Published in final edited form as:

Psychon Bull Rev. 2015 April ; 22(2): 429–436. doi:10.3758/s13423-014-0685-3.

Learned face-voice pairings facilitate visual search

L. Jacob Zweig¹, Satoru Suzuki^{1,2}, and Marcia Grabowecky^{1,2}

¹Department of Psychology, Northwestern University, Evanston, IL

²Interdepartmental Neuroscience Program, Northwestern University, Evanston, IL

Abstract

Voices provide a rich source of information that is important for identifying individuals and for social interaction. During search for a face in a crowd, voices often accompany visual information and they facilitate localization of the sought individual. However, it is unclear whether this facilitation occurs primarily because the voice cues the location of the face or because it also increases the salience of the associated face. Here we demonstrate that a voice that provides no location information nonetheless facilitates visual search for an associated face. We trained novel face/voice associations and verified learning using a two-alternative forced-choice task in which participants had to correctly match a presented voice to the associated face. Following training, participants searched for a previously learned target face among other faces while hearing one of the following sounds (localized at the center of the display): a congruent-learned voice, an incongruent but familiar voice, an unlearned and unfamiliar voice, or a time-reversed voice. Only the congruent-learned voice speeded visual search for the associated face. This result suggests that voices facilitate visual detection of associated faces, potentially by increasing their visual salience, and that the underlying crossmodal associations can be established through brief training.

Keywords

Crossmodal integration; Visual search; Face perception; Spatial attention

Introduction

Searching for a specific face is a common experience, from finding a friend to identifying a security threat. Research investigating visual search for faces with a specific identity has generally used visual stimuli (e.g., Kuehn & Jolicoeur, 1994; Tong & Nakayama, 1999). However, in real-world search for a specific face, voices frequently accompany faces and carry information about the speaker's location, identity, and emotion.

Understanding of how crossmodal information, such as faces and voices, are integrated in the brain continues to evolve. Evidence for crossmodal convergence in parietal (e.g., Schroeder & Foxe, 2002) and temporal lobes (Benevento, Fallon, Davis, & Rezak, 1977; Schroeder & Foxe, 2002) suggests that information converges in multimodal association

areas only after processing in unimodal areas. However, recent studies suggest that crossmodal signals influence processing in primary sensory areas of the brain (e.g., Ghazanfar & Schroeder, 2006; Giard & Peronnet, 1999; Molholm et al., 2002; Schroeder & Foxe, 2005; Shams, Iwaki, Chawla, & Bhattacharya, 2005). For example, Watkins and colleagues have utilized the sound-induced flash illusion to demonstrate auditory modulation of primary visual cortex (Watkins, Shams, Josephs, & Rees, 2007; Watkins, Shams, Tanaka, Haynes, & Rees, 2006).

Studies exploring the consequences of auditory-visual interactions have found that sounds facilitate recognition and search when the sound and visual object are spatially coincident (Bolognini, Frassinetti, Serino, & Làdavas, 2005; Driver & Spence, 1998; Frassinetti, Bolognini, & Làdavas, 2002; Stein, Meredith, Huneycutt, & McDade, 1989). Similarly, sounds facilitate visual processing when beeps and flashes are temporally coincident (e.g., Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008).

Beyond spatial and temporal crossmodal correspondences, experience- or identity-based crossmodal correspondences also influence perceptual processing. These correspondences rely on the co-occurrence of features and/or congruence in identity across stimulus modalities, and influence perceptual processing. For example, congruent audiovisual pairs (e.g., a picture of a dog and a barking sound) elicit faster object-recognition response times compared to identification of unimodal stimuli (Hein et al., 2007; S. Molholm, Ritter, Javitt, & Foxe, 2004), increased negativity in the N1 ERP component (S. Molholm et al., 2004), and increased activation in posterior superior temporal sulcus and superior temporal gyrus (Hein et al., 2007). Similarly, recent studies have shown that presentation of an object-characteristic sound at the onset of search (e.g., jingling keys when looking for keys) that provided no information about target location nonetheless lead to faster target localization, potentially by increasing target salience. This has been demonstrated both as faster reaction times for target localization (Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008), and speeded saccades to targets (within 220 ms, Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010). This suggests a rapid influence of crossmodal correspondences on visual search. Previous studies have relied upon existing associations (e.g., keys and jingling or cows and mooing). They also investigated “type-type” associations, in that any jingling sound should help to find any set of keys. We extended these prior findings in two ways. First, we created specific “token-token” associations by associating a particular face with an arbitrary voice. Secondly, these novel face-voice associations were created through brief training rather than the potentially life-long experience that underlies the crossmodal correspondences used in prior studies.

