MANIPULATING SYNTHETIC VOICE PARAMETERS FOR NAVIGATION IN HIERARCHICAL STRUCTURES

Peer Shajahan

Department of Computer Science
University of Manitoba,
Winnipeg, Canada
peermoh@cs.umanitoba.ca

Pourang Irani

Department of Computer Science
University of Manitoba,
Winnipeg, Canada
irani@cs.umanitoba.ca

ABSTRACT

Auditory interfaces commonly use synthetic speech for conveying information. In many instances the information being conveyed is hierarchically structured, such as menus. In this paper, we describe the results of one experiment that was designed to investigate the use of multiple synthetic voices for representing hierarchical information. A hierarchy of 27 nodes was created (in which 2 of the nodes were not shown to the participants during the training session). A between subjects design (N-16) was conducted to evaluate the effect of multiple synthetic voices on recall rates. Two different forms of training were provided. Participant's tasks involved identifying the position of nodes in the hierarchy by listening to the synthetic voice. The results suggest that 84.38% of the participants recalled the position of the nodes accurately. The results also indicate that multiple synthetic voices can be used to facilitate navigation hierarchies. Overall, this study suggests that it is possible to use synthetic voices to represent hierarchies.

1. INTRODUCTION

In recent years, the mode of conveying information using auditory interfaces has become commonplace. Airport messages, cell phones and Interactive Voice Response (IVR) systems are some examples of systems that take advantage of synthetic voices. For example, in IVR systems, users can access information, such as finding bank details, movie listings and business directories. These systems use synthetic speech as the main medium for information output. In general, synthetic speech is used for presenting information where the graphical display is not available or, limited (such as in telephone-based interfaces, for assisting visually impaired users, in PDAs and in Pocket PCs) and where the eyes and hands are busy (such as while driving a car and while playing video games). In these instances, users heavily rely on voice-based applications in order to acquire the information.

Although the use of voice-based applications has increased, navigation in such interfaces is still considered as a significant design challenge. This is partly due to the underlying navigation structure that is not explicit or visible to the user. For example, in many applications users enter the system from the root node and branch into various dialogs (nodes and levels) in order to extract the necessary information. Since, voice-based applications do not explicitly depict the arrangement of nodes (i.e. the parent-child relationship between dialogs), users can lose track of their position in the navigation hierarchy. This can often lead to frustration [1].

In some respects, navigation can be generalized as the task that requires locating and branching to the appropriate path that leads to the object of interest. Navigation can be facilitated by providing cues about the user's location in the navigation structure. In this paper, we present a method for assisting a user in identifying the location of nodes in the underlying navigation structure of an auditory interface. We describe the results of an experiment which analyzes the use of multiple synthetic voices for assisting users in comprehending hierarchical structures.

2. RELATED WORK

This work is primarily inspired from the literature on earcons and their application to auditory interfaces. We are particularly interested in identifying whether multiple synthetic voices can assist users in recognizing elements organized in a hierarchy. We first describe the results of the literature relevant to this work.

2.1. Supporting Navigation in Hierarchies Using Earcons

The issue of navigation in auditory interfaces has been addressed primarily through the implementation and evaluation of non-speech audio, called "earcons". Blattner et al [2] defined earcons as "abstract, synthetic tones that can be used in structured combinations to create sound messages for representing parts of an interface". Earcons are created by manipulating non-speech audio parameters, such as pitch, rhythm, timbre, register and volume. Initially, Brewster [3] suggested that navigation in hierarchical auditory interfaces can be improved by creating hierarchical earcons. The results of this study suggested that participants could recall 81.5% of the earcons. However, increasing the size of the hierarchy increased the complexity of the earcons.

Brewster et al [4] designed a study to evaluate the use of compound earcons for representing large hierarchies. Compound earcons were constructed by initially creating a set of sounds and then concatenating these sounds (i.e. creating various combinations of these sounds) according to the number of levels and nodes in the hierarchy. The results of this study suggest that 97% of the earcons were recalled accurately.

