Journal of Experimental Psychology: General

What You See Depends on What You Hear: Temporal Averaging and Crossmodal Integration

Lihan Chen, Xiaolin Zhou, Hermann J. Müller, and Zhuanghua Shi

Online First Publication, September 13, 2018. http://dx.doi.org/10.1037/xge0000487

What You See Depends on What You Hear: Temporal Averaging and Crossmodal Integration

Lihan Chen and Xiaolin Zhou
Peking University

Hermann J. Müller
Ludwig Maximilian University of Munich and Birkbeck College, University of London

Zhuanghua Shi
Ludwig Maximilian University of Munich

In our multisensory world, we often rely more on auditory information than on visual input for temporal processing. One typical demonstration of this is that the rate of auditory flutter assimilates the rate of concurrent visual flicker. To date, however, this auditory dominance effect has largely been studied using regular auditory rhythms. It thus remains unclear whether irregular rhythms would have a similar impact on visual temporal processing, what information is extracted from the auditory sequence that comes to influence visual timing, and how the auditory and visual temporal rates are integrated together in quantitative terms. We investigated these questions by assessing, and modeling, the influence of a task-irrelevant auditory sequence on the type of Ternus apparent motion: group motion versus element motion. The type of motion seen critically depends on the time interval between the two Ternus display frames. We found that an irrelevant auditory sequence preceding the Ternus display modulates the visual interval, making observers perceive either more group motion or more element motion. This biasing effect manifests whether the auditory sequence is regular or irregular, and it is based on a summary statistic extracted from the sequential intervals: their geometric mean. However, the audiovisual interaction depends on the discrepancy between the mean auditory and visual intervals: if it becomes too large, no interaction occurs—which can be quantitatively described by a partial Bayesian integration model. Overall, our findings reveal a crossmodal perceptual averaging principle that may underlie complex audiovisual interactions in many everyday dynamic situations.

Keywords: perceptual averaging, auditory timing, visual apparent motion, multisensory interaction, Bayesian integration

Most stimuli and events in our everyday environments are multisensory. It is thus not surprising that our brain often combines a heard sound with a seen stimulus source, even if they are in conflict. One typical such phenomenon, in a performance we enjoy, is the *ventriloquism effect* (Chen & Vroomen, 2013; Occelli, Bruns, Zampini, & Röder, 2012; Recanzone, 2009; Slutsky & Recanzone, 2001): we perceive the ventriloquist's voice as coming from the mouth of a dummy as if it was the dummy that is speaking. Of note in the present context, audiovisual integration has not only been demonstrated in spatial localization, but also in the temporal domain. In contrast to the dominance of vision in audiovisual spatial perception, audition dominates temporal processing, such as rhythm and interval. As an example, think of how we tend to auditorize a conductor's arm movements coordinating a musical passage, or Morse code flashes emanating from a naval ship. In fact, neuroscience evidence has revealed that

Lihan Chen, Center for Brain and Cognitive Sciences and School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, and Key Laboratory of Machine Perception (Ministry of Education), Peking University; Xiaolin Zhou, Center for Brain and Cognitive Sciences and School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Key Laboratory of Machine Perception (Ministry of Education), and PKU-IDG/McGovern Institute for Brain Research, Peking University; Hermann J. Müller, Department Psychologie, Ludwig Maximilian University of Munich, and Department of Psychological Sciences, Birkbeck College, University of London; Zhuanghua Shi, Department of Psychologie, Ludwig Maximilian University of Munich.

Zhuanghua Shi and Lihan Chen contributed equally.

Part of the study has been presented as a talk at the 17th International Multisensory Research Forum (IMRF, June 2016, Suzhou, China). This study was supported by grants from the Natural Science Foundation of China (Grants 31200760, 61621136008, and 61527804), Deutsche Forschungsgemeinschaft Project SH166 3/1 and projektbezogener Wissenschaftleraustausch (ProWA). The data and the source code of statistical analysis and modeling are available at https://github.com/msenselab/temporal_averaging.

Correspondence concerning this article should be addressed to Lihan Chen, School of Psychological and Cognitive Sciences, Peking University, 5 YiHeYuan Road, Beijing 100871, China. E-mail: clh@pku.edu.cn

information for time estimation is encoded in the primary auditory cortex for both visual and auditory events (Kanai, Lloyd, Bueti, & Walsh, 2011). This is consistent with the proposal that the perceptual system automatically abstracts temporal structure from rhythmic visual sequences and represents this structure using an auditory code (Guttman, Gilroy, & Blake, 2005).

Another compelling demonstration of how auditory rhythm influences visual temporal processing is known as the *auditory driving effect* (Boltz, 2017; Gebhard & Mowbray, 1959; Knox, 1945; Shipley, 1964): the phenomenon that variation in auditory flutter rate may noticeably influence the rate of perceived visual flicker. This influence, though, is dependent on the disparity between the auditory and visual rates (Recanzone, 2003). Quantitatively, this influence has been described by a Bayesian model of audiovisual integration (Roach, Heron, & McGraw, 2006), which assumes that the brain takes into account prior knowledge about the discrepancy between the auditory and visual rates in determining the degree of audiovisual integration. Auditory driving is a robust effect that generalizes across different types of tasks, including temporal adjustment and production (Myers, Cotton, & Hilp, 1981) and perceptual discrimination (Welch, DuttonHurt, & Warren, 1986), and it may even be seen in the effect of one single auditory interval on a subsequent visual interval (Burr, Della Rocca, & Morrone, 2013).

It should be noted, however, that auditory driving has primarily been investigated using regular rhythms, the implicit assumption being that the mean auditory rate influences the mean visual rate. On the contrary, studies on *ensemble coding* (Alvarez, 2011; Ariely, 2001) suggest that perceptual averaging can be rapidly accomplished even from a set of various objects or events; for example, we can quickly estimate the average size of apples in a supermarket display, or the average tempo of a piece of music. With regard to the present context, audiovisual integration, it remains an open question how the average tempo in audition quantitatively influences the temporal processing of visual events. This question has become prominent as the mechanisms underlying perceptual averaging processes themselves are still a matter of debate. There is evidence that the mental scales underlying the representation of magnitude (e.g., visual numerosity and temporal duration) are nonlinear rather than linear (Allan & Gibbon, 1991; Dehaene, Izard, Spelke, & Pica, 2008; Nieder & Miller, 2003). It has also been reported that, in temporal bisection (i.e., comparing one interval with two reference intervals), the subjective midpoint between one short and one long reference duration is closer to their geometric, rather than their arithmetic, mean (Allan & Gibbon, 1991). However, it remains to be established whether temporal rate averaging obeys the principle of the arithmetic mean (AM) or the geometric mean (GM), which might have implications for a broad range of mechanisms coding magnitude in perception (Walsh, 2003).

