Attentive Headphones: Augmenting Conversational Attention with a Real World TiVo®

Aadil Mamuji, Roel Vertegaal, Changuk Sohn and Daniel Cheng
Human Media Lab
Queen's University
Kingston, ON K7L 3N6
{mamuji,roel,csohn,dc}@cs.queensu.ca

ABSTRACT
Computer users in public transportation, coffee shop or cubicle farm environments require sociable ways to filter out noise generated by other people. Current use of noise-canceling headsets is detrimental to social interaction because these headsets do not provide context-sensitive filtering techniques. Headsets also provide little in terms of services that allow users to augment their attentive capabilities, for example, by allowing them to pause or fast-forward conversations. We addressed such issues in our design of Attentive Headphones, a noise-cancelling headset sensitive to nonverbal conversational cues such as eye gaze. The headset uses eye contact sensors to detect when other people are looking at the wearer. Upon detecting eye gaze, the headset automatically turns off noise-cancellation, allowing users to attend to a request for attention. The headset also supports the execution of tasks that are parallel to conversational activity, by allowing buffering and fast-forwarding of conversational speech. This feature also allows users to listen to multiple conversations at once.

Author Keywords
Attentive User Interfaces, Eye Tracking, Audio Interfaces.

ACM Classification Keywords
H.5.1 Multimedia Information Systems: Audio input/output.

INTRODUCTION
With the availability of ubiquitous computers and public wireless access points comes an increase in the use of computing systems in public environments. Whether it is on public transportation, in coffee shops or cubicle farms, users increasingly work in environments in which the level of environmental distraction cannot be controlled. Focusing on the task at hand can be highly problematic when working in a public space. We believe it would be beneficial to design interfaces that support the user's attentional strategies for coping with environmental sources of distraction. A prevalent source of noise is presented by the conversational activity of others within the shared space. In busy office and public environments, active ad-hoc and co-located group communications among co-workers may distract others from completing their tasks. This paper explores how co-located communications may be regulated by sensing and augmenting co-worker attention [10].

In general, humans exhibit two ways of coping with interference from multiple conversational sources of information. The first strategy, called conversational turn taking, is only deployed in formal contexts such as during meetings [4]. Here the speaking behavior of others can be controlled through social protocol. By asking only one speaker to be active at any one time, the act of turn taking allows each listener in the meeting to focus the limited attentional resources of their brain onto a single speaker. However, there are many situations in which the use of conversational turn taking is undesirable, or impractical. In public transportation, coffee shop or cubicle farm scenarios, conversational activity of others cannot be controlled, and may thus present interference to others. In these situations, our brain copes by attenuating irrelevant auditory stimuli, a process known as the Cocktail Party phenomenon [2]. Here, the brain uses both environmental and semantic conversational stimuli to tune its attentive system to a single cohesive message from a single conversational source. While the Cocktail Party phenomenon helps us filter extraneous noise, it is by no means a perfect process. According to Gillie and Broadbent, this attentive mechanism is especially sensitive to disruption by information that is semantically related to the ongoing task [6]. Auditory stimuli present a particular challenge because of their omni-directional nature. McFarlane suggested giving users control over the delivery of auditory stimuli. According to him, this considerably improved task performance of users over control conditions. One way of offering users control over the volume of environmental noise in public spaces is to provide them with noise-cancelling headphones [11]. Noise-cancelling head sets are particularly successful at filtering out repetitive environmental noises, such as those produced by fans. While they are generally less successful in filtering out voices, they do attenuate these as well.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

Socially Translucent Technologies

The main problem of today's noise-cancelling headphone is that it creates an attentional barrier between users. This barrier reduces social translucence [5] to the wearer of the headset, as auditory signals for attention by co-workers come to be ignored. When we observed users wearing noise-cancelling headsets in cubicle farms, we noticed these devices essentially offer an all-or-nothing strategy of coping with environmental noise. Users either have their headset engaged and are working on a computer task, or they are in a conversation, with their headphones off. More importantly, we noticed that co-workers frequently have problems approaching a headphone user with sociable requests for attention. Because headsets filter out all environmental stimuli, when users are focused on their computer screen, they may not even notice the presence of a co-worker. As a consequence, we often observed co-workers resorting to shoulder taps and other physical means of requesting attention. The problem with this is that it typically crosses the boundaries of social rules of engagement [5]. In this paper, we discuss the design of a noise-canceling headset that is sensitive to social requests for attention. We augmented a pair of headphones with infrared sensors that detect when someone looks at the wearer, both from behind as well as from the front. The headphones are also equipped with a microphone that picks up the wearer's voice, and an infrared tag that transmits identity information to the infrared sensors on other people's headphones. Upon detecting eye gaze at the wearer, the headsets automatically turn off noise-cancellation. This allows users to decide whether to attend to any request for attention using normal social protocol.

