available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Generalized learning of visual-to-auditory substitution in sighted individuals

Jung-Kyong Kim, Robert J. Zatorre*

Department of Neuropsychology, Montreal Neurological Institute, McGill University, Room 276, 3801 University Street, Montreal, Quebec, Canada H3A 2B4

ARTICLE INFO

Article history:
Accepted 11 June 2008
Available online 20 June 2008

Keywords:
Visual-to-auditory substitution
Generalization
Sighted
Training
Rule learning
Explicit knowledge

ABSTRACT

Visual-to-auditory substitution involves delivering information about the visual world using auditory input. Although the potential suitability of sound as visual substitution has previously been demonstrated, the basic mechanism behind crossmodal learning is largely unknown; particularly, the degree to which learning generalizes to new stimuli has not been formally tested. We examined learning processes involving the use of the image-to-sound conversion system developed by Meijer [Meijer, P., 1992. An experimental system for auditory image representations. IEEE Trans Biom Eng. 39 (2), 112-121.] that codes visual vertical and horizontal axes into frequency and time representations, respectively. Two behavioral experiments provided training to sighted individuals in a controlled environment. The first experiment explored the early learning stage, comparing performance of individuals who received short-term training and those who were only explicitly given the conversion rules. Both groups performed above chance, suggesting an intuitive understanding of the image–sound relationship; the lack of group difference indicates that this intuition could be acquired simply on the basis of explicit knowledge. The second experiment involved training over a three-week period using a larger variety of stimuli. Performance on both previously trained and novel items was examined over time. Performance on the familiar items was higher than on the novel items, but performance on the latter improved over time. While the lack of improvement with the familiar items suggests memory-based performance, the improvement with novel items demonstrated generalized learning, indicating abstraction of the conversion rules such that they could be applied to interpret auditory patterns coding new visual information. Such generalization could provide a basis for the substitution in a constantly changing visual environment.

1. Introduction

Our experience of the world is largely multimodal. We are able to perceive multiple sensory events as a single event not only because the perceptual system can integrate information gathered from different sensory modalities, but also because it allows one to extract information from one sensory modality and apply it to another modality. Sensory substitution involves replacing one sensory input by another, and is based on the idea that information from one sensory modality generates percepts related to those induced by a different sensory modality. Pioneering work in sensory substitution

* *Corresponding author*. Fax: +1 514 398 1338.
E-mail address: jkim@ego.psych.mcgill.ca (J.-K. Kim).

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.06.038

examined visual replacement by touch (Bach-y-Rita et al., 1969; Bach-y-Rita, 1972), and demonstrated that tactile information can be used to transmit visual information (also see Bach-y-Rita et al., 1998; Ptito et al., 2005; Sampaio et al., 2001). More recent work has involved visual substitution via audition, and showed that auditory input can also be used as a suitable replacement for vision (Capelle et al., 1998; Meijer, 1992). However, the mechanisms behind crossmodal substitution learning are still largely unknown, particularly with respect to the question of whether abstract crossmodal rules can be learned. The goal of the present study was to examine visual-to-auditory substitution learning in sighted individuals in a controlled and systematic fashion. Our main focus was to examine the role of explicit knowledge vs. training in substitution learning, and the extent to which generalization occurs, an issue that has not been fully examined in previous studies, but which is critical to understanding whether learning involves abstraction of generalized rules.

Visual-to-auditory substitution involves capturing real-time video images in two-dimensional snapshots and translating them into complex sounds (Capelle et al., 1998; Meijer, 1992). This may provide a suitable replacement for vision because the auditory system is capable of processing complex, rapidly changing acoustical patterns such as music and speech. Visual-to-auditory substitution devices can offer considerably higher resolution than the tactile systems (e.g. Bach-y-Rita, 1972; Bach-y-Rita et al., 1998) because the resolution in the auditory system in time and frequency is higher than that afforded by the somatosensory receptors in the skin or the tongue. In addition, the auditory system is not limited to perispace, and no three-dimensional texture of scenes would be needed to translate visual information to auditory input.

The two major visual-to-auditory substitution devices that have been developed and studied are the Prosthesis Substituting Vision with Audition (PSVA) developed by Capelle et al. (1998) and Arno et al. (1999) and the vOICe developed by Meijer (1992; www.seeingwithsound.com). These devices code pixels in a two-dimensional scene into an auditory signal. The PSVA translates both vertical and horizontal dimensions into frequency while the vOICe codes the vertical dimension into frequency and the horizontal dimension into time. The PSVA mimics the retina by using a higher resolution for the foveal area of the visual scene while the vOICe offers a uniform resolution throughout the entire scene. However, the vOICe provides a much higher resolution than the PSVA (the vOICe being capable of a resolution of 176×144 pixels vs. the PSVA of 8×8 pixels in the periphery and 8×8 in the retina).

Studies of the PSVA and the vOICe have demonstrated the potential suitability of auditory input as visual substitution. Sighted individuals who were trained with the PSVA could use the auditory input to reconstruct relatively simple two-dimensional patterns (Capelle et al., 1998; Arno et al., 1999, 2001; Poirier et al., 2006), detect the Ponzo illusion (Renier et al., 2005a,b), perceive depth from a three-dimensional virtual scene (Renier et al., 2005a,b), and experience the vertical–horizontal illusion (Renier et al., 2006). Sighted subjects who were trained with the vOICe could localize and discriminate objects (Auvray et al., 2007), and match sounds to their corresponding images (Amedi et al., 2007).

The present study examined learning processes involving the use of the vOICe developed by Meijer (1992). The goal was to provide systematic and objective training and testing of the image-to-sound conversion system to sighted individuals, and examine parameters leading to visual substitution learning. One important question we asked was whether the acquired learning after the substitution training is generalized. By generalization, we imply that the image-to-sound mapping rules are learned at an abstract level, such that they could be applied to interpret a new set of patterns. Although previously addressed in a limited way (Arno et al., 1999; Amedi et al., 2007), this issue has never been fully tested. A common paradigm in sensory substitution learning involves training of subjects on certain items and testing of the trained items in order to examine improvement in performance (e.g. Auvray et al., 2007; Sampaio et al., 2001; Ptito et al., 2005). However, the problem with using the same items in testing and training is that the improvement cannot be attributed entirely to generalized learning because it could rather be reflective of an enhanced associative memory with the familiar items. In order to test for generalized crossmodal learning, it would be crucial to evaluate substitution performance with new stimuli.

Another related issue to explore was the role of training in visual substitution performance. We examined the extent and types of visual information that could be extracted from the auditory input given a varying amount of training with the vOICe (Meijer, 1992). In order to establish the minimum amount of training that is necessary for the initial use of the substitution system, we addressed the critical question of whether having explicit knowledge of the image–sound relationship would be sufficient for extracting some visual information from the auditory input. We also examined the effect of extensive training on substitution performance over time, and whether performance improvement varies according to types of visual information.

The present study consisted of two behavioral experiments examining visual-to-auditory substitution performance of sighted individuals. The first experiment was a preliminary study in which baseline visual substitution performance was evaluated after a minimum amount of training was given. The second experiment involved long-term training in visual-to-auditory substitution by extending the training period and increasing the number and types of testing stimuli. By examining the extent of visual information that could be extracted from the auditory input and the patterns of improvement in substitution performance, we evaluated limitations of the conversion system, and considered the feasibility of using the visual-to-auditory converted information as a suitable means of visual substitution.

