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 cross-modal 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 no surprise that our brain often combines a information for time estimation is encoded in the primary auditory cortex for both visual and auditory events (

distinct percepts of visual apparent motion: element or group motion, where the type 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 systematims. 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 proportions 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 participant was 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 onset 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 presen tation 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 frames was preceded by an auditory sequence of 6–8 beeps with a constant interstimulus 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–seventh positions, and followed by 0–2 beeps (number selected at random), to minimize expectations 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 up 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 trains 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 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 sequences 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.

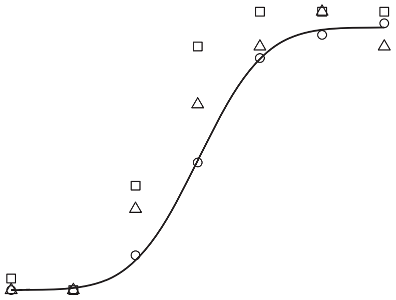
Experiment 4 used three types of auditory sequences, each consisting of six intervals: (a) baseline auditory sequence: three intervals, of 110, 140, and 170 ms, were repeated twice in random order; in this baseline condition, the AM (AM 140 ms) was near-equal to the GM (GM 138 ms); (b) AM-deviated (AriM) sequence: six intervals were constructed from ISI_A of 70, 140, and 280 ms, which were arranged randomly (AM163 ms GM (audio-) visual Ternus apparent motion and for the formal experiments, as well as fitting the corresponding cumulative Gaussian psychometric functions. Based on the psychometric functions, we could then estimate the discrimination variability of Ternus apparent motion (i.e., $_m$) based on the standard deviation of the cumulative Gaussian function. The parameters of the Bayesian models (see Bayesian modeling section below) were estimated by minimizing the prediction errors using the R optim function. Our raw data, together with the source code of statistical analyses and Bayesian modeling, are available at the github repository: https://github.com/msenselab/temporal_averaging

Results

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

We manipulated the intervals between successive beeps (i.e., the ISI$_A$ prior to the Ternus display) to be either regular or irregular, but with their AM being either 70 ms shorter, equal to, or 70 ms longer than the transition threshold (measured in the pretest) between element- and group-motion reports (for both regular and irregular ISI$_A$). Auditory sequences with a relatively long mean auditory interval, as compared with a short interval, were found to elicit more reports of group motion, as indicated by the smaller PSEs (Figure 2), for both regular intervals, F(2, 40) 12.22, p .001, ^2_g 0.112, and irregular intervals, F(2, 42) 8.25, p .001, ^2_g 0.04. That is, the perceived visual interval (which determines the ensuing motion percept) was assimilated by the average of the preceding auditory intervals, regardless of whether the auditory intervals were regular or irregular. Post hoc Bonferroni comparison tests revealed that this assimilation effect was mainly driven by the short auditory intervals in both experiments: ps were 0.001, 0.00001, and 0.57 for the comparisons 70 versus 0 ms, 70 versus 70 ms, and, respectively, 0 versus 70 ms for the regular intervals; and 0.015, 0.0002, 0.77 for the comparisons of the irregular intervals (Figure 2C and 2D).

The fact that a crossmodal assimilation effect was obtained even with irregular auditory sequences suggests that the effect is unlikely due to temporal expectation, or a general effect of auditory entrainment (Jones, Moynihan, MacKenzie, & Puente, 2002; Large & Jones, 1999). In addition, the assimilation effect observed



is unlikely due to a recency effect. To examine for such an effect, we split the trials into two categories according to the auditory interval that just preceded the visual Ternus interval: short and long 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 JNDs (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"), without differing among themselves (all ps 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, without differing among themselves (all ps 0.1). The worsened sensitivities in the three conditions with auditory beep trains 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 tof themaipu-lorystandiz(usey)

These results are 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 weights of the crossmodal temporal integration (Buus, 1999; Shi et al., 2013), with greater variance lessening the influence of the average auditory interval. Accordingly, the slopes of the fitted lines 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 intervals 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 we manipulated. An alternative account of this shthe

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 audiovisual 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: $A \quad N\, I_a, \quad _a, M \quad N\, I_m, \quad _m$ —the expected optimally integrated audio-visual interval, which yields minimum variability, can be predicted as follows:

$$\hat{I}_{full} \quad wI_a \quad (1 \quad w)I_m, \qquad (1)$$

where $w \quad 1 \; ^2_a \; 1 \; ^2_a \quad 1 \; ^2_m$ 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

This can be seen in Figure 7, which illustrates the dynamic changes of the auditory weights across the various 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 intervals are close to each other. For example, the peaks for the relative intervals of 0 ms (i.e., the auditory mean intervals were set 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, they are shifted leftward.

Based on the responses predicted by the partial-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 partial-integration model, as evidenced by the BIC and R^2 scores (see Table 1). Thus, taken together, the partial-integration model can well explain the behavioral data that we observed.

General Discussion

Using an audiovisual Ternus apparent motion paradigm, we

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

Temporal Averaging and Geometric Encoding

The present results indicate that the GM well encapsulates the summary statistics 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 scales underlying the representation of visual numerosity and temporal magnitudes are best characterized as being nonlinear, as opposed to linear, in nature (Dehaene, 2003; Dehaene et al., 2008; Nieder & Miller, 2003, 2004; Rips, 2013). For example, adults from the Munduruku, an Amazonian indigenous tribe with a limited number lexicon, map numerical quantities onto space in a logarithmic fashion (Dehaene et al., 2008; but see Cicchini, Arrighi, Cecchetti, Giusti, & Burr, 2012). A seminal study by Allan and Gibbon also showed that temporal bisection coincided with the GM of the two reference durations (Allan & Gibbon, 1991). Our findings reveal that extraction of the GM also underlies temporal averaging—and this might well be a principle shared by a broad range of mechanisms coding magnitude in perception (Walsh, 2003).

Partial Integration in Cross-Modal Temporal Processing

Research on multisensory integration has shown that the "proximity" and "similarity" of the spatiotemporal structure of multisensory signals—technically, their cross-correlation in time (and space)—is critical for inferring an underlying com-the 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 an audiovisual integration still occurs 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)—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 principles of geometric averaging and partial cross-modal integration demonstrated here (for an audiovisual dynamic perception scenario) generalize 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 visuospatial domain. It allows us to capture our environment at a glance, in summary terms—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 systems. Thus, we designed a paradigm combining a task-irrelevant temporal sequence of auditory events with 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 (apparently stationary) dot. What we see (group vs. element motion) is critically influenced by the temporal interval between the two Ternustotl4ea9g ar0001 Tm T* [umoswn thethethot.levant and,tory thotuences.prompe tTJ T* [(t

quebebin-539.27to quefin3 the-539.5(theporal)-518.5(th

Ernst, M., & Di Luca, M. (2011). Multisensory perception: From integration to remapping. In J. Trommershäuser (Ed.), 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., DutionHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception.