

**Back Talk: An Auditory Environment for Co-presence in Television  
Viewing**

by

**Andrea B. Colaço**

B.E.(Hons.) Electrical and Electronics Engineering  
Birla Institute of Technology and Science, Pilani (2007)

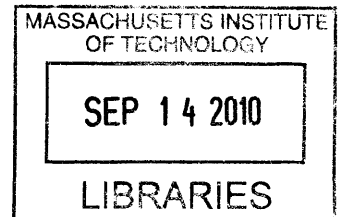
Submitted to the Program in Media Arts and Sciences,  
School of Architecture and Planning,  
in partial fulfillment of the requirements for the degree of  
Master of Science in Media Technology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

© Massachusetts Institute of Technology 2010. All rights reserved.



**ARCHIVES**

Author \_\_\_\_\_

Program in Media Arts and Sciences  
August 6, 2010

Certified by \_\_\_\_\_

Chris Schmandt  
Principal Research Scientist  
Program in Media Arts and Sciences  
Thesis Supervisor

Accepted by \_\_\_\_\_

Prof. Pattie Maes  
Associate Academic Head  
Program in Media Arts and Sciences



# **Back Talk: An Auditory Environment for Co-presence in Television Viewing**

by

Andrea B. Colaço

Submitted to the Program in Media Arts and Sciences,  
School of Architecture and Planning,  
on August 6, 2010, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Media Technology

## **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 content with a group is inherently a sociable one, and often desirable despite the physical distribution of group members. This thesis introduces Back Talk, a system designed to create a sociable television watching experience. We enhance television viewing with an auditory environment around a viewer - constructed from engagement and audio streams of co-viewers in the viewer's micro-social network. We have explored and leveraged the richness of audio to convey presence of remote viewers via a novel framework for capturing and translating engagement of an individual in the viewer's micro-social network into a set of audio cues that are played spatially around the viewer. This work presents the implementation scheme we used, and it also discusses results of a user study that was conducted to examine the impact and effectiveness of the Back Talk system.

Thesis Supervisor: Chris Schmandt

Title: Principal Research Scientist, Program in Media Arts and Sciences



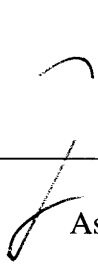
**Back Talk: An Auditory Environment for Co-presence in Television Viewing**

by

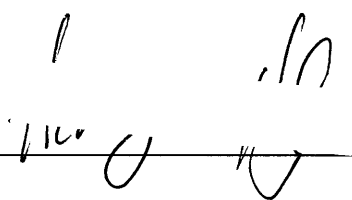
Andrea B. Colaço

The following people served as readers for this thesis:

Thesis Reader \_\_\_\_\_

  
Joseph Paradiso  
Associate Professor of Media Arts and Sciences  
Media Laboratory at MIT

Thesis Reader \_\_\_\_\_

  
Henry Holtzman  
Principal Research Scientist  
Media Laboratory at MIT



# Acknowledgments

My deepest gratitude goes to my advisor Chris Schmandt for being an amazing and supportive mentor. I would like to thank him for giving me the opportunity to go to graduate school. This work exists largely due to his guidance through many discussions, and encouragement.

I consider myself very fortunate to have had an opportunity to interact with Professors Henry Holtzman, Marie-Jose Montpetit and Joseph Paradiso. My work benefitted a great deal from their classes, particularly the SocialTV class. They have been a source of valuable insights that positively influenced my thesis.

I would also like to thank my peers in the Speech+Mobility group who made these past two years a wonderful experience and for always being willing to bounce off ideas. Special thanks to Jaewoo and Matt for sharing “healthy soup” and good times in Seoul; Drew for his encouragement and valuable feedback; Charlie for very promptly and comprehensively sharing his thesis experience with user evaluations and for reviving many a dead laptop; Ig-Jae for making plans A, B & C and for all his assistance with computer vision problems; Dori for being a great office-mate and for introducing me to sushi.

My co-workers and friends at MIT made the experience really enjoyable and memorable – Nan-wei for her assistance in the machine shop, with the GSR sensor and for her great advice; the SocialTV class; Arash Delijani for helping out during the evaluation studies and Santiago for keeping the study alive; Daniel McDuff for his invaluable assistance with face-tracking; Kirmani for making sure I had enough caffeine to keep me going, for renewing my motivation levels; the Starbucks napkin-sketchers - Inna, Keywon, Santiago; Karthik for listening to my vague ideas and presentations; Varun and Jairaj for musical and other entertainment. A special note of thanks to Priya (aka Prixa), my amazing room mate - for being tremendous support, a listening ear for late night rants, many fun activities, and for permitting random mutations of her name.

This thesis would not have been possible without the support and sacrifice of my dad, mum and brother. I’d like to thank them for their love and patience. Without you, I would have never made it this far.

Andrea Colaço  
MIT





# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
<b>2</b>	<b>Background</b>	<b>17</b>
2.1	Co-presence . . . . .	18
2.1.1	Co-presence Around Activities . . . . .	18
2.1.2	Why Co-presence Around Television? . . . . .	19
2.2	Audio Environments: Affordances and Limitations . . . . .	19
<b>3</b>	<b>Related Work</b>	<b>21</b>
3.1	Previous Work in Social Television: Same-Time, Different-Place . . . . .	22
3.2	Previous Work in Audio Environments for Conveying Presence . . . . .	26
3.3	Design Choices Derived from Previous Work . . . . .	27
<b>4</b>	<b>Design</b>	<b>29</b>
4.1	Overview . . . . .	29
4.2	System Design . . . . .	29
4.2.1	Engagement Sensing Module . . . . .	30
4.2.2	Audio Generation . . . . .	30
4.3	System Setup . . . . .	30
4.4	Cell Phone Interface . . . . .	31
4.5	Audio Environment . . . . .	33
4.6	Audio Cues . . . . .	33
4.7	System Design Diagram . . . . .	34
<b>5</b>	<b>Implementation</b>	<b>35</b>
5.1	Overview . . . . .	35
5.2	Back Talk Server . . . . .	36
5.3	Back Talk Client . . . . .	36
5.3.1	Processing Unit . . . . .	37
5.3.2	Engagement Capture Unit . . . . .	37
5.3.3	Output Unit . . . . .	43
<b>6</b>	<b>Evaluation</b>	<b>47</b>
6.1	Overview . . . . .	47
6.2	User Experience Study . . . . .	47

6.2.1	Experimental Setup . . . . .	48
6.2.2	The Experimental Method . . . . .	50
6.2.3	Results . . . . .	53
6.3	Engineering System Performance Study . . . . .	57
6.3.1	Detection of Spoken Comments . . . . .	57
6.3.2	Laughter Detection Accuracy . . . . .	57
6.3.3	Galvanic Skin Response Sensor . . . . .	58
6.3.4	Detection of Number of Faces and Gaze Direction . . . . .	59
<b>7</b>	<b>Future Work and Conclusion</b>	<b>61</b>
7.1	Contribution and Effectiveness of the Prototype . . . . .	61
7.2	Future Directions . . . . .	63
<b>A</b>	<b>Audio Processing Experiments</b>	<b>67</b>

# List of Figures

1-1	<b>Watching television together affordances: peripheral awareness and communication.</b> . . . . .	14
3-1	<b>Interface to the 2BeOn system.</b> . . . . .	23
3-2	<b>(a) An emoticon push in AmigoTV. (b) Buddy mosaic displayed on the television screen</b> . . . . .	24
3-3	<b>NeXtream’s smartphone controller and its feature of accessing one’s social network while watching video.</b> . . . . .	25
4-1	<b>Back Talk system setup with its components.</b> . . . . .	31
4-2	<b>Cell phone interface with audio circle, sonic avatars, and controls.</b> . . . . .	32
4-3	<b>Back Talk System Diagram</b> . . . . .	34
5-1	<b>Visual Module: Camera mounted near the television.</b> . . . . .	38
5-2	<b>Facial points as detected by the tracker.</b> . . . . .	40
5-3	<b>Directional microphone pointed at viewers.</b> . . . . .	41
5-4	<b>Comparison of microphone inputs in the presence of television signal.</b> . . . . .	42
5-5	<b>Galvanic skin response sensor package fitted to the back of the phone.</b> . . . . .	43
5-6	<b>Galvanic skin response sensor electrodes worn on the hand, not attached to the phone.</b> . . . . .	44
6-1	<b>User study locations: a,c - first location, b,d - second location.</b> . . . . .	48
6-2	<b>User study location with directional microphone, GSR sensor, cell phone.</b> . . . . .	50
6-3	<b>User study location with camera and stereo speakers.</b> . . . . .	51
6-4	<b>User study location setup.</b> . . . . .	52
6-5	<b>User study location with directional microphone and speakers.</b> . . . . .	53
6-6	<b>ROC curves for detection of spoken comments using directional and non-directional microphones.</b> . . . . .	58
7-1	<b>Cell phone and television units processing engagement.</b> . . . . .	65
7-2	<b>The television as the local server.</b> . . . . .	65



# Chapter 1

## Introduction

The last few years have seen a surging interest in the area of *social television* (social TV). This trend could partly be attributed to an increase in new content consumption patterns; on-demand options have gained huge popularity. Synchronicity was the glue that broadcast offered in creating shared experiences around content. With availability on one big screen in a typical living room, locality also played a major role in fostering a communal viewing experience. With broadcast on its way out, and an array of portable screens (with content easily accessible) to choose from, consumption of video content in general and television content specifically has become *individualized*. In turn, these new trends have led to innovative solutions and applications that attempt to embed social elements into whatever may be a viewer's consumption preference.

Creating social television experiences, in practice, has for the most part focused on adding communication options, rating content, recommending programs, participating in polls, and one-click shopping. Evident in most of these solutions are aspects of our current web communication practices - chat clients, tweets, thumbing-up or down content. Some of these attempts have proven useful, while others have been discarded for being too intrusive or distracting, or simply for consuming too much screen real estate. However, a key observation that repeatedly emerges is that some content and experiences are consumed better when shared with one's social circle. This leads us to realize the possible importance of fluid and richer presence of people (from our social circle) to share a viewing activity, for instance, a television show. This thesis describes research that addresses the de-

sire of providing seamless access to people with whom we choose to share our viewing experience, in a setting that mimics a group watching content together in the same room.



Figure 1-1: **Watching television together affordances: peripheral awareness and communication.**

Our approach is motivated by the interactions possible when a group of people congregate around a television - this setting affords *communication* and *peripheral awareness* of co-viewers. We have selectively moved away from traditional text-based chat clients - that typically get exported from our web interactions to social television applications. Our aim is to offer free-form interaction that characterizes unmediated person-to-person communication in the setting of television viewing. Central to this design is audio-based *co-presence* of non-located friends which includes an open audio channel for voice communication. Creating an auditory environment around the viewer leverages the rich and fluid nature of audio and an individual's capacity to selectively tune in to or tune out from audio information. The system focuses on augmenting a viewer's auditory space with peripheral awareness of distributed friends. This is achieved by passively sensing viewer activity and engagement in the form of laughter, emotional arousal, and general attentiveness - looking at

the television or not. The primary interface for accessing co-viewers is an application on a mobile phone – it precludes the need for an additional controller, and moreover, is a personal communication device.

Our system, called **Back Talk**, is an attempt at promoting sociable television watching even when participants are non-collocated. The name derives from the fact that it is primarily audio-based and captures conversations and activity in the background of television watching; the “back talk” as it happens is not the main focus of the activity but contributes to the overall experience.

### **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. Matt, is quite scared by the shocking death of the character “Boone”. His galvanic skin response sensor detects sudden emotional arousal as evidenced by his skin conductance values. His friends in the virtual couch hear a mild screech like sound in their auditory environment coming from Matt’s sonic avatar. This prompts a conversation between them. 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.*

### **Thesis Statement**

The fundamental questions this thesis attempts to answer are:

How do we create a sociable television watching experience for distributed viewers?

Does our prototype - Back Talk - convey presence effectively through an auditory environment around the listener?

## **Contribution**

The thesis has a two-fold contribution. In the realm of social television applications, it is an attempt at making television viewing sociable. As compared to previous solutions we use a combination of synthetic and spoken audio to create presence of distributed friends. The second contribution is a novel way of automatically capturing and translating engagement into an auditory environment to create peripheral awareness of remote friends.

