OverHear: Augmenting Attention In Remote Social Gatherings Through Computer-Mediated Hearing

David Smith, Matthew Donald, Daniel Chen, Daniel Cheng, Changuk Sohn, Aadil Mamuji, David Holman and Roel Vertegaal
Human Media Lab, Queen's University
Kingston ON Canada K7L 3N6
smith@cs.queensu.ca

ABSTRACT
One of the problems with mediated communication systems is that they limit the user's ability to listen to informal conversations of others within a remote space. In what is known as the Cocktail Party phenomenon, participants in noisy face-to-face conversations are able to focus their attention on a single individual, typically the person they look at. Media spaces do not support the cues necessary to establish this attentive mechanism. We addressed this issue in our design of OverHear, a media space that augments the user's attention in remote social gatherings through computer mediated hearing. OverHear uses an eye tracker embedded in the webcam display to direct the focal point of a robotic shotgun microphone mounted in the remote space. This directional microphone is automatically pointed towards the currently observed individual, allowing the user to OverHear this person's conversations.

Author Keywords
Attentive User Interfaces, Eye Tracking, Audio Interfaces

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
The idea of creating interfaces that could augment human intellect has been explored early-on by researchers such as Engelbart [5]. The mouse was initially intended as a tool for augmented intelligence. It allows for a reduction of the user's cognitive load by shifting information processing to the user's motor brain. Later on, researchers such as Mann explored augmentation of human senses through mediated reality, or "Humanistic Intelligence" (HI) [6]. Mann designed a pair of EyeTap digital eyeglasses which, coupled with a wearable computer system, could help as a visual intermediary for observing the real world. For example, EyeTap digital eyeglasses recognize and block unwanted distractions, such as advertisements, in the environment, replacing them with computer terminals [6].

In this paper, we explore the use of a media space as a mediated reality interface. Although media spaces were originally designed to support informal gatherings over video conferencing links, there are many issues which prevent them from operating as such. One of the problems with media spaces is that they are typically designed to support only one person per remote location. For example, when used as a video wall that connects groups of individuals between two remote locations, distinct asymmetries appear in interactions between remote and local individuals. One of the factors contributing to this problem is that media spaces do not typically preserve the full richness of 3D visuo-spatial and auditory-spatial cues [9]. As a consequence, when connecting noisy, crowded gatherings of people between two remote places, individual voices fuse into an incomprehensible noise. We addressed this issue in our design of OverHear, a media space that augments the user's attention in remote social gatherings through computer mediated hearing. OverHear uses an eye tracker embedded in the display to direct the focal point of a robotic directional microphone mounted in the remote space. This directional microphone is automatically pointed towards the currently observed individual, allowing the user to attend to this person's conversations.

BACKGROUND
In general, humans exhibit two ways of coping with interference from multiple conversational sources during social gatherings. The first strategy, called conversational turn taking, is typically deployed in formal gatherings, such as 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. Through exchange of nonverbal cues such as eye contact, humans are capable of achieving a very efficient speaker switching process. However, turn taking is only effective in cooperative situations. In public locations like restaurants, subway stations or coffee shops, many speakers may be active simultaneously, without any means of floor control. In these more populated scenarios, humans focus their

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

Figure 1: OverHear remote display with full screen video stream and integrated eye tracker

Figure 2: OverHear robotic directional microphone with webcam.

attentional capacity on a single speaker through what is known as the *Cocktail Party phenomenon* [2]. Here, the brain uses visuo-spatial, auditory-spatial and semantic conversational cues to tune its attentive system to a single cohesive message from a single conversational source. These cues are typically lost in mediated communication systems, making it difficult to support attentional focus on a single speaker when attending a mediated informal gathering [10]. According to Vertegaal et al. [9], eye contact is an excellent predictor of attentional focus in group conversations. According to him, in four-person conversations, the looking behavior of an individual indicates with about 80% accuracy whom that individual is addressing or listening to. This means eye gaze of listeners provides an excellent way to measure conversational attention during mediated gatherings.

