Back Talk: An Auditory Environment for Sociable Television Viewing

Andrea Colaço
Media Lab, MIT
Cambridge, MA 02139
Email: acolaco@media.mit.edu

Ig-Jae Kim
Media Lab, MIT
KIST, South Korea
Email: ijkim@mit.edu

Chris Schmandt
Media Laboratory, MIT
Cambridge, MA 02139
Email: geek@media.mit.edu

***Abstract*—Video content is being consumed in a host of new ways - viewers are no longer restricted to same-time or same-place viewing. However, the experience of watching with a group is inherently social and often desirable despite the physical distribution of group members. This paper introduces Back Talk, a system designed to create a sociable television watching experience. We enhance TV viewing with an auditory environment around a listener. We have explored and leveraged the richness of audio to convey presence of remote viewers. We have developed a novel framework for capturing and translating engagement of an individual into a set of audio cues that are played spatially around a listener. Such auditory enhancements can augment video content consumption in the future.**

I. Introduction

With the advent of the web and on-demand viewing options, consumption of video content in general and television content specifically has been "individualized". But, some content and experiences are consumed better when shared within one's social circle. However, it is not always possible or desirable to be physically present with friends when watching video content. Motivated by typical living room interactions that afford communication and *peripheral awareness* of co-viewers, Back Talk mimics this setting by creating audio based co-presence of non-collocated friends. Our aim is to offer free form interaction that characterizes unmediated person-person communication in the setting of TV viewing. The viewer is provided with co-presence information that includes one's response to TV programs such as laughter, emotional arousal, and gaze direction.

Back Talk uses a combination of sensing modules to capture engagement data and translates this data into pre-defined audio cues. The sound generation process requires two different sources of input: spoken communication and what is sensed beyond.

Consider the following scenario: *Tom and his friends regularly watched the TV series LOST together when in college. Recently, their respective jobs have required them to relocate to different cities. However, they can still catch up together every week on their virtual couch using Back Talk. Tom invites his buddies to their virtual couch (Figure 1). They turn on their televisions and are ready to start. They have a voice channel that allows them to communicate with each other. Half way through the show, Layla who was running late from work joins her friends using Back Talk. Immediately, her buddies hear a set of footsteps indicating her presence. When Matt's friend, John, has to leave twenty minutes into the show, remote co-viewers are signaled with the sound of a door shutting. They communicate frequently during the show and laugh at Tom's futile attempts to defend his favorite character's machinations. Overall, they have an enjoyable experience.*

In sections that follow, we describe our system design and motivation (Section II); prior work (Section III); implementation (Section IV); results of our evaluation study (Section V) and implications and limitations of our system (Section VI).

II. Motivation

In our attempts to foster a sociable experience around a group activity we have to create *co-presence* of remote participants. Co-presence is the participation of a group of people in a common activity or experience, and it can be either *virtual* or *real* [13]. It is characterized as instant and two-way, and is feedback-based. Co-presence is also a vital component of group sociability[12]. We apply this knowledge to plug back social experiences into a group activity: television viewing. By anchoring co-presence information to an activity, it is easier to supply context information.

Since users of our system are physically distributed, it raises questions about what elements we capture from these remote viewers and how we aggregate it into useful presence information. From a range of possibilities, we chose to detect - *1*) number of people watching, *2*) people entering and leaving, *3*) laughter, *4*) arousal (overall activation) and *5*) spoken comments. These were selected to be representative of general activity in a room with viewers and also within technological realization of the prototype. Additionally, since the goal of the prototype was to pick up activity and general attentiveness automatically, we chose elements that met both these conditions.

At the level of the listener, this sensed data is translated into audio information. Each sensed element is played as an audio cue. Table 1 lists triggers and their corresponding audio cues. In order to create a sense of a group of people around the viewer; we spatially place these cues to the left/right of the viewer.