## **Structure of the thesis**

The rest of this thesis is divided into 6 chapters: background, related work, design, implementation, evaluation and conclusion. The **background** chapter introduces various terms key to understanding our system design and some literature related to sociability. In the chapter on **related work** we review previous work in the area of social television and audio environments. The **design** chapter details design aspects of the interface, the auditory environment, overall system design, and server-client architecture. It illustrates how each engagement sensing module fits in the overall system design. We discuss the making of each engagement sensing module, the server, the client cell phone interface and finally the output auditory environment in the chapter on Implementation. The chapter on **evaluation** describes our two-pronged method: evaluation of the completion of the project and our experimental method in answering the main questions this thesis addresses. We also discuss the results of a user study conducted to evaluate the working of the system and the experience of using it. We conclude with lessons learnt from the experience of building the Back Talk prototype, and discuss the implications of our results on possible future iterations of our system in the chapter on **future work and conclusion**.



## Chapter 2

# Background

This chapter introduces the motivation of our research in the area of sociable television watching. There have been a number of solutions that have attempted to address social interaction between distributed co-viewers. Hence, it is important to take a step toward seeing the big picture of applications in social TV with the view of placing our prototype in the bigger scheme of social TV applications. We also introduce and describe terms key to understanding issues of presence of remote participants in shared activities.

The New York Times featured an article titled “Watching TV Together, Miles Apart” (Jan 3, 2010) [1]. It discusses the story of Emma McCulloch and Jennifer Cheek . . .

“(they) used to meet to watch “Dancing With the Stars” together, but that ritual ended when Ms. Cheek moved to Hawaii. ”

Now physically-separated, creating the same level of interaction and experience of watching television, could only be made possible by intervening communication technology. In their case, they set-up Skype to video chat while they watched the show. This article brings out a tacit viewer need and behavior relevant to Back Talk as a sociable television watching prototype: the need to share a television watching experience, and in a micro-social network (friends, family and close acquaintances) participants are comfortable sharing audio and even video and may even go to the extent of setting up ad-hoc solutions toward this end. Such consumer behavior has opened challenges for

researchers to create innovative solutions that bridge distances and allow new forms of shared experiences. At the same time, television manufacturers like LG, Panasonic and Samsung are equipping televisions with webcams and connectivity through Skype to facilitate video communication. In the light of such emerging solutions by consumer electronics organizations, and those put together by consumers, our prototype is positioned somewhere between these two ends.

## 2.1 Co-presence

Co-presence is the participation of a group of people in a common activity or experience, and it can be either virtual or real. It is defined in [2] as:

It (co-presence) is a condition in which instant two-way human interactions can take place. “Instant” human interaction refers to real-time or near real-time human communication, which does not include diachronic exchanges like postal correspondence; and “two-way” human interaction refers to reciprocal or feedback-based human communication, which excludes unidirectional “para-social” behaviors like watching TV or listening to radio . . .

### 2.1.1 Co-presence Around Activities

Co-presence is a vital component of group sociability [3], and traditionally group sociability was grounded in a *shared activity*. For example:

1. A family watching a television program.
2. A set of friends playing or watching a game.

In all of the above activities socializing is an inherent part and usually a major objective of that activity, but with the advent of technology the socializing component moved from being an inextricably linked component to an optional component. For example, consider the following scenarios which support the previous observation:

1. I can more easily watch a TV show the next day in my free time than collect all my friends to watch at the same time.

2. I can more easily play or watch a game online than meet with friends to play it physically.

My thesis seeks to plug back social experiences into a common activity: Television viewing. By anchoring co-presence information to an activity it is more easy to supply context information. For instance, if my friend laughs I can know that she laughed watching a particular show. This enables passive sharing of co-presence information which can enable sharing of non-textual information <sup>1</sup>.

### 2.1.2 Why Co-presence Around Television?

Traditionally, television enabled shared social experiences by virtue of being a communal resource coupled with synchronicity made possible by its broadcast nature, and around which people spend a huge amount of time [4]. In essence, watching television is a social experience, and requires that viewers participate in a *mutual effort* [5] of *understanding and decoding what they see on the screen* [6]. Our work aims at aiding this mutual effort with people that a viewer chooses to share it with, despite their physical separation. We have designed the system to fit a *micro-social network*<sup>2</sup> of people as opposed to a massive multi user experience. This design choice is believed to make sharing the viewing experience easier with audio as the primary means of communication. Since our system is primarily audio-based, familiarity with co-viewers' voices - that comes with knowing participants beforehand - contributes positively to making the experience more sociable. Moreover, viewers would be more willing to share audio and engagement data with an intimate group as opposed to sharing this information with an anonymous group.

## 2.2 Audio Environments: Affordances and Limitations

A survey of previous solutions that create sociable television watching reveals that an audio channel between distributed participants promotes communication in a natural way - discussed in detail in

---

<sup>1</sup>Facebook, Twitter and other forms of social networking services can be seen as providing varying degrees of copresence information usually unanchored to any common activity. Reiterating, while Facebook, Twitter and other social services can be seen as active forms of socialization, I propose a more passive form of sharing social networking via shared co-presence information around an activity.

<sup>2</sup>Here, we refer to one's intimate social circle - close friends, family - as a micro-social network.

chapter 3. Audio environments have also been found to allow the user to simultaneously perform other tasks while listening or speaking [7]. Further, voice is more expressive and efficient than text, as it places less cognitive demands on the speaker and permits more attention to the content of the message [8]. This obvious advantage immediately justifies choosing voice over text. We also anticipate that television audio would dominate the viewers' listening experience, and, so, any auditory environment synthesized for creating sociable interactions would primarily be a background process. It has been discussed in [9] that audio easily fades into the background, but users are alerted when it changes; we use this property to construct our ambient audio environment. Our system is also designed to support multiple viewer audio streams. The "Cocktail Party Effect" [10] provides the justification that listeners can in fact attend to multiple background processes via the auditory channel as long as sounds corresponding to each process are distinguishable. This informs our choice of audio cues; we selected easily distinguishable cues for our set of events that trigger them. Additionally, we add some more scope for distinguishing viewer audio streams using left/right panning where position to the left and right can be manipulated by the primary listener through the cell phone interface. More on selection of audio cues and the auditory environment can be found in chapter 4.

The downside of using purely an audio environment for conveying a large set of actions is that viewers are required to adjust to the library of audio cues used. Moreover, when a cue is played there is a chance that it is not heard by the user, and hence lost - our auditory environment is ephemeral and events are not recorded and played again.

## Chapter 3

### Related Work

Socializing around the television has been around almost as long as television itself [11]. In fact, the television experience was conceived of as a sociable one. Recent years have seen researchers trying to sieve out the most important elements that made television watching sociable; the two most important reasons for this were societal change and technological innovation - increasing distribution of viewers (that once watched television together) and possibilities to watch content on-demand. Consequently, this work led to solutions attempting to redesign viewing to support sociability among viewers, that took into account physical separation of viewers and on-demand viewing behavior. Our work is primarily directed at shared viewing when users are non-located. As such, we will review previous work that has promoted sociable viewing in the synchronous case when participants are distributed. There are many examples to choose from the literature in this space. This chapter however, will focus on two aspects among a subset of these solutions - creating television presence information and channels used to create social presence. More specifically, I will highlight how Back Talk is novel in capturing, sharing and the channel itself for conveying presence. Additionally, this chapter will also review past work on audio environments for conveying activity and connecting distributed groups in both social and work settings.

The first step in connecting distributed people that desire to share a viewing experience is providing a communication channel. With intervening communication technology, this is possible in a variety of forms. Some of the earliest solutions include SMS (short message service) and IM (Instant

Messaging) clients for instant text communication. Audio and video have also been experimented with in some solutions and will be compared with our prototype later in this chapter. The other important step is providing awareness of remote viewers - presence. The importance of presence and conveying a feeling of presence - co-presence - have been discussed earlier in chapter 2.

### 3.1 Previous Work in Social Television: Same-Time, Different-Place

We will survey systems that have aimed at promoting new ways of sociable television watching. This section is not restricted to research explorations only, but, discusses commercial systems that have contributed to this space. We will use pictures depicting the essence of the solution in cases where visual elements of the related work are important to understanding the how the system compares with Back Talk.

- **Reality Instant Messaging** [12] offers presence of remote viewers through their “buddy surfing” option - an awareness that friends are watching the same program; the interface includes an IM client - on the television screen - that facilitates text chats between viewers and provides some conversation promoting information related to the television content playing.
- **2BeOn** [13] also addresses interpersonal communication through IM chat, texts, voice or video. The interface to these communication modes was the television screen as seen in Figure 3-1.
- **AmigoTV** [14] was prototyped and tested by Alcatel to connect viewers real-time through a voice channel. The interface for initiating communication, and viewer avatars are located on the television screen (Figure 3-2). A related feature was expressing emotions via emoticons - these were image, video or audio based (Figure 3-2). Unlike Back Talk which senses engagement and activity passively, this application required viewers to emote manually.
- **Media Center Buddies** [15] developed by Microsoft Research again offered a text-based IM client for co-located viewers. It differs from applications described earlier in that it promotes a new idea of simultaneously allowing multiple viewers to access their online buddies through



Figure 3-1: Interface to the 2BeOn system.

the same interface. This is a modified way of having a group of viewers connect to their respective remote buddies through a common IM client. Again, this solution was primarily text based.

- **Examining presence and lightweight messaging in a social television experience** [16] was conducted by Motorola Labs. They were interested in exploring simpler ways of conveying connectedness to understand if these lightweight options could replace voice chat. Results from this study indicate that participants expressed a strong desire for a free-form communication. This study advised our design to maintain an option for voice communication between distributed viewers.
- **ConnecTV** [17] describes a large-scale field trial carried out to investigate the use of text-based chat through an IM-like interface on the television as a means to connect to remote friends. Their results showed an increase in television consumption by participants.
- **Lycos Cinema**<sup>1</sup> is a virtual cinema experience where viewers can watch a movie together and text-chat about the content.
- **Joost**<sup>2</sup> is a service that offers internet-based television, created by the developers of Skype. Further, it provides options for viewers to chat while they watch and rate content.

<sup>1</sup><http://cinema.lycos.com>

<sup>2</sup><http://labs.joost.com/tv/>

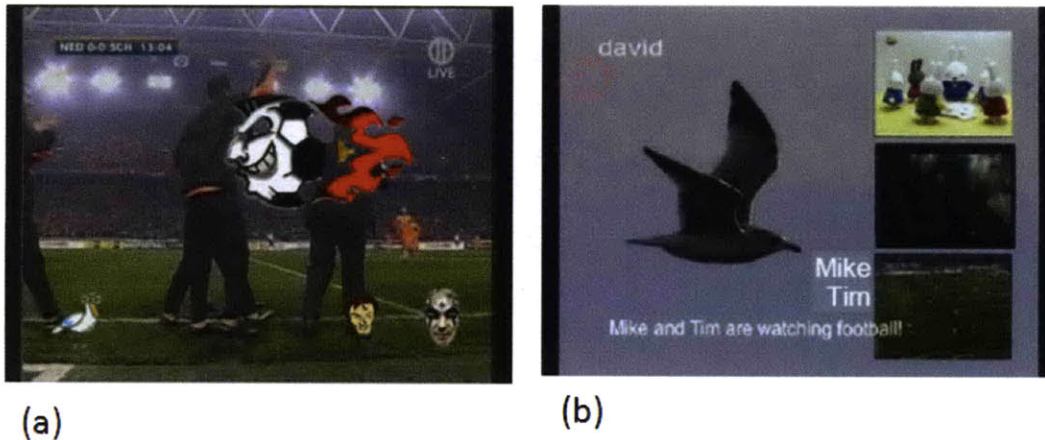


Figure 3-2: (a) An emoticon push in AmigoTV. (b) Buddy mosaic displayed on the television screen

- **The Virtual TV Couch** described in [18] had a similar goal in trying to connect micro-social networks. They use an audio channel to connect distributed people but add *impulsive interactions* like “quick bets” and voting on content; the interface for these actions is presented on the television screen. In contrast, our work propose a similar layer of interactivity, instead we choose to automatically sense engagement and convey it to remote participants.

There have also been applications in this realm developed at the Media Laboratory. Here, we list some of the most relevant ones.

