Published in final edited form as:
Psychon Bull Rev. 2008 June ; 15(3): 548–554.

Characteristic sounds facilitate visual search

Lucica Iordanescu and
Northwestern University, Evanston, Illinois

Emmanuel Guzman-Martinez
Northwestern University, Evanston, Illinois and Universidad Nacional Autónoma de México, Mexico City, Mexico

Marcia Grabowecky and **Satoru Suzuki**
Northwestern University, Evanston, Illinois

Abstract

In a natural environment, objects that we look for often make characteristic sounds. A hiding cat may meow, or the keys in the cluttered drawer may jingle when moved. Using a visual search paradigm, we demonstrated that characteristic sounds facilitated visual localization of objects, even when the sounds carried no location information. For example, finding a cat was faster when participants heard a meow sound. In contrast, sounds had no effect when participants searched for names rather than pictures of objects. For example, hearing "meow" did not facilitate localization of the word *cat*. These results suggest that characteristic sounds cross-modally enhance visual (rather than conceptual) processing of the corresponding objects. Our behavioral demonstration of object-based cross-modal enhancement complements the extensive literature on space-based cross-modal interactions. When looking for your keys next time, you might want to play jingling sounds.

When auditory and visual signals carry redundant sensory information, their integration can increase signal strength and reliability (due to the fact that environmental and sensory noise are often uncorrelated across the two modalities). Prior research on auditory–visual interactions has predominantly focused on the fact that both modalities carry information about location. The superior colliculi and posterior parietal cortex contain neurons that respond to both visual and auditory stimuli with spatially overlapping receptive fields (e.g., Andersen, Snyder, Bradley, & Xing, 1997; Stein, 1998), providing neural substrates for integrated spatial representations. Behavioral relevance of these multimodal neural representations of space has been demonstrated by showing that visual detection is enhanced when a sound is simultaneously presented at the location of a visual target (e.g., Bolognini, Frassinetti, Serino, & Làdavas, 2005; Driver & Spence, 1998; Frassinetti, Bolognini, & Làdavas, 2002; Stein, Meredith, Huneycutt, & McDade, 1989).

Whereas spatial integration of auditory and visual signals is well understood, relatively few studies have examined object-based auditory–visual integration. In real life, auditory and visual signals redundantly indicate object identities as well as object locations. For example, when a cat meows, the spatially coincident sight and sound indicate the location of the cat, but at the same time, meow sounds and the visual features of cats together indicate that the object is a cat. It is thus reasonable to hypothesize that auditory and visual signals are integrated in object processing as well as in spatial processing. Indeed, polysensory areas in the temporal cortex —for example, the superior temporal sulcus (STS)—are strongly activated when pictures of

Correspondence concerning this article should be addressed to S. Suzuki, Northwestern University, Department of Psychology, 2029 Sheridan Rd., Evanston, IL 60208 (e-mail: satoru@northwestern.edu)..

objects and their characteristic sounds are presented simultaneously instead of separately (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004). Furthermore, the fusiform face area is activated during recognition of familiar voices (e.g., von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). Although these brain imaging results suggest the existence of multimodal neural representations for coding objects, few behavioral studies have investigated their perceptual consequences.

In one prior study (Molholm, Ritter, Javitt, & Foxe, 2004), participants responded to a target animal (e.g., a cat) that could be presented visually (e.g., a picture of a cat), auditorily (e.g., a meow sound), or audiovisually (e.g., a picture of a cat and a meow sound together). Responses were fastest (beyond what was predicted by probability summation) when the target was presented audiovisually. This behavioral effect was accompanied by modulations of the ERP component associated with early visual object processing (N1), suggesting that object identification is facilitated when consistent information is provided by both visual and auditory modalities (Molholm et al., 2004). Another study showed that perceived femininity of a face was increased by concurrently presented pure (single frequency) tones in the female fundamental speaking-frequency range, whereas the perceived masculinity of a face was increased by concurrently presented pure tones in the male fundamental speaking-frequency range (Smith, Grabowecky, & Suzuki, 2007). Whereas these results demonstrate perceptual consequences of auditory–visual interactions in the context of identifying single objects, the present study investigated the possibility that object-based auditory–visual interactions might also facilitate the detection and localization of target objects in a cluttered environment.