An obvious question that arises is about the audio environment potentially disturbing or annoying users. We contend that in a typical social television viewing experience one would be

978-1-4244-8790-5/11/$26.00 ©2011 IEEE

TABLE I
CAPTURED ENGAGEMENT VALUES AND CORRESPONDING AUDIO CUES

Trigger	Audio Cue
Entry	Footsteps
Exit	Door closing
Laughter	Canned laughter
Emotional Arousal	Mild to moderate rustling
Gaze Direction (left/right)	Light to prolonged yawn

surrounded by people and hence tolerant to some baseline level of ambient talk and noise. This ambient noise, however, can meaningfully come to the center of our attention whenever our co-viewers seek to communicate with us - our system attempts to create just the right level of auditory cues to enhance the user experience without being annoying. Additionally, by providing users control over how much audio they transmit and receive, they have an option to prevent the acoustic environment from interfering with the viewing experience.

By providing an option for transmitting voice directly to viewers we capture more expressivity than text used for the same purpose[3]. In conveying social presence in as close a fashion as face-face interactions it has been found that often audio/video communication was closer to the audio-only medium despite being higher in ranking in its potential to convey social presence[7] . Based on these studies, we justify our use of an audio environment without using a video channel between participants.

III. RELATED WORK

Presence information of remote viewers in the context of television has been provided using traditional forms - IM chat and voice channels - for example, AmigoTV[4], 2BeOn[2]. Motorola Labs [9], have evaluated multiple iterations of systems that create presence in television viewing. Their solutions for communication between participants ranged from lightweight text messaging to voice and video. Back Talk is similar to these systems in using a voice channel between remote viewers. However, our model attempts to place these voice streams spatially to create a sense of being surrounded by viewers in the same room. A related study by Motorola[6] used ambient displays for augmenting television viewing. It used an orb and Chumby that lit in different colors to indicate the number of friends watching television at any time. NeXtream[8] uses a smartphone as the controller, and for accessing ones social network, albeit through a chat feature. It also provides a social layer through a collaborative filtering model of content selection.

Audio only environments have been explored earlier, but, primarily in the setting of workspaces [10]. These applications also used audio cues to represent events - in a more social and relaxed setting like television watching, one would expect quite a different set of audio cues. Further, our goal is to create a sense of being with as many people as are watching with you and so the intrusiveness condition is a little more relaxed in our setting.

Vision Television[1] creates presence of remote co-viewers by detecting faces of viewers and overlaying them at the bottom of the television video. Back Talk conveys similar information (number of faces, direction of gaze) through appropriate audio cues. By doing so, we do not interfere with the viewer experience of watching video.

IV. SYSTEM ARCHITECTURE

This section describes components of the Back Talk prototype. The main components are the cell phone, the auditory environment, the engagement capture modules, and the Back Talk server.

A. Cell phone Interface

Fig. 1. Cell phone interface showing sonic avatars

Back Talk users access their co-viewing buddies primarily through their cell phones. The selection of the phone precludes the need for an additional controller - it is a portable device that users would almost always carry with them. Figure 1 shows the interface depicting a virtual couch. Each remote co-viewer is represented as a "sonic avatar" [1]. These sonic avatars are movable icons that are associated with the physical listening space around the "primary listener" (represented as a star icon in the interface). When a user wishes to quiet an audio stream from a co-viewers, she can do so by dragging the sonic avatar outside the "audio circle". This results in muting the manipulated avatar (Figure 1). Controls at the bottom of the interface allow a user to limit the amount of audio she transmits to remote co-viewers by selecting "cues-only" mode. If a user turns on "I'm always-on" mode, the system is designed to transmit cues and detect and transmit whenever the user speaks. While playing an audio cue for a particular co-viewer the corresponding sonic avatar is highlighted visually with a volume icon (Figure 1). This additional feedback is intended for the viewer to have a quick glance at the source of the audio cue in case spatial distribution does not provide sufficient disambiguation of the source exuding a particular cue.

