Audio-visual localization of multiple speakers in a video teleconferencing setting

Bill Kapralos[1,3], Michael R. M. Jenkin[1,3], Evangelos Milios[2,3]

[1]Department of Computer Science, York University, Toronto, Ontario, Canada
[2]Faculty of Computer Science, University of Dalhousie, Halifax, Nova Scotia, Canada
[3]Centre for Vision Research, York University, Toronto, Ontario, Canada

{billk, jenkin}@cs.yorku.ca, eem@cs.dal.ca

Abstract

Attending to multiple speakers in a video teleconferencing setting is a complex task. From a visual point of view, multiple speakers can occur at different locations and present radically different appearances. From an audio point of view, multiple speakers may be speaking at the same time, and background noise may make it difficult to localize sound sources without some *a priori* estimate of the sound source locations.

This paper presents a novel sensor and corresponding sensing algorithms to address the task of attending, simultaneously, to multiple speakers for video teleconferencing. A panoramic visual sensor is used to capture a 360^{o} view of the speakers in the environment and from this view potential speakers are identified via a color histogram approach. A directional audio system based on beamforming is then used to confirm potential speakers and attend to them. Experimental evaluation of the sensor and its algorithms are presented including sample performance of the entire system in a teleconferencing setting.

1. Introduction

With the advent of the "Global Village", teleconferencing has found a wide range of commercial applications: From facilitating business meetings to aiding in remote medical diagnoses, teleconferencing is used by corporate, educational, medical, government and military organizations. Teleconferencing enables new operational efficiencies resulting in reduced travel costs, faster business decision making, increased productivity, reduced time to market, and remote classroom environments (Davis, 1999).

Various commercial teleconferencing systems exist, although most have been developed for use by two static participants (one at each end of the connection). For example, Microsoft's NetMeeting™ (Freed, 2000) allows for continuous video and speech transmission, document and "white-board" sharing by the two participants. Systems intended for multiple speakers have also been developed, however, they typically focus on a single speaker at a time and provide limited, if any, automatic speaker tracking. Development of more general teleconferencing systems, systems that can localize multiple parallel speakers, is hampered by the technology used to acquire the visual scene and the complexities involved in attending to multiple speakers in the audio domain. The majority of existing teleconferencing systems utilize conventional cameras, thereby providing a limited number of static or manually tracked views. As a consequence, in a multiple speaker setting, either the speaker must move into the camera's view or a camera operator must be used to manually locate, track and choose between speakers. This is inconvenient for the participants and has deterred many from using such systems. In addition to the cost overhead, the presence of camera equipment and/or an operator during a teleconferencing session can interfere with the group dynamics (Yong et al, 2001). Vision-based systems capable of detecting and tracking human faces exist, however, they also employ

conventional camera lenses, which capture only a narrow field of view. The limited field of view can be overcome by employing multiple cameras (see Wang, et al., 2000) however, in addition to the extra equipment introduced in the room and the associated overhead, the multiple camera system must be accurately calibrated to ensure a correct correspondence between the individual cameras.

Teleconferencing systems must be able to capture and transfer audio (e.g. the speaker's voice) as well as the video signal. In order to capture the speech of multiple participants, popular methods include; having each participant wear their own "tie-clip" microphone and having a human operator determine which microphone to monitor and having speakers physically move to a location where they can speak into a shared microphone. Although automatic sound localization systems exist, most rely on extensive microphone arrays (e.g., Brandstein et al., 1995; Griebel et al., 2001; Rabinkin, 1994; Zotkin et al., 2000) that require expensive specialized equipment, are computationally intensive and are typically, non-portable.

Systems have been developed to deal with multiple speakers in an autonomous manner. For example, the PictureTel 900^{TM} teleconferencing system manufactured by Polycom employs a microphone array capable of localizing a speaker over a 360^{o} field of view. Once the speaker has been localized in the audio domain, a computer controlled pan-tilt-zoom camera is steered to fixate on the audio source. Although this is definitely an improvement over the traditional manually operated camera systems, moving the pan-tilt-zoom camera so that it is focused on the speaker is time consuming and therefore, may interfere with the meeting. Moreover, in order for the sound localization system to be most effective, it must be positioned as close as possible to the speaker. Finally, although the PictureTel 900 system is capable of attending to a single audio source, it is incapable of capturing the speech and video of multiple participants simultaneously.

