Follow That Sound: Using Sonification and Corrective Verbal Feedback to Teach Touchscreen Gestures

Uran Oh[1], Shaun K. Kane[3], Leah Findlater[2]
[1,2] Inclusive Design Lab | HCIL
[1] Dept. of Computer Science, [2] College of Information Studies
University of Maryland, College Park, MD
[3] Human-Centered Computing
Dept. of Information Systems
University of Maryland, Baltimore County (UMBC)
{uranoh@cs.umd.edu, skane@umbc.edu, leahkf@umd.edu }

ABSTRACT
While sighted users may learn to perform touchscreen gestures through observation (*e.g.*, of other users or video tutorials), such mechanisms are inaccessible for users with visual impairments. As a result, learning to perform gestures can be challenging. We propose and evaluate two techniques to teach touchscreen gestures to users with visual impairments: (1) *corrective verbal feedback* using text-to-speech and automatic analysis of the user's drawn gesture; (2) *gesture sonification* to generate sound based on finger touches, creating an audio representation of a gesture. To refine and evaluate the techniques, we conducted two controlled lab studies. The first study, with 12 sighted participants, compared parameters for sonifying gestures in an eyes-free scenario and identified pitch + stereo panning as the best combination. In the second study, 6 blind and low-vision participants completed gesture replication tasks with the two feedback techniques. Subjective data and preliminary performance findings indicate that the techniques offer complementary advantages.

Categories and Subject Descriptors
H.5.2 [User Interfaces]: Auditory (Non-speech) feedback; K.4.2 [Social Issues]: Assistive Technologies for persons with disabilities

General Terms
Design, Experimentation, Human Factors.

Keywords
Blindness, visual impairments, sonification, touchscreen, gestures

1. INTRODUCTION
With the widespread adoption of touchscreen devices, gestural interaction has become a primary means of computer input. Despite being a so-called "natural user interface", touchscreen gestures obey certain conventions that must be learned [17]. For example, to correctly perform a directional swipe gesture, the swipe's location, speed, and angular trajectory all need to fall within expected constraints. Sighted users may learn to perform gestures through observing other users, in-application tutorials, or even television commercials. For visually impaired users, such observation is not accessible and, as a result, learning how to perform gestures can be challenging [15]. While recent commercial and research advances have addressed touchscreen accessibility for users with visual impairments (*e.g.*, [1, [3], [11], [13]]), the gesture learning process has been largely ignored.

ASSETS '13, October 21 - 23 2013, Bellevue, WA, USA

ACM 978-1-4503-2405-2/13/10 $15.00.
http://dx.doi.org/10.1145/2513383.2513442

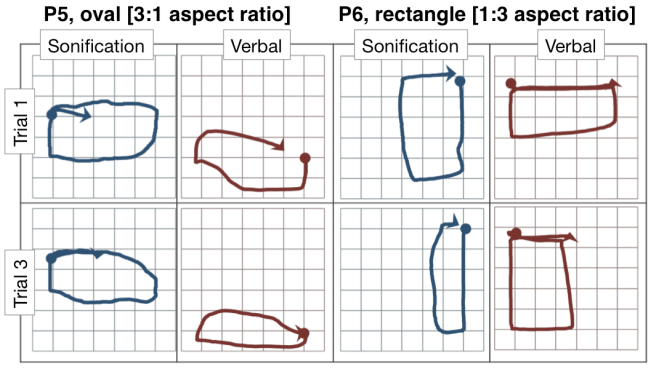

Figure 1. Example shape gestures by two blind participants in Study 2, showing the first and third of three trials. Note details such as lack of closure in one of P5's verbal feedback trials, and changes between the first and third trial for P6.

In this paper, we propose and evaluate two techniques to teach touchscreen gestures to users with visual impairments: *corrective verbal feedback* and *gesture sonification* (Figure 1). The verbal feedback technique automatically analyzes a user's gestures and provides text-to-speech feedback on how to change each gesture to make it more similar to a reference gesture (*e.g.*, "make it longer"). For the sonification technique, sound is generated as the finger touches the screen (*e.g.*, increasing pitch as the finger moves up). Users can compare the sound of their own gesture to that of a reference gesture to determine how to correctly draw it. Sound parameters like pitch, stereo, volume, or timbre can be mapped to the *x*- and *y*-axes of the screen. While prior work has used sonification to convey visual information for blind users (*e.g.*, graphs [4], [25]] or geometric shapes [[8], [19]]), findings from those studies are not necessarily applicable to touchscreen gestures, which can vary not only in shape, but also in location, direction, size and speed (*e.g.*, flick *vs.* drag).

To refine and evaluate these gesture training techniques, we conducted two controlled lab studies. First, to identify the best parameters for sonifying gestures, we compared different sonification mappings with 12 sighted participants in an eyes-free scenario. Among the parameters tested (pitch, volume, timbre, and stereo), stereo on the *x*-axis and pitch on the *y*-axis were significantly more accurate at conveying gestures than any other parameter mapping. For the second study, six blind and low-vision participants compared the two feedback techniques: sonification (using stereo + pitch) and corrective verbal feedback. With each technique, participants performed tap gestures in different locations and of different types, directional swipe gestures, and shape gestures. Subjective data and preliminary performance findings indicate that the two techniques offer sometimes complementary advantages. For example, although the verbal feedback was preferred overall primarily due to the precision of its instructions, almost all participants appreciated the sonification feedback for certain situations (*e.g.*, to convey speed).

This paper makes the following contributions: (1) two straightforward techniques to provide sonified and verbal feedback for blind users' gestures; (2) characterization of the tradeoffs between the two types of feedback; (3) empirical evidence that pitch + stereo is more effective for sonification than alternatives that use volume or timbre on the *x*-axis. These findings provide a basis for the design of future gesture tutorial systems to improve initial touchscreen learnability for blind users.

2. RELATED WORK

Our work builds on accessible gesture design, data sonification, and efforts to train blind users to draw gestures and shapes.

2.1. Accessible Touchscreen Gestures

While early approaches for accessible gesture-based interaction for people with visual impairments combined touchscreen input with physical buttons for confirmation, more recent systems have focused on the touchscreen only [3], [6][7], [11], [12]. *Slide Rule* [11] was a multitouch screen reader for exploring touchscreen applications. *NavTouch* [7], *No-Look Notes* [3], and *BrailleTouch* [6] used touchscreen gestures for entering text. Commercial systems, such as Apple's *VoiceOver*[1] and Google's *Eyes-Free Project*[2], provide access to mainstream mobile applications through accessible gestures. These systems typically rely on a small set of gestures, including single and multi-finger tap, double-tap, tap-and-hold, and directional drags and swipes.

