Perceiving graphical and pictorial information via hearing and touch

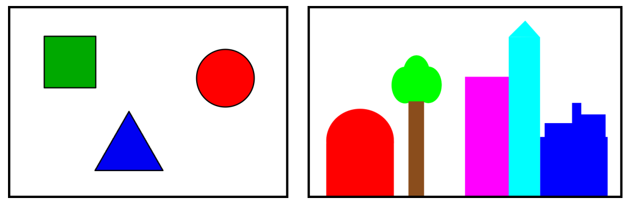
Pubudu Madhawa Silva, *Member, IEEE,* Thrasyvoulos N. Pappas, *Fellow, IEEE,* Joshua Atkins, *Member, IEEE,* and James E. West, *Life Fellow, IEEE*

***Abstract*—We propose a dynamic, interactive system for conveying visual information via hearing and touch to blind and visually impaired (BVI) people. The system is implemented with a touch screen that allows the user to actively explore a two-dimensional layout consisting of one or more objects with the finger while listening to auditory feedback. Sound is used as the primary source of information for object localization, identification, and shape, while touch is used for pointing and kinesthetic feedback. A static overlay of raised-dot tactile patterns can also be added. The head-related transfer function is used for rendering sound directionality, and variations of sound intensity or other features are used for rendering proximity. The main focus is on conveying the shape of an object, but the rendering of a simple scene layout, that consists of objects in a linear arrangement, each with a distinct tapping sound, is also considered and compared to a "virtual cane." We consider a number of acoustic-tactile configurations and use empirical studies with visually-blocked sighted participants to compare their effectiveness. Our findings demonstrate the advantages of spatial sound (directionality and proximity cues) for dynamic display of information (localization, identification, shape), while raised-dot patterns provide the best static shape rendition. We also show that the proposed configurations outperform existing techniques. The proposed approach is also expected to impact other applications where vision cannot be used.**

***Index Terms*—Acoustic-tactile representation of visual signals, semantic mapping, sensory substitution, user interface.**

I. Introduction

THE ubiquity of the Internet and the use of electronic media, rich in graphical and pictorial information, for communication, commerce, entertainment, art, and education, has made it hard for the blind and visually impaired (BVI) community to keep up. While some of this information can be translated to speech or Braille text, the ability to present graphical and pictorial information in acoustic-tactile form will dramatically increase the amount of information that can be made available to the BVI segment of the population. In this paper, we explore the use of hearing and touch to convey such visual information to the BVI.



(a) (b)

Fig. 1. (a) Simple shapes; (b) simple scene.

The main idea is that the user actively explores a two-dimensional (2-D) layout (diagram, graph, chart, map) consisting of one or more objects on a touch screen, using the finger as a pointing device that provides kinesthetic feedback, and auditory feedback for object display and navigation in the virtual space. An overlay of raised-dot tactile patterns or a transparent piezo-active polymer can also be added. The focus of this paper is on conveying the shape of a simple object, e.g., like those shown in Fig. 1(a). However, we also consider the rendering of a simple scene layout consisting of objects in a linear arrangement like the one shown in Fig. 1(b), each with a distinct tapping sound, which we compare to a "virtual cane."

The advantage of visual substitution methods, whereby hearing and touch are used in place of vision [1], is that they do not require invasive approaches, which typically require surgery to restore some degree of functional vision by stimulating the visual cortex, e.g., as in the cortical or retinal electrode matrix displays for partial restoration of vision [2], [3]. A well-established visual substitution approach is Braille, which relies on touch to display a variety of symbols. While not requiring surgery, some visual substitution approaches can still be quite objectionable. For example, the tongue display [4] consists of an array of electrodes that can apply different voltages to stimulate the tongue, which is the most sensitive tactile organ with the highest spatial resolution. While the tongue display has proven to be quite effective in helping BVI people carry out certain visual tasks, the majority of the BVI population find it quite "invasive" and prefer to actively scan/explore with the finger [5], as they are used to doing in Braille. Similar objections apply to the presentation of electrical and other tactile stimuli on other parts of the body

Manuscript received July 22, 2015; revised January 25, 2016 and March 15, 2016; accepted June 16, 2016. Date of publication August 16, 2016; date of current version November 15, 2016. This paper was presented in part at the International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, May 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Martha Larson.

P. M. Silva is with Intel Corporation, Hillsboro, OR 97124 USA (e-mail: pubudu@u.northwestern.edu).

T. N. Pappas is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: pappas@eecs.northwestern.edu).

J. Atkins is with Apple Inc., Culver City, CA 90232 USA (e-mail: joshatkins@gmail.com).

J. E. West is with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: jimwest@jhu.edu).

Color versions of one or more of the gures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identier 10.1109/TMM.2016.2601029

(back, abdomen, forehead) [6], [7]. The advantages of active exploration and kinesthetic feedback are discussed in review papers by O'Modhrain *et al.* [8] and Klatzky *et al.* [9].

The wide availability of dynamic tactile sensing devices (tablets, tablet PCs, cell phones with touch screens) enable the presentation of dynamic acoustic signals in response to finger movements. Thus, an object can be represented as a region in a touch screen associated with a characteristic sound. Static tactile signals can be added by superimposing a raised-dot pattern embossed on paper (using a Braille printer[1]) on the screen, as in the "Talking Tactile Tablet" (TTT) [10]. Recent advances in tactile technology make it possible to combine the touch screen with a variable friction display [11]–[13].[2] An alternative variable friction technology controls the surface friction of a rigid material such as glass [14]–[16]. In addition, vibration feedback can also be integrated in the touch-screen device.[3] There are a number of publications on displays (usually smart phones or tablets with touchscreens) that provide dynamic vibro-tactile feedback, via either motors or Piezo-electric elements, and combine it with kinesthetic and auditory feedback. For example, Poppinga *et al.* [17] used such a setup with speech feedback to convey simplified street maps. Giudice *et al.* [18] used a similar setup with audio feedback to convey bar charts, letters, and line orientations, and later used it to explore non-visual panning methods for vibro-tactile displays [19]. An excellent review of the available technologies and their capabilities can be found in [8]. In this paper, our primary focus is on acoustic display, but we also briefly consider the superposition of raised-dot patterns embossed on paper.

The ultimate goal of the proposed approach would be to use a still or video camera to capture a scene, and then to translate it into tactile-acoustic form. However, due to the limited spatial resolution of touch [20] and hearing [21], we do not expect that we will be able to display all the visual detail. Moreover, the direct translation of visual signals into acoustic and tactile signals cannot be done in an intuitive way, e.g., image intensity to sound intensity, frequency, or other attribute. Thus, the visual to acoustic-tactile translation will have to be based on image segmentation (into perceptually uniform regions [22], [23]) and then mapping of each segment (based on its features or semantics) into a distinct acoustic signal, tactile pattern, or combinations thereof, as shown in Fig. 2. Such a representation will provide key information about the location, shape, and identity of the key objects in the scene. Of course, this approach requires image analysis to produce a region-based representation, or the availability of semantic representation (e.g., as in maps and graphics). However, as we pointed out above, the focus of this paper is on rendering simple object shapes and scene layouts, not on image analysis to obtain the semantic representation.

[1] "ViewPlus Braille embossers," [Online]. Available: http://www.viewplus.com

[1] "Talking Tactile Tablet," [Online]. Available: http://www.touchgraphics.com/

[2] "Senseg Tixel," [Online]. Available: http://senseg.com/technology/senseg-technology

[3] "Immersion," [Online]. Available: http://www.immersion.com

We present and test various configurations for acoustic-tactile display of simple shapes and layouts. In each configuration, the touch screen is partitioned into regions, each with a particular sound field. Each region represents an object, part of an object, background, or other element of a visual scene or graphical display. The auditory feedback, played back on stereo headphones, depends on the finger position on the touch screen. Unlike the TTT [10],[2] where the acoustic signal consists of speech that explains the tactile pattern (typically a block diagram) explored with the finger, we use sound for object identification and for guiding the finger to one or more objects or along their boundary. For the latter, we make use of spatial sound, in the form of head-related transfer functions (HRTFs) [21] for rendering sound *directionality* and variations of sound intensity or other attributes for rendering *proximity.*

The proposed approach assumes that an image of the environment has been captured, and the goal is to present it to the user in acoustic-tactile form. One approach for doing this is by directly translating visual signals into touch and sound. For example, the tongue display [4] we mentioned above relies on direct translation of image intensities captured by a camera into voltages. Another example is "SoundView," developed by Doel *et al.* [24], whereby the user actively explores a color image on a tablet with a pointer. The color of each pixel is translated to a sound, with the pointer acting like a gramophone needle creating sounds as it "scratches" the image surface. Meijer's imaging system named "vOICe" [25], maps a 64×64 image with 16 gray levels to a sequence of tones. The vertical dimension of the image is represented by frequency and the horizontal dimension is represented by time and stereo panning. The loudness of the tone is proportional to the brightness of the corresponding pixel. In contrast to "SoundView," the user does not have active control of the presentation. In the system proposed by Hernandez *et al.* [26] sound rays emanate from the object surface in the same manner as light rays. Finally, among the direct mappings, we should mention the rendition of grayscale values to tactile patterns proposed by Barner *et al.* [27]–[29], using digital halftoning. However, as we saw, direct mapping systems are not very effective because of the limited spatial resolution of touch and hearing and the lack of intuitive mappings from the visual to the acoustic or tactile domain.