An automatic videoconferencing system utilizing sound localization with a set of microphone arrays, and video based face tracking and pose estimation has been developed by Wang et al. (2000). Although this system is capable of tracking multiple speakers in the visual domain, it employs four 4-element microphone arrays and multiple - one per potential speaker - pan-tilt-zoom cameras leading to potentially complex calibration requirements and a significant investment in hardware. Furthermore, the system is has a limited visual field of view, and can only attend to audio events and events within the current field of view of its various cameras.

Because a single standard video camera needs to be fixated on the speaker, it is not an effective sensor for the multi-speaker teleconferencing task. In order to address this issue, a novel hardware device was constructed to provide the raw sensor data. Figure 1 illustrates the hardware components comprising the “Eyes 'n Ears” sensor. The sensor consists of a Cyclovision Paracamera omnidirectional video system (Baker and Nayar, 1999) coupled with a four element microphone array. The sensor is compact, lightweight, portable and is meant to be placed in the center of a table with the participants of the teleconference session seated around it.

Cyclovision's Paracamera omnidirectional camera system consists of a high precision paraboloidal mirror and a specially designed lens. The optics assembly permits the Paracamera to capture a 360^{o} *hemispherical* view of all potential speakers from a single viewpoint. Once the hemispherical view has been obtained, it may be easily un-warped (Peri and Nayar, 1997) producing a panoramic view. From this panoramic view, perspective views corresponding to different portions of the scene can be easily extracted (see Figure 2).

The ability of the Paracamera to capture an image of the entire hemisphere makes it very attractive for a number of different applications. For example, in Stiefelhagen et al. (1999) and Yong et al. (2001) an omnidirectional camera is used to capture the simultaneous video of each

participant in a small group meeting. Omnidirectional devices have also been used in many other vision based applications, including surveillance (Boult et al., 2001, Gutchess et al., 2000), autonomous robot navigation (Zheng and Tsuji, 1992), virtual reality (Yasushi, 1999), telepresence (Yasushi, 1999), remote view from a Dolphin (Boult, 2000) and pipe inspection (Basu and Southwell, 1995). Omni-directional cameras have also been used in a vision based navigation system which allows a robot to determine its own position and orientation (Fiala and Basu 2002) and an array of omnidirectional cameras has been used for the real-time tracking of human movements and faces (Huang and Trivedi, 2001).

In the Eyes ‘n Ears sensor, the Paracamera is used to identify potential speakers in the environment. Using a statistical color model, the video system locates regions of human skin present within the Paracamera's view. These skin regions correspond to faces, arms and other exposed skin regions, as well as other skin-colored distracters in the image. Regions of skin in the camera image are then grouped together to form a cluster. Each cluster of skin regions is assumed to correspond to one particular person. Under the assumption that the face of a person in the Paracamera image is further away from the center of the image relative to the other skin regions, the locations within the camera image of potential speakers are identified. Once each face has been found, an estimate of its direction in the real world relative to the Paracamera is computed and provided to the audio system as the direction to a potential sound source (e.g. voice of a speaking person).

The audio system consists of four omnidirectional microphones (m_1, m_2, m_3 and m_4), mounted in a static pyramidal shape about the base of the Paracamera, which provide an economical and portable acoustic array capable of localizing a sound source in 3-space (Guentchev, 1997). Given the initial estimate of the direction to each speaker's face in the real

world as identified by the visual system, using beamforming (Johnson and Dudgeon, 1993) and sound detection techniques, the audio system detects and validates each speaker and focuses on the speech of each individual thus reducing unwanted noise and sounds propagating from other directions. Sections 2 and 3 provide details of the vision and audio subsystems respectively. Section 4 describes the system in operation and presents experimental results of the two subsystems. Section 5 discusses the approach and presents potential directions for future research.