Although many modern touchscreen devices feature built-in screen readers that support accessible gestures, learning and using these gestures remains problematic. First, there is little consistency between gestures supported by different software platforms [13]. Even when two systems share a common gesture, such as tap-and-hold, the details may be different (*e.g.*, duration of a short *vs.* long tap). Second, current systems provide limited support for teaching gestures. For example, *VoiceOver* and *Eyes-Free* provide textual descriptions of gestures (*e.g.*, "swipe left"), but do not provide detailed information about the dynamics of the canonical gesture, such as size and speed. *VoiceOver* also provides a practice area for users to perform gestures and to hear which gestures are recognized by the system, but no feedback is provided about how to perform a specific gesture reliably.

2.2. Sonification of Spatial Data

Converting spatial information to non-speech audio for blind people has been explored for many years; see Hoggan and Brewster [9] for a general overview of non-speech audio output. Brown *et al.* [4] combined pitch and stereo to represent line graphs with two data series: each series was represented using pitch, while stereo position was used to separate the series. *iSonic* [25] combined pitch and stereo panning to represent two-dimensional map and table data. Walker and Mauney [22] explored sonification mappings for blind and sighted readers of auditory graphs, finding that both blind and sighted individuals typically (with some exceptions) applied similar mappings between audio pitch and other variables, such as size and velocity. Walker and Lindsay [21] found that 3D spatial audio beacons could guide individuals through a map path in a virtual reality environment. While these systems have informed the selection of audio parameters for our studies, sonification to provide an understanding of spatial data is unlike gesture sonification, in that users must also be able to *reproduce* a gesture accurately—with details like location, size, speed, and direction.

[1] http://www.apple.com/accessibility/iphone/vision.html
[2] https://code.google.com/p/eyes-free/

2.3. Training Gestures and Shape Drawing

Sighted individuals have many opportunities to learn gestures through visual observation. For example, Apple provides video tutorials for touchpad gestures in Mac OS X[3], while several research solutions provide continuous gesture recognition and visual guidance in real time (*e.g.*, [2], [14]). These approaches are inaccessible to blind users. As an alternative, several projects have used sonification to teach blind users shapes (though not necessarily gestures). GUESS [10], for example, allowed blind users to explore simple shapes using a stylus and tablet with sonification (pitch + stereo) feedback. Timbremap [20] combined stereo, pitch, and spearcons to guide blind users in exploration of a touchscreen map. Harada *et al.* [8] mapped vowel sounds to radial direction to enable blind people to trace shape contours. These systems enabled tracing of shapes, but were primarily optimized for slow exploration of a shape, rather than aspects such as rotation and speed of a gesture.

Multimodal audio and haptic feedback has also been used to convey shapes. Crossan and Brewster [5] combined pitch and stereo sonification with a force feedback controller to drag the user along a trajectory, and found that performance was higher with audio and haptic feedback than haptic feedback alone. *McSig* [19] used this same combination of sonification and force feedback to teach handwriting to blind children, while *SemFeel* [23], *SpaceSense* [24], and work from Nobel and Martin [16] used primarily tactile feedback to transmit directional and shape data. These systems used custom hardware with multiple actuators, technology that is not available on most touchscreen devices. We have thus focused on training with audio feedback only.

3. STUDY 1: EYES-FREE GESTURE SONIFICATION WITH SIGHTED USERS

To explore possible forms of gesture sonification feedback, we conducted a controlled lab study with 12 sighted participants. We tested different sound parameters (*e.g.*, pitch, timbre) mapped to absolute (*x*,*y*) screen coordinates to assess how effectively each parameter conveyed gesture characteristics such as location, size, speed, direction, and shape. We conducted this initial study with sighted participants to achieve a larger sample than possible with blind participants alone, and to refine the sonification technique before presenting it to blind participants in Study 2. Perception of sound mappings has been shown to be largely consistent between blind and sighted people [22], so we believed that testing with sighted people would provide useful guidance for designing gesture sonification schemes for blind or sighted users.

3.1. Method

3.1.1. Participants

Twelve sighted volunteers (5 female) were recruited through campus email lists. They were on average 26.4 years old (range 20–35). All but one participant owned a touchscreen device; nine reported daily touchscreen use. No participants reported hearing difficulties. Half reported playing a musical instrument.

3.1.2. Apparatus

We used a Samsung Galaxy Nexus running Android 4.2.2 with a display resolution of 124 ppcm. We also built a custom Android application, which used *Pure Data*[4] to generate real-time audio based on the (*x*,*y*) location of fingers on the screen. Study sessions took place in a quiet room and participants wore closed, supra-aural stereo headphones (Sennheiser HD 202 II). Since the Galaxy

[3] http://www.apple.com/osx/what-is/gestures.html
[4] http://puredata.info

Nexus does not have a tactile edge to the screen, we created a physical overlay to demarcate a 700×700px region corresponding to the active input area in the app (Figure 2a); the overlay also covered the experimenter's controls, preventing accidental selections by the participant. To impose eyes-free interaction, the device and hands were shielded from view inside a box (Figure 2b). The software logged all interactions with the touchscreen.

3.1.3. Sound Parameters

To identify which sound parameters would be most useful for gesture sonification, we conducted pilot testing with four sighted participants using a variety of audio filter parameters in the *Pure Data* library. We varied the following sound parameters along the *x*-axis,: pitch, volume, timbre (tone), stereo, vibrato, attack/decay (time to increase to and decrease from a peak sound), and tempo (beats per minute). We also tested different ranges, and determined a comfortable range for each parameter.

After pilot testing, we excluded vibrato, attack/decay, and tempo from further evaluation because they each had a temporal component that interfered with conveying `gesture speed. Of the remaining four parameters, pitch was best for all participants at conveying directionality and the start/end location of swipe gestures. We thus mapped the *y*-axis to pitch for all conditions. Along the *x*-axis we compared three sound parameters: volume, timbre, and stereo. Rather than using a continuous sound change we instead divided each axis into 10 equal-sized, discrete steps, which made it easier to detect auditory changes. We conducted an additional five pilot sessions to identify distinguishable lower and upper ends of the range and step sizes for each parameter, where applicable (*i.e.*, identifying comfortable low and high volume settings). Final configurations were as follows:

Pitch. We varied sound frequency to generate 10 pitch values that correspond to consecutive musical notes near middle C on a piano (261.63Hz). Pitch ranged from a low of B3 (246.94Hz) at the bottom of the screen to D5 (587.33Hz) at the top of the screen. Moving the finger vertically effectively plays a C major scale.