B. Auditory Environment

A key contribution of this system is the auditory environment that provides remote co-viewers a new way of being co-present with friends. In the Back Talk system we have

[1]We refer to these representations as sonic avatars because they act as a source of audio in a user's physical surroundings

two classes of sound sources: natural sounds from users and synthetic sounds (cues) to indicate activity. All these different sound sources are mixed into a stereo signal, where location in one dimension is obtained by left/right panning of each sound source. The location of a sound stream from a sonic avatar is mapped to its location on the phone screen. Audio cues and spoken comments are heard through a set of speakers on either side of the viewer's couch (Figure 2).

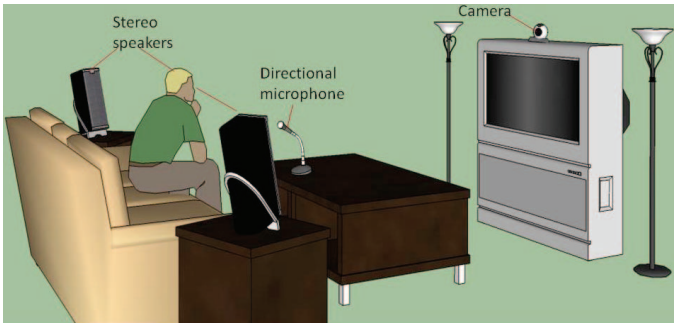

Fig. 2. System set-up

C. Engagement Sensing

Our prototype has three sensing modules that pick up engagement data and convey it to the Back Talk Server. These are the *visual module* - for detection of number of faces and gaze direction; *audio processing module* - for detection of laughter and spoken comments; and *galvanic skin response sensing* - for detection of sudden arousal.

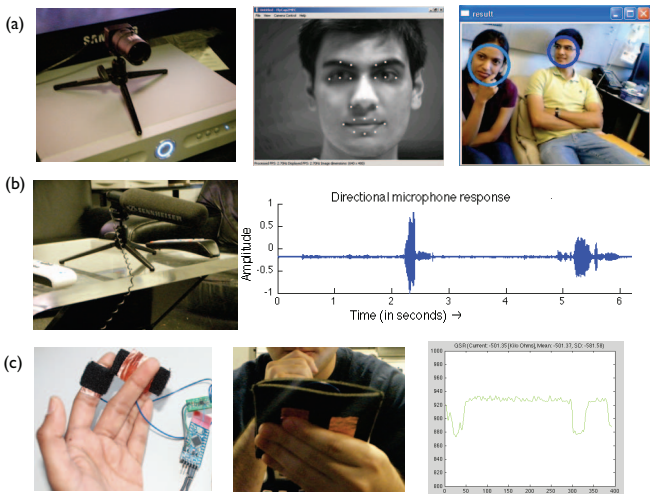

Fig. 3. Engagement sensing modules and their output

1) Visual Module: This module uses a camera to detect the number of people watching television (Figure 3). We use OpenCV (Open Computer Vision) for detecting the number of faces. Further, in order to convey attentiveness of a remote co-viewer, we use Google Tracker[2] for tracking faces. The current implementation is capable of reliably tracking one face. We use facial components detected by the tracker to determine general gaze direction - this tells us if the viewer's head position has moved left/right from the initial position used by the tracker during calibration.

2) Audio processing module: The process of picking up spoken comments from the user is fraught with extraneous audio in a typical television viewing setting. To overcome this issue we used a directional microphone [3] pointed at the user that could detect spoken comments coming from the direction of the viewers and ignore television audio (Figure 3 b). This audio is then processed to detect laughter. The laughter detection module is a Nearest-neighbor classifier trained on 10 laughter/no laughter samples each from 5 users. We use a representation based on the mel-cepstrum coefficients of the speech signal sampled at 8000 Hz. Each instance consists of 12 mel-cepstral coefficients along with the log of the energy, 0th cepstral coefficient, delta and delta-delta coefficients for each frame. Each instance is a window of 2 seconds of audio data split into 256 frames. Dynamic time warping (DTW), a distance metric for sequences based on dynamic programming, was used as the distance metric.

