Article

PERCEPTION

Perception
2017, Vol. 46(7) 793–814

DOI: 10.1177/0301006616682755
journals.sagepub.com/home/pec
SAGE

Catching Audiovisual Interactions With a First-Person Fisherman Video Game

Yile Sun
Volen Center for Complex Systems, Brandeis University, Waltham, MA, USA

Timothy J. Hickey
Department of Computer Science, Brandeis University, Waltham, MA, USA

Barbara Shinn-Cunningham
Department of Biomedical Engineering, Boston University, Boston, MA, USA

Robert Sekuler
Volen Center for Complex Systems, Brandeis University, Waltham, MA, USA

Abstract

The human brain is excellent at integrating information from different sources across multiple sensory modalities. To examine one particularly important form of multisensory interaction, we manipulated the temporal correlation between visual and auditory stimuli in a first-person fisherman video game. Subjects saw rapidly swimming fish whose size oscillated, either at 6 or 8 Hz. Subjects categorized each fish according to its rate of size oscillation, while trying to ignore a concurrent broadband sound seemingly emitted by the fish. In three experiments, categorization was faster and more accurate when the rate at which a fish oscillated in size matched the rate at which the accompanying, task-irrelevant sound was amplitude modulated. Control conditions showed that the difference between responses to matched and mismatched audiovisual signals reflected a performance gain in the matched condition, rather than a cost from the mismatched condition. The performance advantage with matched audiovisual signals was remarkably robust over changes in task demands between experiments. Performance with matched or unmatched audiovisual signals improved over successive trials at about the same rate, emblematic of perceptual learning in which visual oscillation rate becomes more discriminable with experience. Finally, analysis at the level of individual subjects' performance pointed to differences in the rates at which subjects can extract information from audiovisual stimuli.

Keywords

multisensory interactions, temporal comodulation, audiovisual signals, video games

Corresponding author:
Robert Sekuler, Volen Center for Complex Systems, Brandeis University, Waltham, MA, USA.
Email: sekuler@brandeis.edu

Introduction

In our multisensory world, combining information from different sources or sensory modalities yields many benefits including increased response speed and accuracy (Kayser & Remedios, 2012; Tabry, Zatorre, & Voss, 2013). Among the earliest systematic studies of sensory combination was Wundt's (1894) pioneering work with brief visual and auditory stimuli. Following Wundt's lead, many studies of audiovisual interaction have combined a brief visual stimulus, such as a flash, with a brief auditory stimulus, such as a beep or click. Well-known examples include various sound-induced flash illusions (Rosenthal, Shimojo, & Shams, 2009; Shams, Kamitani, & Shimojo, 2000) and sound-induced changes in perceived motion, the "bouncing-streaming" illusion (Donohue, Green, & Woldorff, 2015; Kawachi, Grove, & Sakurai, 2014; Roudaia, Sekuler, Bennett, & Sekuler, 2013; Sekuler, Sekuler, & Lau, 1997), and studies of the temporal binding window (Navarra et al., 2005; Spence & Squire, 2003; Van Wassenhove, Grant, & Poeppel, 2007).

Experiments that use brief stimuli as guides to audiovisual interaction are constrained by the narrow temporal dynamics of such stimuli. After all, audiovisual stimuli in natural environments are usually continuous, extended, and have temporal dynamics of varying rate or intensity. So it is unsurprising that special insights into audiovisual combination have come from studies whose stimuli were extended in time rather than temporally punctate (e.g., Barakat, Seitz, & Shams, 2015; Guttman, Gilroy, & Blake, 2005; Shipley, 1964). Importantly, studies with continuous, time-varying stimuli have identified similarity of temporal structure as a powerful determinant of cross-modal combination (Maddox, Atilgan, Bizley, & Lee, 2015; Parise, Harrar, Ernst, & Spence, 2013; Rainville & Clarke, 2008; Roach, Heron, & McGraw, 2006). The most frequently experienced multisensory situation in our everyday lives is audiovisual speech, that is, speech whose source is seen as well as heard. Research on audiovisual speech has identified temporal coherence as a critical factor in binding auditory and visual speech signals into semantically coherent words and sentences (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; McGurk & MacDonald, 1976; Spence & Deroy, 2012).

Our experiments used a specially designed video game, "Fish Police!!," whose usefulness was proven in a recent field study at Boston's Museum of Science (Goldberg, Sun, Hickey, Shinn-Cunningham, & Sekuler, 2015). The game presented players with concurrent auditory and visual stimuli, which lasted up to several seconds. The visual and auditory stimuli were meant to mimic perceptual aspects of audiovisual speech. Visual stimuli were animated images of a fish whose overall size varied, much as the opening between a speaker's lips would vary. This stimulus was accompanied by a broadband sound that was amplitude modulated, in much the way that a human speaker's volume would be modulated during conversation (Joris, Schreiner, & Rees, 2004). Despite having to test subjects of all ages and to do so in a noisy public setting, Goldberg et al. (2015) demonstrated robust, reliable audiovisual effects. Specifically, when the rate at which a fish's size oscillated matched the rate at which the accompanying, task-irrelevant sound varied, categorization was more accurate and faster.

Goldberg et al.'s study left important questions about audiovisual interaction unanswered. For example, did the difference between performance with matched versus mismatched audiovisual signals reflect a performance benefit from the matched condition, a performance cost imposed by the mismatched condition, or some combination of the two? Additionally, the brief, 5–minute test time allowed for each player in the Museum setting produced just a snapshot of performance, foreclosing the possibility of gauging any learning that might have taken place, including possible differences in learning rates for matched versus mismatched audiovisual stimuli. Here, we report three experiments that confirm Goldberg et al.'s basic result, and go on to address questions that they could not. Our study used visual and auditory stimuli that modulated sinusoidally at temporal rates of 6 or 8 Hz. Pilot testing showed that this difference in rate was discriminable, but not perfectly so. Additionally, these two rates were sufficiently low that the temporal processing capacities of vision or audition would not affect performance (Welch & Warran, 1980; Welch, DuttonHurt, & Warren, 1986).

Experiment 1

The main aim of Experiment 1 was to confirm and extend the audiovisual interactions reported by Goldberg et al. To that end, we adopted their basic design, but with several changes, most notably the addition of a Control condition in which fish were not accompanied by an amplitude modulated sound, and the presentation of many more test trials than were possible in the museum setting. The experiment asked how nominally vision-based judgments were influenced by an accompanying sound. In this and the following two experiments, subjects had to discriminate the rate at which the fish oscillated in size, either at 6 or 8 Hz. In the experimental conditions, fish were accompanied by a broadband sound, amplitude modulated at either 6 or 8 Hz. Subjects were instructed to ignore the sound, which was gated on and off simultaneously with the fish's visual oscillation. Modulation rates of visual and auditory stimuli were either matched (both at 6 Hz or both at 8 Hz) or put into conflict (one at 6 Hz the other 8 Hz). In a Control condition, no sound other than the constant background sound accompanied the fish.

Method

Stimuli. Gameplay was controlled by a Macintosh computer running Java-7 code that was constructed in the open source Eclipse integrated development environment (Eclipse Foundation, Ottawa, Canada). Visual stimuli were presented on a 20-inch LCD monitor with resolution set to 1280 × 960 pixels. The display refreshed at 75 Hz. Subjects were seated with eyes positioned ~70 cm from the monitor. Auditory stimuli were delivered via circumaural headphones (Sennheiser HD 280 Pro).

The size of each fish oscillated sinusoidally over time around a mean length and height of 4.8° and 2.4° visual angle, respectively. Modulation rate was either 6 or 8 Hz, and at either rate, the modulation depth was 25%. Aside from their modulation rates, all fish were identical in appearance.

For each subject, the game designated one rate of visual size modulation as the criterion for a "good" fish and the other rate as the criterion for a "bad" fish. This mapping of size-oscillation frequency (6 or 8 Hz) onto the categories "good" fish and "bad" fish was counterbalanced over subjects. A subject was told that a fish's oscillation rate marked its species membership, but was not told which frequency marked which species, which had to be learned via trial and error using the feedback given after each response. Subjects were instructed that they had to respond within 2 seconds of the fish's appearance, otherwise, the fish disappeared from view and the trial was terminated. Preliminary testing showed that 2 seconds were sufficient for subjects to respond with very few failures to respond in time. An animated progress bar near the top of the game window showed how much time remained until the response deadline was reached.

