Rapid, generalized adaptation to asynchronous audiovisual speech

Erik Van der Burg and Patrick T. Goodbourn

School of Psychology, Brennan MacCallum Building (A18), University of Sydney, Sydney, NSW 2006, Australia

PTG, 0000-0001-7899-7355

The brain is adaptive. The speed of propagation through air, and of low-level sensory processing, differs markedly between auditory and visual stimuli; yet the brain can adapt to compensate for the resulting cross-modal delays. Studies investigating temporal recalibration to audiovisual speech have used prolonged adaptation procedures, suggesting that adaptation is sluggish. Here, we show that adaptation to asynchronous audiovisual speech occurs rapidly. Participants viewed a brief clip of an actor pronouncing a single syllable. The voice was either advanced or delayed relative to the corresponding lip movements, and participants were asked to make a synchrony judgement. Although we did not use an explicit adaptation procedure, we demonstrate rapid recalibration based on a single audiovisual event. We find that the point of subjective simultaneity on each trial is highly contingent upon the modality order of the preceding trial. We find compelling evidence that rapid recalibration generalizes across different stimuli, and different actors. Finally, we demonstrate that rapid recalibration occurs even when auditory and visual events clearly belong to different actors. These results suggest that rapid temporal recalibration to audiovisual speech is primarily mediated by basic temporal factors, rather than higher-order factors such as perceived simultaneity and source identity.

1. Introduction

We combine information from different sensory modalities to create a rich and coherent representation of the external world [1–5]. Multisensory integration conveys numerous benefits: for instance, in a noisy environment, we can better understand a speaker when we can also observe the speaker’s lip movements [6]. Typically, the perceptual benefits of audiovisual integration are optimal when information from the two different senses is presented in close temporal proximity, and declines with increasing temporal asynchrony [2,7,8].

In natural scenes, audiovisual events deriving from the same source are synchronized at their origin. However, owing to physical and neurological constraints, there are likely to be significant cross-modal differences in latency from the perspective of the observer. Sound travels more slowly than light; and conversely, neural transduction is typically faster in the auditory system than in the visual system [9]. Other factors, such as the allocation of spatial attention [10–13], and the intensity of stimuli [14–17], can also affect the speed of neural processing for auditory and visual information. Yet, our brain seems to compensate for the resulting latency differences by adapting to asynchronous multisensory information [18,19]. Vroomen et al. [19], for example, exposed participants to 240 audiovisual events—an auditory tone in combination with a visual flash—with a fixed temporal offset, for a total duration of 3 min. Following this adaptation procedure, participants saw the tone and flash presented across a range of stimulus onset asynchronies (SOAs), and were asked after each trial whether the audiovisual information was presented simultaneously or successively. They estimated the point of subjective simultaneity (PSS) by fitting a psychometric function and found strong evidence of temporal recalibration: the PSS was shifted towards the modality that was leading in the adaptation phase. Note that each experimental trial was preceded by a re-exposure phase (eight audiovisual events with the same temporal offset as in the adaptation phase) to top up adaptation before each trial. Such lengthy top-up procedures
have since been shown to be critical to maintaining audiovisual temporal recalibration in this paradigm [20].

Typically, studies of temporal recalibration have used prolonged periods of asynchronous audiovisual adaptation—often lasting for several minutes—suggesting that adaptation may be rather sluggish [18,19,21–26] (but see [27]). Recently, however, Van der Burg et al. [28] have shown that adaptation to asynchronous audiovisual events occurs far more rapidly than was previously thought. Participants were asked to judge whether or not a flash was synchronized with a tone (that is, they performed a classical simultaneity judgement task) across a range of audiovisual SOAs. They also showed that the PSS on the current trial \( t \) was highly contingent upon the modality order of stimulus on the preceding trial \( t−1 \). More specifically, the PSS on trial \( t \) was shifted towards the leading modality on trial \( t−1 \). These inter-trial effects suggest that adaptation to asynchronous multisensory information is fast, and can occur in the wake of a single audiovisual event, in the absence of an explicit adaptation procedure.

Van der Burg et al. [28] proposed that a fast-acting recalibration mechanism serves to instantaneously realign signals at onset, in order to maximize the perceptual benefits of audiovisual integration. For instance, rapid adaptation to the first asynchronous audiovisual event in a stream of speech would optimize comprehension for the remainder of the stream. However, the perception of audiovisual speech is special: it exhibits enhanced multimodal integration relative to artificial stimuli [29], and relies on specialized neural structures [30]. It is thus important to determine whether rapid recalibration does in fact occur in the case of auditory speech in combination with the corresponding visual lip movements.

