Multiple concurrent temporal recalibrations driven by audiovisual stimuli with apparent physical differences

Xiangyong Yuan · Cuihua Bi · Xiting Huang

Published online: 14 March 2015
© The Psychonomic Society, Inc. 2015

Abstract Out-of-synchrony experiences can easily recalibrate one's subjective simultaneity point in the direction of the experienced asynchrony. Although temporal adjustment of multiple audiovisual stimuli has been recently demonstrated to be spatially specific, perceptual grouping processes that organize separate audiovisual stimuli into distinctive “objects” may play a more important role in forming the basis for subsequent multiple temporal recalibrations. We investigated whether apparent physical differences between audiovisual pairs that make them distinct from each other can independently drive multiple concurrent temporal recalibrations regardless of spatial overlap. Experiment 1 verified that reducing the physical difference between two audiovisual pairs diminishes the multiple temporal recalibrations by exposing observers to two utterances with opposing temporal relationships spoken by one single speaker rather than two distinct speakers at the same location. Experiment 2 found that increasing the physical difference between two stimuli pairs can promote multiple temporal recalibrations by complicating their non-temporal dimensions (e.g., disks composed of two rather than one attribute and tones generated by multiplying two frequencies); however, these recalibration aftereffects were subtle. Experiment 3 further revealed that making the two audiovisual pairs differ in temporal structures (one transient and one gradual) was sufficient to drive concurrent temporal recalibration. These results confirm that the more audiovisual pairs

physically differ, especially in temporal profile, the more likely multiple temporal perception adjustments will be content-constrained regardless of spatial overlap. These results indicate that multiple temporal recalibrations are based secondarily on the outcome of perceptual grouping processes.

Keywords Temporal recalibration · Lag adaptation · Audiovisual · Perceptual grouping · Temporal structure

Perceptual experiences may reshape our subsequent feelings. Persistent exposure to asynchronous inter-sensory stimulus pairs can induce greater simultaneity judgments to exposed temporal offsets. The persistent asynchrony “pulls” an observer's point of subjective simultaneity (PSS) toward the displayed asynchrony (Fujisaki, Shimojo, Kashino, & Nishida, 2004; Stetson, Cui, Montague, & Eagleman, 2006; Vroomen, Keetels, de Gelder, & Bertelson, 2004). Over the past 10 years, evidence for this temporal recalibration phenomenon has been accumulating for almost every combination of sensory modalities (Di Luca, Machulla, & Ernst, 2009; Harrar & Harris, 2008), motor-sensory systems (Cai, Stetson, & Eagleman, 2012; Heron, Hanson, & Whitaker, 2009), and different types of stimuli, such as flashes and tones (Vroomen et al., 2004), stream and bounce scenarios (Fujisaki et al., 2004), and human speeches (Vatakis, Navarra, Soto-Faraco, & Spence, 2007), even for McGurk speeches (Asakawa, Tanaka, & Imai, 2012; Tanaka, Asakawa, & Imai, 2011; Yuan, Bi, Yin, Li, & Huang, 2014).

Although there is evidence that judgments of the simultaneity of occurrence of an inter-sensory pair can be recalibrated by recent experience, in real-world environments there is always more than a single pair of inter-sensory signals that may benefit from heterogeneous independent temporal recalibrations. For example, the temporal recalibration of

X. Yuan · X. Huang (✉)
Key Laboratory of Cognition and Personality, Ministry of Education
and Faculty of Psychology, Southwest University,
Chongqing 400715, China
e-mail: xthuang_swu@163.com

C. Bi
School of Educational Science, and Research Center of
Psychological Development and Education, Sichuan Normal
University, Chengdu, Sichuan, China

one audiovisual pair might be towards auditory leading asynchrony while at the same time the temporal recalibration of another audiovisual pair might be towards an opposing auditory lagging asynchrony. So far, however, research has yet to demonstrate whether concurrent temporal recalibration of multiple inter-sensory inputs is constrained by spatial specific processes or perceptual grouping processes (Heron, Roach, Hanson, McGraw, & Whitaker, 2012; Roseboom & Arnold, 2011; Roseboom, Kawabe, & Nishida, 2013; Sugano, Keetels, & Vroomen, 2014; Yuan, Li, Bi, Yin, & Huang, 2012).

Heron et al. (2012) reported that when two pairs of blobs and white noises with opposing temporal relationships were presented at left and right spatial locations, concurrent temporal recalibrations could be pulled toward their adapted temporal offsets respectively. In contrast, when features corresponding but spatially superimposed Gabors and sounds (e.g., high and low spatial frequency Gabors paired with high and low pitch sounds) were adapted with opposing temporal relationships, the temporal recalibration vanished because the opposing asynchronies negated each other during adaptation. Taken together, these findings suggest a spatial-specific temporal recalibration that is unlikely to arise from perceptual grouping through feature correspondence.

This spatial-specific temporal recalibration, however, weakened when auditory stimuli were presented through headphones rather than loudspeakers, which abolished auditory location cues (Yuan et al., 2012), implying that perceptual grouping through spatial correspondence may affect concurrent temporal recalibration. More directly, another study found that when a light-tone or tone-light sequence repeatedly alternated at left and right positions with a fixed 200-ms interval (e.g., left tone, left light, then right tone, right light etc.), spatial grouping of co-localized audiovisual stimuli led to unambiguous temporal recalibration (Yarrow, Roseboom, & Arnold, 2011). Thus, spatial correspondence may provide a perceptual metric by which audiovisual objects are constructed and subsequently form the basis for temporal recalibration.

The effect of perceptual grouping on multiple concurrent temporal recalibrations is more robust when audiovisual inputs are more complicated and ecologically valid. Roseboom and colleagues (2011, 2013) revealed that concurrent recalibrations of female and male utterances with opposite temporal relationships were contingent on the identities of the speaker regardless of whether their spatial locations overlapped. This content-constrained temporal recalibration is probably based on perceptual grouping through semantic correspondence, as previous research has demonstrated stronger multisensory integration using congruent human speech than incongruent speech (Vatakis & Spence, 2007, 2008).

