Do You Hear Voices?
Exploring Notification Characteristics of Voice

Saurabh Bhatia, Sam Edwards, Joe Stegner, D. Scott McCrickard

Department of Computer Science and Center for HCI
Virginia Polytechnic Institute and State University (Virginia Tech)
Blacksburg, VA 24061-0106
{saurabhb, mccricks}@vt.edu

Abstract

In many situations when our visual faculties are not available, standard notification mechanisms like visual cues are not effective. For this reason, we decided to explore the realm of audio notifications, more specifically, voice notification. Voice shows promise in drawing attention to information that requires appropriate reaction and long-term comprehension. Our research investigated the notification qualities of different voice categories, moving toward interfaces that properly balance the critical parameters of interruption, reaction, and comprehension (IRC) in providing information alerts. We hypothesized that different voice categories would have different notification characteristics. To test the hypothesis we conducted an experiment to determine the interruption, reaction, and comprehension values of three different voice categories: the user's voice, a familiar person's (to the user) voice, and an unfamiliar voice. Initial testing showed some promising results. The unfamiliar voice had the lowest interruption, a user's own voice had the highest reaction, and the comprehension level was the same among all voices.

1 Introduction

Audio is a proficient medium for providing information alerts. Sound has the general characteristic of drawing attention. It has benefit over visual media as it does not require a user to be paying constant attention to a certain focussed visual region. For this reason, most notifications that need immediate attention use some type of audio cue to draw attention. An alarm that you use to wake up is an excellent example. The beeping interrupts sleeping, and users learn to associate beeping with the need to wake up. In today's fast-paced, technologically-enhanced work environment, the alarm clock is not the only audio cue that governs our daily activities. Ubiquitous devices like cell phones, pagers, and PDAs help guide the way we use our time, generally drawing attention by using an audible cue. Users learn to distinguish different audio cues and derive meaning to the notification. These audio cues may also be used to draw attention to a visual artifact which in turn would provide more information about the notification. For example, a PDA beeping to draw attention to a message flashing on the screen that notifies you of a meeting in fifteen minutes.

There are numerous examples of audio, and in particular voice, used in human-computer interaction for notification purposes. For example, the Nomadic Radio project by Sawhney and Schmandt at MIT makes use of a variety of voices and audio cues to provide scalable notifications (1999), and the work of Clifford Nass and his colleagues has empirically explored the utility of voice in interfaces and notification (Nass & Gong, 2000; Lee & Nass, 2003). Nomadic Radio is a wearable audio device that notifies the user of emails, voicemails, and scheduled tasks. The device is context aware and changes the type of notification depending on the user's environment. Nomadic Radio uses audible beeps, natural ambient sounds and pre-recorded human voice to notify the user. These three types of sound notifications were used to achieve different levels of interruption. The work of Nass explores questions related to human reaction to voice, specifically voice generated by a computer or some sort of computing device. It is the open question presented in his Communications of the ACM paper that inspires us: "Will familiarity with a computer-based voice influence users' processing of that voice?" (Nass & Gong, 2000).

Voice, although potentially highly interruptive, has the additional property of being very informative. Voice has an advantage over a generic sound like a beep or other ambient noise because voice can convey information directly and does not require the user to associate different sounds with meanings. Traditionally, voice notification has primarily been restricted to public environments. Devices like the Nomadic Radio have brought voice into the personal notification sphere. Voice notification can also be useful in a semi-public environment. Offices could use a voice notification to notify team members of a meeting or a voice notification could be used to indicate that a meeting is nearing its end and participants should start wrapping things up. Voice is an excellent notification mechanism in these situations as it can reach and inform all concerned people no matter what they are doing or where they are in the office.

These environments allow for a variety of voices that can be used for notification. To harness and understand its potential for notification, we decided to further explore human voice and the notification characteristics different voices produce. As any characteristic would be relative to a user, we decided to group voice into three categories – the user's voice, a familiar person's (to the user) voice and an unfamiliar voice. Each type of voice will have a different notification characteristic. In previous work (McCrickard, Chewar, Somervell & Ndiwala 2003), we established three critical parameters that define the notification system design space – interruption, reaction and comprehension, abbreviated IRC. By measuring user performance for various interface components along these parameters, we can match user goals for a system with appropriate interface components — for example, if a familiar voice was non-interruptive but enhanced comprehension, then that voice should be used in a situation where attention to a primary task is critical but knowledge gained from the voice held high importance as well.