Furthermore, the perceived timing of audiovisual events is known to depend critically on the nature of the stimuli. Dixon & Spitz [31], for example, measured the minimum temporal offset required to detect asynchrony between an auditory and a visual event, and found that participants needed a greater offset in order to detect asynchrony in a clip of audiovisual speech than in a clip of a hammer hitting a peg. They thus proposed that tolerance to audiovisual asynchrony is greater when stimuli are very familiar, as is the case for speech. It is feasible, then, that participants might rapidly adapt to audiovisual asynchrony for unfamiliar, inanimate stimuli (such as a visual flash and an auditory tone; see [28]), but not for familiar speech stimuli.

Temporal recalibration has been found to generalize across different artificial stimuli, whether induced by prolonged [18,25] or brief [32] exposures. By contrast, Roseboom & Arnold [21] found that for audiovisual speech stimuli, participants concurrently maintained two different recalibrations—one with a positive shift in PSS, the other with a negative shift—when each was associated with a different actor. Their result cannot be attributed solely to location-specific recalibration [23,33], because it was still evident when actors switched locations after the adaptation phase. Recalibration for audiovisual speech may thus be uniquely stimulus-specific. However, the extent to which rapid recalibration generalizes across different audiovisual speech stimuli has not been explored previously.

In this study, participants were shown a brief movie clip of an actor saying a single syllable. The voice either preceded or followed the corresponding lip movements across a range of SOAs (0, 50, 100, 200, 400 or 800 ms), and participants were asked to judge whether or not the lip movements were synchronized with the voice. We show that participants rapidly recalibrate to asynchronous audiovisual speech stimuli: the PSS on a given trial was highly contingent upon the modality order on the preceding trial. Moreover, we present compelling evidence that rapid recalibration generalizes across different audiovisual speech stimuli (Experiment 1), and across different actors (Experiments 2 and 3). Finally, we demonstrate that recalibration occurs following auditory and visual events that are not perceived to be simultaneous, and even when auditory and visual stimuli belong to different actors—such as when a male actor’s lip movements are paired with a female actor’s voice (Experiment 3). These results suggest that rapid temporal recalibration is primarily driven by temporal factors, with relatively little influence of higher-order factors such as perceived simultaneity and source identity.

2. Experiment 1

In Experiment 1, we examined whether participants rapidly adapt to asynchronous audiovisual speech stimuli. Participants were shown a brief clip of a male actor (figure 1a) saying a single syllable. The voice was temporally offset from the corresponding lip movement by a range of SOAs, and participants were asked to judge whether or not the lip movement was synchronized with the voice. If participants rapidly adapt to asynchronous audiovisual speech stimuli, then we expect the PSS to be contingent upon the modality order on the preceding trial. We also randomly interleaved two different syllables to examine whether rapid recalibration generalizes across different audiovisual speech stimuli.

(a) Material and methods

(i) Participants

Ten volunteers participated in the experiment (four females; mean age 27 years, ranging from 18 to 38 years). Eight participants were naive to the purpose of the experiment; the other two participants were the authors.

(ii) Stimuli and apparatus

The experiment was programmed and run using E-prime software (Psychology Software Tools, Sharpsburg, PA, USA). Stimuli were presented on a 18-inch Trinitron E400 monitor (Sony Corporation, Tokyo, Japan) running at a spatial resolution of 1024 × 768 pixels and a refresh rate of 85 Hz. Participants were seated at a comfortable distance from the monitor (approximately 60 cm) and wore HD 380 Pro circumaural headphones (Sennheiser, Hanover, Germany), which delivered auditory stimuli with a peak intensity of approximately 70 dB. Movie clips were recorded of an Australian male actor (David;
The mean PSS was presented four times in. Clips were cropped to 2 s in duration, and QT Sync software (Michael Gerbes, Potsdam, Germany) was used to manipulate the SOA between the visual and auditory components. A noise-gating filter was applied to the audio such that the sound level ramped on and off over about 50 ms immediately before and after the syllable, while the remainder of the clip was silent. In total, 11 clips were created for each syllable, corresponding to the 11 SOAs used in the experiment.

