Attention, Perception, & Psychophysics
2010, 72 (7), 1736-1741
doi:10.3758/APP.72.7.1736

Characteristic sounds make you look at target objects more quickly

Lucica Iordanescu, Marcia Grabowecky, and Steven Franconeri
Northwestern University, Evanston, Illinois

Jan Theeuwes
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

and

Satoru Suzuki
Northwestern University, Evanston, Illinois

When you are looking for an object, does hearing its characteristic sound make you find it more quickly? Our recent results supported this possibility by demonstrating that when a cat target, for example, was presented among other objects, a simultaneously presented "meow" sound (containing no spatial information) reduced the manual response time for visual localization of the target. To extend these results, we determined how rapidly an object-specific auditory signal can facilitate target detection in visual search. On each trial, participants fixated a specified target object as quickly as possible. The target's characteristic sound speeded the saccadic search time within 215–220 msec and also guided the initial saccade toward the target, compared with presentation of a distractor's sound or with no sound. These results suggest that object-based auditory–visual interactions rapidly increase the target object's salience in visual search.

Sounds facilitate visual localization on the basis of spatial coincidence. For example, a sound coming from the location of a visual target facilitates its detection (see, e.g., Bolognini, Frassinetti, Serino, & Làdavas, 2005; Driver & Spence, 1998; Stein, Meredith, Huneycutt, & McDade, 1989). A sound also facilitates visual localization on the basis of temporal coincidence when a visual target has unique dynamics (compared with distractors) and a sound is synchronized to the target's dynamics (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008).

In addition to these well-established spatial and temporal auditory–visual interactions, neuroimaging results suggest that auditory–visual interactions also occur in an object-specific manner in polysensory areas in the temporal cortex (see, e.g., Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Molholm, Ritter, Javitt, & Foxe, 2004; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). It is therefore possible that feedback from polysensory areas to visual areas could speed visual processing in an object-specific manner. Consistent with this possibility, behavioral responses to target objects are faster when the target object (e.g., a cat) is presented together with its characteristic sound (e.g., a "meow" sound) for recognizing the visual target (Molholm et al., 2004) and for localizing the target among distractor objects (Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008).

Because these studies used manual responses (via keypresses), however, it was not possible to directly demonstrate that characteristic sounds facilitated perception of the target object. Manual response times include additional processes, such as confirming the identity of the target object, mapping the perceptual decision to an arbitrarily defined motor response, and executing the motor response. The present study was designed to circumvent these confounds associated with manual responses in order to more directly demonstrate that hearing a characteristic sound of an object facilitates its visual localization.

We used saccades as the mode of response in the context of visual search. Because people naturally look at objects of interest, asking participants to quickly fixate targets does not require an arbitrary response mapping. We measured the time it took for participants to saccade to the target object. It has been shown that even when a target location is known, it typically takes 150–350 msec (averaging 200–250 msec) to initiate a saccade (see, e.g., Darrien, Herd, Starling, Rosenberg, & Morrison, 2001; Yang, Bucci, & Kapoula, 2002). Thus, if we could obtain significant speeding of saccades by characteristic sounds for fast saccadic responses (<250 msec), we could reasonably conclude that characteristic sounds rapidly facilitate the process of target selection during the initial engagement of attention. Furthermore, the result would provide an upper estimate of how rapidly object-based

S. Suzuki, satoru@northwestern.edu

auditory–visual neural interactions (potentially mediated by temporal polysensory areas) influence the retinotopic visual processing required for target localization.

METHOD

Participants

Sixteen undergraduate students at Northwestern University gave informed consent to participate for partial course credit. They all had normal or corrected-to-normal visual acuity, had normal hearing, and were tested individually in a normally lit room.

