CHAPTER 3:

CROSSMODAL INTRINSIC MAPPINGS MAKE AUDITORY SENSORY SUBSTITUTION EFFORTLESS

Introduction

Sensory substitution studies have shown that sighted and blind participants can recognize and localize natural and artificial objects with sensory substitution given that participants have extensive training (one week to three months) and use top-down attention (Amedi, et al., 2007; Auvray, et al., 2007; Bach-y-Rita, et al., 1969; Bach-y-Rita, et al., 1998; Chebat, et al., 2011; Poirier, De Volder, & Scheiber, 2007; Proulx, et al., 2008). Whereas visual perception in the sighted is effortless and automatic, the usage of SS has so far been laborious, and this prevents devices from being successful commercially. No studies have investigated whether the processing of sensory substitution can be intuitive, or interpreted by entirely naïve participants with no device experience, training, or instruction. The only study that uses entirely naïve users is Auvray *et al.*, where they test whether distal perception (object perceived externally in perceptual space) can be learned without encoding knowledge of an auditory sensory substitution device, as detailed in Chapter 1 (p. 26) (Auvray, et al., 2005). It should also be noted that a SS visual acuity study used participants not trained with an SS device, but provided with a description of the device's vision-to-auditory encoding algorithm (Haigh, Brown, Meijer, & Proulx, 2013).

The current literature, reviewed in Chapter 1, seems to indicate that sensory substitution interpretation by trained users is a top-down cognitive process with attentive

concentration. Meanwhile, neural imaging studies on SS have so far shown the presence of plasticity, but uncertainty remains as to whether the plasticity is due to a top-down and attention-intensive process, or a bottom-up perceptual process (Amedi, et al., 2007; Poirier, De Volder, & Scheiber, 2007). Further, TMS studies have shown the visual activation from sensory substitution to be causally linked to task performance on the device in blind users (Collignon, et al., 2007; Merabet, et al., 2009). The current study (detailed in this chapter) is the first indication (among behavior or imaging studies) that sensory substitution interpretation (and potentially sensory substitution plasticity) does not always require top-down attention; rather it can rely on an automatic, bottom-up process.

Sensory substitution studies implicitly assume that blind or sighted participants cannot successfully interpret information provided by sensory substitution devices without both knowledge on the device encoding and sensorimotor training with it. However, the crossmodal correspondence literature (also called crossmodal associations, synaesthetic correspondences or associations, or intrinsic mappings) has shown that an intrinsic mapping exists between modalities (Spence, 2011). This intrinsic mapping may allow participants to perform tasks without any training, effort, or knowledge of the device encoding. For example, Figure 3.01 shows the intuitive matching of images to vOICe sounds by just using the amplitude modulation rate of the sound. The crossmodal mapping used in this example (amplitude modulation rate of sound to visual spatial frequency) is well-known, and has been studied in detail by Guzman-Martinez *et al.* (2012).

The crossmodal correspondences could further be used to enhance sensory substitution training by building on intuitive crossmodal features rather than ambiguous and unimodal visual features. Vision and audition correspondences can be generated by a common crossmodal feature, such as amplitude (brightness for vision, intensity for sound). On the other hand, seemingly unrelated modality-specific features have also been found to be matched, and matching can occur even at an abstract level (such as emotional response elicited) (Spence, 2011). It has been argued that these crossmodal mappings are learned priors within a Bayesian framework of crossmodal integration (Ernst, 2007). The encoding of vOICe is based on long-evidenced correspondences across vision and audition, such as the matching of brightness and loudness intensity (Stevens & Marks, 1965), spatial height and pitch height (Pratt, 1930), and scanning from left to right similar to reading written English. Therefore, participants with no knowledge about the vOICe device may in principle be able to use crossmodal correspondences to naïvely match images with their correct vOICe sounds. The device had been designed (either by chance or on purpose) for effortless usage, but somehow this advantage has not been explored. In addition to basic stimuli such as comparing lines of different angles encoded into sound with vOICe, our pilot observations suggest that other stimuli such as textures may have strong intrinsic crossmodal associations, and thus may also be correctly interpreted by naïve participants. This points to a possibility of a radical shift in SS training strategy. The vOICe device is particularly useful at encoding textures, as left-to-right scanning generates a dynamic beat that temporally plays out coarse-to-fine-grained spatial frequencies.

The naïve interpretation of vOICe would indicate that explicit instructions on the audiovisual vOICe encoding are not needed for vOICe interpretation. However, if the users can interpret vOICe without encoding instructions, this indicates that an intrinsic crossmodal mapping is utilized for interpretation, albeit implicitly. Therefore, the automaticity of the interpretation of vOICe naïvely depends on the automaticity of the crossmodal correspondences underlying that interpretation. Crossmodal correspondences can be automatic or require additional attention resources to interpret, depending on the type of mapping and task (Spence & Deroy, 2013). Chapter 1 discussed automaticity in vision, with an emphasis on visual distraction automaticity tests. Distraction tasks evaluate whether the stimuli in question is attention-load insensitive; this is one automaticity criterion. However, there are other criterion of automaticity, such as the "goal independence criterion," "the non-conscious criterion," and the "speed criterion" (Spence & Deroy, 2013). Spence and Deroy's review of crossmodal mappings automaticity indicate that auditory visual correspondences have some evidence of being goal-directed (*i.e.,* not automatic), but in contrast are speeded in the Implicit Associations Test (*i.e.,* automatic) (Parise & Spence, 2012; Spence & Deroy, 2013). The experiments discussed in this chapter will use load-insensitivity criteria for automaticity of vOICe and the crossmodal correspondences therein. The load-insensitivity measure for automaticity will be tested with a distraction task in audition as well as in vision during the vOICe sound interpretation (detailed below). While there are no papers on load-insensitivity of crossmodal mappings, there are studies of load-insensitivity of crossmodal interactions.

Distraction dual task designs have been used in studying the impact of high attention load on crossmodal integration. Alsius and colleagues studied the processing of

auditory and visual speech integration in the McGurk Effect while participants performed a distraction task (Alsius, Navarra, Campbell, & Soto-Faraco, 2005). Results indicated that reduced attention resources limited the McGurk effect. A study performed by Eramudugolla *et al.* indicated that ventriloquist aftereffect can occur under attention load, but that it is modulated by attention load (Eramudugolla, Kamke, Soto-Faraco, & Mattingley, 2011). Helbig and Ernst demonstrated that the weighting of visual and haptic stimuli is independent of attention load (Helbig & Ernst, 2008). These mixed multimodal results on attention load indicate that crossmodal mappings may or may not be independent of attention load. We will study this further in this chapter in application to the crossmodal mappings used in the vOICe device.

We address two crossmodal mapping problems in this Chapter: the engineering issue of optimally encoding vision into audition (V=>A), and the psychological/neural decoding of SS via crossmodal correspondences (A=>V). We began by studying the psychological/neural decoding of SS with the existing vOICe device encoding, to determine if vOICe can be intuitive. The results then suggested optimal methods for the encoding of vision into audition. In other words, once we know what works in vOICe, we can then accentuate those characteristics to make even more intuitive device encodings and training procedures

