Representing Hierarchies Using Multiple Synthetic Voices

Peer Shajahan, Pourang Irani
Department of Computer Science
University of Manitoba, Winnipeg, Canada
{peermoh@cs.umanitoba.ca, irani@cs.umanitoba.ca}

Abstract

This paper reports on ongoing work related to the representation of hierarchical structures using multiple synthetic voices. We manipulated three synthetic voice parameters, average pitch, pitch range and speech rate, to represent nodes in a hierarchy. We created hierarchies containing 10 nodes and three levels deep. A within-subjects design (N=12) was conducted to compare the effect of multiple synthetic voices to single synthetic voices for locating the positions of items in a hierarchy. Subjects were trained with the set of rules we used for constructing the multiple synthetic voices. In a node-finding task, participants identified the position of a previously listened-to node. Our results show that subjects recalled the nodes' positions in the hierarchy significantly better when the hierarchies were equipped with multiple synthetic voices than without.

Keywords--- **Hierarchies, synthetic voice, multiple synthetic voices, telephone-based interfaces, navigation, non-speech audio**.

1. Introduction

Hierarchies are commonly used for structuring and organizing complex information. For example, file structures, disk trees, and many interface elements, such as menus, are organized into hierarchical structures. To extract information organized into hierarchies, interfaces have to adequately facilitate navigational tasks. This is particularly important in interfaces, where the graphical feedback is not available, such as on cell phones, in telephone-based interfaces (TBIs) and in devices for assisting the visually impaired.

In recent years, interfaces with voice capabilities or voice user interfaces (VUIs) have gained popularity. VUIs are used where access to high information bandwidth (i.e. visual displays) is impractical and the task of querying and delivering information is complex. VUIs primarily support tasks where the eyes and hands are busy (such as in driving), and where verbal interaction is the most effective medium of communication (such as in assisting the visually disabled). In some applications, voice output capabilities are essential, for example, in telephony interfaces, on cell phones or in video games. In these instances, the voice is created using synthetic speech, often called computer generated speech.

Recently, synthetic speech has gained widespread popularity as a result of an increased use of automated telephony systems or Interactive Voice Response (IVR) systems. Using IVR systems, end-users are able to perform their day-to-day banking transactions, retrieve ticketing information, lookup driving directions, and listen to movie listings at their convenience. In such interfaces, options (or menus) are presented in an implicit hierarchical fashion. Users typically branch through several nodes in the hierarchy in order to find the element of interest.

Several studies have reported on the use of non-speech audio for emphasizing the hierarchical structure of an interface [4, 10]. The purpose of our investigation is to determine whether we can associate hierarchies in an interface with multiple synthetic voices. We postulate that if this is possible, interfaces represented using multiple synthetic voices could facilitate and improve navigation and data querying performance.

2. Related Work

Our work is inspired by two large bodies of research in the area of audio based interfaces: non-speech audio and synthetic speech. We first applied some of the design principles for developing hierarchies using non-speech audio or earcons. We then considered the results of several studies suggesting that synthetic voices can be created by manipulating synthetic voice parameters to affect users' perception and performance with a given speech interface.

2.1. Using Earcons to Represent Hierarchies

Earcons are a form on non-speech audio. Blattner et al. [1] define earcons as, "abstract, synthetic tones that can be used in structured combinations to create sound messages for representing parts of an interface". Brewster et al. [4] investigated the use of earcons for representing hierarchies. They manipulated several audio parameters such as pitch, intensity, timber, and rhythm in structured combinations. In their initial investigation they limited the size of the hierarchies to a maximum of 9 nodes. Their results show that participants could accurately recall 80% of the nodes' positions in the trees by listening to the structured earcons. In another study,

Brewster [2] tested the use of earcons to improve navigation in TBIs. His results suggest that earcons can be used as good markers for improving navigation in TBIs.

A recent study by Vargas and Anderson [16] suggests that earcons can be combined with synthetic speech to improve navigation in menu-based interfaces. In their study, participants were divided into two groups. One group was given the interface with only the synthetic speech, whereas the other group was presented an interface with a combination of synthetic speech and earcons. During the training session, participants were shown a graphical representation of hierarchical menus and were allowed to click and listen to items in the menus. In the experiment, participants' were required to locate items in the menus. Their results suggest that the tasks were performed better by the subjects using the combination of synthetic speech and earcons than by those using the synthetic speech only.

