Listening to Your Inner Voices: Investigating Means for Voice Notifications

Saurabh Bhatia, D. Scott McCrickard
Department of Computer Science and Center for HCI, Virginia Tech
Blacksburg, VA 24061-0106
{saurabhb,mccricks}@cs.vt.edu

ABSTRACT
Our research investigates notification qualities of different types of voices, moving toward interfaces that support optimal allocation of attention to maximize system utility. We conducted an experiment to determine the interruption, reaction, and comprehension values of three different voice categories: the user's voice, a familiar voice, and an unfamiliar voice. Initial testing showed significant and impactful results: unfamiliar voices are the least interruptive, and a user reacts most quickly to one's own voice. Motivated by these findings, we report on the development and deployment of a notification system that exploits the differences in familiarity of a voice.

Author Keywords
Interruption, voice interfaces, notification systems

ACM Classification Keywords
H.5.2. User Interfaces: Voice Output

INTRODUCTION

In today's fast-paced, technologically-enhanced work environment, audio cues help guide our daily activities. Ubiquitous devices like cell phones, pagers, and PDAs help govern the way we use our time, often drawing attention by using an audible cue. Users learn to distinguish audio cues and derive meaning from notification. For example, Alexanderson, in his study of the auditory environment in a chemical factory, shows the potential in using audio cues for notification [1]. By deriving audio cues from an existing environment users are saved from the burden of re-learning their meaning. Gaver demonstrates that non-speech audio cues can support awareness [2]. His work demonstrates how familiar sounds can keep workers informed of office occurrences. Audio cues may also be used to draw attention to visual artifacts that can provide information about the notification—beeping PDAs draw attention to an on-screen message that notifies of an upcoming meeting.

Voice interfaces convey information directly, without the potentially taxing interpretive stage noted in non-voice audio [1]. To combat this issue, the Nomadic Radio augments audio cues with voices to provide scalable notifications [7]. 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. By augmenting audible beeps and natural ambient sounds with pre-recorded human voice, the Nomadic Radio could achieve different levels of interruption. The work of Clifford Nass and his colleagues have empirically explored the utility of voice in interfaces development, examining societal and other impacts of computer-generated voices [4,6]. 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?" [6].

Devices like the Nomadic Radio have brought voice into the *personal* notification sphere. The work described in this paper explores how voice notification can also be useful in semi-public *group* environments (as defined in [3]), where group members share informational needs and goals. The pervasiveness of voice provides high utility in semi-public environments where users are dispersed within an area and do not necessarily have access to a visual display. These environments could use a voice notification in a break room or on a personal device to notify team members of a meeting, or to indicate that a meeting is nearing its end. In these situations, voice can inform all concerned people in the area regardless of current task. Extending from the concept of semi-public environments, and rising to meet the challenge issued by Nass, we focus on the different voice types that would be commonly heard in such an environment: the listener's own voice, a voice familiar to the listener, and an unfamiliar voice.

While other investigations of the impact of voice on human performance have studied factors such as perceived presence and mood, our work seeks to harness and understand its potential for notification-related goals. Previous work established three critical parameters that describe the goals of notification systems—interruption,

reaction and comprehension [5]. While desire for each of these parameters may vary, each is important to measure and understand in choosing an appropriate notification. By measuring user performance for various voices for 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 holds 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 variance in these parameters when using the different categories of voices. Based on these results we report on our development and deployment of a voice notification system called Notiframe that seeks to provide voice notifications with appropriate notification parameters.

EXPERIMENT

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 key critical parameter levels for each voice. Just as Gaver's work demonstrated the value of familiar non-audio sounds [2], we hypothesize that the three different voices—one's own, that of a familiar person, and that of an unfamiliar person—provide substantially different levels of interruption, reaction and comprehension. The experiment involved the users playing a simple computer game with some falling blocks and a little paddle to catch them. The environment included support for voice playback and recording prior to the study.

Twenty-seven volunteers participated in this experiment. Participants were recruited from an undergraduate class and were given the incentive of extra credit for taking part in the experiment. The experiment was conducted in a quiet computer lab, with each participant wearing a headset to hear the audio. Each session lasted approximately thirty minutes. Participants were required to first record the numbers 0-9 in their own voice. 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 class instructor volunteered to serve as the familiar voice. For the unfamiliar voice, we chose the voice of a person that none of the participants were familiar with—an individual with a French accent (we verified that none of the participants had ever regularly been exposed to a French accent). Since the class instructor's voice would not be as familiar as that of a friend or a co-worker, the choice of an accented voice to broaden the difference between familiar and unfamiliar seems reasonable.

Procedure

Before starting the experiment, the users were asked a series of questions to help us assess their different cultural and social backgrounds. Users were then given four practice rounds to familiarize themselves with the game and environment. The experiment itself consisted of nine rounds. During each round, the game was interrupted by a voice reading out a seven-digit number (the same length as a phone number). The users had to hit the space bar upon hearing the notification, then remember the numbers as they continued playing the game and enter them into a box at the end of the round. The users heard a different voice in each round. A Latin square design was used to control variation among the three voice types, with each user assigned to one of three groups in which the users heard each voice three times in different orders. Each round lasted for one minute, with voice notification approximately 25 seconds into the game and lasting for approximately seven seconds.