On these grounds, the aim of the present study was to quantify temporal rate averaging in a cross-modal, audiovisual scenario using irregular auditory sequences. To this end, we adopted and extended the *Ternus temporal ventriloquism* paradigm (Shi, Chen, & Müller, 2010), which we used previously to investigate cross-modal temporal integration. In the standard Ternus temporal ventriloquism paradigm, two auditory beeps are paired with two visual Ternus frames. Visual Ternus displays (Figure 1) can elicit two

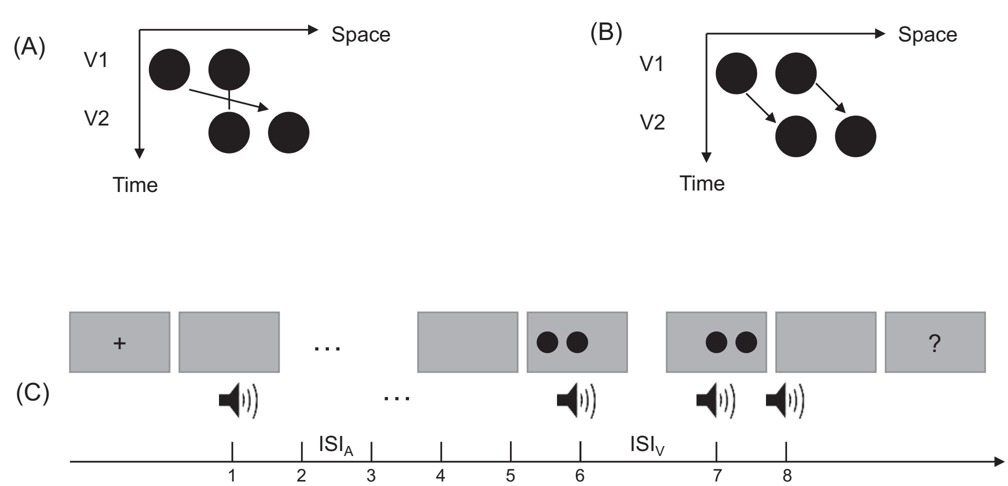


Figure 1. Ternus displays and stimulus configurations. Two alternative motion percepts of the Ternus display: (A) element motion for short interstimulus interval (ISI), with the middle dot perceived as remaining static while the outer dots are perceived to move from one side to the other, and (B) group motion for long ISIs, with the two dots perceived as moving in tandem. (C) Schematic illustration of the stimulus configuration used in the experiments. The auditory sequence consisted of 8–10 beeps. Two of the beeps (the 6th and the 7th) were synchronously paired with two visual Ternus frames, which were separated by a visual ISI (ISI_V) that varied from 50 to 230 ms (for the critical beeps, $ISI_V = ISI_A$). The other auditory ISIs (ISI_A) were systematically manipulated such that the mean of the ISI_A preceding the visual Ternus display was 50–70 ms shorter than, equal to, or 50–70 ms longer than the transition threshold between the element- and group-motion percepts of the visual Ternus events. The transition threshold was first estimated individually for each observer in a pretest session. During the experiment, observers were simply asked to indicate the type of visual motion (element or group) that they had perceived, while ignoring the beeps.

distinct percepts of visual apparent motion: *element* or *group* motion, where the percept of apparent motion is mainly determined by the visual interstimulus interval (ISI_V) between the two display frames (with other stimulus settings being fixed). Element motion is typically observed with short ISI_V (e.g., of 50 ms), and group motion with long ISI_V (e.g., of 230 ms; see Figure 1A and 1B). When two beeps are presented in temporal proximity to, or synchronously with, the two visual frames, the beeps can systematically bias the transition threshold between the two percepts of visual apparent motion: either toward element motion (if the auditory interval, ISI_A, is shorter than the visual interval) or toward group motion (if ISI_A is longer than the visual interval; Shi et al., 2010). Similar temporal ventriloquism effects have also been found with other tasks, such as temporal order judgments (for a review, see Chen & Vroomen, 2013). Here, we extended the Ternus temporal ventriloquism paradigm by presenting a whole sequence of beeps prior to the Ternus display frames, in addition to the two beeps paired with Ternus frames (see Figure 1C; recall that previous studies had presented just the last two beeps) to examine the influence of the temporal averaging of auditory intervals on visual apparent motion.

Experiment 1 was designed, in the first instance, to demonstrate an auditory driving effect using this new paradigm. In Experiment 2, we went on to examine whether temporal averaging with irregular auditory sequences would have a similar impact on visual apparent motion. In Experiment 3, we manipulated the variability of the auditory sequence to examine for (and quantify) influences of the variability of the auditory intervals on visual apparent motion. In Experiment 4, we further determined which type of temporal averaging statistic, the AM or the GM of the auditory intervals, influences visual Ternus apparent motion. And Experiment 5 was designed to rule out a potential confound, namely, a recency effect with the last auditory interval dominating the Ternus motion percepts in the cross-modal temporal averaging. Finally, we aimed to identify the computational model that best describes the cross-modal temporal interaction: mandatory full Bayesian integration versus partial integration (Ernst & Banks, 2002; Roach et al., 2006).

Materials and Method

Participants

A total of 84 participants (21, 22, 16, 12, 12 in Experiments 1–5; age ranging from 18–33 years) took part in the main experiments. All observers had normal or corrected-to-normal vision and reported normal hearing. The experiments were performed in compliance with the institutional guidelines set by the Academic Affairs Committee of the Department of Psychology, Peking University (approved protocol of "#Perceptual averaging [2012-03-01]"). All observers provided written informed consent according to the institutional guidelines prior to participating and were paid for their time on a basis of 20 CNY/hr.

The number of participants recruited for Experiments 1 and 2 was based on the effect size in our previous study of the temporal Ternus ventriloquism effect (Shi et al., 2010), where the pairing of auditory beeps with the visual Ternus display yielded a Cohen's *d* greater than 1 for the modulation of the Ternus motion percepts. We thus used a conservative effect size of 0.25 and a power of 0.8 for the estimation and recruited more than the estimated sample size (of 15 participants). Given that the effect we aimed to examine turned out to be very reliable, we used a standard sample size of 12 participants in Experiments 4 and 5.

