Whispers: An Aural Display

Jonathan Lung
Department of Computer Science
Uiversity of Toronto, Toronto, ON, M5S 2E4
http://www.cs.toronto.edu/~lungj

ABSTRACT

An aural display is examined as an alternative to visual displays. An informal experiment was performed to study the aural equivalent of a visual search task. The results of the experiment included the observation that an increase in distractor voices increases the amount of time required to find a piece of information being spoken. The experiment also identified the effects of learning and voice familiarity as two possible factors affecting the amount of time required. From this, it was concluded that aural interfaces might be suited to situations in which access to a large display is inconvenient or impossible.

KEYWORDS

Cocktail party effect, diotic aural presentation, aural interface, audio, voice, ambient display, text-to-speech (TTS), human aural search, auditory search task

INTRODUCTION

It is sometimes the case that dynamic information is required intermittently, such as the current time or the current temperature. Oftentimes, this information is made available visually to computer users through an operating system feature such as the clock in the system tray in Microsoft Windows or through some third-party application such as Yahoo! Widgets [11]. Due to the limits on screen space, it is almost by necessity that some or all of this secondary information is occluded, hidden, or at least relegated to the periphery. Thus, in order to access this information, users must switch the focus of their visual attention from the active task and sometimes manipulate the current state of the computer, for example by bringing a window to the foreground or activating a utility through some keystrokes [1, 11]. Complicating matters is the fact that it becomes harder to find a desired window as the number of open windows increases. As a result, different methods for window-switching have been devised [2, 4]. For users situated in front of a computer, this may not be a major issue, but for those performing other activities, this represents a significant overhead. Furthermore, this is probably disruptive to tasks such as reading where the information channel is visual.

In this paper, an alternative is proposed that employs the auditory channel. Here, multiple streams of numerical or qualitative information can be presented diotically, i.e., the same signal to both ears, with the intention that users ignore the sound as white noise or attend to a specific channel of interest without needing to interact directly with a computer. A prototype system named Whispers was implemented and informal user tests were performed. Compared to visual methods of presenting information, Whispers has the advantage of presenting information in an ambient manner, freeing the user from attending physically and visually to a computer.

RELATED WORK

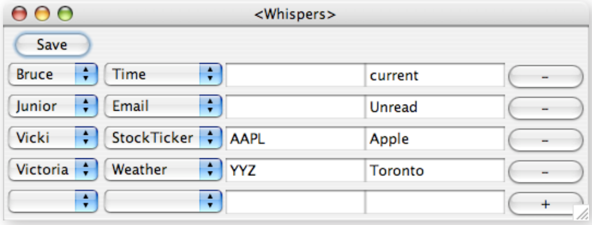
The theoretical basis for this work stems from the "cocktail party effect" which derives its name from a cocktail party where concurrent conversations occur within earshot of each other. Several studies related to the cocktail party effect have been performed demonstrating that humans can isolate and attend to a specific voice from a chorus of voices [5, 9]. Several factors have been shown to affect shadowing performance including the spatial separation [5] and gender of the speaker [9].

One system that leverages the cocktail party effect is AudioStreamer [8]. AudioStreamer relies on a combination of the spatial separation of speakers and proactive suppressing of background channel volume to support audio browsing. More closely related to Whispers is the soundscape of Ishii's ambientROOM in which bottles serve as tangible switches for information [7]. In ambientROOM, different audio channels give approximations of quantitative data.

METHOD

Experiment 1

Two sets of informal tests were performed on Whispers. In the first, an attempt was made to ascertain the effectiveness of the system and, in the second, some basic performance data was gathered. In order to perform the first set of tests, a prototype was implemented that allowed information to be associated with various computer synthesized voices (see Figure 1). Information came from one of four sources: the system clock, a simulated e-mail client, an on-line weather service, and stock quotes on Yahoo! Finance. Whispers monitors its data sources for changes and generates audio clips appropriately using the built-in text-to-speech (TTS) capabilities of MacOS. For example, if the time were 2:15pm and the voice "Vicki"



Bruce	Time		current
Junior	Email		Unread
Vicki	StockTicker	AAPL	Apple
Victoria	Weather	YYZ	Toronto

Figure 1. The voice-channel selection screen in the Whispers prototype.

associated with time, Vicki would continue to repeat, “It is now 2:15”. The author subjected himself to the prototype for several hours while performing an unrelated task. During this time, the current time and three stock prices were monitored.