The research effort described in this paper examines how different types of voices compare in terms of the three critical parameters. Our objective was to investigate if there was any variance in these parameters when using the different categories of voices. Determining these variations will help in building voice based notification systems that will achieve their target levels for interruption, reaction and comprehension.

2 Approach

We designed an experiment to find empirical evidence of the differences in the critical parameters for the different voice categories. The experiment would help isolate the IRC levels for each voice. This would be done in a controlled environment so as to ascertain that any variance in the critical parameters was in fact due to the different voice categories. We state our experiment hypothesis as follows: Hypothesis (Ha) The three different voices — your own, a familiar person's and an unfamiliar person's — provide substantially different levels of interruption, reaction and comprehension.

2.1 Experiment

The experiment involved the users playing a simple computer game of catch the falling blocks (see Figure 1). This game has also been used to study secondary display attributes in our earlier work (Tessendorf et al., 2002). The game was modified to add support for voice playback and recording. TCL script with the SNACK package was used to implement these features.

Twenty-seven volunteers participated in this experiment. Participants were recruited from an undergraduate level class and were given the incentive of extra credit for taking part in the experiment. Each session lasted for approximately thirty minutes. The experiment was conducted in a quiet computer lab, with each participant wearing a headset to hear the audio. Groups of three to five participants performed the experiment at a time. Participants were required to first record the numbers 0-9 in their own voice. This was done one at a time for each participant so that the recordings were clear. Each number was recorded in a span of one second so that there would be uniformity in the way the numbers were read out. The instructor who was teaching their class volunteered to be the familiar voice. For the unfamiliar voice, we chose the voice of a person that none of the participants were familiar with, and had a French accent (we verified that none of the participants had regularly been exposed to a French accent).

The game was simple and involved some falling blocks with a little paddle to catch them. The user could move the paddle left and right using the arrow keys to catch as many blocks as possible. The idea behind using this game was to keep the user engaged so as to simulate a primary task requiring constant attention and concentration. The blocks fell in the same exact order for all participants so that each participant could see the same pattern of blocks falling for each corresponding round.

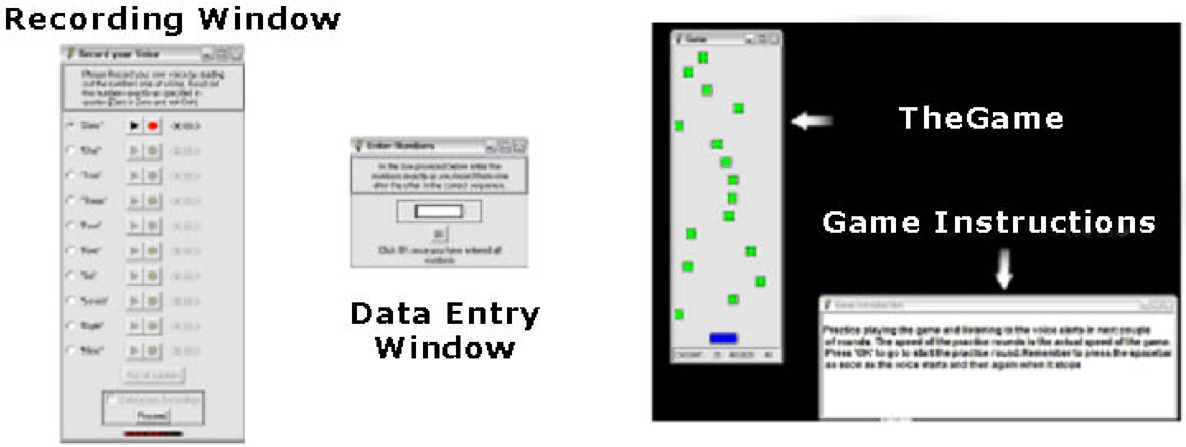
2.2 Procedure

Before starting the experiment, the users were asked a set of questions. These questions give us background about the users and helped us assess their different cultural and social backgrounds. This gave us a sense of the wide range of languages spoken by the users and their familiarity with different accents.

