

EXAMINING THE LEARNABILITY OF AUDITORY DISPLAYS: MUSIC, EARCONS, SPEARCONS, AND LYRCONS

Kay Tislar, Zackery Duford, Brittany Nelson, Madeline Peabody, and Myounghoon Jeon

Department of Cognitive and Learning Sciences
Michigan Technological University
1400 Townsend Dr., Houghton, MI 49931 USA
{cltislar, ztduford, bnelson1, mjpeabod, mjeon}@mtu.edu

ABSTRACT

Auditory displays are a useful platform to convey information to users for a variety of reasons. The present study sought to examine the use of different types of sounds that can be used in auditory displays—music, earcons, spearcons, and lyricons—to determine which sounds have the highest learnability when presented in sequences. Participants were self-trained on sound meanings and then asked to recall meanings after listening to sequences of varying lengths. The relatedness of sounds and their attributed meanings, or the intuitiveness of the sounds, was also examined. The results show that participants were able to learn and recall lyricons and spearcons the best, and related meaning is an important contributing variable to learnability and memorability of all sound types. This should open the door for future research and experimentation of lyricons and spearcons presented in auditory streams.

1. INTRODUCTION

Auditory notifications and representations of tasks, objects, and warnings are becoming ever more prevalent, particularly with screen real estate being at a premium as devices become smaller. There are a host of benefits to presenting information in auditory form, including the fact that it can provide additional information to a user by not overloading the visual channel and by instead capitalizing on the often under-used auditory channel. In addition, when a user is occupied with a task, sound can be a very effective way of capturing a user's attention [1]. For example, one study indicated that people who perform a task while monitoring data perform the task more accurately when the data are presented in auditory rather than in visual form [2].

In order to be effective, auditory displays must be highly learnable. If users cannot learn the sound meanings easily, they may choose not to use the display even if plenty of learning time is available [3].

An important concept in using auditory displays is that of streaming. As defined by Bregman and Campbell, an auditory stream is “a sequence of auditory events whose elements are related perceptually to one another, the stream being segregated perceptually from other co-occurring auditory events” [4]. The auditory events can be variables (sounds) that are mapped to meanings, with auditory dimensions also changing depending on the meaning. The

related auditory dimensions include qualities such as pitch, duration, intensity, and timbre. For example, researchers have experimented with using streams to sonify weather-related data. Schuett, Winton, Batterman, and Walker [5] sonified weather variables including temperature, humidity, wind speed, wind direction, and cloud cover. In some of the experimental conditions, these made up five streams, but in other conditions they only made up three. In the latter case, two of the variables were combined in a stream with an auditory dimension changed to indicate the second variable and its status. For example, temperature and humidity were combined. Temperature was represented by a sound that varied in pitch. Humidity was designated by applying a tremolo effect to the temperature stream, which indicated low or high humidity. Therefore, just one stream was used to represent two variables, each changing in a different auditory dimension.

Prior research has been conducted to compare various sound types. When comparing earcons, auditory icons, and spearcons, users preferred earcons but remembered auditory icons better. When comparing all of those three sound types plus speech, spearcons and speech were both verified to be extremely learnable, and earcons were the most difficult to learn [3]. In comparisons among auditory icons, earcons, and speech, speech was the easiest to learn, auditory icons took the longest to learn, and earcons were the most error prone [6]. According to Ballas, sound identifiability is related to factors such as how easy a listener can form a mental picture of the sound, how familiar the sound is, and sound clarity [7].

Research has also been done to determine how many auditory streams users can be exposed to while still responding correctly to the information they receive. When one dimension changed, users were more accurate when monitoring just one stream. When two dimensions changed, users were just as accurate monitoring one or two streams [8].

The current experiment was intended as a first step in extending this research by examining how many auditory streams containing different sound types users can recall. In addition to music, three other types of sounds were evaluated:

*Earcons (ear + icons) [9] use short musical motives as symbolic representations of objects or events, in which the sound has no relation to the object it represents. Earcons are very flexible in terms of design, but can be difficult to learn because they use arbitrary mappings.

*Spearcons (speech + earcons) [13] consist of compressed speech that may not be comprehensible. They can be nearly as easy to comprehend as speech with only brief training required.

*Lyricons (lyrics + earcons) [10] are a relatively new type of display, in which lyrics and associated earcons are played concurrently. The point is to train users to recall the



This work is licensed under Creative Commons Attribution – Non Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0/>

lyricon so well that the lyrics can be removed, but people will mentally fill in the lyrics themselves.