PREVIOUS WORK

In [1], Basu and Pentland discuss a pair of Smart Headphones that detect and relay sounds in the environment through to the user's headset, but only if they were classified as human speech. Mueller and Karau improved upon this concept with Transparent Headphones: headphones augmented with an amplifier that picked up and modified sound in real time, before user auditioning [12]. One of the applications of this system was to help a user listen to mp3s while still being accessible to surrounding individuals [12]. By mounting proximity sensors on the headphones, the system detected when a person approached a user, presumably to engage in conversation. However, according to Vertegaal et al. [16], eyegaze is a much more accurate predictor of conversational engagement between individuals. In a busy subway station, for example, there may be many people walking in close proximity to the wearer. In such situations, Transparent Headphones would decide to pause content. To allow for social translucence it is critical that information about the orientation of body, head and eyes of co-located individuals is sensed [9,16]. While Mueller and Karau experimented with the use of infrared transceivers, they did not sense eye gaze. More importantly, their headphones did not offer TiVo®-like features such as buffering and fast-forwarding of real-world conversations [12]. To manage periods of distraction in telephone conversations, Deitz and Yerazunis discuss their use of real-time audio buffering techniques [3]. While the phone handset is away from the user's ear, incoming audio is recorded in a circular buffer. Two pointers are used to indicate where to start and stop accelerated audio playback. Using time-compression and pitch preservation algorithms, they allow users to quickly catch up to real-time phone conversations without the loss of information [3].

Look-To-Talk

Eye contact is one of the few nonverbal cues that cross-culturally indicate conversational engagement with another person. In four-person conversations, the looking behavior of an individual indicates with about 80% accuracy whom that individual is addressing or listening to [16]. This means that the sensing of gaze from an onlooker provides an excellent indicator of interest. Wizard of Oz experiments [13,10] also evaluated the mechanics of gaze in speech-enabled environments. Studies found that subjects typically looked at a device before issuing a spoken command, regardless of whether eye gaze was actually processed as input by the device [13]. Devices that use eye gaze to initiate listening are known as Look-to-Talk interfaces. By mounting an eye tracker on a pair of headphones, we designed a similar system, one that disables noise-cancellation upon eye gaze at the headset.

Sensing Eye Contact

An eye contact sensor (ECS) is essentially an inexpensive eye tracker that detects whether a person is looking at the sensor or not. It requires no prior calibration of any kind (see Figure 1). We designed a sensor that can be built cheaply, consisting of a camera that finds pupils within its field of view using a simple computer vision algorithm [14]. The ECS consists of an infrared camera with a set of on-axis infrared LEDs mounted around the camera lens (see Fig. 1). When flashed, these produce a *bright pupil* reflection (*red eye* effect) in eyes within range. Another set of LEDs is mounted off-axis from the camera lens. Flashing these produces a similar image, with black pupils. By synchronizing the LEDs with the camera clock, a bright and dark pupil effect is produced in alternate fields of each video frame. A simple algorithm finds any eyes in front of the camera by subtracting the even and odd fields of each frame [14]. The LEDs also produce a glint on the cornea of the onlooker's eyes. These appear near the center of the detected pupils when the onlooker is looking straight at the camera, allowing the detection of eye contact. When mounted on a device, the eye contact sensor obtains information about the number and location of pupils in its field of view, and whether these pupils are looking at the sensor. It reports this information wirelessly over a TCP/IP connection to a connected server. ECS data is typically filtered by this server, with eye contact reported only when the amount of gaze over time exceeds a user-defined threshold.

Figure 1. Attentive Headphones with embedded eye contact sensors (front and back) and microphone.

Sensing Social Proximity and Identification

In addition to sensing eye contact towards a wearer, ECSes also allow the sensing of social proximity information [8]. According to Hall [8] there are four zones of social proximity: *intimate* (0- .45 m); *personal* (.45- 1.2 m); *social consultative* (1.2- 3 m); and *public* (over 3 m). Most conversational activity occurs within a personal zone of social proximity. ECSes can determine social proximity cues by measuring the distance between detected sets of pupils. The ECS calculates the approximate distance of an onlooker by determining his Interpupillary Distance, and comparing this measure to a known mean of 6.2 cm in the general population [8]. Eye contact sensors can also be used to uniquely identify other eye contact sensors [14]. This is accomplished by using one of the IR LEDs on the ECS to send a unique binary identifier through a pulse code modulated infrared beam. This is used to allow an attentive headphone to detect *who* is looking at their wearer.