In contrast to the direct mapping approaches, the proposed approach assumes that the image has been analyzed to obtain a semantic mapping, which then is presented to the user in acoustic-tactile form. In addition, like "SoundView" [24], in the proposed approach the user actively explores the display. A number of systems have been proposed along these lines. Su *et al.* [30] tested an iPhone display with stereo panned acoustic feedback provided via headphones to convey line drawings and simple indoor floor plans. Cohen *et al.* proposed a tablet PC based display, where the user uses a stylus and auditory feedback (tones with varying pitch and loudness, speech, and pre-recorded sounds) to perceive graphs and relational information. Jacobson [31] implemented an audio enabled map in a touch pad that plays back voice and natural sounds that correspond to the finger position. Parente *et al.* [32] developed an audio-haptic map using 3-D spatial sounds, where the user

Fig. 2. Visual to acoustic-tactile mapping. (left) Original color image; (middle) segmentation; (right) acoustic-tactile display.

explores the map with a mouse, keyboard, or touch screen, querying information about the points visited; the system responds with auditory icons, speech, and haptic feedback. Along these lines, "NOMAD" [33], the "Talking Tactile Maps" [34], and the TTT [10] we mentioned above, provide an embossed surface that the user scans with the finger, while an auditory signal (typically in the form of speech) is played back at certain finger positions. Another semantic approach that uses 3-D spatial sounds, is the personal guidance system proposed by Loomis *et al.* [35], [36], which determines the position and objects or landmarks surrounding the user, and uses a virtual acoustic display in which synthesized speech, played back over headphones, appears to be coming from the correct location of an object within the auditory space of the traveler.

Finally, Ribeiro *et al.* [37] proposed a passive system that sonifies objects in a real-world scene using spatialized three dimensional sounds. The spatialized sound rendition was done using HRTFs from the CIPIC database [38], intensity, and direct-to-reverberant ratio. Their approach is semantic in that computer vision techniques are used to identify objects in the scene. While this is an indirect (semantic) approach like ours, and the acoustic rendition methods are similar to ours, a direct comparison of the results is not possible because of the different tasks in the two experiments.

The goal of this paper is to explore the advantages and limitations of representing visual information, and in particular, object identity (material, shape) and position, as well as overall scene layout, in acoustic-tactile form. Since our primary interest is in dynamic devices, our main focus is on tactile sensing and kinesthetic feedback combined with acoustic display. The reason for also considering a static tactile overlay is to explore the relative advantages of the two modalities, in anticipation of devices that will be able to generate dynamic raised-dot patterns. Note that existing tactile pin arrays are bulky, expensive, and cannot be combined with touch screens [8]. The key distinguishing features of the proposed approach are the active exploration of the object/layout, the use of spatialized sounds, and the semantic, intuitive, but not necessarily realistic, mapping of objects and actions to sounds.

Our experimental results demonstrate that the most effective configurations for the perception of simple object shapes with auditory feedback provide accuracies in the upper 80% range, significantly outperforming existing approaches, like "Soundview," [39], "vOICe" [25], and TeslaTouch [12]. However, the performance degrades as the shapes become more complicated. In addition, we show that tracing of shapes (auditory feedback only) requires a significant amount of time. We believe that both of these shortcomings can be mitigated by user training. When tactile feedback is added in the form of raised-dot pattern overlays, shape perception can reach 100% for more complex, disjointed objects. We also show that the "virtual cane" configuration is quite effective in locating and identifying the objects in a simple scene layout.

In addition to the utility for the BVI community, the proposed techniques are expected to be of use in situations where vision cannot be used, e.g., for GPS navigation while driving, firefighter operations in thick smoke, and military missions conducted under the cover of darkness. Moreover, they can be used to augment visual display with sound and touch.

The paper is organized as follows. The key elements of the proposed approach and testing configurations are presented in Section II. The design of the empirical studies is presented in Section III. The experimental results are discussed in Section IV. The conclusions are summarized in Section V.

II. ACOUSTIC-TACTILE DISPLAY

In this section we present several configurations for acoustic-tactile presentation of simple shapes and layouts. In the proposed mode of exploration, the user scans the 2-D shape or layout sequentially, listening to one sound source at a time, the one that corresponds to the location of the finger. The touch screen typically responds to the centroid of the area that is in contact with the finger tip. The use of the other fingers can be useful, as for example in Braille reading where they are used for looking ahead, is not allowed in our current setup, in order to avoid any distractions while we investigate what can be achieved with one sound at a time. The serial presentation of information in our setup should be contrasted with vision, where there is a mix of sequential high-resolution, limited-view scanning by the fovea and gathering of low-resolution data over a wide receptor field in the periphery [40]. The serial presentation has also been recognized as a significant bottleneck in the perception of raised-line drawings [9], [41]. Interestingly, Loomis *et al.* [42] demonstrated that the limitations imposed by point-by-point serial presentation are not inherent to any one modality. When they narrowed down the visual field of view to the size of a fingertip and asked participants to recognize line drawings, the recognition accuracy was no better than that achieved in recognizing raised-line drawings with touch.

Another key challenge is the amount of cognitive effort that the proposed system requires for guiding the scanning finger to the object or along its boundary, which may interfere with the cognitive effort required for shape and layout perception [9]. This is in contrast to vision, where object and scene

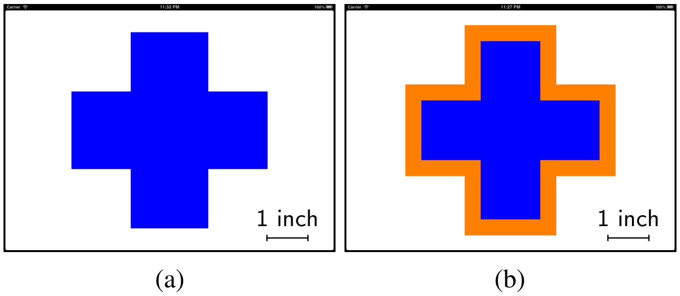


Fig. 3. Training object in the touch screen for different configurations. (a) C1. (b) C2, C4, and C5.

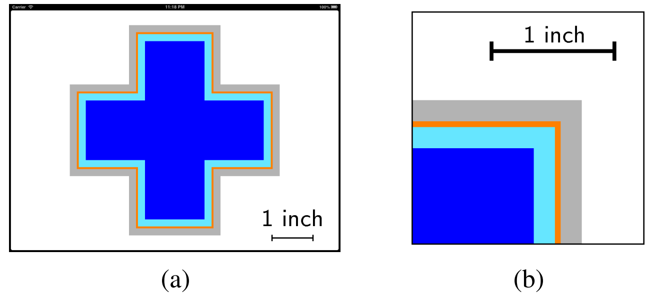


Fig. 4. Training object in the touch screen for Configuration C3. (a) Full screen. (b) Detail.

exploration is typically effortless. To reduce the cognitive effort for exploration, some of the configurations we propose in this section rely on spatial sound cues. As we will see, haptic tracing of raised-dot patterns requires little (if any) cognitive effort, and so does the tracing of simple raised-line drawings [41].

Overall, the perception of graphical information by the proposed system will by necessity be slow and considerably less accurate than vision.

A. Acoustic-Tactile Configurations

We have implemented the proposed concepts in several configurations on an *Apple iPad;* however, any other device with a touch sensitive screen can be used. In all of the configurations, we use sound played back through stereophonic headphones as the primary display and touch as the pointing mechanism for active exploration of the 2-D layout. In the first five configurations we present one object at a time on the touch screen, while in the last two we consider multiple objects in a linear arrangement, and also show that sound and touch can be combined for a more effective display.

Object shape can be represented as a line drawing or as a solid shape (a region with filled interior). Since the exploration is done by finger scanning, our approach can be compared to haptic perception of raised-line drawings [41], [43], [44]. In a study of haptic picture perception, Thompson *et al.* found that solid shapes (object interiors filled with embossed tactile textures) were easier to recognize than raised-line drawings [43]. Based on their results, we have chosen the solid shape representation for our first configuration. However, the two representations can be combined to enhance overall perception, which we have done in Configurations C2–C5.

1) Configuration C1: Shape Representation with Two Constant Sounds: The touch screen is partitioned into two regions, background and object, each represented by a distinct sound, as shown in Fig. 3(a). The advantage of this configuration is that any time the finger is touching the screen, the participant has a clear indication whether it is located inside the object or in the background. The edge of the object can be located at the transition between the two sounds.

2) Configuration C2: Shape Representation with Three Constant Sounds: The touch screen is partitioned into three regions, background, object interior, and object border, each represented by a distinct sound, as shown in Fig. 3(b). During pilot empirical studies with C1, we found that some participants were attempting to trace the edges of the object. For that, they had to move their finger in a zig-zag fashion around the edge, listening to sound transitions, to make sure that it was on the edge. This is quite awkward and confusing. Thus, in order to facilitate edge tracing, we added a relatively thin strip with a distinct sound around the border.

The strip width can have a significant effect on performance. The strip must be wide enough to avoid tracking instabilities that result when the centroid of the contact area of the finger changes due to unintended rolling and turning of the finger. On the other hand, the strip should be narrow to avoid significant variations in the scan direction. By trial and error we selected a 0.38 inch (0.9 cm) wide strip, which corresponds to 50 pixels on the 132 pixels/inch *Apple iPad* screen. This selection agrees with the work of Raja [45] who found that the most conducive line width for vibro-tactile touch screen displays is 0.35 inches. Giudice *et al.* [18] used the same width for vibro-tactile line tracing on a touch screen and obtained accuracy similar to what is achieved with embossed tactile stimuli.

However, a problem remains. Even though the participant has a clear indication that the finger is on the border, the finger may still bounce back and forth between the border and the surrounding segments (background and object interior) as it traces the border. We thus consider alternative strategies for guiding the finger along the border.

3) Configuration C3: Shape Representation with Tremolo: One approach for guiding the finger along the border is by adding proximity feedback near the border via the use of a *tremolo* signal. The idea is that, instead of a thick border defined by a constant sound, we reduce the border to a line, but add strips on each side of the border, as shown in Fig. 4. When the finger is inside the strip on the background (object) side, the background (object) sound is modified to give a clear indication that it is moving toward or away from the boundary. Otherwise, the background or object sound is constant.