Each fish came into view randomly, either from the right or left side of the game window (32.7° wide by 24.6° high). Starting midway up the game window, the fish swan at 13.4°/s along a quasihorizontal trajectory whose direction was randomly perturbed from frame to frame, as described later. When a fish was scheduled to appear, it did not start from offscreen and then swim into the game window. Rather, at the very instant it appeared, 50% of the fish was instantly visible. We chose this starting state because pilot testing revealed that if a fish began entirely off screen, subjects found it difficult to focus on the fish's visual aspect—because oscillation of the accompanying sound was audible before the fish's size oscillation could be seen. Once onscreen, a fish swam on an irregular path across the screen toward the opposite side. The direction in which a fish swam was perturbed by samples drawn from a uniform random distribution whose range was ±2.4°. This variability increased the realism of a fish's swimming movements, and may have made it more challenging for subjects to register the fish's rate of size oscillation. Notional reflecting boundaries excluded fish from zones within 5° of the top and bottom of the game window. Figure 1 shows a pair of screen shots from the game.

The sound accompanying a fish was an inharmonic tone complex comprising 10 random pure tones, summed and then multiplied by an exponentially decaying window that ranged from a value of 1 at onset, down to 1/3 (200 ms time constant). The pure tone components of the sound were selected randomly from frequencies between 150 and 1500 Hz. Their amplitudes were randomized over a 15-dB range. Figure 2 shows the sound's time waveform and spectral components. This basic stimulus was the template for all sounds; as needed, 6 or 8 Hz modulation was applied to the tone complex by multiplication in the time domain to produce a 25% modulation depth. The overall level of the stimulus was set to a comfortable listening level (between 55 and 75 dB SPL).

This auditory stimulus was presented with an interaural time difference (ITD) of ±300 μs, with the leading ear always corresponding to the side of the display at which the fish first appeared. This ITD was meant to diminish the spatial uncertainty that might arise from randomizing the side at which a fish could appear (Eckstein et al., 2013). The sound accompanying each fish was presented against a background sound (39 dB SPL)

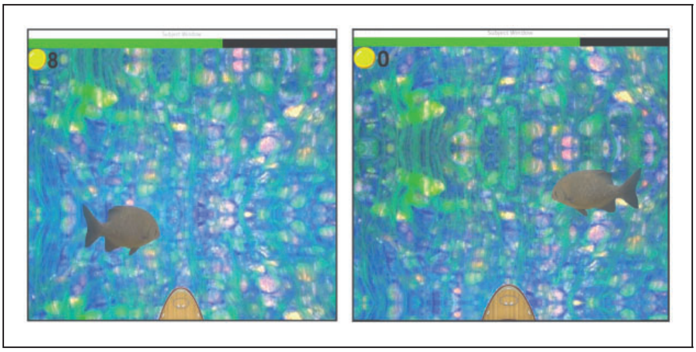

Figure 1. Two screen captures from the game. The game window's background simulates the view that a player might see looking down into the clear water of a shallow river. The player is a fisherman who looks down into the river while standing on the bow of a boat (seen at the bottom of each panel). A flatfish enters the game window randomly from the left or the side, and swims toward the opposite side. Each screen capture shows a fish partway through its journey. A slowly drifting background image of a river bottom simulates the boat's movement on the river. At the window's top, an animated progress bar indicates the time remaining until the 2-second response deadline. Also near the screen's top, the number adjacent to the coin gives the player's current score. The mean luminance of the background and fish were 37 and 23 cd/m^2, respectively.

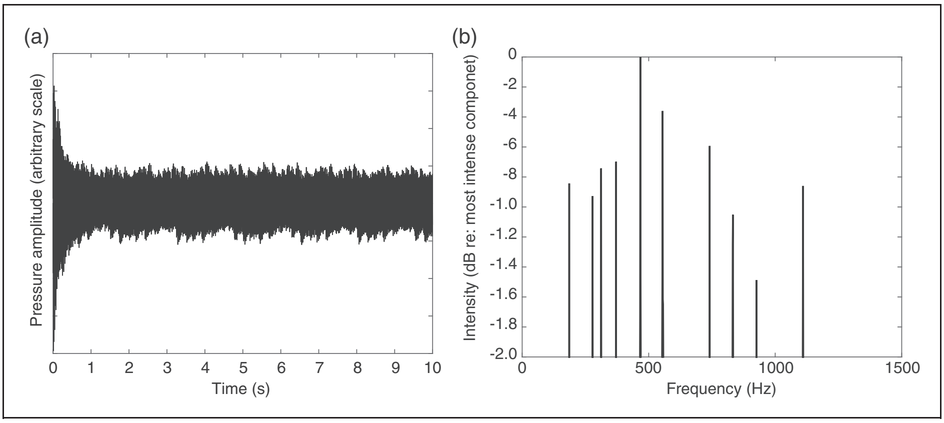

Figure 2. The auditory stimulus from which sounds that accompanied fish were generated. (a) Time domain representation of the inharmonic tone complex, showing the exponential time window of the stimulus. (b) Frequency content of the steady-state portion of the stimulus, which consists of 10 randomly selected pure tones with frequencies chosen from a random distribution between 150 and 1500 Hz. Their relative levels were chosen from a random distribution that spanned a 15-dB range. To create congruent or incongruent stimuli, this basic stimulus was multiplied by an amplitude-modulated envelope with 6 or 8 Hz modulation (depth), as appropriate. The absolute level of the stimulus was set to a comfortable listening level.

synthesized to resemble the sound made by water running over rocks. In a Control condition, fish were not accompanied by any sound other than the constant background sound.

Procedure

Subjects were instructed to categorize as rapidly as possible each fish as "good" or "bad," based only on the rate at which the fish's size modulated. Categorization judgments were to be signaled by pressing one of two buttons on a gamepad (Logitech Dual Action). A response caused the fish to disappear, setting the stage for the next fish to be spawned after a random inter-fish interval of 1.5, 2.0, or 2.5 seconds. As mentioned earlier, from the time of a fish's initial appearance in the game window, subjects had no more than 2 seconds in which to respond.

In Experiment 1, subjects played *Fish Police!!* under three different conditions. One was a condition in which visual oscillation frequency matched, or was congruent, with the rate at which the sound's amplitude modulated (hereafter, AV Congruent); a condition in which the frequency of visual oscillation was different from the rate of the sound's amplitude modulation (hereafter, AV Incongruent); and a no-sound Control condition, which provided a baseline against which performance with AV Congruent and AV Incongruent stimuli could be compared. In this Control condition, fish were mute, with only the background sound of rushing water present as they swam across the screen. AV Congruent, AV Incongruent, and Control conditions were presented in separate 80-trial blocks. Each block was repeated twice, producing a total of 480 trials per condition. The order in which all blocks were presented was randomized anew for each subject. Successive blocks of 80 trials were presented with no break between. Before the experiment, subjects were familiarized with the task and display. For this purpose, each subject received one or more sets of 12 practice trials with AV Congruent fish.

In this experiment and the next, the Control condition was never presented in a subject's initial test block of 80 trials. For five subjects, the first block of trials entailed AV Congruent fish; for the remaining subjects, AV Incongruent fish were presented in the first block. Subjects achieving 75% or more correct responses were allowed to continue on to the actual experiment; subjects failing that criterion received additional practice sets. Six subjects reached criterion in just one set of trials; four subjects required two sets of practice trials.

Subjects

Twelve subjects, 19 to 29 years old, recruited from the Brandeis University community began the experiment. Here, and in each subsequent experiment, all subjects had normal or corrected-to-normal Snellen visual acuity, and reported having normal hearing. Each gave written informed consent to a protocol approved by Brandeis University's Committee for the Protection of Human Subjects, and received $10 for participation. One subject, expressing frustration at the task's difficulty, withdrew from the experiment after just single block of trial. A second subject completed the experiment, but in every condition showed an accuracy that was no different from chance level. We excluded both these subjects' data from further analysis.

