LEARNING RATES FOR AUDITORY MENUS ENHANCED WITH SPEARCONS VERSUS EARCONS

Dianne K. Palladino and Bruce N. Walker

Sonification Lab, School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332
dianne.palladino@gatech.edu, bruce.walker@psych.gatech.edu

ABSTRACT

Increasing the usability of menus on small electronic devices is essential due to their increasing proliferation and decreasing physical sizes in the marketplace. Auditory menus are being studied as an enhancement to the menus on these devices. This study compared the learning rates for earcons (hierarchical representations of menu locations using musical tones) and spearcons (compressed speech) as potential candidates for auditory menu enhancement. We found that spearcons outperformed earcons significantly in rate of learning. We also found evidence that spearcon comprehension was enhanced by a brief training cycle, and that participants considered the process of learning spearcons much easier than the same process using earcons. Since the efficiency of learning and the perceived ease of use of auditory menus will increase the likelihood they are embraced by those who need them, this paper presents compelling evidence that spearcons may be the superior choice for such applications.

[Keywords: Spearcons, Earcons, Auditory Icons, Speech Interfaces, Menu Navigation, Auditory Menus]

1. INTRODUCTION

To support the growing feature sets and shrinking size of mobile consumer devices, there has been an increase in the use of auditory menu-based interfaces. If implemented well, auditory menus have great potential for both sighted and visually impaired users of a variety of devices. Unfortunately, as pointed out by Walker, Nance, and Lindsay [1], relatively little is known about how to make auditory menus effective and usable. Different approaches to enhancing auditory menus have been proposed, and Walker et al. conducted an empirical comparison of menu navigation performance using auditory menus that were enhanced in three different ways. In that study, menus enhanced with *spearcons* [1] outperformed both *auditory icons* [2] and *earcons* [3] for naïve listeners. While spearcons show great promise, it remains to be seen how the different menu enhancements compare in terms of learnability. That is, the earcons, auditory icons, and spearcons are all meant to represent the individual menu items. The more quickly a user can learn the mappings between sounds and menu items, the more usable the interface will be. The present paper reports on a new study that examined how quickly listeners could learn the items in a menu that had either earcon or spearcon enhancements.

1.1. Auditory Menus

In applications as varied as telephone-based reservations systems, mobile phone operating systems, and desktop computing environments, presenting menu options to a listener via sound can greatly enhance the range of uses and users. Generally, menu items are converted from text labels into spoken phrases using automated speech synthesis, or text-to-speech (TTS) software. Often a user navigates through an auditory menu by pressing *up* and *down* navigation keys, and listening to the resulting TTS phrases instead of (or in addition to) reading the menu item text. When the listener hears the desired menu item, a *select* or *return* button (or sometimes a spoken command) is used to choose that item.

The enhancements discussed here are typically accomplished by prepending a brief sound called a *cue* (i.e., an earcon, auditory icon, or spearcon) to the TTS phrase. As soon as the user navigates to a menu item, they hear the cue, and then the TTS phrase. In some systems, the user always hears the TTS phrase. In other systems the user can either select the current item or move to the next item, without hearing all (or in some cases, any) of the TTS phrase. That is, if the cue sound is sufficiently informative, then the user need not listen to the TTS phrase. Since speech can be quite slow and inefficient (even when sped up by expert listeners) learning the mapping between the cue and the full menu text can speed up navigation and increase usability.

Due to the transient nature of sounds, there are a few important usability challenges inherent in auditory menus. Since it takes some time to listen to each menu item, quick and efficient movement through a menu structure can be difficult. Further, as one moves about in a menu hierarchy, it can be difficult to maintain an awareness of which menu or sub-menu is currently active. Finally, since there is considerable memory load for auditory interfaces in general, learning an auditory menu structure—which generally enhances usability—can be difficult. Walker et al [1] addressed the issue of speed and accuracy in menu navigation, and showed that spearcons outperform auditory icons and earcons. However, it remains unclear how learning rates vary for menu items enhanced with different types of sounds.