Systems Deploying Cocktail Party Filters

A number of prior systems exist that deployed eye tracking to provide cocktail party filtering effects. One of the first such systems was Gaze-Orchestrated Dynamic Windows by Bolt [1]. It simulated a composite of 40 simultaneously playing television episodes on one large display. All stereo soundtracks from the episodes were active, creating "a kind of Cocktail Party Effect mélange of voices and sounds". The system sensed via a pair of eye tracking glasses when the user looked at a particular image, turning off the soundtracks of all other episodes. OverHear was also inspired by GAZE-2 [11], a conferencing system that provided cocktail party filtering of remote conversations on the basis of participant looking behavior. GAZE-2 used an eye tracker to determine whom people look at during a remote meeting. It used this information to automatically boost the volume from attended participants during side conversations. Video images of individuals could also be enlarged upon sustained eye contact. The use of cocktail party filtering allowed GAZE-2 to provide a more natural support of side-conversations.

OVERHEAR IMPLEMENTATION

OverHear consists of a video camera and a robotic shotgun microphone located at the remote scene. A local display allows the user to monitor the images from the remote camera. This display contains an eye tracker that monitors the user's eye fixations within the remote scene. Whenever the user fixates his eyes on a person in the remote scene, the system instructs the remote robotic microphone to point in the direction of that person inside the remote space. The system is calibrated such that when the user looks at a 2D coordinate on his display, the robotic microphone picks up sound from its corresponding 3D coordinate at the remote location. This allows OverHear users to single out individual conversations inside the remote space in a manner that is consistent with a naturally occurring Cocktail Party phenomenon.

Architecture

The robotic arm in the remote location consists of a shotgun microphone that is attached to the motorized platform of a Sony EVI-D30 pan-tilt camera. An inexpensive 640 *x* 480 webcam, mounted on the base of this platform, is used to capture a video stream of the remote scene. The motorized platform, microphone, and webcam are connected to a laptop placed in the remote location. The laptop controls the motorized platform via a VISCA connection. The laptop uses the Java Media Framework to broadcast an audio stream from the microphone and a video stream from the webcam to the local computer. This local computer is connected to a 17" Tobii 1750 EyeTracking display [8]. The eye tracker integrated into this display continuously monitors the user's point of gaze, with accuries in the order of .5 degrees of visual angle. Live images from the remote scene are shown in full screen on this Tobii display, with

the sound from the shotgun microphone played on an adjacent pair of speakers.

Robotic Microphone Calibration

During deployment, the webcam and robotic microphone platform are mounted on a wall in the remote location. The webcam is placed such that a suitable image of the remote scene is captured. Subsequently, the robotic microphone is calibrated such that its angular coordinates correspond to the 2D visualization of the remote scene, as captured by the webcam. To allow calibration of the system, a low-powered laser pointer is temporarily mounted on top of the robot arm. This causes a red dot to appear at the coordinate the microphone is pointed at within the remote scene. Prior to calibration, the microphone platform is positioned such that this laser pointer dot appears in the top-left corner of the webcam display when the robot arm is pointed towards its top left coordinate, and in the bottom-right corner of the screen when the robot arm is pointing towards its bottom right coordinate. The laptop subsequently instructs the robot arm to step through a pattern of 15 calibration points spanning the maximum and minimum range within its bounds in both degrees of freedom. At each calibration point, the robot pauses until the user clicks on the 2D location of the laser pointer dot, as shown in the webcam display. This creates a bounding area within the display that the microphone is capable of surveying, and maps the angular coordinates of the robot arm to the 2D coordinates of the webcam display. The Tobii eye tracker in the local space is calibrated in a similar fashion, allowing the local user's eye movements to be mapped to the 2D coordinates of the webcam display. 2D eye coordinates on this display are sent over TCP/IP to the laptop in the remote space, where they are translated into an angular robot arm coordinate through linear interpolation of calibration coordinates. After calibration, the laser pointer is removed from the robot arm.