Calculating Voice Impact

Interruption was measured by the drop in game performance, determined by comparing the percentage of blocks caught before and after the voice notification. The change 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 space bar is reaction time measured in milliseconds. Correctness in remembering the numbers, entered at the end of the game, was used to calculate the comprehension parameter. Due to practical constraints, only this fairly short-term recall value was measured for comprehension.

RESULTS AND ANALYSIS

ANOVA test results suggested that there is a significant difference in the means for the reaction time among the three voices ($F(2,215)=3.74$, $MSE=48785.42$, $p=0.025$). The mean reaction time to the user's own voice ($M=831.83$, $SD=213.92$) was significantly faster than the familiar voice ($M=915.62$, $SD=225.8$), $t(143)=-2.29$, $p=0.023$. User's own voice was also significantly faster compared to the unfamiliar voice ($M=921.31$, $SD=222.78$), $t(144)=-2.47$, $p=0.014$. The t-test between familiar and unfamiliar voice did not reveal any significant differences. Thus, participants reacted most quickly to their own voice.

The slower reaction time for the unfamiliar voice corresponds with the idea that we tend to filter out voices of people we do not know. The quickest reaction time to one's own voice came as a surprise. Listening to one's own voice might have evoked an emotional response that relates to a self-image. This in turn may have triggered the fast reaction time that showed that you are acknowledging yourself. It must also be noted that the instructor's voice was not an extremely familiar voice for the student participants. The students had only been exposed to it in a classroom setting for about thirty hours. Our supposition is that as familiarity with a voice increases its reaction time will have similar reactive characteristics as that of your own voice; that is, very familiar voices like those of close friends, co-workers,

roommates etc. will have characteristics similar to that of your own voice.

An ANOVA test showed near significant difference for interruption for the three voices $F(2,173)=2.33$, $MSE=5.476$, $p=0.099$.This motivated t-tests and the t-test between one's own voice and the unfamiliar voice, revealed a significant difference between performance before and after the notification, with performance calculated by the percentage of balls caught from the total number of balls. The mean reduction in catch rate was significantly larger for the own voice condition ($M=4.06$, $SD=2.64$) than with the unfamiliar voice ($M=3.11$, $SD=2.08$), $t(112)=2.11$, $p=0.036$. Therefore, one's 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. While further study is needed to explore the validity of this result, this initial finding is encouraging.

Recall was consistently high for all three voice types (Own $M=6.76$, $SD=0.50$ Familiar $M=6.65$, $SD=0.69$ Unfamiliar $M=6.74$, $SD=0.54$ for numbers recalled correctly out of 7 numbers), with no statistically significant results to report. The consistently high recall, combined with a cognitively demanding primary task, suggests that voice in general should support adequate short-term recall for many situations.

	Own Voice	Familiar Voice	Unfamiliar Voice
Interruption	*M=4.06, SD=2.64*	M=3.53, SD=2.25	*M=3.11, SD=2.08*
Reaction	*M=831.8,SD=213.92*	*M=915.6, SD=225.8*	*M=921.31, SD=222.78*
Comprehension	M=6.76, SD=0.50	M=6.65, SD=0.69	M=6.74, SD=0.54

Table 1. Means and standard deviations for each condition. Sets with significant difference are italicized.

APPLICATION AREAS

The experiment gives us empirical insight into the variation of the notification characteristics with respect to the critical parameters. The differences in the three voices although significant are very small. This is probably due to the nature of the experimental setup. The empirical data reflects the fast pace of the game which was used as the primary task. We treat this data as mere indicators into the nature of differences between the voices. These differences can be better understood by developing and testing an application that uses the tree voice categories.

Semi-public environments allow for making complete utilization of the differences in own, familiar and unfamiliar voices. Semi-public environments, as introduced by Huang and Mynatt [3], 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 between the different voices. Motivated by these findings we chose to implement a unique semi-public notification system called Notiframe.

Announcing Meetings with Notiframe

To exercise our experiment results in real-world semi-public environments, we developed Notiframe, a voice notification system to notify people about the start and end of meetings. It uses an audio clip with the voice of the person who scheduled the meeting to signal meeting times, leveraging our experimental conclusions in that people are more likely to attend meetings that someone familiar (or they themselves) scheduled. The system was set up in a lab, whose primary users were about 15 students and faculty. The lab has multiple office/lab rooms and one conference room where most meetings took place. Most meetings in the conference room are attended by most lab members, but the room is occasionally used by others outside the lab.