In the present study, the primary goal was to examine the learnability of different types of sounds, namely which specific sound mappings were remembered best and how many streams from a sequence could be recalled. This was done in order to generate a set of highly learnable sounds that could later be manipulated to form streams that incorporate different sound types. We focused on testing non-speech auditory displays, but we excluded auditory icons because many of the real-world items do not have natural sounds.

This study tested the following two hypotheses:

H1—People will remember more sound meanings of all types (music, earcon, spearcon, lyricon) when the meanings are related to the sounds (in other words, they are intuitive).

H2—People will remember more sound meanings when sequences of sounds contain a variety of sound types (music, earcon, spearcon, lyricon), as opposed to sounds of all the same type. This is based on our assumption that varying the sound type would aid in mentally separating the sounds during listening and recalling.

2. METHODS

2.1. Participants

Twenty undergraduate and graduate students volunteered to participate in the study, with some undergraduate students receiving credits in a psychology course and the other students receiving no compensation. Ten female and ten male students participated, all older than the age of 18. Fifteen of the participants were native English speakers.

2.2. Apparatus and Equipment

Windows® desktop workstations were used to deliver the experiment, and participants listened to the sounds through over-the-ear headphones.

As a way of allowing users to train themselves on the sound mappings, a simple HTML file containing a table was developed. Using the file, participants could click a sound button in a row to hear the sound corresponding to the word displayed in that row (See Figure 1.). Thirty-two sounds were mapped to different words.

The experiment itself was written as a script in the Psychology Experiment Building Language (PEBL) [11], [12]. Participants clicked a **Play Sound** button to play a sound sequence; all 32 sounds were used to build 21 audio sequences containing varying numbers and types of sounds (See Table 1.).









| Click Arrow Button to Play Sound | What the Sound Means |
|--|----------------------|
| ▶ 0:00 / 0:04  | Date |
| ▶ 0:00 / 0:06  | Power Off |
| ▶ 0:00 / 0:02  | URL |
| ▶ 0:00 / 0:03  | Word |
| ▶ 0:00 / 0:00  | Fast Forward |
| ▶ 0:00 / 0:01  | Delete |
| ▶ 0:00 / 0:00  | PDF |
| ▶ 0:00 / 0:01  | Play |

Figure 1. Sample HTML training file.

After the sequence finished playing, a list of all of the individual words to which the sounds had been mapped was displayed on the screen as checkboxes. Participants clicked the box corresponding to each word whose sound they remembered hearing in the sequence, clicked the **Done** button, and repeated the process for all 21 auditory sequences (See Figure 2.).

For each participant, PEBL created a data file that contained information on the value of each checkbox (checked or not) for every sound sequence. Later, a researcher converted each data string into correct/incorrect response values, the dependent variables. The file also contained response time data for each sound sequence.

2.3. Stimuli Design

For the stimuli, we selected 32 words representing either tasks or objects and mapped them to sounds, as agreed on by the research team. Eight music sounds, earcons, spearcons, and lyricons were created.

To find music, we searched royalty-free websites to locate songs that could be used to represent abstract ideas, and then spliced phrases as necessary. Earcons were created using the FL Studio Digital Audio Workstation (DAW) from Image Line. To create spearcons, we pre-recorded sound files of spoken words using the online text-to-speech (TTS) engine, and then applied the SOLA algorithm to create the spearcon files [13]. For the lyricons, we searched lyric websites to find keywords that would represent tasks, then isolated the vocal track, and spliced audio around the sung phrase.

| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|-------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| Length | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 8 | 8 | 8 | 8 | 4 | 4 | 8 | 8 | 12 | 12 | 4 | 8 | 12 |
| Type of Sound | E | L | M | S | E | L | M | S | E | L | M | S | C | C | C | C | C | C | C | C | C |
| Related/Unrelated | U | U | U | U | R | R | R | R | C | C | C | C | R | U | R | U | R | U | C | C | C |

Table 1. Number and types of sounds included in each sequence. Row 1 contains the sequence number, and row 2 indicates the number of separate sounds the sequence contained. Rows 3 and 4 show the sound type and whether or not the sounds and their mappings were intuitive. E: Earcon, L: Lyricon, M: Music, S: Spearcon, U: Unrelated (Non-intuitive), R: Related (Intuitive), C: Combination.