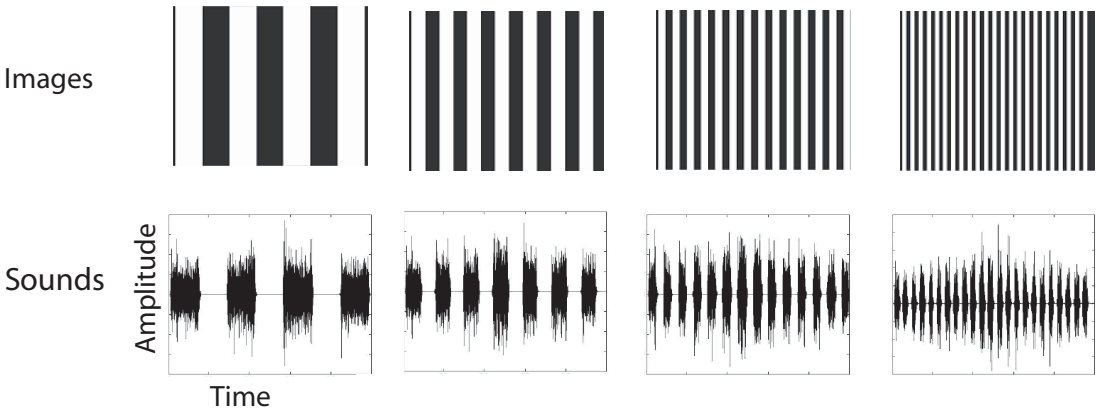


Figure 3.01. Example of intuitive image and vOICe matching. Figure 3.01 shows the example output from the vOICe (row 2) for a given set of images (row 1) used in bimodal matching experiments. Each row in the graphic is a different representation of the set of images: the first row is the visual representation, the second row uses just amplitude of the vOICe sound as a function of time to represent the image. Each column represents the same image or information. It is clear with this particular set of images and vOICe sounds that they have similar structure, and therefore are intuitive to match. In fact, it is clear that it is easy to match the images and sounds even if the positions of images and sounds were jumbled.

We hypothesize that textures will be intuitive with vOICe. Textures have been studied in detail in vision, and are an important element of monocular depth perception, visual segmentation, and automatic visual search (Palmer, 1999). Cues in monocular depth perception such as texture gradient (texture elements become smaller with distance) and texture accretion and deletion (texture elements disappear and reappear with lateral movement) are important elements of monocular depth. In visual search, unique texture elements can be identified in either a parallel or serial manner (Bergen & Julesz, 1983). With parallel search, the unique element pops out and can be identified at the same speed independent of the number of distractors. In the serial search, the unique element localization depends on the number distractors (no pop-out). Textures can also be used in vision for segmenting a scene into different objects and/or visual regions, and can be used in object shape identification (via distortion of texture elements). As an important and prevalent element of vision it is logical that textures would also be valuable to the processing of vOICe stimuli.

Methods

The role of crossmodal correspondences was tested with naïve (N = 5-7) and trained sighted (N = 4) participants in a bimodal matching task (Figure 3.02). First, all stimuli were presented as a preview (all three to four images and then associated vOICe sounds in random order), and then participants heard one sound and were asked to match one of three presented images to the sound (3AFC). Naïve participants were not told the vOICe encoding scheme, nor that the sounds were from the vOICe. Participants were asked to match the image and sound that carried the same information; if uncertain, participants were told to guess. Feedback on performance was not provided to

participants. Images were compared in sets of three or four so that particular image features and types could be tested separately. Image types ranged from natural to artificial images, and from simple to complex images. Images sets included vertical bar textures of different thickness, circular patterns of different element sizes, and images of natural textures (Figure 3.06). All images were presented in grayscale, as vOICe sounds do not convey color information. A total of 24 image sets were tested (all images are included in the supplementary materials). The naïve sighted participants are different participants from the naïve trained participants.

The crossmodal mappings underlying vOICe's interpretation were tested on naïve sighted participants ($N = 8$). Participants performed a bimodal matching task of the same design as the original (detailed above), but with different encoding schemes to test the value of different crossmodal mappings. Different encoding paradigms were generated by altering the images inputted into the vOICe encoding software (for example: The inverted coding of dark regions louder than bright regions was generated by inverting image brightness before inputting the image into the vOICe software). The encoding inversions tested (on top of original; [0]) were: (1) dark regions louder than bright regions, (2) scanning right to left, (3) high frequency on the bottom, and, (4) scanning top to bottom and high frequency on the right (Figure 1.4 has the original vOICe encoding). The order of testing the different encoding inversions on participants was randomized (including the original mapping). All participants completed all five of the different encoding types (four inversions and one original) in one session.

Automaticity of vOICe interpretation via an attention load experiment was tested with a dual task design. In the first experiment, participants counted backward in 7s from

a random number displayed (between 100 and 112), while counting the vOICe sound played (vOICe sound started 10 seconds after counting started) ($N = 8$) (Figure 3.03). Participants then matched the vOICe sound to one image of three images displayed (3AFC, same design and image specifications as the bimodal matching experiment). The same participants also performed the original bimodal experiment (*i.e.,* with no counting) in the same session, which was used for comparison ($N = 8$)(original encoding, *i.e.,* "0" in above list). A subset of the same participants performed a visual search distraction task in a second session ($N = 6$; randomly chosen from the 8 participants above) (Figure 3.04). These participants searched for an F within 50 E's randomly placed in a 100-by-100 location grid in a single image. The E and F locations were jittered vertically and horizontally by up to 50 pixels. The F was present in half of trials, and absent in half. The image to be searched was presented on screen until participants responded to the visual search question. The visual search image was 10 inches by 10 inches, and each letter was 0.25 inches by 0.5 inches on screen. Participants sat about 25 inches from the 27 inch iMac screen where the images were presented. The vOICe sound played at the beginning of the visual search task. The participant was encouraged to continue searching while the sound was played. Participants then matched the vOICe sound to one image of three images displayed (3AFC, same design and image specifications as the bimodal matching experiment).

The tactile auditory mappings were tested via a bimodal matching task (Naïve sighted $N = 2$, Naïve blind $N = 2$, Trained blind $N = 2$ (both late blind)) (Figure 3.05). The set of the experiment was similar to the visual auditory bimodal matching. First, three to four tactile patterns (4 inches by 3.25 inches) were explored and the associated

vOICe sounds were played in random order, as a preview. Then, participants listened to one of the vOICe sounds and matched it to one of three tactile patterns presented on a desk surface (3AFC). Participants were asked to match the image and tactile pattern that carried the same information. The tactile-auditory matching task instructions were read aloud to the blind or blindfolded sighted by the experimenter and the participant's responses (conveyed orally) were inputted by experimenter. Tactile stimuli were placed in front of the participants on a desk surface for exploration by the participant. Tactile patterns used were generated from black and white images containing two brightness levels, by adhering cardstock to the white regions, thereby raising them relative to the black by about 1 millimeter. Images of all tactile relief patterns are presented in Figure 3.11. The trained blind participants are the same participants as the naïve blind participants.

Sighted naïve participants also performed a vOICe memory task (mimicking the vOICe training tasks) ($N = 4$) for a between group comparison. Initially, the sounds from vOICe were played in random order twice, and a label (1-4) was given to each of the sounds. Then, in each trial, one of the sounds would play again and the participant would respond with the number that matches that sound. This memory task was performed on the same sets of images that were used for the bimodal matching task.

Participants performed all tasks at a 27-inch iMac computer station with Sony noise-cancelling headphones (MDR-NC7), and inputting responses into a keyboard. Psychophysics Toolbox and MATLAB were used to code the presentation of instructions and stimuli as well as recording responses. Images were presented in black and white on the iMac screen (image size: 4 inches by 3.25 inches) approximately 25 inches away

from the seated participant. Images were encoded into vOICe sounds using vOICe software from seeingwithsound.com using a 1 Hz scan rate. Screen brightness and audio loudness was set to be comfortable to the participant. Images used were retrieved on the internet or generated by experimenter in Adobe Illustrator. Images retrieved from the internet were occasionally modified in Adobe Illustrator or Adobe Photoshop.

All trained participants were trained for 8 days on the vOICe device on basic object localization and recognition as well as two constancy tasks (rotation and shape constancy). For more details, see Appendix B and Chapter 2 Methods, (p. 62-65). The vOICe device used a camera embedded in a pair of sunglasses or a webcam attached externally to glasses. Sighted participants were requested to close eyes during training and evaluation, and wore opaque glasses and/or mask. A camera provided live video feed of the environment, and we used a small portable computer to encode the video into sound in real time.

