Human Vision and Electronic Imaging XVII, B. E. Rogowitz, T. N. Pappas, and H. de Ridder, eds., Proc. SPIE Vol. 8291, (San Francisco, CA), Jan. 2012.

# Acoustic-Tactile Rendering of Visual Information

Pubudu Madhawa Silva[1], Thrasyvoulos N. Pappas[1], Joshua Atkins[2], James E. West[2] and William M. Hartmann[3]

[1]EECS Department, Northwestern University, Evanston, IL 60201, USA
[2]ECE Department, Johns Hopkins University Baltimore, MD 21218, USA
[3]Physics and Astronomy Department, Michigan State University, East Lansing, MI 48824 USA

## ABSTRACT

In previous work, we have proposed a dynamic, interactive system for conveying visual information via hearing and touch. The system is implemented with a touch screen that allows the user to interrogate a two-dimensional (2-D) object layout by active finger scanning while listening to spatialized auditory feedback. Sound is used as the primary source of information for object localization and identification, while touch is used both for pointing and for kinesthetic feedback. Our previous work considered shape and size perception of simple objects via hearing and touch. The focus of this paper is on the perception of a 2-D layout of simple objects with identical size and shape. We consider the selection and rendition of sounds for object identification and localization. We rely on the head-related transfer function for rendering sound directionality, and consider variations of sound intensity and tempo as two alternative approaches for rendering proximity. Subjective experiments with visually-blocked subjects are used to evaluate the effectiveness of the proposed approaches. Our results indicate that intensity outperforms tempo as a proximity cue, and that the overall system for conveying a 2-D layout is quite promising.

**Keywords:** visual substitution, visually impaired, HRTF, virtual reality, immersive environments

## 1. INTRODUCTION

As electronic media dominate people's daily lives, it is important that the information they provide, which is to a great extent in graphical and pictorial form, be made available to the visually impaired (VI) segment of the population. Raised-line drawings and other tactile patterns have already been utilized, alongside Braille text, to convey graphical and pictorial information in textbooks, maps, and other documents.[1] However, such information is mostly static, and cannot capture the dynamic nature of the information that is made available by electronic media, especially via the Internet. This requires (a) dynamic tactile and acoustic display technologies and (b) the ability to translate visual information into tactile and acoustic form. Another important consideration is the ability to actively explore this information, as the VI typically do using the fingertips to read Braille or a long cane to explore their immediate surroundings. Our focus is on noninvasive approaches, that is, without prosthetic devices that require surgery to restore some degree of functional vision, as for example with cortical or retinal electrode matrix displays.[2] In addition, some noninvasive approaches for visual substitution (VS) are still quite objectionable to the VI because they involve the presentation of electrical and other tactile stimuli on different parts of the body like the back, abdomen, tongue, or forehead.[3, 4] For example, the tongue display consists of an array of electrodes that applies voltages to stimulate the tongue.[5] While such displays can be quite effective for certain visual tasks, a large part of the VI find them quite "invasive," and prefer to scan/explore with the finger.[6] In addition, finger scanning provides valuable kinesthetic feedback.

The focus of our work has been on the dynamic display of visual information via hearing and touch. In Ref. 7, we proposed the use of a touch screen with acoustic feedback for the presentation of simple visual stimuli, such as those shown in Figure 1, whereby the user actively explores the information with the finger while listening to acoustic feedback over headphones. Tactile patterns could also be added, *e.g.,* embossed on Braille paper overlaid on the touch screen (as in Ref. 8), but the availability of devices for dynamic display of tactile patterns is still quite limited. For example, the device described in Ref. 9 controls the surface friction of a rigid glass

Further author information: Send correspondence to T.N.P.: E-mail: pappas@eecs.northwestern.edu

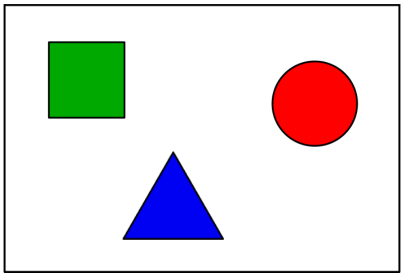


Figure 1. Conveying a simplified visual scene

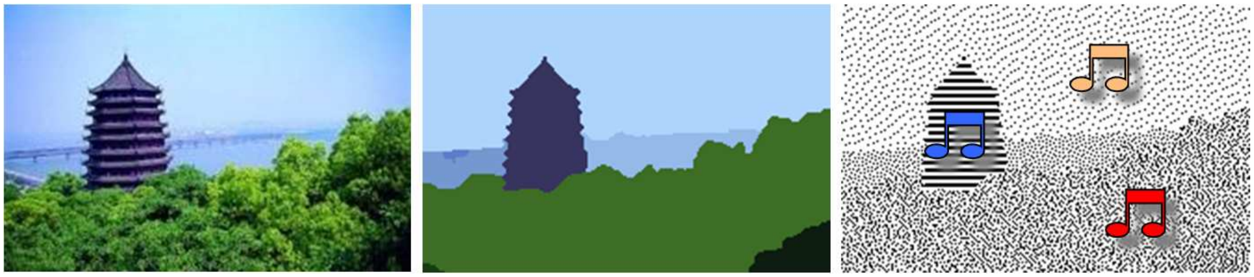


Figure 2. Conveying complex visual scene by acoustic and tactile textures

surface. Other devices have limited capabilities and are quite costly.[10–12] Our ultimate goal is to present a map, diagram, or picture as a collection of segments with perceptually distinct acoustic-tactile textures, as shown in Figure 2.[13]

In Ref. 7, our focus was on conveying object shape (*e.g.,* circle, triangle, square) using a touch screen and acoustic feedback. We demonstrated the importance of directional sound for object localization and boundary tracing. We also considered the rendering of a simple scene layout (a few objects in a linear arrangement, each with a distinct tapping sound), the exploration of which with the finger on a touch screen we compared to a "virtual cane." The focus of this paper is on the sequential exploration of a two-dimensional (2-D) layout of objects and scene perception, in the sense that the subject can identify and localize a number of objects in the scene.