On starting the experiment, the users were first given a list of instructions for playing the game. They were then given four practice rounds to familiarize themselves with the game. The experiment itself consisted of nine rounds. During each round, the game was interrupted by a voice reading out a random seven-digit number. The users had to remember these numbers as they continued playing the game and enter then into a box at the end of the round. The users heard a different voice in each round. The sequence of the different voice types was varied in a Latin Square Design. Each user was assigned to one of three groups, all three of which started the game with a different voice category. In all, the users got to hear each voice three times. Each round lasted for one minute. Voice notification occurred approximately twenty five seconds into the game and lasted for approximately seven seconds.

2.3 Calculating IRC

Interruption was measured by the relative drop in game performance. Each participant's catch rate was determined by calculating the percentage of number of blocks caught to the total number of blocks. The catch rate was calculated before, during and after the voice notification. The drop in catch rate was used as an indicator of the interruption caused by the voice. To measure reaction, participants were asked to hit the space bar as soon as they heard the voice notification. The time difference between the start of voice notification and the user hitting the reaction key was used to calculate reaction time. The correctness in remembering the numbers, as measured by the users entering them at the end of the game, was used to calculate the comprehension.



Practice playing the game and listening to the voice alerts in next couple of rounds. The speed of the practice rounds is the actual speed of the game. Press 'OK' to go to start the practice round. Remember to press the spacebar as soon as the voice starts and then again when it stops

Figure 1: Screenshots of experimental platform. Users begin by recording their own voice. Users are given instructions and four practice rounds before the actual testing begins. At the end of each round they are asked to enter the numbers they heard.

3 Results and Analysis

After collecting data from our twenty-seven participants, we were able parse data files that were created by our TCL experiment script into a database. We ran ANOVA tests on this data to identify differences in interruption, reaction, and comprehension. For reaction, the test results suggested that there is a significant difference in the means for the reaction time; between the three voices ($F(2,215)=3.74$, $MSE=48785.42$, $p<0.03$). Participants reacted most quickly to their own voice, while they reacted slower to the familiar voice and slowest to the unfamiliar voice (as seen in Figure 2). Hence, your own voice gives minimal reaction time and thus provides high reaction.

The slower reaction time associated with the unfamiliar voice corresponds with the idea that we tend to filter out voices of people who we do not know. A simple analogy would involve being in a crowd—one would filter out any unfamiliar voices around you and tend to pay more attention and react quicker to a familiar voice. The quickest reaction time to your own voice came as a surprise. Although we hear ourselves talk all the time, it is not the same as listening to a recording of your voice. Listening to your own voice being played back might have evoked an emotional response that relates to your self-image. This in turn may have triggered the fast reaction time that showed that you are acknowledging yourself. It must be noted that although the difference in the three timings is less than a quarter of a second, this difference may prove significant for more complicated notifications.

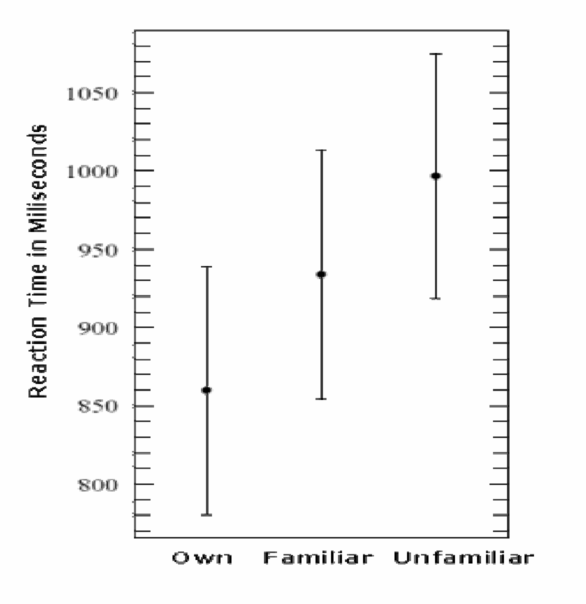


Figure 2: ANOVA test results for reaction time.