Specifically, we hypothesized that hearing characteristic sounds of target objects might facilitate visual search. For example, suppose you were looking for a cat. Multiple objects in the scene would compete for neural activation in the ventral visual pathway mediating object perception (see Kastner & Ungerleider, 2000, and Reynolds & Chelazzi, 2004, for reviews). Even if the cat was within view, it might not attract attention if it happened to be less salient than other objects. If characteristic sounds cross-modally facilitate visual activation for the corresponding objects, hearing the cat meow should enhance the activation of visual neurons responsive to cats, thereby speeding visual detection and localization of the cat.

In order to demonstrate cross-modal enhancements in visual search, however, care must be taken to ensure that the results could not be attributed to a response bias. For example, if we simply ask a participant to respond as to whether or not a cat target is present, a meow sound might bias the participant to prepare for a “cat-present” response, because the sound is closely associated with cats, and this bias might speed the “cat-present” response if a cat happens to be in the search display. In this scenario, characteristic sounds speed target detection by directly biasing a “target-present” response without necessarily interacting with visual representations. To dissociate cross-modal enhancements from a potential response bias, we asked our participants to respond to the location of the target. Because people typically wish to locate the object that they look for in real-life visual search, a target-localization task is potentially more ecologically relevant than a typical target-present/absent task. Because our sounds contained no information about the target's location, they could not generate a response bias for any specific location or response finger. Facilitation of visual *localization* by characteristic sounds (together with the necessary control experiments described in the Results section) would thus provide evidence that auditory processing of object-specific sounds cross-modally enhances visual activation of the corresponding objects.

Participants were told which object was the target (e.g., a cat, a set of keys) prior to each visual search trial. A search display consisted of four common objects presented in the four quadrants of the display (see the upper panel in Figure 1A for an example). One of these objects was the target, and participants indicated its location as quickly as possible. The search display was accompanied by a characteristic sound of the target object, by a characteristic sound of a distractor object, or by a sound unrelated to any of the objects in the search display. If characteristic sounds enhanced visual processing of associated objects, visual search should be faster when the sound is associated with the target than when the sound is associated with a distractor or with none of the objects in the display.

A related question we asked was whether the behavioral effect of auditory–visual facilitation interacted with goal-directed top-down feedback. As in real-life examples, and in most laboratory examples of visual search, the identity of the target object was known in our search task. It was thus expected that the visual processing of the target object would receive facilitative top-down feedback (see Reynolds & Chelazzi, 2004, for a review). Interestingly, an electrophysiological study demonstrated that object-based auditory–visual interactions depended on such goal-directed feedback. Specifically, the N1 component of the visual evoked potential (thought to reflect object processing) elicited in response to viewing an animal picture was modulated by a simultaneous presentation of the animal's characteristic sound. Crucially, however, this object-specific cross-modal effect occurred only when the animal was the target of a behavioral task (Molholm et al., 2004). Thus, the behavioral effect of characteristic sounds in visual search might also occur only for the targets for which cross-modal facilitation is combined with goal-directed top-down feedback. If this were the case, relevant auditory signals should selectively increase the visual salience of target objects but should not affect the visual salience of distractor objects. In other words, even if sounds associated with the target objects facilitated target localization, sounds associated with the distractor objects should not slow target localization, as compared with unrelated sounds.

METHOD

Participants

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

Stimuli

Each search display contained four colored pictures of common objects (scaled to fit within a 6.5° by 6.5° area) placed in the four quadrants at 4.7° eccentricity (center to center) (see the upper panel in Figure 1A for an example). One of these objects was the target, and the remaining objects were the distractors. These objects 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, thunder, toilet, train, and wine glass; see the Appendix for the images), for which we also obtained clips of their characteristic sounds. Some of the pictures had backgrounds; the durations of characteristic sounds also varied due to differences in their natural durations (M = 862 msec with SD = 451 msec, all sounds shorter than 1,500 msec). These heterogeneities, however, should not have affected our measurement of auditory–visual interactions, 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; and they carried no location information for the targets.