Tremolo is a sound effect that is popular among musicians and can be described theoretically as low-frequency amplitude modulation without zero crossings [46]. We picked a tremolo signal with the following form:

$$Tr(t) = [1 - D + D\sin(2\pi f_R t)]\sin(2\pi f_c t) \qquad (1)$$

where f_c is the carrier frequency, f_R is the rate (modulation frequency), D is the depth (modulation degree), and where

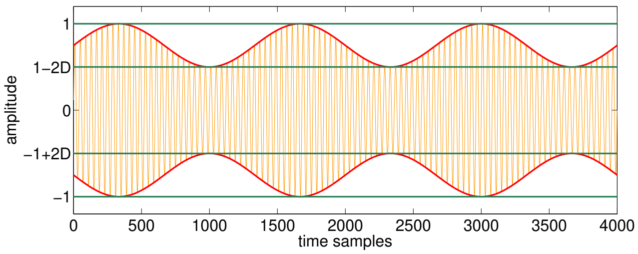


Fig. 5. Tremolo.

$f_c \gg f_R$. An example is shown in Fig. 5. This is a slight variation of the standard amplitude modulation formula, that keeps the signal amplitude within the $[-1, +1]$ range in order to avoid clipping artifacts.

Typical values for the rate are between 3 Hz and 10 Hz. As the rate (f_R) increases, the listener perceives a faster periodic sound [47]. This rate change encodes the proximity feedback to the participant. To distinguish between the background and object (both inside and outside the border strips), we use different carrier frequencies, which relate to pitch, and depths, which control the loudness fluctuations. The tremolo rate is constant within each segment (background, object), except when the finger enters the border strips, where it varies to indicate movement toward or away from the border.

However, in the first set of empirical studies reported in Section IV, we observed lower accuracy for C3 compared to C2, in spite of the additional proximity information near the border. This can be attributed to the fact that changes in the rate of the tremolo are not perceived instantaneously (users have to listen to a few periods of the signal before they can detect rate changes), combined with the relatively small border width and the relatively fast finger movements.

4) Configuration C4: Shape Representation with Three Sounds and Loudness: The fourth configuration is an attempt to combine the best attributes of C2 and C3. The advantage of C2 over C1 and C3 can be attributed to the use of a distinct sound for the border segment, and the use of instantaneous cues (timbre) to distinguish the three sounds. The challenge was thus to maintain the three distinct sounds, and at the same time, to provide a strong, instantaneous, and intuitive proximity cue within the border. To meet all these requirements, we selected a distinct border sound and used loudness variations to indicate proximity within the border. We used an exponential drop in volume with the distance of the finger from the center line of the border. We believe that this provides a close analogy with raised-line tracing, where the relief is maximal at the center and decays rapidly with distance from the center.

5) Configuration C5: Shape Representation with HRTF: This configuration explores the use of spatial sound (proximity *and directionality*) for both locating the object and tracing its border. The advantage of directionality cues is that there is no need to move the finger to find out if it is moving in the right direction. This considerably simplifies the task of finding the object and tracing its boundary, and allows the participant to focus on the perception of object shape. Also, in contrast to the previous two configurations, which provided proximity feedback only in the neighborhood of the object boundary, the idea here is that spatial sound can be used to guide the finger to the object from any point on the touch screen.

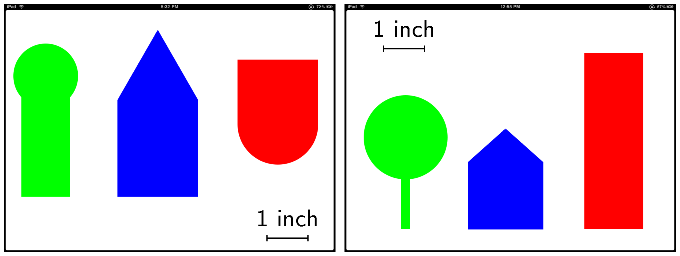
In C5, the screen is divided into three segments (object, background, and border), each with a distinct sound, as in C2 and C4 (Fig. 3(b)). When the finger is inside the object, the sound is constant. In the background and border segments, directionality is introduced via the head related transfer function (HRTF). Given the position of the acoustic source relative to the listener, the HRTF models how diffraction by listener's head, ear, and torso modifies the source sound. Accurate modeling of the sound transfer characteristics is key to synthesizing a realistic and fully immersive acoustic environment that can be presented to the participant via stereo headphones. A key to the effectiveness of sound localization is the use of sounds with sharp onsets, such as tapping or striking an object [48].

When the finger scans the background, to form the 2-D virtual acoustic scene we assume that the listener is in the position of the scanning finger, facing toward the top of the screen. The sound source is assumed to be located inside the object at the point nearest to the finger location (rather than at the center of the object). The sound intensity is proportional to the inverse of the squared distance between the finger and the source (plus a DC offset so that the sound is always audible).

When the finger scans the border, the same assumptions for the virtual listener hold as in the background, but the source that emits the characteristic border sound is placed in the direction that the user must follow in order to track the border in the clockwise direction. Thus, the directional sound guides the participant around the border. In this case, the sound intensity remains constant.

6) Configuration C6: Scene Perception – Virtual Cane: In this configuration, our goal is to render a simple scene consisting of several objects. The goal is to convey the relative position, size, identity, and material composition of each object in the scene. The rendering of shape was considered in the previous configurations, and can be handled separately, e.g., in a zoomed-in mode, one object at a time, as explained below. We thus focus on the layout and other object attributes. For object identification, we can map each object to a unique, distinguishable, but otherwise arbitrary sound. However, a more intuitive assignment is desirable, and since the participant moves the finger on the screen to explore the objects, a characteristic rubbing sound would be the obvious choice. On the other hand, we found that tapping on the objects produces more distinguishing sounds than rubbing, and since this is the virtual world, the interface has to be effective, not realistic. The inspiration comes from the "long cane," the oldest and most widely used visual substitution tool.

BVI people use the long cane for navigation, continuously tapping their surrounds to detect and identify objects, obstacles, and other landmarks. However, it is limited to the immediate vicinity in front of the BVI person. The long cane provides valuable information about the location, shape, size, and even material composition of the objects, not only "feeling" the objects, but also listening to the tapping sounds. In this



(a) (b)

Fig. 6. Test scene in the touch screen for configurations: (a) C6 and (b) C7.

configuration, we imitate the idea of a BVI person exploring a scene (e.g., an outdoor scene outside her/his window) using a *virtual cane* (a very long cane in the case of the outdoor scene) to tap on the objects. This could be realized using a camera to snap a picture, analyze it to obtain meaningful segments (objects or parts thereof), and then represent each segment with a characteristic tapping sound. Of course, obtaining a meaningful region-based representation can be quite difficult. Here, we assume the availability of a semantic representation.

In our initial setup, we assumed that the objects are disjoint, and in a linear arrangement, as shown in Fig. 6(a); However, more complicated layouts with connected (touching) objects, as in Fig. 1(b), and even overlapping objects (occlusion), can also be handled by the same approach. A characteristic tapping sound is assigned to each object, while the background is silent. To further explore the shape of a selected object in finer resolution, the user can enter the zoomed-in mode. In this mode only the selected object is present, zoomed-in at the center of the screen. Any of Configurations C1–C5 can be used here, preferably the most effective. A reserved gesture (*e.g.,* double-tapping on the screen) can be used to go back and forth between the two modes, while a special sound can be used to confirm the mode change.

7) Configuration C7: Scene Perception with Overlaid Tactile Imprint: In this configuration, we test the joint perception of acoustic and tactile signals, by superimposing a raised-dot pattern embossed on paper on the touch screen. In our initial setup, we assumed that the objects are disjoint, as shown in Fig. 6(b), and used a dense tactile pattern to represent the objects on a flat background. The advantage of the tactile overlay is that object shape is much easier to perceive without the need for a zoomed-in mode, while the sound can be used for object and material identification. To discriminate between touching objects, we would have to use perceptually distinct tactile patterns. Different tactile patterns can also be used for object/material identification. Of course, the more patterns we add, the more difficult it will be to tell them apart. The disadvantage of the tactile overlay is that it is static. However, one could use the tactile patterns to display fixed objects (buildings) and sound to display moving objects (cars, people).

III. Empirical Studies

We conducted a series of empirical studies, one for each of the seven configurations described in the previous section. For the first five configurations, the studies were conducted with three different shapes (one trial for each shape), while for the last two configurations, the study was conducted with just one scene layout (one trial).

We conducted the empirical studies in two sets. In the first set, we tested Configurations C1–C3 and C5–C7. As we discuss below, on the basis of the experimental results (discussed in Section IV), we decided to modify C5 and to add C4. We then conducted the second set of studies with these two configurations and, for comparison (since we used a new group of participants), with the original (unaltered) C2.

A. Participants

The participants in our studies were not experts in acoustic or tactile signal processing and perception, and were not familiar with the detailed goals of the studies. They had different educational backgrounds and different degrees of experience with touch-screen devices. There was no financial reward for participation in the studies.

The first set of studies was conducted with 20 participants, 16 male and four female. The age of the participants ranged from 19 to 56 years old (average 29). All except two reported normal or corrected vision and normal hearing. One participant reported nystagmus (uncontrolled eye movement) from birth and another reported tinnitus (ringing in the ears) for the last 20 years; both participants had been treated for their impairments. All 20 participants completed the studies with C1, C2, C3, C5, and C6. Only four of the 20 participants took part in the study with C7. The participants with nystagmus and tinnitus participated in all the studies, except the one with C7. Two additional female participants, aged 21 and 35, carried out one experiment each, with C2 and C7, respectively.

The second set of studies was conducted with a different group of 11 participants, ten male and one female. The original group of participants could not be used because they were already familiar with the shapes and layouts of the studies, which we had to keep the same in order to be able to compare the results with the first set of studies. The participant ages were in the range of 22 to 50 years old (average 33). All except one reported normal or corrected vision and normal hearing. One participant reported a hearing deficiency in the left ear that had not been treated. All 11 participants completed the studies for C2, C4, and the modified version of C5.