Results and Discussion

Each analysis of variance (ANOVA) results presented later were generated by ezANOVA, a component of the *ez* package for *R* (Lawrence, 2013). Entries in each ANOVA were individual subjects' values, either median response time (RT) or accuracy, depending upon the variable under consideration. The ANOVAs generated Type III sum of squares. Additionally, η_G^2, a measure of effect size appropriate for repeated measures designs (Bakeman, 2005; Olejnik & Algina, 2003) is presented. Where 95% confidence limits (CLs) are given, the values were generated from 1,000 bootstrap samples using the adjusted bootstrap percentile (BCa) method (Davison & Hinkley, 1997, Chapter 5).

Across subjects and conditions, fish were categorized with a mean accuracy $\bar{x} = .79$, with 95% CLs [.75,.82]. Across subjects and conditions, RTs associated with correct judgments averaged $\bar{x} = 899.3$, with 95% confidence [840.1, 989.3]. Repeated measures analyses of variance contrasted subjects' mean proportions correct and median RTs for "good" and "bad" fish, and for 6 and 8 Hz modulation rates. None of these variables produced a statistically significant effect, all $p > .25$. So subsequent analyses examined mean proportions correct and median RTs averaged across "good" and "bad" fish, and across both rates of modulation. In computing RT values, only correct responses were included.

Figure 3 shows results for response accuracy (left panel) and RT (right panel) averaged over subjects for each condition. We will consider results for response accuracy first. A repeated measures ANOVA on proportion correct showed a statistically significant difference among AV Congruent, AV Incongruent, and the Control conditions, $F(2,18) = 8.94$, $p < .002$, $\eta_G^2 = .36$. Specifically, subjects categorized AV Congruent fish more accurately than AV Incongruent ones, $\bar{x} = .86$ with 95% CLs [.83,.89], and $\bar{x} = .71$ with 95% CLs [.63,.78], respectively. Control fish were categorized with an accuracy $\bar{x} = .67$, 95% CLs [.62,.71]. Planned contrasts confirmed significant differences between proportion correct responses for AV Congruent fish versus fish from either of the other conditions, $t(9) = 8.69$, $p < .001$ and $t(9) = 2.89$, $p = .02$, for AV Incongruent and Control fish, respectively. The difference between AV Incongruent and Control conditions was not

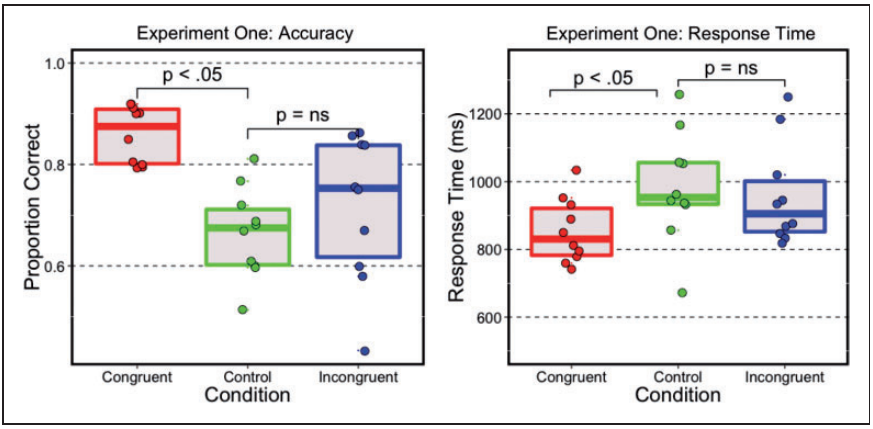

Figure 3. Mean proportion correct (left panel) and median response time (right panel) for each of the three conditions in Experiment 1.

significant, $t(9) = .80$, $p = .44$. Moreover, differences between AV Congruent and AV Incongruent conditions were unrelated to the number of practice blocks subjects required before the experiment (Spearman's $\rho = .28$, $p = .43$).

RTs associated with correct responses showed a pattern similar to that for the accuracy measure. The mean of subjects' median RTs was lowest with AV Congruent fish ($\bar{x} = 853.3$ ms, 95% CLs [807.1, 930.4]), but were similar to one another for AV Incongruent and the mute Control fish, $\bar{x} = 945.3$ ms with 95% CLs [878.7, 1064.1] and 984.0 ms with 95% CLs [898.9, 1060.3], respectively. An ANOVA showed a significant difference among AV Congruent, AV Incongruent, and the no-sound Control conditions, $F(2,18) = 4.43$, $p = .01$, $\eta^2_G = .09$. Planned contrasts confirmed that the difference between AV Congruent and AV Incongruent RTs and the difference between AV Congruent and Control fish were each significant, $t(9) = 2.80$, $p = .02$ and t $(9) = 2.38$, $p = .04$, respectively. However, the difference between Control fish and AV Incongruent fish was not statistically significant, $t(9) = .03$, $p = .98$. Finally, the difference in RTs between AV Congruent and AV Incongruent conditions was unrelated to the number of practice blocks subjects needed prior to the experiment (Spearman's $\rho = .28$, $p = .43$).

As can be seen in Figure 3, matched auditory and visual modulations produced faster and more accurate responses than did mismatched modulations. Although subjects had been instructed to categorize fish solely on the basis of what they saw, response speed and accuracy were both aided by the sounds that accompanied AV Congruent fish. Importantly, for both dependent measures, responses to Control fish were indistinguishable from ones to AV Incongruent fish.

Experiment 1 showed that Control fish and AV Congruent fish produced quite different results. Did that difference arise from the fact that Control fish were accompanied by no sound or from the fact that Control fish were not accompanied by a particular kind of sound, namely, one that was amplitude modulated? Experiment 2 was designed to examine that distinction.

Experiment 2

The control condition of Experiment 1 was meant to provide a baseline against which responses to AV Congruent and AV Incongruent stimuli could be assessed. In the

Control condition, no sound accompanied a fish except for the background sound that was always present. We expected results from the Control condition to clarify whether the effects seen in Experiment 1 reflected an improvement in performance in the AV Congruent condition or a reduction in performance in the AV Incongruent condition. However, we realized after the experiment that its Control condition may have been flawed. Because no sound was coterminous with the Control fish, subjects were deprived of auditory timing markers that might have aided performance. Such markers were, however, available in both other conditions. To rule out the influence of this possible confound, we modified the Control fish for Experiment 2 to make the appearance of a fish coterminous with a sound that was an unmodulated version of the sounds that accompanied fish in other conditions. As in Experiment 1, AV Congruent, AV Incongruent, and Control conditions were presented in randomly ordered 80-trial blocks, two blocks per subject.

Subjects

Eleven subjects whose ages ranged from 19 to 27 years began Experiment 2; none had served in Experiment 1. Before the experiment, each subject completed a short practice with one or more sets of 12 trials in which "good" and "bad" AV Congruent fish were randomly intermixed. Eight subjects reached criterion (75% correct) in just a single set of practice trials; one subject required two sets of practice before proceeding to the main experiment, while one subject needed three sets. One subject's response accuracy was no better than chance in every condition of the experiment; that subject's data were discarded, leaving 10 subjects for analysis.

Results and Discussion

As in the preceding experiment, repeated measures analyses of variances contrasted subjects' mean proportions correct and median RTs for "good" and "bad" fish, and for 6 and 8 Hz modulation rates. Neither the main effect of fish type nor any of the interactions was statistically significant, all $p > .10$. So, to simplify subsequent analyses, we averaged over subjects' proportions correct and RTs for "good" and "bad" fish and over both rates of modulation. Figure 4(a) and (b) shows the mean proportion correct $\bar{x} = .78$ with 95% CLs [.74,.81] and the mean of subjects' median RTs $\bar{x} = 868.9$ with 95% CLs [878.7, 1064.1]. A repeated measures ANOVA on values of proportion correct showed a significant effect of condition, comparable to that seen in Experiment 1: $F(2,18) = 14.96$, $p < .001$, $\eta^2_G = .26$. For AV Congruent, AV Incongruent, and Control fish, mean proportions correct were $\bar{x} = .85,.71$, and .66, with associated 95% CLs = [.79,.89], [.65,.75], and [.63,.68], respectively. Planned contrasts confirmed that the value of proportion correct differed reliably between AV Congruent and AV Incongruent conditions, but not between AV Incongruent and Control conditions, $t(9) = 3.45$, $p = .007$ and $t(9) = 1.27$, $p = .24$. The difference in proportion correct between AV Congruent and AV Incongruent conditions was unrelated to the number of practice blocks subjects required before the experiment (Spearman's $\rho = .14$, $p = .69$).