Complexity quantification was performed in MATLAB. Images were filtered with the Laplacian of Gaussian method (edge function) and then averaged to a single number per image that was averaged across an image set. The resulting number was correlated with the bimodal audiovisual matching performance.

ANOCOVA and correlation analyses were performed in MATLAB using the aoctool and corr functions.

Results

In the original bimodal matching task (matching images to sound with vOICe encoding), naïve sighted participants ($N = 5$ to 7 participants, varied across stimulus sets)

performed significantly above chance (*i.e.,* $p < 0.05$) in 12 of 24 image sets tested, and trained sighted participants ($N = 4$) in 16 of 24 image sets (See Figure 3.06 and Appendix 1; Appendix 1 includes all images tested). Even with the strict Bonferroni multiple comparisons correction (*i.e.,* $p < 0.0021$), 5 of 24 image sets were above chance for naïve, and 8 out 24 for trained.

The image sets tested can be divided into three groups: Artificial images (generally simple and generated by myself; Appendix A, Table A and B), non-modified natural stimuli (such as flowers, forests, natural textures; Appendix A, Table A and B), texture interfaces (natural textures artificially combined to generate interfaces; Appendix A, Table C). In the artificial stimuli, 6 out of 9 image sets (67%) are significantly above chance (*i.e.,* $p < 0.05$) for naïve sighted and 7 (78%) for trained sighted. If just non-modified natural stimuli are counted, of 7 image sets, 2 image sets (29%) were significantly above chance (*i.e.,* $p < 0.05$) for the naïve sighted, and 5 image sets (71%) for the trained sighted. Finally, for the texture interface group, 4 of 8 image sets (50%) are significantly different from chance (*i.e.,* $p < 0.05$) for the trained and naïve. Therefore, the artificial stimuli seem to be the strongest group for matching images and sounds in both naïve and trained, likely due in part to their simplicity (for example: A single line or dot on a black background).

When the naïve and trained are compared directly, only in 1 image set out of 24 was the naïve performance significantly different from the trained performance (row 1 of Table C in Appendix A, $p < 0.01$). The image set is a set of texture interfaces for jeans and wood floor texture. It is useful to note that this image set for naïve *vs.* trained does not survive the Bonferroni multiple comparisons correction (*i.e.,* $p < 0.0021$). When the

results for each image set are averaged across naïve participants and then trained participants, these averages were found not to be significantly different for the naïve *vs.* trained participant groups ($p < 0.30$). Therefore, surprisingly, the naïve and trained groups are quite similar in their bimodal matching performance.

It was an unexpected result that natural stimuli could be intuitive to interpret with sensory substitution. Natural stimuli (such as a natural texture) have more spatial frequencies and brightness variation than the typical simplified lab image (a vertical line, for example). Most participants being trained on sensory substitution as reported in the literature begin with a simplified lab environment, such as an white isolated object on black felt background, and only experience a natural environment with the device after at least several training sessions. Our study indicates that this approach to training could be flawed. We have found that some natural stimuli (such as natural textures) are rich in crossmodal correspondences, and therefore are easy to interpret with vOICe. It might be better to begin training participants with a crossmodal correspondence-rich environment that includes both natural texture tasks and the simplified lab tasks.

Crossmodal mappings underlie the vOICe encoding intuitiveness. While this is a logical conclusion from the results in Figure 3.06, it is not explicitly proven that crossmodal mappings are the critical element that makes vOICe understandable to the entirely naïve. Further, it is unclear which mapping within the vOICe encoding is the most important for accurate interpretation. To address these issues, we reversed each of the primary vOICe encodings or crossmodal mappings, and then tested the new reversals in comparison to the original vOICe encoding. If the encoding or crossmodal mapping reversal significantly reduces the participants' accuracy at matching images and sounds,

then that mapping is important to correctly naïvely interpret vOICe.

Results from 8 sighted naïve participants (in Figure 3.07) indicate that two crossmodal correspondence inversions have a significantly reduced accuracy compared to the original encoding. The correlation of brightness and loudness was significantly less accurate when reversed for two most real-world-like image sets: Interfaces ($p < 0.00$) and Natural Textures ($p < 0.04$) (second and third image sets in Figure 3.07). The *XY* orientation of the encoding (scanning left to right, and high pitch at the top of the image) was also significantly less accurate when reversed (scanning top to bottom, and high pitch on right of image) for one image set: Bars of different thickness ($p < 0.00$) (first image set in Figure 3.07). When all the images are summed together, both the mapping of brightness and loudness ($p < 0.01$) and *XY* orientation ($p < 0.00$) when inverted had significantly less accurate performance than the original encoding (Figure 3.08).

The implications of the crossmodal mapping tests are that two encoding elements are particularly important to image interpretation with vOICe: Brightness correlating with loudness, and the *XY* orientation of the encoding (*i.e.,* the scanning from left to right rather than top to bottom, and high pitch with the top of the image rather than the right). It appears that the reversal of the encoding from top to the bottom or from the left to the right can be tolerated, but the switching of the *Y* and *X* axis encodings is problematic to interpretation. The problem of switching *Y* and *X* axis encodings further emphasizes the anisotropy of the vOICe encoding (unlike vision) and the importance of displaying information on the *X* axis, where the highest resolution occurs (rather than the *Y* axis). In particular, the images that test well with vOICe have information displayed horizontally, and when the *XY* encoding is switched, the information in the *X* direction is

less detectable by the lower *Y*-axis encoding resolution, thereby reducing accuracy. The value of brightness correlating with loudness makes sense, as most bright objects in a dark area are the most interesting (rather than vice versa). However, its value is also fortified by the auditory system's acute ability to recognize the presence of sounds, and its inability to recognize the absence of sounds. Therefore, the combination of these two facts makes the brightness translation to loudness highlight the most important image elements (*i.e.,* the bright elements), whereas the reverse encoding (darkness translates to loudness) obscures the most important elements.

The interpretation of vOICe does not require explicit knowledge of the sound-to-image encoding; however, this doesn't fully prove that vOICe interpretation is effortless. The vOICe interpretation relies upon crossmodal correspondences (as highlighted in the previous experiments), and crossmodal mapping interpretation can be automatic or require attention (discussed in Chapter introduction). Therefore, the automaticity was tested for naïve interpretation of vOICe sounds with an attention distraction experiment. The audio distraction task used for vOICe was counting backward in sevens while the vOICe sound was played (experiment detailed in methods). The visual distraction task was a visual search task, where participants searched for an F within 50 E's. The dual task matching accuracy (both audio and visual) was not significantly different from the original vOICe bimodal matching task for any of the 4 image sets tested (Figure 3.09) ($N = 8$). When the data are summed across image sets, the visual and auditory distraction task accuracy were both still not significantly different from the original bimodal matching task (auditory distractor: $p < 0.08$, visual distractor: $p < 0.31$). Therefore, this result shows that naïve vOICe interpretation is independent of attention load. This fulfills

one important criterion of automaticity and indicates that naïve vOICe interpretation is effortless in at least one measure.

Does image complexity matter to the untrained participants' performance? To examine this, we defined image complexity by an edge metric that quantifies the number of vertical and horizontal edges. The trained and naïve sighted participant performance both weakly anti-correlated with complexity, as measured by the edge metric (Naïve participants: $rho = -0.3491$ $p < 0.09$; Trained participants: $rho = -0.3858$, $p < 0.06$) (Figure 3.10). This result indicates that complexity may make images less intuitive to interpret. However, more importantly, a linear fit to the data indicated a performance above chance at even large complexity values for the naïve and trained participants. The trained and naïve anti-correlations with complexity had slopes and intercepts that were not significantly different from each other (ANOCOVA analysis, $p_{slope} < 0.73$, $p_{intercept} < 0.27$). It is likely that "complexity" can partially mask the crossmodal correspondences or dilute the crossmodally relevant information with unimodal noise. Nonetheless, some of the more "complex" stimuli such as natural textures revealed way-above chance performance that is likely due to direct selection of a high density of crossmodal mappings (such as coarse to fine spatial frequencies) (Figure 3.02 and Appendix A).