Since findings of visual substitution studies would have the most practical implications for blind people, previous studies focused mostly on making the learning conditions comparable to the environment in which blind people would use the visual substitution systems. Training procedures would normally involve visual deprivation (by being blind or blindfolded), and engaging in sensorimotor interactions by using feedback cues received by actively walking around the space, or by head movements with a head-mounted camera (e.g. PSVA: Arno et al., 1999, 2001; Porier at al. 2006; vOICe: Pollok et al., 2005; Auvray et al., 2007). However, substitution learning in the

absence of sight does not inform whether visual deprivation is essential, or whether the relationship between the visual environment and the corresponding auditory input must be learned only through non-visual feedback. Therefore, an examination of sighted individuals who are trained with visual feedback could provide useful information about the basic mechanism behind visual substitution learning, because their learning would suggest that the substituted visual information can be extracted on the basis of learning of the direct relationship between images and the corresponding sounds. Also, most visual substitution studies have included training paradigms where subjects used motor feedback received from the environment. However, the extent to which learning of visual information using the substitution system is necessarily dependent on the sensorimotor contingencies is not known, and therefore, a training paradigm where substitution is learned with direct visual feedback in the absence of sensorimotor cues can provide useful information about the extent to which visual-to-auditory substitution can be accomplished without motor feedback, but based on purely perceptual learning of the relationship between visual images and sounds. In sum, the present study involved giving formal instructions in a controlled laboratory setting where direct visual feedback was provided, and no motor interaction with the visual environment. This approach allowed us to study visual-to-auditory substitution learning in the simplest and most direct way by examining the extent to which the original visual information could be recovered from the auditory information.

2. Experiment 1

2.1. *Introduction: early stage of visual-to-auditory substitution learning*

The purpose of the preliminary experiment was to explore the early stage of visual-to-auditory substitution learning. We investigated the role of explicit knowledge about the image-to-sound conversion rules in the initial use of the vOICe substitution system (Meijer, 1992) by comparing performance of individuals who had no knowledge but received short-term training with feedback (referred to as the 'trained' group) and those who did not receive any training but were given an explicit verbal explanation of the conversion rules (referred to as the 'untrained' group). Subjects were tested on their ability to identify the correct visual images based on their corresponding sound transformations using a forced-choice procedure. In order to examine types of visual information that could be extracted from the auditory input, we presented sounds of abstract images that varied in one of the following three features: position, orientation, and size (see Experiment 1 methods, and Figs. 5D–F). We hypothesized that the discrimination of differentially oriented images would be most difficult to do because the visual change in orientation would not result in a systematic auditory change, unlike the systematic auditory changes that occur as a result of visual changes in position or size. We also tested the difference between familiar items, those presented during training, and novel items, not previously experienced.

2.2. *Results*

The mean proportions of correct answers on the familiar- and novel-item tasks obtained by the trained and untrained groups are shown in Figs. 1A and B. The scores on the 'familiar' items referred to the performance on the stimuli that was part of training for the trained group. The untrained group was tested on the same items, and, of course, these items were entirely new to this group, but this testing condition was referred to as the 'familiar-item' task for the consistency in naming (refer to Experiment 1 methods section for more details). Regardless of training or novelty of the stimuli, both subject groups were able to perform well above chance (.2) for all three variant conditions ($p<.05$).

A 3×2×2 analysis of variance (ANOVA) was performed with variant (position, orientation, size) and novelty of the stimuli (familiar, novel) as repeated measures factors, and training group (trained, untrained) as a between subjects factor. There was a significant main effect of variant ($F(2, 32)=21.52$, $p<.001$) and interaction between variant and training group ($F(2, 32)=4.15$, $p=.02$). No main effect of training group or novelty of the stimuli, or two-way interaction between these two factors was found ($p>.05$). There also was no significant three-way interaction ($p>.05$).

Tukey's honestly significant difference (HSD) multiple comparisons were made at the level of the significant interaction between the variant (position, orientation, size) and training group (trained, untrained) factors. The trained group performed better on position than on orientation and on size, and better on orientation than on size. The untrained group performed better on position than on orientation.

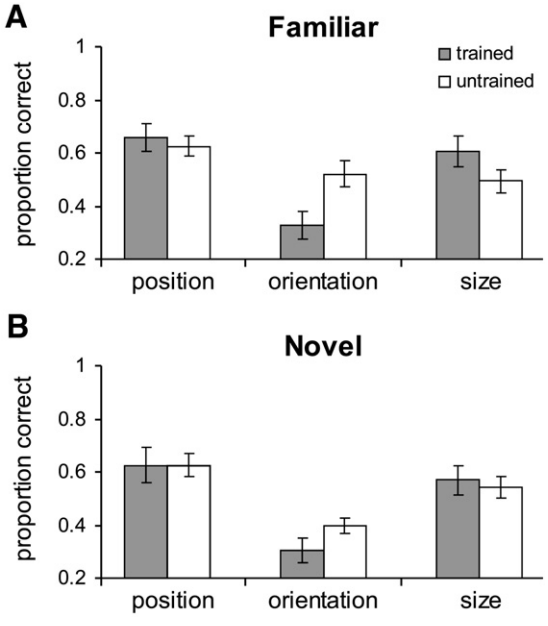

Fig. 1 – Mean performance obtained by trained and untrained groups as a function of three variant conditions (position, orientation, and size). Error bars indicate standard errors of the means. .2=chance performance. (A) Mean proportion of correct responses on familiar items. (B) Mean proportion of correct responses on novel items.

However, when the scores between the two subject groups were compared for each variant (e.g., trained-position vs. untrained-position), no difference was found ($p > .05$).

2.3. Discussion

Regardless of training or familiarity of the testing items, the performance in all the tasks was above chance, suggesting that the image-to-sound conversion is relatively intuitive, and can be applied in the initial use of the system. Furthermore, we found no performance difference in the identification of the correct visual image of a given soundscape between subjects who had no initial knowledge of the conversion rules but received short-term training with feedback and those who were only given the explicit rules without any training. This lack of group difference was true even when the trained group had the advantage of being tested on the familiar items, indicating that having explicit knowledge of the image–sound relationship was as effective as acquiring understanding of the relationship through short-term training. This finding suggests that while improvement in the performance might take place only after extensive training, a mere verbal explanation of the image-to-sound conversion rules can provide a basis for a cognitive strategy in the initial use of the system.

In an attempt to identify types of visual information that could be extracted from the converted soundscapes, we manipulated position, orientation, and size of abstract figures, and examined the performance in the identification tasks for each parameter. Subjects seemed to have relatively good judgments on position and size in particular. Differentiation of the variations in position involves a temporal judgment of the duration between the click and the start of the converted sound (i.e. the left most part of the visual image) and a relative-pitch-shift in the vertical dimension of the visual scene. The high performance in this category suggests that subjects were able to extract the pitch shift associated with the vertical translation of an object in space and the time shift associated with the horizontal translation in space.

The relatively good performance in size differentiation indicates that subjects were able to extract the auditory information that was relevant to the size manipulation. Coding of the visual images of different sizes into sounds involves a use of different frequency ranges on the vertical axis, and different durations on the horizontal axis. It is unclear whether one type of auditory information predominated, however.

The poorest performance was found in the orientation condition. This task seems to be more difficult than the others because unlike the systematic auditory changes resulting from the visual changes in position or size, no consistent correspondence exists between the changes in image orientation and the changes in the converted sounds. The soundscapes that are converted from the same-shaped images in different orientations result in drastically different sound patterns, and are consequently likely to be judged as being different shapes. As supported by the above-chance level of performance, nonetheless, subjects were successful to some extent at using the conversion rules in identifying the correct image among what probably sounded like ‘different shapes’.

To summarize, we demonstrated the relative ease of the initial use of Meijer’s (1992) image-to-sound conversion system in extracting the auditory changes from the changes in visual information such as visual variations in object size and position; having explicit knowledge of the conversion rules seems to be sufficient to extract such information at least to some extent. These findings indicate an intuitive nature to the image-to-sound mapping that could be applied without much training. It is possible that this intuition comes from a natural crossmodal correspondence that might exist between a visual image and the corresponding sound. For example, the higher pitch would be naturally associated with the higher position in space (Evans and Treisman 2005), and the louder sound with the larger object.