- **neXtream** [41] is a recent example of work in social television in the Information Ecology group. Similar to Back Talk, this system also focused on the use of a smartphone as the controller, and for accessing one’s social network, albeit through a chat feature (Figure 3-3). It also provided a social layer through a collaborative filtering model of content selection.
- **VisionTelevision** [19], developed in the Object-based Media Group, is capable of detecting faces of viewers and transmits them to remote locations. The approach focused on visual presence – images of participants placed at the bottom of the screen. Our solution also detects the number of people in the room, but, we use this information in a different way and for a



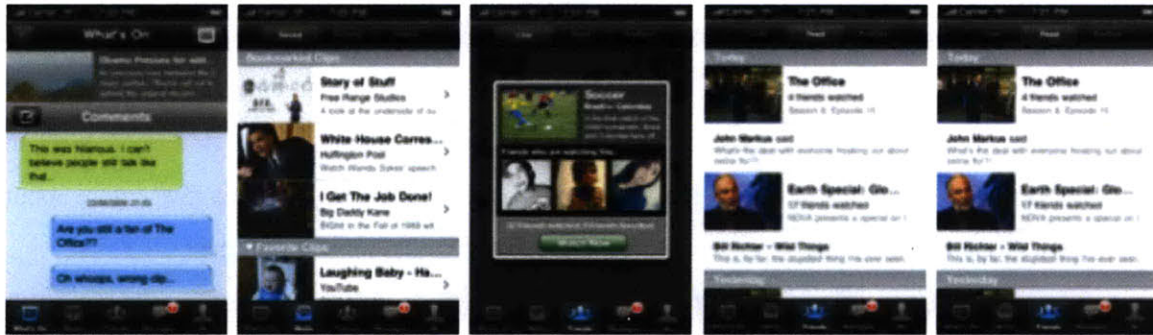


Figure 3-3: NeXtream's smartphone controller and its feature of accessing one's social network while watching video.

different purpose: number of people in the room is conveyed as an audio cue to convey a sense of people coming and leaving, and this information serves as a background indicator of entry/exit patterns that may lead up to conversations between distributed co-viewers.

- **iCom** [42] also from the Object-Based Media group was a media installation that connected different locations. In this solution awareness of remote activity was primarily through a video channel. The focus of this project was to have an always-on presence mechanism between remote locations.
- **Reflexion** [20] is cited in this section to list solutions that convey presence of participants in ways other than text or audio. It is an example of a system connecting remote participants primarily through video. Video streams from each participant are composed into a homogeneous display on the bottom of the screen.
- **Television Meets Facebook** [21] was another related project carried out in the Information Ecology Group. The main portal for socializing around television content is the social network site, Facebook. This work uses elements used in online sharing to enhance the social experience – posting data (recommendations, upcoming viewing schedule, viewed content) to one's profile page for friends to see. sharing content, recommendations and viewing schedule to the site. It presents an interesting technique for making television viewing sociable, but, unlike our system is not restricted to same-time viewing.

- **Telemurals** [43] was a project from the now erstwhile Sociable Media Group. This system consisted of two portals that connected spaces through an audio and a video link. The goal was to facilitate instantaneous social interaction between groups. Our system focused on linking remote groups of people that choose to connect, whereas this system was more about providing a medium and catalysts for serendipitous interactions that could be sustained.

## 3.2 Previous Work in Audio Environments for Conveying Presence

In this section we review examples of audio based solutions aimed at conveying information and presence of distributed people. Among early examples we have **Thunderwire** [22] that created an audio-only media space for workgroups. It connected distributed members of a small group; every member could simultaneously connect to and listen to streams from the rest of the group. The authors' two month long field study showed that the system afforded sociable interactions and a telepresent environment for its users. These findings influenced our choice of an audio-only presence environment (as opposed to using a mix of audio and video).

Our second example **Designing Audio Aura** [23] explored a range of audio cues in the context of a work environment. This research offered us insights into appropriateness and user acclimatization to audio cues. Our setting despite being more recreational benefitted from the knowledge of viewer behavior and perception of audio cues mapping to specific activities.

**Nomadic Radio** [24] was a project developed the Media Laboratory. It is a wearable computing platform for managing communication and information services in a mobile environment primarily through an auditory interface. We examined findings from this research because of its feature for spatial presentation of digital audio. Their results reveal that auditory notifications are useful when the user is engaged in some other requiring a "hands and eyes-free" approach.

### 3.3 Design Choices Derived from Previous Work

We arrived at the design of the Back Talk system after carefully going over directly and indirectly related solutions in connecting distributed groups of people watching television. The examples listed above and heuristics listed in [25] played an important role in helping us understand options that were more likely to be well-accepted by our target audience. This was mostly through studies and user experiences described in these works. Our work distinguishes itself from previous solutions as a result of the following design choices:

- no visual indicators of presence are displayed on the television screen, in order to keep the viewing experience minimally distracting and screen space free of additional artifacts;
- text chat and IM clients are not part of the design – to avoid the look-and-feel of online environments and text input to the system that can potentially distract a viewer, co-viewing buddies are presented through the cell phone interface;
- engagement and activity are captured automatically by the system; consequently it does not require viewers to manually send emoticons or actively emote;
- an auditory environment around the viewer – with synthetic audio cues and viewer audio – is key in enhancing a viewer’s social experience of watching television;
- an auxiliary personal device - a cell phone - is the primary controller of the auditory environment around the viewer.

An important distinguishing element that exists in the design of BackTalk is the feature for automatically sensing a viewer’s engagement and activity without any requirement for active input from the user. The system scans a viewer’s audio track for laughter, measures overall activation through galvanic skin response, detects people coming and going, and gaze direction to capture general attentiveness. This sensed information triggers audio cues that are played for remote viewers.



# Chapter 4

## Design

### 4.1 Overview

This section describes interface design elements, setup of different system components and the overall system design. The cell phone acts as the primary interface to the Back Talk system. We introduce the *couch* metaphor used in designing the cell phone interface. Presence of remote friends is primarily conveyed through an audio environment around the listener; we attempt to recreate the *living room ambience* by spatially distributing audio cues. These elements are key to understanding the system as a whole and will be referred to in sections to follow. We also describe the physical set up of various system components: stereo speakers, directional microphone, camera, galvanic skin response sensor.

### 4.2 System Design

We have designed Back Talk to create a sociable television watching experience. We achieve this using a combination of sensing modules to capture engagement of remote viewers and translate this data into audio cues. This involves two main modules: an engagement sensing module and an audio generation module.

### 4.2.1 Engagement Sensing Module

This module comprises three sub-modules that contribute to the overall engagement capturing process.

- *Visual*: for detection of number of faces and gaze direction.
- *Audio processing*: for detection of laughter and spoken comments.
- *Galvanic skin response sensing*: for detection of sudden arousal.

We have selected this combination of sensing techniques as they indicate presence, general attentiveness (looking at the screen vs. not looking) and convey some understanding of a viewer's reaction to content based on comments. We do not consider this set of sensed values exhaustive enough to give a complete picture of a viewer's engagement; however, they provide sufficient clues for ambient awareness of remote co-viewers to serve as indicators of presence and even conversation starters [11]. Details of how these sub-modules measure data are discussed in the next chapter.

### 4.2.2 Audio Generation

The sound generation process requires two different sources of input: spoken communication and what is sensed beyond. The engagement sensing module is the source of sensed data that triggers the audio module to play iconic audio cues. Spoken communication is transmitted directly to co-viewers. All incoming audio is played through the viewer's cell phone through a set of stereo speakers.

## 4.3 System Setup

The Figure 4-1 sketches out the Back Talk system setup. A typical setting includes a camera mounted on the television for the visual module. A set of stereo speakers is arranged on either side of the viewer's couch to play audio cues around a viewer. Spoken communication and laughter

are picked up by a directional microphone pointed in the viewer's direction. This microphone is connected to a computing device for processing input audio. The directional microphone can also be replaced by a headset with a microphone to capture spoken comments.

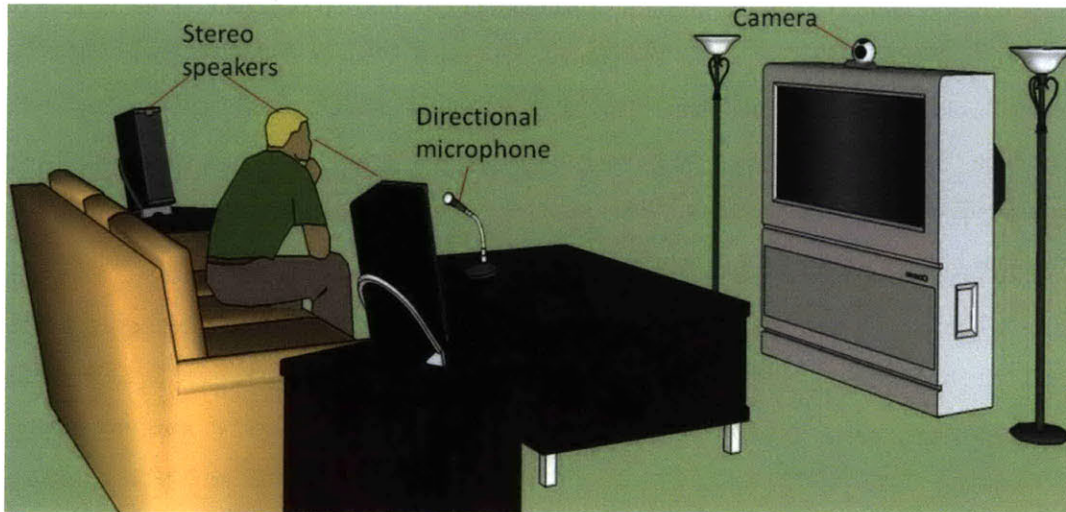


Figure 4-1: **Back Talk system setup with its components.**

#### 4.4 Cell Phone Interface

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 4-2 shows the interface depicting a virtual couch. Each remote co-viewer is represented as a “sonic avatar”. We refer to these representations as sonic avatars because they represent sources of audio in a user’s physical surroundings. These sonic avatars are movable icons that are associated with the physical listening space around the “primary listener”. The primary listener is represented as a star icon in the interface. When a user wishes to quiet an audio stream from a co-viewer, she can do so by dragging the sonic avatar outside the “audio circle”. This results in muting the manipulated avatar (Figure 4-2). 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 4-2). 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. The current prototype also has an option to select “genre” - a selection of comedy, drama, and sports. This action activates a different set of cues depending on the selection. For instance, selecting the *sports* genre activates a *vuvuzela* sound for indicating sudden arousal measured through the galvanic skin response sensor. Likewise, selecting *drama* results in a two-tone beep indicating entry of people.



Figure 4-2: Cell phone interface with audio circle, sonic avatars, and controls.



## 4.5 Audio 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 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 stereo speakers on either side of the viewer's couch Figure 4-1.

## 4.6 Audio Cues

Users of our system are physically distributed. Our process of translating engagement data into an auditory environment requires audio cues representative of the sensed data and also suitably indicative of presence of a co-viewer. 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.

The choice of using these elements derives from our motivation to mimic a real-life viewing experience – we are aware of people watching with us, likewise, when they exit, how they react to the show and so on. At the level of the listener, each sensed element is played as an audio cue. Table 4.1 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.

Table 4.1: Captured engagement values and corresponding audio cues

<b>Trigger</b>	<b>Audio Cue</b>
Entry	Footsteps
Exit	Door closing
Laughter	Canned laughter
Emotional Arousal	Mild to moderate rustling
Gaze Direction (left/right)	Prolonged yawn, boing sound

## 4.7 System Design Diagram

The diagram below (Figure 4-3) illustrates how the different modules and components – hardware and software – are connected. We will refer to this diagram when describing our system implementation.

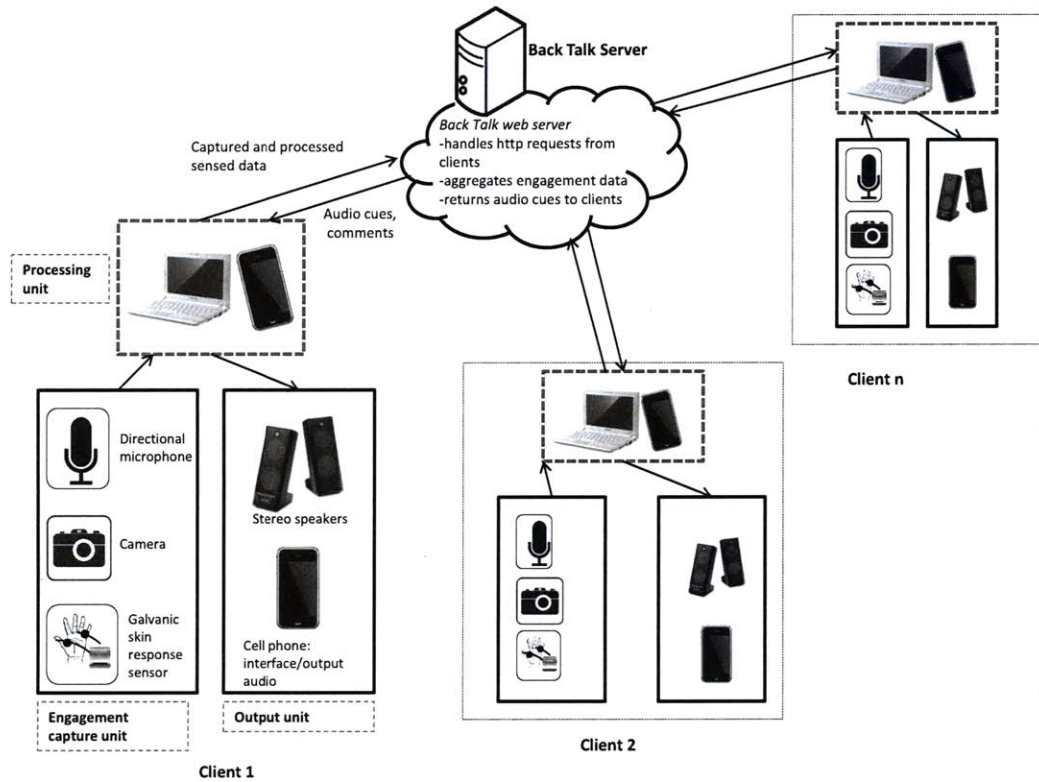


Figure 4-3: Back Talk System Diagram

## Chapter 5

# Implementation

### 5.1 Overview

The Back Talk system prototype is one implementation among many applications related to sociable television watching. This prototype combines a suite of engagement sensing techniques to capture useful information about a viewer and convey it to remote friends. The experience is grounded in a common show being watched by the audience. A key design consideration in building this prototype was avoiding additional visual notifications on the television screen of the viewer. As a result we chose to move presence information to an auxiliary device (in our case, a cell phone) and use the space around a viewer for presence information conveyed fluidly as audio.