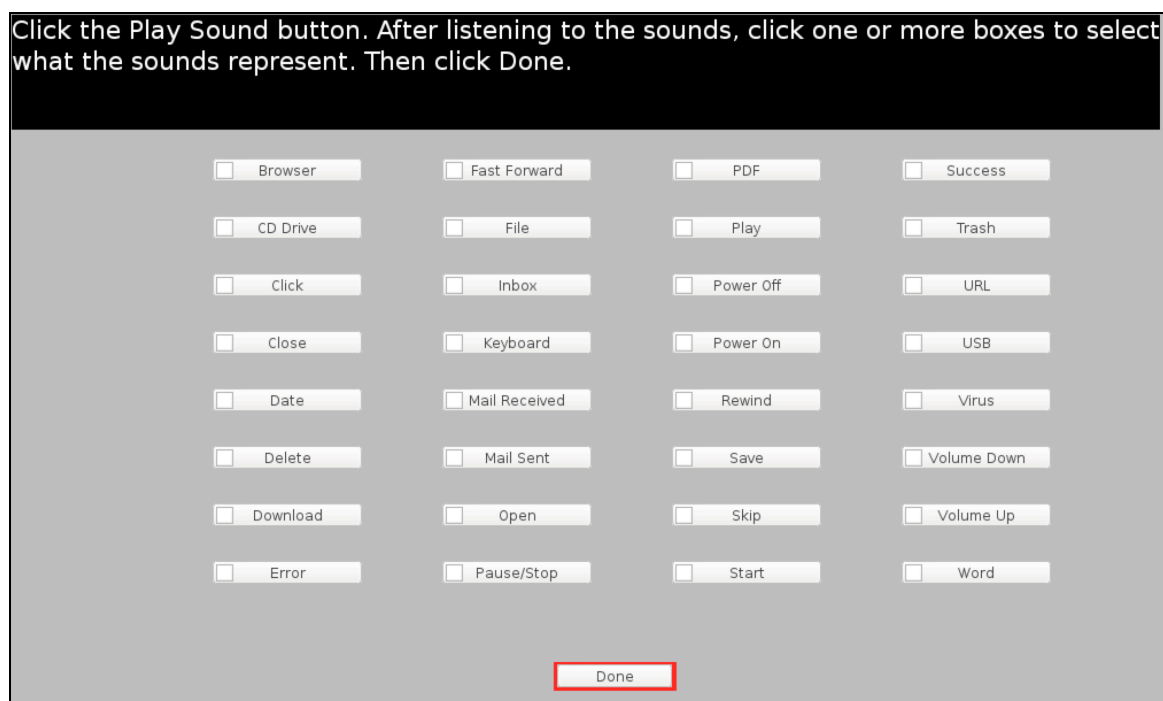


Figure 2. Sample experiment screen in PEBL.

For this study, we did not vary the auditory dimensions and did not join sounds in a stream. A sequence of four sounds therefore represented four streams. The streams were played consecutively, rather than concurrently, to make it easier for users to distinguish which sound they were hearing.

Twenty-one auditory sequences were created using the Audacity® open-source audio software package; we attempted to use each individual sound and sound type approximately the same number of times overall and an equal distribution of related (intuitive) and unrelated (non-intuitive) sounds. A half second time gap was inserted between the sounds in each sequence. All sequences were saved in WAV format.

Table 1 shows the characteristics of each of the 21 sequences, which were delivered in the same order to all of the participants.

2.4. Procedure

Participants were tested one at a time by a single researcher. After a participant signed a consent form and provided demographic information (age, gender, education level, and native language), the researcher asked her or him to be seated at a computer. The researcher then explained the experimental procedure.

Next, the participant was provided with headphones and asked to familiarize herself or himself with the sound mappings using the training file. Users were allowed a maximum of five minutes of training time and could repeat all of the sounds as many times as needed.

Finally, the researcher started the experiment and explained the user interface. The participant then proceeded through the experiment, playing the sound sequences and selecting the sound mappings, until he or she had completed all 21 sequences.

3. RESULTS

As previously noted, some of the participants were not native English speakers. To ensure that native language was not a confounding variable, we tested for a relationship between native language and accuracy. A Spearman's rank-order correlation was run to determine the relationship between native language and overall score accuracy across conditions; it suggested that native language was not a statistically significant predictor of accuracy ($r_s = -.03, p = .89$).