3. Experiment 2

3.1. Introduction: long-term training

We established in the first experiment that at the early stage of using the vOICe substitution system, sighted individuals were able to extract certain visual information from the sound transformations without requiring much training. In the second experiment, we examined the role of more extensive training in visual-to-auditory substitution learning. We provided sighted individuals with greater total duration of training and a larger variety of training stimuli under similar experimental conditions to those of the first experiment. Our goal was to demonstrate that more training results in an overall improvement of the ability to recover visual information from sounds of various visual images above and beyond that which is available via explicit knowledge of the conversion rules, and that this improvement indicates generalized learning due to training. Generalization was tested by presenting soundscapes whose corresponding visual images consisted of new visual information. By including various types of visual images, namely abstract figures, pictures of real-life objects, and pictures of scenes, we also evaluated the extent of visual detail that could be extracted from the sounds of visual images in each of these image categories (see Fig. 6). This allowed us to test learning capacity of more ecological materials. Lastly, we included a control condition where we tested a group of sighted individuals who did not receive any explicit explanations about the conversion rules or exposure to the training stimuli. Having demonstrated in the first experiment the intuitive use of the vOICe system purely on the basis of explicit knowledge, the purpose of this control condition was to exclude the possibility that completely naive subjects can obtain the understanding of the image–sound relationship without receiving any feedback. We only tested this control group on the two easiest tasks, in order to provide the best test of above-chance performance.

3.2. Results

3.2.1. Pretest scores

The mean proportions of the training group’s correct answers for each of the six pretests, *different figures*, *orientation figures*, *similar figures*, *different objects*, *similar objects*, and *scenes* were

.49 (.06), .44 (.07), .36 (.03), .47 (.05), .33 (.03), and .30 (.03) respectively (standard error in brackets). Consistent with the results of Experiment 1, these scores were all significantly greater than the chance level of performance (.20; $p<.05$).

3.2.2. *Overall improvement over time*

The average performance on the familiar and novel items is shown in Fig. 2. A one-way ANOVA on the performance on the novel items yielded a significant effect of test across time ($F(3, 21)=30.08$, $p<.001$). Tukey's HSD tests showed that the mean pretest score was significantly different from the scores of tests 1, 2, and 3 ($p<.05$). No effect of test across time was found for the familiar items.

3.2.3. *Performance based on familiarity and image types*

Fig. 3 shows the average performance on the different image types categorized in terms of the familiar and novel testing items. A 6×2 (image type×familiarity) ANOVA showed a significant main effect of familiarity, indicating better overall performance on the familiar than on the novel items ($F(1, 7)=52.14$, $p<.01$), and a significant main effect of image types, indicating different mean scores across different image types ($F(5, 35)=59.01$, $p<.01$). There also was a significant interaction effect between familiarity and image type ($F(5, 35)=12.57$, $p<.01$). Tukey's HSD tests at the level of interaction between the two factors further showed that subjects performed better on the familiar than on the novel items for the different figures, different objects, and scenes ($p<.05$), whereas no performance difference due to familiarity was found for the orientation figures, similar figures, and similar objects ($p>.05$). For the familiar items, subjects scored better on the different figures than on the orientation figures and the similar figures, higher on the different objects than on the similar objects, and higher on the different figures and the different objects than on the scenes ($p<.05$). For the novel items, subjects scored higher on the different figures than on the similar figures, higher on the different objects than on the similar objects, higher on the different figures than on the different objects and scenes, and higher on different objects than on the scenes ($p<.05$).

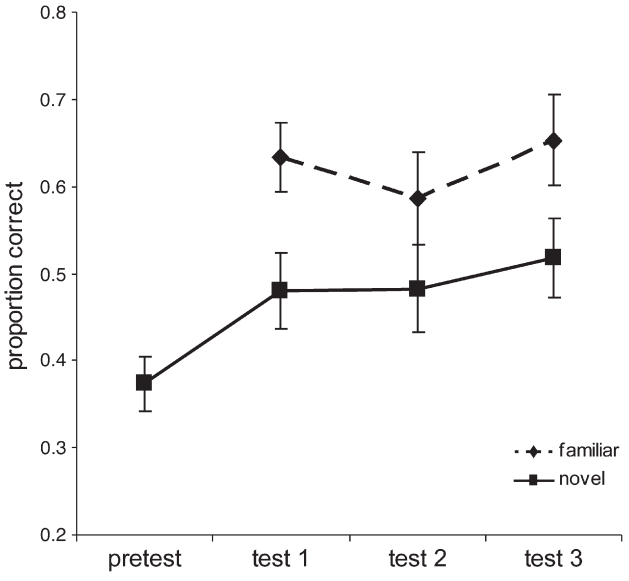

Fig. 2 – Mean performance on novel and familiar items over time measured in terms of proportion of correct responses. Each data point indicates an average of averaged scores of the different-, orientation-, and similar figures tests, averaged scores of the different-, and similar objects tests, and the score of the scenes test (i.e. averages were calculated within each image type first).

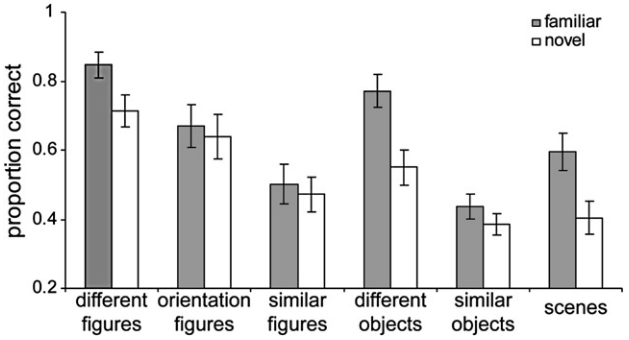

Fig. 3 – Mean performance on familiar and novel items as a function of image types averaged over time.

3.2.4. *Improvement on different images over time*

Average performances for each image type in terms of familiarity are shown in Figs. 4A–F. A one-way ANOVA was performed on the scores for each of the familiar- and novel-item tests for different image types. For the novel-item tests, all the image types except for the similar objects showed a significant effect of time (different figures: $F(3, 21)=16.83$, $p<.001$; orientation figures: $F(3, 21)=8.66$, $p=.001$; similar figures: $F(3, 21)=3.83$, $p=.025$; different objects: $F(3, 21)=5.53$, $p=.006$; scenes: $F(3, 21)=7.35$, $p=.002$). For the familiar tests, only the similar figures scores showed a significant effect of time ($F(2, 14)=6.80$, $p<.001$).

A summary of the Tukey's HSD tests of multiple comparisons for the novel tests that yielded the significant effect of time is shown in Table 1. Overall, subjects made a significant improvement on all image types (with the exception of the scenes which showed a trend toward significance) following the third week of training (i.e. after nine training sessions). After the first week of training, subjects showed a significant improvement on the different figures, orientation figures, and scenes, but no improvement from the first to the second week of training on any of the image types. Subjects made further improvements on the different figures and different objects from the second to the third week of training.

3.2.5. *Performance of control group*

The mean scores of the control group on the different figures and the different objects were .34 (.03) and .17 (.06). The score on the different figures test was significantly greater than chance ($p<.05$), but was significantly lower than the score of the training group ($t(16)=-2.18$, $p<.05$). The score on the different objects test did not exceed the chance level ($p>.05$), and was significantly different from the score of the training group ($t(16)=-5.99$, $p<.001$).