Apparatus and Stimuli

The experiments were conducted in a dimly lit (luminance: 0.09 cd/m^2) cabin. Visual stimuli were presented in the central region of a 22-in. CRT monitor (FD 225P, Qing Dao, China), with a screen resolution of 1,024 × 768 pixels and a refresh rate of 100 Hz. Viewing distance was 57 cm, maintained by using a chin rest.

A visual Ternus display consisted of two stimulus frames, each containing two black disks (10.24 cd/m^2; disk diameter and separation between disks: 1.6 and 3° of visual angle, respectively) presented on a gray background (16.1 cd/m^2). The two frames shared one element location at the center of the monitor, while containing two other elements located horizontally on opposite positions relative to the center (see Figure 1). Each frame was presented for 30 ms; the interstimulus interval (ISI_V) between the two frames was randomly selected from the range of 50–230 ms, with a step size of 30 ms.

Mono sound beeps (1000 Hz, 65 dB, 30 ms) were generated and delivered via an M-Audio card (Delta 1010, BeiJing, China) to a headset (Philips SHM1900, BeiJing, China). To ensure accurate timing of the auditory and visual stimuli, the durations of the visual stimuli and the synchronization of the auditory and visual stimuli were controlled via the monitor's vertical synchronization pulses. The experimental program was written with Matlab (Mathworks, Natick, MA) and the Psychophysics Toolbox (Brainard, 1997).

Experimental Design

Practice. Prior to the formal experiment, participants were familiarized with visual Ternus displays of either typical element motion (with an ISI_V of 50 ms) or typical group motion (ISI_V of 260 ms) in a practice block. They were asked to discriminate the two types of apparent motion by pressing the left or the right mouse button, respectively. The mapping between response button and type of motion was counterbalanced across participants. During practice, when a response was made that was inconsistent with the typical motion percepts, immediate feedback was presented on the screen showing the typical response (i.e., element or group motion). The practice session continued until the participant reached a conformity of 95%. All participants achieved this criterion within 120 trials, given that the two extreme ISIs used (50 and 260 ms, respectively) gave rise to nonambiguous percepts of either element motion or group motion.

Pretest. For each participant, the transition threshold between element and group motion was determined in a pretest session. A trial began with the presentation of a central fixation cross for 300 to 500 ms. After a blank screen of 600 ms, the two Ternus frames were presented synchronized with two auditory tones (i.e., baseline: $ISI_V = ISI_A$); this was followed by a blank screen of 300 to 500 ms, prior to a screen with a question mark prompting the participant to make a two-forced-choice response indicating the type of perceived motion (element or group motion). The ISI_V between the two visual frames was randomly selected from one of the following seven intervals: 50, 80, 110, 140, 170, 200, and 230 ms. There were 40 trials for each level of ISI_V, counterbalanced with left- and rightward apparent motion. The presentation order of the trials was randomized for each participant. Participants performed a total of 280 trials, divided into four blocks of 70 trials each. After completing the pretest, the psychometric curve was fitted to the proportion of group motion responses across the seven intervals (see the Data Analysis and Modeling section). The transition threshold, that is, the point of subjective equality (PSE) at which the participants were equally likely to report the two motion percepts, was calculated by estimating the ISI at the point on the fitted curve that corresponded to 50% of group motion reports. The just noticeable difference (JND), an indicator of the sensitivity of apparent motion discrimination, was calculated as half of the difference between the lower (25%) and upper (75%) bounds of the thresholds from the psychometric curve.

Main experiments. In the main experiments, the procedure of visual stimulus presentation was the same as in the pretest session, except that prior to the occurrence of the two Ternus display frames, an auditory sequence consisting of a variable number of 6–8 beeps was presented (see below for the details of the onset of the Ternus display frames relative to that of the auditory sequence). As in the pretest, the onsets of the two visual Ternus frames (each presented for 30 ms) was accompanied by a (30-ms) auditory beep (i.e., $ISI_V = ISI_A$). A trial began with the presentation of a central fixation marker, randomly for 300 to 500 ms. After a 600-ms blank interval, the auditory train and the visual Ternus frames were presented (see Figure 1c), followed sequentially by a blank screen of 300 to 500 ms and a screen with a question mark at the screen center prompting participants to indicate the type of motion they had perceived: element versus group motion (nonspeeded response). Participants were instructed to focus on the visual task, ignoring the sounds. After the response, the next trial started following a random intertrial interval of 500 to 700 ms.

In Experiment 1 (regular sound sequence), the audiovisual Ternus frame was preceded by an auditory sequence of 6–8 beeps with a constant inter-stimulus interval (ISI_A), manipulated to be 70 ms shorter than, equal to, or 70 ms longer than the transition threshold estimated in the pretest. The total auditory sequence consisted of 8–10 beeps, including those accompanying the two visual Ternus frames, with the latter being inserted mainly at the sixth to eighth position, and followed by 0–2 beeps (number selected at random), to minimize expectation as to the onset of the visual Ternus frames. Visual Ternus frames were presented on 75% of all trials (504 trials in total). The remaining 25% were catch trials (168 trials) to break anticipatory processes. All trials were randomized and organized into 12 blocks, each block containing 56 trials. The ISI_V between the two visual Ternus frames was randomly selected from one of the following seven intervals: 50, 80, 110, 140, 170, 200, and 230 ms.

In Experiment 2 (irregular sound sequence), the settings were the same as in Experiment 1, except that the auditory train were irregular: the ISI_A between adjacent beeps in the auditory train (except the ISI_A between the beeps accompanying the visual Ternus frames) were varied $\pm$20 ms uniformly and randomly around (i.e., they were either 20 ms shorter or 20 ms longer than) a given mean interval (three levels: 70 ms shorter than, equal to, or 70 ms longer than the individual transition threshold).

Experiment 3 introduced two levels of variability in the auditory-interval sequence with 8–10 beeps: a low coefficient of variance (CV, the standard deviation divided by the mean) of 0.1 and, respectively, a high CV of 0.3. For each CV condition, three AM intervals were used: 50 ms shorter than, equal to, or 50 ms longer than the estimated transition threshold. The intervals were randomly generated from a normal distribution with a given mean and CV. The number of the experimental trials was 1,008, and the catch trials totaled 336. All trials were randomized and organized into 24 blocks, each block containing 56 trials.