1.2. Auditory Icons

Auditory icons [2] are representations of the noise produced by, or associated with, the thing they represent. In the case of an auditory menu, the auditory icon would sound like the menu item. Auditory icons are intended to use a very direct mapping, so that learning rates should be almost immediate (i.e, representing the item "dog" with the sound of a dog barking should require almost no learning). This would, in principle, be a great benefit of auditory icons. Unfortunately, the directness of the mapping can vary considerably. For example, the sound of a typewriter could represent the menu item "Print document" in a fairly direct, but not exact, mapping of sound to meaning. In the domains where auditory menus are often useful, such as mobile devices and desktop computers, there is often no real sound available to represent a menu item. For instance, there is really no natural sound associated with deleting a file. Thus, in many cases a metaphorical representation would need to be used, rather than the intended direct iconic representation [see 4]. The mapping can even become completely arbitrary, which requires extensive learning, and opens the door for interference by other pre-conceived meanings for cue sounds. For this reason, genuine auditory icons are of limited utility in practical auditory menu applications. As we move forward with more realistic and ecologically valid studies of auditory menus, such as in mobile phone menus, it is less and less likely that auditory icons will be used systematically. For that reason, the present study compared only earcons and spearcons, and did not include auditory icons.

1.3. Earcons

Earcons [3] are musical motifs that are composed in a systematic way, such that a family of related musical sounds can be created. For example, a brief trumpet note could be played at a particular pitch. The pitch could be raised one semi-tone at a time to create a family of five distinct but related one-note earcons. The basic building blocks of earcons can be assembled into more complex sounds, with the possibility of creating a complete hierarchy of sounds having different timbres, pitches, tempos, and so on. These sounds can then be used as cues to represent a hierarchical menu structure [5; 6; 7].

For example, the top level of a menu might be represented by single tones of different timbres (i.e., a different musical instrument); each timbre/instrument would represent a sub-menu. Then, each item within a sub-menu might be represented by tones of that same timbre/instrument; different items in the sub-menu could be indicated by different pitches, or by different temporal patterns. Users learn what each of the cue sounds represents by associating a given sound with its speech equivalent; users are eventually able (at least in theory) to use the sounds on their own for navigation, without the TTS phrases being required. Participants have been shown to be effective at identifying and understanding this hierarchical information in previous studies [5]. Vargas and Anderson [8] also found that earcons combined with speech can aid in increasing the efficiency and accuracy of menu navigation without increasing workload for the user. Advantages of using earcons as menu item cues include their ability to be applied to any type of menu structure, regardless of menu meaning or domain, and their ability to represent hierarchies by building families of sounds. Earcons are limited, however, by the considerable amount of training that can be required to learn the meanings of the auditory elements, the difficulty involved in adding new items to a hierarchy previously created, and their lack of portability among systems. It seems that the arbitrariness of earcons is potentially both a strength and weakness. A further discussion of these issues is provided by Walker, Nance, and Lindsay [1].

1.4. Spearcons

A spearcon [1] is a brief sound that is produced by speeding up a spoken phrase (often a synthetic TTS phrase), even to the point where the resulting sound is no longer comprehensible as a particular word. Indeed, spearcons need not be recognized as speech at all. Walker et al. [1] liken the spearcon to a fingerprint, because of the acoustic relatedness of the spearcon and the original speech phrase.

When used in an auditory menu, the text of a menu item can be converted to speech using TTS, then a spearcon can be produced from that spoken item. The spearcon is then used as the cue for the menu item from which it was derived. All of this can be automated. Spearcons are also naturally brief, easily produced, and are as effective in dynamic or changing menus as they are in static, fixed menus. It should be pointed out that spearcons, as originally formulated, do not necessarily provide the navigational information (i.e., which menu is active) that hierarchical earcons are designed to provide. However, this can be obtained by using more sophisticated spearcons that vary by, for example, gender of the speaker, or incorporate other kinds of navigational cues. Even without any such extensions, Walker et al. [1] found that hierarchical menu search was faster when using spearcons. If spearcons are also easily learned it will decrease frustration for the user, increase usability, and this interface enhancement will be more likely to be adopted by device manufacturers. Thus, as an initial assessment of learning rates, we examined the average number of trials needed for a user to learn menus of words presented with cues that were either spearcons or earcons.

2. METHOD

2.1. Participants

Participants in the main experiment included 24 undergraduate students (9 male, 15 female) ranging in age from 17 to 27 years (mean = 19.9 years). All reported normal or corrected to normal hearing and vision, and participated for partial credit in a psychology course. Participants were also required to be native English speakers. Five of these participants, plus an additional six participants also participated in a brief follow-up experiment of spearcon comprehension. The age range and gender composition of these additional six participants is included in those mentioned above. Finally, three additional participants attempted the primary experiment but were unable to complete the task within the 2-hour maximum time limit. Data from these individuals were not included in any of the analyses, nor in the demographic information above.