Although, compound earcons are effective in conveying hierarchical information, they possess some limitations. First, the length of the earcons (in time) increases with respect to the number of levels in the hierarchy. Second, the users have to listen to the entire earcon in order to locate the position in the hierarchy. Hence, when these earcons are implemented in speech-based interfaces (where speech is considered as the dominant mode of interaction) users have to listen to the entire earcon in order to navigate the hierarchy. This in-turn increases the overall navigation time. Considering the drawback of using compound earcons in large hierarchies, we investigate the possibility of manipulating the synthetic speech to represent the

hierarchy. We first outline the work that inspired the idea of producing various types of voices by modifying the parameters used for creating of synthetic speech.

2.2. Characteristics of Synthetic Speech in Speech-Based Interfaces

In recent years, the use of synthetic speech in speech-based interfaces has increased considerably. This is because speech-based interfaces are measured based on the intelligibility of the synthetic speech [5]. A study by Beutnagel et al [6] suggests that the currently available TTS systems offer a high level of intelligibility, up to 97% (whereas the intelligibility rate of human speech is 99%).

In addition to the fact that the intelligibility of synthetic speech is equivalent to the intelligibility of human speech, various studies have been conducted to analyze the comprehensibility and the perception of synthetic speech. Lai et al [7] suggest that the comprehensibility of synthetic speech (67%) is approximately equivalent to the comprehensibility of human speech (73%). Another line of research [8, 9] aimed at analyzing users' perception with speech-based interfaces, by manipulating synthetic speech parameters. Brave and Nass [10] suggest that emotions (such as happiness, sadness, anger, etc.) can be created by manipulating synthetic speech parameters, such as pitch, speech rate, stress, pause and volume. Another study by Nass et al [11] suggests that voices representing various personality traits, such as introverts and extroverts, can also be created by manipulating the synthetic speech parameters. The results of these studies suggest that various synthetic voices can be created by manipulating synthetic speech parameters.

2.3. Mixing Synthetic Speech and Earcons in Interfaces

Vargas and Anderson [12] proposed to mix synthetic speech and earcons in order to improve navigation in speech-based interfaces. In their study, participants were divided into two groups, where the first group performed tasks with a speech only condition, and the second group performed tasks with a speech and earcons condition. The results of this study suggest that participants perform node finding tasks comparatively better with the mixing approach than with the speech only condition.

In general, studies by Brewster [3], Brewster et al [4], Vargas and Anderson [12], and Leplâtre, and S. A. Brewster [13] suggest that speech-based interfaces should, in addition to providing information to the users, also assist with the task of navigation. This can be achieved by providing additional cues in these interfaces. We propose to use multiple synthetic voices for representing hierarchies, by assimilating the results of [10, 11]. The motivation behind this approach is in the belief that if synthetic speech can by itself provide both information and navigational cues to users, then a significant amount of navigational time may be reduced.

2.4. Previous investigation on multiple synthetic voices

A recent study by Shajahan and Irani [14] examined the possibility of using multiple synthetic voices to represent hierarchies. Small hierarchies of 10 nodes with 3 levels were created. A within-subjects (N=10) design was conducted to compare the effect of multiple synthetic voices to single synthetic speech. Three synthetic voices parameters (average pitch, pitch range and speech rate) were manipulated to create multiple synthetic voices. These voices were then hierarchically related to represent the nodes in the hierarchy. Participants were trained with the list of rules that were used to create the multiple synthetic voices. The tasks of the participants involved locating the position of the node in the hierarchy by listening to the played voice. The results of this study showed that the participants performed the node finding tasks approximately four times better with multiple synthetic voices than with single synthetic voices.

Although the results of this study suggest that multiple synthetic voices can be used to represent small hierarchical structures, the study did not suggest whether multiple synthetic voices can be used to represent complex hierarchies. Since, these hierarchies contained only a small number of items, we could not conclude that multiple synthetic voices actually facilitated representation of the hierarchical structure. Users may have possibly memorized the location of the nodes during the training phase of the experiment. A follow-up study described here is therefore designed to evaluate the use of multiple synthetic voices on complex hierarchies.

3. EXPERIMENT

This purpose of this experiment was to determine whether multiple synthetic voices can be used to support navigation in complex hierarchies. A hierarchy of about 27 nodes (including A and B) and 4 levels was created to analyze the effect of multiple synthetic voices on larger hierarchies (2.7 times larger than the hierarchy that was previously tested by Shajahan and Irani [14]). Figure 1 shows the structure of the hierarchy used in this experiment, denoting the files and folders on a computer system.