A number of noninvasive visual substitution approaches have been proposed in the literature.[14–19] A more detailed discussion of some of these approaches can be found in Refs. 7 and 20. An important distinction is between approaches that directly translate visual information (image intensities) to an acoustic or tactile signal,[5, 14, 16, 21, 22] and those that utilize a semantic representation of the scene.[8, 13, 23–25] The advantage of the former is that they bypass semantic analysis (*e.g.,* scene segmentation, object identification, etc.), which is a challenging and essentially open research problem. Of course, in many instances (*e.g.,* graphs, maps, building layouts), the semantic information is readily available. The advantage of the latter is that the visual to tactile-acoustic mapping can be more intuitive and can make better use of the relative abilities of hearing and touch compared to vision. It is precisely the understanding of such abilities that is the focus and interest of our research. We will thus, assume that semantic analysis of the visual information is available, and will explore the extent to which it can be displayed in acoustic-tactile form. This, of course, involves not only the perception of size, shape, and relative position of different objects or segments, but also memory, and the task associated with the scene exploration. Since, the shape perception was considered in Ref. 7, this paper will focus on object identification, localization, and perception of a 2-D layout.

While the proposed techniques are aimed at enhancing the ability of VI people to access visual information, they are also expected to be important in instances where sighted people cannot make use of their vision, for example, for GPS navigation while driving, fire-fighter operations in thick smoke, and military missions conducted under the cover of darkness. Our research results also have applications in other disciplines like virtual reality, augmented reality, immersive environments, e-Commerce, and tele-medicine among others.

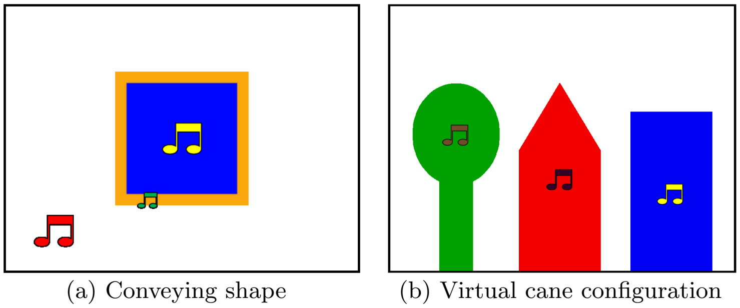

(a) Conveying shape (b) Virtual cane configuration

Figure 3. Configurations used in Ref. 7

The paper is organized as follows. In Section 2 we discuss the basic system setup. The sound design and implementation is presented in Section 3. In Section 4 we discuss our subjective experiments and analyze their results. The concluding remarks are presented in Section 5.

## 2. PROBLEM SETUP AND PREVIOUS WORK

Let us consider the steps involved in conveying a complex scene, such as the one depicted at the left of Figure 2. The task of perceiving visual information can be divided in to two main sub-tasks, the "what" and the "where."[26,27] The "what" involves identification of the different objects/segments in the scene, for example by their shape, size, color, material, sound, *etc.* The "where" involves the determination of the object/segment position in the scene.[26,27] Of course, there are also interactions between the two, for example the perception of relative sizes of two objects depends on their sizes on the retina and their locations relative to the user. This "what" and "where" is in fact common for three modalities under consideration (visual, acoustic, and tactile),[28,29] as well as for olfactory. For a given scene layout, and assuming the availability of a segmentation and associated semantic information, our goal is to examine the best way to convey this information using a combination of the acoustic and tactile modalities. An additional constraint is to accomplish this with existing devices. Thus, due to the lack of a proper dynamic display to render tactile textures, we have opted for a touch screen with acoustic feedback, whereby touch is used as the pointing device, and sound is used for object localization and identification, as we describe below.

Note that while the acoustic signals are used to provide localization feedback, the use of the finger as a pointing device on the touch screen provides kinesthetic feedback, which facilitates scene perception. Initial indications from existing research are that it is possible to perceive scene layout using sound. In particular, Sanchez *et al.*[30] found that with the use of spatialized acoustic stimuli, blind children can construct spatial imagery.

In Ref. 7, the focus was on shape perception. Figure 3(a) shows a simple shape. Two sounds were used to convey each shape: one for the background and one for the interior of the object. We found that including a third sound for the object boundary makes it easier for the subjects to trace the boundary and identify the shape. We also found that using directional sound to guide the subject to the object and along the boundary provided the overall best performance. We also considered a simple layout of a few objects in a linear arrangement, each with a distinct tapping sound, as shown in Figure 3(b); the background was silent. We referred to this configuration as a "virtual cane."

The focus of this paper is on identification and localization of objects in a 2-D scene. Since the perception of size and shape was considered in Ref. 7, we assume that all objects have the same shape and size, and rely on the characteristic sound for identification. An example of a 2-D layout is shown in Figure 4(a). In vision, scene exploration is based on saccadic eye movements, that is, exploration is done sequentially with the order depending on a nontrivial combination of top-down and bottom-up processes and a combination of foveal and peripheral vision.[31–33] The corresponding task in the acoustic-tactile case could be similar, perhaps with the index finger playing the role of the fovea and the other fingers playing the role of periphery. In the "virtual cane" configuration of Ref. 7, we asked the subjects to use their index finger to explore the 1-D layout (Figure 3(b)).