2. Vision System

Skin color is often proposed as an economical and efficient cue to face detection. Color is one of the simplest attributes in a set of pixels comprising the image (see Jones and Rehg, 1998). Color does not require extensive computational processing to compute, the color of an object may be used as an identifying feature that is local to the object and color is largely independent of the view and resolution. In general, color is invariant to partial occlusion, rotation in depth, scale and resolution changes (Raja et al., 1998). Furthermore, there are fast and simple color based human detection and tracking systems available (see Chai and Ngan, 1999; Hans et al., 1999; Howell and Buxton 2000; Herpers et al., 1999; Jones and Rehg, 1998; Phung et al., 2002; Srisuk and Kurutach, 2002; Stiefelhagen et al., 2002; Terrilion and Akamatsu, 1998;).

A major difficulty in using color information as a cue is the lack of *color constancy* (Forsyth, 1990); changes in lighting conditions can lead to variations in the colors of an image (e.g. the color of an object in an image may change under different lighting conditions even when the object itself does not change). The lack of color constancy is most apparent in the RGB (red, green, blue) color space where intensity is distributed equally to all three color channels (Raja et al., 1998). In order to reduce these effects, in this work, color is represented in the HSV

(hue, saturation, value) color space. Furthermore, skin color of different people varies primarily in intensity (Stiefelhagen et al., 2002). Therefore, if value is ignored, the skin color of people, regardless of race, forms a tight cluster in HS (hue-saturation) space (Crowley and Beard, 1997; Raja et al, 1998).

Color histograms (Swain and Ballard, 1991) have been found to be an effective mechanism for representing colored objects in an image. A color histogram is a representation of the distribution of the discrete colors available in an image, i.e. for each possible color a pixel of an image may take on, the color histogram provides a count of the number of pixels in the image (or in an image region) with that corresponding value. Color histograms are invariant to translation, rotation about an axis perpendicular to the image and change only slightly with rotations about other axis, occlusion and change of distance to the object. Furthermore, the color histogram of different objects can differ extensively (Swain and Ballard, 1991). Color histograms have proven to be effective representations of two and three dimensional objects and have been used in a wide variety of computer vision applications, including image matching and retrieval (Mehtre et al., 1995, Swain and Ballard, 1991), human tracking (Lu and Tan, 2001) and skin detection (Jones and Rehg, 1998).

Color histogram models for skin and non-skin color classes for traditional cameras have been constructed previously and are freely available (e.g., Jones and Rehg, 1998). These models are constructed using data from images obtained with normal cameras. However, the non-standard lens and curved mirror assembly of the Paracamera was found to distort the colors of skin tones and thus the Jones and Rehg model was found to be unsuitable for the Paracamera sensor. Instead, two-dimensional hue-saturation model histograms for both skin and non-skin color classes were constructed by manually classifying portions of images obtained with the

Paracamera. Value was ignored, while hue and saturation values were discretized to 32 and eight discrete values respectively (e.g. each color histogram contains a total of $32 * 8 = 256$ possible HS (hue-saturation) pairs).

To construct color histograms for skin and non-skin regions, a total of 154,310 skin pixels were obtained from 35 subjects of various racial groups to ensure a wide variety of skin colors. For each subject, samples of their hands, face (and in several cases legs) were obtained over multiple images. The subjects were asked to change both their pose and their distance relative to the Paracamera to ensure samples in different lighting and orientation conditions were obtained. A total of 307,546 non-skin pixel samples were obtained by sampling portions of Paracamera images that did not contain human skin. Once again, the samples were obtained in different locations and under a variety of lighting conditions. Figure 3 provides a graphical illustration of the resulting histogram distributions for (a) the skin color and (b) non-skin color models.

Given the skin and non-skin histograms, the probability that a particular hue and saturation pair (referred to as HS) belongs to either the skin $P(HS \mid skin)$ or non-skin $P(HS \mid \neg skin)$ classes was estimated (see Jones and Rehg, 1988) as

$$P(HS \mid skin) = \frac{skin[HS]}{T_s}, \; P(HS \mid \neg skin) = \frac{nonskin[HS]}{T_{ns}}$$

where $skin[HS]$ and $nonskin[HS]$ are the counts in bin HS of the skin and non-skin histograms respectively. T_s is the total number of pixels contained in the skin histogram while T_{ns} is the total number of pixels contained in the non-skin histogram.