2.2. Menu Structures and Word Lists

The key research question was whether listeners could learn to associate cue sounds with TTS phrases, and whether the rate of learning would differ for earcons and spearcons. Thus,

Table 1. Menu structure used for the "Noun List" List Type Condition.

Animals	People Sounds	Objects	Nature	Instruments
Bird	Snoring	Car	Ocean	Piano
Horse	Sneeze	Typewriter	Thunder	Flute
Dog	Clapping	Camera	Rain	Trumpet
Cow	Laughing	Phone	Wind	Marimba
Elephant	Cough	Siren	Fire	Violin

Table 2. Menu structure used for the "Cell Phone Menu List" List Type Condition. Items were taken from existing menus on Nokia N91 Mobile Phones.

Text Message	Messaging	Image Settings	Settings	Calendar
Add Recipient	New Message	Image Quality	Multimedia Message	Open
Insert	Inbox	Show Captured Image	Email	Month View
Sending Options	Mailbox	Image Resolution	Service Message	To Do View
Message Details	My Folders	Default Image Name	Cell Broadcast	Go To Date
Help	Drafts	Memory In Use	Other	New Entry

participants were required to learn sound/word pair associations for two different types of lists.

2.2.1. Noun List

The Noun List included exactly the 30 words used by Walker, et al. [1] in their previous auditory menu study. This list included five categories of words, as shown in Table 1, and included a range of items for which natural (auditory icon) sound cues could be created. The words were in a menu structure, with the first word in a column representing the category title for the list of member words shown in that column. This list was used to study performance with brief, single-word menu items that were related within a menu (e.g., all animals), but not necessarily across menus. The identical words were used in an effort to replicate the previous findings.

2.2.2. Cell Phone List

The Cell Phone List included words that were taken from menus found in the interface for the Nokia N91 mobile phone [9]. This list included the menu category in the first position in each column, followed by menu items that were found included in those categories. This list was used to begin to study performance in actual menu structures found in technology. As can be seen in Table 2, these words and phrases tended to be relatively longer, and also were obviously technological in context. As discussed previously, most of these items do not have natural sounds associated with them, so auditory icons are not a feasible cue type.

2.3. Auditory Stimuli

The auditory stimuli included earcon or spearcon cues and TTS phrases, generated from the two word lists already described. During training, when listeners were learning the pairings of cues to TTS phrases, the TTS was followed by the cue sound.

2.3.1. Text to Speech

All TTS phrases of the word lists were created specifically for this experiment using the AT&T Labs, Inc. Text-To-Speech (TTS) Demo program [10]. Each word or text phrase was submitted separately to the TTS demo program via an online form, and the resulting .WAV file was saved for incorporation into the experiment.

2.3.2. Earcons

As discussed, the Noun List words (see Table 1) came from the Walker et al. [1] menu navigation study. Since part of this study was intended as a replication of that previous study, the original 30 earcons from that study were used again here as cues for the Noun List.

For the Cell Phone List (Table 2), 30 new hierarchical earcon cues were created using Audacity software. Each menu (i.e., column in Table 2) was represented with sounds of a particular timbre. Within each menu category (column), each earcon started with a continuous tone of a unique timbre, followed by a percussive element that represented each item (row) in that category. In other words, the top item in each column in the menu structure was represented by the unique tone representing that column alone, and each of that column's subsequent row earcons were comprised of that same tone, followed by a unique percussive element that was the same for every item in that row.

Earcons used in the Noun List were an average of 1.26 seconds in length, and those used in the Cell Phone List were on average 1.77 seconds long.

2.3.3. Spearcons

The spearcons in this study were created by compressing the TTS phrases that were generated from the word lists. In previous studies [1], TTS items were compressed linearly by approximately 40-50%, while maintaining original pitch. That is, each spearcon was basically half the length of the original TTS phrase. While it is a simple algorithm, experience has shown that this approach can result in very short (one word) phrases being

cut down too much (making them into clicks, in some cases), while longer phrases can remain too long. Thus, in the present study, a slightly different compression algorithm was employed. TTS phrases were compressed logarithmically, maintaining constant pitch, such that the longer words and phrases were compressed to a relatively greater extent than those of shorter words and phrases. Logarithmic compression was accomplished by running all text-to-speech files through a MATLAB algorithm. This type of compression also decreased the amount of variation in the length of the average spearcon, because the length of the file will be inversely proportional to the amount of compression applied to the file.