A group of non-located users are connected to the system through their cell phones. We built the Back Talk system prototype to address the question: how can television watching be made sociable with presence of non-located friends. The underlying intent was to explore new ways of conveying presence by sensing engagement.

This chapter describes the two basic parts to the implementation of Back Talk: server-client architecture and the engagement sensing modules.

## 5.2 Back Talk Server

The Back Talk system has a central server that connects cell phone clients associated with a virtual couch (Figure 4-2). These clients are distributed friends that decide to watch a particular television show together. Each couch is assigned a URI and each viewer in the couch streams 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 main role of this server is to aggregate engagement data from each remote viewer and serve this information to querying clients.

The server is a PHP-based web server. It accesses a MySQL database to add and retrieve engagement data for each user. Clients update time-stamped engagement data through HTTP POST requests. In a similar fashion, clients also query the server for engagement data of co-viewers.

## 5.3 Back Talk Client

In our prototype we refer to the combination of the suite of *engagement sensing modules* and the *cell phone interface* as the **Back Talk Client**. Figure 4-3 shows a detailed depiction of the various components that comprise a typical Back Talk client. The client interface in this case is located on the cell phone - an iPhone (it could be any other touch based hand-held device). We further breakdown the client into a *processing unit* and an *input-output unit*. In our implementation, a local computing device (a laptop) and cell-phone comprise the processing unit. The local computing device processes data received from the engagement sensing modules before sending it to the Back Talk server. The cellphone queries the server for co-viewer engagement data and plays a crucial role in translating this data into visual and auditory output. The input-output unit includes the engagement sensing modules and a combination of a stereo speaker system and cell phone. The rest of this section describes these units in detail.

### 5.3.1 Processing Unit

In the current version of our prototype, we use a laptop as the computing device that handles incoming engagement data from the various modules. The cell phone is also part of this unit and its role is primarily to receive incoming engagement data from the Back Talk server. Together these two devices act as intermediaries between the input-output client elements and the Back Talk server (refer to *textitprocessing unit* in Figure 4-3). In future iterations of the Back Talk system, the roles of these two devices can be shared between a television and cell phone. Also, the engagement capture and output units could merge into the processing unit; this can be appropriately achieved by harnessing the cell phone to perform audio and galvanic skin response sensing and using the television as the processing device for the visual module.

### 5.3.2 Engagement Capture Unit

Our prototype has three sensing modules that pick up engagement data (refer to *engagement capture unit* in Figure 4-3). 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.

#### Visual Module

*Detection of number of people in the room and gaze direction of viewer(s).*

**Setup:** This module uses a camera placed near the television, (refer Figure 5-1) to detect the number of people watching television. Our prototype uses a Firefly MV USB Camera <sup>1</sup> and a Tamron 1/3", 5.0-50mm lens. We used this combination of lens and sensor after preliminary experiments revealed that a regular webcam was insufficient for our purpose of detection and tracking of faces of people at distances  $\sim$  8ft away from the camera. A regular webcam (in our case, a Logitech Quickcam Pro 5000) does not have enough resolution and optical zoom to create a feature rich image of the

---

<sup>1</sup>FMVU-13S2C-CS: Color Firefly USB 2.0 Camera, 1/3-Inch CMOS, CS-Mount 13CS

subject(s) to be tracked. Instead an optical zoom lens allows us to obtain a high resolution image of the region of interest.

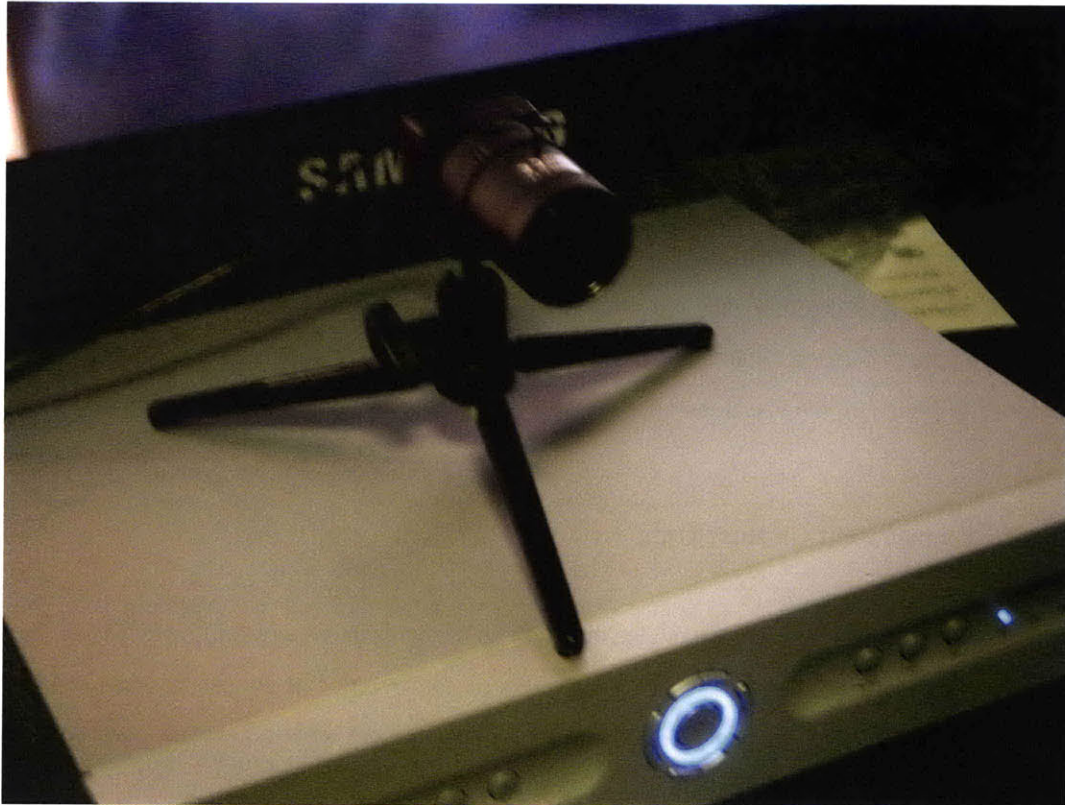


Figure 5-1: **Visual Module: Camera mounted near the television.**

**Processing:** We used OpenCV (Open Computer Vision) [26] for detection of multiple faces as seen by the camera. Face detection is based on the object detection algorithm proposed by Viola and Jones [27] and further improved by Leinhart [28]. Detection is based on Haar-like features that encode the existence of oriented contrasts between different regions of the image. They are called Haar-like features because they are computed similar to the coefficients in Haar wavelet transforms<sup>2</sup>. The classifier in this detection method is first trained with several (a few hundred) sample views of the object to be detected (in our case, faces), called positives and similarly negatives - arbitrary images of the same size as the positives. The classifier is in fact a “cascade” of “boosted” classifiers

---

<sup>2</sup>[http://en.wikipedia.org/wiki/Haar\\_wavelet](http://en.wikipedia.org/wiki/Haar_wavelet)

working with Haar-like features. The word cascade in the name means that the resultant classifier is a combination of several simpler classifiers. A positive or negative result is arrived at only once the region of interest in a candidate image has cascaded through each of the classifiers and has either passed all the stages or has been rejected. Boosted classifier means that classifiers at each stage are complex themselves and using an appropriate boosting technique, in this case Adaboost [29].

Once trained, the classifier can be applied to a region of interest in an input image (same size as training images) to detect if the region is likely to show the object (i.e. a face). An important advantage of this classifier is that it can be easily resized to find objects of interest at different sizes, which proves more efficient than resizing the image itself. OpenCV comes with several cascade files for detecting both frontal and profile faces. We use a cascade file for frontal faces for detection of number of faces in the room. This module is designed to update the Back Talk server with the most recent number of faces detected.

In order to convey attentiveness of a remote co-viewer, we use Google Tracker<sup>3</sup> for tracking faces. The current implementation is capable of reliably tracking one face. The tracker offers 22 feature points for tracking (Figure 5-2). Normally, feature points on the nose are chosen as central points. Likewise, we identify a feature point on the nose as the fiducial point. Points to the left of the fiducial point are referred to as left feature points, similarly points to its right are referred to as “right feature points”. We calculate the following two distances: sum of distances between “left” feature points and the fiducial point and sum of distances between “right” feature points and the fiducial point. When a viewer’s gaze is directed at the television, these two distances are approximately equal. However, when a viewer tilts his face (changes gaze direction) we observe a change in these above two distances calculated. If the ratio of the left sum of distances to the right sum of distances is less (or more) than the predefined threshold,  $L_{lower}$  (or  $L_{upper}$ ), then we say the person is looking left (or right). If the ratio is between these two thresholds, we say the person is looking at the television screen. In our experiments, we’ve set  $L_{lower}$  to 0.85 and  $L_{upper}$  to 1.15. As soon as a change in direction is detected, the software communicates this information to the Back Talk server.

---

<sup>3</sup>Formerly Neven Vision, <http://www.nevenvision.com/>

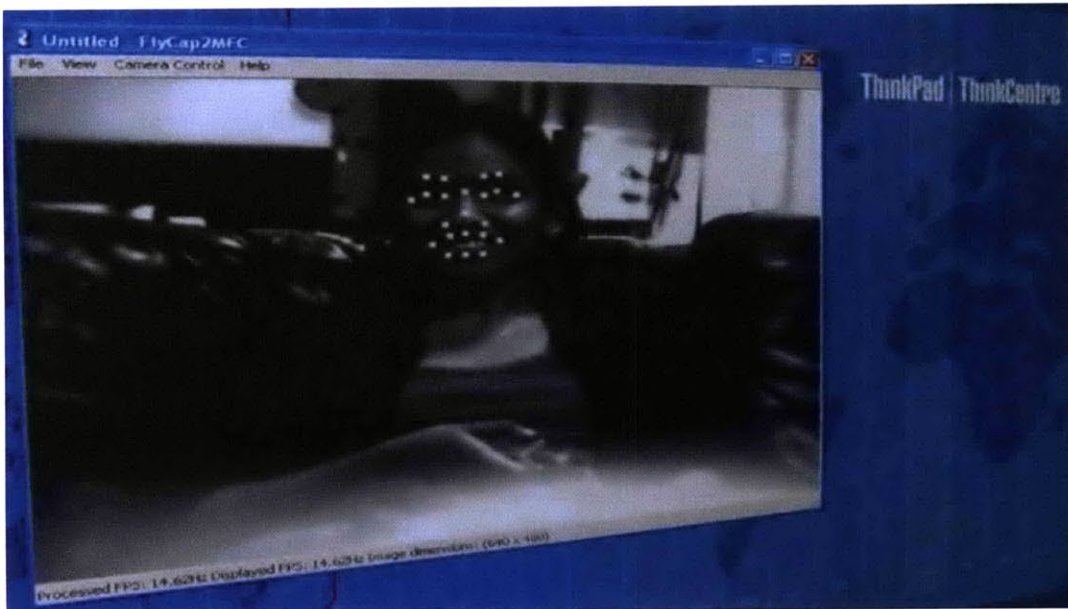


Figure 5-2: Facial points as detected by the tracker.

### Audio Processing Module

*Detection of spoken comments and laughter:*

**Setup:** i) Directional microphone pointed at viewers (Figure 5-3).

ii) Non-directional microphone<sup>4</sup>

**Processing:** Spoken comments detection: The process of picking up spoken comments from the user is fraught with extraneous sources of audio in a typical television viewing setting. While we could choose to transmit everything picked up by the microphone in the viewer's room, we decided that it would deteriorate the auditory experience. This is because the microphone would also be picking up extraneous audio signals contributed to in a large part by television audio. We addressed this issue by using a directional microphone, Sennheiser MKE 300<sup>5</sup> pointed at the user that could detect spoken comments coming from the direction of the viewers and ignore television audio. Figure 5-4 shows the input recordings from a directional microphone and non-directional

<sup>4</sup>Plantronics PC microphone.