3.1. Guidelines for Manipulating Synthetic Voices

DECtalk, version 4.61 (from Fonix Corp., www.fonix.com) was used to create multiple synthetic voices for presenting hierarchical information. In this study, we created multiple synthetic voices by following the two guidelines, "duplication" and "variation", that were suggested by Shajahan and Irani [14]. They have suggested that using these two guidelines, multiple synthetic voices can be created for representing a small hierarchy of about 10 nodes.

1. **Duplication:** Duplicate all the synthetic voice parameters and their values from the parent node. For instance, if a parent node is created with two parameters, speech rate (with value = 110 words-per-minute) and average pitch (with value = 120 Hz), and if the voice for the child node is created with the same parameter values that are used in the parent node, then it is referred to as duplication.
2. **Variation:** Alter the values of one or more synthetic speech parameters between two related nodes. For example, if a parent node is assigned a speech rate of 110 words-per-minute, then its child can be assigned a speech rate of 150 words-per-minute.

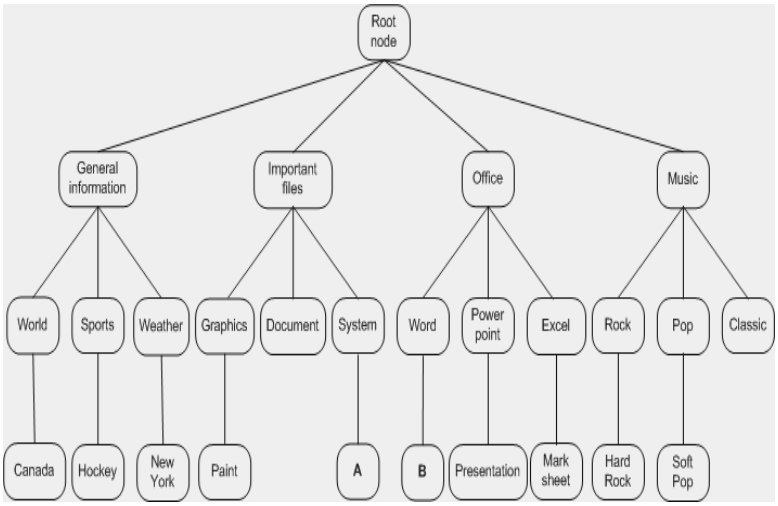

Figure 1 Hierarchy used for testing the effect of multiple synthetic voices on complex hierarchies

However, for representing complex hierarchies, we have devised a new rule called Inclusion. This rule was devised based on the guidelines suggested by Sumikawa et al [15]. Sumikawa et al [15] suggested that earcons can be created for representing complex hierarchies by following the three guidelines: repetition, variation and contrast. We extended the third guideline (contrast) and devised a new rule (inclusion) for creating additional voices for complex hierarchies.

3. **Inclusion:** Include one or more speech parameters to the preceding voice in order to create unique voices. For example, if the child node duplicates all the parameters and the exact values from its parent node (such as speech rate and pitch), then other voice parameters such as laryngealization and breathiness can be added (inclusion rule) to the child node, to make the child node sound different from its parent node.

3.2. Parameters Used for Creating Multiple Synthetic Voices

In the study by Shajahan and Irani [14], multiple synthetic voices were created by manipulating the speech parameters, such as Average Pitch (AP), Speech Rate (SR) and Pitch Range (PR). Average pitch is used to raise or lower the pitch contour for a given synthetic speech. Average pitch is measured in Hz. Speech rate defines the number of words that are spoken per minute (WPM). Pitch range is used to expand or shrink the swings in the pitch. Pitch range was measured in percentage (%), for example if PR=100, then there was no change in the pitch. In addition to using average pitch, speech rate and pitch range, in this study we have also manipulated other speech parameters such as laryngealization, breathiness and gain in volume to create voices for representing hierarchies.

- **Laryngealization (LA):** At the beginning and end of sentences, many speakers turn their voice on and off irregularly. This gives a querulous tone to the voice. This departure from perfect periodicity is called creaky voice quality, which is often referred to as laryngealization. The LA option specifies the amount of laryngealization, in the voice. The value LA=0 specifies (no laryngealization) and LA=100 specifies (maximum laryngealization).
- **Breathiness (BR):** Some voices can be described as breathy. The vocal folds vibrate to generate a breathy noise along with the voice. BR option ranges from 0 dB (no breathiness) to 70 dB (strong breathiness).
- **Gain in Volume (GV):** GV option gives information on the intensity (volume) of the voice. GV ranges from 0 dB (no volume) to 70 dB (full volume).