B. Procedure

All the studies were performed by the first author in a quiet room (student office, no sources of noise from outside, but not soundproofed) to avoid disturbances. The participants interacted with a touch screen and listened to auditory feedback on stereo headphones. The participants were asked to scan the screen using only one finger; however, they could select any finger and were also allowed to switch fingers in order to avoid discomfort.

In all of the studies, we blocked any visual contact of the participants with the touch screen and the scanning finger, in order to eliminate visual cues for the perception of the graphical information presented in acoustic-tactile form. This is because watching the finger movements can provide strong shape identification clues, which would not be available to

BVI participants. To eliminate all visual cues while allowing the participant to have visual contact with the experimenter, we adopted a setup proposed by V. Tartter (City College New York) [49]. The touch screen was placed in a box open in the front, so that the participant could insert her/his hand to access the screen. The box was placed on a table, and the participant was seated in front of the table. The touch screen was placed horizontally in the box in landscape orientation, and participants were asked not to move it during the experiment.

The participants were given written instructions at each stage of the studies. Each time they were given time to ask questions, until the instructions were completely clear. At the very beginning, they were given a general introduction for the entire set of studies, and then at the beginning of each experiment, they were given configuration specific instructions. For each configuration, the participants were shown a training example. The shape or scene layout for the training example was different from those used in the actual experiment. During the training examples, the participants were at first able to see the scanning finger and the shape/layout to be presented on the touch screen; then, they repeated the trial visually blocked as in the actual experiment.

The goal of the studies was to identify an unknown shape, rather than selecting among a set of predetermined shapes. We did not provide any shape information, except for the fact that the boundaries were piecewise smooth. In addition, the written instructions made it clear that each trial is independent of other trials, so that each shape could be presented more than once for a given or multiple configurations. Moreover, we did not provide any feedback to the participants on their responses, until the end of all the studies. At that point, the participants were also asked for comments. The participants were also free to give comments at the end of each trial, and many of them did even though they were not specifically asked to do so.

To get a more accurate representation of what the participants perceived, at the end of each trial, we asked them to draw the perceived shape on a piece of paper, and then asked them to name it or to describe it verbally. Similarly, the participants were asked to draw the object layouts of Configurations C6 and C7, to determine the number of objects, and to name the material of each object. The participants had full visual contact with the paper and the pen while drawing. They were asked to draw something, even if they were not sure. The rating of each response was binary (correct/incorrect) based on the verbal response and the drawing. In cases of ambiguity, the participants were asked to explain. The drawing made the task more demanding, as rotated or transposed versions of a shape were not considered as correct answers. Our performance evaluation criteria did not include any contour matching.

Interestingly, Wijntjes *et al.* [41] found that asking the participants to sketch the shape they perceived improved the identification of raised line drawings. We believe that this relates to the kinesthetic memory of the shape, which helps participants to directly compare the shape they traced with the memory of tracing similar shapes, visually (for sighted participants) and kinesthetically (for BVI participants).

There were no tight time limits for the studies, but the actual time durations to complete the studies (including the time to draw and describe the shapes/layout) were manually recorded. Since the total time for the first set of studies was one to two hours, the participants were given the chance to take a break after the first three configurations. However, all of the participants completed the entire set of studies in one sitting, except one who completed the studies in two sittings on consecutive days. The second set of studies was a lot shorter, so there was no need for a break. In the second set, after the trials for each configuration, we asked the participants to rate the difficulty of the configuration, in a 1–10 scale, with 10 being the most difficult.

C. Equipment and Materials

1) Touch screen: All the configurations were implemented in an *Apple iPad 1,* which has a 9.7-inch (diagonal) 4:3 screen multi-touch display with fingerprint-resistant oleophobic coating. The optical resolution of the screen was 1024×768 pixels at 132 pixels per inch, and the range of the audio playback frequency response was from 20 Hz to 20 KHz. The participants were allowed to adjust the playback volume to a comfortable level, at any point during the studies. All the sounds (except tremolo) were played back as pre-stored sound files in Microsoft/IBM Waveform Audio File Format (WAV).

2) Headphones: We used Sennheiser HD595, around-the-ear stereo headphones (frequency response 12 Hz – 38,500 Hz, sound pressure level 112 dB at 1 kHz and 1 Veff).

3) Tactile Patterns: In C7, a “VersaPoint Duo Braille Embosser” (Model# BP2B-01) was used to emboss tactile patterns on paper. The paper with the patterns was then superimposed on the touch screen so that they are aligned with the corresponding regions on the touch screen.

D. Sound Design

The selection of the sounds for representing the elements of the layout was critical for the success of the studies.

In the studies with C1, C2, C4, and the first set of studies with C5, we used SONAR pings for the background because they have been found to work well in navigation tasks (e.g., [50]). For the representation of the objects, we selected pre-recorded (wood) tapping sounds, as they provide strong clues for the material, emulating the use of a long cane. For C2, C4, and C5 (first set), we picked synthesized chirp sounds for the border, in order to clearly differentiate it from the object and background. The duration of the chirp was 350 ms and the frequency sweep from 100 to 400 Hz. For C1–C4 we used monophonic sound. As we discussed, in C4, we used loudness variations to indicate proximity.

For the studies with C3, we used constant tremolos with $D_1 = 0.25$, $f_{c1} = 200Hz$, and $f_{R0} = 6Hz$ for the background region and $D_2 = 0.42$, $f_{c2} = 150Hz$, and $f_{R0} = 6Hz$ for the object region. In the strips on either side of the border, the tremolo rate f_R was varying from 6 to 22 Hz and the depth and carrier frequency was the same as that of the corresponding (background or object) segment.

In Configurations C5 and C6 we used the HRTF to render directional sounds. Ideally, an individual HRTF should be measured for each participant; however, for practical reasons, we used a best match from a set of pre-measured HRTFs from the CIPIC database [38]. To select an HRTF, we designed a simple application, in which the participant virtually moves around a sound source listening to rendered directional sounds for each of the available HRTFs, and picks the one that works best.

For the studies with C5, we used directional stereophonic sound in the background and border segments, and constant monophonic sound inside the object. In the first set of studies, the participants had to select one of two HRTFs, corresponding to long and short pinnae measured on the KEMAR mannequin [38].

In the second set of studies with C5, we considered two modifications in order to improve the perception of directional sound. First, we reconsidered the sound selection for the background and border segments. We ran a small experiment with six participants, in which we asked them to rate five sounds on the basis of their effectiveness in rendering directionality. The five sounds were: a sonar ping, a low-frequency chirp, a high-frequency chirp, wood tapping, and Gaussian noise. Wood tapping and Gaussian noise proved to be the two best directional sounds, and were assigned to the border and background segments, respectively. The mono chirp signal was used inside the object. Second, we used HRTFs that correspond to humans. We selected five human HRTFs from the CIPIC database [38] to represent a wide range of anthropometric data. Our expectation was that the human HRTFs would provide a closer match to a given participant than those of a mannequin. On the other hand, Fisher and Freedman's studies [51], [52] show that, when users are allowed free head movements, the localization accuracies in three conditions–utilizing their own pinnae, artificial pinnae, and no (occluded) pinnae–converged. Thus, if free listener movement in the virtual space of our studies is analogous to free head movements, the effect of a particular HRTF selection would not matter. However, the actual effects of the various HRTFs can only be determined experimentally.

The sounds used in the configurations were of various durations from 500 ms to 1.25 s, and all were played in a continuous loop. Repeated sounds create onsets, which provide vital cues for sound localization (in C5 and C6) [53].

E. Other Experimental Details

1) Configurations C1–C5: In the studies with the first five configurations, the training shape was a cross and the test shapes were a square, a circle, and an equilateral triangle, presented one at a time. With the exception of the training shape (cross), all of the shapes had roughly the same area in square pixels, and were centered in the touch screen. The width of the border strips for C2, C3, C4, and C5 was 0.38 inches (50 pixels). For each configuration, each of the three shapes was presented once, in random order, even though the participants were told that repetitions were possible. However, the order of the configurations was fixed. In C2–C5, the participants were told that the border strip(s) were introduced to facilitate edge tracing. However, they were free to use any technique they wished (e.g., scanning the screen left to right, top to bottom) in order to perform the task of shape identification.

2) Configuration C6: In the zoomed-out mode, we used prerecorded mono tapping sounds of wood, glass, and metal to represent up to three different objects. The background was silent. The mode change was triggered by double-tapping inside the object. A short-duration sound was used for the mode change notification. The zoomed-in mode was the same as C5 in the first set of studies, except for the choice of object sound. The training scene contained two objects, and the testing scene contained three objects that were different from those used in the training scene. In both cases, the objects were disjoint with roughly equal horizontal spacing.

In the instructions at the beginning of the experiment, the analogy with a virtual cane was provided. The participants were told that the number of objects in the experiment could be different than that of the training example. The participants had no prior information about the set of possible sounds for the objects. They were told that the zoomed-in mode was for shape identification; it could not be used for determination of object size, as the degree of zooming varied with object size.

3) Configuration C7: The sound signals for this configuration were the same as those for C6 (wood, glass, and metal). However, a sheet of paper with embossed tactile patterns was superimposed on the touch screen, so that the tactile and acoustic patterns were aligned. Since the objects were disjoint, the same tactile texture (dot pattern, density, height, and size) was used for all objects, while the background was flat.

The training scene contained the same two objects as in C6 while the testing scene contained the three objects shown in Fig. 6(b), which are different from those in C6. However, in order to address significant errors in material identification observed in studies with C6, we decided to narrow down the number of possible materials. Thus, at the beginning of the experiment, the participants were presented with tapping sounds labeled wood, glass, metal, cardboard, plastic, and composite.

IV. Experimental Results

As we discussed in Section III, we conducted the shape perception experiments in two sets. In the first set, we tested Configurations C1–C3 and C5–C7. On the basis of the results, we then added C4, and conducted a second set of experiments with that, a modified C5, and the unaltered C2. The seven configurations are summarized in Table I.

