Attention, Perception, & Psychophysics (2018) 80:986–998
https://doi.org/10.3758/s13414-017-1475-6

Sound changes that lead to seeing longer-lasting shapes

Arthur G. Samuel[1,2,3] · Kavya Tangella[4]

Published online: 29 January 2018

Abstract
To survive, people must construct an accurate representation of the world around them. There is a body of research on visual scene analysis, and a largely separate literature on auditory scene analysis. The current study follows up research from the smaller literature on audiovisual scene analysis. Prior work demonstrated that when there is an abrupt size change to a moving object, observers tend to see two objects rather than one—the abrupt visual change enhances visible persistence of the briefly presented different-sized object. Moreover, if a sequence of tones accompanies the moving object, visible persistence is enhanced if the tone frequency suddenly changes at the same time that the object's size changes. Here, we show that although a sound change must occur at roughly the same time as a visual change to enhance visible persistence, there is a fairly wide time frame during which the sound change can occur. In addition, the impact of a sound change on visible persistence is not simply matter of the physical pattern: The same pattern of sound can enhance visible persistence or not, depending on how the pattern is itself perceived. Specifically, a change in a tone's frequency can enhance visible persistence when it accompanies a visual size change, but the same frequency change will not do so if the shift is embedded in a larger pattern that makes the change merely a continuation of alternating frequencies. The current study supports a scene analysis process that is both multimodal and actively constructive.

Keywords Multisensory processing · Perceptual organization · Scene perception

There is strong evolutionary pressure to have a functional representation of the world around you; it is important to know what dangers are near and what food can be reached. In short, successful navigation of the environment calls for knowing what is in the scene in front of you and around you. Our sensory and perceptual processes provide the data needed for such scene analysis. Given the particularly detailed information that the human visual system provides, visual input is generally regarded as central to scene analysis. However, even in the 19th century, William James (1890) recognized that analyzing a scene is extremely complex and draws on multiple senses. He argued that "experience, from the very first, presents us with concretized objects, vaguely continuous with the rest of the world which envelops them in space and time, and potentially divisible into inward elements and parts. These objects we break asunder and reunite" (p. 487). In fact, this was the context for his well-known statement about a baby's experience of the world. Speaking of the "impressions" that support perception of an object, and the sensory systems that provide them, James said, "Although they [impressions] separate easier if they come in through distinct nerves, yet distinct nerves are not an unconditional ground of their discrimination. . . . The baby, assailed by eyes, ears, nose, skin and entrails at once, feels it all as one great blooming, buzzing confusion" (p. 488). Thus, in this view, multisensory perception is actually the starting point, with experience needed to develop unimodal object representations. In fact, not only is multisensory processing early in development of the individual, Stein et al. (1996) assert that it is also very early in evolutionary terms.

Ample previous research shows that the senses do work together to optimize perception. For example, Newell, Ernst, Tjan, and Bülthoff (2001) found that in the visual modality, observers recognize objects from the front view best, while in the haptic modality, they recognize objects from the back view best. This difference is grounded in the fact that the hand picks

✉ Arthur G. Samuel
Arthur.Samuel@stonybrook.edu

[1] Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500, USA

[2] Ikerbasque, Bilbao, Spain

[3] Basque Center on Cognition, Brain, and Language, Gipuzkoa, Spain

[4] Stony Brook University, Stony Brook, NY, USA

objects up from the back more often than from the front (Ernst & Bülthoff, 2004). Because of this, when both the haptic and visual senses are used, more detailed and complete information can be gathered than when they are individually used.

There are also many demonstrations of a similarly symbiotic relationship between vision and audition. When auditory and visual cues are correlated, as they often are in the environment, the brain generally assumes a common underlying cause (Parise, Spence, & Ernst, 2012). As a result, audiovisual integration can provide people with more accurate and detailed information than each of these senses individually (Bulkin & Groh, 2006). A nice example of this was reported by Sekuler, Sekuler, and Lau (1997). If a visual display shows two identical disks moving toward each other, merging, and then separating along the original motion axis, there are two quite different perceptual interpretations: The postmerging motion could be seen as each disk continuing through the other, or it could be seen as each disk bouncing back along its original path. Sekuler et al. showed that if a noise was timed to occur at the moment of "contact," observers were much more likely to report that they saw the disks bounce apart. The sound of contact supported the bouncing percept more than the passing-through percept, and perception was driven in that direction.

In many situations, observers may be trying to anticipate when an approaching object (e.g., an oncoming car) will reach them. Information is typically available through both vision and audition for this judgment, and a number of studies have shown that observers can combine the information from the two modalities. Gordon and Rosenblum (2005) investigated the flexibility of this form of multisensory scene analysis by asking observers to make "time to arrival" judgments about an approaching car, based on different types of visual and auditory information that were provided. The authors presented observers with different combinations of the two types of information—for example, by interrupting one source of information while the other continued. They found that observers were essentially as accurate under these conditions as they were when intact information was available, suggesting that the input from the two modalities could be seamlessly combined. They suggested that this result could be a consequence of observers computing "modality neutral information," as would occur if the perceptual system is tracking cues from a common event, regardless of modality.

There are many other demonstrations of audiovisual perception. Perhaps the best known example is the McGurk effect. This occurs when the audio component of one speech sound is presented in conjunction with the visual component of another; the combination can lead to the perception of a third sound. For example, if a video showing a person's mouth saying "ga" is dubbed with the sound "ba," observers often report that the person is saying "da" (McGurk & MacDonald, 1976). This is a kind of "compromise" percept in which the perceived place of articulation is between the lips that are signaled by the sound, and the back of the mouth that is implied by the video. In some cases, the auditory component of a multimodal signal can actually dominate the visual component. For example, when Shams, Kamitani, and Shimojo (2000) presented observers with a single visual flash in conjunction with multiple auditory beeps, participants integrated the multimodal input by consistently perceiving multiple flashes, even though there was just a single flash.

In the current study, we report a series of experiments in which we investigate another situation in which the observer's perception of a visual scene is influenced by certain patterns of auditory input. Our focus is on object perception, and the extent to which perception of visual objects may be influenced by sounds that are presented as part of the multimodal scene. The experiments are grounded in a study by Hidaka, Teramoto, Gyoba, and Suzuki (2010). Their study, in turn, relied on a purely visual effect reported by Moore, Mordkoff, and Enns (2007). Thus, to situate our study, we will first summarize the visual effect reported by Moore et al., and then discuss how Hidaka et al. established an audiovisual extension of the effect.

Moore et al. (2007) were interested in how certain sudden visual changes to an object are interpreted by the processes that support scene analysis. Their theoretical framework was based on the idea of "object files" (Kahneman, Treisman, & Gibbs, 1992). In this framework, the observer represents the current scene in terms of spatially coded object files, with each such file made up of various features associated with a given object (e.g., its color, shape, texture). As objects move, their object files are associated with changing locations. If a property of an object changes a bit (e.g., it darkens under a shadow, or perhaps changes shape a bit due to its motion), the features in the object file are updated. Moore et al. sought to test what happens if the change to an object is more extreme than is likely to happen to a real object in the world. For example, on a computer screen, an object's size can instantly change dramatically. They hypothesized (see also Moore & Enns, 2004) that if the change is too great, a new object file must be created because no single object would be consistent with two such different sizes. In a clever set of experiments, Moore et al. presented a sequence of 80-ms frames in which a disk shifted positions, creating a pattern of apparent motion. They found that observers tended to see two objects (a large and a small disk) rather than one if they introduced a size change that would be very unlikely for a real object, such as the large moving disk suddenly getting much smaller for one 80-ms frame. Interestingly, if the perceptual system was given an "explanation" for such a major size change, such as making it appear as though a disk was passing behind a solid surface with a smaller circular hole in it, observers did not perceive a second object.

Hidaka et al. (2010) hypothesized that the perception of a second disk would be more likely if the sudden size change coincided with a change in an auditory aspect of the scene. In particular, a briefly presented smaller version of the disk might persist perceptually for a longer time if a sound change coincided with the sudden size change, with the constellation of cues attributable to some event in the environment. There was prior evidence for an effect of this sort in a study by Vroomen and de Gelder (2004). Vroomen and de Gelder had shown observers a series of pseudorandom dot patterns, with each flashed dot pattern presented together with a tone. When the frequency of a tone was changed (e.g., a high tone presented after a series of low tones), the corresponding dot pattern seemed to persist longer than dot patterns before or after it, even though the actual durations were matched.

Hidaka et al. (2010) took the sudden size change manipulation of Moore et al. (2007) and introduced the event-coincident sound change of Vroomen and de Gelder (2004). Each 80-ms frame showing a large disk was accompanied by a tone. As the disk followed a circular path around a computer screen, in the penultimate frame the disk was replaced by a smaller version, and then returned to its original size in the final frame; on control trials, both the large and small disks remained on the screen during the final frame. Observers were told to report whether they saw one or two disks in the final frame. The key result was that there were significantly more reports of seeing two disks if the penultimate frame also included a switch in the frequency of the accompanying tone than if there was no such frequency shift (or if there was no accompanying sound at all). Thus, as in Vroomen and de Gelder’s study, changing the pitch of a tone extended the apparent visible persistence of an accompanying visual object. Hidaka et al. demonstrated that the effect was not related to the “synesthetic congruency” effect in which smaller objects are associated with higher-pitched sounds (e.g., Gallace & Spence, 2006; Parise & Spence, 2009), as the likelihood of seeing a second disk was no greater when the tone that coincided with the small disk was high pitched rather than low pitched (see Keetels & Vroomen, 2011, for similar limitations on synesthetic congruency).

In the current study, we report four experiments that are based on Hidaka et al.’s (2010) paradigm, with the goal of better understanding the conditions under which auditory changes can influence visual percepts. The Hidaka et al. study provided some useful additional information about this. In a pair of follow-up experiments, they showed that neither a simple onset (a tone playing only during the presentation of the small disk) nor a simple offset (a constant tone played during all frames except that of the small disk) induced the percept of a longer-lasting second disk. In a final experiment, they showed that if the tone sequence was “captured” by a different part of the display (cf. Bregman, 1990), the effect also was blocked. Collectively, the results from Hidaka et al. suggest that observers were constructing a unified audiovisual scene, with the tones/frames that preceded the critical changes providing an essential context for interpreting sudden changes in the visual and auditory cues.

In our first experiment, we replicate the conditions of Hidaka et al.’s (2010) study to see how robust the basic finding is. The other three experiments are designed to provide important details about the properties of this type of audiovisual integration. Two of the experiments focus on the relative timing of the visual and auditory changes: How synchronous must these be to drive visible persistence? A very tight temporal link might be expected if the sensory information is coordinated at a low level of processing, whereas a wider integration window could be found for higher-level integration. For example, in the McGurk effect, there is a surprisingly wide temporal window (about 200 ms, with substantially more tolerance for the visual information preceding the auditory than vice versa) that still produces the effect (van Wassenhove, Grant, & Poeppel, 2007). In Experiment 2, to test how time linked the multimodal effect is, we associate tones with visual frames at different points during the motion of the disk around the screen. Experiment 3 continues our investigation of the temporal properties by introducing frequency changes just before or just after the visual change. Finally, in Experiment 4, we test whether a frequency change is treated the same way, in this domain, if it comes after a sequence of such changes. This experiment tests whether performance is dominated by the physical pattern given to the observer, or by the perceptual organization that the observer imposes on the physical signal. The overarching goal is to better understand how the perceptual system coordinates visual and auditory information in constructing its analysis of the scene.

Experiment 1

Experiment 1 replicates the conditions of the first experiment reported by Hidaka et al. (2010). We test whether audio frequency changes, in conjunction with visual object movements, can affect visible persistence—the perceptual persistence of a visual object after it has disappeared from the screen. Specifically, we test whether an abrupt change in sound frequency in the penultimate frame of a trial causes visible persistence, when a size change in the visual presentation coincides with the sound change.

Method

Participants

In the original study of this phenomenon (Hidaka et al., 2010, Experiment 1), 12 participants were tested. In Experiment 1 and in each of the following experiments, we tested 20 participants. All of the participants in this experiment, and in the following experiments, were undergraduate students from Stony Brook University who were fulfilling a research participation requirement. All were naïve to the purpose of the experiment, read and signed a consent form, and were debriefed after the experiment. All participants had self-reported normal or corrected-to-normal vision and hearing. The experiments were approved by the Institutional Review Board (IRB) of Stony Brook University in New York.

Apparatus

Participants were tested individually in a small, quiet room. As in the original study, the visual stimuli were presented on a CRT monitor (Viewsonic PS775, operating at 85 Hz), and the auditory stimuli were presented binaurally over high quality closed-ear headphones (Sony MDR-V900). Participants sat approximately 60 cm from the screen, and responded by pressing either "Z" or "M" on a standard computer keyboard, with "Z" indicating that one disk was seen at the end of the trial and "M" indicating that two disks were seen then.

Stimuli

On each trial a white disk (≈1.6 cm, 1.55°) moved around a blue fixation point (≈.56 cm, 0.54°), shifting 15° every 80 ms against a gray background. The diameter of the circle that the white disk traversed was ≈14.9 cm, 14.45°. Pure tones of 600 Hz and 3000 Hz were used as the low and high sounds presented over headphones. Each tone was 50-ms long, including onset and offset amplitude ramps to prevent clicks. The onset of each 50-ms tone occurred at the onset of the 80-ms display of a circle.

Procedure

There were a total of 288 trials in the session. A third (96) of the trials had no sound, a third had a tone that did not change in frequency, and a third had a change in the tone's frequency during the penultimate frame. For the constant audio frequency condition, 48 trials were "low tone" trials (600 Hz; Low; L) and 48 were "high tone" trials (3000 Hz; High; H). For the changing tone condition, 48 had a main tone of 600 Hz, with only the penultimate frame having a 3000 Hz tone; 48 had a main tone of 3000 Hz with only the penultimate frame having a 600 Hz sound. Thus, the tones during the last three frames of half of the changing audio trials were LHL and the other half were HLH.

For the visual stimuli, three factors were varied. The first of these involved the number of disks present in the last frame of a trial. On half of the trials, one disk was present in the last frame, while for the other half, two disks were present. The second factor was the presence/absence of a visual size change. On the no visual size change trials, the size of the disk was constant throughout the entire sequence. On visual size change trials, the size of the disk in the penultimate frame decreased in size (≈.45 cm, 0.44°) and the disk in the last frame went back to the original frame size.

The starting location of the white disk was equally likely to be at 0°, 90°, 180°, or 270°, with 0° referring to the top of the screen. The direction of motion (clockwise or counterclockwise) was also evenly distributed across the 288 trials. The last visual factor was the frame length of the trajectory. There were 13, 19, or 25 frames in each trial. Each frame was presented for 80 ms, and each frame was 15° apart on the trajectory for a total of 195°, 285°, or 375°. The initial position, direction of motion, and length of trajectory were all random on a given trial and counterbalanced across the 288 trials.

Each trial began with a blue dot at the center of the screen for 1,000 ms. Then, the white disk appeared and started moving. At the end of the trial, the participants reported the number of disks that they saw (1 or 2) at the end of the sequence by pushing one of two buttons on the keyboard. The session began with 24 practice trials without any sound. The 24 trials represented the crossing of visual size change or lack of visual size change (2) × the number of disks in the last frame (2) × the frame length of each sequence (3) × repetitions (2). The order of the practice trials was random, and error feedback was provided. After the participants finished the practice trials, they began the 288 trial session in which trials were presented without error feedback.

Results and discussion

As described above, there were 144 experimental trials, trials in which there was a visual change in the penultimate frame (the disk got smaller for 80 ms). In addition, there were 144 control trials, trials in which there was no such visual change. The purpose of Experiment 1 is to determine whether the experimental trials replicate the pattern reported by Hidaka et al. (2010): Does a pitch change that co-occurs with the visual change cause the smaller circle to persist visually, leading to more report of seeing two circles at the end of the trial? Thus, our analyses will focus on whether the different sound conditions led to different outcomes on the experimental trials. Specifically, we will test whether trials with a pitch change during the penultimate trial led to higher rates of reporting two circles, compared to the rates for trials with no sound, and to trials with a constant pitch. For completeness, and to be sure that our results are not due to participants guessing in a particular way, we will report comparable analyses for the control trials. Because there was in fact no visual change during the control trials, we should observe few, if any, reports of seeing two circles when only one was present, and there should not be any influence of the differing sound conditions.

We first divided each participant's data into the 144 experimental trials and the 144 control trials. Within each of these two sets, we sorted the trials on the basis of the three sound conditions. In each of these six cases, we calculated two error rates: trials in which there were actually two disks in the final frame but the participant reported one ("actually 2, guessed 1"), and trials in which there was actually one disk in the final frame but the participant reported two ("actually 1, guessed 2"). The "actually 1, guessed 2" errors are the critical results, as they potentially index visual persistence. The "actually 2, guessed 1" errors provide a baseline for each participant, as participants might vary in their overall tendency to make errors. Therefore, the number of "actually 2, guessed 1" trials was subtracted from the number of "actually 1, guessed 2" trials for each of the audiovisual combinations for each participant.

Conceptually, this difference score is similar to signal detection's d' measure. For our data set, computing d' scores would be inappropriate because of the pattern of the two types of errors.[1] Specifically, our data were clearly unsuitable for d' computations in two related ways. The model underlying d' computations assumes that both error distributions are Gaussian (and that the two distributions have equal variance). The actual error distributions were about as non-Gaussian as one could find: In each case, the peak error rate was near zero, and the incidence of errors declined as the error rate increased. These distributions are also problematic for d' computations because most of the error rates were at or near zero, exactly the portion of the normal distribution where small actual error rate differences produce extreme z-score differences. By using the difference score measure that we employed, these issues do not arise. In fact, the distribution of the difference scores was almost perfectly Gaussian. These difference scores provide a very functional measure of visible persistence, correcting for any overall error tendency (captured by the "actually 2, guessed 1" errors). A positive difference score should be found when the conditions promote visible persistence.

In Experiment 1, and in the subsequent experiments, to identify any outlier participants, we looked at performance in the three sound conditions of the experimental part of the design and the three sound conditions of the control part of the design. If a participant's score was more than 2.5 standard deviations from the mean, in at least two of these six cases, the participant was considered to be an outlier and was not included in the data analyses. In Experiment 1, two of the participants exceeded this threshold, leaving data from 18 usable participants. The data for the experimental conditions for these 18 participants was submitted to a single-factor analysis of variance (ANOVA), with the within-subject factor of auditory condition (no audio, constant audio, or auditory change in the penultimate frame). The left panel of Fig. 1. shows the mean difference scores for the three auditory conditions for the experimental trials.

As the left panel of Fig. 1 shows, participants were more likely to report that two circles were present on trials that included a pitch change that coincided with the appearance of the smaller circle. The difference among the three conditions was significant, $F(2, 34) = 4.834$, $p = .014$ ($\omega^2 = .124$). Recall that the specific question was whether the auditory change condition would produce stronger visible persistence than the other two sound conditions. In fact, it did—the difference scores were significantly higher in the change condition than when there was no sound, $t(17) = 2.787$, $p = .013$, or when there was a constant tone frequency, $t(17) = 2.679$, $p = .016$.

The right panel of Fig. 1 displays the results for the control trials, trials that did not include a brief appearance of a smaller circle. The results are exactly as they should be if participants were making the judgments as they should be. In an ANOVA comparable to that for the experimental conditions, the difference scores overall did not differ from zero, $F(1, 17) < 1$, and there was no difference among the three sound conditions, $F(2, 34) < 1$.

In sum, Experiment 1 provides a strong replication of both Moore et al. (2007) and of Hidaka et al. (2010). The key finding by Moore et al. was that a sudden visual change was likely to cause the creation of a new object file, leading the reports of seeing two disks when in fact there really was only one. The robust effects shown in the left panel of Fig. 1 demonstrate this effect. We also found, as Hidaka et al. (2010) reported, that a concurrent auditory change increased report of multiple objects. The overall similarity between our results and those of the previous studies was sufficient to move forward with further tests, beginning with an investigation of the role of when an audio change occurs relative to a visual change.

Experiment 2

In Experiment 1, as in Hidaka et al. (2010), the timing of the auditory change was carefully matched to the timing of the visual change. While this seems sensible, we do not actually

[1] There are a number of reasons to believe that our results are not being driven by a response bias. Experiment 1 is a replication of Hidaka et al.'s (2010) first experiment, and our difference scores replicate the pattern of d' scores that they reported. If their d' measure accounted for response bias, and our results match theirs, this suggests that our data are not due to response bias. Moreover, in Experiments 1–3 of the current study, we use half of the data to look for any hint of subjects guessing that a second circle was presented on trials that did not include visual changes before the final frame. As Figs. 1, 2, and 3 (right panels) show, there is absolutely no hint of any such guessing—difference scores on these control trials were always almost exactly zero and showed no sign whatsoever of being affected by the auditory conditions. Given the strikingly clean performance on these control trials, an account that invokes response bias would need to assume that the response bias only occurred on the experimental trials and, in addition, that this response bias was systematically affected by the sound conditions.

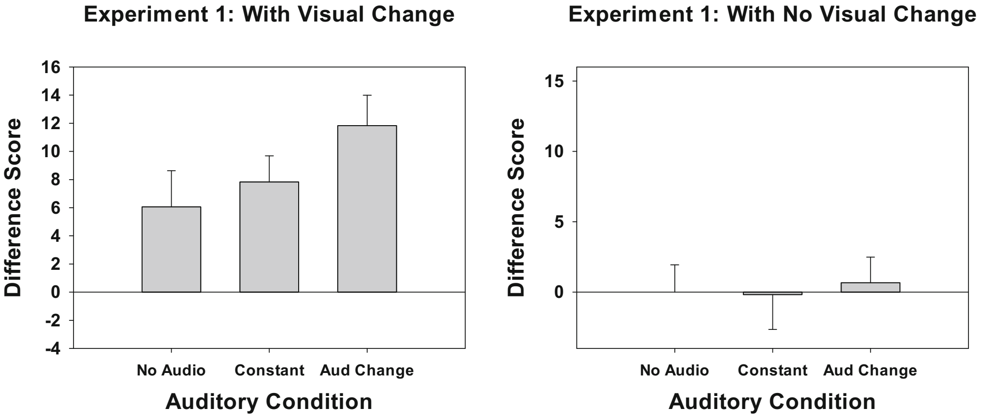

Fig. 1 Results of Experiment 1. Average difference scores for the three auditory conditions for trials that had a visual change (left panel) and for trials that did not include a visual change (right panel). Error bars show standard errors

know how important such a temporal coincidence is. As was noted in the introduction, the well-known audiovisual McGurk effect tolerates a surprisingly large asynchrony between the visual and auditory information (van Wassenhove et al., 2007). Experiment 2 tests whether the auditory and visual changes that drive visible persistence really need to coincide tightly or not. For this test, rather than have a single 80-ms aberrant tone appear in the same frame as the visual change, a tone is introduced two frames before the visual change and accompanies the disk during each frame for the remainder of the trial (i.e., for a total of four frames, 320 ms). Thus, a salient sound change occurs roughly at the same time as the visual change but is intentionally decoupled from it. As a control, there is a condition in which a tone is presented during the first four frames of a trial (320 ms) but then goes silent, relatively far in time from the visual change. Experiment 2 thus tests whether the perceptual system is trying to build a fully coherent multisensory representation or whether the occurrence of change (in any of the sensory systems) is treated as a cue that something needs to be changed in the developing representation, without necessarily converging on a single unified representation.

Method

Participants

There were 20 participants in Experiment 2, drawn from the same population as in Experiment 1. None had participated in the previous experiment.

Apparatus

The apparatus used in this experiment was the same as the apparatus used in Experiment 1.

Stimuli and procedure

All of the visual stimuli were identical to those in Experiment 1. As in Experiment 1, one third (96) of the trials had no audio. Two of the auditory conditions, however, were changed. The 96 trials of the First-4 audio condition had a constant 600 Hz or 3000 Hz tone present only in the first four frames (320 ms) of each trial. In the Last-4 audio condition, a constant 600 Hz or 3000 Hz tone was presented during only the last four frames (320 ms) of a trial. In this case, the first tone onset precedes the visual change by 160 ms and continues beyond the small circle’s disappearance for 80 ms. As in the first experiment, whether the 600 Hz or 3000 Hz tone was presented on a given trial was random; each frequency was presented equally often in each condition.

After completing the same set of practice trials as in Experiment 1, participants completed a 288-trial main session with all the conditions randomized and counterbalanced as before.

Results and discussion

Once again, the number of “actually 2, guessed 1” errors was subtracted from the number of “actually 1, guessed 2” errors for each audiovisual condition. As in Experiment 1, outlier participants were identified as those with scores more than 2.5 standard deviations from the mean in at least two conditions. Two participants were eliminated on this basis. The left panel of Fig. 2 presents the average difference scores for the remaining 18 participants, broken down by the nature of the auditory information that accompanied visual displays that included a visual change in the penultimate frame. The right panel shows the comparable data for the control trials—those that did not have a visual change in the penultimate frame.

The central question of Experiment 2 is whether an auditory change in the general time range of the visual change, but not tightly time linked to it, is sufficient to support the

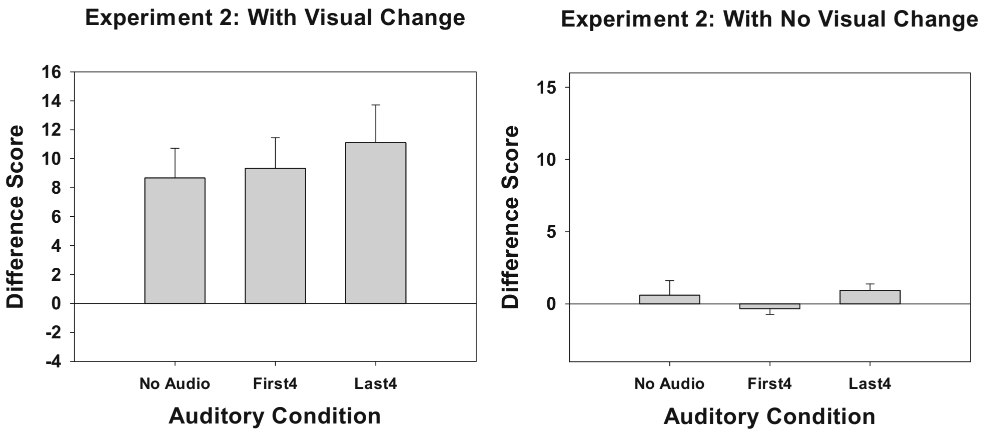

Fig. 2 Results of Experiment 2. Average difference scores for the three auditory conditions for trials that had a visual change (left panel) and for trials that did not include a visual change (right panel). Error bars show standard errors

extended visual perception of the second disk. If so, then the Last-4 condition should produce stronger visible persistence than either the no-sound or the First-4 cases. In a single factor within-subjects ANOVA, the effect of sound condition was marginally significant, $F(2, 34) = 2.676$, $p = .083$ ($\omega^2 = .058$). Similarly, the difference between the Last-4 case and the no-sound case was marginally significant, $t(17) = 1.908$, $p = .074$, as was the difference between the Last-4 and First-4 conditions, $t(17) = 1.798$, $p = .090$.

The right panel of Fig. 2 shows the results for the control trials. As expected, these trials did not produce visible persistence—the difference scores did not significantly diverge from zero, $F(1, 17) < 1$. There was no difference in these scores as a function of auditory condition, $F(2, 34) = 1.080$, $p = .351$ ($\omega^2 = .003$).

The results of Experiment 2 suggest that a late-occurring sound change can support visible persistence, even when the sound change occurs 160 ms before the visual change. However, because this difference was only marginally significant overall, and the individual comparisons were similarly marginal, we cannot yet draw this conclusion with confidence. To provide clarity on this issue, Experiment 4 will include a replication of the comparison between the First-4 and Last-4 conditions. To foreshadow, the results of that comparison will confirm the preliminary conclusion.

The weaker effect found for the Last-4 case in Experiment 2 (an effect size of .058) than the auditory change case of Experiment 1 (an effect size of .124) is consistent with the results of two of the experiments in Hidaka et al. (2010). Recall that they found a robust effect for a sound change in the penultimate frame (our auditory change in Experiment 1), but no effect when a single tone was presented during the penultimate frame, with no preceding or following tones. Our Last-4 condition is something of a compromise between their successful and their unsuccessful cases: There is no sound change, but there are multiple presentations of the tone (four presentations, one in each of the final four frames). The intermediate results for this situation suggest that the perceptual system treats the multiple presentations as an event sufficient to warrant some integration with the visual change, but not as compelling as a major change in tone frequency.

The results in hand suggest that some time locking of sound changes with visual changes may be necessary to drive visible persistence. Experiment 3 focuses on auditory changes that occur during the last 240 ms of each trial, to test whether more precise time locking of the auditory and visual changes will affect the strength of the visible persistence. This window is similar to the 200-ms window of asynchrony that can be tolerated in producing the McGurk effect.

Experiment 3

The results of Hidaka et al. (2010), together with the results of our first two experiments, have shown the strongest effects on visual persistence when the sound change co-occurs with the visual change. In Experiment 3, we compare the effect of such a sound change occurring in the penultimate frame to the same sound change coming one frame (80 ms) earlier, or one frame (80 ms) later. This provides a test of the extent to which the visual and auditory events need to be tightly time locked. If the visible persistence is being driven by low-level sensory integration of the two events, then such time locking is more likely to be necessary.

Method

Participants

Experiment 3 included 20 participants from the same population as before. None had participated in either of the previous experiments.

Apparatus

The apparatus for this experiment was the same as in Experiments 1 and 2.

Stimuli and procedure

As in Experiment 2, only the auditory conditions changed in this experiment—the visual stimuli remained the same as before. On all trials, tones were presented only during the last four frames, and a single frame included a tone that differed in frequency from the tone presented during the other three frames. On one third of the trials, the different tone was presented during the penultimate frame, coinciding with the visual change (on the trials that included a visual change). On one third of the trials, the aberrant tone was presented during the antepenultimate frame—80 ms before the visual change. On the remaining third of the trials, the changed tone was presented during the final frame, 80 ms after the visual change. The same low tone (600 Hz) and high tone (3000 Hz) were used, and each tone was used equally often in each condition. This produced six possible patterns, two for the antepenultimate case (LHLL or HLHH), two for the penultimate case (LLHL or HHLH), and two for the ultimate case (LLLH or HHHL).

After completing a 24-trial practice session identical to the one in previous experiments, participants completed a 288-trial main session with the conditions counterbalanced as they were in the previous experiments.

Results and discussion

The same difference scores were computed as in the previous experiments, for each combination of visual change and aberrant tone timing. The same procedure for identifying outlier participants was used as before. No participants were identified as outliers, leaving all 20 in the statistical analyses. The left side Fig. 3 presents the average difference scores for the experimental trials, and the right side shows the corresponding data for the control trials.

As Fig. 3 shows, the results for the three locations of sound change were extremely similar. A single-factor within-subjects ANOVA found no differences among the antepenultimate, penultimate, and ultimate locations, $F(2, 38) = 1.043$, $p = .362$ ($\omega^2 = .001$). This similarity was reflected in the planned comparisons between the sound change's occurrence in the penultimate frame and the frame preceding it, $t(19) = 0.417$, $p = .681$, or following it, $t(19) = 1.041$, $p = .311$. Clearly, the process that leads to extended visible persistence tolerates an asynchrony between the visual and auditory events across the 240-ms window tested in Experiment 3.

The results for the control trials, shown in the right panel of Fig. 3, look essentially like those seen in the first two experiments. The only difference is that because of exceptionally little variance, the average difference score of 0.67 was significantly different from zero, $F(1, 19) = 7.835$, $p = .011$. As expected, the sound condition had no effect, $F(2, 38) < 1$.

Experiment 3, like Experiment 2, supports the conclusion that a sound change that happens roughly at the same time as a visual change can affect visible persistence. In both experiments the results indicate that the cross-modal interaction does not depend on precise temporal alignment. In the final experiment, we examine whether linking the auditory and visual changes over a longer time span will affect the likelihood of observing an impact of a sound change on visible persistence.

Experiment 4

Hidaka et al. (2010) included an experiment in which they temporally aligned sound changes with changes made to a central fixation circle rather than to the disk that was moving around the screen. That manipulation succeeded in causing the perceptual system to group the sound with the central point

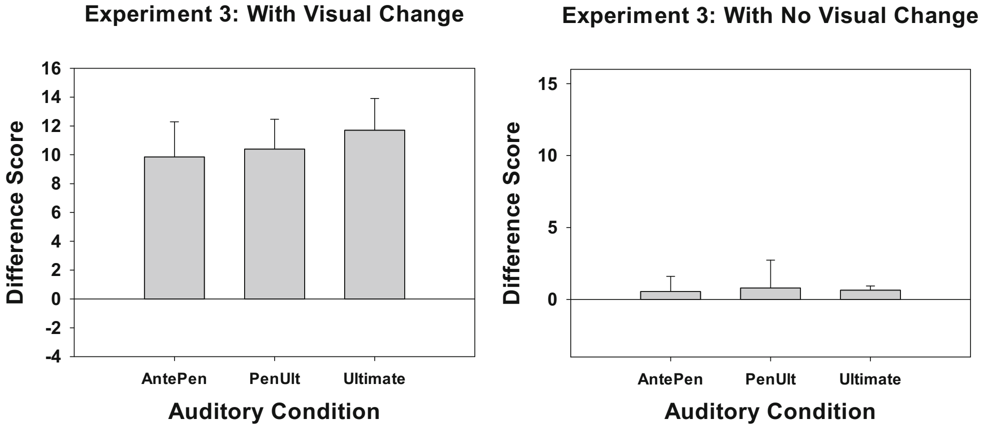

Fig. 3 Results of Experiment 3. Average difference scores for the three auditory conditions for trials that had a visual change (left panel) and for trials that did not include a visual change (right panel). Error bars show standard errors

rather than with the moving disk, thus blocking the effect the sound change had produced on visible persistence of the disk(s). Our final experiment uses a complementary approach, linking changes in tone frequency to changes in the moving disk(s), to see whether such a link affects visible persistence.

Method

Participants

In Experiment 4, 20 participants from the same population as in the other experiments were tested. None had participated in any of the previous experiments.

Apparatus

The apparatus for this experiment was the same as that in the previous experiments.

Stimuli and procedure

As in the previous experiments, there were two visual conditions and three auditory conditions. One of the two visual conditions remained the same as previous experiments—the "visual change" condition in which a disk was a constant size, except in the penultimate frame, when it became smaller. The other visual condition—"alternating visual"—was new. As its name suggests, in this condition, the disk alternated in size from frame to frame. These trials always began with the larger disk (≈1.6 cm, 1.55°), followed by the smaller one (≈.45 cm, 0.44°), then back to the larger one, and so on. Because there was always an odd number of frames, this guaranteed that the last three frames would always include the large disk, the small disk, and, finally, the large disk. Note that this is the same pattern for the last three frames in the visual-change condition, but that in the alternating-visual condition the occurrence of the small disk in the penultimate frame is not a sudden change—it is a continuation of the alternation pattern.

Two of the three audio conditions were identical to two of the audio conditions in Experiment 2. In the First-4 condition, one tone was presented only during the first four frames of each trial; in the Last-4 condition, one tone was presented during only the last four frames of each trial. In both cases, half of the trials used the low tone (600 Hz) and half used the high tone (3000 Hz). In the third audio condition ("alternating audio"), tones were presented throughout the sequence and alternated in frequency. In this condition, the frequency (600 Hz or 3000 Hz) presented during the first frame was randomly selected so that half of the trials were HLHLH . . . and half were LHLHL. . . ... The results of Experiment 2 showed that the Last-4 pattern supported visible persistence more than the First-4 pattern. The new alternating-audio pattern contains frequency changes during the time frame of visual changes (like the Last-4 case), but like the alternating-visual condition, the change is the continuation of a pattern rather than a sudden new occurrence.

Results and discussion

The same difference scores were computed as in the previous experiments. The same procedure for identifying outlier participants was used as before. One participant was identified as an outlier, leaving 19 in the statistical analyses. The left side Fig. 4 presents the average difference scores for the visual-change conditions, and the right side shows the corresponding data for the alternating-visual conditions.

Figure 4 gives a different initial impression than the first three figures because the right panel represents a very different situation than in those cases. In the previous experiments, half of the trials served as a control condition because they included no visual change, and those conditions consistently produced tight distributions of difference scores clustered around

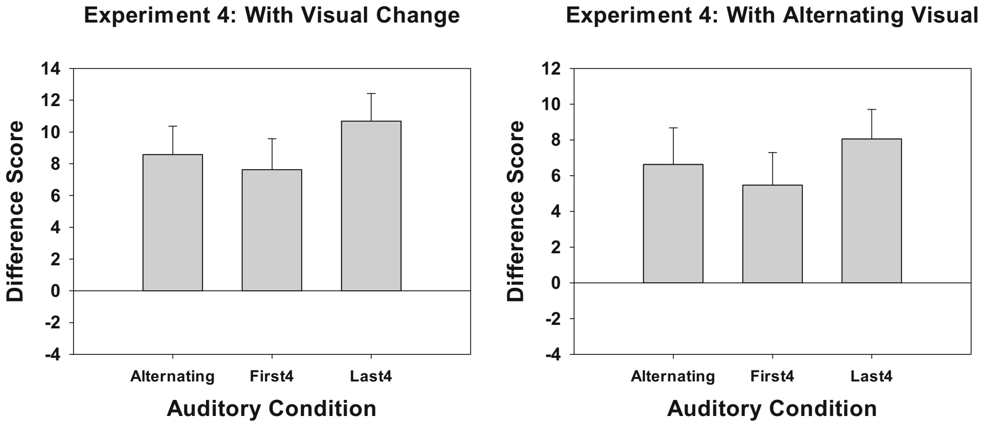

Fig. 4 Results of Experiment 4. Average difference scores for the three auditory conditions for trials that had a visual change in the penultimate frame (left panel) and for trials in which the visual display alternated throughout the trial (right panel). Error bars show standard errors

zero. In Experiment 4, after three demonstrations that with no visual change there is no extended visible persistence, we replaced this control condition with a second experimental condition—the alternating-visual condition. This new condition tests whether linking a visual alternation to an auditory alternation reduces the perceptual system's likelihood of creating a new object file for the smaller disk in the penultimate frame.

Looking first at the visual conditions that were similar to those tested in Experiment 2 (visual change; left panel of Fig. 4), we see results that are similar to those in Experiment 2. In a single-factor within-subjects ANOVA, the effect of sound condition was significant, $F(2, 36) = 3.756$, $p = .033$ ($\omega^2 = .088$). Recall that in Experiment 2, the Last-4 auditory condition produced higher difference scores than the First-4 condition did, but that this difference was only marginally significant. In the current experiment, this difference was significant, $t(18) = 2.196$, $p = .041$. As the left panel of Fig. 4 shows, having the auditory component alternate between high and low tones did not support extended visible persistence despite the occurrence of a sound change in the general time window of a visual change. The Last-4 auditory condition produced a marginally larger average difference score than the alternating-auditory condition, $t(18) = 1.861$, $p = .079$.

Turning to the alternating-visual conditions (right panel of Fig. 4), we see that difference scores were a little lower overall than in visual-change conditions, but this reduction was not significant, $F(1, 18) = 1.104$, $p = .307$. The somewhat dampened effects in the visual-alternation conditions led to weaker differences among the auditory conditions, with the main effect of auditory condition not reaching significance, $F(2, 36) = 2.398$, $p = .105$ ($\omega^2 = .047$). Despite this dampened overall effect, the Last-4 case produced significantly higher difference scores than the First-4 case did, $t(18) = 2.868$, $p = .010$. The stronger effect of the Last-4 case than the alternating sound was not significant, $t(18) = 1.472$, $p = .158$.

The results of Experiment 4 reinforce a previously preliminary finding. We can now say conclusively that a sound change needs to be in the same general time range as a visual change to reinforce visible persistence. Across Experiment 2 and Experiment 4, there are now multiple demonstrations of the stronger effect of the Last-4 case over the First-4 case. A sound change within the last 320 ms of the trial, coupled with a visual change 160 ms before the end of the trial, leads to greater visible persistence than a sound change during the first 320 ms of the trial. Collectively, Experiments 2, 3, and 4 show that there is some tolerance for audiovisual asynchrony, but this tolerance is limited. Differences within the 320 ms window we have tested are tolerated, but the longer mismatch in the First-4 condition significantly reduces visible persistence.

Experiment 4 provides a critical new finding: The cross-modal effect is driven by the perceived change in a sound pattern, not by the physical change. This conclusion is supported by the stronger visible persistence in the Last-4 sound condition than in the alternating-sound condition. Collapsing across the visual-change and the alternating-visual halves of Experiment 4, the Last-4 condition produced more visible persistence than the alternating-auditory condition, $F(1, 18) = 5.070$, $p = .037$ ($\omega^2 = .047$). The alternating-auditory condition, like the consistently effective penultimate condition in Experiment 1 and in Hidaka et al. (2010), has a tone change during the last three frames (either HLH or LHL). However, in the alternating-auditory case, this tone change is a continuation of the alternation that starts at the beginning of the trial, and the perceptual system thus does not treat it like the abrupt change in the penultimate condition (or, the Last-4 condition in Experiments 2 and 4). As a result, the HLH or LHL pattern does not contribute to visible persistence in the former case.

It is worth noting that there is a limit to the dominance of perceptual grouping over physical patterning. If perceptual patterning always dominated, then one might expect that there would be little visible persistence in the alternating-visual condition because the sudden change to a smaller disk in the penultimate frame could be discounted as merely the continuation of the alternating-visual pattern. As the right panel of Fig. 4 shows, the alternating-visual condition systematically led to reporting two disks. It is true that this tendency was weaker than in the visual-change case, but not significantly so. Of course, due to the alternation pattern, there could be a stronger representation of the small disk due to its being presented repeatedly during the alternation.

General discussion

The four experiments of the current study were designed in the context of two broad areas of research—perceptual scene analysis and multimodal perception. The scene-analysis literature includes two bodies of work that have traditionally proceeded separately, with one literature focusing on how observers parse the visual scene (e.g., Kahneman et al., 1992; Levin & Simons, 1997; Rensink, 2000), and another literature focusing on the perceptual organization of the auditory scene (e.g., Bregman, 1990; Gregg & Samuel, 2008; Kubovy & Van Valkenburg, 2001; Shinn-Cunningham, Lee, & Oxenham, 2007). Research on multimodal perception demonstrates that investigating either vision or audition separately runs the risk of missing important cross-modal effects (e.g., Bulkin & Groh, 2006; Ernst & Bülthoff, 2004; Vroomen & de Gelder, 2004).

The specific context for the current study is a property of visual scene analysis described by Moore and his colleagues (Moore & Enns, 2004; Moore et al., 2007), and then brought into the domain of multimodal scene analysis by Hidaka et al. (2010). In the visual domain, an unnatural size change led to the apparent persistence of two objects, rather than only one (Moore et al., 2007). Cross-modally, an auditory change that was presented concurrently with the visual size change strengthened the extended visible persistence of the second object (Hidaka et al., 2010). Moore and his colleagues framed their effect in terms of Kahneman et al.'s (1992) theory that the visual scene is represented by a set of object files, where each object file is a spatially indexed collection of visual features. They argued that if there is a feature change that cannot make sense within a single object file (e.g., a sudden size change), a second object file is generated. This additional object file can then support the perception of a second object, the basic phenomenon in Moore et al.'s experiments, and in the experiments of the current study. The central finding of Hidaka et al. was that a concurrent sound change can further strengthen the perception of a second object. The finding raises the possibility that the features making up an object file are not specifically visual.

The results of our four experiments extend our understanding of the phenomena reported by Moore et al. (2007) and by Hidaka et al. (2010). At the simplest level, our results demonstrate that both of these basic effects are reliable—we consistently replicated the strong effect that a sudden size change had on the visible persistence of a second object, and we have multiple demonstrations that the visible persistence is affected by the auditory stream. Beyond confirming these two effects, the two major new findings of the current study involve the relative timing of cross-modal changes and a distinction between physical changes and perceived changes.

Experiments 2 and 3 provide the new information about the necessary timing of a sound change relative to a visual change in order for the sound change to impact visible persistence. Experiment 2 showed that a sound change in the general time frame of the visual change (the Last-4 condition) was more effective in extending visible persistence than the same sound change occurring earlier in time (the First-4 condition); this difference was replicated in Experiment 4. The type of sound change in these experiments was a bit different than the type of sound change in the Hidaka et al. (2010) study. In their study, the frequency of a tone was changed in the same frame as the visual change, whereas in Experiment 2 (and Experiment 4), the change was from no sound to sound (Last 4) or from sound to no sound (First 4). Hidaka et al. actually included a pair of experiments that in a sense are more like what we did, and failed to find an effect: In one experiment, they only presented a tone during the frame in which a visual change occurred (the penultimate frame), and in another experiment, they only omitted the tone during that frame. In neither case did the sound change strengthen the observed visible persistence. This failure is actually a bit surprising. It suggests that their procedure of a single onset or a single offset did not lead the perceptual system to treat the sound change as relevant to the visual objects. It may be that the manipulation here worked because by presenting a tone four times, with each tone coinciding with a visual frame, the connection between the sound and the visual scene was strengthened. The same pairing of tones with visual frames was present in the condition of Hidaka et al.'s study that found an impact of a frequency change on visible persistence. This interpretation is consistent with the view that there is an active process of building a perceptual scene, as will be discussed shortly.

The onset of the sound change in the (effective) Last-4 condition was two frames before the occurrence of the visual change. This 160ms lead time suggested that the process that builds a multimodal scene may be relatively forgiving of timing misalignment, at least under some circumstances. To test this notion, in Experiment 3, we systematically varied whether a single aberrant tone occurred at the same time as the visual change (the penultimate frame), one frame before that, or one frame after that. Consistent with the idea of there being tolerance for some temporal misalignment across vision and audition, all three of these locations for the aberrant tone were still effective in extending visible persistence of the second disk. Thus, the results of Experiments 2 and 3 show that the perceptual scene-building process is sensitive to the relative timing of the auditory and visual events (hence, the consistently stronger effects for the Last-4 timing than for the First-4 timing) but that there is a tolerance for a mismatch of at least 160 ms. A window of this order of magnitude has been observed for the audiovisual integration found in the McGurk effect (e.g., Massaro, Cohen, & Smeele, 1996; Van Wassenhove et al., 2007).

Perhaps the most intriguing finding in the current study was that the same physical sequence of tones (either LHL or HLH) during the final three frames of a trial promoted visible persistence following one preceding pattern of tones but did not do so following a different preceding pattern. Specifically, when the preceding tones all matched the frequency of the antepenultimate (and ultimate) tones, visible persistence was enhanced; when the preceding tones formed an alternating pattern for which the last three tones were simply a completion of the alternation, visible persistence was not enhanced. In the first case, the tone in the penultimate position is distinct from all other tones in the sequence—it is the auditory change that Hidaka et al. (2010) showed to enhance visible persistence. In the second case, even though locally the penultimate tone is distinct from its neighbors, within the sequence as a whole it is merely one more tone in the alternating pattern. Consistent with Bregman's (1990) classic work on auditory scene analysis, the penultimate tone is perceived differently in these two cases. Experiment 4 shows that it is this perception, rather than the local physical pattern, that determines how the cross-modal scene is constructed.