3.3. Rules for Creating the Hierarchy

The rules that were used to create the hierarchy are described below (Figure 2).

- Nodes belonging to the third level were created by inheriting the same pitch (AP and PR) from their respective parent nodes. The distinctive feature between the second level nodes and the third level nodes was the difference in the speech rate, i.e. for the second level nodes SR=220 wpm was used and for the third level nodes SR=160 wpm was used. At this level, the nodes that belonged to the same family were differentiated using speech rate (SR), Laryngealization (LA) and Breathiness (BR). For all the nodes in the third level, all the parameters and their values, except the speech rate, were inherited from the parent node. The speech rate was set at SR=160 wpm. In addition to these features,
  - For the middle node, a new parameter LA=50 was also added (inclusion rule).
  - For the right node, new parameters (BR=55 dB and GV = 55 dB) were added (inclusion rule) to distinguish this node from the other nodes.
- Nodes that belonged to the forth level were created by inheriting all the parameters and their values, except the speech rate, from the parent node. The speech rate used in this level was SR = 100 wpm.

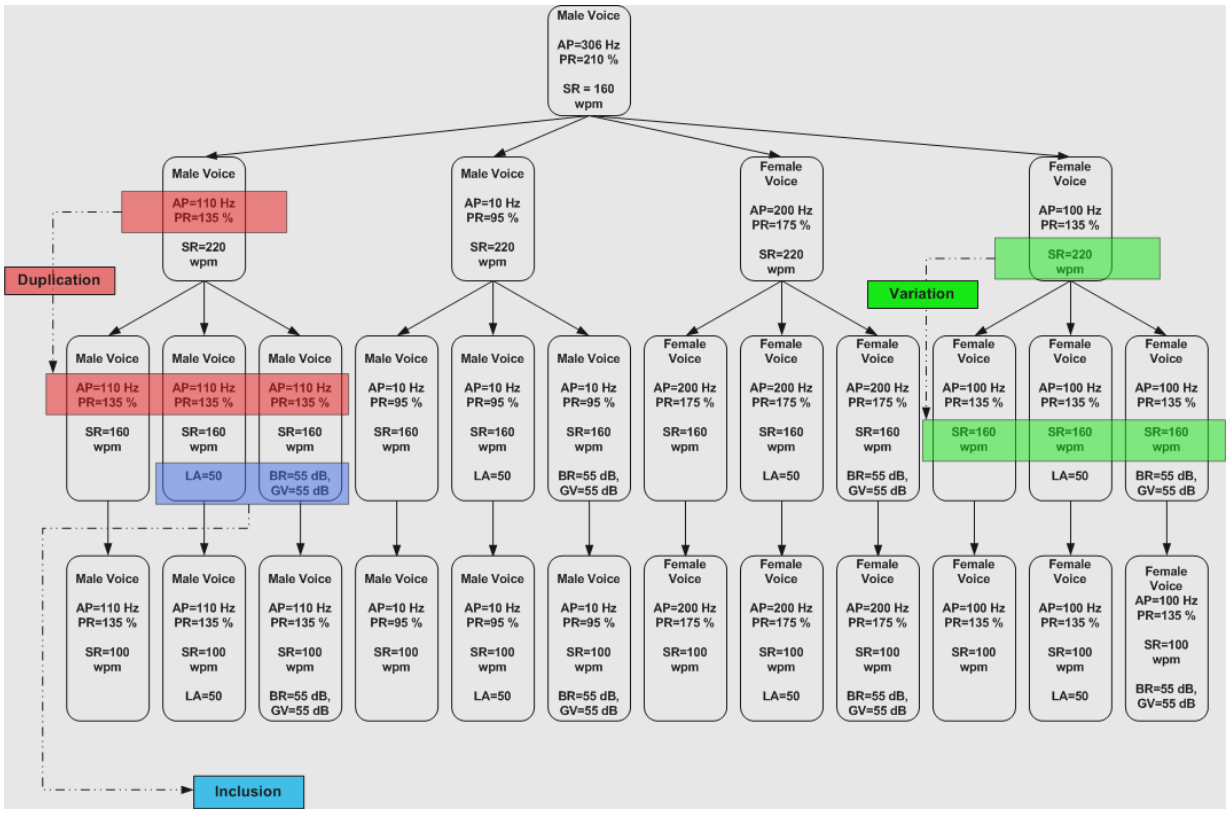

Figure 2 Rules used for creating the hierarchy: Duplication (duplicate exactly the value of a preceding synthetic voice parameter), Variation (alter the values of one or more synthetic speech parameters between two related nodes) and Inclusion (include one or more speech parameters to the preceding voice in order to create unique voices).