Experiment 4 used three types of auditory sequences, each consisting of six intervals: (a) baseline auditory sequence: three intervals of 110, 140, and 170 ms, were repeated twice in random order; in this baseline condition, the AM (AM = 140 ms) was near-equal to the GM (GM = 138 ms); (b) AM-deviated (AriM) sequence: six intervals were constructed from ISI_A of 70, 140, and 280 ms, which were arranged randomly (AM = 163 ms > GM = 140 ms); and (c) GM-deviated (GeoM) sequence: six intervals constructed from ISI_A of 50, 140, and 230 ms, arranged randomly (GM = 117 ms < AM = 140 ms). The audiovisual Ternus frames were appended at the end of the sequences. The number of experimental trials was 504 (there were no catch trials), which were presented randomized and organized in 12 blocks, each block containing 42 trials.

To exclude potential confounding by a recency effect, in Experiment 5, we compared two auditory sequences: one with a GM 70 ms shorter than the transition threshold of visual Ternus motion (henceforth referred to as the short condition), and the other with a GM 70 ms longer than the transition threshold (long condition). Instead of completely randomizing the five auditory intervals (excepting the final synchronous auditory interval with the visual Ternus interval), the last auditory interval before the onset of the Ternus display was fixed at the transition threshold for both sequences. The remaining four intervals were chosen randomly such that the CV of the auditory sequence was in the range between 0.1 and 0.2. This manipulation was expected to minimize the influence of any potential recency effect engendered by the last auditory interval. The audiovisual Ternus frames were appended at the end of the sequences on trials (i.e., 672 out of a total of 784 trials) on which the Ternus display was presented at the end of the sound sequence (the onset of the first visual frame was synchronized with the 6th beep). The remaining (112) trials were catch trials, with 56 trials each in which the Ternus display occurred at the beginning of the sound sequence (i.e., the onset of the first visual frame was synchronized with the second beep) or, respectively, at a middle temporal location (i.e., the onset of the first visual frame was synchronized with the 4th beep). These catch trials were introduced to prevent participants from consistently anticipating the visual event to occur at the end of the sound sequence. The total 784 trials were randomized and organized in 14 blocks, each block containing 56 trials.

Data Analysis and Modeling

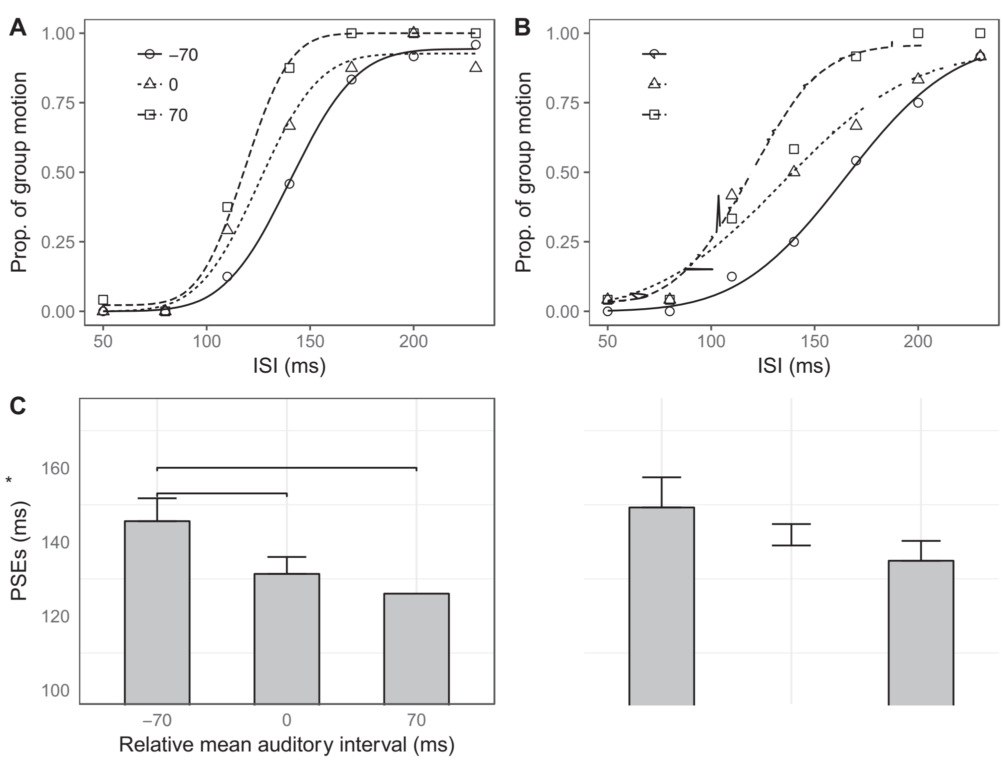
We used the R package Quickpsy (Linares & López-Moliner, 2016) to fit psychometric curves with upper and lower asymptotes, which provide better estimates of the threshold (Wichmann & Hill, 2001). Bayesian modeling was also conducted with R. We first calculated the response proportion for the baseline test with (a dio-) i al Tĕ n a ăen mo ion and fŏ he fŏ mal ex ĕ-imen , a ell a fi ing he cŏ‶e onding c m la i e Ga ian chome‶ ic f nc ion . Ba ed on he chome‶ ic f nc ion , e co ld hen e ima e he di ĉimina ion ăiabili of Tĕ n a ă-en mo ion (i.e., σ_m) ba ed on he andă d de ia ion of he c m la i e Ga ian f nc ion. The ăameĕ of he Ba e ian model (ee Ba e ian modeling ec ion belo) ĕe e ima ed b minimi ing he ‶edic ion ĕ‶ŏ ing he R ŏ lm f nc ion. O‶ ‶a da a, oge hĕ i h he o‶ce code of a i ical anal e and Ba e ian modeling, ăe a ailable a he gi h b‶e o i ŏ : h :// gi h b.com/m en elab/ em ŏ al_a ĕ aging.