3) Galvanic Skin Response (GSR) sensing: The use of the GSR sensor is exploratory in nature. We were interested in looking for non-speech cues of attentiveness to augment the co-viewing experience. An important benefit of a GSR sensor is that it offers a quick way of sensing emotional arousal[11]. In our current prototype we use a sensor fitted with an Arduino Mini micro-controller and Bluetooth Mate to transmit sensor output to the local processing server. The sensor package (Figure 4) is designed to be attached to the back of the cell phone or worn on the user's hand. We calibrate the response of a user for the first ten minutes till the sensor data is stable to obtain our base-line reference. After this initial calibration period, the system detects peaks in data corresponding to sudden arousal in the subject (Figure 3c).

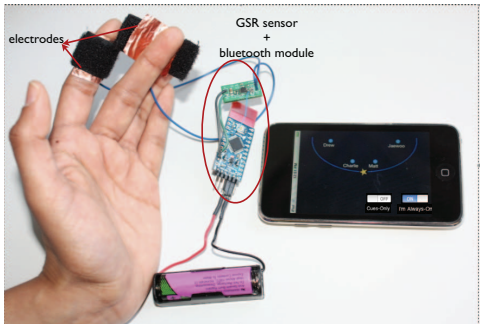

Fig. 4. Galvanic Skin Response sensor

D. Back Talk Server-Client Architecture

The Back Talk system has a central server that connects clients associated with a virtual couch (Figure 5). Each couch is assigned a URI and each viewer in the couch streams

[2]Formerly http://nevenvision.com

[3]Directional microphones are sensitive to audio from a particular direction only

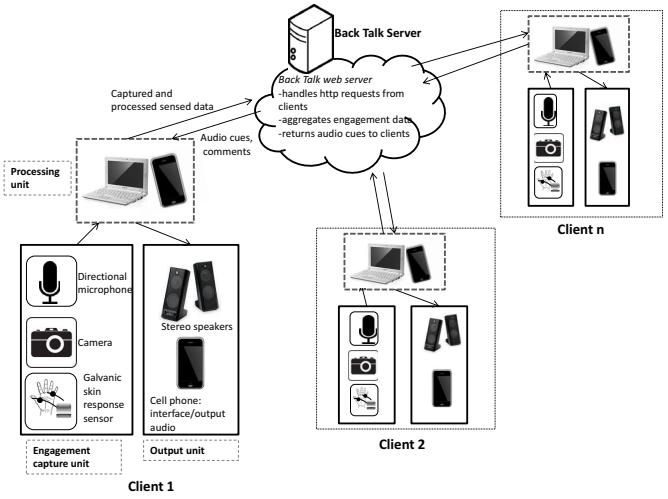

Fig. 5. Server-Client architecture

engagement data to this central server. This server model can easily be extended to have multiple such servers, each connecting clients of a particular couch. Network connections between server-clients are essentially HTTP requests over the Internet. The client in this case is a cell phone - an iPhone (it could be any other touch based hand-held phone). The client periodically queries the server to get most recent engagement cues for each co-viewer. On the cell phone we use OpenAL (Open Audio Library) to create an Open AL listener (the primary viewer represented by a star icon), OpenAL sources (co-viewers or "sonic avatars") and OpenAl buffers (audio cues for different engagement data). For every data point conveying co-viewer engagement, the cell phone plays an appropriate audio cue for the corresponding "sonic avatar".