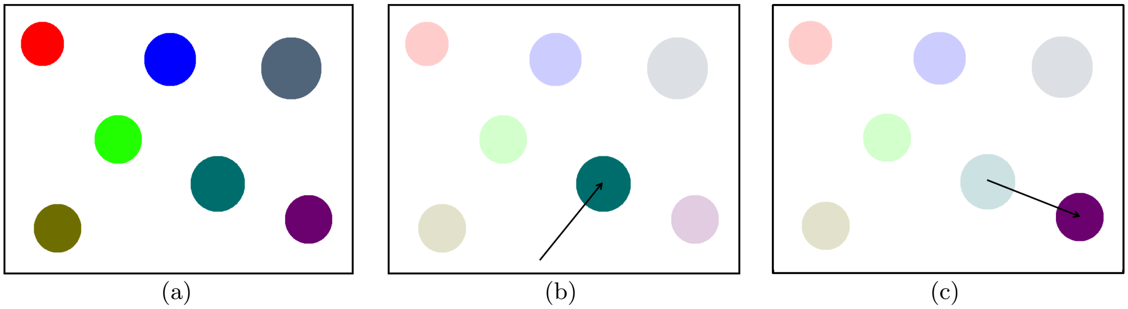


Figure 4. Conveying a simplified two-dimensional layout

Since the 2-D layout can be considerably more complicated, rather than completely unrestricted exploration of the 2-D layout, we opted for sequential, guided, exploration of one object at a time, using the index finger. As shown in Figure 4(b), the finger may be guided (through the use of directional sound and proximity clues) to one object. Once it reaches that object, it is then directed to the next object, as shown in Figure 4(c), and so on. The order could be random or deterministic, for example, each time the subject explores the object closest to the finger. To make sure that all objects are visited, the exploration can be organized in cycles, with each object visited once during each cycle. The subject is also allowed to skip objects whose position she/he is already familiar with.

In our system setup, which we used for the experiments we describe in Section 4, the user explores the layout one object at a time, starting with the object closest to the finger on the screen. Once the finger reaches this object, the object is marked as "inactive" and the user is directed to the next closest object (from the current position of the finger). The process is repeated until all the objects have been explored, *i.e.,* they have become inactive. At this point the cycle is complete, and all the objects are activated again for the next cycle. As we saw above, the user may skip an object that she/he is familiar with and go on to the next object. The signaling between the user and the system can be done via voice commands (by the system and the user) and tapping on the screen and finger gestures (by the user).

The basic elements of the layout exploration are the following:

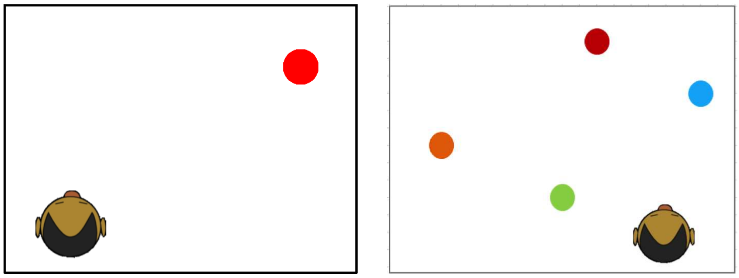
- Each object is characterized by a specific sound.
- The subject explores the scene via the index finger on the screen, one object at a time.
- Directional sound with proximity cues is used to direct the subject (finger) to each object.

The choice of sounds, and details of the scene exploration are discussed in the next sections. Some of the questions we want to address are the following:

- How fast can the subject locate an object at a random position on the screen?
- What is the most natural/intuitive way to guide the finger to an object? It is important that the cognitive load for this task is low, so that it can be conserved for the scene perception.
- How well can the subject perceive a scene layout of a few objects?

To address these questions, in Section 4, we will describe two subjective experiments.

In our system setup, the observer is assumed to be at the position of the finger on the touch screen. To render sound directionality, we assume that the observer is always facing towards the top of the touch screen, as shown in Figure 5. Of course, it is possible to indicate changes in observer orientation, *e.g.,* by changing the finger orientation, but for the purposes of this paper, we will assume constant observer orientation. In sequential object exploration, the observer hears the sound coming from the center of one object plus some sound due to the background. Adding the sounds from other objects is possible, but may interfere with localization, plus there is a limit on the number of objects that a human observer can keep track of.



(a) Experiment 1 (b) Experiment 2

Figure 5. Experimental setup: Observer is assumed to be at the position of the finger, facing towards the top of the touch screen.

## 3. SOUND DESIGN AND IMPLEMENTATION

Since we rely on the acoustic modality as the primary source of feedback to the user, sound selection and sound rendering techniques are of great importance to the overall success of the system. In sequential object exploration, *i.e.,* one object presented at a time, the touch screen is divided into two segments, one inside the object and one in the background. Thus, for each object, we have to select sounds for each of the segments. The selections should facilitate both object identification and localization, the two main tasks of the proposed system. The former is required for both segments, while the latter is only needed in the background.

Object identification can be based on either spectral or temporal cues. For example, as we saw in Ref. 7, humans can easily identify objects by characteristic material sounds. Alternatively, one can use arbitrary sounds that can be assigned to each object, provided they are easy to discriminate/remember. In the present study, we opted for basic sounds for object identification; we used simple sinusoidal tones of different frequency for each object. This is because pitch is naturally a strong identifier due to the fundamental tonotopic organization of the auditory system.[34] To make the tones more distinguishable, we selected tones in different critical bands. In addition, it is important to select tones in the frequency range of 400–1000 Hz, where they can be well localized using ITD, the most powerful localization cue in free field.[35,36] As we will see below, the addition of white noise also provides strong localization. According to the equivalent rectangular band measures by Glasberg and Moore,[37] this frequency range has six critical bands; we can thus have six distinguishable sounds in this range.

As we mentioned above, in the background region, we need both localization – to direct the subject to the object – and identification – to let the subject know what object she/he is trying to get to. According to Tran *et al.*,[38] the acoustic signals that are both pleasant and easily localizable are non-speech broadband sounds with proper balance between low and high frequencies. As Blauert[39] points out, broadband noise provides even stronger localization cues. As we will see below, it also provides some additional advantages for sound localization. By adding a tone to broadband noise, we can then accomplish both goals, localization and identification, *i.e.,* there is a different background sound for each object we are directing the subject to, while the noise makes it clear that the subject is in the background.