On each trial, the sound was consistent with the target object (*target consistent*), consistent with a distractor object (*distractor consistent*), or consistent with 1 of the set of 20 objects not included in the search display (*unrelated*). In the distractor-consistent condition, the relevant distractor object was always presented in the quadrant diagonally opposite from the target across the fixation marker so that any potential cross-modal enhancement of the distractor did not direct attention toward the target. Within a block of 60 trials, each of the 20 sounds was presented once as the target-consistent sound, once as the distractor-consistent sound, and once as the unrelated sound, and each picture was 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. We avoided inclusion of objects with similar sounds (e.g., keys and coins) within the same search display. Aside from these constraints, the objects were randomly selected and placed on each trial. Each participant was tested in four blocks of 60 trials; 10 practice trials were given prior to the experimental trials.

The stimuli were displayed on a color CRT monitor (1,024 × 768 pixels) at 75 Hz, and the experiment was controlled by a Macintosh PowerPC 8600 using Vision Shell software (Micro ML, Inc.). A chinrest was used to stabilize the viewing distance at 61 cm.

Procedure

Participants pressed the space bar to begin each trial. The name of the target (e.g., *cat*) was aurally presented at the beginning of each trial. After 1,070 msec, the search display appeared for 670 msec synchronously with the onset of one of the three types of sounds (target-consistent, distractor-consistent, or unrelated). Participants were instructed to indicate the location of the target as quickly as possible by pressing one of the four buttons (arranged in a square array) that corresponded to the quadrant in which the target was presented. Participants used the middle and index fingers of the left hand to respond to the upper left and lower left quadrants, and used the middle and index fingers of the right hand to respond to the upper right and lower right quadrants (responses were thus ideomotor compatible). Participants were also instructed to maintain eye fixation at a central circle (0.46° diameter) throughout each trial. Response times (RTs) and errors were recorded.

RESULTS

As shown in Figure 1A (lower panel), the target-consistent sounds speeded visual localization of the target objects, as compared with the distractor-consistent sounds [$t(21) = 3.96$, $p < .001$, $d = 0.84$] and unrelated sounds [$t(21) = 4.33$, $p < .0005$, $d = 0.92$]. The overall ANOVA was also significant [$F(2,42) = 13.86$, $p < .0001$, $\eta^2 = .40$]. There was no evidence of a speed–accuracy trade-off, because the error pattern [$F(2,42) = 0.93$, n.s.] mirrored the RT pattern. Thus, characteristic sounds facilitated target localization in visual search, even when the sounds lacked any location information. This suggests that sounds increased the visual salience of the targets in an object-specific manner.

Furthermore, the RT was no slower with the distractor-consistent sounds than with the unrelated sounds [$t(21) = 1.04$, n.s.], indicating that the distractor-consistent sounds did not measurably enhance the salience of the distractor objects. This suggests that object-specific auditory facilitation of visual salience occurs only for target objects that receive goal-directed top-down feedback (consistent with Molholm et al., 2004).

A question remained, however, about the processing level at which this cross-modal enhancement occurred. For example, a meow sound could have enhanced activation of visual neurons responsive to cat-related visual features, as we hypothesized. Alternatively, it could have enhanced activation of higher level semantic representations for the concept of cats. In either way, visual localization of the cat target could be facilitated by a meow sound. To answer this question, we recruited an additional 22 participants and conducted a control experiment identical to the primary experiment except that we replaced the pictures of objects with their names (see the upper panel in Figure 1B for an example of a name-search display). If object-specific cross-modal facilitation occurred at a semantic level, characteristic sounds should produce the same facilitative effects for object names as they did for object pictures. In contrast, if characteristic sounds interacted with visual object processing, the sounds should produce no effects on object names.

We confirmed the latter (see the lower panel in Figure 1B). Characteristic sounds had no influence on the name search [target-consistent sounds vs. distractor-consistent sounds, $t(21) = 0.01$, n.s., for RT, and $t(21) = 0.81$, n.s., for error rate; target-consistent sounds vs. unrelated sounds, $t(21) = 1.17$, n.s., for RT, and $t(21) = 0.35$, n.s., for error rate]. The overall ANOVAs were also not significant [$F(2,42) = 0.61$, n.s., for RT, and $F(2,42) = 0.41$, n.s., for error rate]. Characteristic sounds thus modulate visual activation, rather than semantic activation, of target objects.