Although earcons are adequate in providing navigation cues, they have several disadvantages. When used in conjunction with speech-based interfaces, earcons are played in the background. This requires that users listen to the earcons in addition to the synthetic speech, before moving between nodes in a hierarchy. Thus, in a speech-based interface users would need to listen to both, the synthetic voice and the earcons, in parallel, in order to identify their location in the hierarchy [2, 3, 16]. This in turn increases the workload and navigation time of users.

The results from these studies suggest that non-speech audio can be constructed to enhance navigation in hierarchies. However, in speech-based interfaces this may be achieved at the cost of overloading the user with additional information.

2.2. Characteristics of Synthetic Speech

Synthetic speech is characterized by various speech parameters, such as average pitch, pitch range, loudness, speech rate, breathiness, pauses and stress.

Several studies have been conducted to analyze the effect of modifying the parameters of synthetic speech on user's performance under various conditions and applications. Most of these studies focused on analyzing the effect of synthetic speech on cognitive variables (such as comprehension and adaptation rates) [6, 7, 8, 13]. By manipulating synthetic voice parameters, other investigations have studied the effect of personality types, emotions, and gender on users' perception and performance with speech interfaces [5, 9, 11, 12].

A study by Lai et al. [8] shows that the comprehensibility of synthetic speech is approximately equivalent to the comprehensibility of human speech (67% for synthetic speech and 73% for human speech). The results of this study confirmed the studies [6, 14] that people are able to recognize and adapt to synthetic speech rapidly, i.e. they will quickly learn the perceptual and cognitive strategies that are necessary to improve their comprehension accuracy. Studies by Cahn [5] and Nass et al. [11, 12] suggest that social, psychological, and emotional impressions can be created by manipulating synthetic voice parameters such as pitch, pause, volume, and speech rate.

In general, these studies suggest that manipulating various synthetic voice parameters can influence our perception of the information being presented. However, the ability to manipulate synthetic speech parameters to produce voices that are related in a hierarchical manner has not, to the best of our knowledge, been investigated. If this is possible, then such voices can be used to represent hierarchies in interfaces, where speech is the dominant modal dimension. In the following sections, we first describe the rules that are used for constructing multiple synthetic voices. We then describe the experimental setup that assessed whether our constructed synthetic voices could be used to identify components of a hierarchy.

3. Construction of Multiple Synthetic Voices

To create multiple synthetic voices we used DECTalk version 4.61 (from Fonix Corp., www.fonix.com). DECTalk facilitates the creation of synthetic voices by modifying various synthetic speech parameters (listed in section 2.2). The DECTalk toolkit uses a formant-based synthesizer (uses linguistic rules to generate speech) for producing text-to-speech. At this stage of our research, we used a formant-based engine instead of a concatenative engine (uses human recordings to generate speech) for two main reasons. Firstly, concatenative text-to-speech systems are developed by only a few companies and do not facilitate an easy modification of voice parameters (such as average pitch, pitch range, and breathiness). Secondly, since concatenative speech is perceived as being more natural, we should be able to apply our principles and obtain similar results.

We constructed multiple synthetic voices based on the guidelines that were used in [3] to create hierarchical earcons. To create earcons that are related in a hierarchical manner, Sumikawa et al. [15] proposed three design principles: duplication, variation, and contrast. We extended these basic rules for creating various structured combinations of synthetic speech parameters. We devised the following two guidelines (MSSG stands for Multiple Synthetic Speech Guideline):

- **MSSG1 - Duplication:** Duplicate exactly the value of a preceding synthetic voice parameter. For instance, if a node is assigned a speech rate of 110 words-per-minute, then a child node can be created containing this same value for its speech rate parameter.
- **MSSG2 - 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.

Using the two guidelines described above, we manipulated several synthetic voice parameters to convey hierarchical relationships. We first describe the parameters we selected and then explain our rationale in selecting these parameters. The various voice parameters used for creating multiple synthetic voices to represent nodes in the hierarchies are:

- **Average Pitch (AP):** defines the variation in the pitch contour for a given synthetic voice. For example, increasing the average pitch may result in agitation, where as reducing the pitch will result in calmness of the speaker. Average pitch is measured in Hz.
- **Pitch Range (PR):** is used to expand or shrink the swings in pitch. Using pitch range emotions in a voice can be perceived. For example, increasing the pitch range will increase the level of dynamism projected by the voice. In turn this could lead to a perception of happiness in the voice. Reducing the pitch range will project the image of sadness in the speaker. Pitch range is measured in Hz.
- **Speech Rate (SR):** is defined as the number of words that a system can speak in one minute. It is measured in terms of words-per-minute (wpm).