Turning to subjects' RTs, an ANOVA on the three conditions produced $F(2,18) = 8.62$, $p < .01$, $\eta^2_G = .07$. Mean RTs and 95% CLs for AV Congruent, AV Incongruent, and Control conditions were 810.8 ms [765.8, 850.7], 927.1 ms [851.2, 1001.8], and 959.4 ms [897.5, 1039.1]. Follow-up planned contrasts showed that RTs to AV Congruent fish were significantly shorter than those for either AV Incongruent, $t(9) = 2.55$, $p = .03$, or Control fish, $t(9) = 2.72$, $p = .02$. Moreover, consistent with what was seen in Experiment 1, RTs for

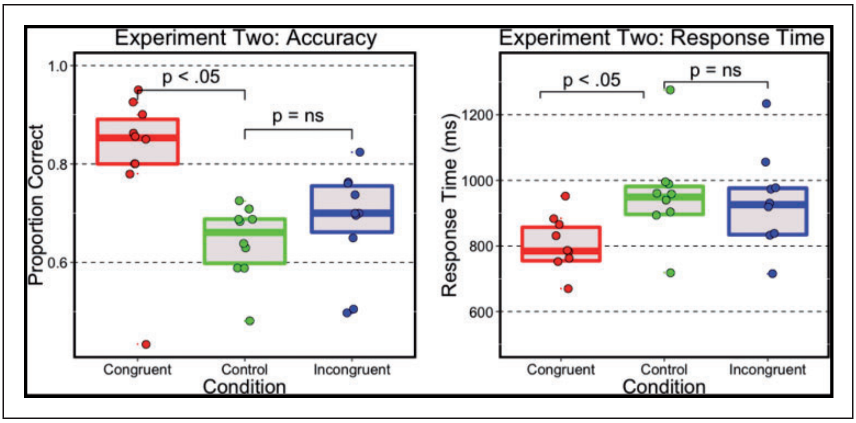

Figure 4. Mean proportion correct (left panel) and median response time (right panel) for each of the three conditions in Experiment 2. Filled symbols represent individual subjects. Box plots span the first and third quartiles for each condition; the horizontal bar within the box represents the mean.

AV Incongruent fish did not differ significantly from those for Control fish, $t(9) = .24$, $p = .82$. The difference in mean RTs for the AV Congruent and AV Incongruent conditions was unrelated to the number of practice blocks subjects required to reach criterion (Spearman's $\rho = .22$, $p = .54$)

Comparing Figures 3 and 4 shows that Experiment 2's results with each dependent measure closely resemble the analogous results from Experiment 1. Relative to the Control conditions in either experiment, audiovisual congruence has a positive effect on the accuracy and speed of categorization. In contrast, the stimulus-response incompatibility (Fitts & Deininger, 1954; Tucker & Ellis, 1998) built into the AV Incongruent condition failed to produce any significant decrease in performance compared with the Control conditions in either experiment, a point considered more fully in the General Discussion. Finally, visual comparison of the right hand panels of Figures 3 and 4 suggests that overall, RTs tend to be somewhat longer in Experiment 1 than in Experiment 2. However, that difference is relatively small, and not statistically significant: $F(1, 18) = 2.13$, $p = .16$. Overall, results from Experiment 2 are consistent with those from Experiment 1, which used a different Control condition. Note that performance with a coterminous, unmodulated sound, in Experiment 2's Control condition, was not noticeably different from performance with no sound, in Experiment 1's Control condition. This means that the result we saw with Control fish in Experiment 1 did not come from the absence of an auditory signal marking a fish's onset. This clarification gives us confidence that relative to the Control condition, the AV Congruent condition produces a benefit in performance, and also that the unmatched sound in the AV Incongruent condition is essentially without effect.

Experiment 3

In the preceding experiments, AV Congruent and AV Incongruent fish were segregated into separate blocks of trials. Although this design decision allowed subjects to adopt a consistent criterion throughout a block, it also opened the possibility that subjects might adopt distinct strategies for the different types of stimuli that appeared in separate blocks. To test this possibility and to gauge how robust our previous results were in the face of changed context and task demands, Experiment 3 randomly intermixed AV Congruent fish and AV

Incongruent fish. Unlike the case in Experiments 1 and 2, random intermixture of conditions made it impossible for a subject to anticipate whether a fish's auditory and visual attributes would match or not. Verifying a match between audiovisual stimuli was impossible until a fish actually appeared. Additionally, in the preceding experiments, sounds were presented with a 300 μs ITD whose leading ear corresponded to the side of the screen from which a fish came into view. One of Experiment 3's aims was to test whether the ITD affected performance. So, we compared performance with the ±300 μs ITD sound used previously to performance when the sound was presented diotically, with zero ITD.

Subjects

Ten subjects between the ages of 19 to 27, recruited from the Brandeis University community, began and successfully completed the experiment. None had served previously, and each was paid $10 for participating.

Procedure

Each subject served in two sessions, each comprising equal numbers of AV Congruent and AV Incongruent fish presented in intermixed random order over the session's 200 trials.[1] In one session, the sound accompanying a fish included a 300-μs ITD consistent with the location, left or right, at which the fish entered the field of view. This replicated the condition used with AV Congruent and AV Incongruent fish in the previous experiments. In the other session, fish sounds were presented with zero ITD. Subjects got no instructions about the auditory localization cues (or lack thereof). Note that both sessions for a subject were run consecutively on the same day, with a minimum of 10 minutes break in between. The order in which AV Incongruent and AV Congruent conditions were run was counterbalanced over subjects. To familiarize them with the task and game controller, subjects received practice trials with AV Congruent fish. Eight out of ten subjects met the criterion of 75% correct in just one 12-trial practice set; the remaining subjects each required two sets of practice in order to reach the same criterion.

Results and Discussion

Figure 5 shows the results of Experiment 3, with proportion correct values in the left hand panel and RT values in the righthand panel. Results for AV Congruent fish are shown in shades of red; results for AV Incongruent fish are shown in shades of blue. Within each panel, results are shown separately for trials on which an ITD was present (darker color bars) and for trials on which it was absent (lighter color bars)

When the sound accompanying a fish included a 300-μs ITD, mean proportions correct for AV Congruent and AV Incongruent fish were $\bar{x} = .85$ with 95% CLs [.83,.88] and.66 with 95% CLs [.60,.71], respectively. The absence of an ITD left these values essentially unchanged: $\bar{x}\text{b} = .86$ with 95% CLs [.81,.91] and.66 with 95% CLs [.59,.73] for AV Congruent and AV Incongruent fish, respectively. An overall ANOVA on proportion correct showed a significant effect of audiovisual congruence $F(1,9) = 49.33$, $p < .0001$ $\eta^2_G = .51$, but neither a significant main effect of ITD, $F(1,9) = .15$, $p = .71$, nor a significant interaction between congruence and ITD, $F(1,9) = .003$, $p = .95$. As in the preceding two experiments, the difference between AV Congruent and AV Incongruent conditions was unrelated to the number of practice blocks subjects had prior to the experiment (Spearman's $\rho = .21$, $p = .55$).

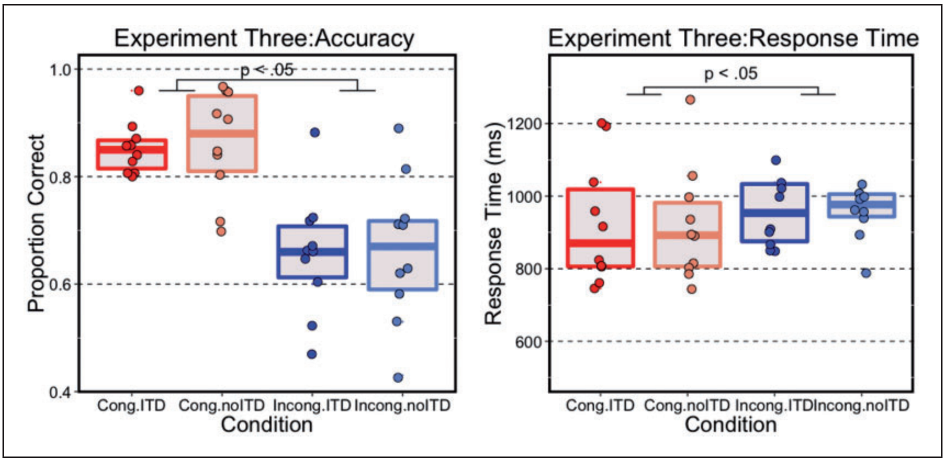

Figure 5. Experiment 3 mean proportion correct categorization (left) and median response time (right) for AV Congruent and AV Incongruent fish presented with and without an ITD. Filled symbols represent individual subjects. Box plots span the first and third quartiles for each condition; the horizontal bar within the box represents the mean.