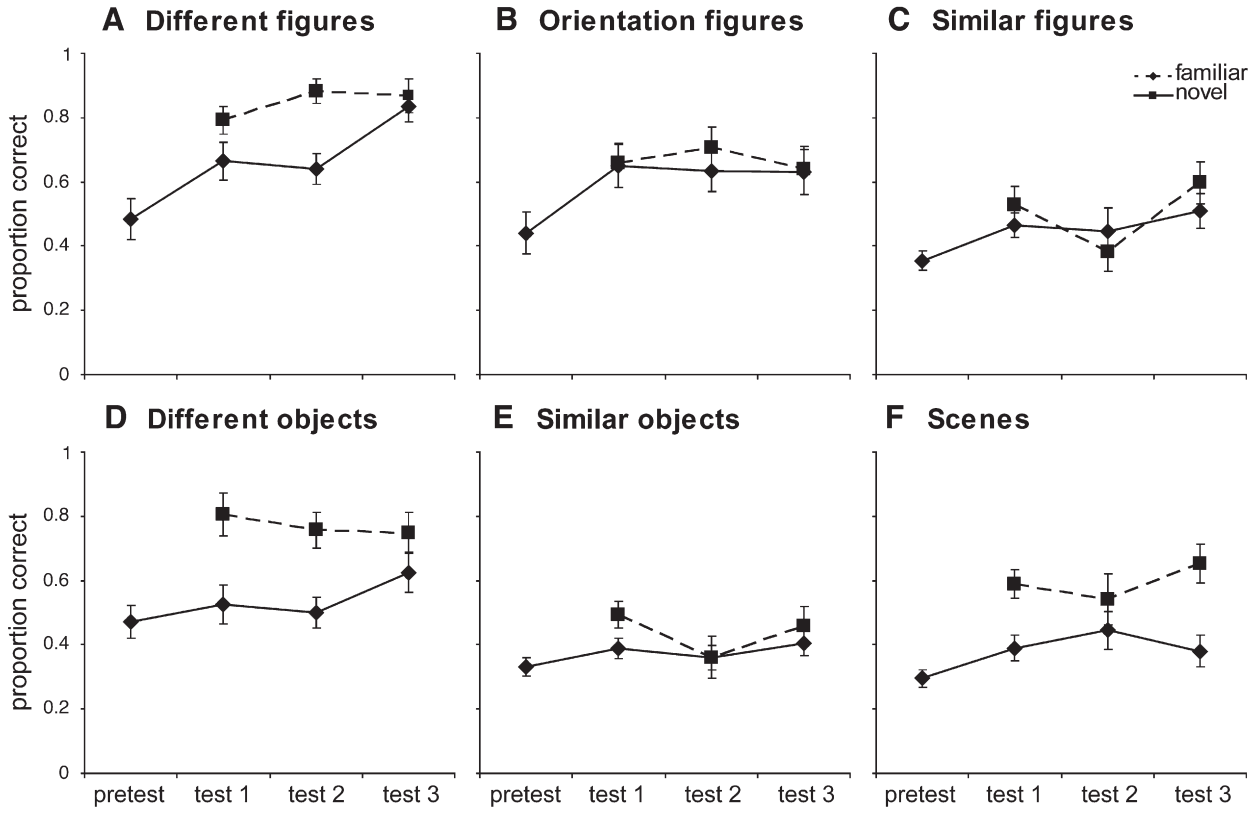

Fig. 4 – Mean performance on (A) different figures, (B) orientation figures, (C) similar figures, (D) different objects, (E) similar objects, and (F) scenes.

According to the post-testing interview on types of strategies used in doing the tasks, all the subjects reported that for the abstract figures, they based their choice of visual images on their association of pitch changes with the shapes of the figures (e.g. if they heard a sound increasing in pitch, they would choose a line that goes up in space). For the real-life objects, they reported their attempts to associate semantic knowledge about the objects with the sounds heard (e.g. if they heard a high pitch sound, they would look for an object that would produce a high pitch sound).

3.3. *Discussion*

In the second experiment, we examined learning processes involving three weeks of visual-to-auditory substitution training in sighted subjects. Consistent with the finding from Experiment 1, the scores on all the pretest tasks were above chance, confirming that some visual information could be extracted from the auditory input purely on the basis of having explicit knowledge of the relationship between the visual and auditory stimuli. But training over the longer time frame of this experiment resulted in significantly higher scores than could be achieved based on explicit knowledge alone. The control group, who was not given any information about the conversion rules, and who did not receive feedback during task performance, scored slightly above chance on the abstract figures, suggesting that they seemed to have developed some intuition about the image–sound relationship after the mere exposure to the relatively simple testing stimuli (post-test interviews confirmed this impression). However, this score was still significantly lower than that of the training group who initially received explanations on the conversion rules. Furthermore, scores on the more difficult real-life object items were no better than chance, and control subjects did not develop a correct intuition about the conversion rules in this case, confirming that deriving complex information from the substitution system is not possible if subjects have absolutely no information about the image–sound relationship.

The overall performance on the familiar items was higher than that on the novel items, but did not result in an

Table 1 – Tukey's HSD tests of multiple comparisons for the novel tests

	Pre vs. t1	Pre vs. t2	Pre vs. t3	t1 vs. t2	t1 vs. t3	t2 vs. t3
Different figures	*	*	*	ns	*	*
Orientation figures	*	*	*	ns	ns	ns
Similar figures	~	ns	*	ns	ns	ns
Different objects	ns	ns	*	ns	~	*
Scenes	*	*	~	ns	ns	ns

* = significant at p of .05. ~ = trend toward significance. Pre = pretest, t1 = test 1, t2 = test 2, t3 = test 3.

improvement over time (Fig. 2). The reason for the better performance on the familiar items without an improvement over time is most likely that rather than solely applying the image-to-sound mapping rules, subjects were likely to have also engaged memory based on an association between the correct images and the corresponding sounds formed during training. On the other hand, the performance on the novel items showed a significant improvement after nine training sessions. Subjects could not have used the association-based strategy on the untrained items because these stimuli had never been presented during training. The performance on the novel items therefore indicates an application of the conversion rules, and the improvement over time demonstrates a true generalization of visual-to-auditory conversion rule learning.

The learning patterns on the novel items showed that much of the improvement occurred in the first week of training (i.e. after three training sessions), and only the performance on the different figures and different objects showed a further improvement beyond the first week (Figs. 4A–F). This suggests that most learning and the steady improvement in learning occurred for the types of visual images that were relatively easier to distinguish from one another. For the items that were more difficult to differentiate, such as the similar figures and the scenes, most learning seems to have occurred in the first week, and no further learning followed throughout the rest of the training period. Subjects were unable to improve beyond a certain point on these visually complex (e.g. the scenes) and similar (e.g. the similar figures) images, suggesting that given the amount of training, the auditory system is probably not capable of keeping track of all the information in the sound patterns, or of resolving the fine differences in the temporal and/or pitch dimension, or perhaps a different type of training (e.g. involving sensorimotor feedback) would be required for further improvement.

When performance is classified into different image types and familiarity of the items, it can be shown that the overall novel-item performance on the abstract figures is better than that on the objects, and better on the objects than on the scenes (Fig. 3). The performance superiority in the order of abstract figures, objects, and scenes suggests that the degree to which the converted sound patterns are interpreted depends on the complexity of the visual images. The scenes were most visually complex because they contained the most visual elements, thus the most visual information. The abstract figures were most simplistic in their compositions because they consisted of only several black-and-white lines and shapes, fewer elements than the real-life objects which contained more lines and shapes in varying grayscale. Therefore, the ability to extract the visual information from the converted sound patterns seems to be reflective of the amount of visual information that the images contain.