The failure of Heron et al. (2012) to report a content-constrained temporal recalibration in their second experiment is probably due to the weak coupling of feature

correspondence, i.e., visual spatial frequency and auditory pitch. In order for multiple arbitrary stimuli like blobs/Gabors and tones to be temporally recalibrated separately (i.e., content constrained), the observers had to be told beforehand that these two audiovisual pairs had opposite asynchronies (Roseboom et al., 2013) so that each observer more easily categorized the audiovisual pairs as “different” rather than “same” before adaptation. This postulation that the perceptual grouping of separate audiovisual pairs into distinctive objects is crucial for multiple concurrent temporal recalibrations is echoed in recently published research (Ikumi & Soto-Faraco, 2014) that found PSS shift to be a function of whether the observer had focused attention to audio-then-flash or to flash-then-audio groupings during the adaptation phase. Such evidence suggests that selective attention provides a useful filter to help resolve which audiovisual stimuli are combined and adjusted in time when multiple pairs with opposing temporal relationships compete for recalibration.

If the perceptual grouping of separate audiovisual pairs into stable and firm objects is indeed required for multiple concurrent temporal recalibrations (which would be constrained by content), apparent physical differences between audiovisual pairs that make them more distinct from each other would be more likely to instigate concurrent temporal recalibration independent of informed categorization (Roseboom et al., 2013) and selective attention (Ikumi & Soto-Faraco, 2014). Here, we tested the hypothesis that apparent physical differences between spatially superimposed audiovisual pairs modulate concurrent temporal recalibration. In particular, we examined whether decreasing the differences between complicated audiovisual stimuli like human speech would diminish content-constrained recalibration (Experiment 1), and whether increasing the differences between simple audiovisual stimuli like flashes and tones would promote content-constrained recalibration (Experiments 2 and 3).

Experiment 1

Experiment 1 tested whether reducing the apparent physical differences between complicated audiovisual pairs would reduce the previously documented content-constrained temporal recalibration (Roseboom & Arnold, 2011; Roseboom et al., 2013). To this end, we manipulated the gender of the speakers. In the two-speaker condition, we replicated the previous finding that concurrent temporal recalibration was constrained by content when a female and a male speaker were located at the same position articulating /di/ and /po/. In the one-speaker condition, however, only one speaker was included (Fig. 1a and b). This meant that a single female or male speaker alternately articulated /di/ and /po/. This manipulation successfully reduced the physical differences between these two

Fig. 1 Procedures for and results of Experiment 1. Each video clip lasted 1,934 ms separated by a uniform random interval of 400–800 ms. The fixation point flashed as a cue of the test stimuli followed by a uniform random blank between 600 ms and 1,200 ms. The particular identity and SOA of test speeches were randomized. **a** In the one-speaker condition, all video clips were from one particular speaker who was alternately articulating /di/ and /po/. Only the female speaker is shown for illustration. **b** In the two-speaker condition, one female and one male speaker were alternately articulating /di/ and /po/. Panels **c** and **e** represent the recalibration aftereffects (i.e., the PSS shifts, calculated by

subtracting the PSS of the auditory-leading adaptation from the PSS of the visual-leading adaptation) for utterances /di/ (left bar) and /po/ (right bar) averaged across participants in the separate conditions, respectively. Error bars indicate standard errors of the mean. Panels **d** and **f** represent the fitted curves of the auditory-leading (red/left curve in each plot) and visual-leading adaptations (blue/right curve in each plot) for the utterances /di/ (left plot) and /po/ (right plot) in two- and one-speaker conditions, respectively, when simultaneity proportions were first averaged across participants. Negative values indicate that audition preceded vision

audiovisual pairs while at the same time retaining their identity differences (always /di/ and /po/).

Methods

Participants The first author and 11 naive participants completed the two-speaker condition (five males, mean age =

21.5 years). The first author and another 11 naive participants completed the one-speaker condition (six males, mean age = 20.9 years). All participants had normal or corrected-to-normal vision and hearing, provided their informed consent for participation, and were paid after the experiment.

Apparatus and stimuli Visual speeches were presented on an LCD monitor with a refresh rate of 60 Hz, and auditory speeches were presented through headphones (Sennheiser

HD 201) set to a comfortable volume. Participants sat approximately 80 cm from the monitor, and were required to keep their eyes on a fixation point (at the location of speaker's lips) and their heads still. The video clips (sample rate: 29.97 frames/s) of /di/ and /po/ uttered by a female and male speaker were recorded and then edited in Ulead Video Studio 10 (Corel Corporation), paired to temporally desynchronized visual and auditory tracks of each utterance with various stimulus onset asynchronies (SOAs). Only the facial areas of video clips were retained within an oval with height = $\sim 6.0^\circ$ and width = $\sim 2.9^\circ$, while the other areas were masked in dark gray (0.96 cd/m^2). E-Prime 2.0 was used to control Experiment 1.

Procedures In the two-speaker condition, we replicated the typical content-constrained temporal recalibration with a female speaker articulating /di/ and a male speaker articulating /po/, both located at the same position (the center of the screen). During the adaptation and top-up phases, these two speakers always had opposite asynchronies, for example, the visual speech of the female speaker led her auditory speech by 400 ms while the visual speech of the male speaker lagged his auditory speech by 133 ms, or vice versa. Participants had no explicit task during these two phases but were instructed to merely focus on the adapted speeches (the mouth area of the speaker). During the test phase, the speaker's visual and auditory speech were varied across 9 SOAs (−400 ms, −267 ms, −133 ms, 0 ms, 133 ms, 267 ms, 400 ms, 533 ms, and 667 ms, with negative SOAs indicating audition leads). Participants were required to judge whether each visual and auditory speech were simultaneous (SJ task). An asymmetric range of adapted temporal lags and SOAs was set because of the fact that our participants were more likely to perceive synchrony when visual speech led auditory speech, which has also been observed in other studies (Maier, Di Luca, & Noppeney, 2011; Roseboom et al., 2013; van Wassenhove, Grant, & Poeppel, 2007). The asymmetric adapted temporal lags (133 ms vs. 400 ms) were each located at about the same temporal distance away from the supposed PSS rather than from the physical synchrony (0 ms) to assure comparable recalibration aftereffects.