Given these selections for the sound signals, we now turn to sound rendering to achieve the localization goals. For object localization we need directionality (angle from which the sound is coming from) and proximity (the distance between the listener and the source) cues. In general, the goal is to present an acoustical environment that consists of one or more sound sources located in a 2-D plane, to a listener who is also located in the same plane. However, in this paper our focus is on one source. We are assuming binaural presentation over headphones. To do this accurately, we need to know the location of the acoustic source and how the sound is modified/diffracted by the listener's head, ears, and torso. The effects of each person's head, ears, and torso can be represented by a *head-related transfer function (HRTF)*, which can be measured in a controlled lab environment or simulated using each person's anatomical data,[40,41] which again requires accurate measurements. In both cases, HRTF estimation is time consuming and requires specialized laboratory settings and equipment. In addition, at sound rendering time, it is costly to calculate the HRTF for points that are not in the measurement grid, as this involves interpolation of the measured HRTF or recalculation of the simulated one. A much simpler approximation of the transfer function is provided by a spherical head model, which predicts *interaural time differences (ITDs)* and *interaural level differences (ILDs)* by approximating the human head as a perfect sphere and ignoring torso effects. This model is capable of rendering directionality but does not sound as natural because the rendered sounds are perceived as if they occur inside the head. The HRTF is a much stronger model for rendering spatialized sound. In addition to externalization (sounds are perceived as coming from the environment), it can provide better localization (directionality and proximity). However, using the HRTF comes at an increased computational complexity that makes it impractical for real-time binaural rendering of many simultaneous sources on mobile computing platforms. On the other hand, Atkins proposed a method[42, 43] that significantly reduces the required computation by representing the HRTF using a basis of spherical harmonics[44–46] that allow the interpolation to be solved offline. Since the HRTF varies from person to person, one approach is to use the HRTF measurements of a KEMAR mannequin, which is based on an "average" human head, torso, and pinna, and which is available in the CIPIC database.[47] This provides a reasonable compromise in directional rendering accuracy, but the CIPIC database does not provide data for rendering proximity information. (It relies only on data points on one circle.) Most importantly, even in a natural environment, human judgment of proximity information is poor,[48, 49] so we have to resort to other means for rendering proximity information. The question of the relative merits of acoustic rendering using HRTF measurement or the spherical head model remains open. In the experiments described in this paper, we used the CIPIC measurements.

One of the key issues is synchronization between acoustic rendering and tactile sensing of the finger position. This is particularly important when one uses portable devices with limited memory and computational power, like the *Apple iPad 1,* which we used for system implementation. Thus, even with the simplifications of Ref. 42, it is not possible to have real-time rendering of directionality via HRTF in an *iPad.* In order to minimize the delay between sensed finger position and sound rendering, we had to quantize the directionality to 30° sections and to preload the sound files for each section.

We now turn to rendering of proximity information, *i.e.,* the distance between the source and listener. In real world conditions, changes in proximity cause several significant changes in the acoustical signal that reaches the listener's ears, thus providing strong proximity cues. There are four main acoustic distance cues for stationary sources and listeners: intensity, direct-to-reverberant energy ratio, spectrum distortions, and binaural differences.[48] For example, in ideal conditions (point source and acoustic free field), the variation of sound intensity with distance follows an inverse-square law.[48] However, the exact relationship between proximity and each of these acoustic cues is largely dependent on the environment and properties of the sound source.[49] Overall, due to the complex and interdependent relationship between physical parameters and perceived acoustic proximity, it is difficult to effectively model human perception of acoustic proximity. More importantly, as we mentioned above, human subjects are proven to be consistently inaccurate in sound proximity judgments.[48, 49] In contrast to proximity, human perception of sound directionality is much more accurate. We thus steer away from realistic models of sound proximity perception, and instead, rely on intuitive conventions for indicating changes in proximity. For relative (because the screen is not the real environment) distances between the objects in the layout, we rely on kinesthetic feedback from the moving finger on the touch screen.

Among several possible conventions for indicating changes in proximity we tried two: intensity and tempo. Note that, since this is a virtual world, there is no need to follow physical laws for signal variations. Instead, we can focus on the most intuitive conventions or those that are the easiest to learn. We describe intensity variation first. Since we have a tone (signal) and noise in background, we found it more intuitive to indicate changes in proximity by changes in the signal-to-noise ratio, that is, as the finger approaches the object, the tone intensity increases and the noise intensity decreases. Ideally, intensity changes should be based on their effect on loudness, but because of the crude playback volume controls of the *iPad* as well as interactions between the tone and noise signals, we could not establish a relationship between loudness and intensity. Instead, we worked directly with the volume controls on the *iPad.* In our implementation, we changed the playback volume instantaneously depending on the position of the finger relative to the object. We tried several functions for varying the volume of the tone and noise with distance, and selected the functions shown in Figure 6. Note that the volume of the white noise decreases rapidly when the finger (subject) is far from the object and slower as it approaches the object, while the volume of the tone increases slowly when the finger is far from the object and more rapidly as it approaches the object. Note that a step in noise volume is necessary to indicate that the finger has entered the object. When the finger is inside the object, the subject only hears a constant, non-directional tone.