Volume. To manipulate perceived volume, we adjusted the gain of the amplifier from 0.1 (0 is absolute silence) to 1 (full gain). A step corresponded to a 0.1 increase/decrease in gain.

Timbre. Timbre refers to tone quality. We varied timbre from a pure sine wave (smooth) on the left side of the screen to a pure triangle wave (jagged) on the right side of the screen; we did not use sawtooth or square waves due to their relatively discomforting sounds. To transition from the triangle wave to the sine wave, we perceptually combined the two sound waves by reciprocally adjusting the gain of each—that is, the triangle wave gain (α) decreased uniformly from 1.0 on the left to 0.0 on the right and the sine wave gain correspondingly increased (always 1-α).

Stereo (pan). To create the perception of sound panning left-right as the finger moves horizontally, we adjusted the gain in the right and left channels. For a touch point on the left of the screen, no sound played (no gain) in the right channel, and the left channel gain decreased by from a high of 1.0 on the far left to 0.2 near the middle (step size of 0.2 gain); vice versa for the right side.

3.1.4. Procedure

The procedure was designed to fit in a single two-hour session. The three different *x*-axis sound conditions (*Volume*, *Timbre*, *Stereo*) were fully counterbalanced and participants were randomly assigned to a presentation order. For each condition, participants began by freely exploring the screen while hearing sound feedback for 30 seconds, then performed several gestures as instructed by the researcher: drawing vertical, horizontal, and diagonal lines, tapping the four corners and center of the screen, and drawing a few patterns of their choice. Following this practice activity, participants completed four tasks in the following order: *Line Length*, *Line Speed*, *Tap Location*, and *Shape*. A fifth task, *Tap Type*, was tested only once at the end of the session (not per condition) because it did not require 2D sonification.

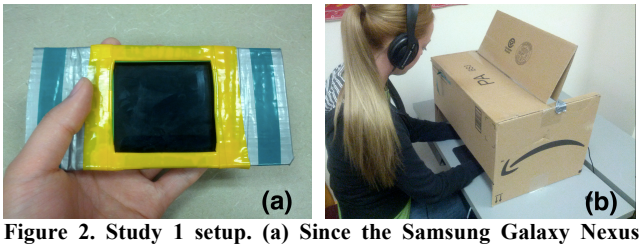

Figure 2. Study 1 setup. (a) Since the Samsung Galaxy Nexus screen has no tactile edge, we used a physical overlay to demarcate the active screen area. (b) To impose eyes-free use, participants placed their hands inside of a box during tasks.

- *Line Length*. Sixteen swipe gestures of varying direction and length: 8 directions (left, right, up, down, and the 4 diagonals) × 2 lengths (short: 315px, long: 630px).
- *Line Speed*. Sixteen swipe gestures of varying direction and speed: 8 directions (same as above) × 2 speeds (fast: 1/2 px/ms, slow: 1/6 px/ms).
- *Tap Location.* Nine tap locations, distributed one per cell across a 3×3 grid filling the entire screen. The location within a cell region was randomly chosen.
- *Shape.* Five single-stroke shapes with varying characteristics (*e.g.*, closed *vs.* open, curved *vs.* straight): circle, diamond, small letter 'e', capital letter 'W' and 'Σ'.
- *Tap Type.* Four tap types: single short (200ms) and long (1000ms) taps, and double and triple short taps (with 400ms gap between taps). Since 2D location and trajectory are not necessary to communicate tap type, we tested this task only with one sound (a mid-range pitch and volume).

For each task, we first gave a description of possible gesture variations (*e.g.*, "we will be testing direction and size of a swipe gesture") and had participants complete a small number of practice trials. Two blocks (repetitions) of the full set of gestures were then given, with trials randomized within a block. For each trial, the software played the *sound prompt* and the participant drew the corresponding gesture. A gesture was deemed to be correct if it was closer to the reference gesture in every characteristic (*e.g.*, direction and length) than to any other gesture in the tested set. After a correct gesture, a chime sound played, while for incorrect gestures, an atonal "thunk" sound played. If the attempt was incorrect, the participant was allowed a single second attempt. The *Shape* task was an exception because its gestures were the most complex: thus, the sound prompt played twice per attempt, participants were *required* to complete two attempts per trial, and no audio feedback on correctness was provided. For all tasks, participants held the device inside a box so that it was shielded from view (Figure 2b). Questionnaires were given after each task and at the end of the study.

3.1.5. Experiment Design and Analysis

This experiment used a within-subjects design with a single factor, *Sound Parameter* (levels: *Volume*, *Timbre*, *Stereo*). The main measures for *Line Length* and *Line Speed* were angular difference (in degrees) and speed difference, respectively, between the reference gesture and drawn gesture. We simplified the analysis for these tasks by calculating average measures for horizontal, vertical, and diagonal directions rather than analyzing all eight directions individually. We then ran separate 3-way

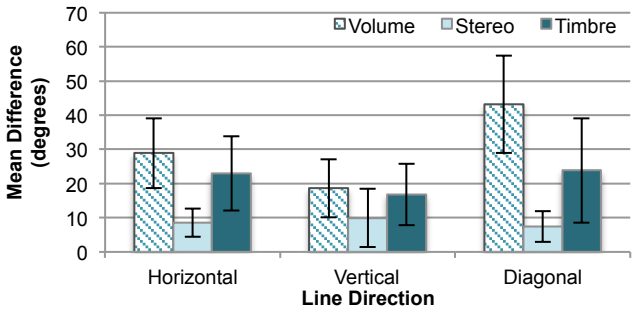

Figure 3. Angular difference for the three sound types for line gestures in horizontal, vertical, and diagonal directions in the *Line Length* task; lower is better. *Stereo* resulted in the lowest angular difference. (*N* = 12; error bars: 95% confidence intervals)

repeated measures ANOVAs (*Sound* × *Direction* × *Speed* or × *Length*) for each of the dependent measures. For the tap location task, we ran separate 2-way repeated measures ANOVAs (*Sound* × *Location*) with horizontal and vertical difference between the reference and drawn gestures as dependent measures. All posthoc pairwise comparisons were protected against Type I error with Holm's Sequential Bonferroni adjustments. Where degrees of freedom are not whole numbers, a Greenhouse-Geisser adjustment was applied to account for violations of the sphericity assumption (tested using Mauchly's test for sphericity).