(iii) Design and procedure
Each trial started with the presentation of a white fixation dot (0.3° radius; 60 cd m$^{-2}$) on a black background (less than 0.5 cd m$^{-2}$) for a duration of 1 s. The fixation dot was replaced by a movie clip (16.8 x 12.0° wide) of 2 s duration, depicting the actor saying the syllable /ba:/ or /em/. The auditory signal was the original voice, which was either advanced or delayed relative to the lip movement across a range of SOAs (0, 50, 100, 200, 400 or 800 ms). After presentation of the clip, a white question mark was displayed. Participants reported whether or not the lip movement and voice were synchronized by pressing the ‘1’ key or the ‘0’ key, respectively. The fixation dot for the next trial appeared after this unspeeded response. Prior to the experiment, participants performed practice trials to gain familiarity with the task. Each combination of modality order (audition leads or vision leads), SOA (0, 50, 100, 200, 400 or 800 ms) and syllable (/ba:/ or /em/) was presented four times in each block, in a random order. Participants performed two sessions, each containing four experimental blocks of 96 trials, leading to 768 trials in total. As our analyses are contingent on trial t − 1, the first trial of each block was excluded.

(b) Results and discussion
We examined whether the probability of a ‘simultaneous’ response on a given trial (t) was reliably affected by the modality order of the previous trial (t − 1). We fit a Gaussian distribution to the proportion of ‘simultaneous’ responses as a function of SOA, separately for each participant, t − 1 modality order (audition leads or vision leads) and syllable congruency (same or different syllable t versus t − 1). There were three free parameters to the fit: the mean, which we take to reflect the PSS; the standard deviation, which we take to reflect the bandwidth of the window of simultaneity; and the amplitude, which reflects the total proportion of ‘simultaneous’ responses. Critically, any effect of temporal recalibration should be evident in a comparison of the PSS parameter between experimental conditions. Accordingly, we focus here on analyses of PSS only. For completeness, we present analyses of the bandwidth and amplitude parameters in the electronic supplementary material; however, none of those analyses reveals effects that would change the interpretation of results presented here. Also in the electronic supplementary material, we report confirmatory non-parametric analyses for the PSS parameter.

The results of Experiment 1 are depicted in figure 2. Figure 2a presents data, collapsed across participants, for trials on which the syllables differed between t and t − 1. The proportion of ‘simultaneous’ responses is shown as a function of SOA and t − 1 modality order, along with Gaussian fits to the data. Figure 2b presents the corresponding data for trials on which the syllables were the same for t and t − 1. The mean PSS, bandwidth and amplitude across all conditions were 41 ms, 208 ms and 0.95, respectively. PSS was entered as the dependent variable in an analysis of variance (ANOVA) with t − 1 modality order and t versus t − 1 syllable congruency as within-subject factors. The ANOVA yielded a reliable main effect of t − 1 modality order, $F(1, 9) = 9.25, p = 0.014, \eta^2_p = 0.51$, reflecting the fact that the PSS was significantly shorter when audition led vision on the previous trial (31 ms) than when vision led audition on the previous trial (52 ms). The main effect of syllable congruency was not significant, $F(1, 9) = 0.17, p = 0.691, \eta^2_p = 0.02$; nor was the two-way interaction, $F(1, 9) = 0.05, p = 0.835, \eta^2_p = 0.01$.

Although we did not use a prolonged adaptation procedure, we found compelling evidence that adaptation to asynchronous audiovisual speech stimuli occurs rapidly: PSS was highly contingent upon the modality order of the preceding trial. Furthermore, we found that rapid recalibration generalizes across different audiovisual speech stimuli: there was no main effect of syllable congruency nor was there an interaction between
syllable congruency and t − 1 modality order. These results indicate that a fast-acting recalibration mechanism serves to instantaneously realign signals at onset. This rapid realignment might maximize the perceptual benefits of audiovisual integration. For instance, generalization across different speech stimuli would be necessary in order for rapid adaptation to an initial asynchronous audiovisual event to produce optimized comprehension for the remainder of the stream.

3. Experiment 2

In Experiment 1, we showed for the first time, to our knowledge, that temporal recalibration generalizes across different speech stimuli. However, the actor was the same on all trials; thus not only was the higher-order factor of actor identity held constant, but lower-level properties of the visual and auditory stimuli also remained highly similar (i.e. lip movements on the same face were paired with syllables spoken in same voice). In Experiment 2, we examined whether rapid recalibration generalizes across different actors. Experiment 2 was identical in design to Experiment 1, with the exception that we used two different male actors (figure 1a, b) saying the same syllable. If rapid recalibration is not specific to a particular stimulus, then we expect transfer across actors: for any given trial, we expect the PSS to be contingent upon the modality order on the preceding trial, regardless of the actor.