Results

Experiments 1 and 2: Both Regular and Irregular Auditory Intervals Alter the Visual Motion Percept

We mani la ed he in ĕ al be een cce i e bee (i.e., he ISI_A ‶ iŏ o he Tĕ n di la) o be ei hĕ ‶eg lă ŏ ĭ‶eg lă, b i h hĕ AM being ei hĕ 70 m hŏ ĕ, e al o, ŏ 70 m longĕ han he ‶an i ion ĥ e hold (mea ‶ed in he ‶e e) be een elemen - and g‶o -mo ion ‶e ŏ (fŏ bo h ‶eg lă and ĭ‶eg lă ISI_A). A di ŏ e ence i h a ‶ela i el long mean a di ŏ in ĕ al, a com ăed i h a hŏ in ĕ al, ĕe fo nd o elici mŏe ‶e ŏ of g‶o mo ion, a indica ed b he mallĕ PSE (Fig ‶e 2), fŏ bo h ‶eg lă in ĕ al , $F(2, 40) = 12.22$, $p < .001$, $\eta_g^2 = 0.112$, and ĭ‶eg lă in ĕ al , $F(2, 42) = 8.25$, $p < .001$, $\eta_g^2 = 0.04$. Tha i , he ĕcei ed i al in ĕ al (hich de ĕ mine he en ing mo ion ĕ ce) a a imila ed b he a ĕ age of he ‶eceding a di ŏ in ĕ al , ‶egă dle of he hĕ he a di ŏ in ĕ al ĕe ‶eg lă ŏ ĭ‶eg lă. Po hoc Bonfĕ-‶oni com ă i on e ‶e ealed ha hi a imila ion effec a mainl d‶i en b he hŏ a di ŏ in ĕ al in bo h ex ĕ imen : p ĕ e 0.001, 0.00001, and 0.57 fŏ he com ă i on : −70 ĕ 0 m , −70 ĕ 70 m , and, ‶e ec i el , 0 ĕ 70 m fŏ he ‶eg lă in ĕ al ; and 0.015, 0.0002, 0.77 fŏ he com ă i on of he ĭ‶eg lă in ĕ al (Fig ‶e 2C and 2D).

The fac ha a‶o modal a imila ion effec a ob ained e en i h ĭ‶eg lă a di ŏ e ence gge ha he effec i n-likel d e o em ŏ al ex ec a ion, ŏ a genĕ al effec of a di ŏ en‶ ainmen (Jone , Mo nihan, MacKen ie, & P en e, 2002; Lă ge & Jone , 1999). In addi ion, he a imila ion effec ob ĕ ed



is unlikely due to a recency effect. To examine for such an effect, we split the trials into two categories according to the auditory intervals that just preceded the visual Ternus intervals: shorter and longer preceding intervals with reference to the auditory mean interval. The length of the immediately preceding interval failed to produce any significant modulation of apparent visual motion, $F(1, 22) = 2.14$, $p = .15$. An account in terms of a recency effect was further ruled out by a dedicated control experiment that directly fixed the last auditory interval (see Experiment 5 below).

Furthermore, in the regular condition, the mean JND ($\pm SE$) for the three ISI_V conditions (34.9 [±3.1], 30.5 [±3.4], and 28.4 [±2.9] ms for the ISI_V 70 ms shorter, equal to, and, respectively, 70 ms longer relative to the transition threshold) were larger than the JND for the threshold (baseline) condition (18.8 [±1.2] ms; $p = .001$, $p = .002$, and $p = .033$ for the shorter, equal, and longer conditions (vs. the threshold), with no differing among themselves (all $p > 0.1$). The same held true for the irregular condition: JNDs of 31.8 (±3.2), $p = .001$, 30.6 (±2.3), $p = .005$, and 27.2 (±2.2) ms compared with the baseline 18.6 (±2.1) ms, with no differing among themselves (all $p > 0.1$). The observed sensitivities in the three conditions with auditory beeps train suggest that the assimilation effect observed here was not attributable to attentional entrainment, as attentional entrainment would have been expected to enhance the sensitivity.

Experiment 3: Variability of Auditory Intervals Influences Visual Ternus Apparent Motion

According to quantitative models of multisensory integration (Ernst & Di Luca, 2011; Shi, Church, & Meck, 2013), the strength of the assimilation effect would be determined by the variability of both the auditory intervals and the visual Ternus interval, assuming that information is integrated from all intervals. According to optimal full integration, high variance of the auditory sequence would result in a low auditory weight in audiovisual integration, leading to a weaker assimilation effect compared with low variance. To examine for effects of the variance of the auditory intervals on visual Ternus apparent motion, we directly manipulated the relative standard deviation of the auditory intervals while fixing their AM. One key property of time perception is that it is scalar (Church, Meck, & Gibbon, 1994; Gibbon, 1977), that is, the estimation error increases linearly as the time interval increases, approximately following Weber's law. Given this, we used CVs, that is, the ratio of the standard deviation to the mean, to manipulate standardized variability across multiple auditory intervals. Specifically, we compared a low CV (0.1) with a high CV (0.3) condition, with an orthogonal variation of the (arithmetic) mean auditory interval: 50 ms shorter, equal to, or 50 ms longer than the predetermined transition threshold.

The main effect of mean interval was significant, $F(2, 30) = 11.8$, $p < .001$, $\eta_g^2 = 0.078$, with longer intervals leading to more reports of group motion (i.e., lower PSEs: mean PSE of 132 ± 4.6 ms), shorter intervals to fewer reports of group motion (i.e., higher PSEs: mean PSE of 147 ± 6.7 ms), and equal intervals to an intermediate proportion of group-motion reports (mean PSE of 138 ± 5.3 ms). Post hoc Bonferroni comparisons revealed this pattern to be similar to that observed in Experiments 1 and 2: significant difference between the short and equal intervals ($p < .01$) and the short and long intervals ($p < .001$), but no between the equal and long intervals ($p = .49$). Interestingly, the main effect of CV was significant (though the effect size is small), $F(1, 15) = 5.29$, $p < .05$, $\eta_g^2 = 0.044$, while the interaction between mean interval and CV was not, $F(2, 30) = 0.31$, $p = .73$, $\eta_g^2 = 0.0008$ (Figure 3). Further examination for a (potentially confounding) recency effect, adopting the same comparison as for the previous experiments, yielded no evidence that the main effect we obtained was attributable to the length of the auditory interval immediately preceding the visual interval, $F(1, 15) = 0.33$, $p = .55$.