V. Results

We evaluated the Back Talk system during the semi-final matches of the FIFA world cup 2010. The choice of sports content for the test derived from findings of [5] that indicate viewers tend to talk while watching sports events. The system connected participants in two different locations in the Media Lab, MIT. The Back Talk system is designed to support multiple distributed viewers in a common viewing experience. The model is ideal for single viewers aggregating through their virtual couch. Our user experience study deviated from this model in that it was primarily a multi-viewer test - based on our design this should have triggered audio activity for only one sonic avatar placed on the primary listeners couch. However, we attempted to create the same effect we designed for, that is, surrounding a primary listener with and auditory environment. We achieved this by mapping each engagement cue to a different sonic avatar. This resulted in cues playing to the left and right of users. In our setting, the primary listener mapped to all participants sitting in the front-center couch in each setting.

One test location was set up with all the sensing modules while the second location had a "coder" observing the participants during the entire viewing session. Engagement and activity as detected by the coder were entered into a web interface that communicated with the server. The study involved a total of 15 participants - the group comprised graduate and under graduate students between 19-35 years of age. The first study consisted of 7 participants - 3 in one location and 4 in the other. The second study - during the second semi-final match - had 8 participants with an equal number in both locations. Prior to the start of each study, each group received an introduction to the system, the controller, the audio cues and were given a few minutes to explore the controls and features.

We found that a large portion of the activity during the evaluation involved viewer comments. Conversations usually peaked around a promising moment in the match. One participant who was particularly interested in hearing comments from the other location, commented "It seems like similar conversations are happening there (the other location) and I like hearing that." We found one other participant asking the study organizer to turn up the volume of the speakers so that comments from the remote group could be heard further out in the room. Overall, participants at both locations agreed that an open audio channel was a positive attribute in the system, especially during an event like a soccer game. As one of our participants described it: "...it is an instantaneous way of connecting to the folks up there (referring to the participants in the second test location)."

A qualitative probe at the end of the evaluation revealed that participants placed this prototype as a sociable one. Overall, the system was received positively by our participants.

- Users appreciated a free-form communication channel, in our case, an audio channel.
- Listening to comments from remote viewers resulted in more engaging conversations. It led to more intra-group interactions as well.
- Participants felt that familiarity with audio cues would help map them more fluidly to the activity they indicate.
- Delay of up to 3 seconds in receiving audio streams was considered tolerable, but, when network lag added an extra 2 - 3 seconds, participants found that undesirable. (The best case performance had delays under 2 seconds)
- Participants opined that the current prototype required in-built audio normalization to match the volume of the audio cues to incoming spoken comments.
- Entry and exit were well understood by participants. However, the gaze direction algorithm updated change in gaze very frequently, which was not desirable.

VI. Discussion

The underlying premise in designing an auditory environment around a primary listener was that it could fluidly fit into the viewing experience, and selectively transit between the center and periphery of one's attention when required. Here we present extensions possible with the current system prototype. We also discuss compact versions of the prototype that can make the system easier to install and use.

- **Customized auditory environments:** The Back Talk prototype presents options for creating a customized listening experience for the primary viewer - by allowing custom picked audio cues based on genre or associated with specific co-viewers. Specific arrival cues could be selected for each friend. Similarly, idiosyncrasies of co-viewers can be mapped to characteristic audio cues and activated by specific triggers from the sensing modules.
- **Imported audio streams:** Auditory environments could be created by importing audio streams from a live event not attached to a particular viewer. We expound this idea with an example of a *sports bar environment*. Back Talk could be modified to create an enhanced sports bar experience for a viewer, even while watching sports content at home. Technically, this would require microphones distributed in the physical location of the sports bar that could stream audio (directly or suitably garbled/modified to de-identify customers) to viewers listening at home. This idea can be extended to other sociable gatherings and reality shows like *American Idol*.
- **Implementation:** Looking ahead, we envisage the system comprising two main processing components – the cell phone and the television (Figure 6). The television and cell phone work together as an engagement capture unit and communicate this data to the central server. The cell phone performs an additional function of creating the output auditory environment. We expect the cell phone to be capable of subsuming the directional microphone + audio processing functionality. The directional feature can be replaced by leveraging the dual-microphone feature of smart phones. Capturing viewer comments will require the phone to stream audio to the server or peer-to-peer. Laughter detection can also be achieved by running a light-weight classification algorithm. The GSR sensor could be a detachable component powered by the cell phone with a micro-controller processing sensed values, and, only communicating sudden arousal to the server. The television fitted with a camera could perform face detection and gaze direction tracking. Figure 7 alludes