Because spearcons are fast-paced speech in English, the relationship between native language and accuracy in the spearcon condition was also tested separately. Results once again suggested there was not a statistically significant relationship between native language and accuracy ($r_s = .24, p = .30$). Therefore, participants who did not speak English as their first language were retained in the following analyses.

A Pearson product-moment correlation coefficient was computed to assess the relationship between training time and accuracy. Results suggested there was not a statistically significant relation between training time and overall accuracy ($r = -.09, p = .7$). However, a visual inspection of the scatter plot suggested there may be a curvilinear relationship to the data. Therefore, a curvilinear regression analysis was performed using hierarchical multiple regression with one bend in the regression line. When a curvilinear relation is accounted for, the model neared statistical significance, $R^2 = .16, F(1, 19) = 3.12, p = .09$. Future testing with a greater number of participants could produce statistically significant results.

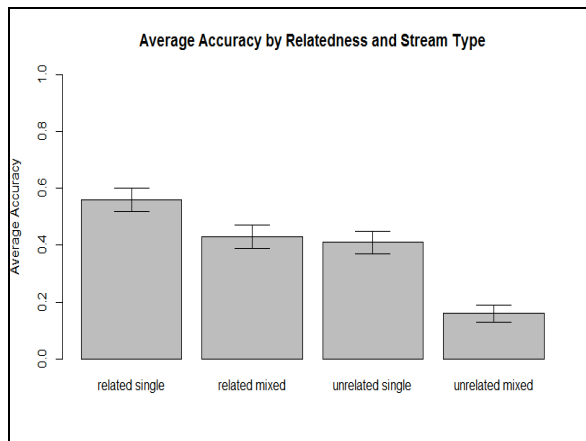


Figure 3. Means and standard errors of accuracy by relatedness and sequence type.

A two-way repeated measures Analysis of Variance (ANOVA) was conducted on the influence of two independent variables (relatedness and single versus mixed sound types in the sequences) on accuracy. Both the main and the interaction effects were statistically significant. Results suggested there was a main effect of relatedness on accuracy, $F(1,19) = 67.91$, $p = .001$. Results also suggested there was a main effect between single versus mixed sound type on accuracy, $F(1,19) = 16.25$, $p = .001$. Finally, there was a statistically significant interaction between both independent variables on accuracy $F(1,19) = 4.93$, $p = .039$ (See Figure 3.). Because the interaction was statistically significant, further analyses were conducted to test if the main effects were still statistically significant across conditions. This was done using paired samples t-tests.

Paired samples t-tests showed that there was a statistically significant difference in accuracy between related and unrelated sounds' scores when sequences consisted of single type sounds ($M = .15$, $SD = .18$), $t(19)=3.862$, $p = .001$, and when sequences contained a combination of different types of sounds ($M = .275$, $SD = .16$), $t(19)=7.678$, $p = .001$. Paired samples t-tests also indicated a statistically significant difference in mean accuracy between single and combined sound sequence types when the sounds were related ($M = .12$, $SD = .23$), $t(19)=2.4$, $p = .027$, and when sounds were unrelated ($M = .243$, $SD = .247$), $t(19)=4.420$, $p = .001$.

To test if there was a statistically significant difference in the accuracy by sound type, a number of paired samples t-tests were performed. A statistically significant difference between scores was found when only spearcons were used ($M = .48$, $SD = .20$) compared to when only earcons were used ($M = .26$, $SD = .15$), $t(19)=6.07$, $p = .001$. A statistically significant difference was also found between the spearcon ($M = .48$, $SD = .20$) and music ($M = .28$, $SD = .13$) conditions, $t(19)=5.44$, $p = .001$. In addition, statistically significant differences were found between the lyricon ($M = .54$, $SD = .14$) and earcon ($M = .26$, $SD = .15$) conditions; $t(19)=7.39$, $p = .001$. The mean accuracy between the lyricon ($M = .54$, $SD = .14$) and music ($M = .28$, $SD = .13$) conditions was also statistically significant, $t(19)=8.77$, $p = .001$.