Stimuli

Each search display (see Figure 1A for an example) contained eight colored pictures of common objects (each confined within a 5.14º × 5.11º rectangular region). The centers of the eight pictures were placed along an approximate isoacuity ellipse (20º horizontal × 15º vertical, an aspect ratio based on Rovamo & Virsu, 1979). One of these pictures was the target, and the remaining pictures were the distractors. Search stimuli (some with backgrounds) and their characteristic sounds were selected from a set of 20 objects (bike, bird, car, cat, clock, coins, dog, door, running faucet, keys, kiss, lighter, mosquito, phone, piano, stapler, lightning, toilet, train, and wine glass; see Iordanescu et al., 2008, for the full set of images). We avoided inclusion of objects with similar characteristic sounds (e.g., keys and coins) within the same search display. The durations of characteristic sounds varied because of differences in their natural durations (M = 862 msec, SD = 451; all sounds < 1,500 msec). These heterogeneities should not have affected our measurement of auditory–visual interactions, however, because our design was fully counterbalanced (see below). The sounds were clearly audible (~70 dB SPL), presented via two loudspeakers, one on each side of the display monitor; the sounds carried no information about the target's location.

On each trial, the sound was consistent with the target object (*target consistent*), consistent with a distractor object (*distractor consistent*), or absent (*no sound*). In the distractor-consistent condition, the relevant distractor object was always presented in the quadrant diagonally opposite the target across the fixation marker, so that any potential cross-modal enhancement of the distractor did not direct attention near the target. Within a block of 60 trials, each of the 20 sounds was presented once as the target-consistent sound and once as the distractor-consistent sound (with sounds absent in the remaining 20 trials), and each picture was presented as the target once in each of the three sound conditions. This counterbalancing ensured that any facilitative effect of target-consistent sounds would be attributable to the sounds' associations with the visual targets, rather than to the properties of the pictures or the sounds themselves. Aside from these constraints, the objects were randomly selected and placed on each trial. Each participant was tested in four blocks of 60 trials. Ten practice trials were given prior to the experimental trials.

The stimuli were displayed on a color CRT monitor (1,024 × 768 pixels) with a 60-Hz refresh rate, and the experiment was controlled by a Sony VAIO computer using MATLAB (The MathWorks, Inc., Natick, MA) and PsychToolbox software (Brainard, 1997; Pelli, 1997). An EyeLink 1000 Tower Mount eyetracker (1000-Hz sampling rate and 0.25º spatial resolution) with a combined chin- and forehead rest was used to monitor eye movements and to stabilize the viewing distance at 81 cm. Onsets and offsets of saccades were detected using the EyeLink software, which uses saccade-detection criteria that are based on thresholds for eye-position shift (0.1º), velocity (30 deg/sec), and acceleration (8,000 deg/sec) in conjunction with the general algorithm described in Stampe (1993).

Procedure

Participants looked at a central circle (1º radius) to begin each trial. The name of the current target (e.g., *cat*) was presented aurally at the beginning of each trial. After 2,000 msec, the search display appeared synchronously with the onset of one of the two types of sounds, target consistent or distractor consistent, or with no sound. Participants were instructed to look at the target as quickly as possible. As soon as the left eye-gaze position reached the 4.03º × 4.03º region of the target, the visual display was terminated, and the saccadic search time (measured from the onset of the search display) was recorded.

