

In summary, we have demonstrated that spatial cues can be used to group bimodal stimuli and bias audio-visual temporal recalibration. Previous data suggested that the mechanism that implements temporal recalibration was injudicious, in that it seemed to respond equally to all possible combinations of sensory events, modulated only by their degree of temporal separation. Our finding is important because it shows that other contextual information can affect audio-visual temporal recalibration. This, of course, fits with the intuition that adaptive behaviour should be smart, not stupid.

Acknowledgements

DHA is supported by an Australian Research Council discovery project grant and fellowship. KY visited DHA's lab thanks to a Royal Society travel grant. The authors thank John Burkardt for making his Matlab Nelder-Mead simplex code available online.

References

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge: MIT Press.

Di Luca, L. 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,* 7-16.

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

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

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

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., Roach, N. W., Whitaker, D., & Hanson, J. V. (2010). Attention modulates the plasticity of multisensory timing. *European Journal of Neuroscience, 31,* 1755-1762.

Holmes, N. P. & Spence, C. (2005). Multisensory integration: space, time and superadditivity. *Current Biology, 15,* R762-R764.

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

Navarra, J., Hartcher-O'Brien, J., Piazza, E., & Spence, C. (2009). Adaptation to audiovisual asynchrony modulates the speeded detection of sound. *Proceedings of the National Academy of Sciences of the U.S.A, 106,* 9169-9173.

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. *Brain Research: Cognitive Brain Research, 25,* 499-507.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7,* 308-313.

O'Neill, R. (1971). Algorithm AS 47: Function minimization using a simplex procedure. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 20,* 338-345.

Paillard, J. (1949). Quelques données psychophysiologiques relatives au déclenchement de la commande motrice [Some psychophysiological data relating to the triggering of motor commands]. *L'Année Psychologique, 48,* 28-47.

Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision, 9,* 4-8.

Schneider, K. A. & Bavelier, D. (2003). Components of visual prior entry. *Cognitive Psychology, 47,* 333-366.

Schroeder, C. E. & Foxe, J. J. (2004). Multisensory Convergence in Early Cortical Processing. In Calvert G.A., C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 295-309). Cambridge, MA: MIT Press.

Stein, B. E. & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9,* 255-266.

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

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. (2010). Perception of intersensory synchrony: a tutorial review. *Attention, Perception & Psychophysics, 72,* 871-884.

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

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics, 63,* 1293-1313.

Yarrow, K., Jahn, N., Durant, S., & Arnold, D. H. (submitted for publication). Shifts of criteria or neural timing? The assumptions underlying timing perception studies.

Footnotes

[1] Simultaneity judgements are often fitted using a Gaussian or truncated Gaussian function, which provides a shape quite similar to the difference of two cumulative Gaussians we employed, but has no detection-theoretic rationale.

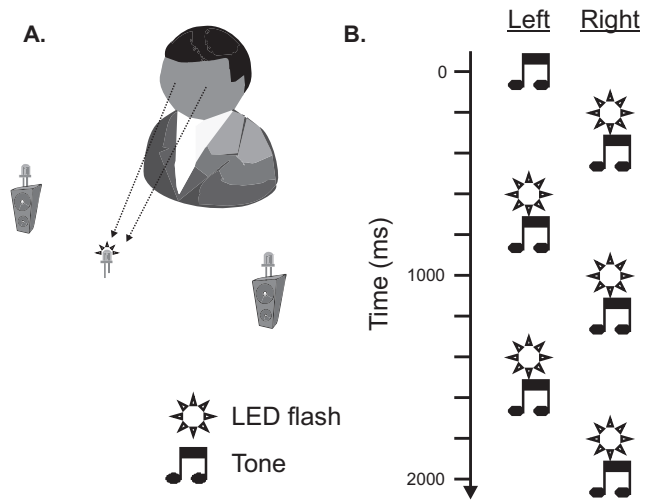
[2] For a fuller discussion of these parameters and their interpretation, see Yarrow, Jahn, Durant and Arnold (submitted for publication).

Table 1. *Additional Parameters Derived From Model Fits in the Light Leading and Light Lagging Conditions*

Condition	Noise		Criterion Extent		Deviance of fit	
	Mean	SE	Mean	SE	Mean	SE
Light Leading	105.0	20.8	191.7	19.4	10.1	1.6
Light Lagging	120.2	26.2	183.4	16.0	12.0	1.5
p (Paired T-Test)	0.065		0.462		0.291	

Note: Under the difference of cumulative Gaussians model, noise can arise from variability in arrival latencies and/or in the trial-to-trial placement of criteria for judging synchrony. Criterion extent is the distance from the PSS to either of two criteria which are used to define a range of central arrival latencies which will be judged synchronous.

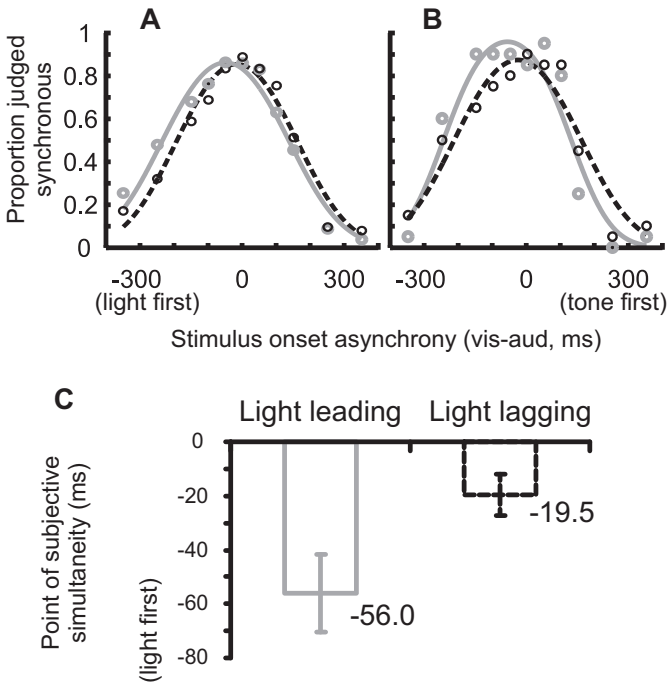
Figure 1



Legend to Figure 1

Schematic showing the position of experimental apparatus (A) and the adaptation procedure (B). Lights and tones were presented in alternation, with grouping implied by the spatial coincidence of each event with just one of the two temporally-adjacent events. This example would imply light-leading grouping. LED = Light emitting diode.

Figure 2



Legend to Figure 2

Results of the spatially-implied asynchrony experiment. (A) Combined data from all participants, alongside the fit provided by a difference of cumulative Gaussian model. Data is shown separately for the light leading (grey solid lines) and light lagging (black dashed lines) conditions. (B) Equivalent data and fit for a single participant. (C) Mean points of subjective simultaneity across all participants (derived from fits like those shown in part B for each adaptation condition and each participant). Error bars show standard error of the mean.