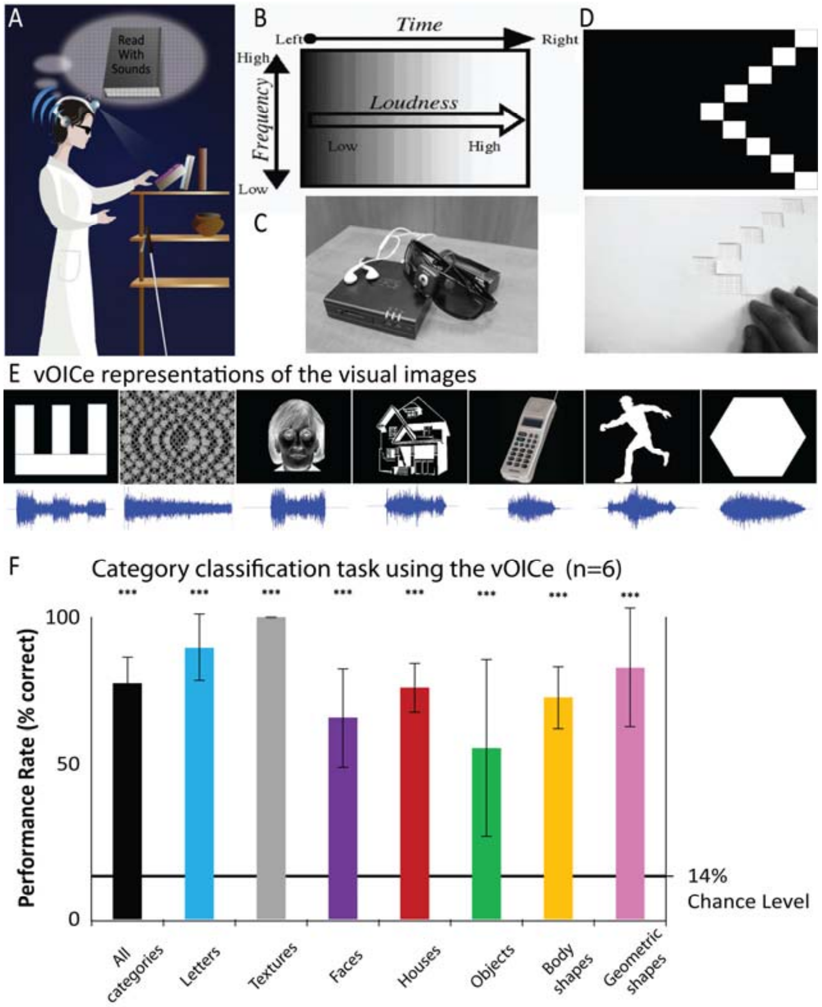
SSD training procedure:

All participants in this study (n=7), as well as two others who could not be scanned due to MR-incompatible implants, were enrolled in a novel unique training program in which they were taught how to effectively extract and interpret high-resolution visual information from the complex soundscapes generated by the vOICe SSD. Each participant was trained for several months in a 2-hour weekly individual training session by a single trainer. The training duration and progress rate varied between participants and were determined by the personal achievements and difficulties (the average training of the participants here was 73 hours, up to 10 of which devoted to learning to read using the SSD). The training program was composed of two main components: Structured 2- dimensional training, in which the participants were taught how to process 2- dimensional still (static) images, and live-view training in which visual depth-perception and training in head-eye-hand coordination were taught, using a mobile kit of the vOICe SSD (similar to that in **Fig. 1C**). The structured 2- dimensional training involved learning to process hundreds of images of seven structured categories: geometric shapes, Hebrew letters and digital numbers, body postures, everyday objects, textures (sometimes with geometric shapes placed over a visual texture, used to teach object-background segregation), faces and houses (see **Fig. 1E**; see also **Movie S1** showing images from all the categories and their SSD transformation to sounds), introduced with increasing complexity and detail. Thus, for teaching letters, we first introduced vertical, horizontal, diagonal and curved lines, and then built the letters from these elements (for a demo in English see http://doc.org.il/voice.ppsx). Some blind subjects, who were accustomed mostly to Braille letters, had to be reminded and trained on the shape of unfamiliar letters. All the participants in the study who completed the entire structured 2-dimensional training were taught the entire Hebrew alphabet, and also taught how to read whole words (up to 5 letters) in a single soundscape (for a demonstration of a blind participant reading a 3-letter word see **Movie S2**), which could be further sequenced to form longer words. The main consideration for choosing these particular categories for training was their functional importance in daily life. Additionally, these categories are known to be processed by distinct specialized brain areas in sighted individuals (Kanwisher, 2010). The participants heard each soundscape, tried to describe what it contained, and were directed through the trainers' questions to a detailed full description of the image. Additionally, whenever possible, we provided the blind subjects with haptic feedback by presenting them with palpable images identical to those they perceived through the SSD (**Fig. 1D**; e.g. for all 2D stimuli and some of the 3D stimuli). Following the structured 2-D training, participants could tell upon hearing a soundscape which category it represented. This required Gestalt object perception and generalization of the category principles and shape perception to novel stimuli. They could also determine multiple features of the stimulus (such as hairstyle in a face image, number of floors and windows in a house image, and body posture in a body-shape image), enabling them to differentiate between objects within categories. A demonstration of a blind participant recognizing emotional facial expressions from soundscapes can be found in **Movie S2**.