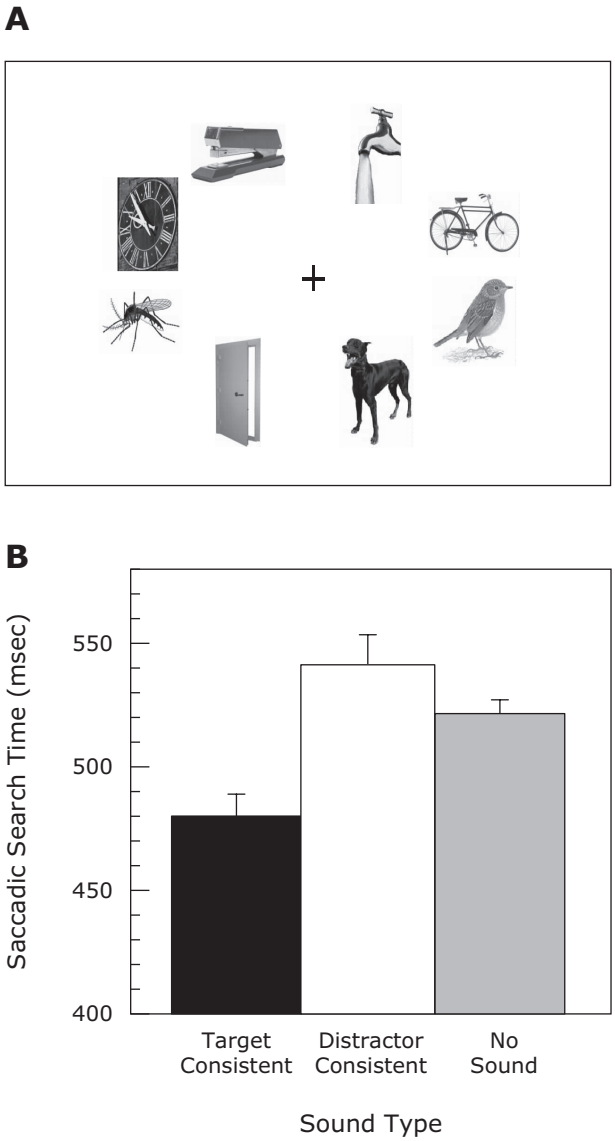

Figure 1. (A) An example of a search display; participants fixated the specified target object as quickly as possible. (B) Saccadic search times when a search display was presented simultaneously with a characteristic sound of the target object (target consistent), with a characteristic sound of a distractor object (distractor consistent), or with no sound. The error bars represent ±1 standard error of the mean (adjusted to be appropriate for the within-subjects design of the experiment).

RESULTS

Saccadic search time was significantly faster in the target-consistent condition (M = 480 msec) than in

both the distractor-consistent condition (M = 541 msec) [$t(15) = 2.967$, $p < .01$, $d = 0.742$][1] and the no-sound condition (M = 521 msec) [$t(15) = 4.113$, $p < .001$, $d = 1.028$]; saccadic search time did not differ between the distractor-consistent and no-sound conditions [$t(15) = 1.203$, n.s., $d = 0.301$] (Figure 1B). Playing the target object's characteristic sound thus speeded eye movements to the target in visual search.

We determined how rapidly characteristic sounds facilitated target selection by computing the proportion of saccadic search times in 5-msec bins and determining the earliest bin at which the target-consistent condition produced a significantly greater cumulative proportion than the distractor-consistent and no-sound conditions. For example, if the cumulative proportion for the target-consistent condition significantly exceeded those for the distractor-consistent and no-sound conditions at the 50th cumulative bin, that would indicate that the target-consistent sounds significantly increased the proportion of search times that were 250 msec and faster. We would then make a conservative inference that the target-consistent sounds facilitated visual search within 250 msec.

As is shown in Figure 2A, the cumulative proportion of fast saccadic search times was greater in the target-consistent condition than in both the distractor-consistent and no-sound conditions, and the distributions do not differ between the distractor-consistent and no-sound conditions; note that the vertical separations in the initial rising portions of the distributions are difficult to discern due to the steep slopes.

To more clearly illustrate how rapidly the object-specific auditory–visual interactions emerged over time, we plotted the difference between the distribution for the target-consistent condition and those for the distractor-consistent and no-sound conditions. The advantages for target-consistent sounds over distractor-consistent sounds (Figure 2B) and those for target-consistent sounds over no sound (Figure 2C) both rose rapidly after 190 msec. To determine how rapidly the advantages became statistically significant, we computed confidence limits using a bootstrapping method. Under the null hypothesis, saccadic search times from all three conditions would come from the same distribution for each participant. To estimate the extent of condition effects expected from sampling error (under the null hypothesis), we combined data from the three conditions into one saccadic search time distribution and randomly sampled from that distribution to simulate the data for the three conditions for each participant. We then pooled the simulated data from all participants in exactly the same way that we pooled the actual data, as shown in Figures 2B and 2C. We repeated this procedure 5,000 times to compute the 2.5th and 97.5th percentile points, which are shown as the lower and upper limits of the 95% confidence intervals (the gray regions) in Figures 2B and 2C. The details of this bootstrapping analysis are provided in the Appendix. The advantages of target-consistent sounds over distractor-consistent sounds and over no sound exceeded the 95% confidence limits at the latencies of 220 and 215 msec, respectively.