Experiment 2

Overview

Participants were given a piece of information to listen for (e.g., “Unread messages”) and to write down the associated value. An audio clip was then played with between one and five voices speaking simultaneously. The amount of time taken for participants to make out the information was recorded. There were two experimental conditions. One group had a 2.5 second silence before the voice(s) began speaking while the other group was cued to the voice that would be speaking the requested information for that trial.

Experimental Apparatus

One 3rd generation iPod (2003 model) along with standard headphones included with the device were used for audio playback and instruction delivery. The included iPod remote control was also used in the experiment.

Participants

Four participants were chosen by convenience sampling. The single requirement was that each participant was required to be a native speaker of English. No compensation was offered for their effort.

Test Cases

Thirty test cases were generated randomly with a simple Python script. There were six test cases for each number of voices between one and five, inclusive. Each of these group of six test cases were split into halves and assigned randomly to be part of one of two blocks. Thus, there were two blocks, each consisting of 15 test cases. An example two-voice case looks like this:

- Voice 1: Temperature
- Voice 3: Price of Sun Microsystems’ stock price
- Target information is weather

In each test case, for an *n*-voice test case, there were *n* distinct voices reading information from *n* distinct data sources.

Sound Generation

For the second experiment, auditory data was not generated using the actual Whispers system/MacOS TTS. Instead, sound clips were generated from randomly generated data using AT&T’s TTS system web demo [3]. This was due to the fact that more voices were available and the voices sounded more natural than those produced by the MacOS TTS software. The nature of the utterances in the sound clips were otherwise identical to the audio produced during a typical Whispers session. The names of the voices used for this experiment were Claire, Crystal, Lauren, Mike, and Rich; each of these voices had an accent not unlike a typical Torontonian. The length of utterances ranged from 1 to 3 seconds. Using Audacity 1.2.6, each clip was looped with 0.5 second gaps inserted after each utterance. Next, the clips were merged to produce the combination of voices as the test cases required. Playback on each channel began at a random point in the corresponding utterance. Following the channel merging, a 2.5 second silence was inserted at the beginning of each audio test case file. These files constituted the test cases for the uncued condition. An otherwise identical copy of these audio files had the phrase “Please listen for my voice” added to the beginning during the first 2.5 seconds, spoken by the voice that would be reading the target information for the trial. These files with the additional introduction constituted the cued condition. Next, each trial was given a randomly generated code. The audio files were named using this code as a prefix. The latter part of the file name was the type of information for which participants were to listen. Thus, a typical file name might be “1cb2 - Price of Sun.mp3”.

Data Collection

The audio files were grouped into four playlists on the iPod: cued condition block A, cued condition block B, uncued condition block A, and uncued condition block B. Each participant experienced both blocks from one condition. The order that blocks were presented was balanced amongst participants in each condition. The “shuffle” playback mode on the iPod was used to randomize trials within blocks. Participants were asked to read the iPod display for the information to be identified. Due to the naming convention used for the audio files, this was simply a matter of reading the name of the file. Participants were also instructed to press the “pause” button on the iPod when they had extracted the information sought and to record that onto a piece of paper. While the participant was recording this information, the trial code and the time elapsed as indicated by the playback status bar on the iPod were also noted. When both the participant and the experimenter were ready, audio playback was resumed by the experimenter using the remote control.

RESULTS

Experiment 1

On a 1.25GHz G4 PowerPC with 1GB of RAM running MacOS X 10.4.8, the maximum number of simultaneous channels that could be played in real-time using Whispers was four. Attempting to play more than this causes silent gaps. However, even with only four, the voices begin to become less obtrusive and more like white noise. With four voices, the system seemed to be useful, providing an awareness of stock prices and the time without drawing undue attention to itself.

Experiment 2

Overall, the average number of times the target information needed to be repeated increased as the number of voices increased (see Figure 2). The number of repetitions required was calculated using the following formula:

$$\text{repetitions} = \frac{(\text{elapsed time}) - (\text{onset delay})}{(\text{target clip length})} = \frac{(\text{elapsed time}) - (2.5\text{s})}{(\text{target clip length})}$$

Note that the number of times a message needs to be repeated may be less than one due to the fact that the entire sentence need not be heard in order to identify the target information. It was also found that, in general, the cued group performed better than the uncued group (see Figure 3). The effects of learning on the speed of identifying target information, if any, was not strongly visible (see Figure 4), but group error rate decreased from a 13% error rate on the first trial block presented to 3% in the second. However, due to the small sample size, none of the results obtained were analyzed for statistical significance. When participants were asked after hearing all the samples how many voices they thought were present during different trials, they knew only that there were trials with one voice, two voices, and the rest with more than two, i.e., they drew no distinction in the number of concurrent voices past two.