To capture more experienced behavior, we focused our analysis on the second block of trials within each task and the final attempt within a trial (a second attempt only occurred if the first attempt was incorrect). The same pattern of results occurs if we examine the first attempt only, largely because *Stereo* required fewer repeat attempts than the other sounds, meaning that the advantages of *Stereo* seen in the next section would likely be magnified by examining the first attempt alone. On average there were 1.24 attempts per trial (*SD* = 0.14) for *Stereo*, followed by 1.45 (*SD* = 0.16) for *Timbre*, and 1.54 (*SD* = 0.13) for *Volume* across *Line Length, Line Speed* and *Tap Location* tasks. For the *Shape* task, participants had to complete all attempts regardless of accuracy.

3.2. Findings

Due to space limits, we report only significant ($p < .05$) main or interaction effects involving *Sound*, our primary factor of interest.

3.2.1. Identifying Direction

Stereo was most effective at conveying direction among the three sound parameters and, as shown in Table 1, resulted in the lowest angular difference in the *Line Length* and *Line Speed* tasks. For both tasks, there was a significant main effect of *Sound* on angular difference (*Length* task: $F_{2,22} = 13.07$, $p < .001$, $\eta^2 = .54$; *Speed* task: $F_{2,22} = 16.94$, $p < .001$, $\eta^2 = .61$). Posthoc pairwise comparisons in each case showed that *Stereo* was significantly more accurate than both *Volume* and *Timbre* (all $p < .05$). Although we did not directly compare the line length and line speed tasks, angular difference may be lower in the *Line Speed* task because participants had more practice at that point.

The positive effects of *Stereo* in the *Line Length* task were strongest for horizontal and diagonal swipes, which is not surprising given that those directions rely on *x*-axis sonification (Figure 3). This result was seen in a significant interaction effect of *Sound* × *Direction* on angular difference ($F_{4,44} = 4.21$, $p = .006$, $\eta^2 = .277$). Posthoc pairwise comparisons showed that *Stereo* was more accurate than *Volume* for horizontal and diagonal swipes ($p < .05$). Additionally, a 3-way interaction effect between *Sound* × *Length* × *Direction* ($F_{4,44} = 5.880$, $p = .001$, $\eta^2 = .348$) was found. Posthoc pairwise comparisons were inconclusive.

Table 1. Mean angular difference the *Line Length* and *Line Speed* tasks. *Stereo* was significantly more accurate than *Volume* and *Timbre*. (*N* = 12)

Task	Sound Parameter		
	Volume	*Stereo*	*Timbre*
Line Length	30.2 (*SD* = 13.4)	8.6 (*SD* = 6.4)	21.2 (*SD* = 13.8)
Line Speed	22.8 (*SD* = 11.4)	4.2 (*SD* = 4.2)	19.8 (*SD* = 17.1)

3.2.2. Line Length

Participants were able to differentiate between the two line lengths and to reproduce lines of each length. On average across all three sound parameters, the drawn gesture lengths were off by 101.7px (*SD* = 17.3px) from the reference gesture, a much smaller amount than the difference between the short and long reference gestures themselves (315px). No significant main or interaction effects were found on length difference.

3.2.3. Gesture Speed

In the *Line Speed* task, participants were generally able to differentiate between the two speeds and to reproduce gestures at each speed. On average across all three sound parameters, the drawn gesture speeds were off by 0.16px/ms (*SD* = 0.22) from the reference gesture, less than the 0.33px/ms between the speeds of the short and fast reference gestures. No significant main or interaction effects on the measure of speed accuracy were found.

3.2.4. Tap Location

Stereo again performed well in the tap location task as compared to the other two sounds. In the horizontal direction, taps in the *Stereo* condition were only off by an average of 4.7mm, or 58.4px (*SD* = 22.6), while *Volume* and *Timbre* were off by 103.5px (*SD* = 31.0) and 82.5px (*SD* = 29.8), respectively. A main effect of *Sound* on *x*-direction difference was significant ($F_{1.36,14.96} = 13.97$, $p = .001$, $\eta^2 = .56$), with posthoc pairwise comparisons revealing that *Stereo* was more accurate than both *Volume* and *Timbre* ($p < .05$). For vertical difference, where pitch was always used on the *y*-axis, no significant main effects were found. There was a significant interaction of *Sound* × *Location* on vertical difference ($F_{16,176} = 1.865$, $p = .026$, $\eta^2 = .145$), although no posthoc pairwise comparisons were significant.

3.2.5. Tap Type

Participants found the *Tap Type* task to be easy, and they were 100% accurate in distinguishing number of taps (recall that we did not compare the three sound parameters for this task). The only errors were in distinguishing tap length (short *vs.* long single tap), where participants sometimes underestimated the length of the long tap. The average duration of short single taps was 155.5ms (*SD* = 94.5), while duration for long single taps was 740.5ms (*SD* = 565.5). These lengths were shorter than the short and long tap durations of the reference gestures (200 and 1000ms).

3.2.6. Shape

For the *Shape* task, we visually inspected the drawn shapes. There were no conclusive differences among the three sounds, with participants exhibiting difficulties in completing shapes regardless of sound type. See Figure 4 for examples.

3.2.7. Subjective Preference

When asked to rank the three sound conditions, all 12 participants ranked *Stereo* first, often citing the ease with which it conveyed horizontal differences. For second place ranking, *Volume* and *Timbre* were roughly equally split (7 and 5 votes, respectively).

3.2.8. Summary

The pitch + stereo combination was the most easily discernable mapping, and was preferred by all participants. It improved angular accuracy in the line tasks and horizontal location accuracy

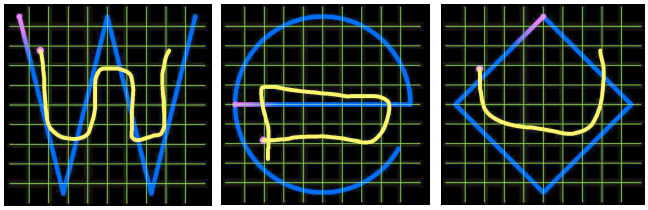

Figure 4. Example shapes from one participant (P8) in the *Stereo* condition, demonstrating both the sporadic success (the 'W' shape here) and common difficulties seen with all sound types. Reference gesture in blue; drawn gesture in yellow.

in the tap location task compared to the other sound combinations. For the shape gestures, visual inspection of the drawn shapes suggested that users need more than only sonification to reliably reproduce a shape. Thus, in Study 2 we use pitch + stereo within a more realistic tutorial procedure, where participants are given a verbal description of the reference gesture before drawing it.