This result fits very nicely with findings both from the original effect reported by Moore et al. (2007) and from Hidaka et al.'s (2010) extension of the phenomenon to a multimodal situation. Recall that in one experiment, Moore et al. presented the same “aberrant” small disk that drives the increase of visible persistence, but did so in a condition in which the larger disk appeared to have passed behind a surface that had a smaller circular cutout. This manipulation exposed observers to the same pattern of a large disk becoming smaller for one (penultimate) frame, but did so in a context that allows the inference that the smaller disk size was not actually a change to the disk—it was merely a consequence of the window through the obstruction. This condition blocked the increase in visible persistence that normally is a consequence of the size change in the penultimate frame. The dependence of this effect on a type of perceptual inference is exactly what one would expect if the effect is a result of actively constructing the perceptual scene. It is also entirely consistent with the “time to arrival” results of Gordon and Rosenblum (2005), who found that observers appear to seamlessly integrate visual cues and auditory cues that stem from the same event—a moving car.

The same kind of cross-modal synthesis is implicated in the experiment in which Hidaka et al. (2010) began each trial with a period in which they coordinated the timing of their tones to changes in a central blue fixation circle rather than to the presentation of the large disk’s movement around the screen. By establishing this connection, they led the observers to group the tone pattern with the central object rather than with the moving disk. As a result, a change in the tone frequency during the penultimate frame did not enhance visible persistence, even though it physically coincided in time with the change to a small disk: The same physical event—a visual change and a tone frequency change co-occurring—led to a different outcome. We thus have recurring findings in this literature that all show that the measure of scene perception (the apparent visible persistence of a small disk) depends on how the observer organizes the multimodal scene rather than on the actual physical properties presented to the eyes and ears. These results are the cross-modal analogs of many demonstrations that Bregman (1990) provided within the auditory domain alone—that the same set of tones will be perceived quite differently as a function of the pattern of other tones that precede, coincide with, or follow the set.

The multimodal situation actually offers the perceptual system the possibility of improving within-modality performance, and there is evidence that this does, in fact, occur. For example, Laurienti, Kraft, Maldjian, Burdette, and Wallace (2004) had observers push buttons to indicate whether a disk was red or blue, and they provided redundant information either within the same modality (by printing the word *red* or *blue* within the colored region) or in a second modality (by playing the words *red* or *blue* aloud). The cross-modal redundancy led to a significant improvement in response times, while the within-modality cueing did not. Studies by Naumer and colleagues (e.g., Alpert, Hein, Tsai, Naumer, & Knight, 2008; Hein et al., 2007) have suggested that there is relatively rapid integration of auditory and visual information in temporal regions superior temporal gyrus/medial temporal gyrus (STG/MTG) for familiar stimuli, and somewhat later integration in more frontal regions inferior frontal cortex (IFC) for stimuli that are less well established.

We began the introduction by pointing out that survival of the organism depends on the ability to accurately represent the world—an organism that does not recognize available food, or approaching predators, will not survive. The results of a growing body of research, including the present study, indicate that organisms have developed representations of the world around them that have two critical properties. First, scenes are represented by combining information from multiple sensory streams. Second, scenes are not represented as simple sets of passively captured physical properties. Rather, there is an active, constructive process that builds a representation of the scene using impressively sophisticated knowledge.

Acknowledgements Support was provided by Ministerio de Ciencia E Innovacion, Grant PSI2014-53277, Centro de Excelencia Severo Ochoa, Grant SEV-2015-0490, and by the National Science Foundation under Grant IBSS-1519908. We thank Kyle Sferrazza for expert programming assistance and Martijn Baart for his helpful suggestions. The article has benefitted from suggestions made by Souta Hidaka and an anonymous reviewer.

References

Alpert, G. F., Hein, G., Tsai, N., Naumer, M. J., & Knight, R. T. (2008). Temporal characteristics of audiovisual information processing. *Journal of Neuroscience*, *28*(20), 5344–5349.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Bulkin, D. A., & Groh, J. M. (2006). Seeing sounds: Visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, *16*(4), 415–419. doi:https://doi.org/10.1016/j.conb.2006.06.008

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. doi:10.1016/j.tics.2004.02.002

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*, 1191–1203.

Gordon, M. S., & Rosenblum, L. D. (2005). Effects of intrastimulus modality change on audiovisual time-to-arrival judgments. *Perception & Psychophysics*, *67*, 580–594.

Gregg, M. K., & Samuel, A. G. (2008). Change deafness and the organizational properties of sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(4), 974–991.

Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience*, *27*, 7881–7887.

Hidaka, S., Teramoto, W., Gyoba, J., & Suzuki, Y. (2010). Sound can prolong the visible persistence of moving visual objects. *Vision Research*, *50*(20), 2093–2099. doi:10.1016/j.visres.2010.07.021

James, W. (1890). *The principles of psychology.* New York, NY: Holt.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219. doi:10.1016/0010-0285(92)90007-O

Keetels, M., & Vroomen, J. (2011). No effect of synesthetic congruency on temporal ventriloquism. *Attention, Perception, & Psychophysics*, *73*, 209–218.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, *80*(1/2), 97–126. doi:https://doi.org/10.1016/S0010-0277(00)00155-4

Laurienti, P. J, Kraft, R. J., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*, 405–414.

Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, *4*, 501–506.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, *100*, 1777–1786.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *246*, 746–748. doi:https://doi.org/10.1038/264746a0

Moore, C. M., & Enns, J. T. (2004). Object updating and the flash-lag effect. *Psychological Science*, *15*(12), 866–871. doi:https://doi.org/10.1111/j.0956-7976.2004.00768.x

Moore, C. M., Mordkoff, J. T., & Enns, J. T. (2007). The path of least persistence: Object status mediates visual updating. *Vision Research*, *47*(12), 1624–1630. doi:https://doi.org/10.1016/j.visres.2007.01.030

Newell, F. N., Ernst, M. O., Tjan, B. S., & Bülthoff, H. H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science, 12*(1), 37–42. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11294226

Parise, C. V., & Spence, C. (2009). When birds of a feather flock together: Synesthetic correspondences modulate audiovisual integration in non-synesthesia. *PLoS ONE*, *4*, e5664.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, *22*(1), 46–49. doi:https://doi.org/10.1016/j.cub.2011.11.039

Rensink, R. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*, 17–42.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature*, *385*, 308.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*, 788. doi:https://doi.org/10.1038/35048669

Shinn-Cunningham, B. G., Lee, A. K. C., & Oxenham, A. J. (2007). A sound element gets lost in perceptual competition. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 12223–12227.

Stein, B. E., London, N., Wilkonson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience, 8*, 497–506.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*, 598–607.

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 513–518.