Mean RTs for AV Congruent and AV Incongruent fish were $\bar{x} = 925.3$ ms with 95% CLs [853.2, 1023.1] and $\bar{x} = 999.7$ ms with 95% CLs [932.0, 1148.6], respectively, when the fish's sound included the ITD; corresponding values when the ITD was zero were essentially unchanged: $\bar{x} = 918.7$ ms with 95% CLs [855.9, 1014.4] and $\bar{x} = 1000.5$ ms with 95% CLs [939.2, 1115.9], respectively (see Figure 5). Confirming these results, an ANOVA showed a significant effect of audiovisual congruence, $F(1,9) = 8.66$, $p < .02$, $\eta^2_G = .05$, but neither a significant main effect of ITD, $F(1,9) = .001$, $p = .92$, nor a significant interaction between congruence and ITD, $F(1,9) = .01$, $p = .75$. In summary, trials with the 300 µsc ITD and trials with zero ITD produced comparable results on both dependent measures, response speed, and response accuracy. As in the two preceding experiments, the difference between RTs for AV Congruent and AV Incongruent conditions was unrelated to the number of practice trials a subject received (Spearman's $\rho = .06$, $p = .87$). In summary, the interaural time difference between left and right ears seemed to give no response advantage, either for proportion correct or for RT.

Experiments 1 and 2 tested AV Incongruent and AV Congruent conditions in separate blocks, which produced a consistent, predictable relationship between the auditory and visual signals within a block. This arrangement allowed subjects to know before a fish appeared whether the auditory modulation accompanying the fish would match or not match the fish's oscillation in size. As a result, with AV Incongruent fish, subjects could have anticipated the value of discounting what they would hear. However, that approach could not have worked in Experiment 3, where the congruence of auditory and visual cues varied randomly from trial to trial. Importantly, for both dependent measures, the difference between responses to AV Congruent and AV Incongruent fish closely matched the corresponding difference from Experiments 1 and 2: .20 for accuracy and 74.45 ms for RTs. The similarity between results obtained with randomly interleaved AV Congruent and AV Incongruent fish (Experiment 3) and results when AV Incongruent and AV Congruent fish were segregated into separate blocks of trials (Experiments 1 and 2) is consistent with idea that subjects in the first two experiments probably did not exploit the predictability of fish types in order to engage systematically different strategies for different conditions.

Differences among individual subjects

Using a somewhat different implementation of *Fish Police!!* from the one used for our experiments, Goldberg et al. (2015) found suggestions of considerable differences among subjects. Although most of their 60 subjects showed reliable differences in response to AV Congruent and AV Incongruent fish, about 10% of the subjects did not. Because each subject was available for only 5 minutes' testing, these apparent individual differences might have come from some uncontrolled sources, such as subjects' imperfect understanding of the task. Results presented here are better suited for evaluating individual differences as the longer testing sessions yielded more data per subject, and practice trials ensured that subjects understood the task. To examine individual differences in our results, we focused on the relationship between a subject's accuracy and that subject's response speed, the well-known speed-accuracy tradeoff (SAT; Heitz, 2014; Henmon, 1911). We worked with the SAT because correct responses in *Fish Police!!* depend upon time-varying (rate) visual information. Gradual accumulation of sensory evidence, including visual evidence about the rate at which a fish oscillates, is fundamental to decision making in a host of situations and in various species (Brunton, Botvinick, & Brody, 2013; Shadlen & Kiani, 2013).

Consider how that approach applies to *Fish Police!!*. From a simple normative perspective, a subject in our experiments should opt to collect as much relevant sensory information as possible before committing to a response. The observation period would be as long as possible, up to the limit imposed by the game's 2-second response deadline. But that depiction does not capture our subject's behavior. In fact, on 75% of all trials, subjects responded correctly after having viewed the fish for less than ~1050 ms, just half the observation time that would have been permitted by the 2-second response deadline. Even more surprising, for only 1% of trials with correct responses did subjects observe a fish for as much as ~1720 ms, a time close to the response deadline. A highly visible animated yellow progress bar was near the top of the game window, only ~5° above the center of the game window. The length of the progress bar gave real-time information about the time remaining until the deadline. Subjects knew, or should have known, how much more time they had to observe and accumulate information from the stimulus. However, on average after availing themselves of just half the information available, their level of confidence was sufficient to support a response (Shadlen & Kiani, 2013).

To optimize the power of our SAT analysis, we combined result from all three experiments. Before doing the analysis, we decided to drop one subject as a clear outlier. That subject's overall mean accuracy and overall mean RT, 45.3% and 405.5 ms, were ~3.3 standard deviations and ~2.9 standard deviations below the mean for all subjects, respectively. For the 29 remaining subjects, mean accuracy was .78 with 95% CLs [.75,.80], while mean RT was 911 ms with 95% CLs [875.5, 967.7]. To evaluate the possible SAT, we took account of subjects' RTs and accuracy. We thought that even if two subjects were equally adept at processing the fish's time-varying visual information, by exploiting additional observation time before responding, one subject could produce more correct responses than another subject. That would produce evidence of a SAT.

Figure 6 plots each subject's mean proportion correct against that same subject's mean RT. Also shown are the best fit linear function, as determined by maximum likelihood, and shaded 95% CLs around that function. If individual differences among performance arose only from a tradeoff between response speed and response accuracy, that is, from differences in how long subjects chose to observe the fish before committing to a response, the path of data points in the figure would have a positive slope, running from lower left to upper right. That is opposite to what the figure shows. Our working hypothesis, then, is that differences

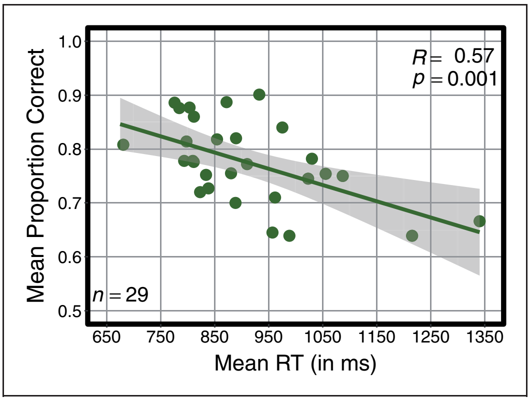

Figure 6. Each data point is the mean for one subject's overall proportion of correct responses and that subject's overall mean response time on correct trials. Also shown are the best fit line (determined by maximum likelihood) and the 95% CLs around the best fit line.

among subjects arise from some factor other than the simple relationship embodied in the SAT. Clearly, differences in accuracy among subjects do not result solely because subjects allow themselves different amounts of observation time. Rather, we believe that individual differences apparent in Figure 6 reflect differences in the rates at which subjects are able to extract information from audiovisual stimulus.

To examine individual differences further, for each subject, we compared two measures of how performance was impacted by a match in audiovisual signals. For this analysis, we took the difference in accuracy between AV Congruent and AV Incongruent conditions and the difference in RT between those two conditions. A positive relationship between the two measures would be a sign of reliable individual differences in the impacts of audiovisual congruence. Figure 7 plots the two sets of differences against one another. The figure also shows the best fit linear function, and the 95% CLs around that best fit. The Pearson product moment correlation between the two variables was $\rho = .33$. With 29 data pairs, this value corresponds to a one-tailed $p = .04$. We think a one-tailed test is justified because of a clear a priori hypothesis that the two measures will be associated positively. The scatterplot shows an association between the two ways of assessing the effect of the sound: a large effect as indexed by RT differences *tends* to be associated with a large effect as indexed by accuracy differences between AV Congruent and AV Incongruent conditions. This association between the two variables suggests that there are reliable differences in how individual subjects are affected by audiovisual congruence.