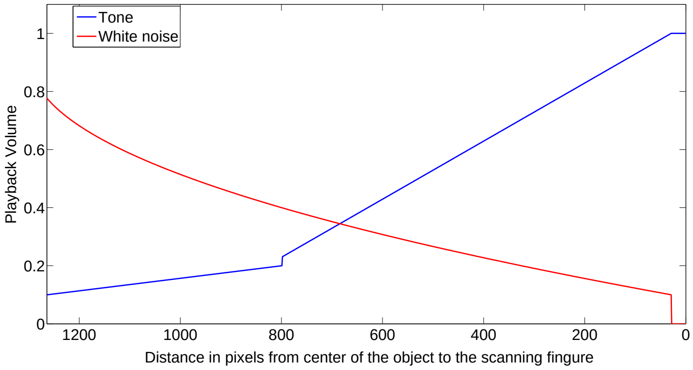


Figure 6. Playback volume variation of tone and white noise with distance from center of the object to the scanning finger

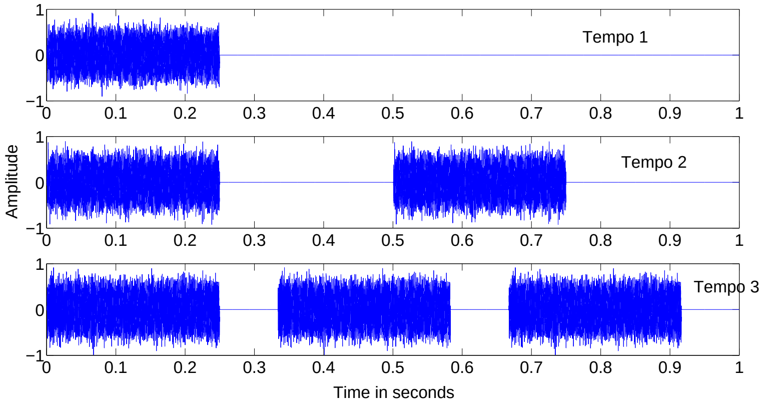


Figure 7. Playback signals (tone and white noise) for the three tempos used in our experiments

We now describe the tempo variation in the background region as the finger approaches the object. In this configuration, we kept the volume of the tone and the noise constant and varied the tempo. We created the tempo by alternating segments of signal and silence. To vary the tempo, we fixed the length of the signal segment and changed the length of the silence segment, as shown in Figure 7. While such tempo variations are not naturally experienced in acoustic proximity judgments, it is easy for subjects to learn to associate the tempo changes with changes in proximity. Moreover, this is analogous to SONAR, and has also been used in other interfaces, *e.g.,* in games. Ideally, we would like to have continuous tempo variation with distance, in direct analogy to the volume variations of the first approach. However, due to computation and memory limitations of the *iPad*, we had to quantize the tempo variations to three levels.

A final consideration is the importance of onset sounds, such as tapping or striking an object, for localization. This should be easy to understand, as any abrupt changes in sound intensity are easy to detect, and emphasize the interaural time differences (ITDs), thus providing strong localization cues.[50] By alternating periodic tone signals with segments of silence to create different tempos in the background region, we are adding strong onset effects, thus adding localization cues. On the other hand, inside the object, where there is no need for localization, we used a continuous tone sound. Due to the advantages of the onsets, we also included them in the intensity approach, picking a constant tempo.

## 4. SUBJECTIVE EXPERIMENTS

To address the questions we posed in Section 2, we conducted two kinds of subjective experiments, one to find the best acoustical cue for object localization (proximity) and one to evaluate the performance of the overall system, which was implemented with the best acoustic cue determined via the first experiment.

All subjective experiments were conducted in an average room with uncontrolled acoustic conditions. We used an *Apple iPad 1* as our touch screen and *Sennheiser HD595* headphones for the acoustic feedback. All subjects were sighted, with normal or corrected vision, and normal hearing. To make sure that they could not see the touch screen or their scanning finger during the experiments, a cardboard box was placed over the touch screen on the table in front of which the subjects were sitting. The subjects were able access the touch screen by inserting their hand through a small opening in the front of the box. This setup was used in order to avoid blindfolding, which could make participants uncomfortable and nervous. The subjects were given written instructions and a chance to ask questions prior to the start of each experiment. All subjects were volunteers and were not financially rewarded for their participation in the experiments. Each subject was asked to sign a consent form before taking part in the experiments.

## 4.1 Experiment 1: Acoustic Cues for Navigation to A Single Object

In the first experiment, we tested the subjects' ability to locate a single object on the touch screen. We tried two approaches for rendering proximity: varying the sound intensity and varying the tempo. The object layout for the experiment is shown in Figure 5(a). A single circular object was placed on the screen; the position was random and the size was fixed in all trials. Each trial started when subject put the finger on the touch screen in a random location. The task was to navigate to the object in the shortest possible time. Once the subject's finger reached the object, *i.e.,* the subject heard the constant tone inside the object boundaries, she/he should notify the system of end of the trial by double tapping anywhere on the screen, at which point the system acknowledged with the phrase "experiment is over." Note that the time for the trial stops only when the subject double taps, not when the finger enters the object. This is necessary to make sure that the subject is fully aware that she/he has located the object, rather than accidentally moving the finger over the object without realizing what happened. A new trial begins when the subject lifts the finger and places it back on the screen. If the subject does not lift the finger after double tapping, the system alerts her/him with the phrase "please lift your finger." When the subject lifts the finger, the word "ready" indicates that a new trial can be started. The next trial starts when the subject places the finger on the touch screen. In the written instructions the subjects are asked to double tap as soon as they locate the object, then to lift their finger and quickly put it down at a random position. The subjects were asked to conduct as many trials as they could until an announcement informing them that the experiment is over is played back over the headphones. The instruction to perform as many trials as they can within a given time provided additional motivation to locate the objects as fast as possible.

We tested two sound configurations. In the first, proximity was rendered via variations in sound intensity and in the second via variations in tempo. The details for each configuration were described in Section 3. In both cases, the sound directionality was rendered using a KEMAR HRTF with quantized directionality as described in Section 3. Each configuration was run for a ten-minute block of time comprising many trials. The configurations were evaluated based on the average or median navigation time over the trials presented in the final three minutes of the test. This was done to negate any learning effects. The system automatically kept track of the time for each trial, as well as the cumulative time for each run. The time between trials was not counted in the summation. The experiment ended when the summation was larger than 10 minutes, that is, each trial was carried to completion. All trials that started or ended within the final three minutes were counted in the calculation of the average and median navigation time.