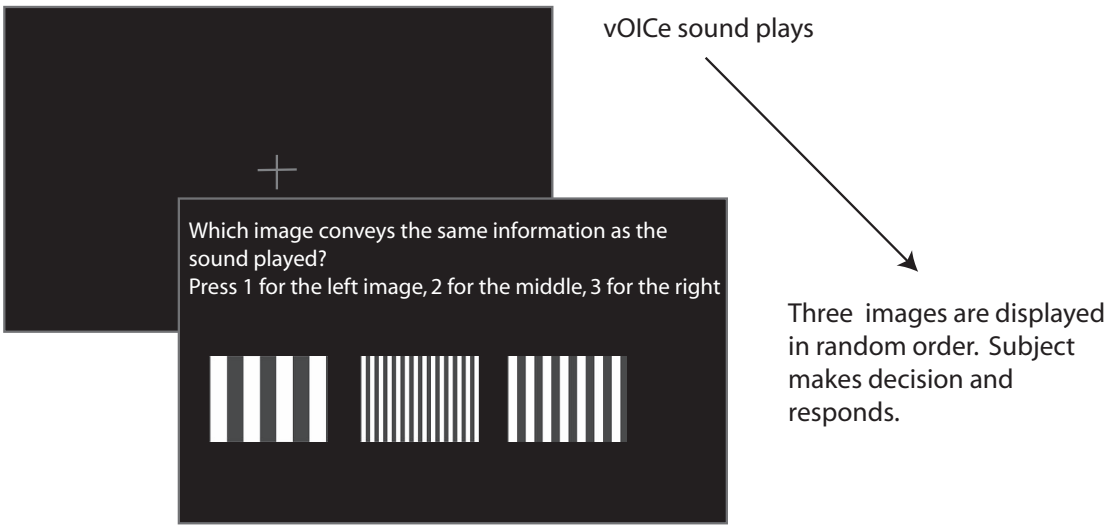


Figure 3.02. Experiment design for visual-auditory matching. As detailed in methods, participants performed matching the images and vOICe sounds while at a computer. First a vOICe sound would play, and then participants would be required to choose an image that seemed to match that sound the best, or contained the same information. Sighted participants responded by inputting a number into the keyboard: 1 for the left image, 2 for the middle image and 3 for the right image.

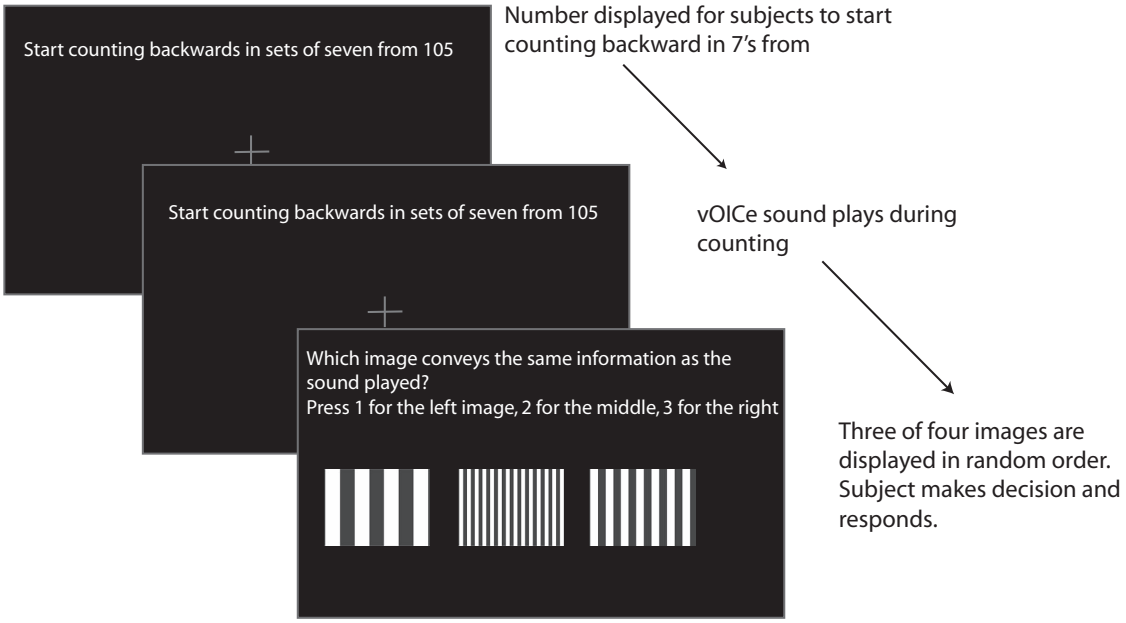


Figure 3.03. Experiment design for auditory distraction during visual-auditory matching. During the auditory distraction version of the auditory-visual matching of images to vOICe, participants were distracted by counting backward in sets of seven. The experiment was designed such that participants count backwards (beginning with the number presented on the screen), and during counting a vOICe sound plays. The final task is for the participants to match the sound heard while counting to one of the three images presented. Participants responded by inputting to the keyboard: 1 for the left image, 2 for the middle image, and 3 for the right image.

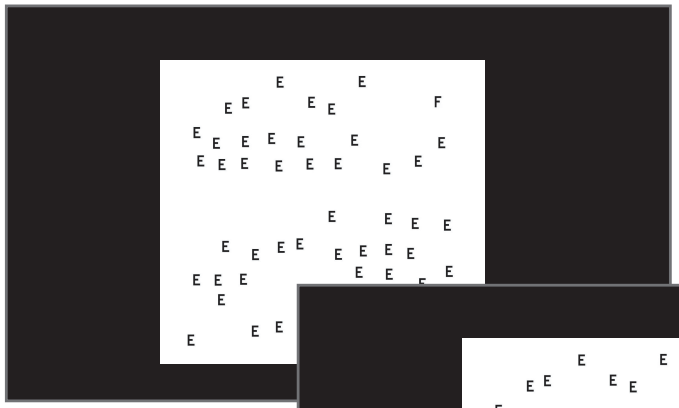
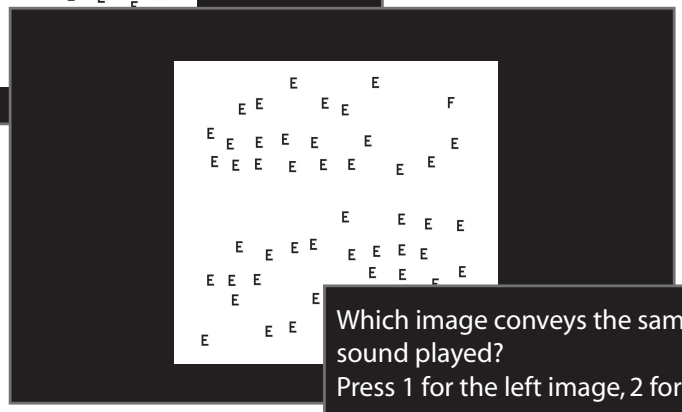
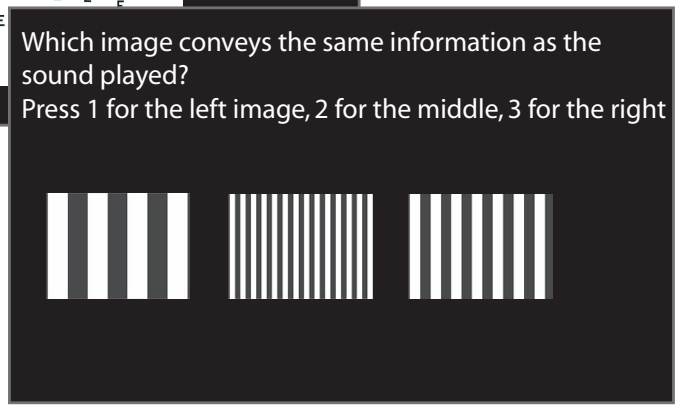


Image of E's and F's is displayed for subject to start searching for the F. vOICe sound plays.