Each block consisted of 72 trials, with a short break after 36 trials. There were four top-up speeches and one test speech in each trial except for the first and 37th trials, which respectively comprised 60 and 30 adaptation speeches (as the adaptation phase) and one test speech (Fig. 1b). The two speakers were alternately presented with a uniform random interval between 400 ms and 800 ms for the adaptation and top-up speeches. The gender and SOA of the speaker in the test speech were randomly selected in each trial. In one block, there were four repetitions for each SOA of each speaker. Participants had to complete two different blocks (i.e., audition led vision for the female speaker while audition lagged vision for the male speaker during adaptation; vision led audition for the female

speaker while vision lagged audition for the male speaker during adaptation) each day, with a total of eight blocks finished across four separate days. The order of block presentation was counterbalanced across days for each participant.

The procedures of the one-speaker condition were almost identical to those of the two-speaker condition (an example illustrated in Fig. 1a), except that only one speaker alternately articulated /di/ and /po/ in the center of the screen. Half the participants observed the female speaker while the other half observed the male. Other aspects of the experiment remained the same as in the two-speaker condition. The two blocks consisted of (a) audition leading vision for the /di/ utterance while audition lagged vision for the /po/ utterance during adaptation, and (b) audition lagging vision for the /di/ utterance while audition led vision for the /po/ utterance during adaptation.

Results and discussion

The one- and two-speaker conditions both included two utterances (/di/ and /po/) which had more physical attributes in common in the former relative to the latter condition. We expected that in the one-speaker condition, the content-contingent aftereffects would vanish while in the two-speaker condition they would persist. As the two utterances had opposing asynchronies during the adaptation and top-up phases, there were a total of four possible combinations of utterance and asynchrony. Each participant's response assessments of the proportion of stimuli that were simultaneous were first fit (using a MatLab fit function, nonlinear least-squares method) as a function of SOAs using a truncated Gaussian function to extract his/her PSS in each combination (Fujisaki et al., 2004; Roseboom & Arnold, 2011): $\text{proportion simultaneous} = \min \{1, \text{maximum_proportion} \times \exp[-(\text{SOA} - \text{PSS})^2 / (2 \times \text{sigma}^2)]\}$. PSS's were then analyzed in a repeated-measures ANOVA with utterances (/di/ and /po/) and temporal lag (auditory speech leading and lagging) included as independent variables for each condition.

In the two-speaker condition, there were significant main effects of temporal lag, $F(1, 11) = 7.917$, $p = .017$, $\eta_p^2 = .419$, and utterance, $F(1, 11) = 26.470$, $p < .001$, $\eta_p^2 = .706$, while the interaction was not significant, $F(1, 11) = 0.853$, $p = .375$, $\eta_p^2 = .072$. As expected based on previous studies (Roseboom & Arnold, 2011; Roseboom et al., 2013), there were significant PSS shifts due to adaptation to auditory leading (132.50 ms, SD = 60.75 ms) and auditory lagging (161.36 ms, SD = 71.68 ms) irrespective of utterances, as shown in Figs. 1c and d. More importantly, in the one-speaker condition, there was no significant effect of temporal lag, $F(1, 11) = 0.193$, $p = .669$, $\eta_p^2 = .017$, utterance, $F(1, 11) = 0.168$, $p = .690$, $\eta_p^2 = .016$, or the interaction, $F(1, 11) = 2.391$, $p = .150$, $\eta_p^2 = .179$. In other words, PSS did not significantly

shift from the auditory leading (144.99 ms, SD = 58.59 ms) to the auditory lagging adaptation (147.20 ms, SD = 45.75 ms) irrespective of utterances, as shown in Figs. 1e and f. This difference between the one- and two-speaker conditions is probably caused by the apparent physical differences between the two adapted audiovisual utterances.

To further illustrate this effect, the PSS shifts of these two conditions were compared. We conducted a repeated-measures ANOVA with utterance (/di/ and /po/) as a within-subject variable and condition (two- and one-speaker) as a between-subject variable. This analysis indicated a significant main effect of condition, $F(1, 22) = 5.447$, $p = .029$, $\eta_p^2 = .198$, and a nonsignificant main effect of utterance, $F(1, 22) = 0.098$, $p = .757$, $\eta_p^2 = .004$, and the interaction, $F(1, 22) = 1.998$, $p = .171$, $\eta_p^2 = .083$. These results confirm that the PSS shifts in the one-speaker condition were significantly smaller than those in the two-speaker condition, implying that the multiple concurrent temporal recalibrations were indeed affected by the physical differences in the two conditions.

The sigma value of the fitted functions was taken as a measure of the width of the temporal window within which participants judged the audiovisual simultaneity more often. Significant and marginally significant effects were found only for the main effect of utterances and its interaction with temporal lag (all $ps < .054$). The sigma values for the /di/ utterance (mean = 257.50 ms and 255.51 ms for the two- and one-speaker condition, respectively) was smaller than that for the /po/ utterance (mean = 320.01 ms and 311.08 ms, respectively). The interaction between temporal lag and utterance was not caused by a separate effect of temporal lag on different utterances as all comparisons between auditory-leading (A-leading) and visual-leading (V-leading) adaptations for each utterance were not significant (multiple comparisons using the least significant difference method, $ps > .08$). Thus, sigma values may not have been a sensitive enough indicator of concurrent temporal recalibration at least in Experiment 1 (detailed statistical analyses can be found in the Appendix).

Experiment 1 suggests that the apparent physical differences between audiovisual pairs can modulate concurrent temporal recalibrations. These results also explain why it is so natural for multiple complicated stimuli pairs to be temporally recalibrated in a content-contingent way (Roseboom & Arnold, 2011; Roseboom et al., 2013). If our hypothesis is true, however, merely increasing physical differences should instigate content-constrained recalibration as well. We noticed that the arbitrarily simple audiovisual pairs used in previous studies typically differ in a single one-stimulus attribute and do not consistently show content-constrained temporal recalibration. Their visual components have different spatial frequencies (Heron et al., 2012), orientations (Roseboom et al., 2013, supplemental Experiment 1), or colors (Yuan et al., 2012, Experiment 1a), while their auditory components always have different frequencies (from 300–3,500 Hz). It is interesting to speculate whether endowing these simple

audiovisual pairs with differences in more attributes/features would promote further content-constrained temporal recalibration.