General Discussion

In three experiments, a first-person fisherman game provided a vehicle for examining how a temporal match or mismatch between auditory and visual signals impacted judgments that were nominally vision based. Notwithstanding encouragement to judge only on the basis of a fish's visual characteristics, performance was clearly affected by the presence of auditory signals. Subjects were consistently more accurate and faster to respond when the frequency of an accompanying amplitude modulated sound matched the frequency with which a fish oscillated in size. Compared with Control fish, an unmatched sound seemed to have no detectable impact, either on RT or accuracy.

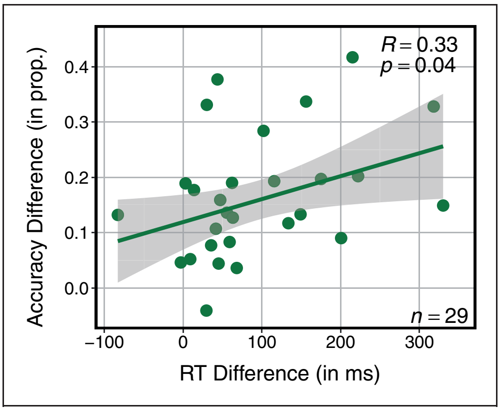

Figure 7. Each data point is one subject's accuracy difference between AV Congruent and AV Incongruent and that subject's response time difference between AV Congruent and AV Incongruent on correct trials. Also shown are the best fit line (determined by maximum likelihood) and the 95% CLs around the best fit line. Note: The *p* value is for a one-tailed test.

Differences among types of fish

All three experiments agreed that categorization of AV Incongruent fish was poorer than that of AV Congruent fish. Was this difference a result of enhanced performance with AV Congruent fish or diminished performance with AV Incongruent fish? If diminished performance with AV Incongruent fish were to blame, how did that diminished performance come about? One possibility is a sound-induced shift in perceived visual modulation, away from its actual rate and toward the rate at which the sound modulates. On that view, errors with AV Incongruent fish occur because the fish's visual oscillations are perceptually entrained by the mismatched auditory modulation rate (Guttman et al., 2005; Shipley, 1964), a form of temporal ventriloquism. Alternatively, diminished performance with AV Incongruent fish could have come from stimulus-response incompatibility (Fitts & Deininger, 1954; Tucker & Ellis, 1998). To appreciate this point, consider a subject for whom a fish oscillating in size at 6 Hz must be categorized as a "good" fish. When such a fish is accompanied by an 8 Hz amplitude modulated sound, the visual signal for which a "good fish" response is correct is accompanied by an amplitude modulated sound that promotes the opposite "bad fish" response. Competition between the two could degrade accuracy and slow response. We can reject both these possible explanations for diminished performance with AV Incongruent fish. After all, fish that were unaccompanied by a sound or were accompanied by an unmodulated sound (Control fish) produced about as many errors as fish accompanied by the mismatched sound (AV Incongruent fish); additionally, RTs for those conditions did not reliably differ from one another. With no concurrent sound or with a concurrent unmodulated sound, Control fish would suffer neither temporal ventriloquism nor stimulus-response incompatibility. So, the similar outcomes with AV Incongruent and Control fish make it unlikely that diminished performance with AV Incongruent fish was to blame for the consistent large difference in results from AV Incongruent and AV Congruent fish.

As just discussed, in Experiments 1 and 2, neither mean RTs nor mean accuracies differed for AV Incongruent and Control fish. In turn, for AV Congruent fish, the means for each dependent measure differed significantly from the corresponding measures in the other conditions. It is worth noting that for AV Incongruent and for Control fish, the order relationships among results from individual subjects was somewhat variable: For each experiment, about half the subjects showed results with AV Incongruent fish slightly above those for Control fish, while other subjects showed the opposite relationship. In contrast, for every subject, the dependent measures with AV Congruent fish were above the corresponding values for both of the other conditions. Our working hypothesis is that detection of a strong cross-correlation (Parise et al., 2013) between auditory and visual signals with AV Congruent fish initiates integration of visual and auditory signals. Such integration might take place in heteromodal regions of the cerebral cortex, as has been observed with other audiovisual paradigms (e.g., Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Calvert, Campbell, & Brammer, 2000; Talsma, Doty, & Woldorff, 2007). On this view, in the absence of a sufficiently strong cross-correlation, integration would not occur. Although frankly post hoc, this formulation could explain the relative performance levels with AV Congruent, AV Incongruent, and Control fish. The modulation rates of our fish were sufficiently low and stable to enable rapid and reliable computation of the requisite cross-correlation, perhaps based on temporal features that were independently extracted from a stimulus' auditory and visual components and then compared (Pollack, 1974; Pollack, 1975; Fujisaki & Nishida, 2005). Moreover, the short latency of cortical responses to a mismatch between the auditory and visual aspects of an audiovisual signal (Winkler, Horváth, Weisz, & Trejo, 2009) suggests that the presence or absence of audiovisual correspondence in our stimuli could have detected pretty quickly. Of course, a test of this possible explanation for our results will require coordinated behavioral and electrophysiological assays.

Note that we do not view audiovisual integration as some all-or-none process, that is, a process that is unfailingly triggered to full size by the detection of a match between signals from the two sensory streams. For example, detection of audiovisual correspondence varies with the temporal offset between the auditory and visual streams (Denison, Driver, & Ruff, 2013) and varies in a continuous fashion with small changes in the relative rates at which visual and auditory pulses are delivered (Roach et al., 2006). Finally, two groups of researchers, using quite different stimuli and tasks, demonstrated that audiovisual integration varies with the details of each sensory stream's own temporal structure (Denison et al., 2013; Keller & Sekuler, 2015).

Changes in performance over trials

All the results presented to this point have entailed mean values aggregated across trials. Those values naturally elide evidence of something that might be of considerable interest, namely, any changes in performance that might occur over successive trials. Unfortunately, it was not feasible to examine changes on a trial-by-trial basis because with so few subjects, a change in just a single binary response would produce a large swing in the mean across subjects. For example, the mean for each trial in one of our experiments would reflect the outcome of just 10 occurrences, 1 from each of 10 subjects. Based on such a small number of responses, a change in just a single response (from correct to incorrect or vice versa) would shift the mean proportion correct by.20 (±.10). Therefore, we turned to a measure that was a compromise between a stable result and ability to capture possible changes in performance over trials. Specifically, we examined mean accuracy within successive sets of 10 trials for AV Congruent fish and for AV Incongruent fish. Recall that each of these conditions was presented in block randomized order, with two 80-trial blocks per condition. We examined successive 10-trial sets over the 160 AV Congruent and 160 AV Incongruent trials obtained for from subject. Visual inspection showed that results from Experiments 1 and 2 were similar, so we increased the reliability of our analysis by combining the two sets of results. Figure 8 shows the result of this averaging process. Mean proportions correct for AV Congruent fish are represented in the upper set of points; mean proportions correct for AV Incongruent fish shown as the lower set of points. Note that Experiment 3's results were omitted from this analysis because, unlike the first two experiments, its design lacked a Control condition.

In many situations, the trajectory of learning can be described by a power function (Kahana, 2012). So, in the absence of a strong a priori expectation for how performance would change over trials, a simple power function was fit to each data set. The best fitting power model parameter values and their associated confidence intervals were found using the *fit* function (in Matlab's Curve Fitting Toolbox), using unweighted least squares. The goodness of fit from a simple power series model, $f(x) = a * x^b$, was significantly better than that one produced with a linear fit, $f(x) = a * x$, as visual inspection of Figure 8 confirms. To determine whether an additional parameter would significantly improve the power function's fit, an F test compared the fit produced with the simple, single-term power function, $f(x) = a * x^b$, against the fit from a power function with an additional parameter, $f(x) = a * x^b + c$. Including that extra parameter produced no clear improvement in goodness of fit, either for the AV Congruent condition or the AV Incongruent condition, $F(1,3) = 4.17$, $p = .13$ and $F\ (1,3) = .54$, $p = .59$, respectively. As we do not claim that the simple power series model is *the* optimum model, we decided against a more extensive model selection process.