Subject responds if an F is present



Three of four images are displayed in random order. Subject makes decision and responds.

Figure 3.04. Experiment design for visual distraction during visual-auditory matching. During the visual distraction version of the auditory-visual matching of images to vOICe sounds, participants were distracted by searching for an F within a field of 50 E's. While searching for the F, a vOICe sound is played. The participants finished the searching task by inputting to the keyboard 1 if an F is present, and 2 if an F is absent. The second task then appears, wherein the participants are required to match the vOICe sound played while searching to one of three images presented. To complete the matching task, participants input to the keyboard 1 for the left image, 2 for the middle image, and 3 for the right image.

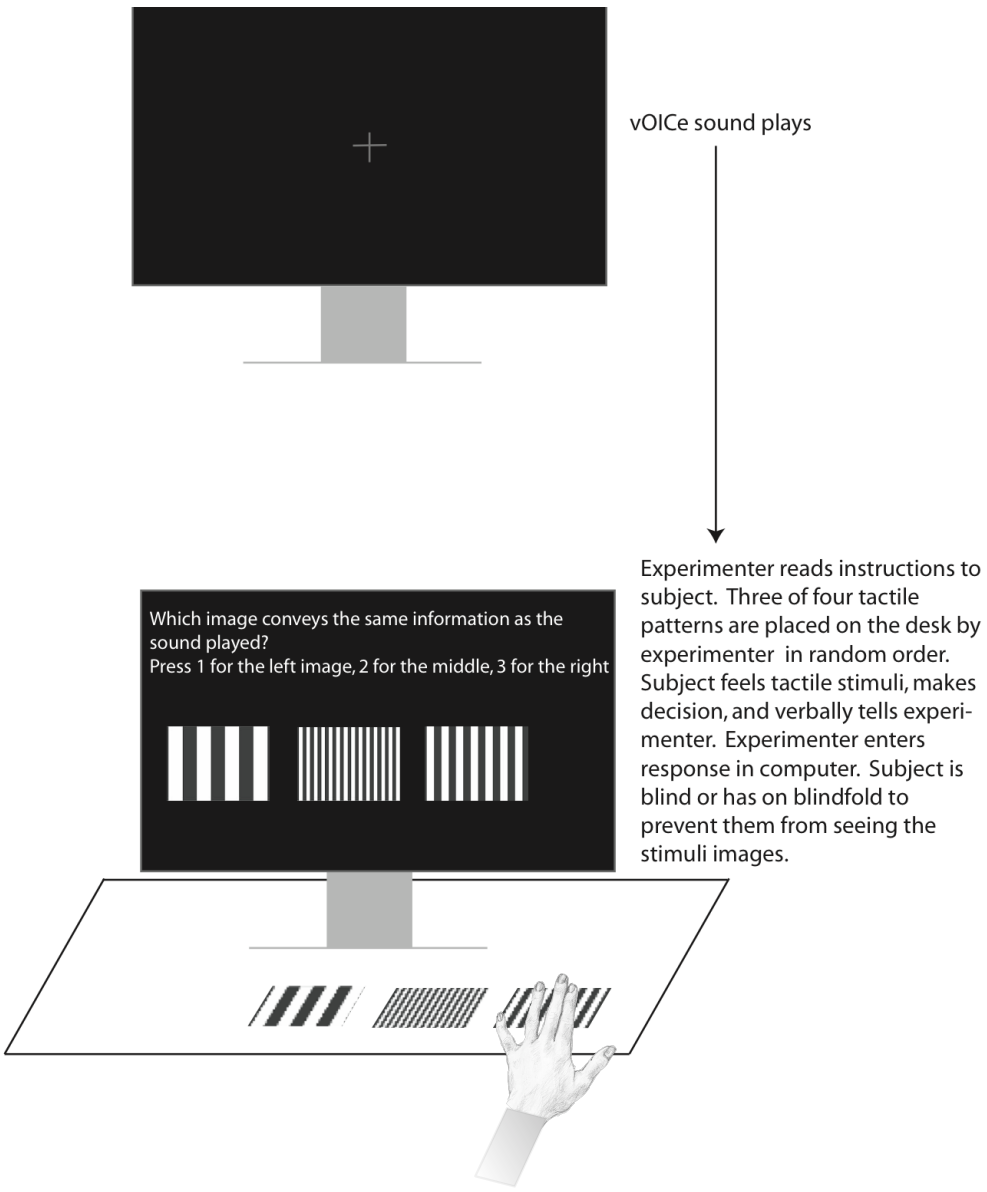


Figure 3.05. Experimental design for tactile-visual matching. Blind and blindfolded sighted participants were read the instructions for the task by the experimenter. The task began with a vOICe sound playing in headphones; then, three tactile patterns would be placed in front of the participant for tactile exploration. The participant indicates the chosen pattern, and the experimenter enters the corresponding number in the computer.