Experiment 2

Experiment 2 tested whether increasing the physical differences between simple audiovisual pairs by complicating the stimulus would promote content-constrained temporal recalibration. The experimental design still included a concurrent exposure to opposing asynchronies at the same location. In the orientation/color conditions, visual stimuli differed in orientation (horizontal or vertical) and also in color (red or green), while the accompanying auditory stimuli were complicated tones generated by multiplying two pure tones of different frequencies (either a 500 Hz or 2,000 Hz tone each multiplied by a 1,250 Hz tone which was set as a moderate baseline frequency). In two other contrasting conditions, the visual stimuli differed in either orientation or color and the auditory stimuli differed in frequency (500 Hz or 2,000 Hz pure tone).

Methods

Participants The first author and 11 naive participants were recruited for the orientation/color condition (five males, mean age = 21.6 years). The first author and another nine naive participants completed the orientation condition (four males, mean age = 21.8 years) and ten participants also including the first author took part in the color condition (five males, mean age = 21.6 years). All participants had normal or corrected-to-normal vision and hearing, provided informed consent, and were paid after the experiment.

Apparatus and stimuli Visual stimuli were presented on a CRT monitor with a refresh rate of 85 Hz, and, as before, auditory stimuli were presented through headphones (Sennheiser HD 201) set to a comfortable volume. Participants sat approximately 60 cm from the monitor and were required to keep their eyes on a fixation point (a black spot on the screen) with their heads still. All audiovisual stimuli lasting 35 ms were presented on a gray background (13.4 cd/m², $x = .272$, $y = .314$). In the orientation condition, two disks with vertical and horizontal gratings (diameter: 3.1°, spatial frequency of gratings: 0.40°, average luminance: vertical, 21.0 cd/m², $x = .275$, $y = .311$; horizontal, 21.9 cd/m², $x = .272$, $y = .309$) were each paired with two tones of low and high pitches respectively (500 Hz and 2,000 Hz, with 5-ms ramp-on and -off). In the color condition, green (12.5 cd/m², $x = .276$, $y = .538$) and red disks (12.0 cd/m², $x = .587$, $y = .330$) of the same size were each paired with these same low- and high-pitched tones, respectively. Greater changes to the audiovisual stimuli were made in the orientation/

color condition, wherein the audiovisual stimuli included (a) a disk with green and vertical gratings accompanied by a complicated tone generated by multiplying a frequency of 500 Hz by 1,250 Hz and (b) another disk with red and horizontal gratings paired with another tone generated by multiplying a frequency of 2,000 Hz by 1,250 Hz (see Fig. 2). MatLab version 2009b (The MathWorks) equipped with PsychToolbox 3.0 was used for stimulus and protocol control. Temporally precise presentation of auditory stimuli was achieved by ASIO drivers in PsychToolbox that ensure a low-latency presentation of auditory stimuli.

Procedures All conditions had similar procedures in which all disks and tones were presented at the center of the screen. During the adaptation and top-up phases, the two disk-tone pairs always had opposing asynchronies. For example, the vertical grating/green disk led the lower tone by 118 ms while the horizontal grating/red disk lagged the higher tone by 118 ms, and vice versa. Participants had no explicit task during these two phases but were instructed to merely focus on the stimuli pairs. During the test phase, the test disk and tone were separated by one of the 7 SOAs (± 247 ms, ± 165 ms, ± 82 ms, 0 ms, with negative SOAs indicating audition leads). Participants were required to judge whether the presented disk and tone occurred simultaneously.

Each block consisted of 112 trials, with a short break after 56 trials. There were six top-up stimuli pairings and one test stimuli pair in each trial except that the first and 57th trials comprised 60 and 30 adaptations, respectively (as adaptation phases), and one test (Fig. 2). The two stimuli pairs were alternately presented with a uniform random interval of 706–1,176 ms during the adaptation and top-up phases. For each trial, the stimuli pair presented and their SOA during the test phase were chosen randomly. In one block, there were eight repetitions for each SOA of each stimuli pair. Participants had to complete two different blocks each day (i.e., the lower tone leading the vertical grating/green disk while the higher tone lagged the horizontal grating/red disk during adaptation, and the lower tone lagging the vertical grating/green disk while the higher tone led the horizontal grating/red disk during adaptation), with a total of four blocks finished across 2 separate days. The test order of the adaptation blocks was counterbalanced across days within participants.

Results and discussion

According to our hypothesis, if apparent physical differences modulate content-constrained temporal recalibrations, more obvious recalibration aftereffects should have been observed

Fig. 2 Stimuli and protocol for Experiment 2. Panels a, b, and c demonstrate example stimuli and procedures for the orientation, color, and orientation/color conditions, respectively. The fixation point flashed to cue the test stimuli followed by a random blank of 706–1,176 ms in duration. The stimuli pairing and SOA used in the test phase were randomly selected in each trial

in the orientation/color condition relative to the other two conditions. Each audiovisual pair had two possible asynchronies during the adaptation and top-up phases, so that four combinations were generated regardless of condition. Each participant's assessments of proportion of stimuli pairings that occurred simultaneously were again fit to truncated Gaussian functions to extract his/her PSS for each combination. PSS's were then analyzed in a repeated-measures ANOVA with stimuli (vertical grating/green disk and lower tone vs. horizontal grating/red disk and higher tone) and temporal lag (tone leading vs. lagging) as independent variables. The results for the orientation and color conditions were quite similar, with both revealing no significant main effect for temporal lag: the orientation condition, $F(1, 9) = 0.114, p = .743, \eta_p^2 = .013$; the color condition, $F(1, 9) = 0.293, p = .601, \eta_p^2 = .032$. All other effects did not reach significance ($ps > .5$), except a marginally significant main effect of stimuli in the orientation condition, $F(1, 9) = 4.492, p = .063, \eta_p^2 = .333$. The averaged PSS shifts in the two conditions were 1.61 ms and 2.25 ms, respectively (Fig. 3a and c), therefore replicating the established conclusion that without explicit instructions and selective attention it is unable to observe content-constrained temporal

recalibration at the same spatial location (Heron et al., 2012; Roseboom et al., 2013).