4. General discussion

The present study explored learning processes involving visual-to-auditory substitution in sighted individuals using Meijer's vOICe system (Meijer, 1992). At the early stage of the visual-to-auditory substitution learning, subjects were able to use the system merely on the basis of their explicit knowledge of the image-to-sound conversion rules. The ease of the initial use of the vOICe was reflected in the relatively accurate judgments made on size and position of the visual images with little training, suggesting an intuitive correspondence between visual images and their corresponding sounds. We further showed that extensive training allowed improvement in identifying shapes and patterns of visual images using sounds. Overall, the ability to discriminate visual objects using converted auditory patterns demonstrates that the auditory system is capable of extracting and processing visual information that is coded in the auditory input. This is consistent with the anecdotal reports of the blind users of the vOICe system who described their learned abilities to make visual judgments, such as localization and pattern recognition, on objects present in various visual environments (refer to www.seeingwithsound.com). Given the amount of training, the improvement on the tasks involving different abstract figures and pictures of various real-life objects suggests that objects could be analyzed at the level of their global shapes, whereas the rather slower improvement in interpreting the similar images and the pictures of scenes indicate that more extensive training might be necessary for decoding the sound patterns of more complex images. Consistent with the findings of the previous visual substitution studies (e.g. Arno et al., 1999; Auvray et al., 2007), the present study demonstrated the feasibility of using visual-to-auditory converted information as a suitable strategy for visual substitution.

Our most important finding in the examination of learning processes involving extensive training was generalized learning of the visual-to-auditory substitution system. As the improved ability to discriminate the sounds of novel stimuli suggests, the performance was not simply based on a mere association between the visual images and their corresponding sounds, but on the application of the image-to-sound conversion rules to novel stimuli. The improved performance on novel items indicates that the mechanism behind the improvement is learning of the fundamental relationship between the visual and auditory patterns and the ability to apply the visual–auditory relationship in order to decode the novel visual information via the auditory input.

The fact that the scores on the familiar items were not entirely based on the conversion rule learning, but rather confounded by the memory-based strategy indicates that the performance on the trained items was not reflective entirely of generalized learning. According to this finding, a clear distinction should be made between the memory- vs. rule-application-based performances in the examination of visual substitution learning because the latter would be required for the use of the substitution system in the constantly changing visual environment. Despite the confounding factor, the distinction between performances due to the different strategies has often been neglected in previous studies of sensory substitution. If the same testing items were repeatedly presented during the training and evaluation phases, it would be difficult to conclude that the improvement in the tasks actually is due to true learning of the relationship between the visual and auditory patterns. Our finding demonstrates the importance of testing of novel stimuli as a criterion for true generalization of learning.

Our finding of generalized crossmodal learning suggests crossmodal plasticity, and provides a basis to study neural substrates associated with such learning in visual-to-auditory substitution. It is now well established that the deafferented visual cortex in the blind is engaged in processing of different types of information (e.g. tactile: Sadato et al., 1996; Hamilton and Pascual-Leone, 1998; auditory: Kujala et al., 1995; Weeks et al., 2000; Gougoux et al., 2005; verbal: Burton et al., 2002; Amedi et al., 2003). However, much still remains unknown about the organization of this crossmodal information processing, and the study of brain activity associated with learning to decode image-to-sound converted information provides a unique opportunity to study the cortical re-organization. So far, findings of the functional imaging studies examining visual-to-auditory substitution suggest that cortical changes are likely to follow after practice in using a visual substitution system (PSVA: Poirier et al., 2006; Poirier et al., 2007; vOICe: Amedi et al., 2007; Pollock et al., 2005). However, since these studies vary greatly in their training and testing methods, much still remains to be discovered about the mechanism of the plastic changes. One important question that should be asked is about the relationship between the cortical changes and generalization of learning. Since these studies involved presenting only the highly trained items in the scanner, the performance might not be indicative of generalized learning, and thus it would be difficult to infer a causal association between the cortical changes and true abstract learning (and not memory-based association). In order to resolve this issue, future imaging studies should examine in-scanner performance on novel items as an index of generalized learning.

Another question should be addressed about the role of training in the cortical changes. Since the previous studies used training paradigms that varied highly in their training environments (ranging from free, unsupervised training- to controlled experimental settings) and the nature of their experimental tasks (e.g. pattern recognition, localization, orientation discrimination, passive listening), it is difficult to identify the aspects of training that lead to the plastic changes, we also do not know whether the cortical changes that occur after training in sensory substitution are specific to the training method of choice, or whether they can be attributed to training in general. Therefore, in order to make a more specific interpretation about the neural substrates associated with visual substitution learning, the type of training method and the nature of experimental tasks would need to be thought out carefully for future functional imaging studies.

The current training paradigm was tested in a controlled static environment where direct visual feedback was given instead of naturalistic settings where blindfolded or blind subjects would engage in sensorimotor interactions by using feedback cues received from moving around with the substitution system. By eliminating the motor feedback, we sought to examine whether learning of the visual substitution system could be achieved solely based on perceptual learning of the relationship between visual images and sounds. We demonstrated that it is possible to learn to use the substitution system without necessarily depending on sensorimotor contingency information, suggesting that visual-to-auditory substitution in sighted individuals can be achieved to a large degree through perceptual learning of the relationship between visual images and sounds. However, the fact that performance did not reach the ceiling level by the end of the training period allows us to question factors that could help to achieve further improvement. The presence of motor feedback could be one such factor along with the role of visual deprivation. These will therefore be appropriate topics for future studies to address.

5. Experimental procedures

5.1. *Experiment 1 methods*

5.1.1. *Participants*

Eighteen sighted subjects with normal hearing (12 women; 21 to 48 years of age, M= 26.72, SD=6.23) were recruited and gave written consent for the study approved by the McGill University Research Ethics Board. They were randomly assigned to 'trained' and 'untrained' experimental groups. The trained subject group was given no explanations on the image-sound relationship, but was simply told that the sounds and images that were going to be presented were systematically related. Instead of receiving any training, the untrained group was explained the image-to-sound conversion rules, and were asked to use these explicitly given rules in their tasks.

5.1.2. *Stimuli*

Visual stimuli consisted of 39 abstract shapes consisting of various lines and geometrical figures, and 819 variants of these shapes that were modified in terms of position, orientation, and size. The shapes were adapted from the Aggie Figure Learning Test developed by Majdan, Sziklas, and Jones-Gotman (1996; see Fig. 6(A) for examples).

Auditory stimuli were converted sounds of the visual images. They were generated using the vOICe developed by Meijer (1992; www.seeingwithsound.com). The vOICe runs as software on a personal computer, and translates video input into auditory output using an algorithm varying in three parameters. It maps the vertical dimension of the input visual scene into a range of sound frequencies, the horizontal displacement into time (from left to right, with the leftmost part of the scene sounding the earliest), and visual intensity into sound intensity. The algorithm was applied such that for the vertical coding, the visual scene was divided into 144 pixel rows with each row corresponding to a certain sound frequency from a range of 500–5000 Hz in the increasing order of the exponential frequency distribution with a higher row in a higher spatial position. For the horizontal coding, the visual scene was divided into 176 pixel columns and was scanned from left to right at a rate of 2 s per frame with the pixels that were part of the same column sounding simultaneously (~.011 s per column). The soundscapes were in stereo, creating a perception of sound panning from left to right. Visual intensity was coded such that black was translated into the most intense sound whereas white was silence, and grayscale in the middle intensity range between white and black (16 shades of gray were used; see Meijer, 1992 for details). For the purpose of the first experiment, the visual images were only in black and white. A 50 ms click was added at the beginning of each auditory stimulus in order to cue the start of the sound.