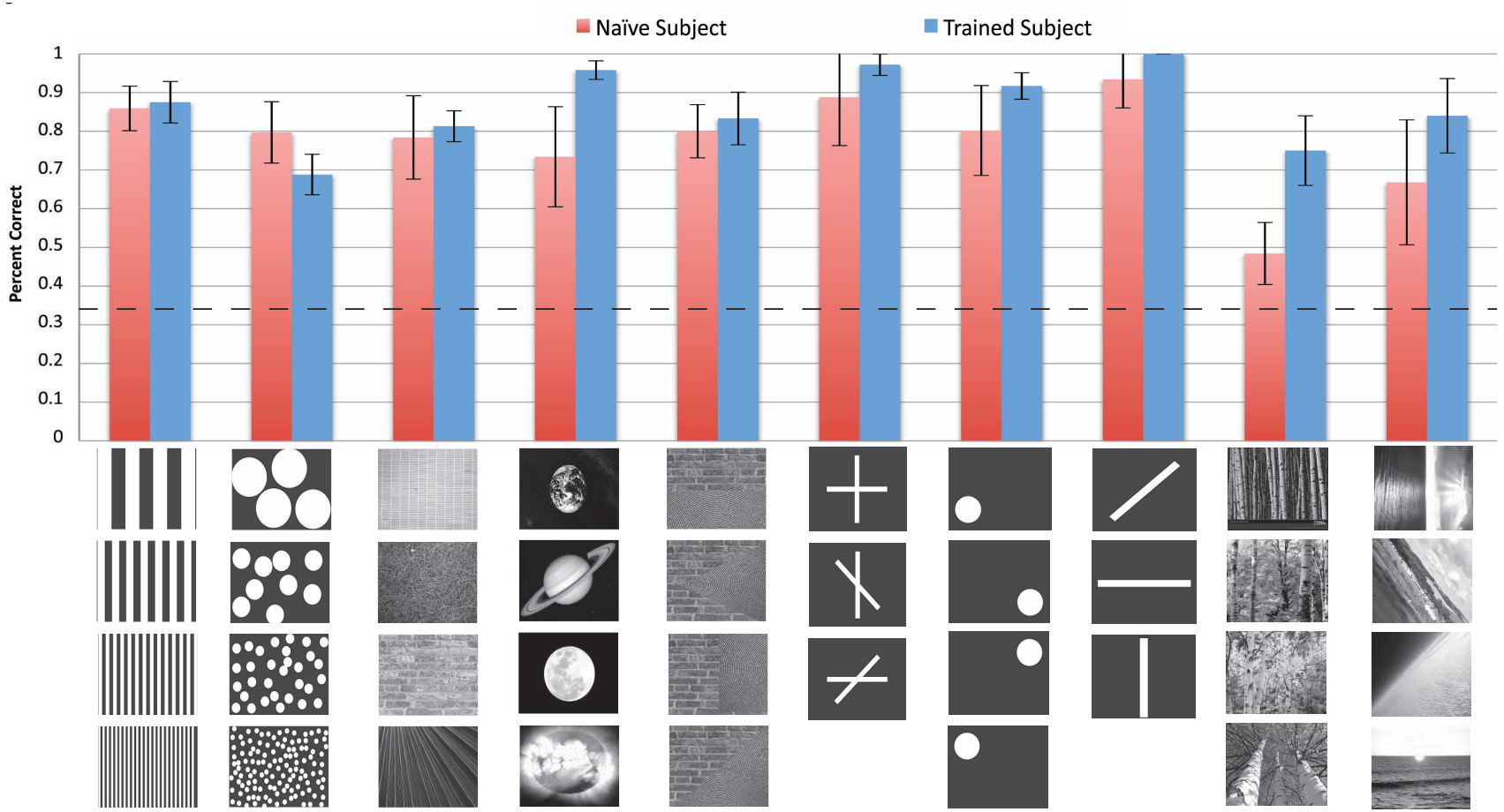

Naïve Subject
Trained Subject
Percent Correct
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure 3.06. Select vOICe data and images. Data and images from a select set of images encoded into vOICe sounds and tested on naïve and trained sighted participants. Participants were tested at matching a vOICe sound to the corresponding image out of three presented. The error bars are the standard deviation across participants. All data presented in Figure 2B is significantly different from chance ($p < 0.05$), except the naïve percent correct for the last two image sets on the right (*i.e.,* trees and horizon images).

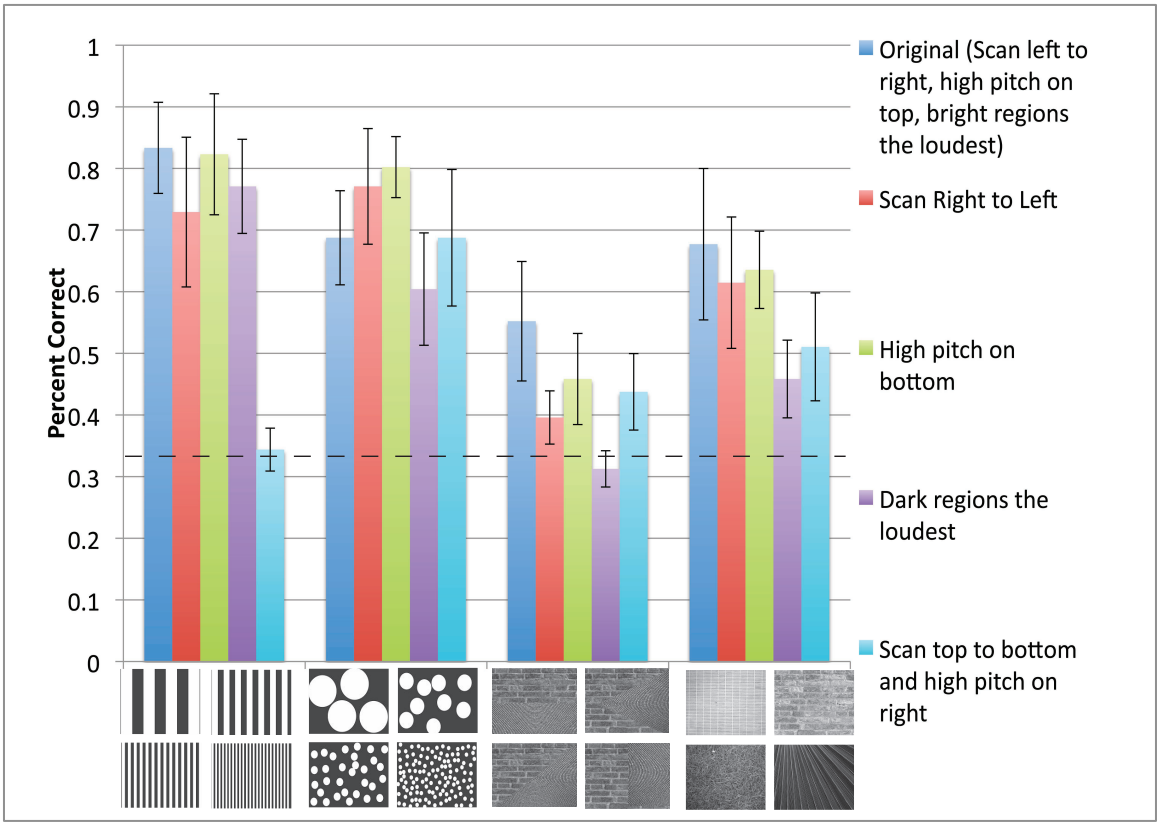


Figure 3.07. Tests of vOICe crossmodal mappings. Modifications in the vOICe auditory to visual mapping were tested with naïve participants to determine each of the crossmodal mappings' importance. The error bars are the standard deviation. The dashed line is chance.

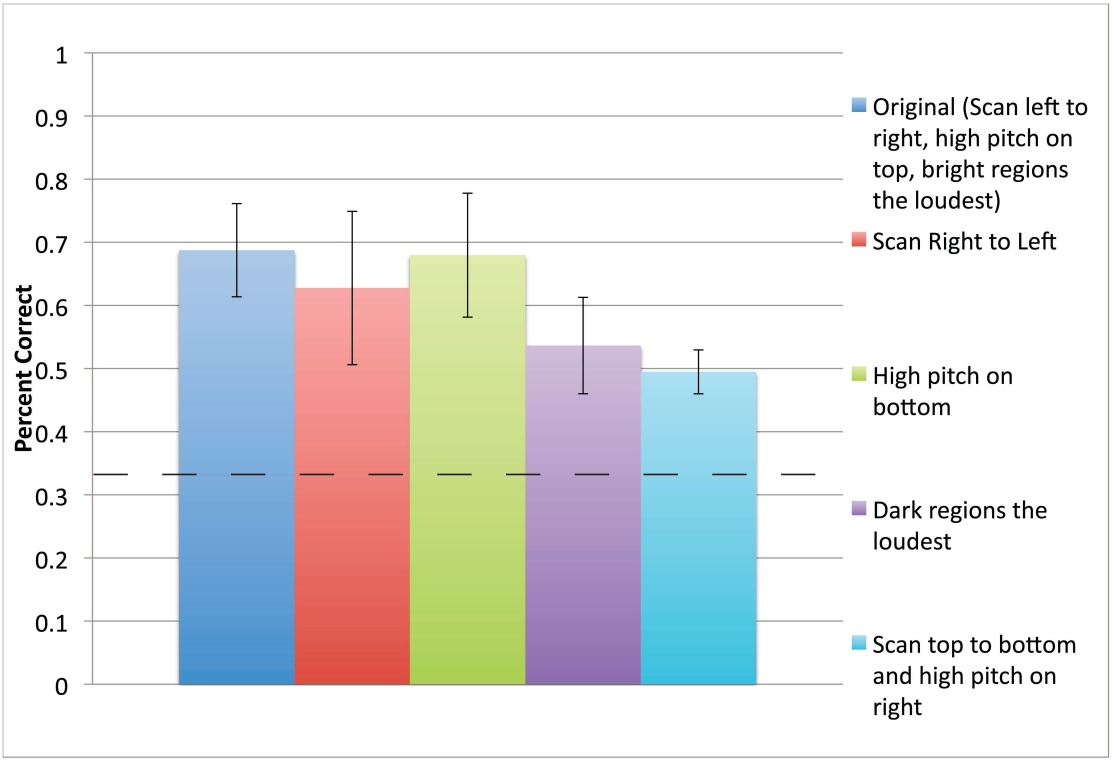


Figure 3.08. Tests of vOICe crossmodal mappings summed across images. Modifications in the vOICe auditory to visual mapping were tested with naïve participants to determine each of the crossmodal mappings' importance. The images sets were averaged together to generate a generalized percent correct for all four image sets tested (Figure 3.07 shows individual image set data). The error bars are the standard deviation. The dashed line is chance.