When $skin[HS]$ is greater than a predefined threshold value δ, the probability

$P(skin \mid HS)$, that a pixel in the hue-saturation color space is skin color is determined using Bayes rule, otherwise, $P(skin \mid HS)$ is set to zero

$$P(skin \mid HS) = \begin{cases} 0 & if\ skin[HS] < \delta \\ \dfrac{P(HS \mid skin)P(skin)}{P(HS \mid skin)P(skin) + P(HS \mid \neg skin)P(\neg skin)} & otherwise \end{cases}$$

where $P(skin) = \dfrac{T_s}{T_s + T_{ns}}$ and $P(\neg skin) = \dfrac{T_{ns}}{T_{ns} + T_s}$ are the probability of a pixel belonging to the skin and non-skin color classes respectively. Once the probability that

a particular HS pixel value corresponds to skin has been calculated, the pixel is classified as skin if $P(skin \mid HS) \geq \Delta$, where Δ is a pre-defined threshold

In order to enhance the efficacy of the color histogram skin detection process a number of pre- and post-filtering stages are applied. These are:

i) Prior to applying the color histogram skin detection process, an image differencing process is used to remove static (background) image locations from further processing. A time-averaged background image is computed and portions of the image that do not differ significantly from this background are ignored.

ii) Once individual pixels have been labeled as skin pixels, these pixels are combined into skin regions. Holes or concavities that may be present in these skin regions are reduced by applying dilation and erosion operations.

Following the dilation and erosion operations, a unique identifier is assigned to each region or connected component. Components smaller than a predefined threshold are ignored. Components in close proximity are grouped together, to form a cluster. The component in each cluster furthest from the center of the Paracamera image is identified as a potential face. Figure 4

shows (a) a sample image of users in a teleconferencing setting, and (b) the identified skin regions (yellow) and potential faces (red crosses).

Given the centre $(\bar{x},\bar{y})$ of a face or skin region as identified by the visual system, the direction to $(\bar{x},\bar{y})$ is computed in world co-ordinates. Since the focus of the paraboloidal mirror is also its centre of projection, it is straightforward to convert camera pixel locations to 3D direction vectors (Gutchess et al., 2000). Once potential speaker directions have been determined, each direction is probed by the audio system to determine if the speaker direction has a corresponding signal in the audio domain.

3. Audio System

The audio system is responsible for confirming the presence of speakers in the directions estimated by the video system. Since the participants of a teleconference session will generally be speaking, speaker confirmation is achieved by detecting the presence of a sound source within a small set of directions centered on the direction obtained by the video system.

Due to various acoustical factors including reverberation, a propagating speech signal may be degraded or altered extensively before reaching a microphone. Furthermore, this degradation increases as the distance between the sound source and the microphone becomes greater (Renomeron, 1997). As a result, the information captured by a microphone and transferred to the other parties may be incomplete and may lead to confusion amongst the participants. This has made teleconferencing a poor substitute for person-to-person contact (Zotkin et al, 2000). To reduce this problem the microphone should be positioned close to the person speaking ensuring the Signal-to-Noise Ratio (SNR) of the captured speech is very large (e.g. the power of any captured noise is negligible relative to the power of the speech signal) (de Mori 1998). Unfortunately, this is not always practical, especially in large rooms where the

distance between the participants and the microphone is large and when participants at different locations in the room need to speak and be heard. Traditional methods to overcome this problem involve providing each participant with their own microphone. However, an operator is then needed to determine who is speaking and turn on the speaker's microphone while turning all other microphones off (Rabinkin, 1994). The task of determining who is speaking and turning off all other microphones may be performed using speaker detection software (e.g. the signal of each microphone containing the greatest energy may be used to indicate the speaker). Such an approach is limited by the number of audio input channels available and since the microphone of each participant requires its own input channel such an approach is certainly not scalable. Alternatively, a directional microphone can be focused in the direction of the person speaking. However, directional microphones are rather bulky, are limited to a fixed beam pattern and therefore do not allow for the tracking of multiple speakers (Wilson, et al., 2001).