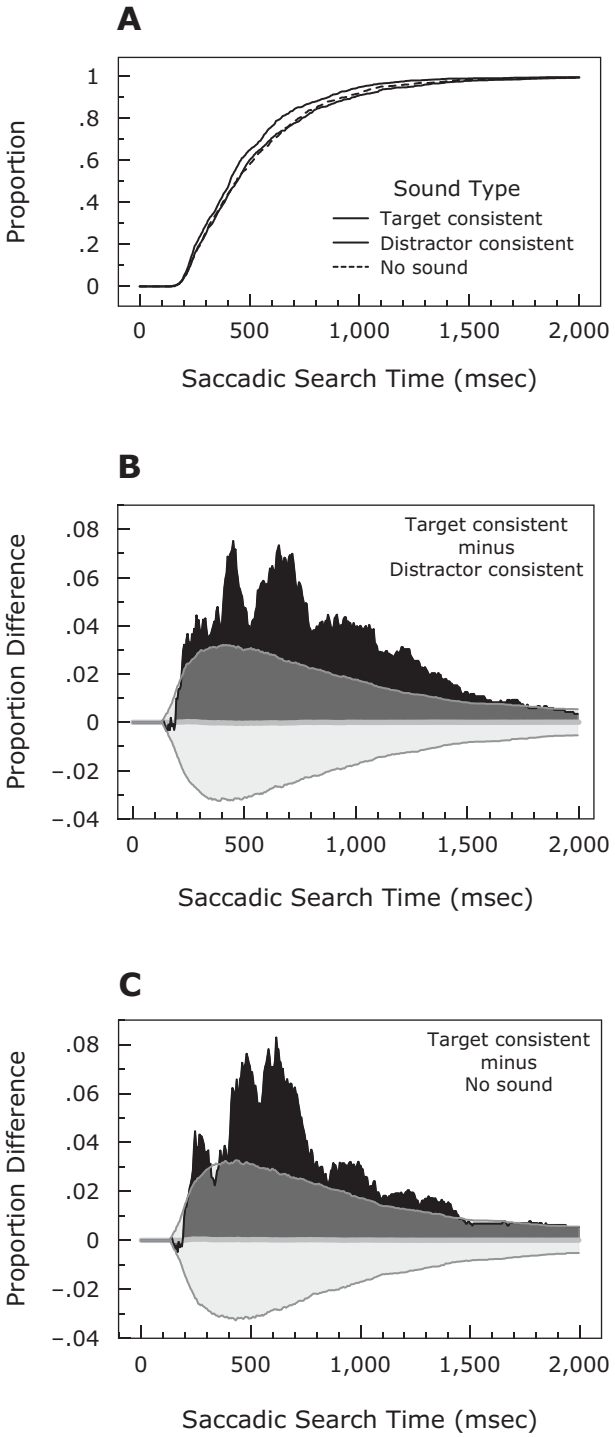

Figure 2. (A) Cumulative distributions of saccadic search times for trials with target-consistent sounds (thick solid curve), distractor-consistent sounds (thin solid curve), and no sound (thin dashed curve). (B) The difference between the cumulative distribution for trials with target-consistent sounds and the cumulative distribution for trials with distractor-consistent sounds. (C) The difference between the cumulative distribution for trials with target-consistent sounds and the cumulative distribution for trials with no sound. In B and C, the translucent gray regions indicate the 95% confidence limits (see the main text and the Appendix for details).