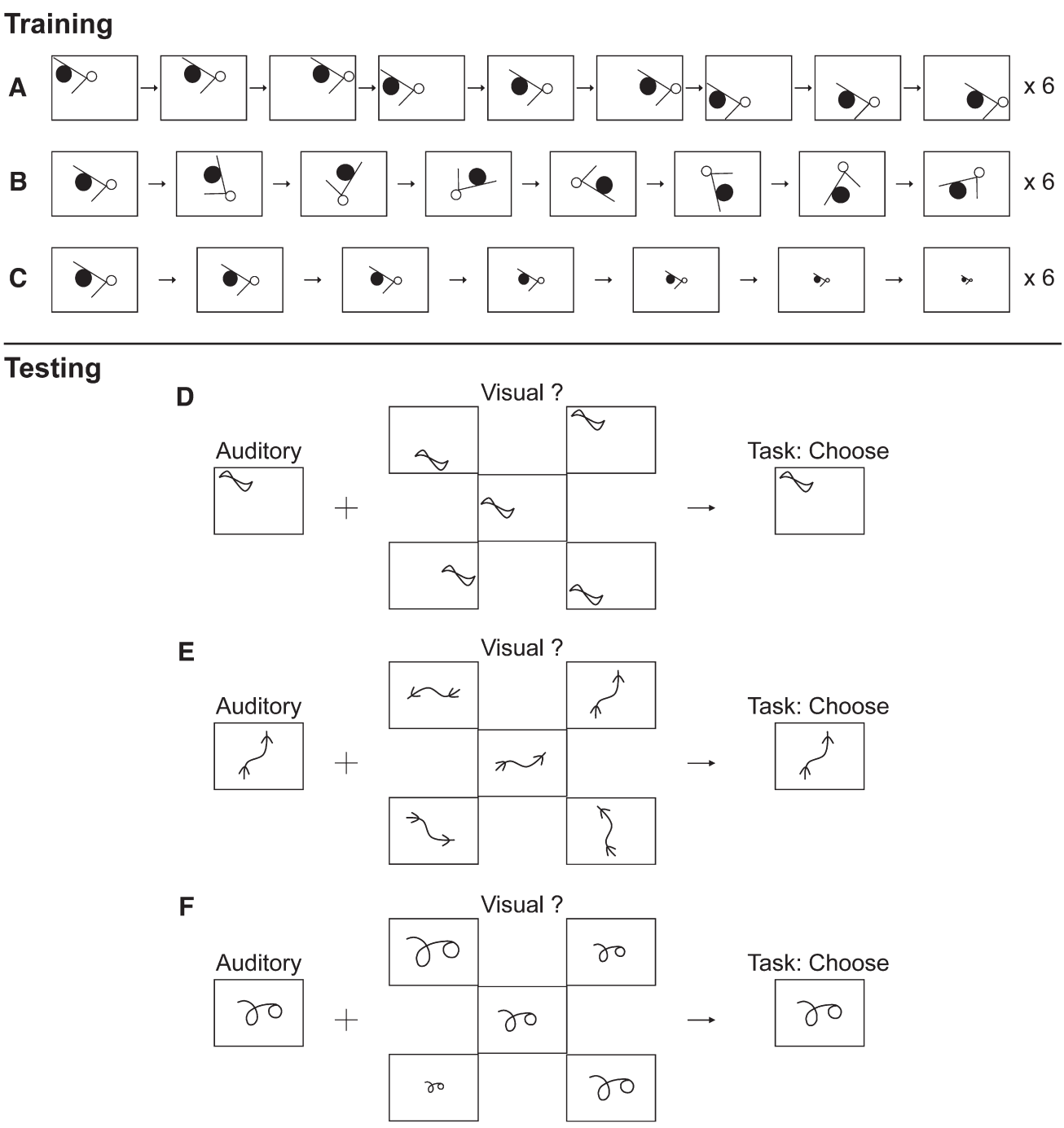

Fig. 5 – Illustrations of the training and testing procedures in Experiment 1. Training involved simultaneous presentations of visual images and their converted sounds (for the trained group only). Each training trial consisted of repetitive presentations of image-and-sound pairs of an abstract figure that changed in (A) position, (B) orientation, or (C) size. Testing involved a forced-choice task where subjects were presented with one sound and five visual images among which the correct image of the sound was to be chosen. In each testing trial, the five visual images differed in their (D) position, (E) orientation, or (F) size. The trained group was tested on the stimuli that had been used for training as well as stimuli never seen before (whereas all the testing stimuli were new to the untrained group).

5.1.3. *Training (for the 'trained' group only)*

The study consisted of two training sessions and one evaluation session spanning over a minimum of three- to a maximum of five days, each of which took place on a separate day. The three sessions lasted approximately 3 h in total. The two training sessions comprised three counterbalanced conditions ('position', 'orientation', and 'size') in each of which eight pairs of visual images and their corresponding sounds were presented. On the first day the subjects belonging to the 'trained' group participated in one of the three conditions, and

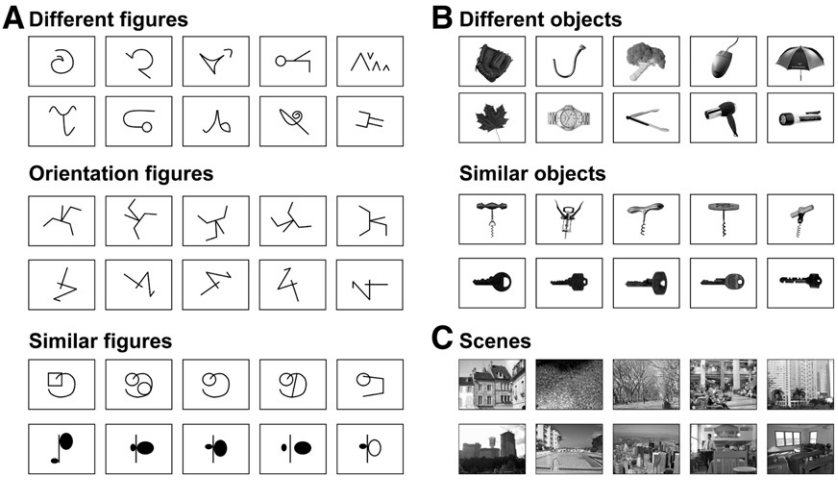

Fig. 6 – Examples of (A) abstract figures, (B) pictures of real-life objects, and (C) pictures of scenes used in Experiment 2.

on the second day in the remaining two conditions. The visual images were presented on a computer screen and the corresponding sounds through a set of headphones.

Throughout the training sessions, every visual image was presented simultaneously with its corresponding sound. In the 'position' condition, a visual image appeared in nine different positions of the screen in a systematic order (Fig. 5A). Once the image appeared in all nine positions, it was presented in the reverse order. The procedure of the image appearing in- and reverse order was repeated three times (i.e. each position was presented six times in total).

In the 'orientation' condition, each image appeared in the middle of the screen, and changed its orientation in steps of 45°in angle (Fig. 5B). A total of eight orientations (0, 45, 90, 135, 180, 215, 270, and 315°) were presented first clockwise, then counter-clockwise. This procedure was repeated three times such that each orientation was presented six times in total.

In the 'size' condition, each image was presented in seven different sizes. Subjects were first presented with a visual image, and then saw the image in six smaller sizes in the decreasing order (Fig. 5C). Each image was smaller than its previous image by 12.5%, thus the smallest image was 25% of its original size. After the decreasing order, the image was presented in the increasing order starting with the smallest. The procedure of the image appearing in the decreasing and increasing size was repeated three times (i.e. each size was presented six times).

Following each presentation of image-and-sound series, subjects were given four forced-choice evaluation trials. Upon hearing a sound, they saw five of the visual images from the previously presented image-and-sound series, and were asked to choose the correct image of the sound. Subjects were allowed up to three chances to choose the correct answer. Correctness of the given answer was provided, and regardless of the answer, the correct pair of the image and the sound was presented at the end of each evaluation trial.

5.1.4. *Testing (for the 'trained' and 'untrained' groups)*

The evaluation session consisted of two tests, one using previously presented stimuli, and the other using novel stimuli. Each testing trial involved choosing the correct image of a sound from a pool of five visual images. The 'familiar' test consisted of three blocks of 16 trials, with each trial involving choosing of the correct visual image that appeared in a particular position, orientation, or size (16 trials tested for each condition; see Figs. 5D–F). The 'novel' test followed the same procedure as the familiar test. The only difference was that it used 31 new visual images that had not been previously presented. The untrained group were administered the same tests. The testing stimuli were identified as being 'familiar' or 'novel' items according to the familiarity to the trained group, and for the sake of naming, these terms were kept consistent for the untrained group although all the stimuli were new to them.