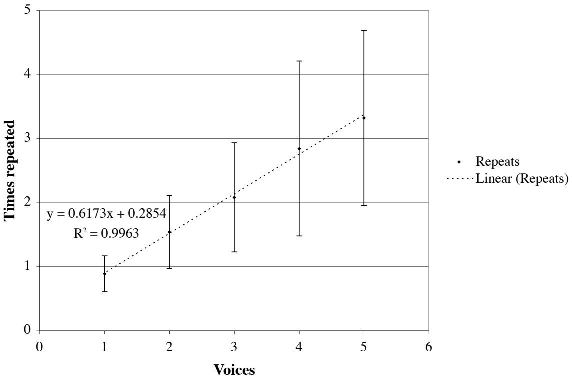


Figure 2. The effect of the number of voices on the number of times information had to be repeated.

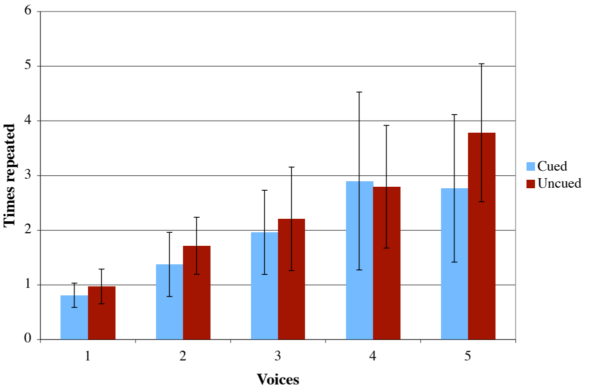


Figure 3. The effect of cuing on the number of times information had to be repeated for a given number of voices.

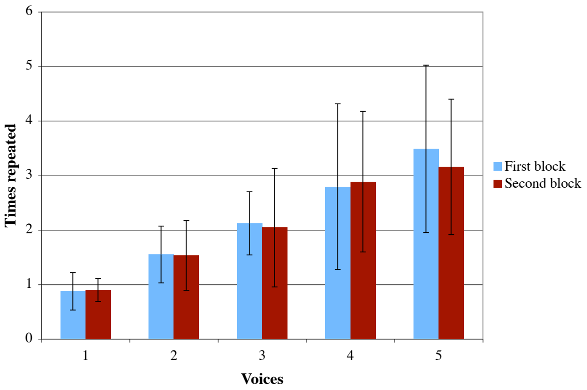


Figure 4. The effects of experience on the number of times information had to be repeated for a given number of voices.

DISCUSSION

The results of the first experiment indicate that more computing power is necessary to play back a sufficient number of synthesized voices for this application to become a source of white noise. More powerful computers are presently available which may solve the problem. In any event, this is purely a technical challenge that can be easily overcome with additional amounts of processing power. While the results of the first experiment are not particularly interesting, the second experiment warrants deeper discussion on both the interpretation of the data and the implications.

In the second experiment, similar to a conjunctive visual search task, the average number of times the target information had to be repeated increased as the number of distractors increased (see Figure 2) [6]. However, the number of times required for even a few voices seems to make Whispers unattractive as a display mechanism to a user already interacting with a visual display. In that situation, activating a window with the desired information should be faster since three repetitions of information corresponds to 6 – 9 seconds.

Other results from the second experiment hint at some ways that the number of repetitions can be reduced. The data in Figure 3 suggests that knowing in advance the sound of the voice uttering the target sentence helps reduce time. If this is indeed the case, one possible corollary from the observation is that real-world usage might see improved user performance: while configuring a TTS ambient display, the user can select a voice to which he or she is accustomed and, in doing so, be able to more quickly pick out the desired information by attending to a specific and familiar voice. Indeed, for some large number of voices, given sufficient familiarity with a voice, the number of times information needs to be repeated may approach unity [8]. This is the case in real world environments where many conversations can be occurring simultaneously without the need for every sentence to be repeated. Of course, this might be due to some form of top-down processing like phoneme restoration [10]; such top-down processing would likely be inappropriate for a repetitive message where there are only one or two key phrases in the entire sentence.

Even without the benefits of becoming familiar with a voice, there appears to be some learning taking place during the experiment (Figure 4). Due to the limited number of participants, it was not possible to determine if the learning effect might be stronger for the cued or uncued condition, if either. There was also evidence that accuracy improves with practice. It seems likely that, even with listening errors, Whispers can provide a more precise quantification of numerical data than "natural" sounds in ambientROOM. For example, a blue chip stock usually varies in price, up or down, by only a few fractions of a percent a day. On other days, it may fluctuate by several percent. Intensity (e.g. traffic density) and/or amplitude (i.e., quiet versus loud) are not good candidates for conveying this information in absolute terms.