To find converging evidence for the rapid influences of object-based auditory–visual interactions, we also analyzed the impact of characteristic sounds on the trajectory of initial saccades (defined as the first saccade that the participant made following the onset of a search display). We determined whether target-consistent sounds guided initial saccades toward the target compared with distractor-consistent sounds and with no sound. We quantified the degree to which an initial saccade moved the eyes toward the target by computing the projection of its vector on the axis determined by the fixation point and the target. A larger positive value would indicate that the eyes initially moved closer to the target, and a larger negative value would indicate that the eyes initially moved farther away from the target. If a target-consistent sound had an impact on the direction of the initial saccade, the projection value should be significantly greater in the target-consistent condition than in the no-sound condition. Because a distractor-consistent sound was always associated with the distractor placed diagonally opposite the target, if a distractor-consistent sound had an impact on the direction of the initial saccade, the projection value should be significantly smaller in the distractor-consistent condition than in the no-sound condition.

The average projection values were positive for all conditions, indicating that an initial saccade moved the eyes toward the target overall. The projection value was significantly greater in the target-consistent condition than in both the no-sound [$t(15) = 2.912, p < .02, d = 0.728$] and distractor-consistent [$t(15) = 3.740, p < .002, d = 0.935$] conditions, whereas the projection values were not significantly different between the distractor-consistent and no-sound conditions [$t(15) = 1.137$, n.s., $d = 0.284$]. Thus, target-consistent sounds guided initial saccades toward the targets, whereas distractor-consistent sounds had no significant impact on the trajectory of initial saccades.

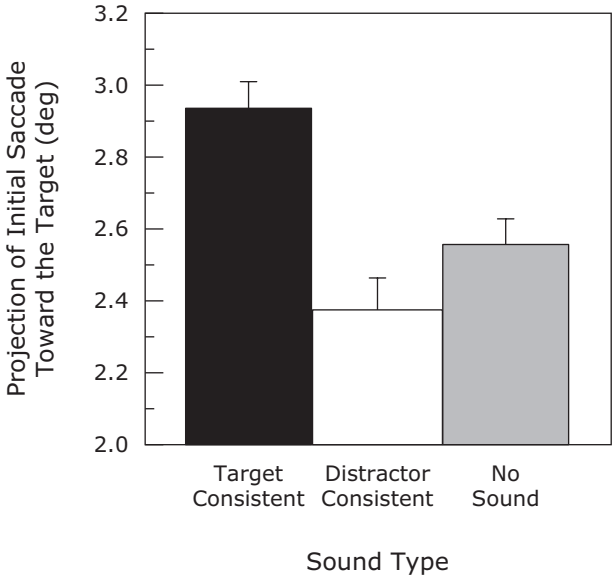

Figure 3. Average projections of initial-saccade vectors in the target direction when a search display was presented simultaneously with a characteristic sound of the target object (target consistent), with a characteristic sound of a distractor object (distractor consistent), or with no sound. A larger positive value indicates that the initial saccade moved the eyes closer to the target (see the main text for details). The error bars represent ±1 standard error of the mean (adjusted to be appropriate for the within-subjects design of the experiment).

DISCUSSION

We investigated how quickly people looked at a target object presented among distractor objects when a sound characteristic of the target object, a sound characteristic of a distractor object, or no sound was presented concurrently with the search display. All of our measures—mean saccadic search times (Figure 1B), cumulative distributions of saccadic search times (Figure 2), and trajectories of initial saccades (Figure 3)—provided converging evidence, indicating that playing a characteristic sound of a target object guides and speeds saccades to the target, whereas playing a sound associated with a distractor has little impact.