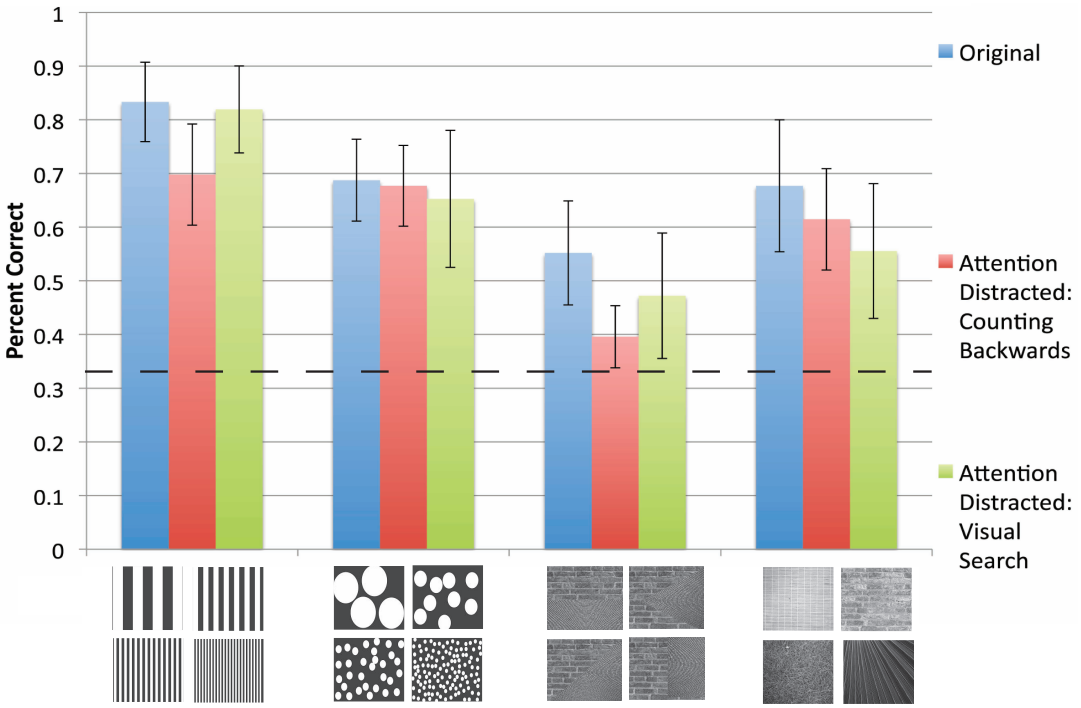


Figure 3.09. Auditory and visual attention distraction vOICe data. Naïve (untrained, and no encoding knowledge) participants matched vOICe sounds with images while performing a distraction task (either counting backward in sets of 7 from a random number [$N = 8$] or visual search [$N = 6$]). Participants then matched the sound heard to 1 of 3 images displayed. The attention distraction data is compared to the original matching of sounds to images without distraction in the same participants. Error bars are the standard deviation, and the dashed line is chance.

The naïve sighted participants can perform marvelously well matching visual images to sounds, but the real question relevant to sensory substitution should be whether the same (multimodal mappings) can be applied to, say, auditory and tactile modalities in naïve blind participants. Thus, we tested blind participants on matching sounds to tactile (relief) patterns that corresponded to the visual patterns described above for lines of different thicknesses and circle patterns of different sizes, and they also performed above chance (Figure 3.11, Bars of different thickness: Late Blind Naïve ($N = 2$) 50%, Late Blind Trained ($N = 2$) 71%, Sighted Naïve ($N = 2$) 67%; Dots of different sizes: Late Blind Naïve ($N = 2$) 50%, Late Blind Trained ($N = 2$) 71%, Sighted Naïve ($N = 2$) 63%; Chance 33%). Although the late-blind data for tactile-auditory matching is weaker than the sighted data for auditory-visual matching, the late-blind will also likely have a hidden and untestable vision-audition intrinsic mapping from past visual experience that does not appear on the tactile-audition matching test performance. Such a hidden visual-auditory mapping may assist or facilitate in the learning of vOICe by the late blind.

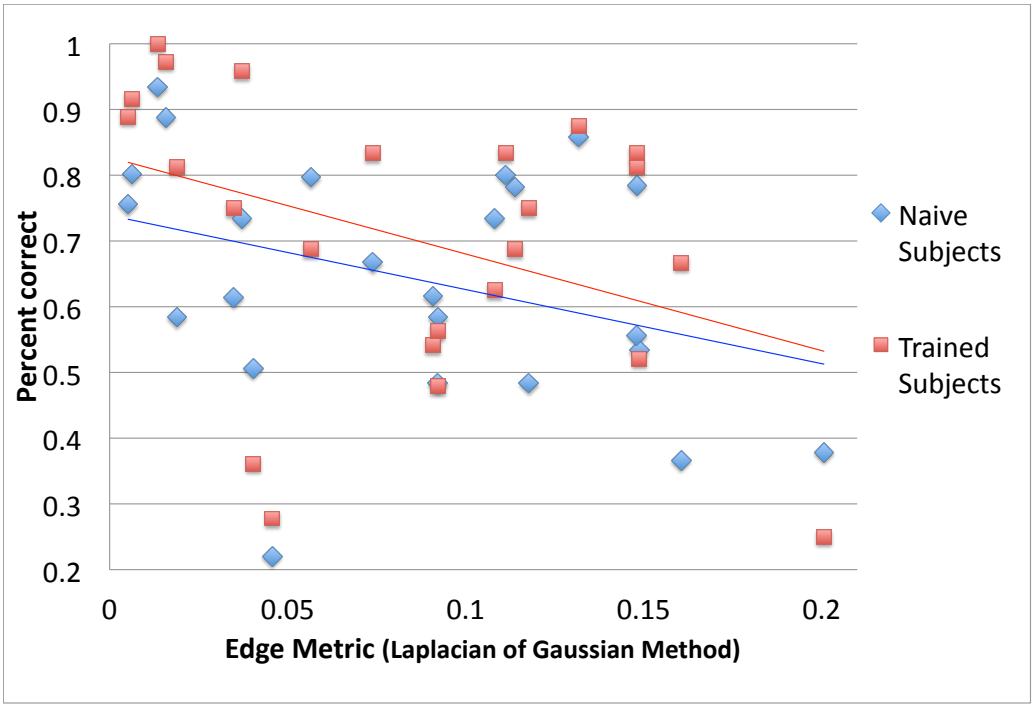


Figure 3.10. Correlation between bimodal matching data and edge metric. Correlation data: Naïve Participants: $rho = -0.3491$, $p < 0.09$; Trained Participants: $rho = -0.3858$, $p < 0.06$. Edge metric calculated in MATLAB by filtering images for edges and then averaging all pixels.