3.4. Hypotheses

Based on the results of Brewster [3] and Shajahan and Irani [14], we anticipated the following results:

- ❶ **Hypothesis 1:** Participants should be able to perform the node finding tasks with high accuracy by listening to the multiple synthetic voices, if the rules that were used to create the multiple synthetic voices were easy to remember and recall.
- ❷ **Hypothesis 2:** Participants should also be able to locate the position of the unheard voices (nodes A and B) in the hierarchy by recalling the rules that were used to construct the multiple synthetic voices.

3.5. Design

3.5.1. Method

A two-condition (group-1 and group-2) between-subject experiment was conducted, where the type of training varied between the groups. The participants were selected and randomly assigned to one of the two groups. The accuracy rate for locating nodes in a hierarchy was measured.

3.5.2. Materials

In this experiment, the hierarchies were presented using PowerPoint files, with appropriate .wav files (sample rate = 11.025 KHz, 16-bit Mono) on the various nodes. The PowerPoint slides were shown using a Dell Inspiron 8600 laptop with a 15.4" display using Intel® Integrated laptop audio speakers. Lower quality voice resolution was used in this experiment in order to simulate the output of several real-time environments, such as those found in TBIs.

3.5.3. Participants

16 students from a local university volunteered for this study. In order to avoid any potential difficulty in understanding synthetic speech, we selected only native English speakers who did not exhibit any auditory disorder.

3.5.4. Training

In this experiment, the type of training received by the participants varied between the groups. At the start of the experiment, all the participants were given a write-up, which explained the rules that were used to create multiple synthetic voices for presenting hierarchical information. During the training session, the participants were shown the structure of the

hierarchy (Figure 1) and were allowed to click on a node to listen to its associated synthetic voice.

Participants belonging to group-1 received the training directly from the experimenter. The experimenter showed the hierarchy to the participants and explained the rules that were used to represent the hierarchical information. Participants belonging to group-2 were given three minutes to learn the rules by themselves, and did not receive any help from the experimenter during the training session

3.6. Task

The experiment began when participants felt comfortable listening to the nodes in the system. During this experiment, fourteen voices were randomly selected and played to the participants. These voices were selected from the set of 29 possible voices, in which 3 voices were from level-2 (where root node is at level-1), 5 from level-3 and the remaining 6 were from level-4. Out of the fourteen voices, two voices (A and B) were new voices, which had not been shown or played to the participants during the training session (see Figure 1).

After playing each voice, the participants were asked to answer a basic question: "locate the position of the node in the hierarchy". The participants were given a sheet, which included a hierarchy, similar to Figure 1. However, the labels associated with the nodes in the hierarchy were removed, in order to avoid any confounding effects. The participants were asked to label the nodes in the hierarchy based on the order of the presentation (i.e. the first node played was labeled 1, etc.).

The labeling provided by each subject was used to measure the accuracy rate. Upon completion of the first twelve questions (Q1-Q12), the participants were informed that Q13 and Q14 are new voices and had not been shown to them during the training session. After playing Q13 and Q14, the participants were asked to identify the position of these two voices in the hierarchy.

3.7. Results and Discussion

The overall results show that 84.38% of the voices (nodes) were recalled accurately by the participants. Consistent with hypothesis-1, the results suggest that the rules that were used to create multiple synthetic voices were easy to remember and recall. The results also support hypothesis-2 and reveal that the position of unheard voices was located by the participants with an accuracy rate of 96.88% (A=100%, B=93.75%) (Figure 3).