Several studies have suggested that object and person-identity representations may be inherently multisensory. For example, image sequences of faces producing human speech movements (Calvert, 1997) or of monkey vocalizations (Ghazanfar, Maier, Hoffman, & Logothetis, 2005) evoke responses in auditory cortex. Similarly, voices of familiar speakers activate the fusiform face area (von Kriegstein & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). Fewer studies have explored the perceptual consequences of multisensory-person-identity associations. Experiments by Schweinberger and colleagues have demonstrated faster and more accurate recognition of a familiar voice

when the corresponding face was presented compared to when the voice was presented alone (e.g., Schweinberger, Robertson, & Kaufmann, 2007; Schweinberger, Kloth, & Robertson, 2011; see Schweinberger, 2013 for a review).

If person-identity representations are multisensory then we would expect that identity congruence between a presented face and a presented voice will facilitate visual processing. We hypothesized that presentation of a spatially uninformative voice, trained to be associated with a particular face, would speed visual search for that face compared to when an unassociated voice was presented. In order to begin to understand the nature of any effects of face-voice associations, across three experiments we compared learned face-voice associations to other combinations of faces and voices. The learned face-voice association was tested with a change in utterance between learning and visual search to test the generalizability of the vocal cue. This condition was compared to an unlearned (novel) face-voice combination, to presentation of a familiar voice and a familiar face that were not previously paired (a familiar but incongruent pairing), and to a learned face-voice association, but with the voice utterance played backwards. This latter condition disrupted linguistic qualities such as formant transitions and syllable shape but left other acoustic features such as fundamental frequency and vocal timbre relatively preserved (e.g., Sheffert, Pisoni, Fellowes, & Remez, 2002). Finally, a face-alone condition allowed us to assess whether presentation of a voice facilitated visual search beyond a face-alone condition.

Experiment 1: Congruent voices facilitate visual search for associated faces

Method

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

Stimuli—Color photographs of faces were randomly selected from a set of 33 males from the Karolinska face database (Lundqvist, Flykt & Ohman, 1998). Each subtended 3.18° (width) by 4.24° (height) of visual angle and was presented centrally. Faces appeared on a color CRT monitor (1024 × 768 pixels) with a 60-Hz refresh rate. A MacBook Air computer using MATLAB (The MathWorks, Inc., Natick, MA) with PsychToolbox software (Brainard, 1997; Pelli, 1997) controlled the experiment.

Auditory stimuli were pre-recorded voices uttering a brief introduction (e.g., “Hi, my name is Michael and I’m 26”). Three male speakers were used throughout the experiment. The voices used did not correspond to any of the faces used during the experiment. All voices were presented at a clearly audible level (~70dB SPL), via Sennheiser HD 280-Pro headphones.

Procedure

Training face-voice associations: Participants completed two phases of training to associate face-voice pairs: an associative-learning phase (Fig. 1a) followed by a two-

alternative matching task (Fig. 1b). During the associative-learning trials, participants were instructed to imagine that they were meeting new individuals for the first time. For each participant, two voices were each associated with one of two faces randomly selected from a set of 33 male faces. A third voice was later used as an unlearned voice in the search task. Each display contained a color photograph of one of the two selected faces appearing synchronously with the onset of the associated voice. The same voice was always paired with the same face. Participants completed 20 trials (10 trials for each face-voice pair).

Following associative-learning trials, participants completed a two-alternative forced-choice matching task. Each trial of the matching task presented two faces and a simultaneous voice previously paired with one of the trained faces. Participants indicated the face associated with the voice by pressing a button corresponding to the side of the screen where the associated face appeared. Participants used the left index finger to respond to left-side faces and the right index finger to respond to right-side faces. On approximately 2/3 of trials, a previously-learned face and a randomly selected unlearned-distractor face appeared; on the other 1/3 of trials both faces were previously-learned. Unlearned-distractor faces were randomly selected for each trial. The matching task continued until participants reached 25 consecutive correct responses.