At the outset, we should point out that the experiments were very time consuming and were carried out with sighted, visually-blocked, unpaid volunteers. As a result, we could only conduct a limited number of trials for each configuration. For this reason, as we will see, some of the results are not statistically significant at the alpha level of 0.05.

In the shape experiments, performance was measured by the accuracy of the participant response and the time it took to complete the experiment. As we discussed, the rating of each response was binary (correct/incorrect), based on the verbal description and the drawing. Since the accuracy is a categorical variable (nominal data), we used the chi-square test of independence (TOI) and goodness of fit (GOF) for categorical analysis of the accuracies [54]. The timing data, on

TABLE I
SUMMARY OF CONFIGURATIONS

C1	2 constant sounds
C2	3 constant sounds
C3	2 tremolo sounds with varying border rate
C4	3 sounds with varying border intensity
C5	3 Sounds with HRTF in border and background
C6	Virtual cane - acoustic display with zoomed-in mode
C7	Virtual cane - acoustic display with tactile overlay

TABLE II
FIRST SET OF EXPERIMENTS: PERFORMANCE OVER ALL PARTICIPANTS (SQUARE, CIRCLE, TRIANGLE)

	C1			C2		
	SQ	CI	TR	SQ	CI	TR
Accuracy (%)	85	40	75	86	76	81
Aver. Accuracy (%)	66.7			80.9		
Time (s)	221.5	186.5	172.5	158	191	185
Overall time (s)	196			180		
	C3			C5		
	SQ	CI	TR	SQ	CI	TR
Accuracy (%)	80	55	80	95	60	85
Aver. Accuracy (%)	71.7			80.0		
Time (s)	193.5	170	124.5	129	195	170
Overall Time (s)	161.5			147.5		

the other hand, is continuous but, due to limited sample sizes, high variance, and outliers, a normal distribution cannot be safely assumed. Therefore, we resorted to a non-parametric method, the Kruskal-Wallis H test (KWHT) [55] for the statistical analysis of the timing data.

A. *First Set of Experiments with Shape Configurations*

The results of the first set of experiments with C1, C2, C3, and C5 are summarized in Table II, which shows accuracy and time averaged over all participants for each shape and configuration. The data were analyzed using the chi-square TOI and it was shown that the accuracy is independent of the configuration ($\chi^2(3, 243) = 4.55$, $p > 0.2$), that is, there are no significant differences in accuracy among configurations. However, in a pairwise comparison using chi-square GOF, the performance of C2 is shown to be significantly better than C1 ($\chi^2(1, 123) = 7.94$, $p = 0.005$). C5 also performs significantly better than C1 ($\chi^2(1, 120) = 6.67$, $p = 0.01$.). These observations justify the addition of a border strip with a distinct sound. On the other hand, the difference between C3 and C1 is not significant ($\chi^2(1, 120) = 0.74$, $p = 0.39$), which indicates that the tremolo is not effective.

Fig. 7 shows box plots for the time it took the participants to identify each shape for each configuration, as well as combined across the different shapes. The red line indicates the median, the box edges indicate the 25th and 75th percentiles, each whisker extends the box by 1.5 times its length, and the crosses show the outliers (outside the range defined by the whiskers). KWHT showed significant differences among the configurations ($H = 7.91$, $dof = 3$ (degrees of freedom), $p = 0.048$). In particular, paired comparisons between C1 and C5 ($H = 5.28$, $dof = 1$, $p = 0.02$) and between C1 and C3 ($H = 5.78$, $dof = 1$, $p = 0.02$) showed that C5 and C3 require significantly shorter time compared to C1. Thus, C5 is the only configuration that is significantly better than C1 in terms of both accuracy and timing. This can be attributed to the addition of spatial sounds in C5.

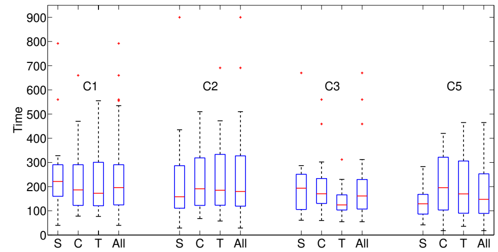


Fig. 7. First set of experiments: Time distribution for each shape of C1, C2, C3 and C5 (S: square, C: circle, and T: triangle).

Based on the statistical analysis of the results, we have established that adding a distinct border sound improves accuracy and that the addition of spatial sounds (in C5 and C3) improves timing. However, we have no direct statistical evidence of the superiority of C5 over C3 or C5 over C2. On the other hand, the feedback from the participants after the experiments provides an indication of the ineffectiveness of C3; there were numerous negative comments, such as, “inside/outside of shapes were not differentiable by assigned tremolos,” “tremolo rate changes very fast within a small area,” and “tremolo rate changes were not noticeable.” Given the negative feedback, the explanation for the relatively fast time of C3 may be that, when faced with an ineffective interface, the participants give up or make a haphazard guess. We expect that more extensive studies will establish the superiority of C5 over C3 in terms of accuracy.

Overall, out of the four configurations, C5 received the most positive feedback, with special emphasis on the ease of use. According to the comments, the addition of directional sounds in the border segment was quite helpful for tracing the edges, and also provided cues about edge orientation. On the other hand, the participants reported that spatial sound did not help much in the background region. This is not surprising as the object was relatively large and at the center of the screen, which made it easy to locate even without spatial sound.

It is important to point out that the accuracies are much greater than what would be achieved by mere guessing. The results are strengthened by the fact that the participants did not have any prior knowledge about the shapes and were not given any feedback on their performance during the test. Fig. 8 shows some interesting drawings from both sets of the shape experiments. The first row shows drawings that were marked as correct. All the remaining ones were marked as wrong. The diversity of the drawings is a clear indication that the participants did not have any prior knowledge of the shapes.

Fig. 9 shows the relationship between time and accuracy for each participant. As can be seen, there were wide performance variations. Note also that about a third of the participants have over 80% average accuracy and median time of less than 3 minutes. In addition, only three of the participants (14%) have median time over four minutes and only three have accuracy less than 50%.

The data in Table II also indicate that the accuracy varies with shape. Indeed, statistical analysis shows that the par-

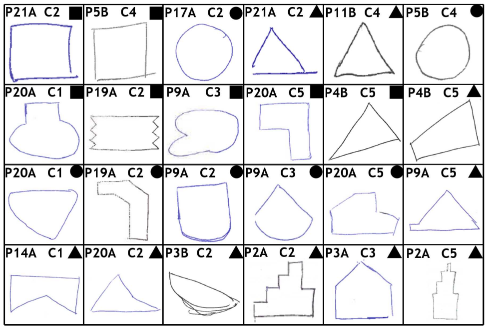


Fig. 8. Selected participant drawings for C1–C5 (P21A: Participant 21, first set of experiments; P11B: Participant 11, second set of experiments).

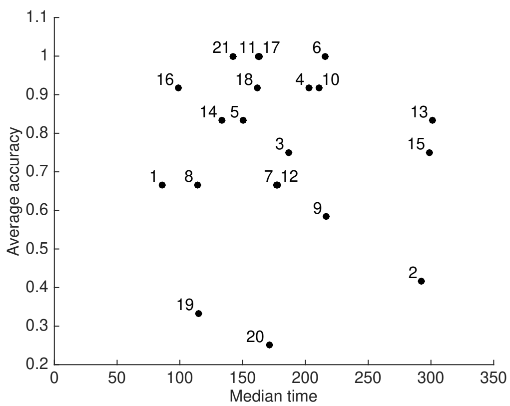


Fig. 9. First set of experiments: Average accuracy *vs.* median time (s) for each participant in C1–C3 and C5 (Points labeled with participant number).

ticipants were significantly less accurate in identifying the circle compared to the other two shapes ($\chi^2(2, 243) = 19.22$, $p < 0.001$). Overall, one would expect that the detection of curved edges is more difficult than that of straight edges and corners. Fig. 10 shows some interesting drawings by participants in response to the circle stimulus for various configurations, all of which were marked wrong. The first four were drawn by Participant 7 in the first set of experiments (P7A). The participant called the first a hexagon and the other three octagons; all of these drawings can be considered as straight line approximations of a circle. The next four, drawn by Participant 2 (P2A), can be considered as pixel approximations of a circle. Note that for both participants the crudest approximation corresponds to C1. The remaining drawings in Fig. 10 were drawn by six different participants and include line and pixel circle approximations. Actually, the triangle approximations by Participant 2 in Fig. 8 are also pixel approximations. The use of these approximations (line and

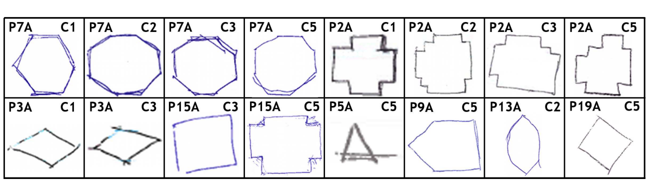


Fig. 10. First set of experiments: Selected participant drawings, all marked wrong, in response to circle stimulus in C1–C3 and C5

TABLE III
RESULTS OF EXPERIMENTS WITH C6 AND C7

Accuracy (%)	Configuration C6			Configuration C7		
	Obj. 1	Obj. 2	Obj. 3	Obj. 1	Obj. 2	Obj. 3
Number of Objects	100			100		
Material Identification	90	80	70	100	60	60
Shape Identification	20	30	20	100	100	100
Shape (aver.)	23.3			100		
Overall Median Time (s)	709			185.5		

pixel) is due to the difficulty of tracing the object boundaries. However, it could have also been influenced by the use of a training shape (cross) with only horizontal and vertical lines.