A different pattern of results occurred in the orientation/color condition, revealing a significant main effect of temporal lag, $F(1, 11) = 5.891, p = .034, \eta_p^2 = .349$, and a significant main effect of stimuli, $F(1, 11) = 7.189, p = .021, \eta_p^2 = .395$, while the interaction was not significant, $F(1, 11) = 0.869, p = .371$. The PSS shifts averaged across audiovisual pairs were demonstrated as the difference between the tone leading (-12.34 ms, $SD = 36.28$ ms) and lagging adaptation (-6.88 ms, $SD = 38.10$ ms). However, as seen in Fig. 3e, despite no significant interaction, the PSS shift mainly occurs for the disk with green and vertical grating paired with the low-pitch complicated tone, with a mean PSS shift of 9.82 ms ($p = .033$). This asymmetric PSS shift result is puzzling, but perhaps due to the particular properties of the stimuli used in the present experiment. It is premature to give a firm explanation.

To further assess the aftereffects, PSS shifts of the three conditions were compared similarly to Experiment 1. We conducted a repeated-measures ANOVA with stimuli (vertical grating/green disk and lower tone vs. horizontal

Fig. 3 Results for Experiment 2. Panels a, c, and e demonstrate the recalibration aftereffects, i.e., the PSS of the A-leading adaptation subtracted from the PSS of the V-leading adaptation averaged across participants for each audiovisual pair and condition. Error bars represent standard errors of the mean. Panels b, d, and f illustrate fitted curves in the A-leading (red/left curve in each plot) and V-leading (blue/right curve in each plot) adaptations as a function of SOAs for

each audiovisual pair in the orientation, color, and orientation/color conditions, respectively, when simultaneity proportions were first averaged across participants. The left plots illustrate the vertical grating, green disk, and their combination accompanied by the lower tone, and the right plots illustrate the horizontal grating, red disk, and the combination accompanied by the higher tone, respectively. Negative values indicate that the auditory stimulus precedes the visual stimulus

grating/red disk and higher tone) as a within-subject variable and condition (orientation, color, or combination) as a between-subject variable. This analysis did not reveal any significant effect, $ps > .5$. Thus, there is little evidence that increasing physical differences by adding stimulus attributes strongly altered concurrent temporal recalibration, although it had subtle effect.

The temporal window indicated by the sigma values of the fitted functions were again analyzed by repeated-measures ANOVAs with stimuli (vertical grating/green disk and lower tone vs. horizontal grating/red disk and higher tone) and temporal lag (tone leading vs. lagging) as independent variables. All effects were nonsignificant ($ps > .09$) except for the main effect of temporal lag in the color condition, $F(1, 9) = 7.384$, $p = .024$, $\eta_p^2 = .451$, which indicated that the sigma value for the A-leading adaptation (125.77 ms) was significantly smaller than that for the V-leading adaptation (140.74 ms) regardless of colors (more details can be found in the [Appendix](#)). Given that larger temporal window without an accompanying PSS shift is more common when comparing an asynchrony against synchrony adaptation (Navarra, Soto-Faraco, & Spence, 2007; Navarra et al., 2005), this result is probably not as persuasive and informative as the PSS when comparing A-leading and V-leading adaptations.

From the PSS results, it seems that increasing the physical differences between stimuli pairs has subtle rather than dramatic effects on concurrent temporal recalibration. This outcome is possibly caused by the conservative instructions adopted here that did not require participants to attend to the temporal dimension, which has been shown to enlarge the recalibration aftereffects (Heron, Roach, Whitaker, & Hanson, 2010). We also used only a moderate manipulation of stimuli attributes, taking into consideration only two attributes. In essence, however, if we scrutinize the contrast between complicated and simple audiovisual stimuli used in Experiment 1 and 2, we find that complicated stimuli usually have distinctive temporal structures as well. For instance, the utterances /di/ and /po/ by two speakers not only include different identities but also evolve in different ways along the timeline (depicted by the temporal profile of each utterance in Fig. 1). Given that audiovisual components with greater temporal correlation are more likely to be bound together (Parise, Harrar, Ernst, & Spence, 2013; Parise, Spence, & Ernst, 2012), those simple audiovisual stimuli commonly used in previous studies that always have identical temporal structures (abrupt signals) may not be distinguished as different objects when observers' judgments focus on their temporal dimension. Therefore, we propose that altering the temporal structures of these simple audiovisual pairs with all other non-temporal features unchanged would probably have more power to increase their physical differences and promote content-constrained recalibration.

Experiment 3

Experiment 3 also aimed to verify the hypothesis that increasing physical differences between stimuli pairs would promote concurrent temporal recalibration, but this time we manipulated the temporal structure of the audiovisual stimuli. One audiovisual pair was the commonly used abrupt white blob paired with a pure tone, while the amplitude of the other blob and tone were modulated by a triangular wave. This manipulation increased the apparent physical differences while retaining all other non-temporal features. We expected a more robust content-constrained temporal recalibration.

Methods

Participants Fourteen naive participants were included (seven males, mean age = 21.3 years), all with normal or corrected-to-normal vision and hearing. They all provided informed consent and were paid after the experiment.

Apparatus and stimuli The apparatus and control software were identical to those in Experiment 2. The visual stimulus was a Gaussian blob ($\sigma = 0.91^\circ$) with a maximal luminance of 54.5 cd/m^2 . The auditory stimulus was a pure tone of 2,000 Hz. The simple audiovisual event lasted only 35 ms, as before, while the other with luminance and loudness unfolding following a triangular temporal profile lasted 165 ms. The temporal profile of the audiovisual stimulus is depicted in Fig. 4b.