(a) Material and methods

(i) Participants

Eight volunteers participated in the experiment (three females; mean age 26 years, ranging from 20 to 38 years). Six participants were naive as to the purpose of the experiment; the other two participants were the authors.

(ii) Stimuli and apparatus

Materials were the same as in Experiment 1, except that we used two different Australian male actors (David, figure 1a; and Sam, figure 1b) saying the syllable /ba:/.

Figure 3. Results of Experiment 2. (a) Proportion of ‘simultaneous’ responses as a function of SOA and modality order on the previous trial (t − 1), for trials in which the actor was different from that on trial t − 1. Data are collapsed across participants. Curves show the best-fitting Gaussian in a least-squares sense. (b) Proportion of ‘simultaneous’ responses as a function of SOA and modality order on the previous trial (t − 1), for trials in which the actor was the same as that on trial t − 1. (c) Mean PSS as a function of t versus t − 1 actor congruency, and t − 1 modality order. In all panels, error bars represent ± 1 s.e.m. (Online version in colour.)

(b) Results and discussion

The procedure was the same as for Experiment 1. Modality order (audition leads or vision leads), SOA (0, 50, 100, 200, 400 or 800 ms) and actor (David or Sam) were selected randomly on each trial. Participants performed two sessions, each containing four experimental blocks of 96 trials, leading to 768 trials in total. As for Experiment 1, the first trial of each block was excluded from analysis.

(iii) Design and procedure

We fit a Gaussian distribution to the proportion of ‘simultaneous’ responses as a function of SOA, separately for each participant, t − 1 modality order and actor congruency (same or different actor t versus t − 1). The results of Experiment 2 are depicted in figure 3. Figure 3a presents data, collapsed across participants, for trials on which the actors differed between t and t − 1. The proportion of ‘simultaneous’ responses is shown as a function of SOA and t − 1 modality order, along with Gaussian fits to the data. Figure 3b presents the corresponding data for trials on which the actor was the same for t and t − 1.

The mean PSS, bandwidth and amplitude across all conditions were 37 ms, 205 ms and 0.97, respectively. PSS was entered as the dependent variable in an ANOVA with subject factors. The ANOVA yielded a reliable main effect of t − 1 modality order, F(1, 7) = 5.93, p = 0.045, ηp² = 0.50. This indicates that the PSS was shorter when audition led vision on the previous trial (24 ms) than when vision led audition on the previous trial (50 ms). The two-way interaction between t − 1 modality order and t versus t − 1 actor congruency was not significant, F(1, 7) = 1.04, p = 0.342, ηp² = 0.13; nor was the main effect of t versus t − 1 actor congruency, F(1, 7) = 3.15, p = 0.119, ηp² = 0.31.

Experiment 2 thus replicated our earlier finding that rapid recalibration to asynchronous audiovisual speech stimuli occurs online, based on a single asynchronous audiovisual event. Furthermore, the results demonstrate that adaptation generalizes across different actors, suggesting that rapid recalibration is not specific to particular speech stimuli.
4. Experiment 3

Multisensory interactions depend on both basic temporal factors, such as temporal proximity between the auditory and visual information [7,8,34,35], and higher-order factors, such as whether or not a participant believes that the auditory and visual stimuli belong together [36–38]. The latter factor—the participant’s belief that a single audiovisual event, rather than two separate unisensory events, has occurred—has been dubbed the unity assumption [39]. In Experiments 1 and 2, the unity assumption was never violated as the auditory information was always paired with the matching visual information. That is, a specific voice always coincided with the lip movements of the actor to whom the voice belonged.

In Experiment 3, we manipulate the congruency between the auditory and visual events within single trials, in order to examine whether the unity assumption must be satisfied to produce rapid temporal recalibration (see also [36,40], for a similar methodology). Participants viewed either a male or female actor (figure 1b,c) in combination with either a male or female voice saying a single syllable. The actor producing the voice, and the actor producing the lip movements, were manipulated independently. Stimuli were presented such that half of the trials contained congruent audiovisual pairings (the male voice paired with the male actor’s lip movements, or the female voice paired with the female actor’s lip movements), and the remaining trials contained incongruent audiovisual pairings (the male voice paired with the female actor’s lip movements, or the female voice paired with the male actor’s lip movements). If rapid recalibration is largely due to temporal factors rather than higher-order factors—like the unity assumption—then we expect to observe rapid recalibration regardless of the congruency between auditory and visual stimuli.