B. Experiments with Layout Configurations

Table III summarizes the results of the experiments with C6 and C7. In both configurations, the participants had no problem determining the number of objects in the scene. In C6, there were significant errors in material identification, which prompted us to narrow down the number of possible materials in C7. However, material identification errors persisted, this time confined to confusions between glass and metal, which were difficult to distinguish. This calls for more distinguishable sounds, even if they are not as realistic. In the zoomed-in mode of C6, which all of the participants used for object shape identification, the accuracy was remarkably low compared to the simple shapes we used for testing in C1 to C5. As we explained in Section II-A6, the zoomed-in mode is the same as C5. However, the objects in the layout configurations are considerably more complicated, which is most probably the reason for the lower performance. It is notable, however, that two of the participants had perfect responses.

Fig. 11 shows interesting drawings from the experiments with C6. The first row shows successful drawings and the second row shows failed attempts to identify the shapes in the first row. Again, these can be considered linear or pixel approximations of the actual shapes. Note, however, that there is also a curved line approximation of a piecewise linear segment (roof). It is interesting to note that, in this case, the training samples did include curved lines, yet the participants still used pixel and line approximations. Note that the pixel approximations were made by Participant 2, who also drew the pixel approximations of the circle in Fig. 10 and the triangles in Fig. 8. All these shape deformations can be explained by the "temporally extended exploration" with the finger, which may turn curved into straight lines [9], [56] and may cause other haptic illusions [57].

The median time for the experiments with C6 was 11.8 minutes. The recorded time included the exploration in both modes and the time taken to draw and label the scene on paper. The median time corresponds to 4 minutes per shape, which is higher than the 2.5 minutes per shape for C5. As in the case of accuracy, this may be explained by the fact that the objects in C6 are considerably more complicated.

In contrast, C7 resulted in perfect shape identification and much faster median time (a little over 3 minutes). Statistical analysis showed that C7 has significantly better performance

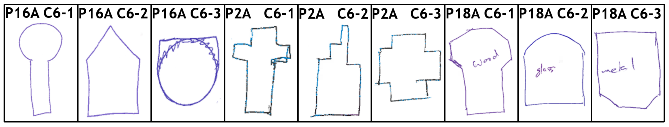


Fig. 11. Selected participant drawings in experiments with C6: First row marked correct, second row marked wrong (C6-1: First object in C6).

TABLE IV
SECOND SET OF EXPERIMENTS: PERFORMANCE OVER ALL 11 PARTICIPANTS (SQUARE, CIRCLE, TRIANGLE)

	C2			C4			C5		
	SQ	CI	TR	SQ	CI	TR	SQ	CI	TR
Accuracy (%)	100	45	64	100	64	82	91	64	64
Aver. Accuracy	69.7 %			81.8 %			72.7 %		
Time (s)	131	372	240	167	274	186	65	122	141
Overall Time	202 s			186 s			109 s		
Difficulty (1-10)	6			6			3		

than C6 in both accuracy ($\chi^2(1, 72) = 25.48$, $p < 0.001$) and timing (KWHT $H = 7.35$, $dof = 1$, $p = 0.007$). This is because the use of raised-dot patterns in a flat background is much better suited for shape identification than sound rendition, thus, eliminating the need for the time consuming zoomed-in mode. As we discussed, the performance is expected to decrease when two or more distinct raised-dot patterns are used to differentiate adjacent objects.

Unfortunately, the tactile feedback is static, which is not well-suited for interactive applications. The development of dynamic tactile devices is expected to remedy this problem in the not too distant future. An alternative approach is to provide feedback through the use of vibrations or variable friction. However, neither of these is expected to have the shape rendition accuracy of the raised-dot patterns for reasons that will be discussed in Section IV-D.

C. Second Set of Experiments with Shape Configurations

Table IV summarizes the results of the second set of experiments, conducted with C2, C4 and modified C5. Fig. 12 shows box plots for the time it took the participants to identify each shape for each configuration, as well as combined across the different shapes. As we discussed in Section II, we introduced C4 in order to provide a strong, instantaneous, and intuitive proximity cue within the border. In Section III, we discussed the reasons for the sound design modifications of C5. We also explained the reason that we had to use a different set of participants. Thus, in order to be able to compare the performance of the two sets of participants, we included C2 in both experiments.

Comparing with Table II, note that there is a 12.5% drop in accuracy and a 12% increase in time for C2 in the second set. The difference in accuracy was marginally statistically significant ($\chi^2(1, 96) = 3.78$, $p = 0.05$), while the difference in time was not significant (KWHT $H = 1.2$, $dof = 1$, $p = 0.27$). The variability of C2 performance in the second set also seemed to be high. Even though the statistical evidence was not too strong, we nevertheless tried to find an explanation.

The drop in performance could be attributed to the natural abilities of the participants or the prior experience with touch-screen devices. Indeed, a closer look at the background of the participants revealed that six of the participants owned and used touch-screen devices (smart phones, tablets, or tablet PCs) in their day-to-day life (experienced group), while the remaining five had very little experience with touch-screen devices (inexperienced group). In contrast, in the first set of experiments, only one participant belonged to the latter category. In fact, the performance of the inexperienced group (mean accuracy 62.2%; median time 273 s) was significantly lower than that of the experienced group (mean accuracy 85.2%; median time 113 s), with $\chi^2(1, 99) = 12.11$ and $p = 0.0005$ for accuracy and KWHT $H = 12.29$, $dof = 1$, $p = 0.00001$ for time.

TABLE V
SECOND SET OF EXPERIMENTS: PERFORMANCE OVER 6 OF THE 11 PARTICIPANTS (SQUARE, CIRCLE, TRIANGLE)

	C2			C4			C5		
	SQ	CI	TR	SQ	CI	TR	SQ	CI	TR
Accuracy (%)	100	50	83.3	100	66.7	100	100	83.3	83.3
Aver. Accuracy	77.8 %			88.9 %			88.9 %		
Time (s)	121	200	166	126	111	129	57	96.5	82
Overall Time	166 s			121 s			69 s		
Difficulty (1-10)	7.5			6			2.75		

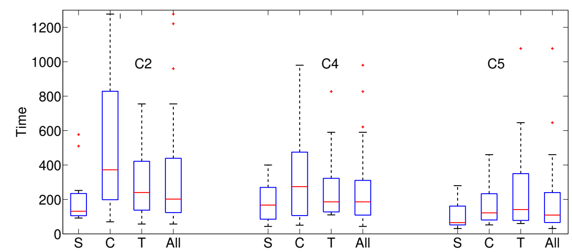


Fig. 12. Second set of experiments: Time distribution for each shape of C2, C4, and C5 (S: square, C: circle,, and T: triangle)

Table V summarizes the results of the experiments for the experienced group of six participants. Note that the overall performance for C2 is now comparable in the two sets of experiments, with both C4 and C5 doing better in terms of accuracy and time. In particular, the modified C5 reported a 53% reduction for the experienced group in average time compared to the C5 in the first set of experiments (KWHT $H = 9.95$, $dof = 1$, $p = 0.002$), which justifies the modifications.

We now focus on the second set of experiments and the performance comparison of the three configurations. In our statistical analysis we use all 11 participants, as it is difficult to draw significant conclusions with just six participants. When the accuracies of three configurations were analyzed using the chi-square TOI, it was shown that the participant performance was independent of configuration ($\chi^2(2, 99) = 1.4$, $p = 0.5$). Paired comparisons did not show any significant differences either; in particular, they showed no significant advantage of C4 over C2 ($\chi^2(1, 66) = 3.26$, $p = 0.07$). On the other hand, the time differences among configurations, were found to be significant (KWHT $H = 8.49$, $dof = 2$, $p = 0.01$). Our statistical analysis demonstrates that C5 significantly outperforms C2 in terms of timing (KWHT $H = 7.81$, $dof = 1$,

$p = 0.005$). It also shows significant differences in timing for C5 over C4 (KWHT $H = 4.13$, $dof = 1$, $p = 0.004$). In addition, we asked the participants to rate the difficulty of the task and KWHT showed significant differences in the ratings of the three configurations ($H = 7.56$, $dof = 2$, $p = 0.02$). C5 was the easiest, followed by C4, and then C2 (for C5 over C2 $H = 6.00$, $dof = 1$, $p = 0.01$ and for C5 over C4 $H = 4.92$, $dof = 1$, $p = 0.03$).

The faster times achieved with the C5 modifications justify the better sound selection and closely matched HRTFs. Our results establish that C5 is the fastest and easiest to use configuration, and that C4 is better than C2. Thus, the importance of utilizing instantaneous acoustic cues (directionality in C5 and distance in C4) for shape exploration becomes apparent. Klatzky *et al.* [9] point out that the lack of natural guidance of the finger along the object boundary may demand considerable cognitive effort, and failure to maintain contact with the boundary causes transient loss of information, memory loading, and distortions of the kinesthetically acquired mental map. The spatialized acoustic cues of Configurations C4 and C5 help mitigate such problems, resulting in faster response and ease of use. In fact, it may be beneficial to combine the two, for example, using directional sound to guide the finger along the border and loudness to keep it on the border. Overall, spatial sound can be used to provide a natural, intuitive interface for exploration, that allows the user to focus on the perception of object shape and scene layout.

Further improvements can be obtained by using individually calibrated HRTFs, as well as finer angular quantization and interpolation. Such improvements are costly and time consuming, but may be worthwhile, especially for BVI participants.

More importantly, the fact that participants familiar with touch-screen interfaces significantly outperformed participants without such experience, points to the importance of training. Since none of the participants in either experiment received any systematic training, it should be clear that there is a lot of room for performance improvement with extended training and experience. Indeed, O'Modhrain *et al.* [8] emphasize the need for training by users of multimodal haptic/audio interfaces.