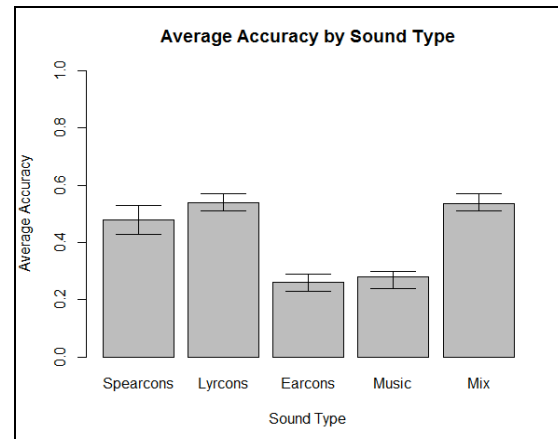


Figure 4. Means and standard errors of accuracy by sound type.

Between mixed and single sound sequences, there was a statistically significant difference between mixed sequences ($M = .54$, $SD = .16$) and single-sound earcon conditions, $t(19)=5.69$, $p < .001$. Finally, there was a statistically significant difference between mixed-sequence accuracy ($M = .54$, $SD = .16$) and single-sound sequence music ($M = .28$, $SD = .13$) conditions, $t(19)=6.24$, $p < .001$. No other t-tests were statistically significant. Overall, accuracy in the mixed sequences was equivalent to accuracy in the spearcon-only and lyricon-only conditions, but was better than in the earcon-only and music-only conditions. (See Figure 4.).

To test if there was a statistically significant difference in the accuracy by number of sounds per sequence, additional paired samples t-tests were performed. There was a statistically significant difference between scores when sequences contained four sounds ($M = .45$, $SD = .13$) compared to sequences that contained eight sounds ($M = .53$, $SD = .16$); $t(19)= -2.02$, $p = .05$. A statistically significant difference was also found between sequences when they contained four sounds versus 12 sounds ($M = .30$, $SD = .10$); $t(19)=4.63$, $p = .001$. Finally, statistically significant differences were found between conditions using sequences of length eight and length 12; $t(19)=5.25$, $p = .001$. All other relationships were not statistically significant (See Figure 5.).

The differences in accuracy according to number of individual sounds could be at least partially due to the fact that the shorter sequences (four and eight sounds) were presented early in the experiment, with most of the four-sound sequences coming at the beginning; on the other hand, the 12-sound sequences came near the end. Participants may have become familiar with the sounds in the four-sound sequences and remembered them when encountering the 8-sound sequences. However, by the time they reached the 12-sound sequences, they may have exceeded their working memory capacity, leading to a decrease in performance.

To test if there was a statistically significant difference in accuracy based on whether the mapped words were related to the sounds (the relationship was intuitive), further paired samples t-tests were performed. There was a statistically significant difference between related ($M = .42$, $SD = .10$) and unrelated sound ($M = .32$, $SD = .13$) conditions; $t(19)=3.4$, $p = .001$. (See Figure 6.)

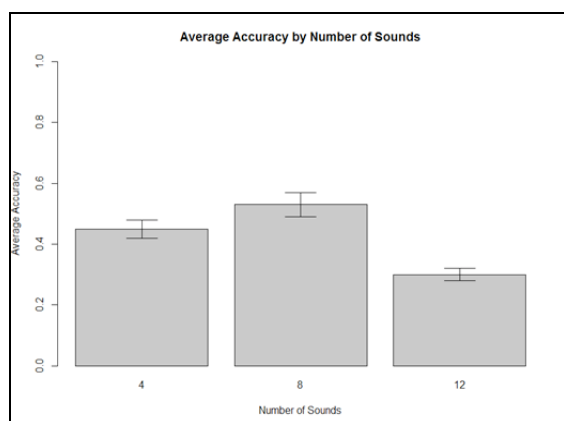


Figure 5. Means and standard errors of accuracy by number of sounds.

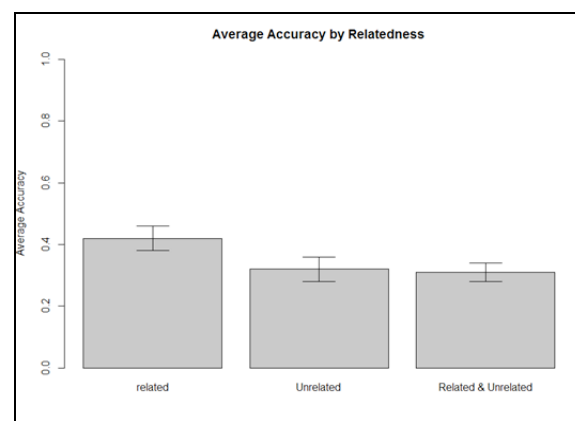


Figure 6. Means and standard errors of accuracy by sound relatedness.

4. DISCUSSION

While limited by design, the current study yielded several significant results that prompt further experimentation in the area of auditory streams and memory. The results concluded that eight sounds was the optimal length when compared to sequences of four or 12 sounds, regardless of sound type and relatedness. Furthermore, related sounds were recalled significantly more often than unrelated sounds or mixed related/unrelated sound sequences. Contrary to our second hypothesis, participants did not always perform better when sequences contained a variety of sound types; in certain cases, they performed as well with a single sound type.

The present results showed that lyricons and spearcons are recalled more frequently than music and earcons when people are given a short amount of time to remember many sounds. This is likely due to the fact that both are based on speech and, therefore, are more intuitive and processed more easily than are the arbitrary mappings of earcons and music. One limitation of the current study is that the design did not enable us to accurately measure training time for individual sounds so that we could compare training time for the various sound types.

Lyricons and spearcons also have the highest learnability; these findings suggest that future experimentation could separate them from music and earcons, and examine sequences and streams containing mixes of lyricons and spearcons. In addition, relatedness of sounds is important for user recall. Sounds should be related to the meaning they are conveying to the user; inter-rater reliability proved effective in the current study to judge relatedness of sounds.

Another limitation of this study is that the audio sequences used are not necessarily representative of what a person would encounter in an actual task. Future research is needed to create streams that convey an overall meaning to the participant, instead of sequences that contain related and unrelated individual word meanings. The present study identifies that certain types of sounds perform better in sequences and relatedness is a key component to memorability and learnability. Further experiments should create streams that can be used in different task environments, such as surgical operating rooms. Research is needed on the effectiveness of multiple streams being played at once, and the effectiveness of streams containing sounds that vary in dimensions such as pitch, rhythm, and register.

5. CONCLUSIONS

The present study identified that lyricons and spearcons have the highest recall when presented in varied sound sequences. Relatedness of meaning and sound is also a key component in memorability and learnability. Further research is needed to examine streams containing only lyricons and spearcons, and to integrate these sound sequences into streams.

6. REFERENCES

- [1] S. A. Brewster, P. C. Wright, and A. D. Edwards, "An evaluation of earcons for use in auditory human-computer interfaces," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, pp. 222-227, April 1993.
- [2] M. Watson and P. Sanderson, "Sonification supports eyes-free respiratory monitoring and task time-sharing," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 3, pp. 497-517, 2004.
- [3] T. Dingler, J. Lindsay, and B. Walker, "Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech," in *Proceedings of the 14th International Conference on Auditory Display*, 2008.
- [4] A. Bregman and J. Campbell, "Primary auditory stream segregation and perception of order in rapid sequences of tones," *Journal of Experimental Psychology*, vol. 89, no. 2, pp. 244-249, 1971.
- [5] J. H. Schuett, R. J. Winton, J. M. Batterman, and B. N. Walker, "Auditory weather reports: demonstrating listener comprehension of five concurrent variables," in *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*, ACM, p. 17, 2014.
- [6] P. A. Lucas, "An Evaluation of the Communicative Ability of Auditory Icons and Earcons," in *Proceedings of the 2nd International Conference on Auditory Display*, Addison-Wesley, 1994.
- [7] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, no. 2, pp. 250-267, 1993.
- [8] J. H. Schuett and B. N. Walker, "Measuring comprehension in sonification tasks that have multiple

- data streams,” in *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*, ACM, p. 11, 2013.
- [9] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, “Earcons and icons: Their structure and common design principles,” *Human-Computer Interaction*, vol. 4, pp.11–44, 1989.
- [10] M. Jeon, “Lyrics (Lyrics+ Earcons): Designing a new auditory cue combining speech and sounds,” in *Proceedings of the International Conference on Human-Computer Interaction*, Springer Berlin Heidelberg, pp. 342-346, 2013.
- [11] S. T. Mueller and B. J. Piper, “The Psychology Experiment Building Language (PEBL) and PEBL Test Battery,” *Journal of Neuroscience Methods*, vol. 222, pp. 250–259, 2014.
- [12] S. T. Mueller, The Psychology Experiment Building Language, Version 0.14. Retrieved from <http://pebl.sourceforge.net>, October 2016.
- [13] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon, “Spearcons (Speech-based earcons) improve navigation performance in advanced auditory menus,” *Human Factors*, vol. 55, no. 1, pp. 157-182, 2013.