5.2. *Experiment 2 methods*

5.2.1. *Participants*

Eight sighted individuals were assigned to the training group (5 women; 20 to 45 years of age, M=25.63, SD=8.03) and ten sighted individuals were assigned to the control group (7 women; 19 to 30 years of age, M=24.50, SD=4.70). None of them had participated in the previous study, and each gave written consent at the beginning of the first session.

5.2.2. *Stimuli*

Three types of visual images were converted into sounds: abstract figures, pictures of real-life objects, and pictures of scenes (see Fig. 6 for examples). We created a total of 129 sets of five black-and-white abstract figures. Each set of five abstract figures was created such that the figures in the

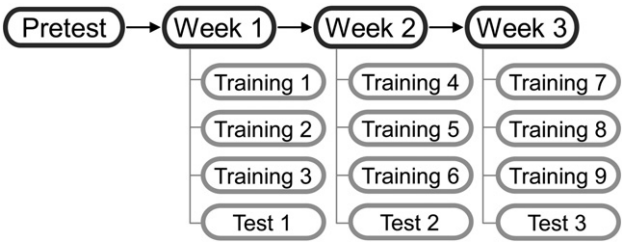

Fig. 7 – Time schedule of Experiment 2.

same set were different only by a small number of elements, thereby preserving the global shapes of the five figures (the figures within each set were referred to as 'similar abstract figures'). We used a total of 129 sets of five pictures of real-life objects (e.g. pencils, chairs, mugs, hammers, and so on) found on the internet (the objects within each set were referred to as 'similar objects'). The pictures were edited so that backgrounds (if any) were removed in order to isolate the objects, and the objects were kept similar in size. Finally, we used 347 pictures of scenes that were found on the internet. The themes of the pictures were various and random (e.g. city/nature sceneries, people, animals). All the pictures were edited in the width-to-height aspect ratio of 4:3. The colors of the scene and real-life object pictures were changed to grayscale. All the visual images were converted into sounds using the vOICe in the same way as in the first experiment.

5.2.3. *Time schedule*

The study took place over the span of approximately three weeks (mean of 19.13 days; see Fig. 7 for the time schedule). The study began with a pretest session, and there were three training sessions followed by one test session each week (totaling 13 sessions for the entire study). Subjects were required to attend every session with the number of days between two sessions being no more than three (mean of 1.67 days). Each session lasted approximately 2 h.

5.2.4. *Pretest and test sessions 1, 2 and 3*

Prior to the pretest, the rules of image-to-sound conversion were explained to the subjects. The pretest consisted of six 25-trial tests, with each test presenting different types of visual images. For each trial, subjects were presented with one sound and five visual images, and were asked to choose the correct visual image of the sound heard (Fig. 8A). In the different figures test, the five images for each trial were abstract figures that were selected from different sets of 'similar abstract figures'. In order to increase the difficulty of the task, the figures that were noticeably different in their global shapes (e.g. one figure consisting of round elements whose overall shape is round vs. another figure consisting of lines whose overall shape is oriented in one particular direction) were chosen not to be in the same trial. In the orientation figures test, five images were of one abstract image in five different orientations (one abstract figure in 0, 72, 144, 216, and 288°). In

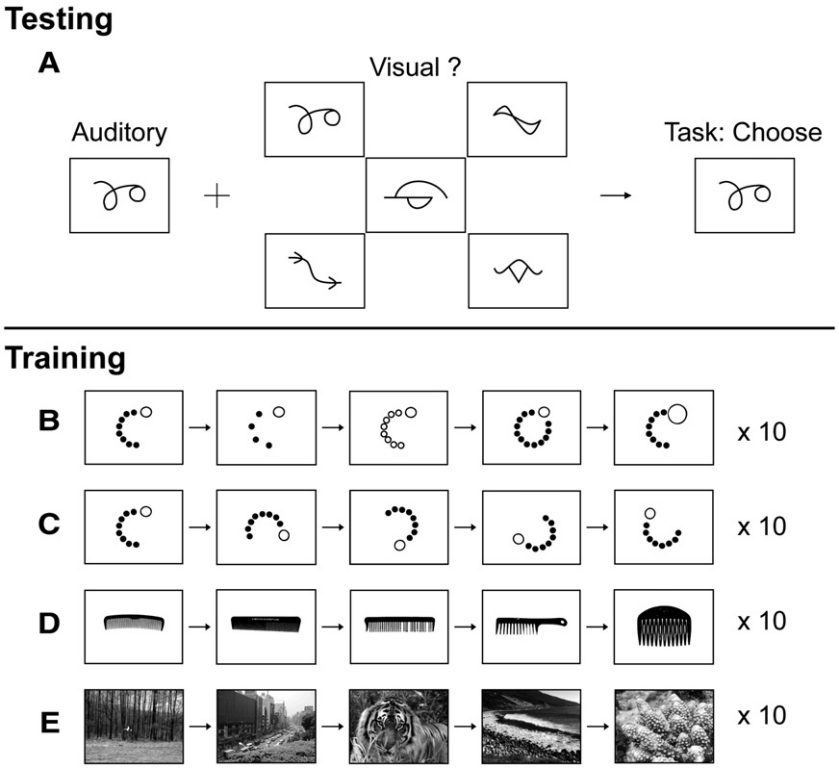

Fig. 8 – Illustrations of the testing and training procedures in Experiment 2. Similar to the procedures of Experiment 1, each testing trial involved a presentation of one sound and five visual images among which the correct image of the sound heard was to be chosen. The type of five visual images presented depended on the category of images subjects were being tested on among six different categories. The categories included abstract figures that differed in shape (different figures), same-shaped abstract figures in five orientations (orientation figures), abstract figures that were similar in shape (similar figures), various real-life objects (different objects), real-life objects belonging to the same object category (similar objects), and scenes (scenes). (A) An example of a testing trial for the different figures test. Training consisted of simultaneous presentations of visual images and their converted sounds. This experiment included a larger variety of training stimuli than in Experiment 1. The types of images included (B) abstract figures similar in shape, (C) abstract figures in different orientations, (D) real-life objects belonging to the same category, and (E) scenes. Each trial consisted of ten repetitions of five image-sound pairs.

Table 2 – A complete list of tests

Pretest	Tests 1, 2, and 3
Different figures	Different figures — novel
Orientation figures	Different figures — familiar
Similar figures	Orientation figures — novel
Different objects	Orientation figures — familiar
Similar objects	Similar figures — novel
Scenes	Similar figures — familiar
	Different objects — novel
	Different objects — familiar
	Similar objects — novel
	Similar objects — familiar
	Scenes — novel
	Scenes — familiar

Different names of the tests indicate different types of visual images presented. Each of the tests 1, 2, and 3 included the items that were used in training each week as well as new items.

the similar figures test, each trial presented five 'similar abstract figures'. In the different objects test, each trial presented pictures of five different real-life objects. As in the different figures test, the inclusion of objects that were visibly different in their global shapes (e.g. a ball vs. a golf club) was avoided in the same trial. In the similar objects test, each trial involved presenting pictures of five 'similar objects'. In the scenes test, five scene pictures were pseudo-randomly chosen from the pool of the scene pictures. For each trial, we avoided having scenes that were noticeably different in terms of their basic patterns of elements (e.g. a scene with many vertical lines vs. a scene with many horizontal lines). See Fig. 6 for examples of the images used in the tests. Subjects were allowed to repeat the sound in each trial as many times as they wished, but encouraged to limit to four repetitions.

Test sessions 1, 2, and 3 took place at the end of each training week, and consisted of tests that involved the same task as in the pretest (different figures, orientation figures, similar figures, different objects, similar objects, scenes tests). Each forced-choice test consisted of 25 trials which involved presenting items that had not been previously presented during training (referred to as a 'novel' test), and 18 trials which involved presenting items that were used for training during each relevant week (referred to as a 'familiar' test). See Table 2 for a complete list of tests. There was no pretest measurement for the familiar items because by definition 'familiar items' have to be those that subjects have been exposed to.