4. STUDY 2: COMPARING FEEDBACK TECHNIQUES WITH BLIND USERS

The second study evaluated two gesture feedback techniques for blind users: (1) *gesture sonification* using the pitch + stereo combination that was best in Study 1, and (2) *corrective verbal feedback* using text-to-speech and generated by automatically analyzing the drawn gesture. While the software recorded performance data as in Study 1, the primary focus here was to collect subjective responses on the feedback techniques.

4.1. Method

We used the same device and physical overlay as in Study 1, except that we did not use a box to shield the device from view, since the Study 2 participants were blind.

4.1.1. Participants

Six visually impaired volunteers (3 female, 3 male) participated in this study. The average age was 36.1 (*SD* = 16.2; range 24–62). All participants were totally blind, except for one who had low vision (20/200). On average, participants had 15.8 years of experience with computers (*SD* = 6.2; range 9–25). All participants owned a touchscreen device and four participants reported daily touchscreen use. One participant reported a mild auditory disorder but was able to complete the study tasks. Five participants played at least one musical instrument, one of whom reported having perfect pitch.

4.1.2. Feedback Techniques

4.1.2.1 Gesture Sonification Feedback

The gesture sonification condition was based on the pitch + stereo combination found to be best in Study 1. Study 2 incorporated sonification in the following ways: (1) a *sonified preview* of the reference gesture, presented before the first gesture trial and accompanied by a text-to-speech description of the gesture; (2) *sonification feedback* produced when the user touched the screen; (3) upon an error, a *replay* of the reference gesture sound for comparison to what had been generated by the user. For sonification, the screen was divided into a 9×9 grid, where each row mapped to a different pitch and each column mapped to a different stereo position. As with Study 1, pitch was set to D_5 (587.33Hz) in the topmost row and dropped by one musical note on the C major scale per row. Gain for stereo panning (*x*-axis) was also manipulated similarly to Study 1, with the exception that the middle column in the grid was set to a gain level of 0.1 in both the left and right channels to create a perceptually smooth horizontal transition.

4.1.2.2 Corrective Verbal Feedback

The corrective verbal feedback condition consisted of a text-to-speech description of the gesture, presented before the gesture trial, and text-to-speech corrective feedback after errors. To generate the corrective feedback, the software compared the drawn gesture to the reference one as follows:

Speed. For gestures that required a specific speed, the software told the user whether the gesture needed to be "faster" or "slower". This feedback was used for swipes and double taps (to distinguish short/fast double taps).

Size and aspect ratio. For swipes of different lengths, feedback was provided to make the gesture "longer" or "shorter". For more complex shapes, feedback was provided on the aspect ratio. When the reference gesture had an aspect ratio of 1:1, feedback consisted of "try wider" or "try taller", as appropriate. For other aspect ratios, both width and height feedback was provided. For example, if the expected aspect ratio was 1:3 but the drawn gesture was 1:2, the feedback would be: "try taller and narrower".

Direction. Directional feedback was based on the angle of rotation. For swipes in this study, we tested only the horizontal direction (right/left), so the feedback was: "opposite direction".

Location. When location was important (for the *Tap Location* task), feedback was given based on the four cardinal directions (*e.g.*, "higher" or "more to the right and higher").

Repetition. To correct single or multiple taps, the system asked the user to try more or fewer taps.

When a drawn gesture was erroneous in multiple ways, the feedback was concatenated. For example, a swipe that was too short and slow would result in: "longer and faster".

4.1.3. Procedure

Study sessions were designed to last 90 minutes. Order of presentation for the feedback techniques (*Sonification* and *Verbal)* was fully counterbalanced, and participants were randomly assigned to an order. For each feedback condition, the procedure mimicked a gesture tutorial scenario. The following tasks were tested in random order, with gestures presented randomly within each task. Participants were asked to replicate each gesture.

- *Swipe.* Twelve swipe gestures: 2 directions (left, right) × 3 lengths (short: 157.5px, medium: 315px, long: 630px) × 2 speeds (fast: 1/6 px/ms, slow: 1/2 px/ms).
- *Tap Location.* Nine tap locations, one in each cell of a 3×3 grid covering the screen. Locations were described to participants as "top-center of the screen", "bottom-right", *etc*.
- *Shape.* Six shapes: 2 base shapes (circle, rectangle) × 3 aspect ratios (large: [1:1], short and wide: [3:1], and tall and narrow: [1:3]). As an example, a circle that has a 1:3 ratio was described as "a tall, narrow oval".
- *Tap Type.* Three tap types: single tap, and slow and fast double taps. Each tap lasted 100ms, with 200ms and 500ms gaps between taps for slow and fast double tap, respectively.

Participants performed three trials of each gesture. For the first trial, the software played a verbal description using text-to-speech (*e.g.*, "a tap in the top-left of the screen"). This description was followed immediately by a sonified preview of the reference gesture in the *Sonification* condition (like Study 1). After a correct gesture trial, a chime sound played. After an incorrect trial, the system played a "thunk" sound, followed by either corrective verbal feedback or a replay of the audio prompt, depending on the feedback condition. Participants were asked about their subjective experience after each task and at the very end of the study.

As in Study 1, swipe gestures and tap type were deemed to be correct if they were closer on all characteristics (*e.g.*, length, speed, direction) to the reference gesture than to any other gesture in the set. For tap locations, the drawn gesture was correct if it was within 4.8mm of the reference gesture's location, equivalent to falling within the bounds of a reasonably sized touch target centered at that location using a 9.6mm target size [18]. Finally, for shapes, an aspect ratio between 4:3 and 2:3 was considered a correct square (1:1), narrower than 2:3 was considered tall and narrow, and wider than 4:3 was considered short and wide.

4.2. Findings

Due to the sample size, we focus primarily on subjective findings, descriptive statistics, and individual user differences. However, we also report on statistically significant findings where applicable—these should be considered preliminary, but will be useful for informing the design of a future gesture tutorial system.