ATTENTIVE HEADPHONES IMPLEMENTATION

An Attentive Headphone consists of a Bose™ noise-cancelling headphone augmented with two eye contact sensor, one pointing to the front, and one pointing to the back, as well as a microphone (see Figure 1). We modified the headset with a circuit that allows noise-cancellation to be switched on or off wirelessly through an X10 interface [17]. When the headset is turned off, this allows wearers to hear each other normally. When the headset is turned on, ambient sound is attenuated by –20 Db. Sound from the microphone is sent through a wireless connection to a server that buffers and relays it to other headsets.

Monitoring User Attention with Multiple ECSes

When the wearer is engaged in a computer task, visual requests for attention outside the wearer's field of view are detected by an eye contact sensor on the back of the headset (Figure 1). Griffin and Bock showed that participants tended to fixate on a given entity in a scene roughly 900 milliseconds before verbally referring to it [7]. To avoid unintentional triggering, the back ECS only reports fixations that are longer than 1 second. This time interval can be adjusted manually according to user preference. Similarly, the headset can detect when the wearer is in a conversation by polling the second ECS, mounted toward the front of the headset (see Figure 1). This ECS scans the eyes of individuals standing in front of the wearer in order to predict when the wearer is likely to be engaged in conversation [14]. The front ECS reports only on pupils within about 1.2 meters, the *personal* social proximity zone [8]. The information from multiple ECSes is integrated through the user's personal attention server [14]. This EyeReason server determines which device or person the user is likely to be engaged with, by polling all eye contact sensors associated with that user. To altogether avoid interference from other eye contact sensors, the EyeReason server turns off the illuminators on the front of the speaker's headset upon detecting the presence of speech by its wearer. A single LED on the speaker's headset continues to blink a unique binary identification signal. This essentially allows ECSes in conversational groups to take turns in determining the speaker's ID, in a process driven by actual conversational turn taking, allowing each headset to uniquely identify the members in the current social network.

Attentive Headphones Operation

When the headphones detect eye contact by an onlooker, the EyeReason server responds by temporarily turning off noise cancellation on the headset, pausing any audio currently playing on the headset. This allows the voice of the potential interlocutor to be heard by the wearer, and the request to be serviced according to social protocol. It also functions as a subtle ambient notification of the pending request for attention. The user's EyeReason server determines when the user responds to the request by detecting eye contact with people in front of the user within a user-defined interval. When eye contact is not detected within that period, noise-cancellation is again engaged, and any audio playing on the headset is resumed. When eye contact is detected, noise cancellation is left off instead, allowing the wearer to have a normal conversation with the interlocutor. When the user ends the conversation and returns to his task, this is detected by loss of eye contact with the frontal ECS. When this occurs, headset noise cancellation is engaged. Any content previously playing in the headphones smoothly fades in, and continues where it was paused.

Augmenting User Attention with TiVo® for the Real World

Even without noise cancellation, the headphones tend to attenuate sound from the outside world. To alleviate this issue, sound from the microphone mounted on the headset can be relayed to other headsets within the current social network, as determined by eye contact between headset wearers. This further improves the signal to noise ratio of sound from attended individuals.

We were particularly interested in experimenting with ways in which we could boost the user's attentional capacity. To achieve this, we experimented with the use of buffering techniques similar to those of a TiVo® personal video

recorder [15]. For this purpose, each wearer's EyeReason server continuously records audio streams from individuals engaged within his or her social network. A button on the headset allows users to pause live conversational audio, for example upon receiving a cell-phone call. This allows them to attend to the call without loosing track of the ongoing conversation. Pressing the button a second time plays back the recorded conversation at double speed, without affecting its apparent pitch [3]. Buffering can be set to trigger automatically upon servicing an incoming phone call.