Visual-Search Task: One of the faces from the associative learning trials was selected as the visual-search target for each participant. The display was divided into four quadrants. Each search display consisted of the target and 15 distractors randomly selected from the 31 remaining male faces. The target-face location was pseudo-randomly counterbalanced to occur at every position with approximately equal probability for each condition. At the onset of each visual-search display, a saying “I’m over here!” was presented. The voice was the one paired with the target during learning (*congruent-learned*), a novel, unlearned voice (*unlearned*), or was the same phrase and voice played in reverse (*reversed*). As the utterance differed from the trained utterance, any facilitation may suggest that characteristics of the voices, such as pitch, timbre, or phonemic characteristics may influence perceptual processing. To separately evaluate non-linguistic effects of the voice, the reversed voice condition was included to preserve timbre and fundamental frequency without linguistic cues (Sheffert et al., 2002). Each participant was tested in 3 blocks of 120 trials (40 trials per sound condition), with the voice conditions randomly intermixed.

Participants were instructed to locate the target by pressing a button corresponding to the quadrant of the target as quickly and accurately as possible. Response times (RTs) and accuracy were recorded.

Results

Response accuracy was high (91.5%). No significant differences in accuracy across the three voice conditions were detected (Figure 2a) [$F_{(2,22)} = 1.77, p = 0.19, \eta^2 = 0.14$], indicating that RT differences were not due to speed-accuracy trade-offs. Error trials were not further analyzed.

A repeated-measures ANOVA revealed differences in reaction times across the three sound conditions (Figure 2b) [$F_{(2,22)} = 6.99, p = 0.004, \eta^2 = 0.39$]. RTs for congruent-learned-voice

trials were faster than RTs for unlearned-voice trials [$t_{(11)} = 3.01, p = 0.01, d = 0.49$], and RTs for congruent-learned-voice trials were faster than for reversed-voice trials [$t_{(11)} = 4.06, p = 0.002, d = 0.29$]. No significant differences in RTs between the reversed-voice and unlearned-voice conditions were observed [$t_{(11)} = 1.79, p = 0.10, d = 0.27$], suggesting that features such as fundamental frequency or timbre did not produce a search advantage.

These results suggest that the visual-search advantage from simultaneously presented voices depends on a learned auditory-visual relationship, and perhaps learned linguistic features. However, it is possible that the facilitation may have been due to the familiarity of the learned voice compared to the unfamiliar unlearned and reversed voices. Schweinberger and colleagues (2007) reported a significant familiarity effect, with faster “familiar” than “unfamiliar” responses in a voice familiarity judgment task. Similarly, other studies have shown faster response times for familiar faces compared to unfamiliar faces (e.g., Balas, Cox, & Conwell, 2007; Bruce, 1986).

Facilitation in visual search may occur for any familiar voice, regardless of face-voice congruence. To investigate this in a second experiment we added a condition with a previously learned (thus familiar), but target-incongruent voice. If search facilitation depends solely on the familiarity of the voice, a learned but target-incongruent voice should speed RTs during visual search relative to reversed and unlearned voices. Alternatively, if the search facilitation depends on face-voice identity congruity, a learned but target-incongruent voice should produce facilitation relative to reversed and unlearned voices.

Experiment 2: The visual search advantage due to voices relies on identity congruity

Method

Participants—A new group of 15 undergraduate students at Northwestern University gave informed consent to participate for partial course credit. All had normal or corrected-to-normal visual acuity, normal hearing, and were tested individually in a normally lit quiet room.

Stimuli—The stimuli were identical to Experiment 1, with the addition of a previously learned but target-incongruent voice condition (*familiar-incongruent*).

Procedure—The procedure was identical to Experiment 1 (see Fig. 1), except that each participant was tested in 3 blocks of 160 trials (40 trials per sound condition randomly intermixed).