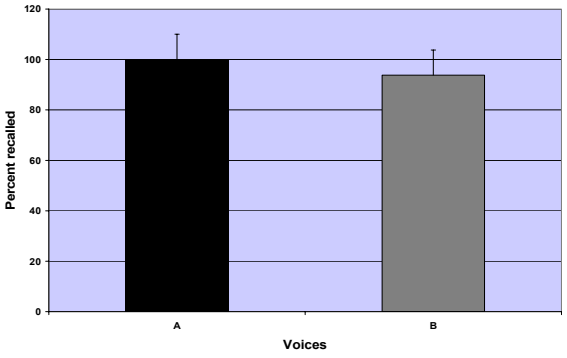

Figure 3 Recall rates of unheard voices

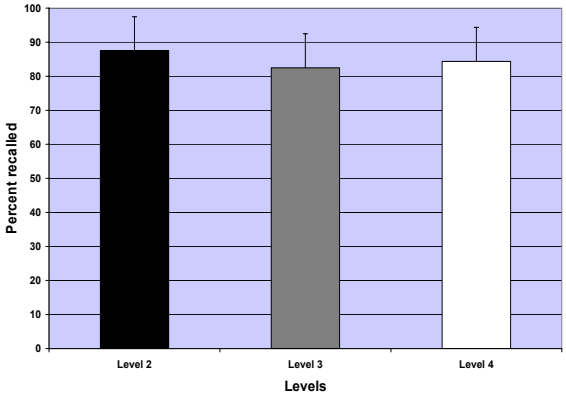

Figure 4 Recall rates of voices at each level

In order to analyze in which level of the hierarchy the most number of errors had occurred, the tasks (voices) were grouped based on the location (different levels) of the voices in the hierarchy, i.e. out of 14 questions, 3 are from level-2, 5 from level-3 and 6 from level-4. The overall recall rates of multiple synthetic voices at each level are shown in Figure 4.

The results as summarized in Figure 4 show that the errors occurring at each level are evenly distributed. Hence, we have decided to analyze the recall rates for each family (sub-tree), where Family-1 consists of nodes inclusive and belonging to "General Information", Family-2 referring to the sub-tree "Important Files", etc… (Figure 1). The overall recall rates for each family are shown in Figure 5. From Figure 5, we can clearly say that the highest number of errors occurred in Family-4. This may due to the fact that the voices in Family-4 were not highly distinct. Therefore, better parameter values have to be chosen to represent the nodes in Family-4, in order to improve the accuracy rate.

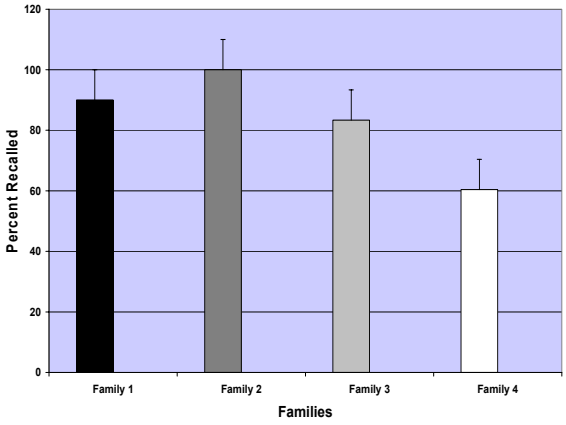

Figure 5 Recall rates of voices for each family

Overall, the results of this study corroborate with our hypotheses that multiple synthetic voices can be used to present hierarchical information. However, additional investigation is required to identify the best manipulation of parameters for representing wide and deep hierarchies.

4. CONCLUSIONS AND FUTURE WORK

An experiment was conducted to analyze the effect of multiple synthetic voices on complex hierarchies. A hierarchy of 27 nodes was created similar to the hierarchy that was used to evaluate the effect of earcons [3, 4]. Multiple synthetic voices were used to represent nodes in the hierarchy. These voices were created by manipulating speech parameters (such as average pitch, pitch range, speech rate, laryngealization, breathiness and gain in volume). The voices were related in a hierarchical manner by applying the three guidelines: duplication, variation and inclusion. The results of this study showed that participants recalled 84% of the voices accurately. The results also suggested that 96.88% of the unheard voices were located accurately, and the effect of training did not have a significant effect on the recall rates.