The null effects of the sounds on the name search also ruled out the possibility that the effect of characteristic sounds on the picture search might have been due to their effects on working memory. Because the target was different on each trial, participants needed to briefly store the target identity in working memory on each trial. It was possible that the target-consistent sounds could have facilitated working memory for the target (e.g., reminding participants what target to look for), whereas the distractor-consistent and unrelated sounds could have disrupted this working memory (e.g., causing momentary confusion as to what target to look for). Because the working memory aspect of the experiment was identical for the picture search and name search, the lack of sound effects on the name search indicates that the effect of characteristic sounds obtained for the picture search cannot be due to an effect of the sounds on working memory.

Although we provided evidence of object-based auditory–visual interactions in visual search, a question still remained about whether the effect of characteristic sounds was due to facilitation of target localization by the target-consistent sounds, or to potentially distracting effects of the distractor-consistent and unrelated sounds. To evaluate these possibilities, we recruited an additional 22 participants to replicate the picture-search experiment, with the exception that the unrelated sounds were replaced with no sounds; thus, on each trial the participants heard the target-consistent sound, the distractor-consistent sound, or no sounds at all. If characteristic sounds facilitated target localization, the RT should be faster with the target-consistent sounds than with no sounds. In contrast, if distractor-consistent sounds interfered with target localization, the RT should be slower with the distractor-consistent sounds than with no sounds.

In addition, to confirm that the cross-modal effect generalizes to a larger search display, we increased the number of pictures in each display from four to eight. The centers of the eight pictures were placed along an approximate iso-acuity ellipse (21° horizontal by 16° vertical, the aspect ratio based on Rovamo & Virsu, 1979), with 2 pictures presented in each quadrant (the backgrounds of some of the pictures were cropped to avoid overlap of adjacent pictures). The participant's task was still to indicate the quadrant in which the target appeared. Other experimental details were also the same as in the primary experiment.

As shown in Figure 2, the target-consistent sounds speeded visual localization of the target objects, as compared with no sounds [$t(21) = 6.02$, $p < .0001$, $d = 1.28$] and the distractor-consistent sounds [$t(21) = 4.04$, $p < .001$, $d = 0.86$], but the distractor-consistent sounds did not slow visual localization of the target, as compared with no sounds [$t(21) = 0.23$, n.s.]. The overall ANOVA was significant [$F(2,42) = 12.81$, $p < .0001$, $\eta^2 = .38$]. There was no evidence of a speed–accuracy trade-off, because the error pattern [$F(2,42) = 1.88$, n.s.] mirrored the RT pattern.

We thus replicated the effect of characteristic sounds using a larger display size. Crucially, we demonstrated that the target-consistent sounds substantially speeded target localization, whereas the distractor-consistent sounds had no effect, as compared with no sounds.

DISCUSSION

We have demonstrated that characteristic sounds facilitate visual localization in an object-specific manner, and that this facilitation occurs at the level of visual object processing rather than at the level of semantic processing. Whereas both the neural mechanisms (e.g., Andersen et al., 1997; Stein, 1998) and perceptual consequences (e.g., Bolognini et al., 2005; Driver & Spence, 1998; Frassinetti et al., 2002; Stein et al., 1989) of auditory–visual interactions have been well established with respect to representations of space, our results add to the growing evidence that object representations are also fundamentally multimodal (e.g., Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp, Argall, et al., 2004; Beauchamp, Lee, et al., 2004; Molholm et al., 2004; Smith et al., 2007; von Kriegstein et al., 2005).