(a) Material and methods

(i) Participants

Eight volunteers participated in the experiment (five females; mean age 29 years, ranging from 23 to 38 years). Six participants were naive as to the purpose of the experiment; the other two participants were the authors.

(ii) Stimuli and apparatus

Materials were the same as for Experiments 1 and 2, except that we used an Australian male actor (Sam; figure 1b) and female actor (Rosalind; figure 1c) saying the syllable /buk/. Furthermore, the voices and lip movements were independently manipulated such that on half of the trials, participants were presented with congruent pairings (the male voice paired with the male actor’s lip movements, or the female voice paired with the female actor’s lip movements), and on the remaining trials participants were presented with incongruent pairings (the male voice paired with the female actor’s lip movements, or the female voice paired with the male actor’s lip movements).

(iii) Design and procedure

The procedure was the same as for Experiments 1 and 2. Modality order (audition leads or vision leads), SOA (0, 50, 100, 200, 400 or 800 ms), voice actor (Sam or Rosalind) and lip movement actor (Sam or Rosalind) were selected randomly on each trial. Participants performed four sessions, each containing four experimental blocks of 96 trials, leading to 1536 trials in total. As for Experiments 1 and 2, the first trial of each block was excluded from analysis.

(b) Results and discussion

We fit a Gaussian distribution to the proportion of ‘simultaneous’ responses as a function of SOA, separately for each participant, $t-1$ modality order, audiovisual congruency on trial $t-1$ and audiovisual congruency on trial $t$. The results of Experiment 3 are shown in figure 4. Figure 4a shows PSS as a function of $t-1$ audiovisual congruency and $t-1$ modality order for incongruent trials; figure 4b shows the same results for congruent trials.

The mean PSS, bandwidth, and amplitude were 49 ms, 205 ms and 0.99, respectively. PSS was entered as the dependent...
variable in an ANOVA with t-1 modality order, t-1 audiovisu-
    al congruency and t audiovisual congruency as within-
    subject factors. The ANOVA yielded a significant effect of t-1 
    modality order, F (1, 7) = 18.80, p = 0.003, η² = 0.89, indicat-
    ing that the PSS was shorter when audition led vision on the prece-
    ding trial (41 ms) than when vision led audition (57 ms) on the 
    preceding trial. The main effect of audiovisual congruency on 
    trial t was not significant, F(1, 7) = 2.23, p = 0.179, η² = 0.24; 
    nor was the main effect of audiovisual congruency on trial 
    t-1, F(1, 7) = 0.02, p = 0.888, η² = 0.00. None of the two-way 
    interactions was significant, all F < 1.37, p > 0.280, η² < 0.16, 
    suggesting that temporal recalibration was independent 
    of the congruency on the preceding (t-1) and current (t) 
    trial. Similarly, the three-way interaction was not significant, 
    F(1, 7) = 2.17, p = 0.184, η² = 0.24.

In follow-up analyses, we examined whether PSS varied 
    as a function of the SOA on trial t-1 (see also [28]). We 
    expect the PSS on trial t to be smaller (i.e. vision must lead 
    by a smaller amount to achieve perceived simultaneity) 
    when audition is leading on trial t-1, compared with the 
    PSS on trial t when the SOA on trial t-1 is zero (baseline 
    conditions). By contrast, we expect the PSS on trial t to be larger 
    (i.e. vision must lead by a greater amount to achieve per-
    ceived simultaneity) when vision is leading on trial t-1, 
    compared with the PSS under baseline conditions. Figure 4c 
    illustrates how the mean PSS varies as a function of the 
    SOA on trial t-1. An ANOVA with PSS as the dependent 
    variable, and SOA on trial t-1 as the within-subject factor, 
    yielded a reliable effect, F(10, 70) = 4.8, p = 0.001, confirming 
    our expectation that PSS varies as a function of SOA on trial 
    t-1. In further exploratory analyses, we compared the PSS 
    for each t-1 SOA with the PSS under baseline conditions 
    in a series of one-tailed t-tests. PSS was significantly shorter 
    when audition led vision on trial t-1 by 100 ms, t(7) = 2.2, 
    p = 0.034, but not for shorter or longer SOAs (all p > 0.29). 
    By contrast, PSS was significantly larger when vision led on 
    trial t-1 by 200 ms, t(7) = 3.0, p = 0.01; 400 ms, t(7) = 2.1, 
    p = 0.036 and 800 ms, t(7) = 4.1, p = 0.002, but not for shorter 
    SOAs (all p > 0.09).