There are many situations in which a physical signal present in our environment is monitored by some appropriate sensor, providing us with information about the surroundings. To increase the effectiveness of the sensor apparatus, rather than relying on a single sensor, an array of sensors may be used instead. A sensor array is a set of sensors placed at different locations to spatially sample a propagating wave-field (e.g., sample a wave-field emanating from a particular direction), which may be electromagnetic, acoustic, seismic, etc. (Johnson and Dudgeon, 1993). Using the array allows a speech signal to be captured from any location in the room while greatly attenuating background noise, reverberation and sounds from other sound sources. Most importantly, however, it allows for the automatic speech acquisition with minimal noise and detection and the tracking of multiple active speakers regardless of their position in the room (de Mori, 1998), thus alleviating the participants the need to wear and carry their own microphone.

Although various methods of beamforming have been developed, the simplest and perhaps most common is referred to as *delay and sum beamforming* (Johnson and Dudgeon, 1993). Consider a sound source located at some position x_s in three dimensional space. Furthermore, consider an array of *M* microphones (each microphone is denoted by m_i for $i = 1...M$) where each microphone is at a unique position x_i and each is in the path of the propagating waves emitted by the sound source. In general, the time taken for the propagating sound wave to reach each microphone will differ and the signals received by each of the microphones will not have the same phase.

The differences in the time of arrival of the propagating wave at the sensors depend on the direction from which the wave arrives, the positions of the sensors relative to one another, and the speed of sound v_{sound}. Beamforming takes advantage of these time differences between the time of arrival of a sound at each sensor (Rabinkin, 1994), and allows the array signals to be aligned after applying suitable delays to steer the array to a particular sound source direction of arrival. Beamforming consists of applying a delay Δ_i and amplitude weighting w_i to the signal received by each sensor $s_i(t)$ and then summing the resulting signals

$$z(t) = \sum_{i=0}^{M-1} w_i s_i(t - \Delta_i)$$

where $z(t)$ is the beamformed signal at time *t* and *M* is the number of sensors (Johnson and Dudgeon, 1993).

Since the output of the beamformer will be maximized when it is steered in the direction of the source, the beamformer may also be used to determine the direction to a sound source(s). This may be accomplished by focusing the beamformer to every possible direction and recording the direction corresponding to the strongest beamformer output. The direction with the

maximum output corresponds to the location of a propagating source. Unfortunately, such an approach is not feasible in general. It will take far too much time and processing effort to actually focus the beamformer to every possible direction. However, when the number of potential sound source directions is restricted in some manner, beamforming is an effective method of audio detection. Essentially, this is the purpose of the video system. For this application the speech of the participants is the signal of interest. By locating the face of potential speakers present in the Paracamera's view, the video system reduces the number of potential sound sources from many thousands to a very small number (e.g. under 10), making the audio system's task tractable.

In voice recognition and detection applications, the signals of interest fall within a small frequency region. Although humans are capable of perceiving sounds from 20-20000Hz (Adler, 1996), most speech falls within the frequency range of 200-4000Hz (Hioki, 1990). By filtering the signal received at each of the microphones, energy in frequencies that do not lie in the region of interest can be attenuated. This reduces noise present in the original signal and leads to more accurate sound detection. Filtering is accomplished entirely in software using a band pass digital FIR filter.

If the direction of propagation between the sound source and each microphone of the array are assumed to be the same (far-field assumption), then to focus the array in some direction $\vec{\beta}$, the delays Δ_i are computed as:

$$\Delta_i = -\frac{\vec{\beta} \bullet \vec{x_i}}{v_{sound}}$$

where $\vec{\beta}$ is the unit vector denoting the direction of propagation relative to the array's origin, $\vec{x_i}$ is the vector from the array reference point towards the i'th microphone and the speed of sound v_{sound} is assumed to be constant at 345m/s.

Geometrically, the time delay is determined by projecting $\vec{x_i}$, onto the unit vector $\vec{\beta}$. The length of this projection gives the difference in distance the sound must travel to reach the array reference and the i'th microphone. Dividing this distance by the speed of sound v_{sound}, gives the associated time delay.

When the delays have been determined correctly (e.g. to accurately match the source location relative to the array) and applied to each microphone signal, the resulting microphone signals will be in phase. As a result, when the signals are summed together to form the beamformed signal, they will reinforce each other causing the energy of the beamformed signal to reach a maximum (see Figure 5). Delays that do not correspond to a true sound source direction (e.g. the beamformer is incorrectly steered to some other direction), will cause a reduction in the energy level of the beamformed signal.