Spearcons used in the Noun List were an average of 0.28 seconds in length, and those used in the Cell Phone List were on average 0.34 seconds long.

2.4. Apparatus and Equipment

Participants were tested with a computer program written with Macromedia Director to run on a Windows XP platform listening through Sennheiser HD 202 headphones. Participants were given the opportunity at the beginning of the experiment to adjust volume for personal comfort.

2.5. Procedure

2.5.1. Main Experiment

The participants were trained on the entire list of 30 words in a particular list type condition by presenting each TTS phrase just before its associated cue sound (earcon or spearcon). During this training phase the TTS words were presented in menu order (top to bottom, left to right). After listening to all 30 TTS + cue pairs, participants were tested on their knowledge of the words that were presented. Each auditory cue was presented in random order, and, after each, a screen was presented displaying all of the words that were paired with sounds during the training in the grids illustrated in Tables 1 and 2. The participant was instructed to click the menu item that corresponded to the cue sound that was just played to them. Feedback was provided indicating a correct or incorrect answer on each trial. If the answer was incorrect, the participant was played the correct TTS + cue pair to reinforce learning. The number of correct/incorrect answers was recorded. When all 30 words had been tested, if any responses were incorrect, the participant was "retrained" on all 30 words, and retested. This process continued until the participant received a perfect score on the test for that list. Next, the participant was presented with the same training process, but for the other list type. The procedure for the second list type was the same as for the first. The order of list presentation to the participant was counterbalanced.

After the testing process was complete, participants completed a demographic questionnaire about age, ethnicity, and musical experience. They also completed a separate questionnaire pertaining to their experience with the experiment (see the Appendix), such as how long it took them to recognize the sound patterns, and how difficult they considered the task to be on a six point Likert scale.

2.5.2. Follow-up Spearcon Analysis Experiment

Spearcons are always made from speech sounds. Most spearcons are heard by listeners to be non-speech squeaks and chirps. However, some spearcons are heard by some listeners as very fast words (that is, after all, what they are). It is important to remember that it does not matter whether a given spearcon is heard as speech or non-speech, but it is still interesting to examine the details of this still-new audio cue type. To this end, an additional exploratory study was completed in conjunction with the main experiment. After completing the main experiment, five participants assigned to the spearcon condition were also asked to complete a recall test of the spearcons they had just learned in the main experiment. For this, a program in Macromedia Director played each of the 60 spearcons from the main experiment one at a time randomly to the participant. After each spearcon was played, the participants were asked to type in a field what word or phrase they thought the spearcon represented. We also asked six naïve users (new individuals who had had no exposure to the main experiment in any way) to complete this same follow-up experiment. These six naïve listeners would presumably allow us to determine which spearcons were more "recognizable" as spoken words. Note that all participants were informed on an introduction screen that spearcons were compressed speech, in order to control for any possible misinterpretation of the origin of the sounds. Naïve participants did not then participate in the main experiment.

3. RESULTS

3.1. Main Experiment of Learning Rates

A 2x2 mixed design repeated measures ANOVA was completed on the number of training blocks required for 100% accuracy on the recall test. The first independent variable was a between-subjects measure of cue type (earcons vs. spearcons), and the second independent variable was a within-subjects manipulation of list type (Noun List vs. Cell Phone List). The means and standard deviations of numbers of trial blocks for each of the four conditions are shown in Table 3, and illustrated in Figure 1. Overall, spearcons led to faster learning than earcons, as supported by the main effect of cue type, $F(1,22) = 42.115$, $p < .001$. This is seen by comparing the average height of the two left bars in Figure 1 to the average of the two right bars. It is also relevant to mention that the three individuals who were unable to complete the experiment in the time allowed (two hours), and whose data are not included in the results reported here, were all assigned to the earcons group. This suggests that even larger differences would have been found between earcons and spearcons, if those data had been included.

Overall, the Cell Phone List was easier to learn than the Noun Words, as evidenced by the main effect of list type $F(1,22) = 7.086$, $p = .014$. These main effects were moderated by a significant interaction of cue type and list type, in which the Cell Phone List was learned more easily than Noun Words for the earcon cues (Figure 1, left pair of bars), but there was no difference in word list learning in the spearcons condition (Figure 1, right pair of bars), $F(1,22) = 7.086$, $p = .014$. Interpreting this interaction is difficult with the results available here, because it may be attributed to a floor effect apparent for results in the spearcons condition.