The control group was given the different figures and the different objects tests that the training group received before training (as part of the pretest). The control group was instructed to perform the tasks without any explicit explanations of the conversion rules or any kind of feedback on the image–sound relationship. They were simply told that they were going to be presented with a sound and five images, and were asked to choose a visual image that they thought best matched the sound heard. The two tests were chosen because they were anticipated to be the easiest among the six tests that the training group was given, and therefore would provide the most conservative measure of comparison to the performance of the training group.

5.2.5. *Training*

Each training session involved a presentation of six sets of the following image types: (1) similar abstract figures (similar figures), (2) an abstract figure in five orientations (orientation figures), (3) similar objects, and (4) scenes (see Figs. 8B–E). The presentation of each image-and-sound pair always started with a click that cued the beginning of the 2-second sound.

The training procedure was similar to that of Experiment 1. For each set of the similar figures, subjects were presented with a series of five similar abstract figures and their corresponding sounds. Followed by the presentation of all five image–sound pairs, the series were presented in the reverse order. The in- and reverse-ordered series were repeated four additional times such that each pair was presented ten times in total. During the presentation, subjects were asked to study the relationship between the visual images and sounds and observe differences in the sounds according to the differences in the visual images. At the end, they were given three forced-choice trials, for each of which the task was to choose the correct visual image of the sound heard among five visual images (all from the set that they had just been trained on). Subjects were given up to three chances to choose the correct image. Whether correct or not, the correct pair of the image-and-sound pair was provided at the end of each trial. The same procedure was used for the other image types (orientation figures, similar objects, and scenes).

The testing and training sessions also included a brief drawing task in which subjects drew the visual images of the sounds heard. However, since the drawings did not yield meaningful interpretations, the relevant data are not reported in the present study.

Acknowledgments

We thank Dr. Peter Meijer for his helpful discussions, Marc Bouffard and Nick Foster for their technical assistance, and Dr. Marilyn Jones-Gotman for giving permission to modify her stimuli. This research was supported by grants from the Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Amedi, A., Raz, N., Pianka, P., Malach, R., Zohary, E., 2003. Early 'visual' cortex activation correlates with superior verbal memory performance in the blind. Nat. Neurosci. 6 (7), 758–766.

Amedi, A., Stern, W.M., Camprodon, J.A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P., Pascual-Leone, A., 2007. Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nat. Neurosci. 10 (6), 687–689.

Arno, P., Capelle, C., Wanet-Defalque, M.-C., Catalan-Ahumada, M., Veraart, C., 1999. Auditory coding of visual patterns for the blind. Perception 28, 1013–1029.

Arno, P., Vanlierde, A., Streel, E., Wanet-Defalque, M.-C., Snanbria-Bohorquez, S.S., Veraart, C., 2001. Auditory substitution of vision: pattern recognition by the blind. Appl. Cogn. Psychol. 15, 509–519.

Auvray, M., Hanneton, S., O'Regan, J.K., 2007. Learning to perceive with a visuo-auditory substitution system: localisation and object recognition with 'The vOICe'. Perception 36 (3), 416–430.
Bach-y-Rita, P., 1972. Brain Mechanisms in Sensory Substitution. Academic Press.
Bach-y-Rita, P., Collins, C.C., Saunders, F.A., White, B., Scadden, L., 1969. Vision substitution by tactile image projection. Nature 221 (5184), 963–964.
Bach-y-Rita, P., Kaczmarek, K.A., Tyler, M.E., Garcia-Lara, J., 1998. Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note. J. Rehabil. Res. Dev. 35 (4), 427–430.
Burton, H., Snyder, A.Z., Diamond, J.B., Raichle, M.E., 2002. Adaptive changes in early and late blind: a fMRI study of verb generation to heard nouns. J. Neuropsyhiol. 88 (6), 3359–3371.
Capelle, C.C., Trullements, C., Arno, P., Veraart, C., 1998. A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. IEEE Trans. Biomed. Eng. 45 (10), 1279–1293.
Evans, K.K., Treisman, A., 2005. Crossmodal binding of audio-visual correspondent features. J. Vis. 5 (8), 874a.
Gougoux, F., Zatorre, R.J., Lassonde, M., Voss, P., Lepore, F., 2005. A functional neuroimaging study of sound localization: visual cortex activity predicts performance in early-blind individuals. PLoS. Biol. 3 (2), e27.
Hamilton, R., Pascual-Leone, A., 1998. Cortical plasticity associated with Braille learning. Trends Cogn. Sci. 2 (5), 168–174.
Kujala, T., Huotilainen, M., Sinkkonen, J., Ahonen, A.I., Alho, K., Hämäläinen, M.S., Ilmoniemi, R.J., Kajola, M., Knuutila, J.E., Lavikainen, J., Salonen, O., Simola, J., Standertskjöld-Nordenstam, C.-G., Tiitinen, H., Tissari, S.O., Näätänenet, R., 1995. Visual cortex activation in blind humans during sound discrimination. Neurosci. Lett. 183 (1–2), 143–146.
Majdan, A., Sziklas, V., Jones-Gotman, M., 1996. Performance of healthy subjects and patients with resection from the anterior temporal lobe and matched tests of verbal and visuoperceptual learning. J. Clin. Exp. Neuropsychol. 18 (3), 416–430.
Meijer, P., 1992. An experimental system for auditory image representations. IEEE Trans. Biom. Eng. 39 (2), 112–121.
Poirier, C., De Volder, A., Tranduy, D., Scheiber, C., 2007. Pattern recognition using a device substituting audition for vision in blindfolded sighted subjects. Neuropsychologia 45, 1108–1121.
Poirier, C., Richard, M.-A., Duy, D.T., Veraart, C., 2006. Assessment of sensory substitution prosthesis potentialities in minimalist conditions of learning. Appl. Cogn. Psychol. 20 (4), 447–460.
Pollock, B., Schnitzler, I., Stoerig, P., Mierdorf, T., Schnitzler, A., 2005. Image-to-sound conversion: experience-induced plasticity in auditory cortex of blindfolded adults. Exp. Brain Res. 167, 287–291.
Ptito, M., Moesgaard, S.M., Gjedde, A., Kupers, R., 2005. Crossmodal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. Brain 128 (Pt 3), 606–614.
Renier, L., Bruyer, R., De Volder, A.B., 2006. Vertical–horizontal illusion present for sighted but not early blind humans using auditory substitution of vision. Percept. Psychophys. 68 (4), 535–542.
Renier, L., Collignon, O., Poirier, C., Tranduy, D., Vanlierde, A., Bol, A., Veraart, C., De Volder, A.G., 2005. Crossmodal activation of visual cortex during depth perception using auditory substitution vision. NeuroImage 26 (2), 573–580.
Renier, L., Laloyaux, C., Collignon, O., Tranduy, D., Vanlierde, A., Bruyer, R., De Volder, A.G., 2005. The Ponzo illusion with auditory substitution of vision in sighted and early-blind subjects. Perception 34 (7), 857–867.
Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M.P., Dold, G., Hallet, M., 1996. Activation of the primary visual cortex by Braille reading in the blind subjects. Nature 380 (6574), 526–528.
Sampaio, E., Maris, S., Bach-y-Rita, P., 2001. Brain plasticity: 'visual' acuity of blind persons via the tongue. Brain Res. 908 (2), 204–207.
Weeks, R., Horwitz, B., Aziz-Sultan, A., Tian, B., Wessinger, C.M., Cohen, L.G., Hallett, M., Rauschecker, J.P., 2000. A positron emission tomographic study of auditory localization in the congenitally blind. J. Neurosci. 20 (7), 2664–2672.