The lack of a measurable effect of distractor-consistent sounds in this study is consistent with previous results (see, e.g., Iordanescu et al., 2008; Molholm et al., 2004; von Kriegstein et al., 2005), suggesting that object-based auditory–visual enhancements occur in a goal-directed manner. Because neurons in the prefrontal cortex selectively respond to task-relevant stimuli (see, e.g., Duncan, 2001; Miller & Cohen, 2001), and some neurons there respond to both auditory and visual stimuli (see, e.g., Watanabe, 1992), the locus of the cross-modal effect in our study might have been the prefrontal cortex; however, object selectivity in the prefrontal cortex might be too weak (see, e.g., Warden & Miller, 2007) to guide the search mechanisms to specific objects. Moreover, because responses of prefrontal neurons are task dependent (see, e.g., Asaad, Rainer, & Miller, 2000; Rainer, Asaad, & Miller, 1998), it is unclear how their responses would be affected by the characteristic sounds in our study, in which the sounds were task irrelevant in that they were uninformative of target location and were consistent with target identity only one third of the time. Alternatively, a target-specific auditory–visual enhancement might arise from a combination of top-down sensitization and cross-modal interaction. For example, a top-down signal, likely from the prefrontal cortex (see, e.g., Desimone & Duncan, 1995; Duncan, 2001; Miller & Cohen, 2001; Reynolds & Chelazzi, 2004), could sensitize visual representations of the target object, and a target-consistent sound would cross-modally boost activation of this sensitized representation. A distractor-consistent sound would have little effect, because the corresponding visual representation would not be sensitized by the top-down signal. The locus of the sensitized representation might be visual object-processing areas or polysensory areas in the temporal lobe (see, e.g., Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp, Argall, et al., 2004; Beauchamp, Lee, et al., 2004).

For a characteristic sound to influence saccadic latency and trajectory, the complex auditory signal must be pro-

cessed at the level of encoding sounds of common objects, the auditory and visual processing must interact at the level of object-based processing (potentially in temporal polysensory areas or the prefrontal cortex), and then feedback interactions must enhance the retinotopic representation of the target object to facilitate an eye movement to it. These processes would be time consuming if they proceeded serially. An electroencephalographic study examining auditory–visual interactions in visual object recognition showed that a characteristic sound (e.g., a "moo" sound presented with a picture of a cow) enhanced a visual-selection-related ERP signal within 210–300 msec (Molholm et al., 2004). The fact that we demonstrated the effect of target-consistent sounds on saccades within 215–220 msec suggests that object-based auditory–visual interactions influence behavior as rapidly as they modulate an ERP correlate. The rapid impact of characteristic sounds on saccades is even more impressive if one considers the fact that eye movements to even a single predictable target take 150–350 msec (see, e.g., Darrien et al., 2001; Yang et al., 2002). Our results are thus consistent with the emerging view that sensory processing is fundamentally multimodal, with cross-modal neural interactions influencing all levels of sensory processing, including those that were traditionally thought to be unimodal (see, e.g., Kayser & Logothetis, 2007; Schroeder & Foxe, 2005; Sperdin, Cappe, & Murray, 2010).

AUTHOR NOTE

This research was supported by National Institutes of Health Grant R01 EY018197 and National Science Foundation Grant BCS0643191. Correspondence concerning this article should be addressed to S. Suzuki, Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208-2710 (e-mail: satoru@northwestern.edu).

REFERENCES

Amedi, A., von Kriegstein, K., van Atteveldt, M. N., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research*, **166**, 559-571.

Asaad, W. F., Rainer, G., & Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, **84**, 451-459.

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.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, **41**, 809-823.

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. *Experimental Brain Research*, **160**, 273-282.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.

Darrien, J. H., Herd, K., Starling, L.-J., Rosenberg, J. R., & Morrison, J. D. (2001). An analysis of the dependence of saccadic latency on target position and target characteristics in human subjects. *BMC Neuroscience*, **2**, 13.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, **18**, 193-222.

Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, **2**, 254-262.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, **2**, 820-829.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sound facilitates visual search. *Psychonomic Bulletin & Review*, **15**, 548-554.

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure & Function*, **212**, 121-132.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167-202.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, **14**, 452-465.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.

Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, **393**, 577-579.

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, **27**, 611-647.