Some of the reasons for choosing the above discussed voice parameters to create hierarchical relationships between items are as follows:

- Several studies suggest that average pitch, pitch range, and speech rate are the three main voice parameters that play an important role in extending personalities and emotions in speech interfaces [5, 11, 12].
- Other voice parameters, such as breathiness and laryngealization, do not allow us to distinctly differentiate between voices.
- Although, other parameters such as quality of voice can be used for differentiating between multiple voices, their utility in real-world applications may not be practical.

The choice of values for each voice parameter was based on how strong a contrast between two different voices was achievable. In general we assigned values to the parameters in a structured and systematic manner. We first created voices with the lowest and highest possible values for each of the three parameters without degrading users' comprehension of the text being vocalized (*comprehensible range*). Comprehension of the text was tested during several pilot trials. In the case of speech rate, the upper limit was set to less than 250 wpm. This was based on the results of Slowiaczek and Nusbaum [14] suggesting that a user's comprehension with synthetic speech will decrease when the speech rate is greater than 250 wpm.

The lowest and the highest values in the comprehensible range were then selected for representing two extreme nodes in a given sub-tree. For example, the left-most node and the right-most node would be assigned a speech rate of 90 or 210 wpm if this was the range defined for the speech rate parameter. The nodes in between the two extremes were assigned values that ranged between the two extreme values.

4. Experiment

We conjectured that embedding navigation cues in the voice itself will result in lowering cognitive loads for tasks that involve identifying nodes and navigating hierarchies. In this experiment, we tested the effectiveness of multiple synthetic voices to represent hierarchies. However, evaluating the use of multiple synthetic voices for navigation tasks was not investigated in this experiment.

To evaluate the effectiveness of earcons in representing hierarchies, Brewster et al. [4] began their investigation on hierarchies with a small number of nodes. They initially tested their idea on hierarchies containing 9 nodes (excluding the root), three parent nodes with each containing two child nodes. Following the methodology adopted by Brewster et al. [4], we evaluated the synthetic voices on a hierarchy of 10 nodes (including the root, see Figure 1).

For the purpose of this experiment, each node was assigned a sentence produced using the DECTalk synthetic voice engine described earlier. All sentences contained the same number of words and did not include any information to suggest or hint at the child-parent relationships that existed in the hierarchies. Some of the sample text sentences were "the big dog slept on the floor last night", "the temperature is very high to go out today". The sentences were not repeated on any of the three hierarchies. We also ensured that the sentences were significantly different from one another and contained unrelated words.

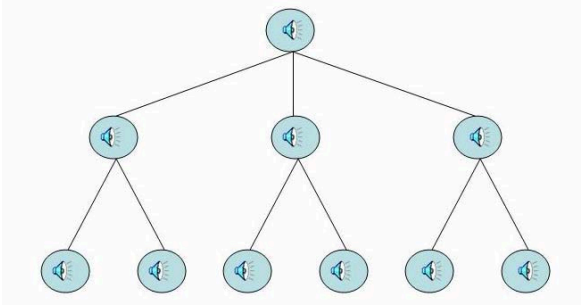

Figure 1 Hierarchy used for analyzing the effect of single synthetic speech versus multiple synthetic voices.

Three separate conditions were created for testing our hypothesis: Single Synthetic Voice (SSV), Multiple Synthetic Voice-1 (MSV-1) and Multiple Synthetic Voice-2 (MSV-2). The rules that were used to create the hierarchies are described below:

Single Synthetic Voice (SSV):

- All the nodes have the same pitch, pitch range and speech rate (AP=110, PR=135, SR=170). The only perceivable difference between nodes was the content of the sentences.

Multiple Synthetic Voice-1 (MSV-1): The rules that are used to create the MSV-1 hierarchy are shown in Figure 2.

- The root node is assigned a "neutral" synthetic voice. We used the following values to represent this node: AP=306 Hz, PR=210 Hz and SR=160 wpm.
- All the nodes in the second level were created with the same pitch (AP=110 Hz, PR=135). We assigned different values of speech rate to each node in the second level to create a sufficient contrast between them. For example, the left node was created with a low speech rate (SR=90 wpm), the middle node with a medium speech rate (SR=150 wpm), and the right node with a high speech rate (SR=210 wpm).
- The nodes at the third level were created with the same speech rate as that used in their parent node. In this case speech rate was inherited (MSSG1), i.e. children nodes of the left most sub-tree in the hierarchy inherit a SR=90 wpm. The distinctive dimension at the leaf nodes was the pitch (MSSG2). The left child had a low pitch (AP=10 Hz, PR=90 Hz), and the right child a high pitch (AP=200 Hz, PR=200 Hz).