Because spatial cross-modal enhancements are mediated by multimodal neurons with spatially overlapped auditory and visual receptive fields (e.g., Bolognini et al., 2005; Frassinetti et al., 2002; Stein, 1998; Stein et al., 1989), our demonstration of object-specific cross-modal enhancements might imply the existence of multimodal neurons with overlapped object selectivity for visual and auditory stimuli. For example, a multimodal neuron might selectively respond to both an image of a cat and a meow sound. The polysensory areas in the temporal cortex (e.g., STS) may contain such object-selective multimodal neurons (e.g., Amedi et al., 2005; Beauchamp, Argall, et al., 2004; Beauchamp, Lee, et al., 2004). Alternatively, our results might reflect excitatory cross-modal neural interactions that are object specific (e.g., von Kriegstein et al., 2005). For example, auditory encoding of a meow sound might enhance visual processing of cat-related features via long-range excitatory neural connections. In either case, object-specific auditory–visual associations are likely to develop due to repeated bisensory coincidence. For example, you are likely to look at a cat when it meows; you are likely to look at the toilet when you flush it; you are likely to see two touching faces when you hear a kissing sound; and so on. It is plausible that visual responses to complex patterns and the coincident auditory responses to the characteristic accompanying sounds get associated through a Hebbian-type associative learning process (e.g., Hebb, 1949).

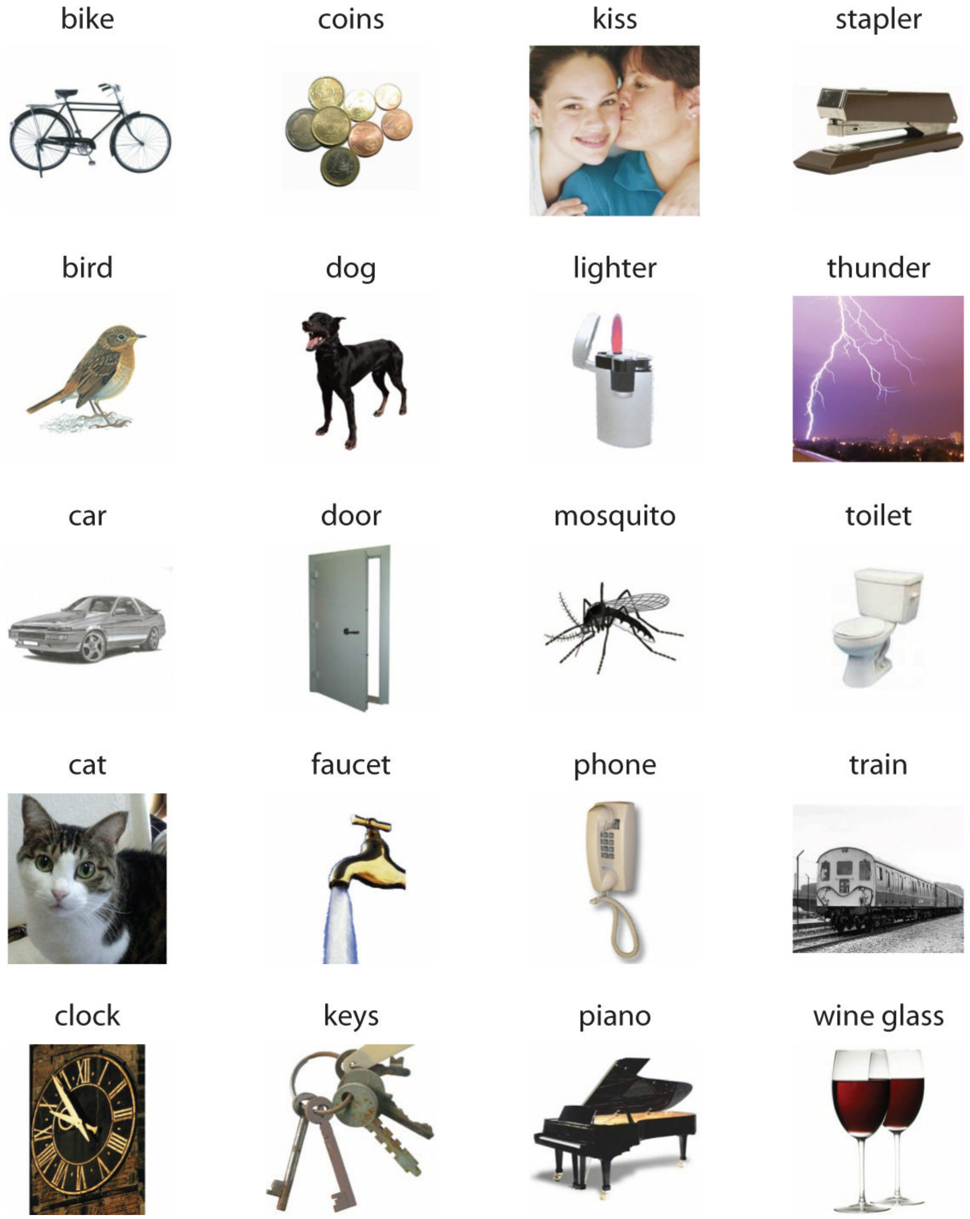
Once the high-level visual representation of the target is enhanced by the associated auditory signal, localization of the target could be facilitated via the extensive cross-connections between the ventral (thought to mediate object processing) and dorsal (thought to mediate spatial and action-related processing) cortical visual pathways (e.g., Felleman & Van Essen, 1991), and/or via the feedback connections from high-level polysensory areas and object-processing visual areas to low-level retinotopic visual areas (e.g., Rockland & Van Hoesen, 1994; Roland et al., 2006). For example, within about 100 msec after presentation of a visual stimulus, a wave of feedback activation from high-level visual areas selectively enhances the low-level retinotopic (spatially selective) responses to the stimulus (Roland et al., 2006). It is plausible that when the high-level object-based visual responses to the target (e.g., a cat) are cross-modally enhanced by the simultaneously presented characteristic sound (e.g., a meow), those enhanced high-level visual responses in turn strengthen the feedback enhancement of retinotopic responses to the target stimulus in low-level visual areas. Once neural responses to the target stimulus are enhanced in retinotopic visual areas with spatial selectivity, target localization will be facilitated.