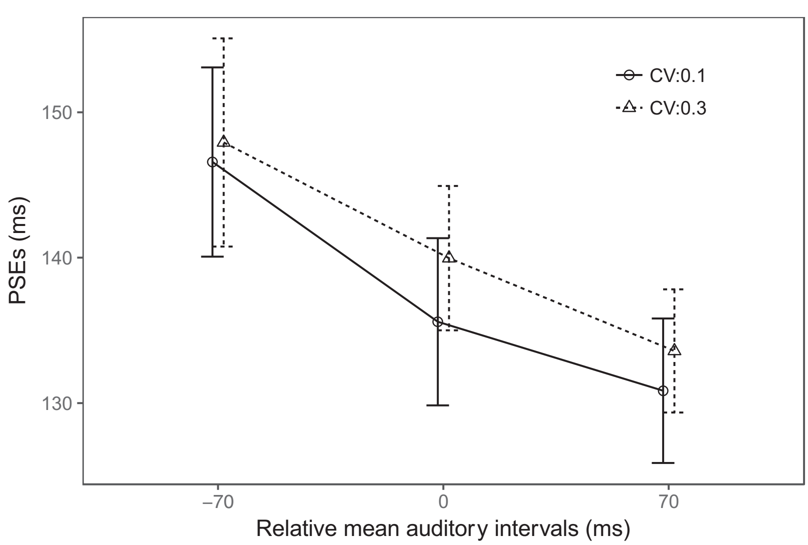


Figure 3. Points of subjective equality (PSEs) between element- and group-motion reports for auditory beep trains with a low and a high coefficient of (auditory-interval) variance (CV, 0.1 or 0.3), as a function of the (arithmetic) mean auditory interval (50 ms shorter, equal to, or 50 ms longer than the pretest transition threshold).

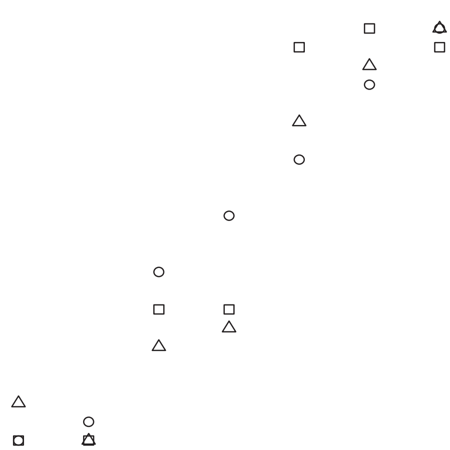
The results were interesting in two respects. First, according to mandatory, full Bayesian integration (see the Bayesian Modeling section below for details), auditory-interval variability should affect the weight of the cross-modal temporal integration (Burr, 1999; Shi et al., 2013), with greater variance lessening the influence of the average auditory interval. Accordingly, the slope of the fitted line in Figure 2 would be expected to be flatter under the high compared with the low-CV condition, yielding an interaction between mean interval and CV. The fact that this interaction was nonsignificant suggests that the ensemble mean of the auditory interval is not fully integrated with the visual interval (we will return to this point in the Bayesian Modeling section). Second, the downward shift of the PSEs in the low, compared with the high, CV condition indicates that the perceived auditory mean interval (that influences the audio-visual integration) is actually not the AM that was manipulated. An alternative account of this shift may derive from the fact that the auditory sequence with higher CV has a lower GM than the sequence with low variance, that is: the perceived ensemble mean is likely geometrically encoded. Experiment 4 was designed to address this (potential) confound by directly comparing the effects of ensemble coding based on the GM versus the AM.

Experiment 4: Perceptual Averaging of Auditory Intervals Assimilates the Visual Interval Toward the GM Rather Than the AM

In Experiment 4, we compared three types of auditory sequence in our audiovisual Ternus apparent motion paradigm: a baseline sequence, an AriM sequence, and a GeoM sequence. The PSEs were 136 (±5.46), 148 (±6.17), and 136 (±6.2) ms for the AriM, the GeoM, and the baseline conditions, respectively, $F(2, 22) = 8.81$, $p < .05$, $\eta_g^2 = 0.08$ (Figure 4). Bonferroni-corrected comparisons revealed the transition threshold to be significantly larger for the GeoM compared with the baseline condition, $p < .01$, whereas there was no difference between the AriM and the baseline condition, $p = 1$. This pattern indicates that ensemble coding of the auditory intervals assimilates the visual interval toward the GM rather than the AM.

Experiment 5: Auditory Sequences With the Last Interval Fixed

In Experiments 1–3, we split the data according to the last interval (i.e., the interval preceding the visual Ternus display) of the auditory sequence into two categories (short vs. long), which failed to reveal an influence of the last interval. In Experiment 5, we formally manipulated the last interval by fixing it at the respective transition thresholds for the short and long auditory sequences (i.e., sequences with the smaller and, respectively, larger GMs). Figure 5 depicts the responses of a typical participant from Experiment 5. The PSEs were 153.1 (±7.3) and, respectively, 137.9 (±9.1) for the short and long conditions, respectively, $t(11) = 3.640$, $p < .01$. That is, reports of element motion were more dominant in the short than in the long condition, replicating the findings of the previous experiments. In other words, it was the mean auditory interval, rather than the last interval (prior to the Ternus frame), that assimilated visual Ternus apparent motion. Given this, the audio-visual interactions we found here are unlikely to be attributable to a recency effect.



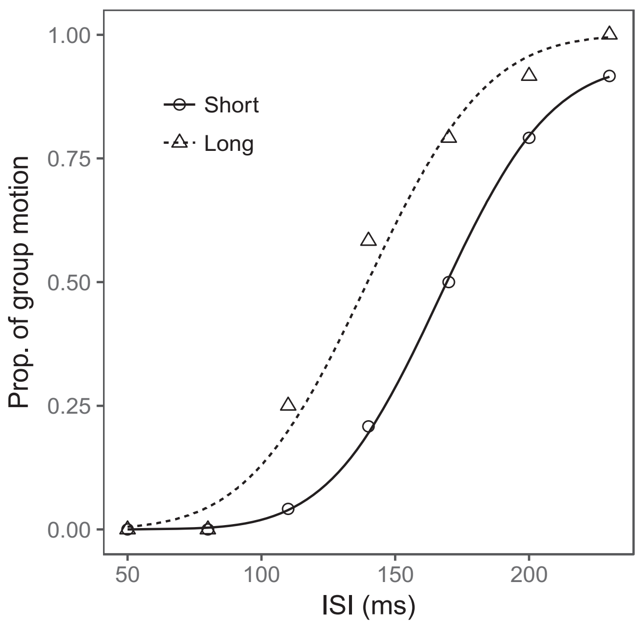


Figure 5. Mean proportion of group-motion responses from a typical participant as a function of the probe visual interstimulus interval (ISI), and fitted psychometric curves, for the two geometric mean conditions: the short sequence (with the smaller geometric mean) and the long sequence (with the larger geometric mean).