Mediating User Attention

To avoid continuous operation of the robotic arm, user fixations on the Tobii display are filtered using a 500 ms. dwell-time threshold [3]. After detecting a fixation inside the remote scene, the local OverHear client sends a command over the network to the remote server, which directs the motorized platform to point the microphone towards the object of interest in the remote scene. Tests over our wireless campus network suggest a minimum latency between remote capture and local presentation of about 500 ms. The average latency between detection of a user eye fixation and positioning of the robotic microphone is in the order of 500 ms. This means the overall latency between observing an event on the local display and observing the sound from that location is in the order of 2 seconds, which is well within the boundaries of conversational turn taking behavior [4]. The directional microphone does not exclusively detect the sound coming from the object the user is looking at. Instead, the effect is quite subtle. The directional microphone raises the sound pressure level of sound from the desired location, while attenuating ambient noise present in the remote scene. The horizontal coordinate of the motorized controller is used to determine the stereo presentation of audio on a pair of speakers attached to the local client. This allows for an experience that is very similar to that of a naturally occurring Cocktail Party phenomenon.

APPLICATION SCENARIOS

As a media space technology, OverHear allows for a redefinition of the boundaries of connected spaces. More specifically, OverHear extends the attentive system of an onlooker to the remote space by establishing a virtual tunnel of attention between a local listener and remote speaker. As such, it can be viewed as an auditory version of Auramirror [7]. Auramirror is a video mirror that uses eye gaze to determine when two interlocutors are talking to one another. When interlocutors are looking at each other, Auramirror paints a fluid dynamic tunnel between them, thus symbolizing the window of attention that connects the two interlocutors. However, it is important to note that the tunnel of attention provided by OverHear only operates in one direction. By contrast, in face-to-face situations, the attentive connection provided by the Cocktail Party phenomenon is reciprocal.

We will illustrate the operation of OverHear using two distinctively different scenarios. The first scenario demonstrates OverHear's ability to connect two separate locations via a unidirectional attentive tunnel. In the second scenario illustrates OverHear's application as a surveillance interface.

Scenario 1: Sampling Cocktail Party Atmospheres

Jeff is attending a reception that celebrates the opening of his art exhibit in San Francisco. While Jeff is enjoying a glass of wine, a second exhibit of his work begins its opening reception 800 miles away in Seattle. To connect the exhibits, the gallery owners have deployed OverHear displays at both galleries. Jeff chose the San Francisco gallery because it is closer to home. However, he wonders about the opinions on his work in Seattle. Jeff sits down in a quiet side room of the gallery, where an OverHear display shows the live broadcast from Seattle. He notices two attendees in Seattle standing in front of what he considers to be one of his best works. Due to the noise levels in the Seattle gallery, he cannot hear what the two attendees are saying. As he looks at their image on the OverHear display, the remote robot arm points its shotgun microphone towards the two interlocutors. About a second later, sound from the two interlocutors is boosted in San Francisco, allowing Jeff to overhear their conversation. He is pleased to find out the two attendees are enjoying the artwork, and returns to the local gallery to sample the atmosphere there.

Scenario 2: Augmented Surveillance

Mark is a professor at a university in Canada. Unfortunately, the final exam for his undergraduate course is during a conference in Santa Fe, where he will be demoing a new eye tracking display. While his teaching assistants were so kind to offer proctoring of the exam, Mark is concerned there might be some questions they cannot answer. As part of his research program, he has deployed an OverHear system in the remote classroom. It is connected to the eye tracking display set up in his hotel room in Santa Fe. To answer questions, Mark uses an instant messaging system that is connected to a projector screen in the exam room. As Mark scans the room, he notices one of the students puts up her hand. As he looks at her image on the OverHear display, the remote robot arm points its shotgun microphone towards the student. Its platform is mounted high on the wall of the classroom, and allows no tampering. The student notices the robotic arm pointing towards her from a distance. So as not to disturb other students, she whispers her question. The sound is picked up by the shotgun microphone and broadcast to her instructor. Mark answers the question with an instant message that is projected on the classroom screen. After answering the question, Mark notices two of his students, Mike and Daniel, bending towards each other on the far side of the classroom. He looks at them to overhear their conversation, and is relieved that Mike is only asking Daniel how much time is left. After the exam is done, a teaching assistant comes in to pick up the essays. Mark returns to the conference, satisfied that he was able to support his students' exam effort as if he had been there.