4.2.1. Performance

4.2.1.1 Swipe

For the *Swipe* task, *Verbal* was particularly effective for correcting line length. All six participants improved in length accuracy after receiving corrective verbal feedback, from being on average 102.0px (*SD* = 13.8) off from the reference gesture length in the first trial to only 73.0px (*SD* = 15.5) off in the third trial. With *Sonification*, only half the participants improved on this measure from the first to third trials (first trial: *M* = 89.9px, *SD* = 37.3; third trial: *M* = 100.0px, *SD* = 27.8). There was no evidence that either feedback type had an impact on line speed (*Sonification* trial 1: *M* = 0.17px/ms, *SD* = 0.09; and trial 3: *M* = 0.20px/ms, *SD* = 0.13; *Verbal* trial 1: *M* = 0.23px/ms, *SD* = 0.10; and trial 3: *M* = 0.21, *SD* = 0.15). Finally, participants always correctly replicated the direction (right/left) of the reference gesture.

Examining where drawn swipes were located, we found that participants exhibited a tendency to begin gestures close to the edge of the device. Participants had been told that swipe gestures were centered on the screen (at an *x*-axis location of 350px). However, the midpoints of the drawn gestures per participant were offset: for left-to-right swipes the average midpoint was left of center at 275.1px (*SD* = 34.6), and for right-to-left swipes it was right of center, at 398.3px (*SD* = 59.2). A two-way repeated measures ANOVA (feedback type × direction) showed that direction (left-to-right or right-to-left) had a significant impact on the midpoint location ($F_{1,5} = 24.79$, $p = .004$, $\eta^2 = .83$). No other main or interaction effects were significant.

4.2.1.2 Tap Location and Type

Both feedback types had a positive impact on tap location. Calculating the Cartesian distance between drawn tap locations and the reference gesture location, participants were off by on average 43.3px (*SD* = 15.0) with *Sonification* and 42.5px (*SD* = 21.7) with *Verbal*, across all three trials. From the first to the third trial, all participants improved in accuracy with *Sonification* (improvement in px: *M* = 38.7, *SD* = 46.8) and 5/6 improved in accuracy in *Verbal* (improvement in px: *M* = 24.8, *SD* = 18.6).

For *Tap Type*, participants made no errors in recognizing and performing single tap gestures compared to double taps. For double taps, the difference between the gap lengths of the drawn gestures and those of the reference gestures was similar for both feedback conditions (*Sonification*: *M* = 118.7ms, *SD* = 51.8; *Verbal*: *M* = 129.9ms, *SD* = 46.8). These differences are less than the 300ms difference between the slow and fast double tap reference gestures (500ms *vs.* 200ms).

Table 2. Swipe and tap average difference between reference gesture and drawn gestures across all three trials per gesture in Study 2. Smaller numbers are better; 1px = 0.08mm. (*N* = 6)

Measure	Sonification		Verbal	
	Mean	*SD*	*Mean*	*SD*
Swipe speed (px/ms)	0.18	0.10	0.26	0.18
Swipe length (px)	92.7	28.5	90.7	9.3
Tap location *x* (px)	37.8	19.1	39.4	16.3
Tap location *y* (px)	33.0	5.29	30.5	9.8
Double tap gap duration (ms)	119.0	55.2	142.8	69.6

Table 3. Average aspect ratios of drawn shapes in Study 2 for both circles/ovals and squares/rectangles, showing change from trial 1 to trial 3 with tall and wide shapes. (*N* = 6)

Reference Gesture Aspect Ratio	Sonification		Verbal	
	Trial 1 *M* (*SD*)	Trial 3 *M* (*SD*)	Trial 1 *M* (*SD*)	Trial 3 *M* (*SD*)
Tall [1:3]	0.74 (0.28)	0.55 (0.15)	0.69 (0.22)	0.55 (0.10)
Wide [3:1]	1.40 (0.36)	1.61 (0.14)	1.38 (0.17)	2.06 (0.88)
Uniform [1:1]	0.95 (0.09)	0.90 (0.07)	1.05 (0.19)	0.95 (0.07)

4.2.1.3 Shapes

Both *Verbal* and *Sonification* feedback appeared to help with the accuracy of drawing tall and wide shapes (Table 3). The effect was particularly strong with the tall shapes (aspect ratio of [1:3]), where all participants improved with *Sonification* and all but one participant improved with *Verbal*. For perfect squares and circles—that is, an aspect ratio of [1:1]—participants did not have much trouble completing the gestures accurately on the first trial.

We also assessed how the feedback impacted form closure for the shapes. To do so, we calculated the Cartesian distance between the start and end points of a gesture; a distance of 0px represents perfect closure, while anything greater is either an open shape or has overlapping start/end points (Figure 1). The gap between start and end points was lower with *Sonification* (*M* = 209.4px, *SD* = 83.5) than for *Verbal* (*M* = 285.6px, *SD* = 150.9). Although not a statistically significant difference, this data suggests that it would be useful to further explore if *Sonification* is particularly effective at conveying complex shape features such as closure.

4.2.2. Subjective Experience

Differences between *Sonification* and *Verbal* were clearer in participants' subjective experiences than in the performance data. When asked about overall preference, four of six participants preferred *Verbal*, one wanted both types of feedback, and one preferred *Sonification*. These sentiments were also reflected in ratings of overall satisfaction with the feedback conditions. *Verbal* received more positive ratings than *Sonification* using a 7-point scale (1: 'I like it very much' to 7: 'I don't like it at all'): *Verbal*'s median was 2 (range 1–4) and *Sonification*'s median was 3 (range 2–5). This difference was statistically significant using a Wilcoxon Signed-Rank test ($Z = 2.33$, $p = .02$, $r = .67$). Despite the overall preference for *Verbal*, however, participants' comments highlighted tradeoffs between the two techniques.

Importance of sonified preview. Overall, participants were positive about the sonified preview that played at the beginning of each trial in the *Sonification* condition. Five reported that it was helpful, particularly for conveying time-related characteristics such as speed and duration between two taps, for example:

> *"[The] sound example was helpful for speed"* (P2)
>
> *"Faster and slower are better with audios [audio]"* (P3)
>
> *"That's good [the sonification]. You can tell how fast, how slow [for taps]"* (P5)

There were a smaller number of negative comments, two focused on the utility of *Sonification* for conveying shapes. One of the

participants who generally found the sonified preview helpful also felt it provided too much information when used for a shape. The only participant who did not find the sonified preview helpful noted that it was "unnecessary" and required extra time to listen to before completing the shape tasks (P6).