Bayesian Modeling

To account for the above findings, we implemented, and compared, two variants of Bayesian integration models: mandatory full Bayesian integration and partial Bayesian integration. If the ensemble-coded auditory-interval mean (A) and the audio visual Ternus display interval (M) are fully integrated according to the maximum likelihood estimation (MLE) principle (Ernst & Banks, 2002), and both are normally distributed (e.g., fluctuating due to internal Gaussian noise) that is: $A \sim N(I_a, \sigma_a)$, $M \sim N(I_m, \sigma_m)$ the expected optimally integrated audio-visual interval, which yields minimum variability, can be predicted as follows:

$$\hat{I}_{full} = wI_a + (1 - w)I_m, \tag{1}$$

where $w = (1/\sigma_a^2)/(1/\sigma_a^2 + 1/\sigma_m^2)$ is the weight of the averaged auditory interval, which is proportional to its reliability. Note that full optimal integration is typically observed when the two cues are close to each other, but it breaks down when their discrepancy becomes too large (Körding et al., 2007; Parise, Spence, & Ernst, 2012; Roach et al., 2006). In our study, the Ternus interval and the mean auditory interval could differ substantially on some trials (e.g., visual interval of 50 ms paired with mean auditory interval of 210 ms). Given this, a more appropriate model would need to take a discrepancy prior and the causal structure (Körding et al., 2007) of audio-visual temporal integration into consideration. Thus, similar to Roach et al. (2006), here we assume that the probability of full integration P_{am} depends on the discrepancy between the mean auditory and Ternus intervals:

$$P_{am} \sim e^{-(I_a - I_m)^2 / \sigma_{am}^2}, \tag{2}$$

where σ_{am}^2 is the variance of the error measure of the discrepancy between the ensemble mean of the auditory intervals and the visual interval. P_{am} will vary from trial to trial, depending on the discrepancy between the mean auditory interval and the visual interval. Thus, a more general, partial integration model would predict:

$$\hat{I}_{av} = P_{am}\hat{I}_{full} + (1 - P_{am})I_v. \tag{3}$$

Combined with Equation 1, Equation 3 can be simplified as follows:

$$\hat{I}_{av} = (1 - wP_{am})I_v + wP_{am}I_a. \tag{4}$$

To compare the full-integration and partial-integration models, we took into account the data from those of our experiments that manipulated the auditory-interval regularity and variability (Experiments 1–3; we excluded Experiments 4 and 5, as these did not include a baseline task of Ternus apparent-motion perception; see the Materials and Methods sections). Given that the baseline task provided an estimate of σ_m, there is one parameter σ_a for the full-integration model and two parameters σ_a and σ_{am} for the partial-integration model, which are free parameters for fitting. This was carried out using the optimization algorithm L-BFGS in R (see our source code at https://github.com/msenselab/ensemble_averaging). We assessed the goodness of the resulting fits by means of coefficients of determination (R^2) and Bayesian information criteria (BIC). The BIC and R^2 scores are presented in Table 1. As can be seen, the BIC differences between the partial- and full-integration models are large for all experiments, clearly favoring the partial-integration model (Kass & Raftery, 1995). The R^2 values also confirm this finding.

To visualize how well the partial-integration model predicts behavioral performance, we calculated the predicted mean responses based on the partial-integration model for individual visual ISI across all experimental conditions. Figure 6 illustrates the predictions, indicated by curves, together with the observed mean responses, indicated by the points. As can be seen, the predicted mean responses are within one standard error of the observed mean responses (see Figure 6).

The key difference between the full- and partial-integration models is that the latter takes the probability of cross-modal integration into account; accordingly, the weight of the auditory ensemble interval (i.e., wP_{am}) depends on the difference between the ensemble mean of the auditory intervals and the visual interval.

Table 1
Model Comparison Using BIC and R^2 for the Partial- and Full-Integration Model

	Partial integration		Full integration		
Experiment	BIC	R^2	BIC	R^2	ΔBIC
Irregular	−1,859	.86	−1,392	.63	**467**
Regular	−1,932	.91	−1,772	.88	**160**
Variance	−2,894	.91	−2,878	.91	**16**

Note. The differential Bayesian information criterion (BIC) scores revealed the partial-integration model to outperform the full-integration model across all experiments (strong evidence in all experiments: ΔBIC >10). The absolute values of bold values are the difference between BIC score by partial-integration model and BIC score by full-integration model.

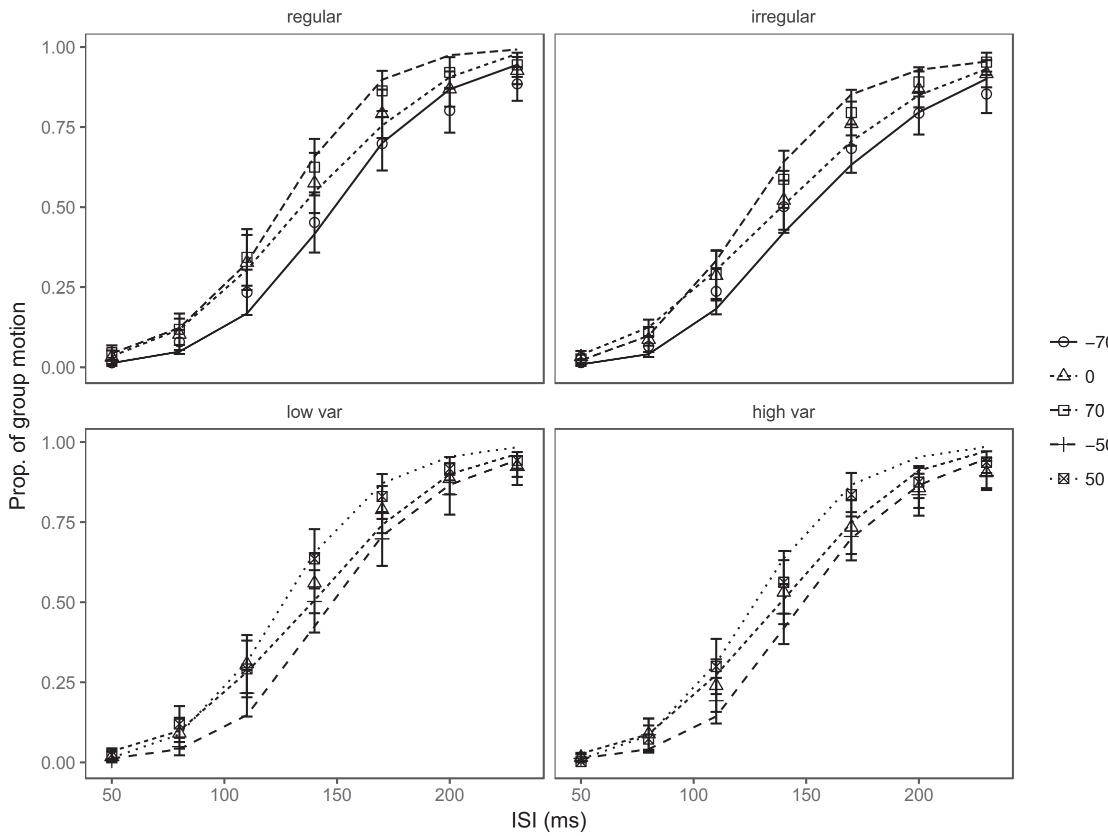


Figure 6. Mean behavioral responses (proportion of group-motion responses, indicated by shape points) and responses predicted by the spatial-integration model (indicated by curves) as a function of the visual interstimulus interval (ISI) of the Ternus display, separately for auditory sequences with different (arithmetic) mean interval relative to the individual transition thresholds. The relative-interval labels (−70, −50, 0, 50, and 70 ms) denote the magnitude of the difference between the mean auditory interval and the transition threshold. Error bars denote standard errors of mean (±*SEM*).