<sup>5</sup>Directional microphones are sensitive to audio from a particular direction only.



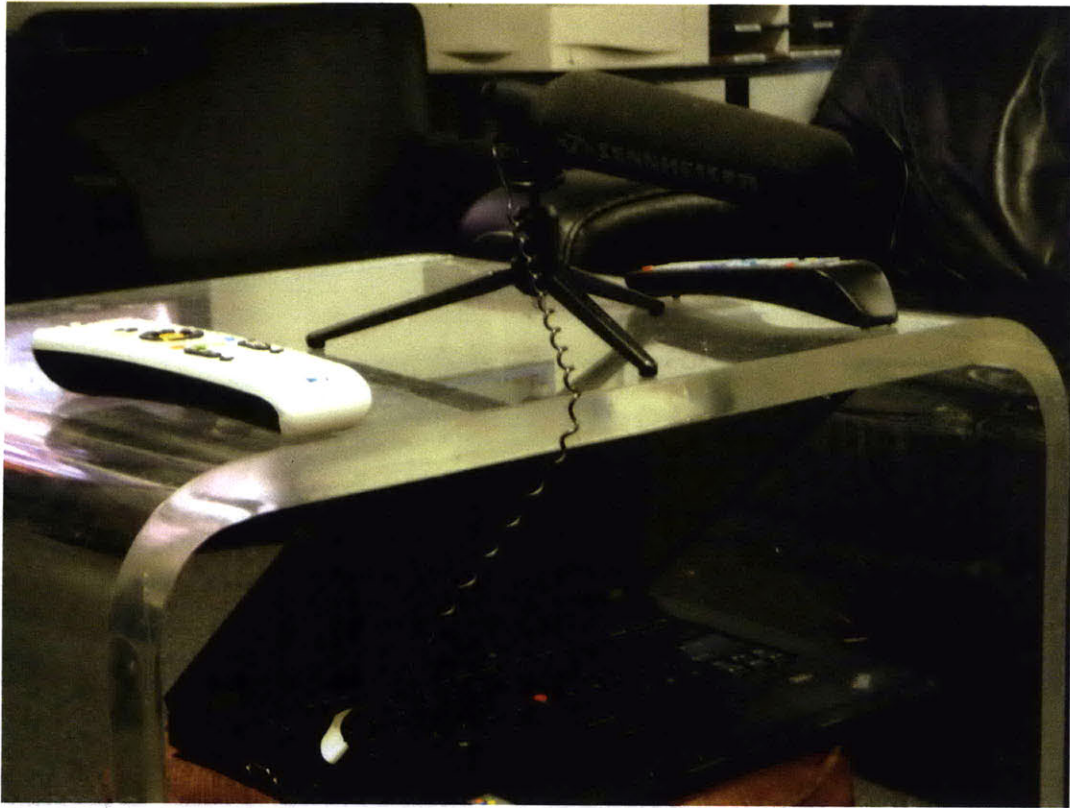


Figure 5-3: **Directional microphone pointed at viewers.**

microphone. Both these recordings are in the presence of television audio playing at a volume comfortable for viewers<sup>6</sup>. Input captured by the directional microphone by virtue of the properties of the device, tends to have portions of speech significantly distinguishable from background noise. It is this difference that we leverage in computing speech segments from the microphone input.

As a first step in the process, we measure the energy of samples in every window of one second length. The first five seconds of the start of the viewing experience is usually assigned as a calibration period. During this time, the energy of the microphone input samples is averaged to obtain the average level of ambient noise in the room with the television playing. Energy calculation is done based on amplitude of samples – sampling rate 8000 Hz, 16 bits per sample, PCM encoding. The process is designed to detect the onset of spoken comments by the viewer as seen by a

---

<sup>6</sup>Each recording had 3 speech segments with the last one spoken at a lower volume.

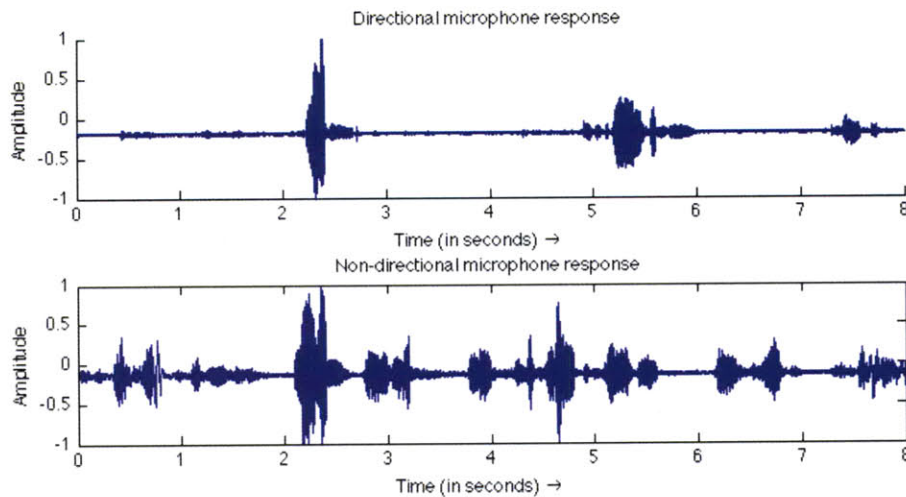


Figure 5-4: **Comparison of microphone inputs in the presence of television signal.**

sudden change (significant increase) in energy in a given window. The algorithm keeps flagging windows as “speech segments” until the energy calculated is below the set threshold. Finally, the Back Talk server makes available the buffered speech segments for all clients that register at the start of the viewing session. Prior to arriving at this process for capture of viewer comments, we experimented and explored a number of different techniques. However, these were not ideal for real-time processing. Refer to appendix A for details of related experiments.

Laughter Detection: Whenever speech segments are detected, they are 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<sup>7</sup> along with the log of the energy, 0th cepstral coefficient, delta and delta-delta coefficients for each frame. Each instance is a window of 1 second 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. We chose to use mel-cepstral coefficients as features based on our survey of previous work in laughter detection. Knox’s experiments [30] clearly indicate that MFCC features outperform pitch related features. Similar results were also seen in [31].

<sup>7</sup>[http://en.wikipedia.org/wiki/Mel\\_Frequency\\_Cepstral\\_Coefficients](http://en.wikipedia.org/wiki/Mel_Frequency_Cepstral_Coefficients)

## Galvanic Skin Response (GSR) Sensing

*For measuring overall activation.*

**Setup:** The sensor module is designed to be fitted easily on the back surface of the cell phone or worn detached from the cell phone.

**Processing:** 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 [32]. In our current prototype we use a small sensor fitted with an Arduino Mini micro-controller and Bluetooth Mate to transmit sensor output to the processing unit. The sensor package is designed to be attached to the back of the cell phone, Figure 5-5, or worn on the user's hand Figure 5-6. 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 5-5: Galvanic skin response sensor package fitted to the back of the phone.

### 5.3.3 Output Unit

A set of stereo speakers and a cell phone comprise the output unit (Figure 4-3). Together, they create the output auditory environment around the primary viewer. The Back Talk system has a

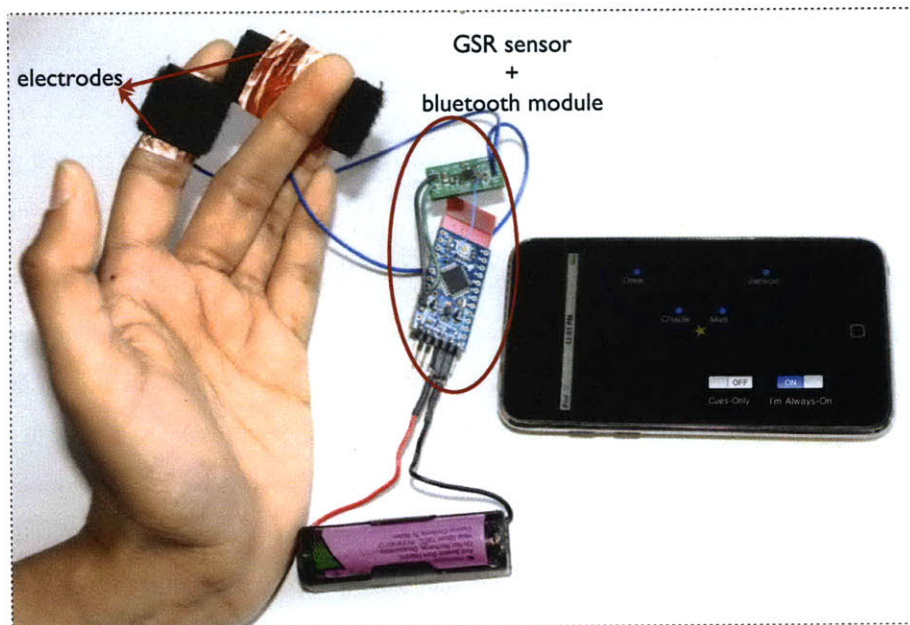


Figure 5-6: Galvanic skin response sensor electrodes worn on the hand, not attached to the phone.

central server that connects cell phone clients associated with a virtual couch (Figure 4-2). Network connections between server-clients are essentially HTTP requests over the Internet. The client in this case is an iPhone (it could be any other touch based hand-held phone). The client periodically queries the server to get most recent engagement cues and spoken comments 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 client plays an appropriate audio cue for the corresponding “sonic avatar”. Audio cues are pre-loaded on the client. The set of stereo speakers are connected to the cell phone to play the audio cues. We use OpenAL to position sonic avatars on either side of the primary listener and achieve left/right panning of the sound source. The positions of these avatars can be manipulated by moving them around on the screen.

In order to play spoken comments, the iPhone streaming client receives audio data in MP3 format from the server and funnels this data to the Audio File Stream Services [33], part of the iOS Audio Toolbox framework [34]. The Audio File Stream Services component is used to parse the data as we receive it continuously from the server. This data is then sent to Audio Queue Services [35], another component in the Audio Toolbox framework, which can handle low-level playing and recording of audio data. The Audio Queue comprises of multiple buffers which are filled at one end by parsed data from the Audio File Stream Services and played at the other end. As the buffers in the Audio Queue are filled the system plays the data in the buffers. The streaming process runs in a separate thread to keep the GUI responsive at all times.



## Chapter 6

# Evaluation

### 6.1 Overview

In order to evaluate the success of this thesis, we conducted two different kinds of user studies. The first was a **User Experience Study** and the second was an **Engineering System Performance Study**. As their names suggest the former was conducted to get a holistic idea if the experience we designed for meets the expectation of users. The second study had a focus on evaluating performance of all engineering components that were built as part of this thesis. The aim was to highlight areas that contributed positively and robustly to the overall system and areas that did not perform as expected. We will also discuss directions for improvements in future iterations based on our system evaluation.

### 6.2 User Experience Study

This study was conducted during the semi-final matches of the FIFA World Cup 2010. These sports events rendered themselves ideal for a television watching session that could potentially have a lot of conversation between groups. We picked two locations in the Media Lab where the soccer matches were being screened. It was observed that these two locations attracted quite a few viewers

and hence chosen. Figure 6-1 show the set up in these two locations - MIT Media Lab, E-15 344 and E-14 5th floor cafe <sup>1</sup>.