Though the results of the experiments were largely inconclusive, by increasing the number of participants, trials, and blocks used in future experiments, it is quite possible that relevant and statistically significant results will be found leading to a better understanding of the human ability to attend to specific voices. This knowledge can be leveraged to create ambient aural displays. Additionally, the possibility of using voices that are more distinct (e.g. different accents) and using sounds from different spatial locations should be explored. The effects of using of real, familiar voices would also be interesting to study as the odd emphasis placed on certain words by the TTS software may be hampering recognition.

On the surface, then, the results from the second experiment make the prospect of an aural display unappealing. There may, however, be some situations when it is appropriate or necessary. Consider a foreground visual task demanding vigilance, e.g. watching a security camera video. In such a situation, visual attention should remain fixed on the display and a display utilizing another sense may in fact be necessary to convey additional information. Another situation where an ambient aural display may be useful is when a person is not situated in front of a computer terminal such as while reading a book. In that case, getting to a display to view information on a screen may take more time and is less convenient. Yet another situation in which aural displays may be particularly well suited is for portable electronic devices. The resolution and size of displays on personal digital assistants (PDAs) and cellular telephones is only a fraction of what is commonplace for desktop computers. Here, the amount of room available for peripheral information and task switching interfaces is minimal. The use of an aural display in that situation may be beneficial.

CONCLUSION

Because of the amount of time required to access information being read by aural displays, they are relatively inefficient in environments with ample display space. In other cases, aural displays may be an improvement if there are several pieces of information that must be readily accessible with little or no display space available for the purpose. Through practice and better methods of making voices distinct, the usefulness of aural displays can be increased as users become adept at working with many simultaneous streams of information. Thus, aural displays may prove to be a practical alternative to visual displays in some situations.

ACKNOWLEDGEMENTS

Thank to Professors Ravin Balakrishnan and Khai Truong for acting as a sounding board for various ideas and to those who volunteered their time as participants in the experiments.

REFERENCES

1. Apple Computer, Inc. (2006). Apple – Mac OS X – Dashboard. Accessed from http://www.apple.com/macosx/features/dashboard/ on December 9, 2006.

2. Apple Computer, Inc. (2006). Apple – Mac OS X – Exposé. Accessed from http://www.apple.com/macosx/features/expose/ on December 9, 2006.

3. AT&T. (2006). AT&T Labs Text-to-Speech: Demo. Accessed from http://www.research.att.com/~ttsweb/tts/demo.php on December 15, 2006.

4. Card, S., Henderson, A. (1987). A multiple, virtual-workspace interface to support user task switching. CHI + GI 1987, pages 53-59. Accessed electronically through http://library.utoronto.ca on December 9, 2006.

5. Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. The Journal of the Accoustical Society of America, volume 25, pages 975-979. Accessed electronically

through http://library.utoronto.ca on December 14, 2006.

6. Horowitz, T. and Wolfe, J. (2003). Memory for rejected distractors in visual search? Visual Cognition, volume 10(3), pages 257-298. Accessed electronically through http://library.utoronto.ca on December 21, 2006.

7. Ishii, H., Wisneski, C., Brave, S., Dahley, A., Gorbet, M., Ullmer, B., & Yarin, P. (1998). ambientROOM: Integrating ambient media with architectural space. Conference Summary of CHI '98. Accessed from http://tangible.media.mit.edu/content/papers/pdf/ambientROOM_CHI98.pdf on December 20, 2006.

8. Mullins, A. (1996). AudioStreamer: Leveraging the cocktail party effect for efficient listening. Accessed from http://www.media.mit.edu/speech/papers/1996/mullins_thesis96_AudioStreamer.pdf on October 27, 2006.

9. Treisman, A. (1964). Effect of irrelevant material on the efficiency of selective listening. The American Journal of Psychology, volume 77, pages 533-546. Accessed electronically through http://library.utoronto.ca on December 14, 2006.

10. Warren, R. (1970). Perceptual restoration of missing speech sounds. Science, volume 167, pages 392-393. Accessed electronically through http://library.utoronto.ca on December 15, 2006.

11. Yahoo! Inc. (2006). Yahoo! Widgets - Get Weather, Photos, Calendar, and More on Your Mac or Windows Desktop. Accessed from http://widgets.yahoo.com/ on December 9, 2006.