Regardless of the exact neural mechanisms underlying the object-specific auditory–visual interactions, our results, combined with prior results on spatial auditory–visual interactions, suggest that auditory processing facilitates visual search through both spatial and object-based interactions. Having these separate modes of auditory–visual enhancements is likely to offer behavioral benefits. For example, if you expect to encounter a rattlesnake, a rattling sound will enhance visual detection of the snake, both by locally enhancing visual processing near the sound source (via spatial interactions) and by globally enhancing the processing of snake-related visual features (via the object-specific interactions). Furthermore, whereas spatial cross-modal enhancements occur relatively automatically (e.g., Driver & Spence, 1998), object-based enhancements occur in a goal-directed manner (see our results and Molholm et al., 2004, and von Kriegstein et al., 2005). The two modes of cross-modal enhancement might thus be complementary; the location-based interactions might facilitate detection of unexpected objects, whereas the object-based interactions might selectively facilitate detection of goal-related objects. Finally, our results may find an interesting practical application. During a search for your keys, for example, playing jingling sounds might help you by increasing the visual salience of key-related features.

Acknowledgements

This work was supported by National Institutes of Health Grant EY14110 and NSF Grant BCS 3643191 to the last author.

APPENDIX



REFERENCES

Amedi A, von Kriegstein K, van Atteveldt MN, Beauchamp MS, Naumer MJ. Functional imaging of human crossmodal identification and object recognition. Experimental Brain Research 2005;166:559–571.

Andersen RA, Snyder LH, Bradley DC, Xing J. Multimodal representation of space in the posterior parietal cortex and its use in planning movements. Annual Review of Neuroscience 1997;20:303–330.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. Nature Neuroscience 2004;7:1190–1192.

Beauchamp MS, Lee KE, Argall BD, Martin A. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 2004;41:809–823. [PubMed: 15003179]

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

Driver J, Spence C. Attention and the crossmodal construction of space. Trends in Cognitive Sciences 1998;2:254–262.

Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1991;1:1–47. [PubMed: 1822724]

Frassinetti F, Bolognini N, Làdavas E. Enhancement of visual perception by crossmodal visuo-auditory interaction. Experimental Brain Research 2002;147:332–343.

Hebb, DO. The organization of behavior. Wiley; New York: 1949.

Kastner S, Ungerleider LG. Mechanisms of visual attention in the human cortex. Annual Review of Neuroscience 2000;23:315–341.