Meetings are scheduled throughout the day through a web-based scheduling system. Although most meetings were regular events repeated on a weekly basis, users in the lab still often would lose track of time and miss the start of the meeting. Also, since there is little impetus to check the schedule regularly, ad-hoc meetings would often be interrupted by scheduled meetings. There was need for notification to provide timely and appropriate information about upcoming meetings. This would help users finish working and prepare for the meeting.

The existing online reservation system allowed users to reserve the conference room. The reservation system categorized meetings as internal and external. An internal meeting would involve lab members while an external meeting would involve users outside the lab. The booking system provided the necessary scheduling mechanism and data that Notiframe would need to make appropriate notifications. The internal and external events provided an ideal situation to use the differences between the voice types. For internal events the familiar voices of the people in charge could be used to attract attention of all the users in the lab. The "own voice" condition would hold for the person in charge. For external events, an unfamiliar voice of a person external to the lab was used. Thus, notifications made in the familiar voice would be relevant to the users in the lab while the notifications in the unfamiliar voice would not. The voices could announce the type of meeting and the time left before it begins thus providing all the necessary information in the notification itself.

Notiframe made three announcements—fifteen, ten and five minutes before the start of a meeting. Although Notiframe primarily uses a voice interface it was decided that it needed a physical or visual counterpart to accentuate its presence in the lab and to provide additional information. The visual interface shows the meeting scheduler's picture and day's schedule on a monitor next to the meeting room.

User Feedback

Notiframe was run in the lab for a period of two months, at the end of which user feedback was collected from six users in the form of a survey.

Users were asked to rate the various effects of the Notiframe system like annoyance and disturbance on a 5 point Likert scale. In addition users were also asked to give anecdotal feedback of any incident regarding Notiframe. Due to the limited number of users from whom feedback could be collected we focused on the subjective feedback received and present an aggregation of the observations from the user stories.

The unfamiliar voice gave rise to a certain amount of confusion, most notably when an external person scheduled significant time in the lab to run a user study. Most lab members were curious as to whose voice it was. They understood that the announcement was probably not meant for them but their inquisitiveness led to the disruption of normal activity. Impromptu conversations would break out between the users as they tried to guess who the voice belonged to. Confusion also arose due to ambiguity in the contents of the announcements. Because of implementation limitations wherein Notiframe had to be built on top of an existing system, not all voice notifications described what or who the meeting was for, leading to generic announcements. These announcements confused users and they usually had to check Notiframe's visual display or their personal schedule to determine whether the notification was meant for them. There was added confusion when the ambiguous announcement was made by the unfamiliar voice.

The goal of the voice notifications made by Notiframe was to generate appropriate reaction in the user. This goal was realized when users recognized that they had to attend a meeting and started preparing for it. The familiar voice was successful in this endeavor as it attracted attention from the users and provided them with enough information to elicit the right reaction.

OTHER APPLICATION AREAS

Numerous possibilities for applications exist for interfaces that take advantage of differences in reactions to voices, particularly for shared semi-public spaces where users of the area have common preconceptions of voice types and characteristics. Interfaces in break rooms could start and guide conversations around shared interests. Interfaces in hallways could opportunistically remind people of errands or deliverables.

We also see utility in the application of our results for on-the-move users. Building on previous successes of voice interfaces in in-vehicle systems, a next step is to integrate voice notifications into mobile devices to guide users, particularly in an unfamiliar environment. In our own campus guide project, we plan to use voice notifications to alert visitors to campus of interesting labs, contextual information about the area, and upcoming meetings—altering voices based on importance and relevance of the notification.

CONCLUSIONS

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 in semi-public environments. An experiment and case study point to interesting characteristics of voice notifications. Voice familiarity seemed to result in minimally interruptive notifications that prompted rapid and appropriate reaction.

As discussed in this paper's introduction, there are many more inherent social aspects of speech that may affect voice notifications. 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.

REFERENCES

1. Alexanderson, P. Peripheral Awareness and Smooth Notification: the Use of Natural Sounds in Process Control Work. In *Proc. NordiCHI* 2004, 281-284.
2. Gaver, W. W., Smith, R, & O'Shea, T. Effective Sounds in Complex Systems: The ARKola Simulation. In *Proc CHI* 1991, 85-90.
3. Huang, E. M. & Mynatt, E. D. Semi-Public Displays for Small, Co-located Groups. In *Proc. CHI* 2003, 49-56.
4. Lee, K. M. & Nass, C. Designing Social Presence of Social Actors in Human Computer Interaction. In *Proc. CHI* 2003, 289-296.
5. McCrickard, D. S., Chewar, C. M., Somervell, J. P., & Ndiwalana, A. A Model for Notification Systems Evaluation—Assessing User Goals for Multitasking Activity. *Transactions on CHI 10* (4), 2003.
6. Nass, C. & Gong, L. Speech Interfaces from an Evolutionary Perspective. *Communications of the ACM 43* (9), 2000, 36-43.
7. Sawhney, N., & Schmandt, C. Nomadic Radio: Scaleable and Contextual Notification for Wearable Audio Messaging. In *Proc.CHI* 1999, 96-103.