While the ANOVA test showed no significant difference for interruption, a t-test between your own voice and the unfamiliar voice however revealed a significant difference in means. These means were the difference between performance before and after the voice notification. Performance was calculated by the percentages of balls caught divided by balls dropped. The mean reduction in catch rate was significantly larger with your own voice (M=4.06, SD=2.64) than with the unfamiliar voice (M=3.11, SD=2.08), t(112)=2.11, p=0.03. Therefore your own voice has a higher interruption level than the unfamiliar voice. The high interruption possibly arises from the same reasoning that causes high reaction for your own voice. Further study is needed to explore the validity of this result.

ANOVA and T-Tests were run on data collected for comprehension. There was no significant variance in the comprehension levels between the three voices. Comprehension remained high in general for all three voices. This shows that voice in general has a high comprehension and is good for application where high comprehension is desired.

4 Conclusions

Our direction for this work was to identify notification characteristics of different categories of voices. Our results suggest that the user's own voice has high interruption, the highest reaction and high comprehension. A familiar voice has high interruption, comparatively moderate reaction and high comprehension. Lastly, an unfamiliar voice has low interruption, the lowest reaction and high comprehension.

The characteristic of high IRC values of the user's voice can be used to design numerous applications, such as an alarm clock that tells the user to wake up in their own voice. This would certainly be more effective than a regular alarm clock. It would also be feasible for users to record their own voice when using personal devices. Such devices will greatly benefit from emerging technologies in wearable computing as showcased in the Nomadic Radio project.

Sound can directly be focussed into the user's ear such that only that person may hear it, thereby maintaining privacy.

On the other hand, semi-public environments allow for making complete utilization of the differences in familiar and unfamiliar voices. Semi-public environments, as introduced by Huang and Mynatt (2003), by typically involve fifteen to twenty people who know each other. As everyone in the group is familiar with each other's voice, identifying different voices within the group is not as much of an issue. Scalable notifications can be created by choosing the familiar and unfamiliar voices appropriately. For instance, an unfamiliar voice could be used in a semi-public setting that requires low interruption and high comprehension. The idea of a meeting announcer discussed earlier in the paper, which notifies participants that the meeting is nearing its end, will work best if an unfamiliar voice is used to make the announcement. The meeting announcer generates a comparatively low interruption, thereby ensuring that people are notified but the meeting is not disrupted. The unfamiliar voice fits this scenario. Alternatively, a familiar voice would be suitable for announcing an impromptu meeting. In this case the notification has to be highly interruptive as people have to stop what they are doing and go to the meeting.

Our study presents initial results regarding notification using voice, building on several established efforts in the exploration of computerized voice and integrating with efforts in the design, building, and testing of notification systems. Clearly more empirical study is needed to understand the potential use of voice in the construction of notification systems, and the results from such studies must be used to develop a variety of systems that would apply the results. As the efforts continue through lab-based and real-world study, we will better understand the potential role of voice in notification.

5 References

Huang, E. M. and Mynatt, E. D. (2003). Semi-Public Displays for Small, Co-located Groups. In *Proceedings ot the ACM Conference on Human Factors in Computing Systems (CHI 2003).*

Lee, K. M. and Nass, C. (2003). Digital Sociability: Designing Social Presence of Social Actors in Human Computer Interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2003).*

McCrickard, S., Chewar, C. M., Somervell, J. P., & Ndiwalana, A. (2003). "A Model for Notification Systems Evaluation--Assessing User Goals for Multitasking Activity." *ACM Transactions on Computer-Human Interaction*, 10 (4), 312-338.

Nass, C., Gong, L. (2000). Speech interfaces from an evolutionary perspective. *Communications of the ACM*, 43 (9), 36-43.

Sawhney, N., & Schmandt, C. (1999). Nomadic radio: scaleable and contextual notification for wearable audio messaging. In *Proceedings of the ACM Conference on Human factors in Computing Systems*.

Tessendorf, D., Chewar, C. M., Ndiwalana, A., Pryor, J., McCrickard, S., & North, C. (2002). "An Ordering of Secondary Task Display Attributes." In *Conference Companion of the ACM Conference on Human Factors in Computing Systems (CHI 2002).*