Additionally, general principles of visual perception which were not familiar to the congenitally blind participants were demonstrated and taught using stimuli from the various classes. Such principles were (i) the conversion of 3- dimensional objects to two-dimensional images (and back) depending on viewpoint and on object position, (ii) the transparency of objects and the ability to see through parts of objects, and (iii) the occlusion of parts of objects. Other visual perception principles which required active sensing were demonstrated using the live-view training technique. These included principles such as (i) visual field-of-view, (ii) orienting their heads (on which the webcam was mounted, placed on sunglasses) to the objects at hand in order to scan a visual scene (iii) variation of apparent object size depending on distance, and (iv) the use of monocular depth cues such as occlusion between different objects and perspective to estimate the depth and distance of seen objects. The scan was conducted following the completion of the structural 2-dimensional training (although most participants continued their live-view training further).

In order to minimize sensory-motor artifacts, no recording of performance was conducted during the fMRI scan. Prior to each scan we verified the subjects were able to easily recognize learned stimuli from the tested categories. In a complementary experiment within the scanner on the same day (not reported here) the subjects were instructed to decide whether a presented letter was one of 3 pre-specified target letters (3 target letters out of 10 different letters presented; each presented 3 times), and performed at 77.5% (±6.4% standard deviation), significantly better than chance (t-test p<0.0005).

Figure 1: "Visual" performance in blind users of "The vOICe" sensory substitution device following training.



A. Visual-to-auditory sensory substitution is used to convey visual information to the blind via their intact auditory modality.

B. The transformation algorithm of the vOICe (Meijer, 1992): Each image is scanned from left to right, such that time and stereo panning constitute the horizontal axis in its sound representation, tone frequency makes up the vertical axis, and loudness corresponds to pixel brightness (see also www.seeingwithsound.com).

C. The mobile kit for SSD usage includes a lightweight inexpensive webcam worn on eyeglasses, a computing device (such as a netbook computer or smartphone), and earphones.

D. Tangible image feedback (lower panel), identical to that presented using the vOICe (upper panel) was provided to the blind participants to help them further understand the images during training.

E. The structured 2- dimensional training program consisted of hundreds of stimuli organized in order of complexity and grouped into structured categories: Hebrew letters, textures, faces, houses, tools & everyday objects, body postures and geometric shapes.

F. Success in discriminating between object categories was assessed upon completion of the structured 2-dimensional training (n=6). Object discrimination differed significantly from chance (mean percent correct 78.1%±8.8% SD, $p<0.00005$, student t-test), and no difference was observed between performance in the letter category and any of the other object categories ($p > 0.05$, corrected for multiple comparisons). Error bars denote standard deviation. Significance refers to difference from chance level: * $p<0.05$, ** $p<0.005$, *** $p<0.0005$.