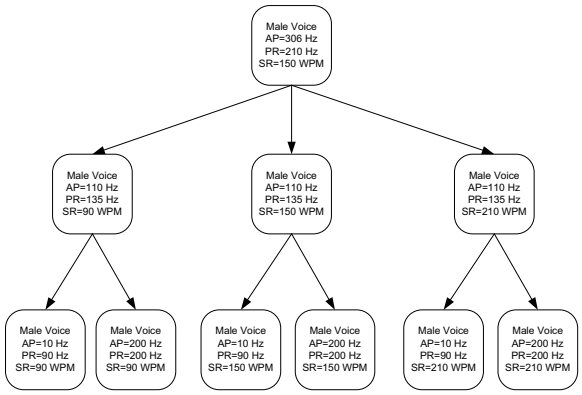

Figure 2 Values for the various parameters used in creating the Multiple Synthetic Voices-1 (MSV-1) hierarchy.

Multiple Synthetic Voice-2 (MSV-2): Figure 3 shows the values that were used for creating the MSV-2 hierarchy. The rules that are used for creating this hierarchy are the same as those used in creating the previous hierarchy (MSV-1) with the exception that the variation on the pitch and the speech rate dimension is interchanged.

- Similar to MSV-1, the root node is assigned a "neutral" synthetic voice. We used the following values to represent this node: AP=306 Hz, PR=210 Hz and SR=160 wpm.
- All the nodes in the second level were created using the same speech rate. The speech rate selected for this level was SR=160 wpm. The differentiating parameter between the nodes in this level is pitch. For example, the left node has a low pitch (AP=10 Hz, PR=90 Hz), the middle node has medium pitch (AP=110 Hz and PR=135 Hz), and the right node has high pitch (AP=200 Hz, PR=200 Hz).
- The nodes in the third level were created by using the same pitch as used in their parent node (i.e. pitch is inherited). The leaf nodes are differentiated by changing the speech rate. For example, the left child of the sub-tree has a low speech rate (SR=90 wpm), and the right child of the same sub-tree has a high speech rate (SR=210 wpm).

4.1. Hypothesis

Based on the results of earlier studies on non-speech sounds and on the perception of synthetic speech at the interface, as discussed in section 2, we anticipated the following effect:

- Multiple synthetic voices will lead to higher performance in tasks requiring identification of node position in a hierarchy than single synthetic voice.

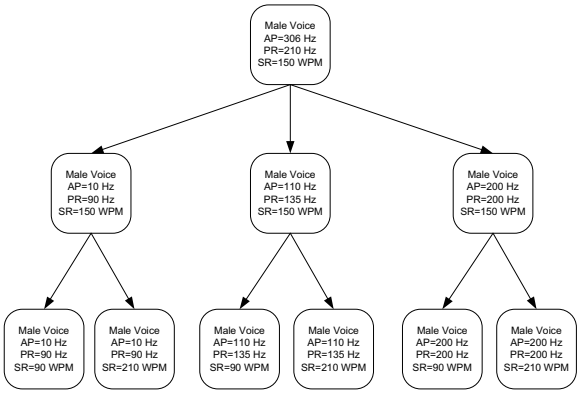

Figure 3 Values for the various parameters used in creating the Multiple Synthetic Voices-2 (MSV-2) hierarchy.

4.2. Method

4.2.1. Design. A within group experiment was designed with the three conditions described earlier: Single Synthetic Voice (SSV), Multiple Synthetic Voice-1 (MSV-1), and Multiple Synthetic Voice-2 (MSV-2). To reduce the learning effect, a fully balanced latin-square design was used. Each subject was randomly assigned to one of the three conditions. We measured the accuracy of locating a node that was previously listened to.

4.2.2. Materials. The hierarchies described above were presented using PowerPoint files with appropriate .wav files (sample rate = 11.025 Khz, 16-bit Mono) on the various nodes. The slides were shown using an x86 laptop with a 15.4" display and using the built-in Intel® Integrated 16-bit speakers.