3.1 Implementation Details

The delay derivation described above assumes a continuous time signal and a delay of any arbitrary time. In reality however, this luxury does not exist! The signal obtained by each of the microphones is actually a sampled version of the original propagating signal and therefore, only integer delays may be achieved easily. As a result, the beamformer may not be able to focus directly to locations corresponding to non-integer delays. For this work, the delays are calculated as described above, assuming a continuous time signal. After computing the continuous time delay, the discrete time delay $\Delta_{i_{discrete}}$ is obtained by rounding the continuous time

delay to the nearest integer. Once the discrete delays have been calculated, the beamformed signal $s_{beamform}$ can be constructed as

$$s_{beamform} = \sum_{j=0}^{N} s_1[j + \Delta_{1_{discrete}}] + s_2[j + \Delta_{2_{discrete}}] + s_3[j + \Delta_{3_{discrete}}] + s_4[j + \Delta_{4_{discrete}}]$$

where $N = 2048$ is the size of the sample window.

Two measures are used to determine if the beamformed signal corresponds to a speaker, the signal variance and signal difference.

Signal Variance Criterion: The variance of the beamformed signal must be substantially higher than that of the background noise. Generally, the signal variation associated with a signal emanating from sound sources such as speech, music and impulsive sounds, is greater than the variance of background noise. Following the approach of Guentchev (1997), the variance is computed based on k = 32 sample sub-windows. The variance of each sub-window is computed and finally, the variance for the entire window is estimated by the mean variance of the 52 sub-windows.

$$V_{signal} = \frac{\sum_{i=0}^{M-1} \sum_{j=i*k}^{[(i+1)*k-1]} \frac{\left\| s[j] - \bar{s} \right\|^2}{k-1}}{M}$$

During system calibration, the variance of the background noise is computed, and this is used to establish a threshold value (V_{thresh}) for the background noise. A potential speaker is only confirmed when the variance of the beamformed signal V_{signal} is greater than V_{thresh}.

Signal Difference Criterion: The difference in magnitude levels between the "correctly steered" beamformed signal and an "incorrectly steered" beamformed signal that is obtained by averaging the signal of each microphone with zero delay must exceed a threshold. An average

signal s_{avg} is computed by summing the signals received by each of the microphones without introducing any delay. The energy of this signal (E_{avg}) is compared with the energy associated with the beamformed signal (E_{beam}):

$$s_{dif} = \begin{cases} \dfrac{E_{beam} - E_{avg}}{E_{avg}} & if\ E_{beam} > E_{avg} \\ 0 & otherwise \end{cases}$$

The presence of a sound source is confirmed only when both the signal variance and signal difference exceed specific thresholds, (V_{thresh} and Dif_{thresh} respectively).

$$s_{dif} \geq Dif_{thresh} \wedge V_{signal} \geq V_{thresh}$$

3.2 Calibration

Beamforming with a far field acoustical model requires that the position of each microphone $\{\vec{x}_i\}$ relative to the array origin must be known. Rather than using a global coordinate system, the coordinate system defined by the Paracamera system is used. This eliminates the need to calibrate to some arbitrary global reference frame, however the audio and video systems must to be calibrated with respect to each other. The vertical distance between the plane of the lower three microphones and the optical centre of the Paracamera system is measured manually. The direction to each of these three microphones is easily measured as they appear in the view of the Paracamera. Under the assumption that the plane defined by these three microphones is perpendicular to the Paracamera optical axis permits the microphones' position to be easily estimated. The fourth microphone is mounted in line with the optical axis of the Paracamera system and its position is measured directly.

4. System Performance

The video system is responsible for detecting visible skin regions within the visual view of the Paracamera; grouping together the regions corresponding to each person; finding the region of each group that corresponds to the face; and finally, determining the direction to each potential face. Given this information, the audio system is focused to the direction of each potential face in order to detect sound (speech) emanating from the person when they are speaking. Either the audio and video systems can be used alone to locate potential speakers in a teleconference setting, however, by combining the two systems, each one is able to complement the other and therefore overcome some of the limitations inherent in each system when used alone. This provides a much more robust system of automatic speaker detection. This section describes two experiments that illustrate the effectiveness of the combined system to detect and focus on potential speakers in a teleconferencing session.