It is interesting to compare the difference in recall rates for multiple synthetic voices to those for earcons in a similar study [3]. In the study designed by Brewster, the recall rate for hierarchical earcons was approximately 81.5% and for compound earcons was 97% [3]. The results of our experiment fall in between these two values (84%). However, for unheard earcons, users were able to locate their position in the hierarchy with an accuracy of 91.5% in comparison to 97% with multiple synthetic voices. A more rigorous experiment could be designed to compare earcons to multiple synthetic voices on dimensions other than locating elements in hierarchies, such as whether users are able to comprehend information better or quicker with one system over the other.

The results also reveal that the rules that were used to create the multiple synthetic voices were easy to remember and recall. However, in this study we have not established the set of manipulation rules that facilitate the best performance. An additional study will be designed to evaluate the effectiveness of multiple parameter configurations for representing depth and width in hierarchies.

In future, we are planning to evaluate the effect of multiple synthetic voices on more complex hierarchies, such as those found in file systems. We believe that this could be done either by integrating earcons and multiple synthetic voices to represent the nodes in the hierarchy or by making use of other vocal cues such as richness, smoothness, etc.

We are also planning on implementing and evaluating the effect of multiple synthetic voices in real-time applications that contain hierarchical structures, such as in video games and TBIs. We believe that this approach will provide significant insight to designers once we are able to devise a firm set of guidelines for manipulating the various synthetic voice parameters.

5. ACKNOWLEDGEMENTS

We thank Fonix Corporation for providing the DECtalk development toolkit for this project. We also extend our appreciation to Jennifer Lai from IBM Watson Research Labs for her guidance in the initial phases of the study and Dean Slonowsky for helping with the statistical analysis. This project is supported by an NSERC grant.

6. REFERENCES

[1] C. Wolf, L. Koved, and E. Kunzinger. Ubiquitous mail: Speech and graphical user interfaces to an integrated voice/E-mail mailbox. In Interact, pp 247–252, 1995.

[2] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg. Earcons and icons: Their structure and common design principles. Human Computer Interaction, 4(1):11– 44, 1989.

[3] S. A. Brewster. Using nonspeech sounds to provide navigation cues. ACM Transactions on Computer-Human Interaction (TOCHI), 5(3):224–259, 1998.

[4] S. A. Brewster, A. Capriotti, and C.V. Hall. Using compound earcons to represent hierarchies. HCI Letters, 1(1), pp 6–8, 1998.

[5] T. Dutoit. An Introduction to Text-to-Speech Synthesis, volume 3. Kluwer Academic Publishers, 1997.

[6] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-gen TTS system. In Joint Meeting of ASA, EAA, and DAGA, pp 15–19, Berlin, Germany, 1999.

[7] J. Lai, D. Wood, and M. Considine. The effect of task conditions on the comprehensibility of synthetic speech. In Proceedings of the SIGCHI conference on Human factors in computing systems, pp 1–6, Seattle, Washington, United States, 2000. ACM Press.

[8] C. Nass, E. Robles, H. Bienenstock, M. Treinen, and C. Heenan. Voice-based disclosure systems: Effects of modality, gender of prompt, and gender of user. International Journal of Speech Technology, 6(2):113–121, 2003.

[9] L. M. Slowiaczek and H. C. Nusbaum. Effects of speech rate and pitch contour on the perception of synthetic speech. Human Factors, 27(6):701–712, 1985.

[10] S. Brave and C. Nass. Emotion in human-computer interaction. pp. 251–271 in J. Jacko & A. Sears (Eds.), Handbook of human-computer interaction. New York: Lawrence Erlbaum Associates, 2002.

[11] C. Nass and K. Lee. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied, 7(3):171–181, 2001.

[12] M. L. M. Vargas and S. Anderson. Combining speech and earcons to assist menu navigation. In Proceedings of the 2003 International Conference on Auditory Display. Addison-Wesley, 2003.

[13] G. Leplâtre, and S. A. Brewster. An Investigation of Using Music to Provide Navigation Cues. In Proceedings of the International Conference on Auditory Display. Addison-Wesley, 1998.

[14] P. Shajahan and P. Irani. Representing hierarchies using multiple synthetic voices. In 8th International Conference on Information Visualisation, pp 885–891. IEEE Computer Society, 2004.

[15] D. A. Sumikawa. Guidelines for the integration of audio cues into computer user interfaces. Technical Report UCRL 53656, Lawrence Livermore National Laboratory, Livermore, California, United States, 1985.