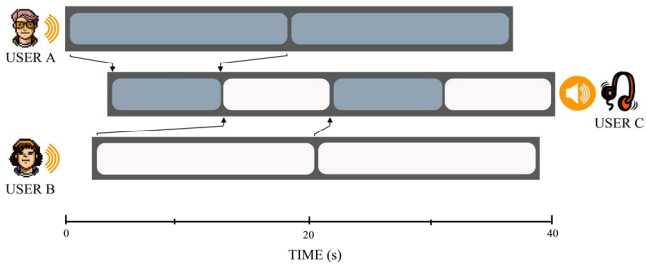

Figure 2 – Time-multiplexing buffered speech from two simultaneous conversations

Attending to Two Simultaneous Speakers

We are also experimenting with the use of time multiplexing techniques that would allow users to attend to two speakers at once. When two individuals A and B, within a user C's current social network begin talking simultaneously, user C's EyeReason server begins an automated turn taking process in which it plays back recorded speech from A and B at twice the speed in a time-multiplexed fashion. Since the voices from user A and B are recorded separately on user C's EyeReason server, they can be time shifted and relayed independently to user C's headset. After a user-specified buffering delay, first user A's recorded speech is played back at double speed to user C, after which user B's speech is played back at double speed, allowing user C to listen to both contributions in real time. This process is stopped when either user A or B falls silent (see Figure 2). When this happens, the remaining buffer is first played back, after which user C can respond. Initial experiences suggest this time-multiplexing technique is most advantages in cases where two individuals simultaneously request the attention of a third, for example, to ask that person a question. It is less appropriate during group conversations, where user A and B might both be interested in hearing each other's contributions. However, in such cases, user A and user B may choose to use their pause button to buffer each other's speech for playback after they have finished speaking. We are currently investigating the implications of the above scenarios on comprehension, as well as the conversational turn taking process in small groups.

CONCLUSIONS

Attention management can be problematic in public places such as cubicle farms, where many co-workers share the same space. While the use of static noise-cancelling headsets may help alleviate distraction by ambient noise, they also place constraints on co-worker awareness of social interactions. In this paper, we presented Attentive Headphones, a noise-cancelling headset sensitive to social requests for attention of its user. The headset uses eye contact sensors to detect when another person is looking at the wearer. Upon detecting eye gaze, the headset automatically turns off noise-cancellation, while pausing any content played, allowing users to attend to the request. The headset also supports the execution of parallel tasks by allowing buffering and fast-replaying of live conversations. One interesting use of this feature is that it allows users to listen and respond to multiple conversations at once.

REFERENCES

1. Basu S. and Pentland A. Smart Headphones. In *Extended Abstracts of CHI 2001*. Seattle, 2001, pp. 267-268.
2. Cherry, C. Some experiments on the reception of speech with one and with two ears. In *Journal of the Acoustic Society of America 25*, 1953, pp. 975-979.
3. Deitz, P., Yerazunis, W.: Real-Time Audio Buffering for Telephone Applications. In *Proceedings of UIST 2001*. ACM Press, New York, 2001, pp. 193–194.
4. Duncan, S., Some Signals and Rules for Taking Speaking Turns in Conversations. In *Journal of Personality and Social Psychology* 23/2, 1972, pp.286-288.
5. Erickson, T., and Kellogg, W. Social Translucence: An approach to designing Systems that Support Social Processes. In *ACM Transaction on Computer-Human Interaction*, 7, No. 1, March 2000, pp. 59-83
6. Gillie, T. and Broadbent, D. What makes interruptions disruptive? A study of length, similarity and complexity. In *Psychological Research 50*, 1989, pp. 243-250.
7. Griffin, Z. M. and Bock, J. K., What the Eyes Say About Speaking. In *Psychological Science,* 11, 2000, pp. 274-279.
8. Hall, E.T. *The Hidden Dimension*. Garden City, NY, USA. Doubleday, 1966.
9. Hudson S. et al., Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of CHI 2003*, ACM Press, pp. 257 – 264.
10. Matlock, T., et al. Designing feedback for an attentive office. In *Proceedings of INTERACT'01*, 2001.
11. McFarlane, D. Coordinating the interruptions of people in human-computer interaction. In *Proceedings of INTERACT '99*, pp. 295-303.
12. Mueller F. and Karau M. Transparent Hearing. In *Extended Abstracts of CHI'02*. Minneapolis, 2002, pp. 730 – 731.
13. Oh, A. et al. Evaluating Look-to-Talk. In *Extended Abstracts of CHI 2002*, Minneapolis: ACM Press, 2002.
14. Shell et al. ECSGlasses and EyePliances: Using Attention to Open Sociable Windows of Interaction. In *Proceedings of ACM ETRA'04*, 2004.
15. TiVo® Inc., http://www.tivo.com, 2001
16. Vertegaal, R., Slagter, R., Van der Veer, G., and Nijholt, A. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of CHI'01*, 2001, pp. 301-308
17. X10 Home Solutions, http://www.x10.com, 2003