This can be seen in Figure 7, which illustrates the dynamic change of the auditory weights across the ratio audio-visual interval discrepancy conditions. All three experiments exhibit a similar pattern: weights are at their peak when the visual interval and the auditory mean interval are close to each other. For example, the peak for the relative intervals of 0 ms (i.e., the auditory mean interval set equal to the individual visual thresholds) are around 140 ms, close to the mean visual transition threshold (134.6 ms for regular and 135.3 ms for irregular sequences, and 139.0 ms for low and 144.8 ms for high variance). For relative intervals of 70 ms, the peaks are shifted rightward; and for relative intervals of −70 ms, the peaks are shifted leftward.

Based on these responses predicted by the spatial-integration model, we further calculated the predicted PSEs. Figure 8 shows a linear relation between the observed and predicted PSEs for all experiments. Linear regression revealed a significant linear correlation, with a slope of 0.978 and an adjusted R^2. The full-integration model, by contrast, produced flat psychometric curves for 6% of the individual conditions in Experiments 1 and 2 (due to the weight of the mean auditory interval approaching 1), which yielded unreliable estimates of the corresponding PSEs. This led to lower predictive power compared with the spatial-integration model, as evidenced by the BIC and R^2 scores (see Table 1). Thus, taken together, the spatial-integration model can well explain the behavioral data that we observed.

General Discussion

Using an audiovisual Ternus apparent motion paradigm, we conducted five experiments on audiovisual temporal integration with regular and irregular auditory sequences presented prior to the (audio-) visual Ternus display. We found that perceptual averaging of both regular (Experiment 1) and irregular auditory sequences (Experiments 2 and 3) greatly influenced the timing of the subsequent visual interval, as expressed in systematic change of the transition threshold in visual Ternus apparent motion: longer mean auditory intervals elicited more reports of group motion, whereas shorter mean intervals gave rise to dominant element motion. In Experiment 4, we further found that the GM of the auditory intervals can explain the audiovisual interaction better than the AM. Further (post hoc) analyses and a purpose-designed experiment (Experiment 5) effectively ruled out an explanation of these findings in terms of a recency

averaging of the auditory sequence (regardless of its regularity) has exerted a great influence on the visual interval.

Temporal Averaging and Geometric Encoding

The present results indicate that the GM well encapsulates the summary statistic of the temporal structure hidden in a complex multisensory stream (Hanson, Heron, & Whitaker, 2008; Heron, Roach, Hanson, McGraw, & Whitaker, 2012). Previous work on numerosity had already suggested that the mental scale underlythe percept of the last auditory interval is assimilated by the preceding intervals (Nakajima, ten Hoopen, Hilkhuysen, & Sasaki, 1992; Nakajima et al., 2004), as well as in audiovisual interval judgments when auditory and visual intervals are presented sequentially (Burr et al., 2013). The present study demonstrated that such audiovisual integration will occur even when participants are explicitly told to ignore the (task-irrelevant) auditory sequence, suggesting that processes of top-down control cannot fully shield visual motion perception from audiovisual temporal integration.

Conclusion

It has long been known that auditory flutter drives visual flicker (Shipley, 1964) as a typical phenomenon of audiovisual temporal interaction with regular auditory sequences. Here, in five experiments, we demonstrated that irregular auditory sequences also capture temporal processing of subsequently presented visual (target) events, measured in terms of the biasing of Ternus apparent motion. Importantly, it is the geometric averaging of the auditory intervals that assimilates the visual interval between the two visual Ternus display frames, thereby influencing decisions on perceived visual motion. Further work is required to examine whether the principle of geometric averaging and partial cross-modal integration demonstrated here (for an audiovisual dynamic perception scenario) generalizes to other perceptual mechanisms underlying magnitude estimation in multisensory integration.

Context of the Research

Perceptual averaging of sensory properties, such as the mean number, size, and spatial layout of objects in a scene, has been documented extensively in the visual domain. It allows us to capture our environment at a glance, in summary form, overcoming attentional and working memory capacity limitations. This phenomenon prompted us to ask whether and, if so, how processes of perceptual averaging may also be applied in the temporal domain, specifically in (cross-modal) scenarios involving multiple interacting sensory streams. Thus, we designed a paradigm combining a task-irrelevant temporal sequence of auditory events with a task-relevant Ternus apparent motion, a phenomenon where we see two aligned dots either move together (e.g., to the left or right) or only one dot jumping across the other (a stationary) dot. What we see (group vs. element motion) is critically influenced by the temporal interval between the two Ternus display frames. What we found is that the irrelevant auditory sequence preceding the visual Ternus display alters the visual interval, thus biasing observers to see either more group motion or more element motion, depending on the GM of the preceding auditory intervals. This interaction depends on the distance between the (mean) auditory and the visual interval: if the distance becomes too large, no interaction occurs. Conceptually, the finding of temporal averaging over a sequence of auditory intervals and its subsequent influence on the visual interval

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences, 17,* 556–564. http://dx.doi.org/10.1016/j.tics.2013.09.009

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science, 145,* 1328–1330. http://dx.doi.org/10.1126/science.145.3638.1328

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport: For Rapid Communication of Neuroscience Research, 12,* 7–10. http://dx.doi.org/10.1097/00001756-200101220-00009

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7,* 483–488. http://dx.doi.org/10.1016/j.tics.2003.09.002

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics, 39,* 294–300. http://dx.doi.org/10.3758/BF03204939

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63,* 1293–1313. http://dx.doi.org/10.3758/BF03194544

Received August 30, 2017
Revision received May 16, 2018
Accepted June 29, 2018 ■