Complexity of sonification feedback. Some participants commented on the complexity of interpreting the sonification:

> *"I need to pay attention"* (P2)
>
> *"Focusing on two things [pitch and stereo] at the same time was hard"* (P3)
>
> *"You have to listen to the feedback multiple times to make a correction"* (P4)

Precision of verbal feedback. Participants who preferred *Verbal* overall appreciated its preciseness, because it provided clear directions for what to correct. For example:

> *"[I like] telling exactly what you need to do when you messed up"* (P2)
>
> *"Easy to correct gestures by hearing feedback only once"* (P4)
>
> *"It gives more accurate description, more in detail"* (P6)

Three participants (P3, P5, P6), however, commented that a downside of the verbal feedback is that it does not quantify *how much* to adjust a gesture when it makes a suggestion. In this respect, *Sonification* offers additional cues. For example:

> *"Audio [sonification] feedback was more helpful for tap location [than verbal feedback]"* (P1)
>
> *"It [verbal feedback] says it's not narrow enough, but how narrow?"* (P5)

Effects of individual differences. Participants varied in level of visual ability and in musical training, which could impact subjective experience. For example, there was no visual guidance in the study interface—simply a blank screen—yet a user with limited vision may have different preferences than a user who is completely blind. Only one of the six participants had low vision (P1) and was able to see his fingers on the screen when the device was held at a short distance. This same participant reported having a "slight" auditory processing disorder and, ultimately, preferred the corrective verbal feedback. However, he felt that *Sonification* was more helpful for tapping than for swiping and suggested that combining the two forms of feedback would be useful.

Another participant reported having perfect pitch (P5). She had the most extensive musical experience of all participants, and had earned a college degree in music. She was the only participant who preferred *Sonification* overall, reporting that it was useful for conveying many kinds of information, including width, length, and height. She also felt that the pitch was particularly helpful. For example, she could tell based on the pitch change that her swipes were not perfectly straight even though she was attempting to draw a straight line. In contrast to some other participants who felt the sonification was too complex for conveying shapes, this participant reported that the sonified previews for shapes were useful: *"…it was easy to tell the direction, that's neat"*.

5. DISCUSSION

5.1. Parameters for Gesture Sonification

We tested a variety of sound parameters for mapping two-dimensional touchscreen gestures to sound. Based on our collected data, we recommend using pitch to represent movement along the *y*-axis, and stereo panning to represent movement along the *x*-axis. This combination resulted in the best performance on a gesture replication task and was unanimously preferred by the sighted participants in Study 1. In cases where stereo is not usable, such as when the user does not wish to wear headphones, either volume or timbre could be used to represent movement along the *x*-axis; no differences were found between these two combinations. During pilot tests before the full study, we tested and excluded several additional sound parameters that were not appropriate for mapping temporal gesture characteristics: vibrato and tempo, for example, are both periodic and thus interfere with communicating gesture speed. Finally, participants had difficulty replicating shapes based purely on asonification of the shape, suggesting that a more realistic training scenario should also include verbal descriptions of the gesture (as in Study 2).

While our recommendations are based on a study with sighted participants, Walker and Mauney [22] have shown that perception of sound mappings are usually consistent between blind and sighted people. Our findings should thus be useful in informing future work with both sighted and visually impaired participants.

5.2. Toward a Gesture Training System

As shown in our second study, both corrective verbal feedback and gesture sonification offered performance and subjective advantages, suggesting that a combination of the two may ultimately be useful for a gesture training system. Either one or both of the feedback techniques improved gesture accuracy for participants from the first to the third trial in Study 2 in terms of swipe length, tap location, and shape aspect ratio. Overall preference was skewed toward the verbal feedback, though almost all participants also appreciated sonification for some tasks. While participants considered the verbal feedback to be precise and easy to understand, they perceived the sonification to be useful for conveying speed (*e.g.*, slow *vs.* fast taps) and magnitude of change—that is, in communicating not only that a correction needs to be made but by how much.

It is important to note that the verbal feedback and sonification techniques we tested were not informationally equivalent. The verbal feedback reflected relatively simple analysis of shapes, focusing on location, size, speed, and aspect ratio, but did not assess other shape characteristics, such as shape closure or the "roundness" of a circle versus a square. Although not statistically significant, the data in Study 2 suggests that gesture sonification may be superior to verbal feedback at communicating shape closure. Shape characteristics such as roundness and closure may be especially important for performing complex gestures, and we intend to study these tradeoffs further in future work.

Our findings suggest that sonification and corrective verbal feedback could be useful for helping blind users to replicate gestures, but we did not evaluate whether this effect remains after the feedback is removed. A comprehensive gesture tutorial system should extend to these learning contexts. As well, unlike corrective verbal feedback, sonification could be employed during regular use and it would be interesting to explore whether there is a benefit in doing so. Kane *et al.* [13] have shown that blind and sighted users exhibit differences in performing gestures, such as in size and variability, which could cause gesture recognition problems particularly for the blind users. Sonification feedback could potentially address this problem.

We used an *absolute* mapping in this research to provide location information of a gesture, but *relative* sonification has also been used previously to communicate shape trajectories [8]. Location information is important both because some interfaces designed for blind users make use of location-specific gestures and because location can help users position their finger appropriately for the start of a gesture—for example, so the entire gesture fits within the screen bounds. Relative sonification, however, may also be beneficial, especially for gestures that are mostly location

insensitive (*e.g.*, gestures used for scrolling or screen transition). Considering these trade-offs, a comprehensive gesture tutorial system may need to include both absolute and relative sonification depending on the type of gesture being taught.

To build a comprehensive gesture tutorial system would require several extensions to our approach. One participant in Study 2 commented that she would have found feedback useful in learning the two-finger rotation gesture used in iOS *VoiceOver*. However, how to effectively provide feedback for multitouch gestures is an open question. Both of our feedback techniques would need to be extended, perhaps using chords to sonify multiple fingers on the screen. Another possible extension is to provide haptic feedback. Crossan and Brewster [5], for example, used pen-based haptic feedback and stereo + pitch sonification to teach shape trajectories to people with visual impairments. It may be worthwhile, though not necessarily straightforward, to adapt their approach for touchscreen gestures (including location, size, speed) and to use the simpler vibration motor found on most touchscreen devices.