Eight subjects, four male and four female, with ages between 23 and 35 years old took part in the experiments. At the beginning of each experiment, the subjects were given a trial run with conditions similar to the test conditions. During the trial runs, the subjects could see both the scanning finger and object in the touch screen. In both configurations, we used a 652 Hz sinusoidal tone to identify the object. In the variable tempo configuration, we used a constant tone and noise intensity, fixed the length of the signal segments at 250 ms, and changed the tempo by varying the length of the silence segments. We used three different lengths: 750, 250, and 83 ms. Thus, the periods were 1 s, 500 ms, and 333 ms (1, 2, and 3 Hz). In the variable intensity configuration, we used constant tempo with 250 ms signal segments and 250 ms silence segments. The intensity variations were as described in Section 3. In all cases, a constant tone (no silence) was played in the interior of the object.

The results are shown in Figure 8, which shows the distribution of the time that was required for each trial in the two configurations of Experiment 1. The overall mean (across all observers) for the intensity configuration

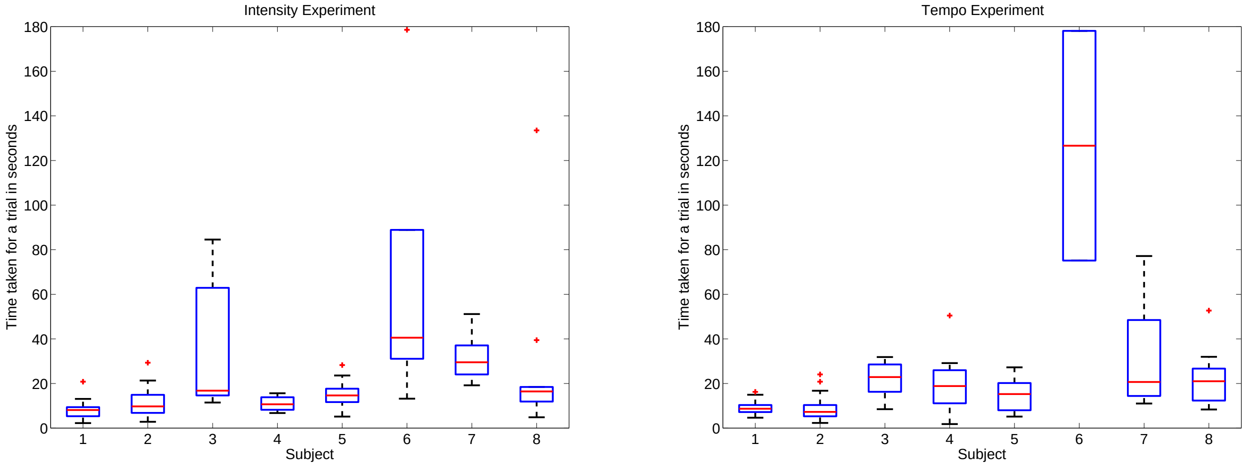


Figure 8. Results of Experiment 1: (left) Intensity; (right) Tempo

was 25.6 s and the standard deviation was 17.9 s. The corresponding numbers for the tempo configuration were 32.0 s and 36.5 s. Note that there are a few outliers, so the median is a better representative of the distribution than the mean. The medians for the intensity and tempo configurations were 15.6 s and 19.8 s, respectively.

According to both statistics, our results suggest that rendering proximity via intensity is more effective than via tempo, when the time to reach the object is used as the performance criterion. One explanation for this may be that, in our experimental setup, intensity variations were continuous while the tempo variations were quantized to just three levels. Another factor is the fact that the perception of intensity changes is instantaneous, while temporal cues like tempo are not. In addition, intensity is a natural auditory cue for source proximity. On the other hand, interviews with the subjects after the completion of the experiments indicate that tempo may provide a more comfortable interface that requires less concentration from the user, who may thus devote more attention to comprehending scene organization. Regarding the quantization of tempos to three levels, some of the subjects used it to their advantage, adjusting their scanning speeds based on the tempo, *i.e.,* their perceived distance from the object. (The further from the object the faster the scanning speed.)

There are several issues to be investigated with respect to navigation based on intensity. Most of the subjects complained that they could not perceive significant intensity changes in background sound for small finger movements and had to move a larger distance to feel a change in intensity. This appears to nullify the advantage of continuous variation. Another issue is the use of two sounds (tone and noise). Three of the subjects complained about having to monitor two sounds instead of one. Two other subjects liked the idea of having two sounds in the background, but suggested that we increase the volume of noise and decrease that of the tone as the finger reaches the object. Finally, the relationship between the intensity and loudness of the tone and noise signals is complicated by masking effects;[51] an in depth exploration of such effects is beyond the scope of this paper.

## 4.2 Experiment 2: Overall System Evaluation

In the second experiment, we tested the subjects' ability to perceive a 2-D layout of simple objects on the touch screen. Based on the results of the first experiment, we used sound volume to indicate changes in proximity. The proposed system was implemented with four circular objects as shown in Figure 5(b). As we discussed, each object was represented by a sinusoidal tone with white noise added in the background region. We assigned tones of 452, 652, 852, and 1052 Hz to the four objects. As described above, we used a constant tempo in the background region, with alternating segments of signal (tone and noise) and silence, each of 250 ms duration. The interior of the objects was represented with a continuous pure tone signal of the same frequency.