4.2.3. Participants. 12 undergraduate students volunteered for this study. None of the participants reported a history of auditory disorder or exhibited any hearing problems. All the subjects also stated having previous experience listening to synthetic speech. They all spoke English fluently. Their exposure to synthetic speech was either through playing games or working with telephony based applications.

4.2.4. Training. At the start of the experiment, participants were given a write-up that described the rules that were used for creating the nodes in each of the three hierarchies. Participants were shown the structure of the hierarchy (Figure 1) and were asked to click on all the nodes. The node that was clicked on played the sentence assigned to it with the associated synthetic voice parameters for that node's position. For each subject the order presentation of each hierarchy type was random. They were allowed a maximum of three repetitions for each hierarchy. During the training session participants received very limited help from the experimenter.

4.2.5. Testing. After completing the training session participants were asked if they had understood the procedure and the rules that are used to create the hierarchies. The experiment began only when the participants were comfortable with the system. During the experiment, six voices were randomly selected from the set of 10 voices in the hierarchy and were played to the participants. Two of the voices were from level 2 (root is at level 1) and the remainder were leaf nodes. The participants were not able to see the computer screen during the experiment. After playing each voice, participants were asked to locate the position of the node in the hierarchy. Participants were given a sheet of paper with a hierarchy containing unlabelled nodes. They were asked to label the nodes in the hierarchy based on the order of presentation (i.e. the first node played was labeled 1, etc.). We used the labeling provided by each subject to compute their accuracy in identifying the position of the nodes in the hierarchy.

4.3. Results and Discussion

Results are summarized in Table 1, which reports error rates by the type of mapping. The results are obtained by averaging all subjects' scores. A Kruskall-Wallis test (non-parametric ANOVA) statistically shows that subjects performed differently with all three types of hierarchies ($p < 0.0001$). The mean error rate shows that subjects are four times more accurate with MSV-1 than with SSV and 2.8 times more accurate with MSV-2 than with SSV. A Mann-Whitney test statistically shows that subjects performed significantly better with both multiple synthetic voice conditions than the single synthetic voice ($p < 0.0001$ and $p < 0.005$ respectively for MSV-1 and MSV-2). However, the results do not show that there is a statistically significant difference between MSV-1 and MSV-2 (Mann-Whitney test, $p = 0.277$).

	SSV	*MSV-1*	*MSV-2*
Error Rate	62.5%	15.3%	22.2%

Table 1 Average error rates for finding the appropriate node in the hierarchy.

Error rates per voice type and per level in the hierarchy are summarized and are shown in Figure 4. From the entire set of six nodes that users were required to locate in each hierarchy, two nodes were on level 2 (level 1 is the root) and four nodes were on level 3. We observe that users' accuracy with the single synthetic voice condition degrades with the depth of the hierarchy. A Kruskall-Wallis test on level-2 of the hierarchy statistically shows that there is no significant difference in performance across all three conditions ($p=0.297$). Pair wise Mann-Withney tests show that there is no significant difference in performance between SSV and MSV-1 ($p=0.219$), between SSV and MSV-2 ($p=0.932$), and between MSV-1 and MSV-2 ($p=0.266$).

On level-3, a Kruskall-Wallis test shows that there is a significant difference in performance across all three conditions ($p<0.0001$). Pair wise Mann-Whitney tests show that there is a significant improvement with MSV-1 over SSV ($p < 0.0001$) and with MSV-2 over SSV ($p < 0.0001$). However, there is no statistical difference between MSV-1 and MSV-2 ($p=0.799$).

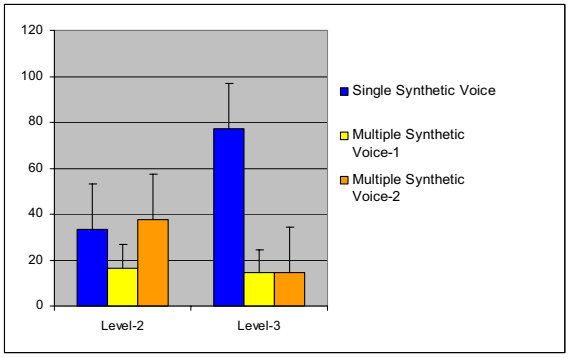

Figure 4 Average error rates for the three synthetic voice conditions for finding the appropriate node in the hierarchy based on the level in the tree.