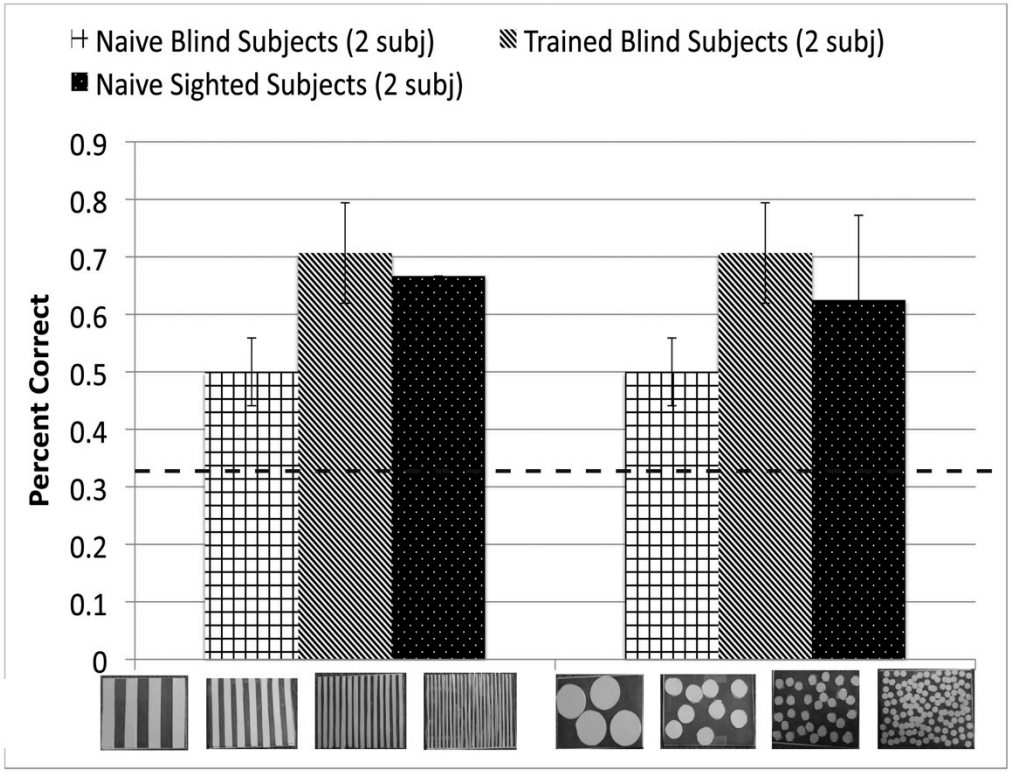


Figure 3.11. Data and images from matching of vOICe sound and tactile patterns. The tactile patterns are derived from image textures previously tested. Participants were tested at matching a vOICe sound to the corresponding tactile pattern out of three presented. The error bars are the standard deviation across participants. The white regions of the tactile patterns are raised relative to the black regions.

The matching experiments demonstrated that participants have the ability to crossmodally match vOICe sounds and images. It was yet unclear whether this crossmodal ability affects more conventional, unimodal (*i.e.,* just auditory) training with the device. To demonstrate the relationship between vOICe training and crossmodal matching ability, naïve sighted participants also performed a memory task with the same stimuli as in the bimodal matching task (detailed above). Participants were told a label (1-4) to remember for each sound, and then asked to recall the label when a random one of vOICe sounds was played. The memory task format is similar to most sensory substitution training tasks. There, participants are presented with an object or stimulus and allowed to explore or listen to it, and then told a label such as "pencil" or "square." The participant would be asked later whether they could identify the objects when presented in random order. Such a memory-based label task is in the same format as our memory task with the intuitive sensory substitution stimuli. Participant performance on this auditory memory task (chance: 25 percent) correlated significantly with the performance on the crossmodal matching task (chance: 33 percent) with a *rho* of 0.7139 ($p < 8.8 \times 10^{-4}$) (Figure 3.12). The result therefore indicates that the participants' ability to remember and interpret sensory substitution stimuli correlates significantly with the density of crossmodal mappings (as measured by our crossmodal matching task). Therefore, crossmodal intrinsic mappings provide a common basis for sensory substitution training as well as adaptive behavior and scene perception in the real world with the device. Crossmodal correspondences are the unrecognized common key to the relative intuitiveness/ease of existing vOICe training tasks.

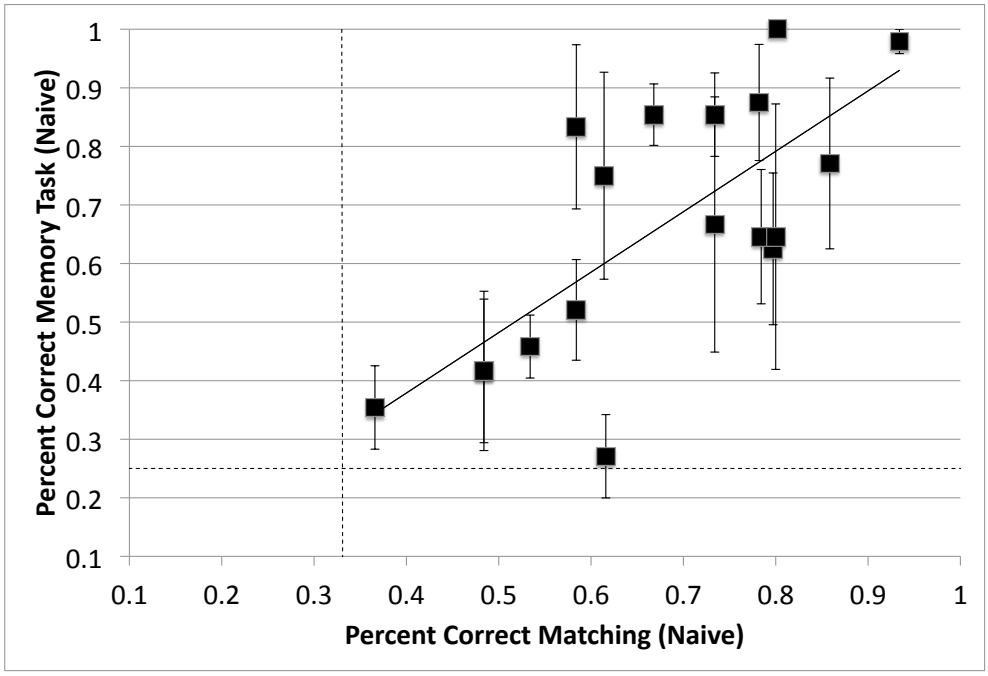


Figure 3.12. Correlation between the bimodal and unimodal tasks. In the bimodal matching task, the participant matches vOICe sounds to images, and in the unimodal memory task, the participant indicates the remembered label for each vOICe sound. The memory task is the same as most vOICe training tasks. Dashed lines are chance for each of the tasks.

Discussion

Sensory substitution training has a hidden assumption that the primitives of sensory substitution perception will be the same as the primitives of vision, such as dots, lines and intersections. While sensory substitution is vision-like, it may have crossmodally intuitive primitives that are different from the classical visual primitives, and should not be overlooked. Training protocols that are specially designed to access intrinsic mappings as primitives may enable faster training and more ease of use. If intuitive stimuli such as textures are the starting point of vOICe training, followed by the gradual increase of image complexity (but also closer to the real-world), participants may be able to learn to use devices more effectively and effortlessly with a shorter training period. Training could also use image-processing filters to heighten textures in the natural images (such as a high pass filter), thereby making them more intuitive. Note that this is a grossly different approach from the conventional (more effort-demanding) training, where trainees are forced to learn geometric primitives and then more natural cluttered scenes constructed from these primitives.

This study indicates that participants can interpret vOICe stimuli with no knowledge of the audiovisual encoding. The strongest crossmodal correspondences that underlie this naïve vOICe interpretation were found to be brightness to loudness mapping and the *XY* mapping orientation. Finally, the naïve interpretation of vOICe was shown to be automatic (attentional load insensitive) with a dual task design.

Sensory substitution interpretation and functional ability is generated by multimodal interaction and crossmodal plasticity. Crossmodal mappings are the foundation of sensory substitution interpretation, and if used intelligently in device

training and design, could dramatically improve functional outcomes. The fundamental bottleneck towards a commercial product may be removed by vigorous crossmodal plasticity kick-started from such an advantageous start point.