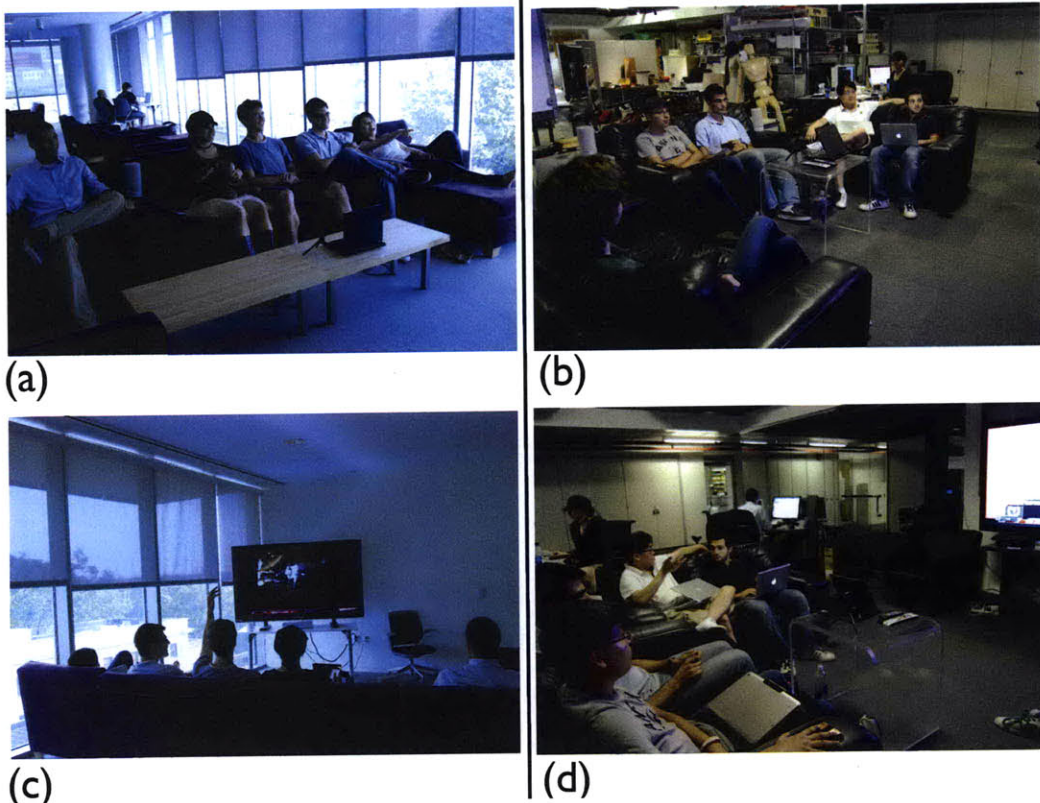


Figure 6-1: User study locations: a,c - first location, b,d - second location.

### 6.2.1 Experimental Setup

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 listener's couch. However, we attempted to create the same effect we designed for, that is,

<sup>1</sup>These pictures are representative of the audience setting during the days of the study. We did not take pictures of the audience in these locations during the study as we did not obtain prior permission for the same.



surrounding a primary listener with an 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.

The first location (E15 - 344) was set up with all three modules: visual module, audio module and galvanic skin response sensing. A set of speakers was placed on either side of the main couch in this area. Figure 6-2 highlights the microphone, galvanic skin response sensor and cell phone interface held by the viewer. Figure 6-3 shows the camera monitoring people coming and leaving, and one speaker from the stereo set placed at the side of the couch. The microphone (non-directional) was placed on the table in front of the viewers on the main couch. However, this microphone picked up audio even of participants sitting on couches flanking the main couch. The camera was positioned close to the television in a non-intrusive fashion such that activity on the main couch was visible to it. The viewers were also provided with an iPod touch that provided access to the Back Talk application interface. This device was also connected to the speakers to play audio cues and comments from the other location. The galvanic skin response (GSR) sensor module was placed on arm rest of the main couch instead of strapping it to the rear of the portable device (in this case, an iPod touch). This setting was adopted so that participants could share the iPod touch during the television viewing experience without having to worry about the GSR electrodes. It ensured that skin response was collected uninterrupted from a viewer and also that the viewer was positioned comfortably while making contact with the GSR electrodes.

The second location (E14, 5th floor cafe) had the audio module setup to pick up audio and laughter from the viewers, Figure 6-4. However, we did not setup the two other modules in this location. Instead we had a “coder” (an organizer of the study) observing the participants during the entire viewing session. Engagement and activity as detected by the coder were entered into a web interface that served as a portal for capturing engagement as observed by the coder. As seen in Figure 6-5, we had a directional microphone placed on the coffee table just in front of the viewers in this location. It was observed on days prior to the experiments that the audience in this location preferred the television volume turned up higher than the audience in the other test location. Consequently, we chose to place a directional microphone in this location, so as to capture audio free of television

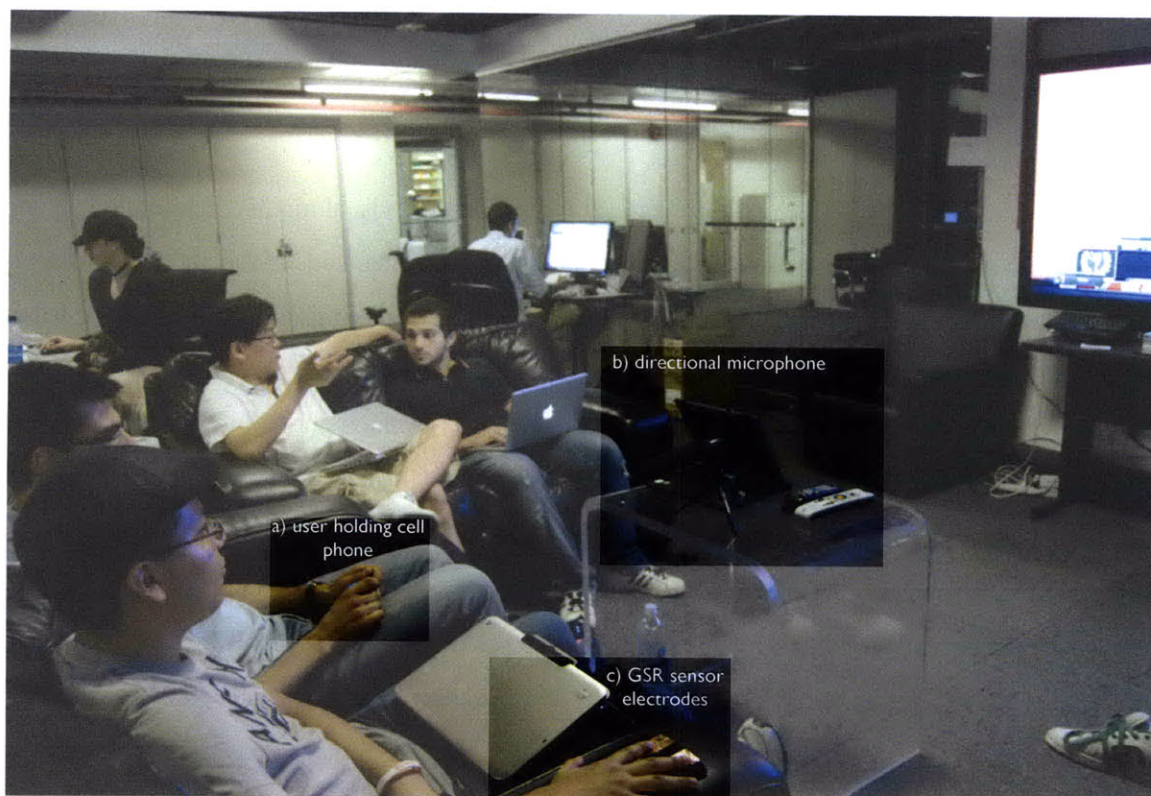


Figure 6-2: User study location with directional microphone, GSR sensor, cell phone.

signal. Similar to the setting in the other location, we placed a set of stereo speakers on either side of the main couch in this area; the viewers were also provided with an iPhone that ran the Back Talk application. The device was connected to the set of speakers to create the auditory environment around the viewers.

### 6.2.2 The Experimental Method

Prior to the start of each experiment, participants were introduced to the Back Talk system and informed about the design of the study. After a quick introductory demo to the interface and its components we encouraged participants to use the touch device interface. The introduction included an explanation of the “virtual couch” on the handheld device, the audio circle outside of which



**Figure 6-3: User study location with camera and stereo speakers.**

a remote viewer gets muted, the choice of audio cues based on genres (default, comedy, sports, drama), the modes: always-on and cues-only. Participants were shown the different modules and informed that no audio or video was being recorded, but, only transmitted to the other test location. As an introduction to the audio environment, we played a sample of each audio cue and explained its correspondence to engagement and activity of the remote viewer. We also explained the choice of audio cues and the translation step: captured engagement or activity to audio cue. The idea that Back Talk places an audio stream spatially around the primary listener was also expounded.

Viewers were encouraged to interact with the system and the interface before the game started. Volume of the television was adjusted by the participants to suit their listening. The system started shortly after the beginning of the semi-final match between Uruguay and Netherlands. The first ten minutes were spent in calibrating the audio module to threshold off ambient television noise and



Figure 6-4: User study location setup.

capture only viewer comments. This time was also used in calibrating the galvanic skin response sensing module. Participants were asked to watch the game just the way they would have in a non-experimental setting.

Each study location had one organizer observing participants and recording their reactions to the system and use of it. The representative was also around to make sure components functioned as required and to answer any questions participants had about the Back Talk system and using its components. In the location not equipped with the engagement sensing modules, the representative also served the role of a “coder”: observing viewer reactions, entry and exit patterns, and general attentiveness. This information was then fed into a web interface as *laughter*, *entry*, *exit*, *looking away*, *overall arousal* in lieu of the modules that automatically sense these values. Appropriate audio cues corresponding to these values were received at the remote location.



**Figure 6-5: User study location with directional microphone and speakers.**

The study was carried out for the duration for the first semi-final and the first-half of the second semi-final. At the end of the study, participants were asked questions about the overall experience. During the course of the study an investigator in each location also noted the performance of the system components - unanticipated engineering or design problems. We also took care to inform participants that they could stop the study at any point and manipulate avatars of remote viewers to mute/unmute their audio streams as desired.

### **6.2.3 Results**

The study involved a total of 15 participants. 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. We will now discuss some of the salient observations recorded during the evaluation - these are from the perspectives of the participants and the investigators that conducted the study.

### **Viewer comments**

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 with each group narrating their version and assessment of the system. 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. During some portions viewers would strain out to the microphone to speak their comments directly into it. This however proved counter-productive as the audio was unpleasant and distorted at the receiving end. 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 5th floor test location).”

**Sociable watching:** We conducted a quick survey at the end of the viewing session to gauge participants’ reactions to the system, and to evaluate if the system had created a more sociable viewing experience. Here are results we compiled from this phase.

- When asked if participants would have preferred a text channel in addition to or without an audio channel, the study participants unanimously agreed that the voice channel made it easier to communicate. A few said that the text channel would have distracted them from the game.
- Participants found that the audio channel sparked conversations among them, particularly in reaction to interesting comments from the remote group.
- Quality of audio channel: received audio sometimes got choppy. Participants pointed this out and expressed that the experience could benefit from higher quality audio. We also followed up with a question about how much the quality of the audio channel affected the listening

experience. Responses revealed that choppiness made the comments end abruptly sometimes, but, were understood whenever the remote users made intelligible comments.

#### Issues

i) Audio lag: One problem we constantly faced on the first day of the user study was with lag in transmitting audio across. In tests before the evaluation we had measured a lag between 2 - 3 seconds. However, during the evaluation lag frequently went up to 6-7 seconds and frequently ended up with severed connection to the remote audio stream. This led to a deteriorated user experience as portions of buffered audio data would play as soon as the remote end re-established connection. The cause of the problem was an overloaded wi-fi network. In addition to our study participants (that were seated in the front row of couches) the 5th floor cafe area had a good number of other viewers, this led to network congestion and increased audio lag. This delay was noticeable and participants pointed it out as one of the areas that needed improvement. To quote a user, “We heard the excitement about the near-goal after almost everything had calmed down. The audio cues on the other hand felt more instantaneous.”

ii) Spoken comments detection: Our algorithm for sending viewer comments relied on detection of start and stop of speech. However, the initial calibration period did not prove sufficient at acquiring an adequate threshold. For our system evaluation we had to re-adjust the threshold twice during the study. This problem could be overcome with some amount of learning in the system to readjust the threshold depending on volume of viewer comments. Another solution could be a mechanism for the system to prompt the user - two or three times after the initial calibration - to categorize a moment in the viewing experience as background noise or spoken comments. Additionally, a slider could be provided to adjust the threshold with feedback to the user whenever the system detects a segment with speech. The current version of our system provides text notification to the viewer while recording detected spoken comments.

#### Corrective measures:

Before the second test, we attempted to address this audio lag problem. We installed a Wi-fi access point at the location that was prone to more network congestion. Audio transmission rates improved, though, we recorded some severed connections a few times.

### **Visual Module**

This module mainly detected the number of faces in the room and general gaze direction. In the current version of our system we are able to track gaze of one viewer at a given time. Over the course of our user evaluation, we found detecting number of people in the room was reasonably accurate and provided a sense of people entering and leaving. However, the algorithm for tracking gaze detected even slight tilt in orientation and every time it occurred. Even though this was not false detection, it triggered an audio cue each time. This was aggravated by the fact that while watching the match viewers repeatedly turned around to talk to each other and gestured in disapproval or agreement by nodding their head every so often. Remote viewers that heard the audio cue play frequently found it disturbing and we had to turn off the stream coming from the sonic avatar. Moreover, we had mapped change in gaze direction to a *yawn* audio cue and this confused viewers. As one participant commented in jest, “Why is that guy always yawning?”