D. Comparison With Existing Techniques

We first compare our results with those of *Soundview* [39], which was implemented in a graphical tablet using a pointing device for scanning. Due to the differences in the setups and the fact that the authors report only the mean values for their results, we cannot perform statistical analysis for the comparison. *Soundview* used two sounds, one inside the object and one in the background, as in C1. However, the sound played to the participant at a given time depended on both the location and the velocity of the pointer. In addition, they used six shapes (square, circle, and triangle, with and without a hole in the middle). In contrast to our experiments, they allowed participants to have visual contact with the tablet (shape visually hidden) and scanning pointer, thus indirectly using vision for shape identification, which is unrealistic for BVI participants. They used three different experimental setups. In the first, the participants did not know the shapes they were going to be tested on, but they were told that they would be simple shapes, and had to draw the shape after each trial. This is the closest to our experimental setup. In the second, the participants were asked to choose the shape they perceived from a set of 18 shapes. In the third, they had to pick one of 6 possible shapes. Each shape was presented once, and the participants were not explicitly told whether shapes could be repeated or not. The overall accuracy for the three experiments was 30%, 38% and 66%, respectively. In all of their setups, there was a time limit of 90 seconds for perceiving a shape and 90 seconds to record their response. We believe that the poor performance of *Soundview* can be attributed to the dependency of acoustic stimuli on the velocity of the pointer, which makes it too complicated for participants to decode. In addition, they tested the performance of the *vOICe* system (described in Section I) in their six-shape setup, and found that the overall accuracy was only 31.0%. Overall, our experimental setup was harder (no indirect use of vision) than the first and hardest of their setups, yet, our experiments demonstrated significantly better performance than the easiest of their setups.

We also compare with the TeslaTouch experiments reported in [12], as they also attempted to convey simple shapes (triangle, square, and circle). As in the third *Soundview* experiment [39], the participants had to select one out of a small set of shapes (three). They used three types of shape rendering: solid (as in C1), outline only, and solid with outline (as in C2). They tested these configurations with three blind participants. There were no time limitations for the experiments. They reported just below 80% accuracy for the solid rendering, and just above 40% for the other two configurations. The average time per trial was less than two minutes. While their results for the solid configuration were about as good as those of our best configuration, one should keep in mind that the task of discriminating among three known shapes is a lot easier than that of identifying (and drawing correctly) an unknown shape.

It is interesting that in the variable friction rendition, the solid shape outperforms the solid with outline configuration, while in our acoustic renditions, the solid with outline performs best. In fact, the authors also report that participants had difficulty following the object edge in the outline configurations. This can be attributed to the fact that there is no stimulation gradient across the finger to provide directionality or distance cues; this is because at any point in time the friction of the entire display is fixed and changes only with the position of the finger. Thus, at any point in time, the entire fingertip feels the same friction. This argument was made by Klatzky *et al.* [9] for vibration displays, but applies equally well to friction displays. This is also the reason for the limited spatial resolution of friction and vibration displays; an edge is sensed only when the friction of the entire display changes. This is in contrast to line drawings and raised-dot tactile patterns, where the fingertip is simultaneously exposed to the raised and flat parts of the surface, and can thus feel the orientation of the ridges in line drawings and edges in raised-dot patterns, as is elegantly explained in [9].

The limited spatial resolution argument also applies to sound renditions on a touch screen. However, sound signals offer a multiplicity of dimensions (intensity, frequency, directionality, timber, etc.) for conveying directionality and distance, as well as other information. In contrast, friction offers only intensity, and vibration offers intensity and frequency. Finally, the fact that the use of a third friction level for the object outline does not improve performance is also an indication that the finger is sensitive to changes in friction rather than absolute levels.

V. Conclusions

We have proposed a new approach for conveying graphical and pictorial information via hearing and touch, and showed that it can be applied to the perception of basic geometric shapes, significantly outperforming existing approaches. We have also shown that our approach can be used to locate and identify the objects in a simple scene layout, using what we call a "virtual cane," an interface that plays back tapping sounds as the user explores the scene on a touch screen.

Empirical studies with visually-blocked sighted participants and a number of acoustic-tactile configurations demonstrated the advantages of spatial sound (directionality and proximity cues) for dynamic display of information, and that raised-dot patterns provide the best static shape rendition. Our studies also showed the limitations of acoustic-tactile interfaces, namely that exploration is slow and performance degrades with layout complexity, but also emphasized the importance of training; both are in agreement with the existing literature [8], [9].

Our empirical studies were very time consuming and were carried out with unpaid volunteers, with very little training. Thus, our experimental data were limited compared to what typical studies produce, and as a result, some of our conclusions are not statistically significant. However, by exploring a wide variety of design alternatives and focusing on different aspects of the acoustic-tactile interfaces, our results offer many valuable insights for the design of future systematic tests, utilizing the most effective configurations, with paid (and therefore more committed) blind and visually impaired as well as visually blocked participants.

Looking forward, our results indicate that we should combine the best of what each of the modalities can offer. As dynamic tactile devices become available, raised-dot patterns can be used for shape recognition, while sound can be used for navigation in the virtual space (providing directionality and proximity information) and object/material recognition.

In future work, we also plan to explore the use of multiple fingers, which will increase the perceptual field [41], [58], as well as the use of multiple simultaneous sounds to convey information about neighboring or occluded objects. The interface can also be augmented with GPS, accelerometer, camera, and GIS enabled maps. The proposed approach is expected to have a significant effect on map perception, imaging, and navigation.

VI. Acknowledgments

The authors would like to thank all the empirical study participants. They also want to thank Dr. Karen Gourgey and Prof. Vivien Tartter for discussions on the design of experiments with visually-blocked participants, Prof. Jack Tumblin for discussions on the design of the user interface, Prof. Bill Hartmann for discussions on acoustics, and Philbert Lin for generating the tactile overlays.

References

[1] J. M. Loomis, R. L. Klatzky, and N. A. Giudice, "Sensory substitution of vision: Importance of perceptual and cognitive processing," in *Assistive Technology for Blindness and Low Vision*, R. Manduchi and S. Kurniawan, Eds. Boca Raton, FL: CRC Press, 2012, pp. 162–191.

[2] W. H. Dobelle, M. G. Mladejovsky, and J. P. Girvin, "Artificial vision for the blind: Electrical stimulation of visual cortex offers hope for a functional prosthesis," *Science*, vol. 183, no. 4123, pp. 440–444, 1974.

[3] G. J. Chader, J. Weiland, and M. S. Humayun, "Artificial vision: needs, functioning, and testing of a retinal electronic prosthesis," in *Neurotherapy: Progress in Restorative Neuroscience and Neurology*, ser. Progress in Brain Research, J. Verhaagen, E. M. Hol, I. Huitenga, J. Wijnholds, A. B. Bergen, G. J. Boer, and D. F. Swaab, Eds. Elsevier, 2009, vol. 175, pp. 317–332.

[4] P. Bach-y-Rita and K. A. Kaczmarek, "Tongue placed tactile output device," US Patent # 6430450, Aug. 2002.

[5] K. Gourgey, "Personal communication," Jan. 2006.

[6] P. Bach-y-Rita, *Brain Mechanisms in Sensory Substitution*. New York: Academic Press, 1972, ch. 1 and 2, pp. 1–37.

[7] H. Kajomoto, Y. Kanno, and S. Tachi, "Dorehead electro-tactile display for vision substitution," in *Proc. Int. Conf. Eurohaptics*, 2006, pp. 75–79.

[8] S. O'Modhrain, N. A. Giudice, J. A. Gardner, and G. E. Legge, "Designing media for visually-impaired users of refreshable touch displays: Possibilities and pitfalls," *IEEE Trans. Haptics*, vol. 8, no. 3, pp. 248–257, Jul-Sep 2015.

[9] R. L. Klatzky, N. A. Giudice, C. Bennett, and J. M. Loomis, "Touch-screen technology for the dynamic display of 2d spatial information without vision: Promise and progress," *Multisensory Research*, vol. 27, no. 5-6, pp. 359–378, 2014.

[10] S. Landau and L. Wells, "Merging tactile sensory input and audio data by means of the Talking Tactile Tablet," in *EuroHaptics*. IEEE Computer Society, 2003, pp. 414–418.

[11] O. Bau, I. Poupyrev, A. Israr, and C. Harrison, "TeslaTouch: Electrovibration for touch surfaces," in *Proc. 23nd Annual ACM Symp. User Interface Software and Technology (UIST'10)*. New York, NY: ACM, 2010, pp. 283–292.

[12] C. Xu, A. Israr, I. Poupyrev, O. Bau, and C. Harrison, "Tactile display for the visually impaired using TeslaTouch," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. ACM, May 2011, pp. 317–322.

[13] A. Israr, O. Bau, S.-C. Kim, and I. Poupyrev, "Tactile feedback on flat surfaces for the visually," in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. ACM, May 2012, pp. 1571–1576.

[14] L. E. Winfield, J. Glassmire, J. E. Colgate, and M. A. Peshkin, "T-PaD: Tactile pattern display through variable friction reduction," in *Proc. Second Joint Eurohaptics Conf. and Symp. on Haptic Interfaces for Virtual Environment and Teleoperator Systems.*, Tsukuba, Japan, Mar. 2007.

[15] E. C. Chubb, J. E. Colgate, and M. A. Peshkin, "Shiverpad: A device capable of controlling the shear force on a bare finger," in *World Haptics Conference*, Salt Lake City, Utah, Mar. 2009, pp. 18–23.

[16] N. D. Marchuk, J. E. Colgate, and M. Peshkin, "Friction measurements on a large area TPaD," in *IEEE Haptics Symposium*, Waltham, MA, Mar. 2010, pp. 317–320.

[17] B. Poppinga, C. Magnusson, M. Pielot, and K. Rassmus-Gröhn, "TouchOver Map: Audio-tactile exploration of interactive maps," in *Proc. 12th Int. Conf. Human Comp. Interaction with Mobile Devices*, ser. MobileHCI'11. ACM, 2011, pp. 545–550.

[18] N. A. Giudice, H. Palani, E. Brenner, and K. M. Kramer, "Learning non-visual graphical information using a touch-based vibro-audio interface," in *Proc. 14th Annual. ACM SIGACCESS Conf. Comp. Access. (ASSETS)*. ACM, Oct. 2012, pp. 103–110.