Molholm S, Ritter W, Javitt DC, Foxe JJ. Multisensory visual–auditory object recognition in humans: A high-density electrical mapping study. Cerebral Cortex 2004;14:452–465. [PubMed: 15028649]

Reynolds JH, Chelazzi L. Attentional modulation of visual processing. Annual Review of Neuroscience 2004;27:611–647.

Rockland KS, Van Hoesen GW. Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. Cerebral Cortex 1994;4:300–313. [PubMed: 8075534]

Roland PE, Hanazawa A, Undeman C, Eriksson D, Tompa T, Nakamura H, et al. Cortical feedback depolarization waves: A mechanism of top-down influence on early visual areas. Proceedings of the National Academy of Sciences 2006;103:12586–12591.

Rovamo J, Virsu V. Visual resolution, contrast sensitivity, and the cortical magnification factor. Experimental Brain Research 1979;37:475–494.

Smith E, Grabowecky M, Suzuki S. Auditory–visual crossmodal integration in perception of face gender. Current Biology 2007;17:1680–1685. [PubMed: 17825561]

Stein BE. Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. Experimental Brain Research 1998;123:124–135.

Stein BE, Meredith ME, Huneycutt WS, McDade LW. Behavioral indices of multisensory integration: Orientation to visual cues is affected by auditory stimuli. Journal of Cognitive Neuroscience 1989;1:12–24.

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud A-L. Interaction of face and voice areas during speaker recognition. Journal of Cognitive Neuroscience 2005;17:367–376. [PubMed: 15813998]