Table 3. Number of training blocks necessary to obtain a perfect recall score, for each of the four experimental conditions.

Condition	Mean	SD
Spearcons: Cell Phone List	1.08	0.28
Spearcons: Noun List	1.08	0.28
Earcons: Cell Phone List	6.55	3.30
Earcons: Noun List	4.55	2.25

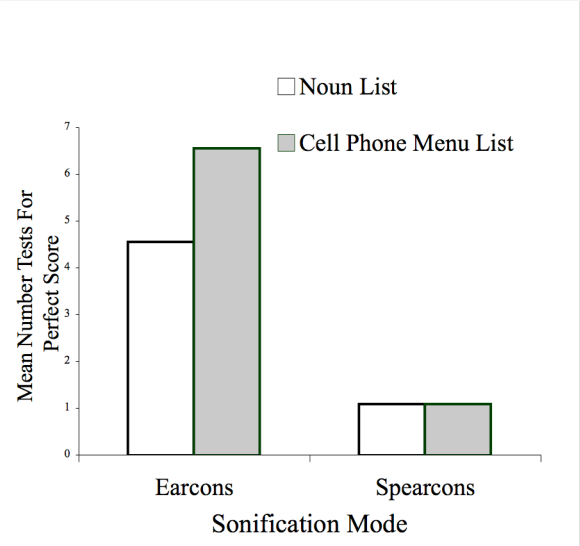

Figure 1. *Mean number of trials necessary for participants to obtain perfect score on sound recall for both earcons and spearcons for Noun and Cell Phone word lists.*

3.2. Debriefing and Follow-up Study Results

Debriefing questions included a six point Likert scale (1="Very Difficult"; 6="Very Easy") on which participants were requested to rate the difficulty of the task they had completed. Participants found the earcons task ($M = 2.91$, $SD = 0.831$) significantly more difficult than the same task using spearcons ($M = 5.25$, $SD = 0.452$), $t(21) = -8.492$, $p < .001$.

Finally, the spearcons analysis follow-up experiment data revealed that the training that the participants received on the word/spearcons associations in these lists led to greater comprehension. Out of a possible 60 points, the mean performance of individuals who had completed the spearcons condition in the main experiment before the spearcons recall test ($M = 59.0$, $SD = 1.732$) was significantly better than that for naïve users ($M = 38.50$, $SD = 3.782$), $t(9) = -11.115$, $p < .001$). No significant main effect was found for list type in the follow-up experiment.

4. DISCUSSION

The difference in means between sonification modes was as expected, as spearcons clearly outpaced earcons in learning rates. The effect of list type, however, was the opposite of what was expected. Since earcons do not provide cues to the word itself, and need to be trained in order for associations to items on a menu to exist, it was not expected that the words included in a menu would make a difference. The spearcons conditions, however, were expected to lead to a significant difference between the two list types, mainly due to the increased contextual information provided by spearcons because they are created directly from the word that they represent. The menu items that were derived from the cell phone menu were generally longer, and therefore provided more remnants of the original TTS to use for recognition purposes. Perhaps the nature of the earcons used in the Cell Phone list were inherently easier to remember due to the particular sounds used, thus leading to faster rates of learning to discriminate among the various sounds. The lack of significant difference in list type for the spearcons condition may also have been due to the floor effect apparent in the results. If the rates of learning had not turned out as fast on average, we may very well have seen more variability in the spearcons condition, and perhaps the interaction would not have been significant. In general, however, these results, combined with the participants' perceptions that learning the spearcons task was significantly easier than for the same task with earcons, and the findings that spearcons used in this study indeed were more recognizable on the whole after training all provide strong empirical evidence of the superior nature of spearcons for use in auditory menus.