For the second day of our user evaluation, we replaced this audio cue with a more subtle *boing*<sup>2</sup> sound - the sound also faded out towards the end of play. We noted that viewers found this cue less intrusive when compared to the earlier yawn cue.

### **Galvanic Skin Response Sensing Module**

During the first study we found participants very enthusiastic to watch their skin conductance values plotted while they watched the game. Participants even switched seats so as to spend equal time having their overall activation measured. Prior to the start of the experiment we familiarized participants with the audio cue that would play in response to sudden skin conductance changes. However, we noted that a lot of this information was lost in the auditory environment particularly due to engaging conversations between participants. Further, participants could not map the cue to any particular kind of reaction, and found that the audio channel provided clues of how engaged the group at the other end was. We have more details about the use of this sensor in section 6.3.

---

<sup>2</sup>A sound representing the noise of a compressed spring suddenly released.



## **6.3 Engineering System Performance Study**

The purpose of this study was to record the performance of the Back Talk prototype implemented as part of this thesis. This assessment is important to understand the performance of each component; this would advise the design and implementation of future iterations of the system. We carried out tests to investigate how well the sensing environment works. We achieved this by running the system on a test subject while they used it - one module at a time, recording all sensor data, and then comparing it with hand labelled time data as noted by the conductor of the experiment.

### **6.3.1 Detection of Spoken Comments**

This study focused on obtaining an optimum operating point for the audio processing module that detects whenever a user has spoken. We studied the performance by running the module with television audio in the background and having a test subject speak a fixed set of comments at regular intervals. The goal was to plot this module's Receiver Operating Characteristic (ROC) curve. We varied the threshold to obtain points on the ROC curve. For each threshold value we recorded the number of true positives and false positives. We maintained the same experimental conditions while varying threshold. The experiment was first carried out using a directional microphone, and repeated with a non-directional microphone. The plot in Figure 6-6 compares the ROC curves for detection of spoken comments using the two different microphones as discussed earlier. The highlighted blue region indicates the ideal region of operation that maintains accuracy of detection around 90% for the directional microphone and between 85-90% for the non-directional case.

### **6.3.2 Laughter Detection Accuracy**

The goal of this test was to measure the accuracy of the laughter detector under two test conditions: with television audio playing in the background and without any television audio. The detector outputs a decision whether laughter was detected in the last one second at the end of every second of input. For each case - with and without television audio - these decision labels were compared with hand labels for data consisting of 120 seconds obtained from two users. These were measured 4 sets of 30 seconds each. It was observed that the performance of the detector dropped in the presence

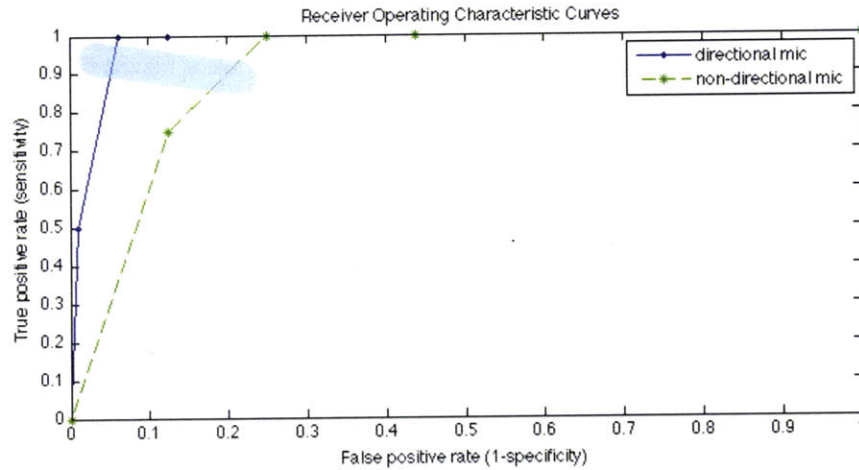


Figure 6-6: **ROC curves for detection of spoken comments using directional and non-directional microphones.**

of television audio. We measured accuracy from the observed set of true and false outcomes.

$$Accuracy = \frac{(number\ of\ true\ positives + number\ of\ true\ negatives)}{(numbers\ of\ true\ positives + false\ positives + false\ negatives + true\ negatives)} \quad (6.1)$$

In the absence of television audio playing this accuracy was found to be 81%. However, in the presence of television audio this value dropped to ~65%. This accuracy can be improved by using neural networks for classification.

### 6.3.3 Galvanic Skin Response Sensor

We evaluated the sensor performance by running it on 3 different users while they watched content. We had labeled data of sudden changes in reaction of users - for instance, laughter, sudden onset of

excitement about an event etc. We used a cartoon show *Tom and Jerry*, an episode of the television show *House MD* and an episode of *The Big Bang Theory*. The skin response was recorded for all three users and an observer noted down time-labeled sudden reaction to content. We found that relative to the baseline values (computed from the first 5 minutes of viewing) the sensed values gave an indication of onset of sudden changes. However, the relative difference in sudden changes between participants was significantly different, that is, the magnitude of change in skin conductance values for users was different. Our algorithm was designed to process sudden changes relative to the baseline values by computing the number of peaks in data every minute. Further, the algorithm recognizes a peak if it exceeds the baseline value by a certain pre-decided threshold. Through this evaluation test we found that this method could not accurately capture sudden changes for participants whose response did not vary significantly from the baseline values. In future iterations of the system we could improve the sensitivity of the module by factoring in magnitude of change in deciding the threshold for recognizing peaks.

#### **6.3.4 Detection of Number of Faces and Gaze Direction**

For multiple-face detection we used OpenCV's multiple face detection algorithm<sup>3</sup>, and for gaze tracking we used Google Tracker<sup>4</sup>. We found that the performance of this system to be quite sufficient in terms of accuracy and real-time performance. Since this subsystem is ancillary to the main audio-based Back Talk system, we did not perform any in-depth accuracy or performance testing. Though we did not evaluate the performance of the system, it was used in user experience studies. One observation that was also pointed out in the results from the user experience study was that the gaze direction detection updated the server every time tilt was detected. This resulted in the audio cue corresponding to "looking away" getting fired too often. The problem can be overcome in future iterations by experimenting with different threshold values (refer chapter 5 for details of the algorithm), and by training the system to update only significant gaze changes. Additionally, the system would also require some feedback mechanism in place to prevent triggering of audio cues too frequently.

---

<sup>3</sup><http://opencv.willowgarage.com/wiki/FaceDetection>

<sup>4</sup>Formerly Neven Vision, <http://www.nevenvision.com/>



## Chapter 7

# Future Work and Conclusion

In this chapter we will summarize future directions of the ideas and prototype presented in this thesis. We will also discuss conclusions based on our preliminary evaluation. Additionally, we will introduce possible applications related to Back Talk. The work in this thesis started with a goal of creating a sociable television watching experience, we will assess the effectiveness of our prototype in fulfilling this vision. This step is intended to highlight design and implementation choices that were more successful than others. We will also analyze the contributions of this thesis so as to position our work in the larger scheme of social television applications.

### 7.1 Contribution and Effectiveness of the Prototype

The primary problem this thesis addresses is that of connecting a distributed micro-social network watching television at the same time. The focus was on *peripheral awareness* of remote co-viewers and *natural communication* in the group. The solution was expected to create a sociable ambience without encumbering the participant with disruptive screen displays, text input channels or a mechanism for manually ‘emoting’ to the system.

**Prototype implementation contributions:**

- The prototype retained this automatic nature of capturing engagement and activity - we consider this a major shift from existing solutions. We promoted natural communication with an open audio channel. Further, to prevent television audio from sneaking into the listening experience we designed an algorithm to select only those segments that contained speech.
- The auditory experience contributed a new way of creating “surround presence” - by playing audio cues and audio streams to the left and right of the primary listener - achieved by left/right panning.
- Our prototype captured laughter, gaze direction, number of people in the room and overall arousal automatically without requiring the user to input these values into the system or convey this information as emoticons.

**Evaluation learnings:**

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. However, some interesting issues emerged from the user experience study that bring out pointers for future iterations of this prototype.

- 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 their 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.
- Participants opined that the current prototype required in-built audio normalization to match the match the volume of the audio cues to incoming spoken comments.
- The gaze direction algorithm updated change in gaze very frequently, which was not desirable. This module will need to update changes at longer intervals in order to

### **Summary:**

In conclusion, Back Talk can potentially enhance a viewing experience. From the first prototype of the system, we have directions to improve the listening experience and improve the automated modules for capturing user engagement and activity.

## **7.2 Future Directions**

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 list modifications in implementation of the prototype that can make such a 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. Future iterations can incorporate this feature by allowing custom picked audio cues - based on genre or associated with specific co-viewers. This will promote personalization of the system. This process would be akin to users selecting a particular ringtone for a person in their phone contact list to indicate arrival of a call. Implications of this option are plenty - custom cues can be selected for arrival of friends to be instantly notified of presence - idiosyncrasies of a co-viewers can be mapped to characteristic audio cues and activated by specific triggers from the engagement sensing modules.
- **Imported audio streams:** Presence in our system is primarily conveyed through an auditory environment. A possible option is importing audio streams from a remote location 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 television (especially 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 - reality shows like *American Idol*. The core selling point of this feature, is a customized auditory environment from a live event.

- **Implementation:** We suggest alternatives for implementing this system in a more compact form as opposed to the present distributed modules. Looking ahead, we envisage the system comprising two main processing components – the cell phone and the television. Figures 7-1 illustrates these units. The television and cell phone work together as an engagement capture unit - together they 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. This can be achieved if developers are granted hardware access to make use of the two microphones so as to cancel out incoming extraneous television sounds (similar to noise canceling already present in cell 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 processing captured audio with a light-weight classification algorithm. Additionally, the galvanic skin response sensor can be a detachable component powered by the cell phone with a micro-controller processing sensed values, and, only communicating sudden arousal - events - to the server. The television would perform the remaining engagement capture functions - face detection and gaze direction tracking.

Figure 7-2 alludes to the vision of designing the television as a local *home server* that subsumes all the engagement sensing modules. Televisions equipped with cameras could serve the role of detecting number of people entering and leaving, and general gaze direction. Further, it offers an option for carrying out the necessary audio processing – the bezel of the television could be fitted with an array of microphones to capture viewer audio. The galvanic skin response sensor could communicate sensed arousal - ostensibly via bluetooth - to the television that in turn could communicate with the central server.



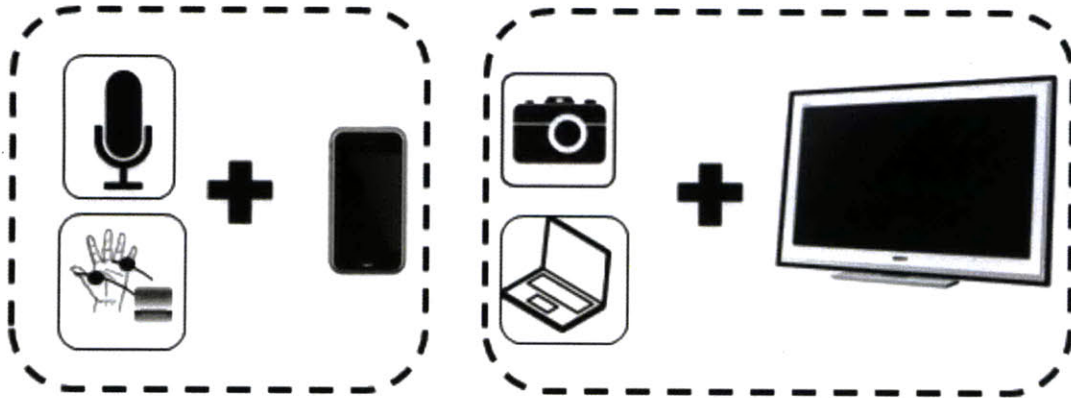


Figure 7-1: Cell phone and television units processing engagement.

- 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 in 1D by left/right panning of the sources of audio and distance based volume control of the sound streams. However, a more *surround* experience can ostensibly be created by using a 5.1 speaker system. Such a system would require modifications in audio capture, and multi-channel input, but, could lead to better spatial positioning of the audio source.

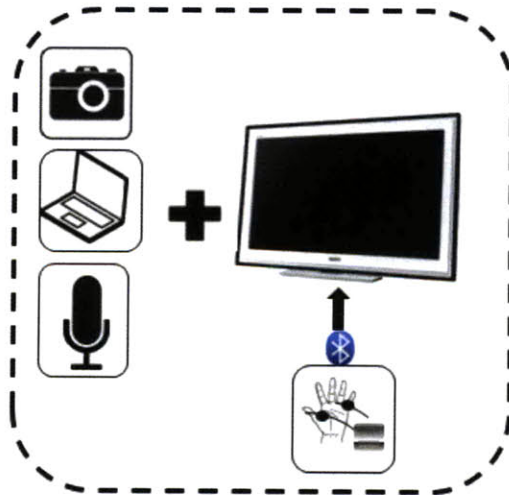


Figure 7-2: The television as the local server.