Results

As shown in Figure 3a, no significant differences in accuracy across the four voice conditions were detected [$F_{(3,42)} = .64, p = 0.59, \eta^2 = 0.04$], therefore error trials were not further analyzed. Overall response times did not significantly differ between experiments 1 and 2 [$t_{(25)} = .74, p = .46, d = .29$]. Consistent with the results of Experiment 1, a repeated-measures ANOVA revealed significant differences across the four voice conditions [$F_{(3,42)}$

= 3.40, $p = 0.03$ $\eta^2 = 0.20$] (Figure 3b). RTs for congruent-learned-voice trials were faster than RTs for unlearned-voice trials [$t_{(14)} = 2.42$, $p = 0.03$, $d = 0.34$] and reversed-voice trials [$t_{(14)} = 2.14$, $p = 0.05$, $d = 0.23$]. Additionally, RTs for congruent-learned trials were faster than for familiar-incongruent trials [$t_{(14)} = 2.663$, $p = 0.02$, $d = 0.48$].

We observed no significant differences between unlearned- and reversed-voice trials [$t_{(14)} = 1.10$, $p = 0.29$, $d = 0.11$], unlearned- and familiar-incongruent-voice trials [$t_{(14)} = .96$, $p = 0.35$, $d = 0.20$], or reversed- and familiar-incongruent-voice trials [$t_{(14)} = 1.48$, $p = 0.16$, $d = 0.29$]. These results suggest that familiar but target-incongruent voices produce no search advantage. This suggests that the facilitation in visual search observed in Experiment 1 was not due to the familiarity of the voices presented during search. However, the facilitation of response times for congruent voices may still represent either facilitation relative to visual search without an auditory cue or instead reduced interference with congruent-learned-voice compared to all other auditory-visual conditions.

Experiment 3: Congruent voices facilitate visual search compared to visual search for faces with no auditory cue

Method

Participants—A new group of 15 undergraduate students at Northwestern University gave informed consent to participate for partial course credit. All had normal or corrected-to-normal visual acuity, normal hearing, and were tested individually in a normally lit quiet room.

Stimuli—The stimuli were identical to Experiment 1, with the addition of a condition with no auditory cue in place of the reversed voice condition.

Procedure—The procedure was identical to Experiment 1 (see Fig. 1), with the no-auditory-cue condition replacing the reversed-voice condition.

Results

No significant differences in accuracy across the three voice conditions (congruent-learned-voice, unlearned-voice, no-auditory-cue) were detected [$F_{(2,28)} = .54$, $p = 0.59$, $\eta^2 = 0.04$], therefore error trials were not further analyzed. A one-way ANOVA revealed that response times did not significantly differ among experiments 1, 2, and 3 [$F_{(2,39)} = 2.81$, $p = .07$, $\eta^2 = 0.13$]. As shown in in Figure 4b, consistent with the results of Experiments 1 and 2, a repeated-measures ANOVA revealed significant differences across the three voice conditions [$F_{(2,28)} = 5.43$, $p = 0.01$ $\eta^2 = 0.28$]. RTs for congruent-learned-voice trials were faster than RTs for unlearned-voice trials [$t_{(14)} = 2.26$, $p = 0.04$, $d = 0.14$] and than RTs for no-auditory-cue trials [$t_{(14)} = 3.15$, $p = .007$, $d = 0.20$]. No significant differences in RTs between the no-auditory-cue and unlearned-voice conditions were observed [$t_{(14)} = 1.16$, $p = 0.27$, $d = 0.06$] indicating that learned-congruent voices facilitate visual search beyond a visual-only baseline.

General Discussion

These results demonstrate that visual search for faces can be facilitated by recently associated voices that do not spatially cue the face location. Whereas multisensory spatial interactions have been extensively studied (e.g., Buchtel & Butter, 1988; Driver & Spence, 1994, 1998; Eimer, Cockburn, Smedley, & Driver, 2001; Eimer & Driver, 2000), our demonstration of auditory enhancement of visual search is consistent with previous studies exploring the perceptual consequences and neural correlates of crossmodal object representation (e.g., Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Guzman-Martinez, Ortega, Grabowecky, Mossbridge, & Suzuki, 2012; Iordanescu et al., 2008; Iordanescu et al., 2010; Iordanescu, Grabowecky, & Suzuki, 2011; Smith, Grabowecky, & Suzuki, 2007; Taylor, Stamatakis, & Tyler, 2009).