In Experiment 3, we again found evidence for rapid reca-
    libration to asynchronous audiovisual speech. We found that 
    recalibration occurred irrespective of audiovisual congruency 
    on both trial t and trial t-1. This suggests that the recalibra-
    tion is largely due to temporal factors, and not susceptible to 
    higher-order cognitive factors such as the unity assumption.

Results are consistent with those of Green et al. [40], who 
    found that the McGurk effect was equally strong for 
    incongruent audiovisual pairings (e.g. when a male voice 
    was paired with female lip movements) as for congruent 
    pairings. Experiment 3 also indicates that perceived simultane-
    ity on trial t-1 is not necessary to produce recalibration: strong 
    recalibration effects were evident even when vision led audi-
    tion by 800 ms, under which conditions participants hardly 
    ever perceived the auditory and visual events as simultaneous (1.9% of trials).

5. General discussion

The present study provides the first demonstration, to our 
    knowledge, that recalibration to asynchronous audiovisual 
    speech stimuli occurs rapidly, on the basis of a single audio-
    visual event. Although we did not use an explicit adaptation 
or top-up procedure, the effect sizes reported here are com-
    parable to—or even greater than—the effect sizes reported 
in studies using prolonged adaptation procedures [21,41–43].

Few previous studies have investigated whether temporal 
    recalibration to asynchronous audiovisual events, either induced 
    rapidly or using prolonged adaptation, generalizes across dif-
    ferent stimuli. Here, we demonstrate for the first time, to our 
    knowledge, that rapid recalibration generalizes across different 
    speech sounds and actors: we found compelling evidence that 
    the PSS on a given trial was contingent on the modality order 
    on the preceding trial, even when the sounds or actors differed 
    between the two successive trials. These findings are consistent 
with those of Navarra et al. [25], who showed that prolonged 
    adaptation to asynchronous audiovisual stimuli generalized 
    across different frequencies of an auditory tone. Likewise, 
    Heron et al. [23] found that recalibration generalized across 
    stimulus features (auditory and spatial frequencies), though it 
    did not generalize across spatial locations. The results are also 
    consistent with a recent study by Harvey et al. [32], who found 
that rapid recalibration generalizes across different colours and 
auditory frequencies.

Is the inter-trial effect due to a purely auditory or visual 
    recalibration, to a combination of independent auditory and 
    visual recalibrations or to a unified audiovisual recalibration?

Recently, Van der Burg et al. [44] investigated this by apply-
    ing a simultaneity-judgement task across all combinations of 
visual, auditory and tactile modalities. They observed a 
    strong inter-trial recalibration effect when participants per-
formed an audiovisual task (a beep in combination with a 
flash), but not when the same visual or auditory event was 
presented in combination with a tactile pulse in a visuotactile 
or audiotactile task. The authors concluded that rapid recalib-
    ration is uniquely audiovisual: if recalibration was driven by 
separate, purely visual or auditory components, then it 
    should also be observed in visuotactile and audiotactile tasks.

In simultaneity-judgement tasks, it is difficult to deter-
    mine whether a shift in the psychometric function reflects a 
genuine change in the experienced temporal relationship of 
the stimuli, or rather a change in the criteria used to judge 
whether or not a particular temporal relationship constitutes 
synchrony [45]. The absence of audiotactile and visuotactile 
inter-trial recalibration effects [44], however, supports the 
    notion that rapid recalibration is not due to a shift in criterion, 
or response bias. That is, if response bias underpinned recalib-
    ration, one should expect the same effect when participants 
perform an audiotactile or visuotactile task. Furthermore, 
    in the present study, recalibration was more consistent when 
vision led audition on the previous trial than when audition 
led vision (figure 4c; see also [28]). A response bias derived 
from the modality order of the previous trial should yield a 
symmetrical pattern of results. We thus contend that the reca-
    libration effects reported here are unlikely due to response bias.