Table 1 gives the exponent values and their 95% CLs for the functions represented by the curves in Figure 8. Note that the exponents in both best fit functions have positive signs, with CLs that exclude zero. This confirms the increase in accuracy over trials. Importantly, the

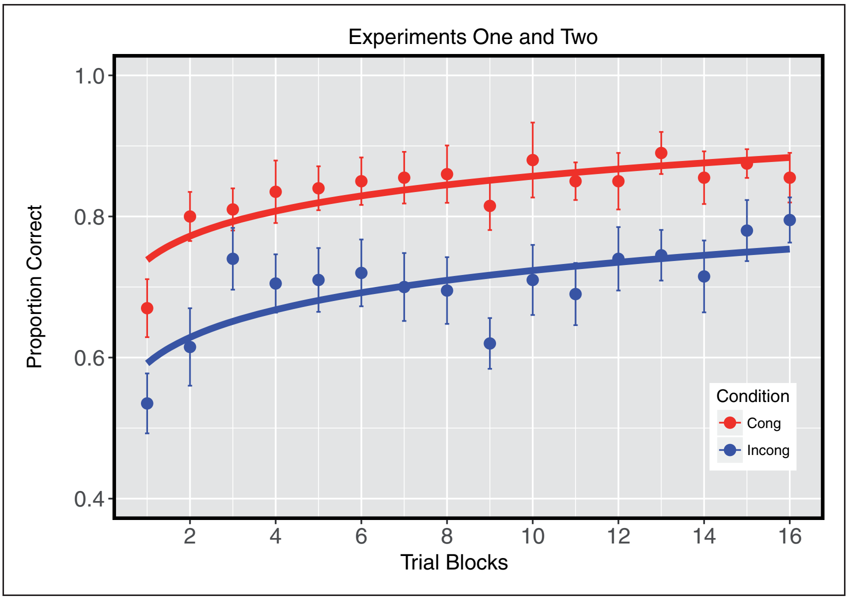

Figure 8. Proportion correct from successive 10-trial blocks for subjects in Experiments 1 and 2. Results are shown separately for AV Congruent fish (●) and for AV Incongruent fish (●). Error bars represent within-subject standard errors of the mean.

Table 1. Exponents From the Best-Fitting One-Parameter Power Functions.

Condition	Exponent	Exponent CLs	R^2
AV Congruent	.065	[.039,.090]	.673
AV Incongruent	.087	[.038,.136]	.499

Note. Also shown are the confidence limits on the exponents and Adjusted R2 values.

overlap between the CLs associated with the exponents for AV Congruent and for AV Incongruent conditions suggests that learning rates probably did not differ reliably between conditions. We examined this possibility further with an ANOVA that included orthogonal polynomial contrasts.

Factors in the ANOVA included 16 successive Trial Sets (10 trials in each set), Conditions (AV Congruent and AV Incongruent), and Subjects ($n = 20$). To increase the sensitivity of our analysis, the ANOVA isolated orthogonal polynomial contrasts (linear and quadratic) for the main effect of Trial Sets, and the interaction between that effect and Conditions. Table 2 summarizes the outcome of this analysis. Mauchly's test showed significant violations of sphericity for the main effect of Trial Sets, but not for any interaction involving that term. The Huynh–Feldt correction was applied to the main effect of Trial Sets and to its linear and quadratic contrasts. First, as expected, the difference between Conditions is highly significant, $F(1, 19) = 23.49$, $p < .0001$. Second, as also expected, the omnibus effect of Trial Sets is significant, $F(9.40,178.67) = 5.79$, $p < .001$. When that omnibus effect is decomposed into linear and quadratic components, the former is highly significant ($p < .003$), while the latter is not significant ($p < .26$). Importantly, the interaction Condition × Trials was not significant ($p = .77$). We can conclude, therefore, that the rates of learning for AV Congruent and for AV Incongruent conditions do not differ significantly.

To interpret the results shown in Figure 8, consider what information is required for correct responses (leaving aside lucky guesses). First, subjects must know the binary rule that links frequency of visual modulation to the response categories "good" fish and "bad" fish, and, of course, the keyboard responses assigned to each category. Second, for AV Congruent fish, the amplitude modulated sound that accompanies the fish must be associated with its visual modulation. Third, subjects certainly had to exploit visual information in order to categorize a fish's visual modulation rate as 6 or 8 Hz.

The first source of information mentioned earlier, information about response mapping, is unlikely to have played a major role in improved performance over trials. For one thing, during practice and before any trial represented in Figure 8, every subject had to satisfy the criterion of 75% success with a series of AV Congruent fish. AV Congruent fish were used in these practice trials because we knew they would make rate of oscillation easiest to discriminate (Goldberg et al., 2015). The level of success achieved by every single subject during the practice trials would have been unlikely unless subjects understood the binary response rule, that is, the mapping of oscillation rate onto response category ("good" or "bad" fish). Undoubtedly, the feedback that immediately followed each response helped subjects acquire that response rule.

As Figure 8 showed, performance with AV Congruent fish was consistently better than performance with AV Incongruent fish: accuracy was higher and RTs were shorter. To understand this result, it is important to recall that in both Experiments 1 and 2, performance with AV Incongruent fish and with Control fish did not differ from one another. This pattern of results with all three conditions makes clear the origin of

Table 2. ANOVA on Learning Results.

Effect	MS	*df*	*F*	*p*
Trial Set	0.075	9.403	5.787	.0001
Trial Set:linear	0.556	0.627	42.64	.0009
Trial Set:quadratic	0.073	0.627	5.596	.2150
Condition	2.957	1	23.49	1.12 e-5
Trial Set:Condition	0.014	15	0.714	.7700

superior performance with AV Congruent fish. In particular, superior performance represents a benefit from the combination of audio and visual signals whose rates of modulation are matched and does not represent a performance cost when signals are mismatched (with AV Incongruent fish).

We believe that the changes in performance seen in Figure 8 reflect increasing discriminability of the rate at which fish size oscillates. This improved discriminability may be a form of perceptual learning. Such learning has been studied for decades, with many different unimodal stimuli and tasks (Hussain, McGraw, Sekuler, & Bennett, 2012; Karni & Sagi, 1993; Watanabe & Sasaki, 2014), but only recently has it been examined in a multisensory context. Several groups have demonstrated that when subjects are trained with stimuli that comprise both auditory and visual signals, discrimination of visual stimuli improves more than it does when subjects are trained with stimuli with unisensory, visual stimuli (Barakat et al., 2015; Kim, Seitz, & Shams, 2008; Seitz, Kim, & Shams 2006; Zilber, Ciuciu, Gramfort, Azizi, & Van Wassenhove, 2014). Although the stimuli and tasks used by those researchers differ from one in *Fish Police!!*, perceptual learning in all these multimodal context might reflect the influence of some general, supramodal principle. As Zilber et al. (2014) noted, such a supramodal principle could fit well within the framework of the Reverse Hierarchy Theory of learning (Ahissar, Nahum, Nelken, & Hochstein, 2009).

Future Research

In all three of our experiments, subjects were instructed to base their judgments solely on a fish's *visual* behavior. That instruction notwithstanding, performance with AV Congruent fish revealed that a concurrent sound could impact subjects' judgments. That result raises the question of whether there is a comparable effect in the opposite direction. That is, can the same *visual* attribute (rate of size modulation) alter judgments of the same *auditory* attribute (rate of a sound's amplitude modulation)? An early study of audiovisual interaction showed that a train of clicks dramatically altered the perceived frequency of visual flicker, producing as much as a two-fold change in perceived flicker rate (Shipley, 1964). Others subsequently demonstrated an effect in the opposite direction, but one that was smaller than what Shipley reported (Roach et al., 2006; Welch & Warren, 1980; Welch et al., 1986).

These and other studies remind us that the magnitude of AV interactions depends upon many variables, including the degree of match between modulation rates of auditory and visual signals (Roach et al., 2006), the apparent colocation of signals in the two modalities (Heron, Roach, Hanson, McGraw, & Whitaker, 2012), the statistical reliability of each signal (Ernst & Banks, 2002; Sheppard, Raposo, & Churchland, 2013), as well as differences in the weights that individual subjects place on one modality versus the other (Keller & Sekuler, 2015).