The first experiment is a demonstration of the combined system to detect multiple speakers in a typical group meeting scenario while the second experiment quantifies the accuracy of the system's ability to locate and focus on the face of a participant.

4.1 Experimental Room Set-Up

Both experiments were conducted in a meeting room at York University. The room is 3.7m long, 5.8m wide, with a height of 3.0m and is used for group meetings. As illustrated in Figure 6, one wall of the room contains several windows facing a hallway, while the other wall contains two windows facing outdoors. The windows facing outdoors are covered with black curtains. There is a counter running the length of one wall approximately 0.8m in height, a white-board (with a wooden cover that was closed during the duration of the experiments) mounted on one wall and a corkboard mounted on the opposite wall. The room also contains a black filing cabinet and a

large table with several chars in the middle of the room. Furthermore, the room contains standard flooring tiles, concrete ceiling and normal fluorescent lighting. No effort was made to limit audio reverberation or modify the lighting conditions in any manner. This room is a good representation of the environment of a typical teleconferencing room.

Both the audio and video systems rely on several pre-defined (and dynamically computed) threshold values. Tables I and II summarize these threshold values for both the audio and video systems. Both experiments described in this section were conducted with these threshold values.

4.2 Experiment One – Demonstration

The purpose of this experiment is to demonstrate the ability of the Eyes 'n Ears sensor to detect and focus on the participants who are speaking in a typical multiple person group meeting setting. As illustrated in Figure 7, four male subjects (where the i'th subject is denoted by s_i, for i=1…4), were seated around the sensor in the meeting room described above. Subjects were instructed to perform several tasks (the actual tasks are listed in Table III) while the system was operating. The tasks involved different scenarios of the participants speaking in a meeting. For example, two of the scenarios involved having all subjects remain silent for the duration of the test, while in another scenario, two of the subjects would speak at the same time while the others remained silent. The subjects that did speak were free to choose their own words and phrases for the duration of the experiment and spoke in a normal voice (e.g. they did not speak purposely louder than they normally do). In addition, the subjects were free to change their pose as they wished (e.g. raise their hands, move their head etc.). In each scenario, the system was started, the appropriate subjects began speaking, and 15 seconds later, the system would locate the speakers.

In addition to the participants, a radio was also placed in the room. The radio was tuned to an all news radio station (CFTR 680 News, Toronto, Ontario, Canada) and several different announcers, both male and female, were heard. The volume of the radio was set to a moderate level (6 out of the 10 possible volume settings) and was very similar, in loudness, to the level of the subject's speech. The radio loudspeaker was placed on the counter close to the sensor (and close to subject s_1) and was meant to simulate audio distracters (noise) that might be present in a regular meeting.

The video system correctly detected all participants in the six scenarios. Furthermore, given the real world direction to each detected face (relative to the Paracamera coordinate system), the audio system was capable of localizing participants that were actually speaking in each of the scenarios. It did incorrectly determine that one participant was speaking (s_1 in scenario one and scenario six), when in fact this participant was not speaking. However, subject s_1 was close to the loudspeaker and audio from the radio was associated with the nearby silent participant. In addition to the correctly classified faces, the video system incorrectly classified several non-face skin regions as faces, however, in each of these cases, the audio system correctly determined there was no speech emanating from the direction of the incorrectly detected face.

4.3 Experiment Two – Accuracy

This experiment investigates the accuracy with which the Eyes 'n Ears sensor locates a speaker in the environment. In order to quantify the results of the experiments, rather than using a human subject, a test “dummy” speaker was used. Using a test dummy ensured audio consistency over the entire duration of the experiment, and permitted much longer experimental trials to be conducted. After being placed in the appropriate position, the face of the test speaker did not

move in any manner ensuring its height above the table and its pose relative to the Paracamera remained static, allowing quantitative measurements to be made.

The test face used in this experiment was a color image of a face. The picture was mounted on an L-shaped wooden stand and the region of the image corresponding to the mouth on the picture was cut out. A small audio speaker was mounted directly behind it. The speaker was connected to a radio that was tuned to an all news radio station (CFTR 680 News, as in experiment one) for the entire duration of the experiment. This allowed the output of the speaker to emanate from the opening in the mouth simulating a person talking. The volume of the radio was set to 6 (out 10 possible volume levels, where 10 is the loudest setting).