Multisensory interactions depend on both lower-level factors, such as temporal proximity between events in different 
    modalities [7,8,34,35], and higher-level factors such as the unity assumption [36–38]. In this study, we find compelling 
evidence that rapid recalibration still occurs when the audiovisual speech stimuli obviously belong to different 
actors (i.e. the unity assumption is violated). Furthermore, 
we find that recalibration occurs even following cross-modal delays (800 ms) that almost never (1.9% of trials) produce an experience of simultaneity (see also [28]).
We note that the current study cannot rule out the existence of subtle congruency effects that modulate the exact magnitude of recalibration. Rather, it serves as a basic demonstration that rapid recalibration to asynchronous audiovisual speech stimuli is driven primarily by temporal factors, and occurs even when higher-order principles, such as the unity assumption, are violated.

Although our results are unambiguous, they stand in apparent opposition to results from at least one previous study of temporal recalibration. Roseboom & Arnold [21] found that observers could maintain concurrent recalibrations, which differed in the direction in which the PSS was shifted, for two different actors. Their study, however, used a conventional, prolonged adaptation procedure. This suggests that rapid and prolonged recalibration may be mediated by different mechanisms; or that rapid recalibration drives prolonged recalibration, but higher-order associations are possible only for the latter. In the above study, each actor was consistently paired with a spatial location during adaptation phases; but during isolated test trials, the PSS was shifted in a direction consistent with adaptation for the particular actor, regardless of presentation location. One possibility, then, is that the identity-specific recalibration they observed was developed over many instances of location-specific rapid recalibration, which transferred across spatial locations after it became associated with the higher-order factor of actor identity.

While the transfer we observed across different actors may not be adaptive in some cases, transfer of adaptation across different consecutive speech stimuli is a useful device for improving comprehension in verbal communication. Rapidly adapting to a cross-modal delay that is detected during the first syllable of an utterance is likely to enhance the intelligibility of the remainder of the speech stream [6]. The asymmetrical adaptation we observed in Experiment 3 (see figure 4c) is also highly consistent with a mechanism whose purpose is to realign auditory and visual signals originating from the same source [28]. The difference in perceptual latency between synchronized (at origin) auditory and visual signals arises from two primary factors: the different air propagation speeds of sound (approx. 340 m s⁻¹) and light (approx. 3.0 x 10⁸ m s⁻¹); and the slower (by approx. 30 ms) process of neural transduction in vision compared to audition [9]. In addition, with increasing distance from the source, the intensity of an auditory stimulus usually decreases more rapidly than the intensity of a visual stimulus. For short distances, over which the difference in speed of propagation and stimulus intensity has little effect, audition is perceived before vision; however, the difference in perceptual latency is never greater than what is caused by neural transduction. Beyond about 10 m, at which point the difference in propagation speed balances the difference in neural transduction time, vision is perceived before audition; in this case, the difference in perceptual latency can continue to increase with distance indefinitely. There is thus a marked asymmetry in what we might term ecologically plausible differences in audiovisual timing: technological manipulations notwithstanding, audition can lead vision by only fractions of a second, yet considerably longer delays are possible when vision leads audition. Accordingly, we observed no effect on the PSS for trial t when audition led by more than 100 ms on trial t – 1; but a significant effect on PSS for trial t when vision led by up to 800 ms (the largest asynchrony tested). Curiously, past studies using prolonged adaptation procedures have found clear effects of adaptation to stimuli in which audition led by 200–300 ms, for simple tone–flash pairs [18,19] as well as for more complex speech stimuli [21]. Perhaps, a small rapid recalibration effect—too small to be detected in the present experiment—occurs at these SOAs, and prolonged procedures allow detection of the cumulative effect over many trials.

We conclude that temporal recalibration to asynchronous audiovisual speech stimuli occurs rapidly. It appears to be induced only by stimuli with ecologically plausible differences in audiovisual timing, and may serve to rapidly realign auditory and visual signals to maximize speech comprehension. It generalizes across different stimuli and actors, and can be induced by auditory and visual events that are perceived to be asynchronous, suggesting that it is driven primarily by low-level temporal characteristics rather than higher-order cognitive factors.

Data accessibility. The full dataset is available online from the Open Science Framework (https://osf.io/yxkkz/).

Acknowledgements. We thank the authors (David, Sam and Rosalyn) for their contributions.

Funding statement. This research was supported by a grant from the Australian Research Council to E.V.d.B. (Discovery Early Career Research Award, DE130101663). P.T.G. was supported by a grant from the John Templeton Foundation to H. Price, A. Holcombe, K. Miller and D. Rickles.

References