The subjects' task was to explore the layout to learn the locations of the four objects until they are confident that they can reproduce the scene on a sheet of paper. There were no time limitations for this experiment but the time taken was recorded. At the end of the experiment, subjects were first asked to say how many objects

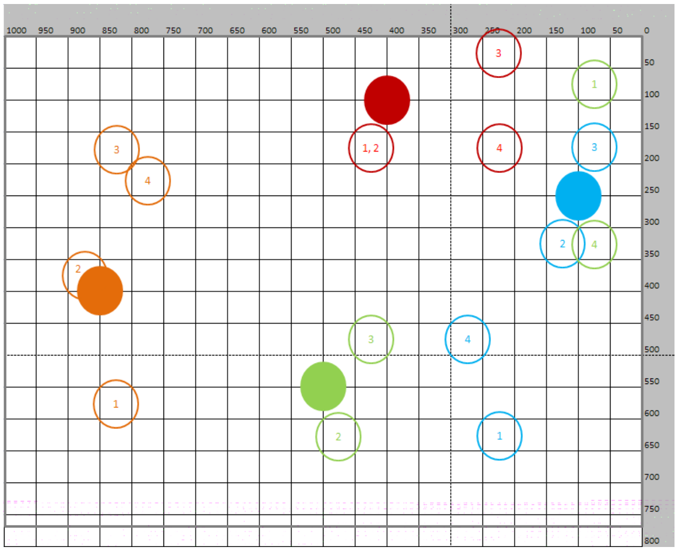


Figure 9. Results of Experiment 2

were in the layout. Then they were given a customized graph paper of the same size as the active area of the *iPad*. (At the *iPad* 768 × 1004 resolution, the grid spacing was 50 pixels.) To eliminate the additional burden of associating the tone of each object with a label, the subjects were given the ability (another *iPad* application) to play the tone for each object before they were asked to place it on the graph paper.

Four out of the eight subjects of the first experiment participated in this experiment; two were male and two female, with ages ranging from 23 and 35 years old. In order to familiarize the subjects with the system, before the start of the experiment, they were presented with a training example that consisted of a layout of three objects. During the training period, the subjects were able to see both their finger and the layout on the touch screen. The results are shown in Figure 9, which shows a composite drawing of object placement by the four subjects. The solid circles denote the actual object positions and the empty circles show the placement by the four subjects, marked by the subject number in the center of the circle. Note that two of the subjects confused the green and blue objects. This can be attributed to the fact that the objects were represented by adjacent frequencies (452 and 652 Hz). Aside from this, it appears that Subject 2 had considerably better placement accuracy than the other subjects, but overall the performance is reasonable, at least in the relative placement of the objects, if not in the accuracy of placement. The average time for scene exploration was seven minutes.

Overall, the second experiment is a composite task that involves the subject's ability (a) to locate objects, (b) to integrate relative object locations and kinesthetic feedback from finger movements into a perception of 2-D space, (c) to identify individual objects, and (d) to remember all that. The first of these tasks was considered in the first experiment. A detailed understanding of each of the other tasks is beyond the scope of this paper. The results of the second experiment provide strong encouragement that the whole concept is promising.

## 5. CONCLUSIONS

We have proposed a non-invasive system whereby subjects can actively explore a 2-D object layout by scanning a touch screen with the finger while listening to spatialized acoustic feedback. We discussed sound selection and its effect on identification and localization, the two main tasks of the system. We relied on HRTF to render sound directionality and considered variations of sound intensity and tempo as two alternative approaches for rendering proximity. Subjective experiments with visually-blocked subjects indicate that intensity outperforms tempo as a proximity cue, and that the overall system for conveying a 2-D layout is quite promising.

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under Grant No. IIS-1049001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and

do not necessarily reflect the views of the NSF.

## REFERENCES

[1] Ladner, R. E., *et al.,* "Automating tactile graphics translation," in [*Proc. 7th Int. ACM SIGACCESS Conf. on Computers and Accessibility*], *Assets '05*, 150–157, ACM, New York, NY, USA (2005).

[2] Meijer, P., "Sensory substitution - vision substitution," (Oct. 2010).

[3] Bach-y-Rita, P., [*Brain Mechanisms in Sensory Substitution*], Academic Press, New York (1972).

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

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

[6] Gourgey, K., "Personal communication."

[7] Silva, P. M., Pappas, T. N., Atkins, J., and West, J. E., "Perceiving graphical and pictorial information via touch and hearing," in [*Proc. Int. Conf. Acoustics, Speech, and Signal Processing*], 2292–2295 (May 2011).

[8] "Talking tactile tablet." http://www.touchgraphics.com/.

[9] Winfield, L. E., *et al.,* "T-pad: Tactile pattern display through variable friction reduction," in [*Proc. 2nd Joint Eurohaptics Conf. and Symp. Haptic Interfaces Virtual Environment and Teleoperator Systems.*], (Mar. 2007).

[10] Wall, S. A. and Brewster, S., "Sensory substitution using tactile pin arrays: Human factors, technology and applications," *Signal Processing* **86**, 3674–3695 (2006).

[11] Yang, T.-H., Kim, S.-Y., and Kwon, D.-S., "Applications of a miniature pin-array tactile module for a mobile device," in [*Int. Conf. Control, Automation and Systems*], 1301–1304 (Oct. 2008).

[12] Zhao, F., Fukuyama, K., and Sawada, H., "Compact braille display using SMA wire array," in [*18th IEEE Int. Symp. Robot and Human Interactive Communication*], 28–33 (Oct. 2009).

[13] Pappas, T. N., Tartter, V., Seward, A. G., Genzer, B., Gourgey, K., and Kretzschmar, I., "Perceptual dimensions for a dynamic tactile display," in [*Human Vision and Electronic Imaging XIV*], Rogowitz, B. E. and Pappas, T. N., eds., *Proc. SPIE* **7240**, 72400K–1–12 (Jan. 2009).

[14] Heyes, A. D., "The sonic pathfinder: A new electronic travel aid," *Journal of Visual Impairment and Blindness* **78**, 200–202 (May 1984).

[15] Zawrotny, K., *et al.,* "Fingertip vibratory transducer for detecting optical edges using regenerative feedback," *Int. Symp. Haptic Interfaces for Virtual Environment and Teleoperator Systems,* **0**, 57 (2006).