Procedures The protocol was almost the same as Experiment 2 with a few adjustments as outlined here: During the adaptation and top-up phases, the two waveforms always had opposing asynchrony, for example, the triangular blob led the triangular tone by 165 ms while the abrupt blob lagged the abrupt tone by 165 ms, or vice versa. During the test phase, the range of the SOAs now included $\pm 282 \text{ ms}$, $\pm 188 \text{ ms}$, $\pm 94 \text{ ms}$, and 0 ms , with negative values indicating audition leads. Each block still consisted of 112 trials with a short break after 56 trials, but the first and 57th trials comprised 40 and 20 adaptations, respectively, and one test stimuli pair, while in other trials there were six top-up stimuli pairs and one test. The triangular and abrupt stimuli were alternately presented with a uniform random interval between 941 and 1,412 ms during the adaptation and top-up phases. The stimuli presented and SOA during the test phase were randomly selected in each trial.

Results and discussion

As in previous experiments, two possible asynchronies for each pair generated four possible combinations. The same

Fig. 4 Stimuli and results for Experiment 3. Panel **a** illustrates a white Gaussian blob following a triangular temporal profile. Panel **b** demonstrates the luminance-modulated blobs (upper rows) and amplitude-modulated tones (bottom rows) varying as a function of time. Panel **c** shows the recalibration aftereffects calculated by subtracting the PSS of the A-leading adaptations from the V-leading adaptations averaged across participants for each stimuli pair. Error bars represent

standard errors of the mean. Panel **d** indicates the average proportion of stimuli assessed as occurring simultaneously and its corresponding fitted curves as a function of SOAs in the A-leading (red/left curves in each plot) and V-leading (blue/right curves in each plot) adaptations for each stimuli pair (left plot: the triangular; right plot: the abrupt). Negative values indicate that auditory stimuli preceded visual stimuli

analyses were conducted as described for Experiments 1 and 2. PSS's for all participants were analyzed in a repeated-measures ANOVA with stimuli (triangular vs. abrupt) and temporal lag (audition leading vs. lagging) included as independent variables. Results indicated a significant main effects of temporal lag, $F(1, 13) = 8.863$, $p = .011$, $\eta_p^2 = .405$, and stimuli, $F(1, 13) = 15.738$, $p = .002$, $\eta_p^2 = .548$, but their interaction did not reach significance, $F(1, 13) = 0.015$, $p = .903$. As illustrated in Fig. 4c and d, the average PSS shifted from the audition leading (6.08 ms, SD = 39.33 ms) to lagging adaptations (16.95 ms, SD = 43.76 ms), irrespective of stimuli.

The repeated-measures ANOVA conducted on the temporal window revealed that all effects were significant or

approaching significance, $ps < .063$. Since the interaction term was significant ($p = .021$), we compared the sigma values for the A-leading and V-leading adaptations on each stimuli level separately. Results indicated that only for the abrupt stimuli did the sigma increase from the A-leading (133.57 ms) to the V-leading adaptation (150.03 ms; multiple comparison with least significant difference method, $p = .001$). For the triangular-wave stimuli, there was no significant difference (172.14 ms vs. 167.45 ms, $p = .441$). Given that the A-leading and V-leading adaptations were not compared against a synchrony adaptation or a pre-adaptation, the observed temporal window differences between them might not be as informative about temporal recalibration as PSS shifts.

Even when non-temporal features were identical, different temporal structures successfully induced large enough discrepancies between these two audiovisual pairs that correspondent perceptual grouping was facilitated and led to content-constrained temporal recalibrations. Considering the results of Experiments 2 and 3, it seems that a more efficient way to increase the apparent physical difference between two audiovisual pairs is to alter their temporal structures rather than complicating their non-temporal dimensions. However, we still have no evidence to completely exclude the possibility that increasing physical differences by introducing more attributes would constrain temporal recalibration to the same extent as altering temporal structure.

General discussion

The present study investigated the effect of one crucial factor, i.e., the apparent physical difference between spatially superimposed audiovisual pairs, on multiple temporal recalibrations. Experiment 1 found that decreasing the physical difference between two utterances by using a single speaker rather than two different speakers (male and female) of two utterances with opposing asynchronies at the same location did not promote multiple temporal recalibrations. Experiments 2 and 3 revealed that increasing the physical differences between arbitrary simple audiovisual pairs by complicating their non-temporal dimensions (Experiment 2) or altering their temporal structures (Experiment 3) can promote multiple temporal recalibrations after concurrent adaptation to opposing asynchronies. However, the effect of complicating stimuli in non-temporal dimensions (Experiment 2) was weak and more subtle.

It has already been shown in other experimental paradigms that apparent physical differences between stimuli can affect perceptual timing. First, if sparse distinctive audiovisual stimuli (red flashes and high-pitch pips) were embedded in dense pulse trains made of other features (white flashes and low-pitch pips), synchrony perception was significantly facilitated even when the total stimulus density was too high to reliably judge the audiovisual synchrony (Fujisaki & Nishida, 2007, Experiment 8). Second, when observers were required to judge the duration of an “oddball” stimulus after several stimuli repetitions, increasing the degree of apparent physical difference between the repetitive and oddball stimuli dilated the perceived duration of oddballs to a greater extent (Schindel, Rowlands, & Arnold, 2011, Experiment 3; Pariyadath & Eagleman, 2012, Experiment 2).

The present finding that the apparent physical difference between audiovisual pairs plays a determining role in the multiple possible concurrent temporal adjustments may be attributed to its promotion of the perceptual formation of distinct audiovisual “objects.” Perceptual grouping is found to impair

temporal order judgment of two spatially separated stimuli (Nicol & Shore, 2007), or to reorganize the perceived order of a sound with periodic interruptions filled with noise (Koenderink, Richards, & van Doorn, 2012). We agree with this, proposing that the brain prioritizes perceptual grouping over correct timing, implying that perceived timing may be a secondary effect of grouping and integration (Holcombe, 2013; Koenderink et al., 2012; Nicol & Shore, 2007; Roseboom, Nishida, Fujisaki, & Arnold, 2011). From this view, it is probable that only when separate audiovisual stimuli are perceptually grouped into independent objects are they subject to multiple temporal relationships in a content-constrained way. In a similar fashion, ambiguous groupings may not be followed by obviously multiple temporal recalibrations.