- **Future evaluation studies:** In successive prototype implementations and evaluation studies it would be meaningful to attempt to answer the question - does repeated use of the Back Talk system facilitate better ambient use of activity of remote viewers? Additionally, the audio cues used in the current implementation are quite literal - for example, footsteps to indicate entry of a viewer and a door slamming to indicate people leaving. We would need to evaluate test subjects' responses to various other cues - our current selection of cues was intended to be literal so as to evaluate if mapping these cues to a particular trigger was easier. Some example variant cues are two-tone beeps indicating people entering and leaving (similar to IM clients).

## Appendix A

# Audio Processing Experiments

We list and describe relevant experiments that were carried out in order to address the problem of separation of television audio from a viewer's spoken comments.

The first set of experiments focused on source separation by simple background noise subtraction. For the case in which the list of possible television programs is known, we also have knowledge of the television audio that is being played a priori<sup>1</sup>. We therefore performed quick tests to determine how well this information could be used.

In the first set of experiments we recorded a segment containing viewer comments and television audio playing in the background. At the start of the show we also played a series of square pulses that were prepended to the original source of television audio. These square pulses were useful in cross-correlating the original source of audio with the recorded segment so that comparisons made between these two signals start at the same initial point. Next we attempted to subtract out the television audio from the recorded audio (containing viewer comments). The resultant audio was not a significant improvement over the otherwise noisy viewer comments and in some parts had little or no (positive) effect on the audio.

The second part of this experiment included recording only the television audio (recording A) as heard by the microphone in the room and under similar conditions - distance and orientation of

---

<sup>1</sup>This would not be the case if viewers decide to pick a show we do not have beforehand, for instance, a live soccer game.

microphone from the television, volume level of television audio - we recorded for the same time length viewer comments while the show was playing (recording B). We then followed the same procedure of cross-correlation to have both the streams start at the same initial point. The television audio-only recording was subtracted from the audio recording that had both viewer comments and television audio. This experiment was designed as a quick way to assess how useful recording A would be in reducing background television audio. In order to get recording A without actually playing television audio in the room and recording it, we would have had to measure the impulse response of the room and then convolve this signal with our original television audio signal. Our results did not indicate significant background noise cancellation and resulted in muffled viewer comments. Further, this option is not easily extensible and is heavily dependent on the characteristics of the room, and orientation of the microphone.

Dynamic noise thresholding: In this method we attempted to detect the start and stop of viewer comments based on amplitude difference in portions with viewer comments and without relative to the original television audio amplitude levels. For this purpose the first 5 seconds was maintained without any comments to obtain an average ambient noise level. Similarly, average energy level in the first 5 seconds of purely the television audio was also computed. A ratio of these two values gave a measure by which the pure television audio signal had to be scaled by. Following this initial calibration and scaling, every window ( 30 ms) of incoming audio is compared with a corresponding segment in the scaled television audio. Comparison is made by calculating a ratio of the two values and deciding if it exceeds a value obtained from the initial calculation. This process is dynamic in that at the end of a speech segment we analyze the next few windows to obtain an average ambient noise value and compare it with the current ambient noise value; the higher of these two values is set as the new ambient noise value.

An interesting observation from applying this technique was that the method was prone to recognizing some segments of television audio that were relatively louder than others as viewer comments. This was particularly the case for canned laughter tracks played during a show. Further, another frequently occurring problem with this technique was the true end of spoken comments being miscalculated. Analysis revealed that towards the end of speech the energy in the samples significantly drops and becomes almost comparable to the television audio.

rithm. We based our implementation on the algorithm described in [36] and modified its parameters to suit our detection purposes. Since our aim was to obtain only those segments that contained speech, detection of pauses is important. Aligned with the settings described in the paper, our experiments also used a single microphone. The technique described in the paper is intended to identify speech pauses. We used the underlying mechanism for identifying segments without viewer comments as pauses and extracted the remaining segments as spoken comments. This method calculates the signal's temporal power envelope, a low-pass band power envelope and high-pass band power envelope. The maximum and minimum values for each of these envelopes are updated and their difference is computed. The values are compared against certain threshold values to detect pauses. Refer [36] for details of the algorithm.

While this technique yielded relatively good spoken comments detection as compared to the previous methods it worked best when the signal with viewer comments was pre-recorded. As a result, it could not be easily adapted to work real-time.



# Bibliography

- [1] “Watching TV Together, Miles Apart”, [http://www.nytimes.com/2010/01/04/technology/internet/04couch.html?\\_r=1&hpw](http://www.nytimes.com/2010/01/04/technology/internet/04couch.html?_r=1&hpw).
- [2] S. Zhao, “Toward a taxonomy of copresence”, *Presence: Teleoperators & Virtual Environments*, vol. 12, no. 5, pp. 445–455, 2003.
- [3] J. Short, E. Williams, and B. Christie, *The social psychology of telecommunications*, John Wiley & Sons, 1976.
- [4] C. Harrison and B. Amento, “CollaboraTV: Using asynchronous communication to make TV social again”, *Adjunct Proceedings of EuroITV2007*, pp. 218–222, 2007.
- [5] T.A. Rasmussen, “Interactive Television—social use or individual control?”, in *Paper to be presented at the 2nd European Conference on interactive television: Enhancing the Experience*, 2003.
- [6] S. Goldenberg, “Digital video recorders and micro-social networking: Recreating the shared watching experience of television”, in *Adjunct Proceedings of the European Conference on Interactive Television*, 2007.
- [7] G.L. Martin, “The utility of speech input in user-computer interfaces.”, *INT. J. MAN MACH. STUD.*, vol. 30, no. 3, pp. 355–375, 1989.
- [8] B.L. Chalfonte et al., “Expressive richness: a comparison of speech and text as media for revision”, in *Procs. of the SIGCHI conf.* ACM, 1991, p. 26.

- [9] J. Cohen, "Monitoring background activities", in *SANTA FE INSTITUTE STUDIES IN THE SCIENCES OF COMPLEXITY-PROCEEDINGS VOLUME-*. ADDISON-WESLEY PUBLISHING CO, 1994, vol. 18, pp. 499–499.
- [10] B. Arons, "A review of the cocktail party effect", *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.
- [11] J. Lull, "The social uses of television", *Human communication research*, vol. 6, no. 3, pp. 197–209, 1980.
- [12] M. Chuah, "Reality instant messenger: The promise of iTV delivered today", *Ardissono & Buczak*, 2002.
- [13] J. Abreu, P. Almeida, and V. Branco, "2BeOn-Interactive television supporting interpersonal communication", in *Multimedia 2001: proceedings of the Eurographics Workshop in Manchester, United Kingdom, September 8-9, 2001*. Springer Verlag Wien, 2002, p. 199.
- [14] T. Coppens, L. Trappeniers, and M. Godon, "AmigoTV: towards a social TV experience", in *Proceedings from the Second European Conference on Interactive Television "Enhancing the experience"*, University of Brighton, 2004, vol. 36.
- [15] T. Regan and I. Todd, "Media center buddies: instant messaging around a media center", in *Proceedings of the third Nordic conference on Human-computer interaction*. ACM New York, NY, USA, 2004, pp. 141–144.
- [16] Crysta Metcalf, Gunnar Harboe, Joe Tullio, Noel Massey, Guy Romano, Elaine M. Huang, and Frank Bentley, "Examining presence and lightweight messaging in a social television experience", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 4, pp. 1–16, 2008.
- [17] E. Boertjes, J. Klok, and S. Schultz, "ConnecTV: Results of the Field Trial", *Adjunct Proc. EuroITV*, pp. 21–22, 2008.
- [18] S. Goldenberg, "Creating augmented and immersive television experiences using a semantic framework", in *Proceeding of the 1st international conference on Designing interactive user experiences for TV and video*. ACM, 2008, pp. 45–48.



- [19] “Vision television”, <http://web.media.mit.edu/~stefan/vtv/>.
- [20] “Reflexion”, <http://web.media.mit.edu/~stefan/hc/projects/reflexion/>.
- [21] M. Baca and H. Holtzman, “Television meets Facebook: Social Networks through Consumer Electronics”, in *Workshop on Sharing Content and Experiences with Social Interactive Television, co-located with the European Interactive TV Conference (EuroITV2008), Salzburg, Austria, 2008*.
- [22] D. Hindus, M.S. Ackerman, S. Mainwaring, and B. Starr, “Thunderwire: a field study of an audio-only media space”, in *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. ACM, 1996, pp. 238–247.
- [23] E.D. Mynatt, M. Back, R. Want, M. Baer, and J.B. Ellis, “Designing audio aura”, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1998, p. 573.
- [24] N. Sawhney and C. Schmandt, “Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments”, *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 7, no. 3, pp. 383, 2000.
- [25] David Geerts and Dirk De Grooff, “Supporting the social uses of television: sociability heuristics for social tv”, in *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, New York, NY, USA, 2009, pp. 595–604, ACM.
- [26] “OpenCV”, <http://opencv.willowgarage.com/wiki/>.
- [27] P. Viola and M.J. Jones, “Robust real-time face detection”, *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [28] R. Leinhardt and J. Maydt, “An extended set of Haar-like features”, in *Proc. Int. Conf. on Image Processing*, 2002, pp. 900–903.
- [29] P. Viola and M. Jones, “Fast and robust classification using asymmetric adaboost and a detector cascade”, *Advances in Neural Information Processing Systems*, vol. 2, pp. 1311–1318, 2002.

- [30] M. Knox and N. Mirghafori, “Automatic laughter detection using neural networks”, in *Proceedings of INTERSPEECH*, 2007, pp. 2973–2976.
- [31] L. Kennedy and D. Ellis, “Laughter detection in meetings”, in *NIST ICASSP 2004 Meeting Recognition Workshop*. Citeseer, 2004, pp. 118–121.
- [32] R.W. 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.
- [33] “Apple Developer Library: AudioQueue Reference”, <http://developer.apple.com/iphone/library/documentation/MusicAudio/Reference/AudioQueueReference/Reference/reference.html>.
- [34] “Apple Developer Library: CAAudioToolbox”, <http://developer.apple.com/iphone/library/documentation/MusicAudio/Reference/CAAudioToolboxRef/index.html>.
- [35] “Apple Developer Library: Audio Stream Reference”, <http://developer.apple.com/mac/library/documentation/MusicAudio/Reference/AudioStreamReference/Reference/reference.html>.
- [36] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics”, *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, 2002.
- [37] Jim Hollan and Scott Stornetta, “Beyond being there”, in *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, 1992, pp. 119–125, ACM.
- [38] David Geerts, “Comparing voice chat and text chat in a communication tool for interactive television”, in *NordiCHI '06: Proceedings of the 4th Nordic conference on Human-computer interaction*, New York, NY, USA, 2006, pp. 461–464, ACM.
- [39] Gunnar Harboe, Noel Massey, Crysta Metcalf, David Wheatley, and Guy Romano, “The uses of social television”, *Comput. Entertain.*, vol. 6, no. 1, pp. 1–15, 2008.

- [40] L. Oehlberg, N. Ducheneaut, J.D. Thornton, R.J. Moore, and E. Nickell, “Social TV: Designing for distributed, sociable television viewing”, in *Proceedings of the European Conference on Interactive Television-EuroITV 2006*, 2006, pp. 251–262.
- [41] R.D. Martin, A.L. Santos, M. Shafran, H. Holtzman, and M.J. Montpetit, “neXtream: a multi-device, social approach to video content consumption”, in *Proceedings of the 7th IEEE conference on Consumer communications and networking conference*. IEEE Press, 2010, pp. 779–783.
- [42] S. Agamanolis, “At the intersection of broadband and broadcasting: How ITV technologies can support human connectedness”, in *Proceedings of the European Conference on Interactive Television-EuroITV 2006*. Citeseer, 2006, pp. 17–22.
- [43] K. Karahalios and J. Donath, “Telemurals: linking remote spaces with social catalysts”, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 615–622.
- [44] K. Isbister and K. Höök, “Supple interfaces: designing and evaluating for richer human connections and experiences”, in *CHI’07 extended abstracts on Human factors in computing systems*. ACM, 2007, p. 2856.
- [45] G. Harboe, N. Massey, C. Metcalf, D. Wheatley, and G. Romano, “Perceptions of value: The uses of social television”, *Lecture Notes in Computer Science*, vol. 4471, pp. 116, 2007.
- [46] R.D. Putnam, *Bowling alone: The collapse and revival of American community*, Touchstone Books, 2001.