Facilitation of visual search following face-voice learning may rely on both anatomical connectivity and functional coupling. Support for direct connectivity between face and voice recognition areas has been provided by probabilistic tractography (Blank, Anwender, & von Kriegstein, 2011). Additional studies have demonstrated changes in functional connectivity between the temporal voice area and the fusiform face area following face-voice associative learning (von Kriegstein & Giraud, 2006; von Kriegstein et al., 2005), suggesting that person-identity recognition may rely on such plasticity in connectivity. Alternatively, these results may suggest the existence of a neural substrate that responds maximally to multisensory person-identity representations, which may include areas proposed to be involved in multisensory object recognition (Amedi, von Kriegstein, van Atteveldt, Beauchamp, & Naumer, 2005; Beauchamp et al., 2004; Dahl, Logothetis, & Kayser, 2009; Zheng, Wild, & Trang, 2010).

Following enhancement of person-identity representation, facilitation in visual search may rely on top-down modulation of spatial-orienting networks that influence early visual processing (e.g., Kelly, Gomez-Ramirez, & Foxe, 2008; Poghosyan & Ioannides, 2008; Saalman, Pigarev, & Vidyasagar, 2007). For example, identity-based multisensory associations may serve to increase the strength of feedback to spatial-orienting networks, increasing the salience of locations with potential targets and decreasing the salience of locations with potential distractors (e.g., Roland et al., 2006).

Alternatively, the observed facilitation might rely on modification of thresholds for person-identity recognition. Congruent multisensory information may increase the strength of the person-identity representation, thereby lowering the threshold for face-identity recognition (e.g., Tsakiris, 2008), while incongruent information may conversely increase the threshold for face-identity recognition. Consequently, congruent information may speed identification of a target face upon initial fixation, whereas incongruent information may increase initial misses during scanning, thereby delaying target detection. Modulation of spatial processing and recognition threshold are not mutually exclusive and further research is necessary to determine the degree to which each may be affected by feature-based multisensory information.

Whereas previous studies have demonstrated a benefit from consistent face information for recognition of voices (Schweinberger, Kloth, & Robertson, 2011; Schweinberger et al.,

2007), the experiments reported here demonstrate behavioral facilitation in face localization resulting from voice-identity congruence. Our results also demonstrate the rapid formation of crossmodal person-identity representations that facilitate visual search. Whereas Iordanescu and colleagues (2008; 2010) explored the role of normative “type-type” crossmodal associations (e.g., a cat's meow or keys jingling) based on extended prior experience, we have demonstrated facilitation in visual search derived from rapidly learned “token-token” associations between a particular individual's face and the unique linguistic features of his/her voice. Our results suggest that visual perception is influenced by these rapidly-learned identity-based crossmodal associations and strengthened through repeated spatial and/or temporal coincidence.

References

- Amedi A, von Kriegstein K, van Atteveldt NM, Beauchamp MS, Naumer MJ. Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*. 2005; 166(3-4):559–71.
- Balas B, Cox D, Conwell E. The Effect of Real-World Personal Familiarity on the Speed of Face Information Processing. *PLoS ONE*. 2007; 2:5.
- Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience*. 2004; 7(11): 1190–2.
- Benevento LA, Fallon J, Davis BJ, Rezak M. Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental Neurology*. 1977; 57(3):849–872. [PubMed: 411682]
- Blank H, Anwender A, von Kriegstein K. Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*. 2011; 31(36):12906–15. [PubMed: 21900569]
- Bolognini N, Frassinetti F, Serino A, Làdavas E. “Acoustical vision” of below threshold stimuli: interaction among spatially converging audiovisual inputs. *Experimental Brain Research Experimentelle Hirnforschung Expérimentation Cérébrale*. 2005; 160(3):273–82.
- Bruce V. Influences of familiarity on the processing of faces. *Perception*. 1986; 15:387–397. [PubMed: 3822723]
- Buchtel HA, Butter CM. Spatial attentional shifts: implications for the role of polysensory mechanisms. *Neuropsychologia*. 1988; 26(4):499–509. [PubMed: 3405396]
- Calvert GA. Activation of Auditory Cortex During Silent Lipreading. *Science*. 1997; 276(5312):593–596. [PubMed: 9110978]
- Calvo MG, Nummenmaa L. Detection of emotional faces: salient physical features guide effective visual search. *Journal of Experimental Psychology. General*. 2008; 137(3):471–94. [PubMed: 18729711]
- Dahl CD, Logothetis NK, Kayser C. Spatial organization of multisensory responses in temporal association cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*. 2009; 29(38):11924–32. [PubMed: 19776278]
- Driver, J.; Spence, C. Spatial synergies between auditory and visual attention.. In: Umiltà, C.; Moscovitch, M., editors. *Attention And Performance*. Vol. XV. Erlbaum; 1994. p. 311-331.
- Driver J, Spence C. Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*. 1998; 2(7):254–62. [PubMed: 21244924]
- Eimer M, Cockburn D, Smedley B, Driver J. Cross-modal links in endogenous spatial attention are mediated by common external locations: evidence from event-related brain potentials. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*. 2001; 139(4):398–411.
- Eimer M, Driver J. An event-related brain potential study of cross-modal links in spatial attention between vision and touch. *Psychophysiology*. 2000; 37(5):697–705. [PubMed: 11037046]