[19] H. P. Palani and N. A. Giudice, "Evaluation of non-visual panning operations using touch-screen devices," in *Proc. 16th Annual ACM SIGACCESS Conf. Comp. Access. (ASSETS)*. Rochester, NY: ACM, Oct. 2014, pp. 293–294.

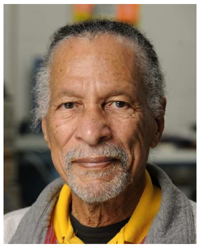
[20] J. M. Loomis, "Tactile pattern perception," *Perception*, vol. 10, pp. 5–27, Feb. 1981.

[21] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.
[22] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Process.*, vol. SP-40, no. 4, pp. 901–914, Apr. 1992.
[23] J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz, "Adaptive perceptual color-texture image segmentation," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1524–1536, Oct. 2005.
[24] K. van den Doel, "SoundView: sensing color images by kinesthetic audio," in *Int. Conf. Auditory Display*. IEEE, July 2003, pp. 303–306.
[25] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 2, pp. 112–121, Feb. 1992.
[26] A. F. R. Hernández, C. M. Gracia, O. C. González, C. M. Pascual, M. Ángel Torres Gil, R. Montserrat, E. B. Gutiérrez, and J. L. González-Mora, "Computer solutions on sensory substitution for sensory disabled people," in *Proc. 8th WSEAS Int. Conf. Comp. Intelligence, Man-machine Systems and Cybernetics (CIMMACS)*, 2009, pp. 134–138.
[27] S. E. Hernandez and K. E. Barner, "Tactile imaging using watershed-based image segmentation," in *ASSETS'00*, Arlington, VA, Nov. 2000, pp. 26–33.
[28] A. Nayak and K. E. Barner, "Optimal halftoning for tactile imaging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 12, no. 2, pp. 216–227, June 2004.
[29] T. C. Aysal and K. E. Barner, "Stochastic and deterministic models for haptic pseudo-textures," in *Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS'06)*, Alexandria, VA, Mar. 2006, pp. 469–476.
[30] J. Su, A. Rosenzweig, A. Goel, E. de Lara, and K. N. Truong, "Timbremap: Enabling the visually-impaired to use maps on touch-enabled devices," in *Proc. 12th Int. Conf. Human Computer Interaction with Mobile Devices*, ser. MobileHCI'10. ACL, 2010, pp. 17–26.
[31] R. D. Jacobson, "Navigating maps with little or no sight: An audio-tactile approach," in *Workshop on Content Visualization and Intermedia Representations (CVIR)*, Montreal, Quebec, Canada, 1998, pp. 95–102.
[32] P. Parente and G. Bishop, "BATS: the blind audio tactile mapping system," in *ACM South Eastern Conference (ACMSE)*, Mar. 2003, pp. 11–16.
[33] D. Parkes, "NOMAD - an audio tactile tool for the acquisition, use and management of spatially distributed information by partially sighted and blind people," in *Proceedings Second Int. Conf. Maps Graphics for Visually Disabled People*, 1988, pp. 24–29.
[34] P. Blenkhorn and D. G. Evans, "A system for reading and producing talking tactile maps," in *Proceedings 9th Int. Conf. Technology and Persons with Disabilities*, 1994, pp. 24–29.
[35] J. M. Loomis, R. G. Golledge, R. L. Klatzky, J. M. Speigle, and J. Tietz, "Personal guidance system for visually impaired," in *Proc. First Annual ACM Conf. Assistive Technol. (ASSETS)*. Marina del Rey, CA: ACM Press, 1994, pp. 85–91.
[36] J. M. Loomis, J. R. Marston, R. G. Golledge, and R. L. Klatzky, "Personal guidance system for people with visual impairment: A comparison of spatial displays for route guidance," *J. Vis. Impair. Blind*, vol. 99, no. 4, pp. 219–232, 2005.
[37] F. Ribeiro, D. Florencio, P. Chou, and Z. Zhang, "Auditory augmented reality: Object sonification for the visually impaired," in *IEEE 14th Int. Workshop Multimedia Signal Proc. (MMSP)*, Sept. 2012, pp. 319–324.
[38] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.
[39] K. V. D. Doel, D. Smilek, A. Bodnar, C. Chita, R. Corbett, D. Nekrasovski, and J. McGrenere, "Geometric shape detection with soundview," in *Int. Conf. on Auditory Display*, 2004, pp. 1–8.
[40] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 498–504, 2003.
[41] M. W. A. Wijntjes, T. van Lienen, I. M. Verstijnen, and A. M. L. Kappers, "Look what I have felt: Unidentified haptic line drawings are identified after sketching," *Acta Psychologica*, vol. 128, no. 2, pp. 255–263, June 2008.
[42] J. M. Loomis, R. L. Klatzky, and S. J. Lederman, "Similarity of tactual and visual picture recognition with limited field of view," *Perception*, vol. 20, no. 2, pp. 167–177, 1991.
[43] L. J. Thompson, E. P. Chronicle, and A. Collins, "The role of pictorial convention in haptic picture perception," *Perception*, vol. 32, pp. 887–893, 2003.
[44] M. W. A. Wijntjes, T. van Lienen, I. M. Verstijnen, and A. M. L. Kappers, "The influence of picture size on recognition and exploratory behaviour in raised-line drawings," *Perception*, vol. 37, pp. 602–614, 2008.
[45] M. K. Raja, "The development and validation of a new smartphone based non-visual spatial interface for learning indoor layouts," Master's thesis, University of Maine, Orono, ME USA, Dec. 2011.
[46] E. Zwicker, "Audibility limits for amplitude and frequency modulation of a tone," *Acustica*, vol. 2, pp. AB125–AB133, 1952, in German, available in English as NASA Report No. NASA TT F-15, 597.
[47] R. J. Reedstrom, "Tremolo effect," (date last viewed 04/01/2013). [Online]. Available: http://cnx.org/content/m15497/latest/lfmod_tremolo-eqn.html
[48] A. Ihlefeld and B. G. Shinn-Cunningham, "Effect of source spectrum on sound localization in an everyday reverberant room," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 324–333, 2011.
[49] V. C. Tartter, "Personal communication," Mar. 2008.
[50] T. V. Tran, T. Letowski, and K. S. Abouchacra, "Evaluation of acoustic beacon characteristics for navigation tasks," *Ergonomics*, vol. 43, pp. 807–827, 2000.
[51] H. G. Fisher and S. J. Freedman, "The role of the pinna in auditory localization," *Journal of Auditory research*, 1968.
[52] R. M. Lambert, "Dynamic theory of sound-source localization," *The Journal of the Acoustical Society of America*, vol. 56, no. 1, pp. 165–171, 1974.
[53] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, ii: The effects of a single reflecting surface," *The Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 524–533, Aug. 1985.
[54] W. J. Steinberg, *Statistics Alive!*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2011.
[55] D. Downing and J. Clark, *Business Statistics*, 5th ed. Barron's Educational Series, 2010.
[56] A. F. J. Sanders and A. M. L. Kappers, "Haptically straight lines," *Perception*, vol. 36, no. 11, pp. 1682–1697, 2007.
[57] S. J. Lederman and L. A. Jones, "Tactile and haptic illusions," *IEEE Trans. Haptics*, vol. 4, no. 4, pp. 273–294, oct-dec 2011.
[58] R. L. Klatzky, J. M. Loomis, S. J. Lederman, H. Wake, and N. Fujita, "Haptic identification of objects and their depictions," *Perception and Psychophysics*, vol. 54, pp. 170–178, Aug. 1993.

Pubudu Madhawa Silva (M'06) received the B.Sc (Hons) degree in electronics and telecommunication engineering from University of Moratuwa, Colombo in 2006, and the M.S. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, Evanston IL in 2010 and 2014, respectively. He joined Intel Corporation, Hillsboro, OR in 2014 as a senior software engineer where he works on image processing, computer vision and machine/deep learning and heterogeneous computing. His research interests are in tactile and multimodal interfaces, sensory substitution, perceptual models for multimedia processing, virtual/augmented reality and machine/deep learning.

Thrasyvoulos N. Pappas (M'87, SM'95, F'06) is Professor of Electrical and Computer Engineering and Computer Science at at Northwestern University. From 1987 until 1999, he was a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. Prof. Pappas is a Fellow of SPIE and IS&T. He is Vice President-Publications for the Signal Processing Society (SPS) of IEEE and co-chair of the Human Vision and Electronic Imaging Conference. His research interests are in image analysis, quality, and compression, content-based retrieval, model-based halftoning, and tactile and multimodal interfaces.

Joshua Atkins (M'03) received the M.S. and Ph.D. degrees in Electrical Engineering from Johns Hopkins University, Baltimore, Maryland, in 2006 and 2011, respectively. From 2011 to 2014 he was Senior Research Engineer and then Signal Processing Research Manager at Beats Electronics, Santa Monica, CA, working on speech and audio algorithms for headphones, speakers, and mobile devices. Since 2014, he has led a signal processing team at Apple, Culver City, CA, focused on spatial audio and multichannel speech processing algorithms. His current research interests include spatial audio capture and reproduction, multichannel speech enhancement for telephony and speech recognition applications, and applications of machine learning to speech and audio processing.



James E. West (IEEE Life Fellow) is Professor of Electrical and Computer Engineering at Johns Hopkins University. From 1957 to 2001 he was at Bell Laboratories, Murray Hill, NJ, where he was named Bell Laboratories Fellow, the organization's highest honor. Prof. West is member of the Communications Society of IEEE. He has authored numerous journal and conference proceedings papers, and holds over 250 patents. He is a member of the National Academy of Engineering, Fellow of the Acoustical Society of America, and was inducted into the Inventors Hall of Fame in 1999. Dr. West received the 2006 National Medal of Technology and Innovation, and in 2010, the Benjamin Franklin Medal in Electrical Engineering. His research interests are in transducers, acoustics, and medical devices.