Rovamo, J., & Virsu, V. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, **37**, 475-494.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology*, **15**, 454-458.

Sperdin, H. F., Cappe, C., & Murray, M. M. (2010). The behavioral relevance of multisensory neural response interactions. *Frontiers in Neuroscience*, **4**, 9-18.

Stampe, D. M. (1993). Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavior Research Methods, Instruments, & Computers*, **25**, 137-142.

Stein, B. E., Meredith, M. A., Huneycutt, W. S., & McDade, L. (1989). Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli. *Journal of Cognitive Neuroscience*, **1**, 12-24.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception & Performance*, **34**, 1053-1065.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, **17**, 367-376.

Warden, M. R., & Miller, K. (2007). The representation of multiple objects in prefrontal neuronal delay activity. *Cerebral Cortex*, **17**, i41-i50.

Watanabe, M. (1992). Frontal units of the monkey coding the associative significance of visual and auditory stimuli. *Experimental Brain Research*, **89**, 233-247.

Yang, Q., Bucci, M. P., & Kapoula, Z. (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative Ophthalmology & Visual Science*, **43**, 2939-2949.

NOTE

1. Each effect size was computed by dividing the mean difference by the standard deviation of the difference scores, consistent with the within-subjects design of our experiment.

APPENDIX
Bootstrapping Analysis of Cumulative Distributions of Saccadic Search Times

In order to determine how rapidly characteristic sounds facilitated saccades to target objects, we compared the cumulative saccadic search time distribution for the target-consistent condition with those for the distractor-consistent and no-sound conditions.

Comparing cumulative distributions neither imposes limits on temporal resolution (beyond the measurement error) nor introduces a potential artifact of bin size. A general disadvantage of using cumulative distributions is that beyond the earliest point at which distributions for different conditions diverge, subsequent differences are difficult to interpret because they include earlier differences. Thus, cumulative distributions would not be suitable, for example, for determining whether saccadic search times differ among conditions for a specific time interval (say, between 400 and 500 msec). However, cumulative distributions are ideal for determining the earliest time point at which saccadic search time distributions from different conditions begin to diverge.

To statistically evaluate the distribution differences, we computed confidence intervals. Note that it would be inappropriate to compute confidence intervals in a conventional way, on the basis of the interparticipant variability at each time point. Although the overall shape of a probability distribution is free to vary, different time points along each distribution are "yoked." For example, if values in the lower half of the distribution are frequent, values in the upper half of the distribution must be infrequent. Consequently, it would be inappropriate to assume that each time point contributes an independent source of variability when comparing probability distributions. We thus evaluated the experimental differences in the distribution shapes between the target-consistent condition and the distractor-consistent and no-sound conditions against the range of random differences expected under the null hypothesis, using a bootstrapping method.

For each participant, we combined all of his or her saccadic search times into a single distribution, assuming the null hypothesis that the sound conditions made no difference. We then randomly sampled from this distribution (with replacement) as many times as the number of saccadic search times in each condition, to simulate the distribution of the participant's saccadic search times for each of the three sound conditions under the null hypothesis. We converted these simulated saccadic search time distributions into cumulative distributions. We then computed the differences between these simulated cumulative distributions, one between the simulated distributions for the target-consistent and distractor-consistent conditions, and the other between the simulated distributions for the target-consistent and no-sound conditions. These two difference distributions were calculated for all participants and were then averaged across participants to generate a pair of average difference-distribution curves (expected under the null hypothesis) comparable to those shown in Figures 2B and 2C.

To estimate the confidence limits on the variability of these difference distributions under the null hypothesis, we repeated the above procedure 5,000 times, yielding 5,000 simulated average difference distributions of each type (i.e., target-consistent condition minus distractor-consistent condition, or target-consistent condition minus no-sound condition). The 97.5th and 2.5th percentile values of these simulated distributions were used as the upper and lower limits of our 95% confidence intervals.

(Manuscript received March 21, 2010;
accepted for publication May 27, 2010.)