Overall, the results confirm our hypothesis that multiple synthetic voices facilitate recall of component positions in a hierarchy. However, the mappings we selected for the multiple synthetic voices on the hierarchy did not reveal any significant difference. Overall, we can say that subjects performed slightly better with MSV-1 than with MSV-2. This may suggest that pitch is a better dimension than speech rate for identifying depth in the hierarchy, and identifying nodes within a level may be better achieved by varying the speech rate. Further experimentation is required to determine the best rules for manipulating the voice parameters such that deep and wide hierarchies can be adequately represented.

5. Conclusions and Future Work

We conducted an experiment to test whether multiple synthetic voices can be used to represent hierarchies. Following the methodology employed in a related study [4], we tested our hypothesis on a small and fully balanced hierarchy containing ten nodes. We manipulated three synthetic voice parameters, average pitch, pitch range, and speech rate to create multiple synthetic voices that were hierarchically related. The results of our experiment show that multiple synthetic voices can be used to represent nodes in the hierarchy. The manipulation of synthetic voice parameters followed similar guidelines (repetition, variation, contrast) used in the design of hierarchical earcons. Our results show that these rules facilitated the recall of node positions within a hierarchy.

We believe that the results described here lay the groundwork for further investigation. We will next determine the size of the largest possible hierarchy that can be represented using multiple synthetic voices. Certain speech parameters such as speech rate cannot be extended beyond a certain range before comprehension is degraded. In such cases, a different mapping of the parameters is required.

The experiment described in this paper was conducted on a fully balanced hierarchy. In many real-world applications hierarchies are far from symmetric. Further investigation is needed to determine whether the guidelines that are used to create the hierarchy tested in our experiment are applicable to more generic structures.

A primary objective in this line of investigation will be to derive a set of rules to guide the design of multiple synthetic voices for representing hierarchies. Other synthetic speech parameters such as breathiness can be taken into account to create variations between elements in a hierarchy. This will allow us to create voices for hierarchies that may grow or shrink dynamically.

Finally, we plan on investigating various applications for which multiple synthetic voices can be successfully developed. For example, in virtual environments, multiple synthetic voices could be a source of navigation cue to the users. The set of conventions for creating the synthetic voices could be different from one context to another. Identifying these rules will be necessary if multiple synthetic voices are to be used in real-world applications.

6. Acknowledgements

We thank Jennifer Lai from IBM Watson research labs for her ideas and guidance in this phase of the project. We are grateful to Fonix Corporation for providing us support with the DECTalk development toolkit. Finally we thank our colleague Christel Kemke for reviewing and suggesting improvements to this article. This research is supported by a NSERC grant.

References

[1] 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.

[2] S. A. Brewster. Navigating telephone-based interfaces with earcons. In Proceedings of BCS HCI'97, pages 39–56, 1997.

[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, P. C. Wright, and A. D. N. Edwards. An evaluation of earcons for use in auditory human-computer interfaces. In INTERCHI Conference Proceedings, pages 222–227. ACM Press, 1993.

[5] J. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, May 1989.

[6] S. L. Greenspan, H. C. Nusbaum, and D. B. Pisoni. Perception of synthetic speech produced by rule: Intelligibility of eight text-to-speech systems. Behavior Research Methods, Instruments, & Computers, 18:100–107, 1988.

[7] J. Lai, K. Cheng, P. Green, and O. Tsimhoni. On the road and on the web?: Comprehension of synthetic and human speech while driving. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 206–212, Seattle, Washington, United States, 2001. ACM Press.

[8] 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, pages 1–6, Seattle, Washington, United States, 2000. ACM Press.

[9] E. J. Lee, C. Nass, and S. Brave. Can computer-generated speech have gender?: An experimental test of gender stereotype. In CHI '00 extended abstracts on Human factors in computer systems, pages 289–290. ACM Press, 2000.

[10] G. Leplatre. The design and evaluation of non-speech sounds to support navigation in restricted display devices. PhD thesis, University of Glasgow, Nov 2002.

[11] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and D. C. Dryer. Can computer personalities be human personalities? International Journal of Human-Computer Studies, 43(2):223–239, 1995.

[12] C. Nass, J. Steuer, and E. Tauber. Computers are social actors. In Conference companion on Human factors in computing systems, page 204. ACM Press, 1994.

[13] E. C. Schwab, H. C. Nusbaum, and D. B. Pisoni. Effects of training on the perception of synthetic speech. Human Factors, 27:279–291, 1985.

[14] 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.

[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.

[16] 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.