- Frassinetti F, Bolognini N, Làdavas E. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*. 2002; 147(3):332–43.
- Ghazanfar, A. a; Schroeder, CE. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*. 2006; 10(6):278–85. [PubMed: 16713325]
- Ghazanfar AA, Maier JX, Hoffman KL, Logothetis NK. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*. 2005; 25(20):5004–12. [PubMed: 15901781]
- Giard MH, Peronnet F. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*. 1999; 11(5):473–90. [PubMed: 10511637]
- Guzman-Martinez E, Ortega L, Grabowecky M, Mossbridge J, Suzuki S. Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Current Biology : CB*. 2012; 22(5):383–8. [PubMed: 22326023]
- Hein G, Doehrmann O, Müller NG, Kaiser J, Muckli L, Naumer MJ. Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*. 2007; 27(30):7881–7. [PubMed: 17652579]
- Iordanescu L, Grabowecky M, Franconeri S, Theeuwes J, Suzuki S. Characteristic sounds make you look at target objects more quickly. *Attention, Perception & Psychophysics*. 2010; 72(7):1736–41.
- Iordanescu L, Grabowecky M, Suzuki S. Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. *Acta Psychologica*. 2011; 137(2):252–9. [PubMed: 20864070]
- Iordanescu L, Guzman-Martinez E, Grabowecky M, Suzuki S. Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*. 2008; 15(3):548–554. [PubMed: 18567253]
- Kelly SP, Gomez-Ramirez M, Foxe JJ. Spatial attention modulates initial afferent activity in human primary visual cortex. *Cerebral Cortex*. 2008; 18(11):2629–36. [PubMed: 18321874]
- Kuehn SM, Jolicoeur P. Impact of quality of the image, orientation, and similarity of the stimuli on visual search for faces. *Perception*. 1994; 23(1):95–122. [PubMed: 7936979]
- 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(4):452–65. [PubMed: 15028649]
- Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ. Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*. 2002; 14(1):115–128. [PubMed: 12063135]
- Poghosyan V, Ioannides AA. Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron*. 2008; 58(5):802–13. [PubMed: 18549790]
- Saalmann YB, Pigarev IN, Vidyasagar TR. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science (New York, N.Y.)*. 2007; 316(5831):1612–5.
- Schroeder CE, Foxe J. Multisensory contributions to low-level, “unisensory” processing. *Current Opinion in Neurobiology*. 2005; 15(4):454–8. [PubMed: 16019202]
- Schroeder CE, Foxe JJ. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Research. Cognitive Brain Research*. 2002; 14(1):187–98. [PubMed: 12063142]
- Schroeder CE, Lindsley RW, Specht C, Marcovici A, Smiley JF, Javitt DC. Somatosensory input to auditory association cortex in the macaque monkey. *Journal of Neurophysiology*. 2001; 85(3):1322–7. [PubMed: 11248001]
- Schweinberger, SR. Integrating Face and Voice in Person Perception. Springer; New York: 2013. Audiovisual integration in speaker identification.; p. 45-69.
- Schweinberger SR, Kloth N, Robertson DMC. Hearing facial identities: brain correlates of face–voice integration in person identification. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*. 2011; 47(9):1026–37.
- Schweinberger SR, Robertson D, Kaufmann JM. Hearing facial identities. *Quarterly Journal of Experimental Psychology (2006)*. 2007; 60(10):1446–56. [PubMed: 17853250]