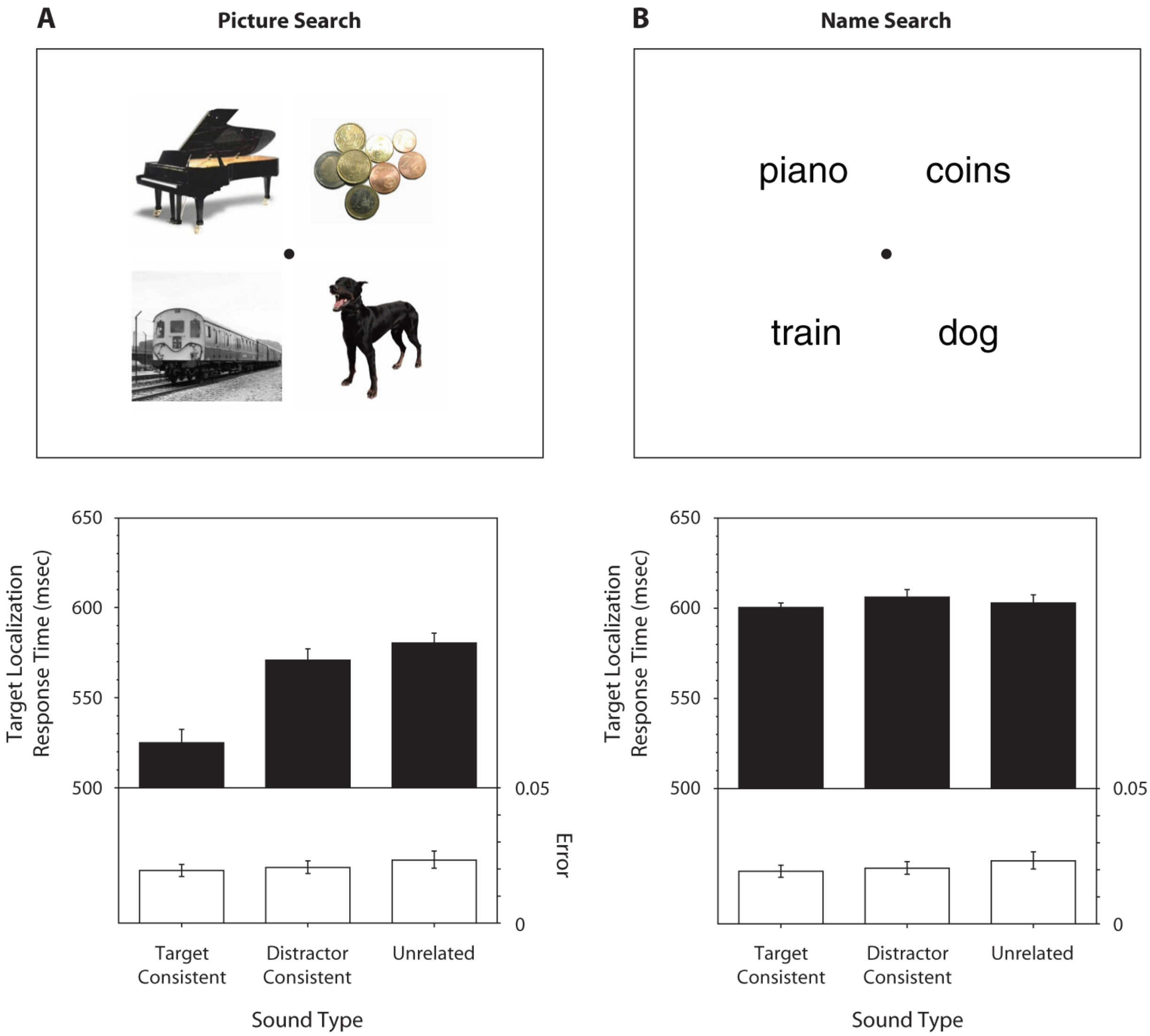


Figure 1.

Effects of characteristic sounds on four-item visual search. (A) The upper panel shows an example of a picture-search display. The lower panel shows target localization response times (RTs; filled bars) and error rates (open bars) when the search displays were presented with the target-consistent sounds, distractor-consistent sounds, or unrelated sounds. The error bars represent ±1 *SE* (the variance due to differences in the overall RT or error rate among the participants was removed before computing *SE*). (B) The corresponding information for the name-search experiment.

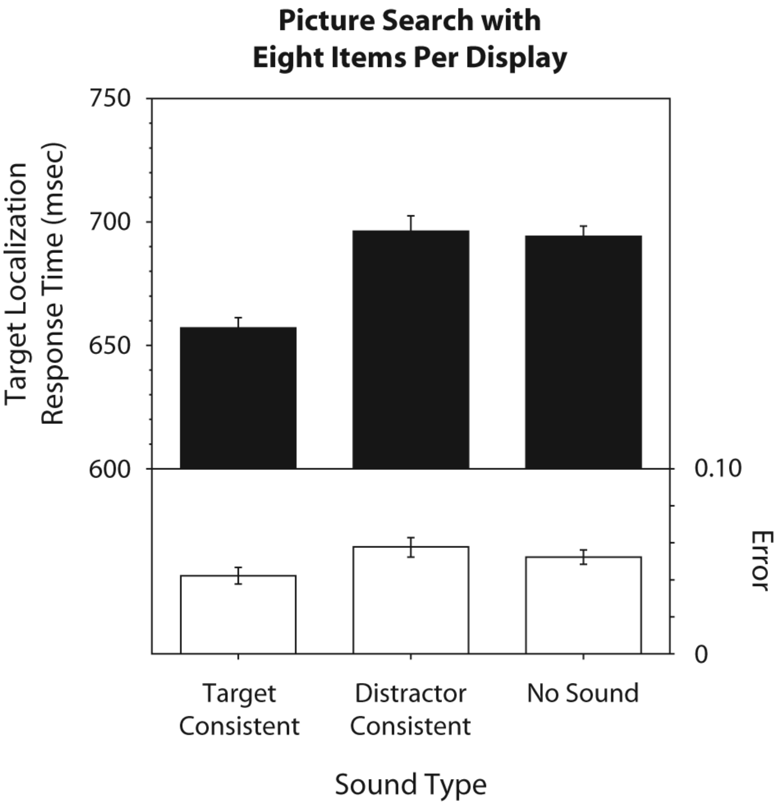


Figure 2.
Effects of characteristic sounds on 8-item visual search. The graph shows target localization response times (RTs; filled bars) and error rates (open bars) when the search displays were presented with the target-consistent sounds, distractor-consistent sounds, or no sounds. The error bars represent ±1 *SE* the variance due to differences in the overall RT or error rate among the participants was removed before computing *SE*).