Like other video games constructed for psychophysical purposes (e.g., Abramov et al., 1984; Anguera et al., 2013; Greenwood et al., 2012; Wade & Holt, 2005), elements of *Fish Police!!* were designed to enhance subjects' enjoyment and engagement. We believe the game succeeded in that at least to some degree. For example, when a handheld tablet-based version of *Fish Police!!* was deployed at Boston's Museum of Science, potential subjects were willing to endure a long wait for a chance to play, were eager to compare their scores against those of other players, and many asked for a chance to play again (Goldberg et al., 2015).

However, we recognize that our implementation of *Fish Police!!* lacks key features that make video games compelling and engaging (Morris, Croker, Zimmerman, Gill, & Romig, 2013). For example, *Fish Police!!* violates a principle of good game design by failing to insure "that the difficulty level varies so the players experience greater challenges as they develop mastery" (Stráát, Rutz, & Johnansson, 2014, p. 49). Holding task difficulty constant throughout an experiment does make it possible to gauge learning over trials, but that design decision likely fails to maximize the engagement that could have come from systematic, subject-driven titration of task difficulty. Researchers who want to use games for psychophysical purposes must rely on ad hoc decisions about how to balance the requirements of strict experimental control and repeatability of test conditions, on one hand, against the advantages of a task that engages subjects by introducing stimulus variability and by allowing subjects control over their own test conditions, on the other. We expect that over time, with trial and error, guidelines will be developed for achieving the right balance.

Acknowledgements

We are grateful to Fatima Abu Deeb, Nick Moran, and Jonathan Chu for programming assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article was supported by CELEST, a National Science Foundation Science of Learning Center (SBE-0354378).

Note

1. The game controller divided each 200-trial session into four nominal blocks, which were presented without interruption in between. As this arrangement made the block structure entirely transparent to subjects, it was ignored for purposes of analysis.

References

Abramov, I., Hainline, L., Turkel, J., Lemerise, E., Smith, H., Gordon, J., & Petry, S. (1984). Rocketship psychophysics. Assessing visual functioning in young children. *Investigative Ophthalmology & Visual Science*, *25*, 1307–1315.

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *364*, 285–299.

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., . . . ,Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*, 97–101.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384.

Barakat, B., Seitz, A. R., & Shams, L. (2015). Visual rhythm perception improves through auditory but not visual training. *Current Biology*, *25*, R60–R61.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*, 1190–1192.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*, 95–98.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*, e1000436.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, England: Cambridge University Press.

Denison, R. N., Driver, J., & Ruff, C. C. (2013). Temporal structure and complexity affect audio-visual correspondence detection. *Frontiers in Psychology*, *3*, 619.

Donohue, S. E., Green, J. J., & Woldorff, M. G. (2015). The effects of attention on the temporal integration of multisensory stimuli. *Frontiers in Integrative Neuroscience*, *9*, 32.

Eckstein, M. P., Mack, S. C., Liston, D. B., Bogush, L., Menzel, R.Krauzlis, R. J. (2013). Rethinking human visual attention: Spatial cueing effects and optimality of decisions by honeybees, monkeys and humans. *Vision Research*, *85*, 5–19.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fitts, P. M., & Deininger, R. L. (1954). S-R compatibility: Correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology*, *48*, 483–492.

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research*, *166*, 455–464.

Goldberg, H., Sun, Y., Hickey, T. J., Shinn-Cunningham, B. G., & Sekuler, R. (2015). Policing fish at Boston's Museum of Science: Studying audiovisual interaction in the wild. *iPerception*, *6*, doi:10.1177/2041669515599332

Greenwood, J. A., Tailor, V. K., Sloper, J. J., Simmers, A. J., Bex, P. J.Dakin, S. C. (2012). Visual acuity, crowding, and stereo-vision are linked in children with and without amblyopia. *Investigative Ophthalmology & Visual Science*, *53*, 7655–7665.

Guttman, S. E., Gilroy, L. A., & Blake, R. (2005). Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science*, *16*, 228–235.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*, 186–201.

Heron, J., Roach, N. W., Hanson, J. V. M., McGraw, P. V., & Whitaker, D. (2012). Audiovisual time perception is spatially specific. *Experimental Brain Research*, *218*, 477–485.

Hussain, Z., McGraw, P. V., Sekuler, A. B., & Bennett, P. J. (2012). The rapid emergence of stimulus specific perceptual learning. *Frontiers in Psychology*, *3*, 226.

Joris, P. X., Schreiner, C. E., & Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological Reviews*, *84*, 541–577.

Kahana, M. J. (2012). *Foundations of human memory*. New York, NY: Oxford University Press.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, *365*, 250–252.

Kawachi, Y., Grove, P. M., & Sakurai, K. (2014). A single auditory tone alters the perception of multiple visual events. *Journal of Vision*, *14*, 16.

Kayser, C., & Remedios, R. (2012). Suppressive competition: How sounds may cheat sight. *Neuron*, *73*, 627–629.

Keller, A. S., & Sekuler, R. (2015). Memory and learning with rapid audiovisual sequences. *Journal of Vision*, *15*, 7.

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS One*, *3*, e1532.

Lawrence, M. A. (2013). *ez: Easy analysis and visualization of factorial experiments*. Retrieved from http://cran.r-project.org/web/packages/ez/

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *Elife*, *4*. doi:10.7554/eLife.04995

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Morris, B. J., Croker, S., Zimmerman, C., Gill, D., & Romig, C. (2013). Gaming science: The ''gamification'' of scientific thinking. *Frontiers in Psychology*, *4*, 607.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W.Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, *25*, 499–507.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447.

Parise, C. V., Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisensory Research*, *26*, 307–316.

Pollack, I. (1974). Within- and between-modality correlation detection. *The Journal of the Acoustical Society of America*, *55*, 641–644.

Pollack, I. (1975). Identification of temporal coherence between auditory and visual channels. *Perception & Psychophysics*, *17*, 277–284.

Rainville, S., & Clarke, A. (2008). Distinct perceptual grouping pathways revealed by temporal carriers and envelopes. *Journal of Vision*, *8*, 9.1–9.15.

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings. Biological Sciences/The Royal Society*, *273*, 2159–2168.

Rosenthal, O., Shimojo, S., & Shams, L. (2009). Sound-induced flash illusion is resistant to feedback training. *Brain Topography*, *21*, 185–192.

Roudaia, E., Sekuler, A. B., Bennett, P. J., & Sekuler, R. (2013). Aging and audio-visual and multi-cue integration in motion. *Frontiers in Psychology*, *4*, 267.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, *16*, 1422–1427.

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

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*, 791–806.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*, 788–788.

Sheppard, J. P., Raposo, D., & Churchland, A. K. (2013). Dynamic weighting of multisensory stimuli shapes decision-making in rats and humans. *Journal of Vision*, *13*, pii: 4. doi: 10.1167/13.6.4

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science*, *145*, 1328–1330.

Spence, C., & Deroy, O. (2012). Hearing mouth shapes: Sound symbolism and the reverse McGurk effect. *i-Perception*, *3*, 550–552.

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology*, *13*, R519–R521.

Strååt, B., Rutz, F., & Johnansson, M. (2014). Does game quality reflect heuristic evaluation? *International Journal of Gaming and Computer-Mediated Simulations*, *6*, 45–58.

Tabry, V., Zatorre, R. J., & Voss, P. (2013). The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in Psychology*, *4*, 932.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*, 679–690.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 830–846.

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

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *Journal of the Acoustical Society of America*, *118*, 2618–2633.

Watanabe, T., & Sasaki, Y. (2014). Perceptual learning: Towards a comprehensive theory. *Annual Review of Psychology*, *66*, 197–221.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*, 638–67.

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics*, *39*, 294–300.

Winkler, I., Horváth, J., Weisz, J., & Trejo, L. J. (2009). Deviance detection in congruent audiovisual speech: Evidence for implicit integrated audiovisual memory representations. *Biological Psychology*, *82*, 281–292.

Wundt, W. (1894) *Lectures on Human and Animal Psychology*. Translated by Creighton, J. E. and Titchener, E.B. New York: Macmillan & Co.

Zilber, N., Ciuciu, P., Gramfort, A., Azizi, L., & Van Wassenhove, V. (2014). Supramodal processing optimizes visual perceptual learning and plasticity. *Neuroimage*, *93*, 32–46.