The principles of perceptual grouping as depicted by the “unity assumption” propose that two or more sensory inputs will be more likely to be treated as indicating the same multi-sensory event rather than separate unimodal events if they are highly consistent in one or more dimensions such as space, time, or semantic content (Vatakis & Spence, 2007; Vatakis & Spence, 2008). Previous research has revealed that perceptual grouping through spatial and semantic correspondence can drive unambiguous temporal adjustments after adaptation when there are multiple inputs (Roseboom & Arnold, 2011; Roseboom et al., 2013; Yarrow et al., 2011). Here, the present study adds to this evidence that when there are no spatial cues, physical differences, especially distinctive temporal structures, can be as important as subjective categorization and selective attention to define different audiovisual objects (Ikumi & Soto-Faraco, 2014; Roseboom et al., 2013).

Regarding the prediction that perceptual grouping precedes inferences regarding timing, the failure of Heron et al.’s (2012) Experiment 2 (and a similar result in our orientation condition in Experiment 2) to report significant content-constrained temporal recalibration may not represent a spatially specific mechanism, but simply be due to poor object-forming grouping processes when there are (a) no top-down mechanisms such as subjective categorization and selective attention and (b) no stimulus-driven processes such as apparent physical differences to facilitate the forming of separate “objects”. Although synesthetic congruent audiovisual stimuli, e.g., loud or high-pitch tones corresponding to brighter, lighter, smaller, or high-spatial-frequency visual stimuli, are more likely to be bound together (Spence, 2011), this perceptual grouping may not be as strong as congruent human speeches (Roseboom & Arnold, 2011; Roseboom et al., 2013; the two-speaker condition in our Experiment 1) and temporal correlation (our Experiment 3) during multiple concurrent temporal recalibrations.

Finally, it is worth discussing why simply making two audiovisual pairs differ in temporal structure is such an effective means to promote multiple temporal recalibrations. First, high

temporal correlations between audiovisual inputs can lead to greater multisensory integration (Parise, Harrar, Ernst, & Spence, 2013; Parise, Spence, & Ernst, 2012). Second, the multisensory integration of human speech is not only attributed to semantic congruency but also to the temporal correlation of lip movement and pronunciation. Different temporal structures may contribute to the robust temporal recalibrations and as suggested by our Experiment 1, decreasing these differences can weaken the recalibration aftereffects. Third, temporal correlations may evoke better coupling for arbitrary stimuli. In contrast, the non-temporal auditory feature will be sometimes bound with an irrelevant visual feature even when they are not presented together. This inference is partially evidenced by the self-reports of our participants after Experiment 2, a small portion of whom incorrectly reported that, for instance, the green disk was accompanied by a higher tone.

In summary, the current study reveals that apparent physical differences between audiovisual stimuli pairs can independently drive multiple concurrent temporal recalibrations regardless of spatial overlap. When these physical differences were decreased by using a single speaker instead of two distinctive speakers, there was no concurrent recalibration. When these physical differences are increased by presenting asynchronous stimulus pairs that combine two stimulus attributes/features instead of only one, or which were generated by different temporal profiles, this appears to promote the concurrent temporal recalibration. These findings imply that multiple concurrent temporal recalibrations are preceded by perceptual grouping processes that construct distinctive audiovisual objects.

Acknowledgments This research was funded by a grant from the Mechanism and Application of Temporal Range/Synthetic Model (TR201201-1) and a key fund project of Sichuan Provincial Department of Education (15SA0003).

Appendix: The analysis of sigma values in our three experiments

Experiment 1

The sigma values were submitted into a repeated measures ANOVA with utterances (/di/ and /po/) and temporal lag (auditory speech leading and lagging) as independent variables for each condition. The pattern looked similar. The results of the two-speaker condition showed that the main effect of temporal lag was not significant, $F(1, 11) = .142$, $p = .713$, while the main effect of the utterance was significant, $F(1, 11) = 40.170$, $p < .001$, $\eta_p^2 = .785$, and the interaction was marginally significant, $F(1, 11) = 4.718$, $p = .053$, $\eta_p^2 = .300$. The results of the one-speaker condition showed that the main effect of temporal lag was not significant, $F(1, 11) = .012$, $p = .915$,

while the main effect of the utterance, $F(1, 11) = 7.165$, $p = .022$, $\eta_p^2 = .394$, and the interaction was significant, $F(1, 11) = 5.538$, $p = .038$, $\eta_p^2 = .335$.

The main effect of utterance indicated that the sigma value for the /di/ utterance (mean = 257.50 ms and 255.51 ms for the two- and one-speaker conditions, respectively) was much smaller than that for the /po/ utterance (mean = 320.01 ms and 311.08 ms for the two- and one-speaker conditions, respectively). The significant interaction was largely due to the differences of the two utterances, as all comparisons between A-leading and V-leading adaptation for each utterance were not significant (multiple comparison with the least significant difference method, $ps > .08$).

Experiment 2

The sigma values were submitted to a repeated measures ANOVA with stimuli (vertical grating/green disc and lower tone vs. horizontal grating/red disc and higher tone) and temporal lag (tone leading and lagging) as independent variables. The results of the color condition revealed that the main effect of temporal lag was significant, $F(1, 9) = 7.384$, $p = .024$, $\eta_p^2 = .451$, with all other effects not significant, $F_s < 1$, $ps > .4$. The sigma value for the A-leading adaptation (125.77 ms) was significantly smaller than that for the V-leading adaptation (140.74 ms). The results of the orientation condition never revealed any significant effect, $F_s < 3$, $ps > .15$. The results of the orientation/color condition did not reveal any significant effect either, $F_s < 3.5$, $ps > .09$. As can clearly be seen, all effects were not significant except the main effect of temporal lag in the color condition.