INITIAL USER EXPERIENCES

The second scenario discusses one of the more intriguing applications of OverHear, that of unobtrusive surveillance of conversations in a remote scene. Because of the ethical implications of this application, we were particularly interested in the attitudes of those surveilled. We deployed OverHear on our university campus to survey a remote classroom. Initial observations of students using the system suggest the user experience is sufficiently similar to the naturally occurring Cocktail Party phenomenon for it to be perceived as transparent. Evaluations show OverHear boosting the sound level of conversations in the direction of the microphone by 5 dB, allowing conversational speech to be picked up from up to 10 m., depending on noise levels. As we expected, students were highly concerned when we suggested the system could be deployed for remote surveillance of classrooms by professors. Their concern led us to perform a wider survey on campus, which solicited the response of 15 randomly selected individuals. Results suggest a large majority of students (91% of respondents) felt uncomfortable when presented with the surveillance scenario. 75% of respondents said they did not like the idea of someone overhearing their conversations from a distance, with or without their knowledge. We considered the possibility that students' discomfort could be alleviated using a sousveillance device [6], such as a personal tape recorder. 41% of respondents indicated they would feel more comfortable being surveyed with a system such as OverHear if they had their own copy of the surveyed conversation. We conclude that deployment of the OverHear system should be considered only in cases where permission can be obtained of those surveilled.

CONCLUSIONS

In this paper, we presented OverHear, a media space that augments the user's attention in remote social gatherings through computer-mediated hearing. OverHear uses an eye tracker embedded in the display to direct the focal point of a robotic directional microphone mounted in the remote space. This directional microphone is automatically pointed towards the currently observed individual, allowing the user to OverHear this person's conversations.

REFERENCES

1. Bolt, R. A. "Gaze-orchestrated dynamic windows". In Proceedings of ACM SIGGRAPH Computer Graphics Conference (Dallas, Texas, August 3-7). ACM, New York, 1981, pp. 109-119.
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. Duchowski, A. *Eye Tracking Methodology: Theory & Practice*. Berlin: Springer-Verlag, 2003.
4. Duncan, S. "Some Signals and Rules for Taking Speaking Turns in Conversations". Journal of Personality and Social Psychology 23, 1972.
5. Engelbart, D.C. "A Conceptual Framework for the Augmentation of Man's Intellect". Vistos in Information Handling, P.D. Howertown and D.C. Weeks, Eds, Spartan Books Washington, D.C., 1962.
6. Mann, S. "Wearable Computing: Towards Humanistic Intelligence". Guest Editor's Introduction, IEEE Intelligent Systems Special Issue. Vol 16, No. 3. 2001.
7. Skaburskis, A. W., Shell, J.S., Vertegaal, R., and Dickie, C. "AuraMirror: Artistically Visualizing Attention". In Extended Abstracts of ACM CHI 2003 Conference on Human Factors in Computing Systems, 2003.
8. Tobii Technology website: http://www.tobii.se.
9. Vertegaal, R. "The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration". In Proceedings of ACM CHI'99 Conference on Human Factors in Computing Systems. Pittsburgh, PA USA: ACM, 1999.
10. 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.
11. Vertegaal,R., Weevers,I., Sohn, C. and Cheung, C. "GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction". In Proceedings of CHI 2003 Conference on Human Factors in Computing Systems, 2003.