[16] Meers, S. and Ward, K., "A vision system for providing 3d perception of the environment via transcutaneous electro-neural stimulation," in [*Eigth Int. Conf. Information Visualisation*], 546–552 (July 2004).

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

[18] Hernández, *et al.,* "Computer solutions on sensory substitution for sensory disabled people," in [*Proc. 8th WSEAS Int. Conf. Comp. Intelligence, Man-machine Systems and Cybernetics*], 134–138 (2009).

[19] Pressey, N., "Mowat sensor," *Focus* **11**(3), 35–39 (1977).

[20] Silva, P. M., *Perceiving Graphical and Pictorial Information via Touch and Hearing*, Master's thesis, Northwestern University, Evanston, IL USA (2010).

[21] Meijer, P. B. L., "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.* **39**, 112–121 (Feb. 1992).

[22] Nayak, A. and Barner, K. E., "Optimal halftoning for tactile imaging," *IEEE Trans. Neural Syst. Rehabil. Eng.* **12**, 216–227 (June 2004).

[23] Doel, K. V. D., "Soundview: Sensing color images by kinesthetic audio," in [*International Conference on Auditory Display*], 303–306, IEEE (2003).

[24] Jacobson, R. D., "Navigating maps with little or no sight: Anovel audio-tactile approach," in [*Workshop on Content Visualization and Intermedia Representations*], 95–102 (1998).

[25] Parente, P. and Bishop, G., "BATS: the blind audio tactile mapping system," in [*ACM South Eastern Conference (ACMSE)*], 11–16 (Mar. 2003).

[26] Sagi, D. and Julesz, B., ""Where" and "What" in Vision," *Science* **228**, 1217–1219 (June 1985).

[27] Das, A., Roy, A., and Ghosh, K., "Proposing a cnn based architecture of mid-level vision for feeding the where and what pathways in the brain," in [*Swarm, Evolutionary, and Memetic Computing*], Panigrahi, B., *et al.,* eds., *Lecture Notes in Computer Science* **7076**, 559–568, Springer Berlin / Heidelberg (2011).

[28] Mortimer and Mishkin, "Analogous neural models for tactual and visual learning," *Neuropsychologia* **17**(2), 139–151 (1979).

[29] Kaas, J. H. and Hackett, T. A., "'what' and 'where' processing in auditory cortex," *Nature Neuroscience* **2**, 1045–1047 (Dec. 1999).

[30] Sánchez, J., Lumbreras, M., and Cernuzzi, L., "Interactive virtual acoustic environments for blind children: computing, usability, and cognition," in [*CHI '01 extended abstracts on Human factors in computing systems*], *CHI '01*, 65–66, ACM, New York, NY, USA (2001).

[31] Noton, D. and Stark, L., "Scanpaths in eye movements during pattern recognition," *Science* **171**(3968), 308–311 (1971).

[32] Noton, D. and Stark, L., "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision Research* **11**, 929–942 (1971).

[33] Henderson, J. M., "Human gaze control during real-world scene perception," *Trends in Cognitive Sciences* **7**(11), 498–504 (2003).

[34] Hartmann, W. M., "Pitch, periodicity, and auditory organization," *J. Acoustical Society of America* **100**(6), 3491–3502 (1996).

[35] Wightman, F. L. and Kistler, D. J., "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoustical Society of America* **91**, 1648–1661 (1992).

[36] Hartmann, W. M. and Wittenberg, A. T., "On the externalization of sound images," *J. Acoustical Society of America* **99**(6), 3678–3688 (1996).

[37] R.Glasberg, B. and Moore, B. C. J., "Derivation of auditory filter shapes from notched noise data," *Hear. Res.* **47**, 103–138 (1990).

[38] Tran, T. V., Letowski, T., and Abouchacra, K. S., "Evaluation of acoustic beacon characteristics for navigation tasks," *Ergonomics* **43**, 807–827 (2000).

[39] Blauert, J., "Sound localization in the median plane," *Acustica* **22**, 205–213 (1969-70).

[40] Hartmann, W., "How we localize sound," *Physics today* **52**, 24 (1999).

[41] Blauert, J., "Spatial hearing: the psychophysics of human sound localization," *Book* , 494 (Jan 1997).

[42] Atkins, J. D., "Binaural reproduction of spherical microphone array signals.," *J. Acoust. Soc. Am.* **126**(4), 2156–2156 (2009).

[43] Atkins, J. D., *Spatial acoustic signal processing for immersive communication*, PhD thesis, Johns Hopkins University, Baltimore, MD (2011).

[44] Zhang, W., Abhayapala, T. D., Kennedy, R., and Duraiswami, R., "Modal expansion of hrtfs: Continuous representation in frequency-range-angle," *ICASSP 2009* , 285—288 (Oct 2009).

[45] Duraiswami, R., Zotkin, D., and Gumerov, N., "Interpolation and range extrapolation of hrtfs [head related transfer functions]," *ICASSP 2004* **4**, iv–45– iv–48 vol.4 (2004).

[46] Evans, M. and Angus, J., "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Am.* (Jan 1998).

[47] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., "The CIPIC HRTF database," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* , 99–102 (2001).

[48] Zahorik, P., "Auditory display of sound source distance," in [*In: Proceedings of the 2002 International Conference on Auditory Displays*], 326–332 (2002).

[49] Zahorik, P., "Assessing auditory distance perception using virtual acoustics," *Acoustical Society of America* **111**(4), 1832–1846 (2002).

[50] Litovsky, R. Y., Colburn, H. S., Yost., W. A., and Guzman, S. J., "The precedence effect," *J. Acoustical Society of America* **106**, 1633–1654 (1999). Review and Tutorial.

[51] Zwicker, E., Flottorp, G., and Stevens, S. S., "Critical band width in loudness summation," *Journal of the Acoustical Society of America* **29**, 548–557 (1957).