5.3. Limitations

As with any small set of studies, our work has limitations. Study 1 included only sighted participants, although past work [22] suggests that the findings should also apply to blind participants. Additionally, we did not control for previous touchscreen experience in either study. That Study 2 showed an improvement in gesture performance even for these more experienced participants is promising. For completely novice participants, we expect to see a similar or larger effect, although more work is required to confirm this prediction. Another limitation is that we used corrective verbal feedback for only a limited set of differences between the reference gesture and the performed gesture; in future work, we plan to extend this approach with automatically generated feedback. Finally, the findings from Study 2 are useful to inform the design of a comprehensive gesture tutorial system, but it will be important to evaluate such a system with a larger number of users.

6. CONCLUSION

While accessible interfaces have improved touchscreen-based devices for blind users, challenges to true equal access remain. Improving the ability for blind users to learn to use their devices independently will more fully establish touchscreen devices as an option for users of all abilities. The techniques proposed and evaluated in this paper—corrective verbal feedback and gesture sonification—show promise toward this goal. While overall preference was skewed toward the verbal feedback technique, the two techniques appear to offer complementary benefits for supporting gesture replication, both in terms of user performance and subjective experience. A fruitful direction for future work will be to integrate both techniques into a full tutorial system for evaluation with novice touchscreen users.

7. ACKNOWLEDGMENTS

YooJin Kim & Judith Odili provided valuable help and feedback. This research was supported in part by a Google Research Award.

8. REFERENCES

[1] Azenkot, S., Wobbrock, J. O., Prasain, S., and Ladner, R. E. 2012. Input finger detection for nonvisual touchscreen text entry in Perkinput. *Proc. Graphics Interface '12,* 121-129.

[2] Bau, O., Mackay, W. E. 2008. OctoPocus: a dynamic guide for learning gesture-based command sets. *Proc. UIST '08*, 37-46.

[3] Bonner, M., Brudvik, J., Abowd, G., and Edwards, W.K. 2010. No-Look Notes: accessible eyes-free multi-touch text entry. *Proc. Pervasive '10*, 409-427.

[4] Brown, L. M., & Brewster, S. A. 2003. Drawing by ear: Interpreting sonified line graphs. *ICAD '03*, 152-156.

[5] Crossan, A. & Brewster, S. 2008. Multimodal trajectory playback for teaching shape information and trajectories to visually impaired computer users. *ACM TACCESS '08*, *1*(2), Article 12, 34 pages.

[6] Frey, B., Southern, C., and Romero, M. 2011. Brailletouch: mobile texting for the visually impaired. *Proc. HCI International '11*, 19-25.

[7] Guerreiro, T., Lagoá, P., Nicolau, H., Gonçalves, D., and Jorge, J. A. 2008. From tapping to touching: Making touchscreens accessible to blind users. *IEEE MultiMedia '08*, 48-50.

[8] Harada, S., Takagi, H., Asakawa, C. 2011. On the audio representation of radial direction. *Proc. CHI '11*, 2779-2788.

[9] Hoggan, E. and Brewster, S. 2012. Nonspeech Auditory and Crossmodal Output. In: *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. 211-236.

[10] Kamel, H. M., Roth, P., and Sinha, R. R. 2001. Graphics and user's exploration via simple sonics (GUESS): providing interrelational representation of objects in a non-visual environment. *Proc. ICAD '01*, 261-266.

[11] Kane, S. K., Bigham, J. P., Wobbrock, J. O. 2008. Slide rule: making mobile touchscreens accessible to blind people using multi-touch interaction techniques. *Proc. ASSETS '08,* 73-80.

[12] Kane, S.K., Morris, M.R., Perkins, A.Z., Wigdor, D., Ladner, R.E., and Wobbrock, J.O. 2011. Access Overlays: Improving non-visual access to large touchscreens for blind users. *Proc. UIST '11*, 273-282.

[13] Kane, S.K., Wobbrock, J.O. and Ladner, R.E. 2011. Usable gestures for blind people: Understanding preference and performance. *Proc. CHI '11*, 413-422.

[14] Kristensson, P. O., and Denby, L. C. 2011. Continuous recognition and visualization of pen strokes and touch-screen gestures. *Proc. SBIM '11*, 95-102.

[15] Leporini, B., Buzzi, M. C., and Buzzi, M. 2012. Interacting with mobile devices via VoiceOver: usability and accessibility issues. *Proc. OzCHI '12*, 339-348.

[16] Noble, N., and Martin, B. 2006. Shape discovering using tactile guidance. *Proc. EuroHaptics '06*, 561-564.

[17] Norman, D. A. 2010 Natural user interfaces are not natural. *Interactions*, *17*(3), 6-10.

[18] Parhi, P., Karlson, A. K., and Bederson, B. B. 2006. Target size study for one-handed thumb use on small touchscreen devices. *Proc. MobileHCI '06*, 203-210.

[19] Plimmer, B., Reid, P., Blagojevic, R., Crossan, A., and Brewster, S. 2011. Signing on the tactile line: A multimodal system for teaching handwriting to blind children. *ACM TOCHI '11*, *18*(3), Article 17, 29 pages.

[20] Su, J., Rosenzweig, A., Goel, A., de Lara, E., and Truong, K. N. 2010. Timbremap: enabling the visually-impaired to use maps on touch-enabled devices. *Proc. MobileHCI '10,* 17-26.

[21] Walker, B. N., and Lindsay, J. 2005. Navigation performance in a virtual environment with bonephones. *Proc. ICAD '05*, 1-26.

[22] Walker, B. N., and Mauney, L. M. 2010. Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *TACCESS '10*, *2*(3), Article 12, 16 pages.

[23] Yatani, K., and Truong, K. N. 2009. SemFeel: a user interface with semantic tactile feedback for mobile touch-screen devices. *Proc. UIST '09,* 111-120.

[24] Yatani, K., Banovic, N., and Truong, K. 2012. SpaceSense: representing geographical information to visually impaired people using spatial tactile feedback. *Proc. CHI '12*, 415-424.

[25] Zhao, H., Plaisant, C., Shneiderman, B., and Lazar, J. 2008. Data sonification for users with visual impairment: a case study with georeferenced data. *ACM TOCHI '08*, *15*(1), Article 4, 28 pages.