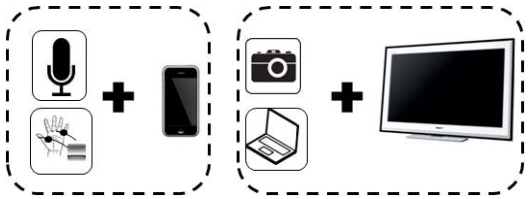

Fig. 6. **Cell phone and television units processing engagement.**

to the vision of designing the television as a local *home server* that subsumes all the engagement sensing modules - a more compact version of the design just described.
- **True surround experience:** The current implementation of the Back Talk system uses a set of stereo speakers to create the auditory environment. We are able to achieve position by left/right panning of the sources of audio and distance based volume control of the sound streams. However, a more *surround* experience could be created by using a 5.1 speaker system. This would require changes in audio capture, and multi-channel input, but, could lead to better spatial positioning of the audio source.

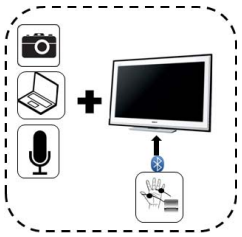

Fig. 7. **The television as the local server.**

VII. Conclusion

Back Talk, is a system that promotes near free-form communication and peripheral awareness of distributed viewers watching television at the same time. This paper investigates how various consumer electronic devices can be integrated meaningfully to enhance connectedness in a distributed micro-social network. The primary medium of achieving this is via an audio back channel created using the cell phone and engagement sensors Results from our user experience study indicate positive acceptance towards the auditory environment for facilitating presence of distributed co-viewers.

References

[1] Vision television. http://web.media.mit.edu/~stefan/vtv/.

[2] J. Abreu et al. 2BeOn-Interactive television supporting interpersonal communication. In *Multimedia 2001: procs. of the Eurographics Workshop in Manchester, UK*, 2002.

[3] B. Chalfonte et al. Expressive richness: a comparison of speech and text as media for revision. In *Procs. of the SIGCHI conf.*, page 26. ACM, 1991.

[4] T. Coppens et al. AmigoTV: towards a social TV experience. In *Procs. from the Second Euro Conf. on Interactive TV, Univ. of Brighton*, 2004.

[5] D. Geerts et al. The implications of program genres for the design of social TV systems. In *Proc. of the 1st Int'l conf. on Designing interactive user experiences for TV and video*. ACM, 2008.

[6] G. Harboe et al. Ambient social tv: drawing people into a shared experience. In *Proc. of the 26th annual SIGCHI conf. on Human factors in computing systems*. ACM, 2008.

[7] J. Hollan and S. Stornetta. Beyond being there. In *Proc. of the SIGCHI conf. on Human factors in computing systems*, 1992.

[8] R. Martin et al. neXtream: a multi-device, social approach to video content consumption. In *Procs. of the 7th IEEE conf. CCNC*, 2010.

[9] C. Metcalf et al. Examining presence and lightweight messaging in a social tv experience. *ACM Trans. Multimedia Comput. Commun. Appl.*, 4(4):1–16, 2008.

[10] E. Mynatt et al. Designing audio aura. In *Procs. of the SIGCHI conf.*, page 573. ACM Press, 1998.

[11] R. Picard et al. The galvactivator: A glove that senses and communicates skin conductivity. In *Procs. from the 9th Int'l Conf. on Human-Computer Interaction*, 2001.

[12] J. Short et al. *The social psychology of telecomm.* John Wiley & Sons, 1976.

[13] S. Zhao. Toward a taxonomy of copresence. *Presence: Teleoperators & Virtual Environments*, pages 445–455, 2003.