From a practical standpoint, the support for spearcons as a preferred sonification mode for menu enhancement is fourfold. First, spearcons are very easy to create, so it is feasible that with the proper technological enhancement, they could be created on the fly for ease of use in any language or application. Secondly, using spearcons does not restrict the structure of a menu system. Their use in a menu hierarchy can be as fluid as necessary, because they do not require fixed indications of grid position. For this reason, they also can be considered a strong candidate for any imaginable menu system, not just for the standard hierarchical menu common in today's applications. Thirdly, this study has shown that spearcons are very easy to learn, and therefore will minimize frustration and training time for new users. Finally, spearcons are short in length. With the average size of the earcons used in this study over one and a half seconds, and the average spearcons size less than one third of a second, spearcons are poised to provide greater efficiency for users of electronic menus. Once learned, it is feasible that the time to reach a menu item will be much less with menus using spearcons than earcons, and, therefore, will provide a faster, less frustrating user experience.

The uses of small electronic devices are increasing and becoming more integrated into our lives on a daily basis. More and more, these devices are becoming essential not only for business use, but also for communication and information seeking in countless occupations. It is essential that these devices be accessible to all who could benefit from them, including those who rely on auditory cues exclusively, such as the blind and those with temporarily obstructed vision, such as firefighters and soldiers. The ability to use these devices with minimum frustration and efficient rates of learning will stem directly from the characteristics of the auditory cues that are provided by these devices. Spearcons clearly are capable of fulfilling these needs.

5. ACKNOWLEDGEMENTS

Portions of this research are supported by an award from Nokia. We would also like to thank Mark Godfrey for the algorithm and advice relating to spearcon creation, and Jeff Lindsay for input on statistical analysis.

6. REFERENCES

[1] B.N. Walker, A. Nance, and J. Lindsay, "Spearcons: Speech-based earcons improve navigation performance in auditory menus," Proceedings of the International Conference on Auditory Display, London, U.K., 2006.

[2] W. W. Gaver, "Using and creating auditory icons," in *Auditory display: sonification, audification, and auditory interfaces*, G. Kramer, Ed. Reading, MA: Addison-Wesley, 1994, pp. 417-446.

[3] M.M. Blattner, D. A. Sumikawa, and R.M. Greenberg, "Earcons and icons: Their structure and common design principles, *Human-Computer Interaction*. Vol 5, 1989, pp. 11-44.

[4] B.N. Walker and G. Kramer, "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making," in *Ecological Psychoacoustics,* J.G. Neuhoff, Ed. New York: Academic Press, 2004, pp. 150-175.

[5] S. Brewster, V. Raty, and A. Kortekangas, "Earcons as a method of providing navigational cues in a menu hierarchy," Proceedings of the HCI'96 Conference, Imperial College, London, U.K., 1996.

[6] G. LePlâtre and S. Brewster, "Designing non-speech sounds to support navigation in mobile phone menus," presented at International Conference on Auditory Display (ICAD2000), Atlanta, USA, 1998.

[7] S.A. Brewster, P.C. Wright, and A.D.N. Edwards, "An evaluation of earcons for use in auditory human-computer interfaces," Proceedings of the SIGCHI Conference on Human Factors in Computing systems, Amsterdam, 1993.

[8] M.L.M. Vargas and S. Anderson, "Combining speech and earcons to assist menu navigation," Proceedings of the International Conference on Auditory Display, Boston, MA, USA, 2003.

[9] Nokia N91 Cell Phones, "Nokia NSeries, http://www.nokia.com/nseries/index.html?loc=inside,main_n91"

[10] AT&T Research Labs, "AT&T Text-to-Speech Demo, http://www.research.att.com/projects/tts/demo.html."

7. APPENDIX

7.1. Sample Debriefing Questionnaire (Earcons)

1. Did you recognize that the sounds were organized in a hierarchical manner, with a single tone for the menu category, and the same percussive element for each item underneath? (Circle one, please)

   a. Yes
   b. No

2. If yes, about how long do you think it took you to notice this pattern?

   a. I noticed it right away during the first **training** session.
   b. I noticed this toward the end of the first **testing** session.
   c. I did not notice until I had been trained and tested several times.
   d. I never noticed that there was a pattern – I just memorized the sounds.

3. Do you think that seeing and selecting a word from the menu after hearing the sound, rather than being asked to type what you heard helped you make correct decisions?

   a. Yes
   b. No
   c. Not Sure

4. Please write your reason for answering question 3 the way that you did.

5. How difficult do you think this task was to complete? (Circle one, please)

1	2	3	4	5	6
Extremely Difficult	Very Difficult	Somewhat Difficult	Somewhat Easy	Very Easy	Extremely Easy

Thank you for participating in this research study. Your data will assist us to increase usability in auditory menus. Feel free to make any additional comments that you have not already expressed below.