Experiment 3

The sigma values were again submitted to a repeated measures ANOVA with stimuli (triangular and abrupt) and temporal lag (audition leading and lagging) as independent variables. The results revealed that the main effect of temporal lag was marginally significant, $F(1, 13) = 4.161$, $p = .062$, $\eta_p^2 = .242$, the main effect of stimuli was significant, $F(1, 13) = 9.876$, $p = .008$, $\eta_p^2 = .432$, and the interaction was also significant, $F(1, 13) = .6845$, $p = .021$. The sigma value for the triangular-wave stimuli (169.80 ms) was significantly larger than that for the abrupt stimuli (141.80 ms). Since the interaction was significant, we compared the sigma values for the A-leading and V-leading adaptation on each stimuli level; the results showed that only for the abrupt stimuli was the sigma value increased from the A-leading (133.57 ms) to the V-leading adaptation (150.03 ms, multiple comparison with the least significant difference method, $p = .001$) whereas for the triangular-wave stimuli, there were no significant differences for the two adaptations (172.14 ms vs. 167.45 ms, $p = .441$).

References

- Asakawa, K., Tanaka, A., & Imai, H. (2012). Audiovisual temporal recalibration for speech in synchrony perception and speech identification. *Kansei Engineering International Journal*, 11(1), 35–40.
- Cai, M., Stetson, C., & Eagleman, D. M. (2012). A neural model for temporal order judgments and their active recalibration: A common mechanism for space and time? *Frontiers in Psychology*, 3, 470. doi:10.3389/fpsyg.2012.00470
- Di Luca, M., Machulla, T., & Ernst, M. (2009). Recalibration of multisensory simultaneity: Cross-modal transfer coincides with a change in perceptual latency. *Journal of Vision*, 9(12), 1–16.
- Fujisaki, W., & Nishida, S. (2007). Feature-based processing of audiovisual synchrony perception revealed by random pulse trains. *Vision Research*, 47(8), 1075–1093.
- Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience*, 7(7), 773–778.
- 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(4), 517–524.
- Heron, J., Hanson, J. V. M., & Whitaker, D. (2009). Effect before cause: Supramodal recalibration of sensorimotor timing. *PLoS One*, 4(11), e7681.
- Heron, J., Roach, N. W., Whitaker, D., & Hanson, J. V. M. (2010). Attention regulates the plasticity of multisensory timing. *European Journal of Neuroscience*, 31(10), 1755–1762.
- Heron, J., Roach, N., Hanson, J., McGraw, P., & Whitaker, D. (2012). Audiovisual time perception is spatially specific. *Experimental Brain Research*, 218(3), 477–485. doi:10.1007/s00221-012-3038-3
- Holcombe, A. (2013). The temporal organization of perception. In J. Wagemans (Ed.), *Handbook of perceptual organization*. Oxford, UK: Oxford University Press.
- Ikumi, N., & Soto-Faraco, S. (2014). Selective attention modulates the direction of audio-visual temporal recalibration. *PLoS One*, 9(7), e99311. doi:10.1371/journal.pone.0099311
- Koenderink, J., Richards, W., & van Doorn, A. J. (2012). Space-time disarray and visual awareness. *i-Perception*, 3(3), 159–165.
- Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance*, 37(1), 245–256.
- 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(2), 499–507.
- Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Adaptation to audiotactile asynchrony. *Neuroscience Letters*, 413(1), 72–76.
- Nicol, J., & Shore, D. (2007). Perceptual grouping impairs temporal resolution. *Experimental Brain Research*, 183(2), 141–148. doi:10.1007/s00221-007-1034-9
- Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, 22(1), 46–49.
- Parise, C. V., Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisensory Research*, 26, 307–316.
- Pariyadath, V., & Eagleman, D. M. (2012). Subjective duration distortions mirror neural repetition suppression. *PLoS One*, 7(12), e49362. doi:10.1371/journal.pone.0049362
- Roseboom, W., & Arnold, D. H. (2011). Twice upon a time: Multiple concurrent temporal recalibrations of audiovisual speech. *Psychological Science*, 22(7), 872–877. doi:10.1177/0956797611413293
- Roseboom, W., Nishida, S., Fujisaki, W., & Arnold, D. H. (2011). Audiovisual speech timing sensitivity is enhanced in cluttered conditions. *PLoS One*, 6(4), e18309.
- Roseboom, W., Kawabe, T., & Nishida, S. (2013). Audio-visual temporal recalibration can be constrained by content cues regardless of spatial overlap. *Frontiers in Psychology*, 4, 189. doi:10.3389/fpsyg.2013.00189
- Schindel, R., Rowlands, J., & Arnold, D. H. (2011). The oddball effect: Perceived duration and predictive coding. *Journal of Vision*, 11(2). doi:10.1167/11.2.17
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4), 1–25. doi:10.3758/s13414-010-0073-7
- Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, 51(5), 651–659.
- Sugano, Y., Keetels, M., & Vroomen, J. (2014). Concurrent sensorimotor temporal recalibration to different lags for the left and right hand. *Frontiers in Psychology*, 5. doi: 10.3389/fpsyg.2014.00140
- Tanaka, A., Asakawa, K., & Imai, H. (2011). The change in perceptual synchrony between auditory and visual speech after exposure to asynchronous speech. *Neuroreport*, 22(14), 684–688.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3), 598–607.
- Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics*, 69(5), 744–756.
- Vatakis, A., & Spence, C. (2008). Evaluating the influence of the ‘unity assumption’ on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica*, 127(1), 12–23. doi:10.1016/j.actpsy.2006.12.002
- Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Temporal recalibration during asynchronous audiovisual speech perception. *Experimental Brain Research*, 181(1), 173–181.
- Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audiovisual asynchrony. *Cognitive Brain Research*, 22(1), 32–35.
- Yarrow, K., Roseboom, W., & Arnold, D. H. (2011). Spatial grouping resolves ambiguity to drive temporal recalibration. *Journal of Experimental Psychology: Human Perception and Performance*, 37(5), 1657–1661.
- Yuan, X., Li, B., Bi, C., Yin, H., & Huang, X. (2012). Audiovisual temporal recalibration: Space-based versus context-based. *Perception*, 41(10), 1218–1233.
- Yuan, X., Bi, C., Yin, H., Li, B., & Huang, X. (2014). The recalibration patterns of perceptual synchrony and multisensory integration after exposure to asynchronous speech. *Neuroscience Letters*, 569, 148–152. doi:10.1016/j.neulet.2014.03.057