- Shams L, Iwaki S, Chawla A, Bhattacharya J. Early modulation of visual cortex by sound: an MEG study. *Neuroscience Letters*. 2005; 378(2):76–81. [PubMed: 15774261]
- Shams L, Wozny DR, Kim R, Seitz A. Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology*. 2011; 2:264. [PubMed: 22028697]
- Sheffert SM, Pisoni DB, Fellowes JM, Remez RE. Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology. Human Perception and Performance*. 2002; 28:1447–1469. [PubMed: 12542137]
- Smith EL, Grabowecky M, Suzuki S. Auditory-visual crossmodal integration in perception of face gender. *Current Biology : CB*. 2007; 17(19):1680–5. [PubMed: 17825561]
- Stein BE, Meredith MA, Huneycutt WS, McDade L. Behavioral Indices of Multisensory Integration: Orientation to Visual Cues is Affected by Auditory Stimuli. *Journal of Cognitive Neuroscience*. 1989; 1(1):12–24. [PubMed: 23968407]
- Taylor KI, Stamatakis EA, Tyler LK. Crossmodal integration of object features: voxel-based correlations in brain-damaged patients. *Brain : A Journal of Neurology*. 2009; 132(Pt 3):671–83. [PubMed: 19190042]
- Tong F, Nakayama K. Robust Representation for Faces: Evidence from Visual Search. *Journal of Experimental Psychology* 1999
- Tsakiris M. Looking for myself: current multisensory input alters self-face recognition. *PloS One*. 2008; 3(12):e4040. [PubMed: 19107208]
- Van der Burg E, Olivers CNL, Bronkhorst AW, Theeuwes J. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology. Human Perception and Performance*. 2008; 34(5):1053–65. [PubMed: 18823194]
- Virsu V, Rovamo J. Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*. 1979; 37(3):475–94. [PubMed: 520438]
- Von Kriegstein K, Giraud A-L. Implicit multisensory associations influence voice recognition. *PLoS Biology*. 2006; 4(10):e326. [PubMed: 17002519]
- 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(3):367–76. [PubMed: 15813998]
- Watkins S, Shams L, Josephs O, Rees G. Activity in human V1 follows multisensory perception. *NeuroImage*. 2007; 37(2):572–8. [PubMed: 17604652]
- Watkins S, Shams L, Tanaka S, Haynes J-D, Rees G. Sound alters activity in human V1 in association with illusory visual perception. *NeuroImage*. 2006; 31(3):1247–56. [PubMed: 16556505]
- Zheng ZZ, Wild C, Trang HP. Spatial organization of neurons in the superior temporal sulcus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*. 2010; 30(4):1201–3. [PubMed: 20107047]

Figure 1.

(a) Training sequence. During associative learning trials, a color photograph of a face appeared simultaneously with a pre-recorded voice introduction (e.g., “Hi, my name is Michael and I’m 26”) for a duration of 10 seconds. During the two-alternative forced-choice matching task (b), participants matched the presented voice to the congruent face until criterion of 25 consecutive correct responses. (c) An example search display. One of three voices was played (*congruent-learned*, *unlearned*, or *reversed*; see text for details). Participants indicated the quadrant in which the target face was presented. Each search display contained 16 color photographs of faces, arranged on two approximate iso-acuity ellipses (based on Virsu & Rovamo, 1979).

Figure 2.

(a) Response accuracy for the three voice conditions. Overall response accuracy was high (91.5%) and there were no significant differences in accuracy. (b) Response times (RTs) for the three voice conditions. RTs for congruent-learned-voice trials were significantly faster than those for unlearned-voice trials and reversed-voice trials. The error bars represent ± 1 SEM adjusted for within-participant comparisons.

Figure 3.

(a) Response accuracy for the four voice conditions; There were no significant differences in accuracy. (b) Response times (RTs) for the four voice conditions. Consistent with the previous experiment, RTs for congruent-learned-voice trials were significantly faster than those for reversed-, unlearned-, and familiar-incongruent-voice trials. The error bars represent ± 1 SEM adjusted for within-participant comparisons.

Figure 4.

(a) Response accuracy for the three voice conditions; There were no significant differences in accuracy. (b) Response times (RTs) for the three voice conditions. Consistent with the previous experiment, RTs for learned-voice trials were significantly faster than those for unlearned-voice and no-auditory-cue trials. The error bars represent ± 1 SEM adjusted for within-participant comparisons.