A 10cm x 10cm grid was laid out on the table shown in Figure 8. The Eyes 'n Ears sensor was then centered on the table. The test face was then placed at specific grid locations and face localization was performed at 18 discrete locations. For each location, the image coordinate of the face in the Paracamera was obtained by identifying the nose of the test face in the video image. For each test position, the test face was then located three times using the audio and video detection algorithm and these values were averaged.

Figure 9(a) shows the actual (measured) direction to each face position (denoted by $\vec{\beta}_{actual}$ vs. the computed direction to the face (denoted by $\vec{\beta}_{computed}$) in the test (provided that the face was detected). Figure 9(b) plots the difference ε between the actual and computed directions for each face location ($\varepsilon = 1 - \vec{\beta}_{actual} \cdot \vec{\beta}_{computed}$).

The video system correctly detected the skin region present in each of the (18*3=54) tests (100%). A total of 9 false positive potential faces were identified by the vision system. The audio system confirmed 45 of the 54 true faces (83% of the true faces were correctly identified by both

the audio and video systems), and 1 false positive face was incorrectly confirmed by the audio system. In Figure 9, only those faces that were confirmed by the audio system are plotted.

5. Summary and Future Work

This paper describes a lightweight and portable face detection system that utilizes audio and video cues. The system is intended primarily for use in teleconferencing applications in which the sensor (consisting of the combined audio and video components) is placed on a table and the participants are seated around it. An omnidirectional camera (Paracamera), allows the video component to capture dynamic views of all participants from a single viewpoint, eliminating the need for a camera operator or having each speaker move within the camera's view. Using a statistical color model the pixels of each incoming Paracamera image are classified as either skin or non-skin. Skin classified pixels are grouped into labeled regions. Skin labeled regions which are spatially close are further grouped into clusters. Assuming there is an appropriate amount of space between the people in view, each cluster corresponds to one particular person. The region in each cluster furthest from the center of the Paracamera image is chosen as the face and an estimate of its direction in the real world, relative to the Paracamera coordinate system, is made and provided to the audio system. Beamforming and sound detection techniques with a small, compact fixed microphone array permits the audio system to be steered in the direction of potential faces, and then to focus on the speech of each participant. The detection of sound allows the audio system to confirm and therefore validate the presence of speakers in the directions determined by the video system.

Experiments performed in a normal meeting room environment indicate that by working together the audio and video system are capable of overcoming some of the limitations inherent in each component. Various factors may negatively affect each component, but these factors are

usually specific to either the audio or video system. For example, a reverberant environment may result in the incorrect localization of a sound source, but will not affect the video system. Similarly, the color of objects in the environment has no bearing on the audio system whereas it may negatively affect the video system and lead to the incorrect classification of non-skin regions as skin (e.g. certain yellow objects, such as a cardboard box or a standard corkboard may be incorrectly classified as skin). Furthermore, although techniques exist to allow for the localization of sound sources using the array solely, their performance quickly degrades in a reverberant environment and with multiple sources (Wilson et al., 2001). Furthermore, such techniques can be computationally expensive given the potentially large auditory space, however, by locating the direction to potential faces within the Paracamera's view, the video system essentially reduces the workspace of the audio system from potentially many thousands of directions to only a few, making the audio system's task tractable.

Various improvements could be made to the Eyes 'n Ears sensor. Faster computers, and especially more sensitive microphones would enhance system speed and performance. One specific advantage of more powerful computers would be the ability to probe, via audio beamforming, every detected large skin region as an audio source, rather than only probing the skin region in each cluster most distant from the centre of the video image.

Future extensions to the sensor include incorporating the system within an audio-video tracking system to permit speakers to be attended to and then tracked as they participate in teleconferencing applications, and to apply the system to distance learning applications. Specifically, to use the sensor as a device to enable a remote instructor to interact with his or her remote class.

6. Acknowledgements

The financial support of NSERC (Natural Sciences and Engineering Research Council of Canada), CRESTech (Centre for Research in Earth and Space Technology), IBM CAS and CITO (Communications and Information Technology Ontario) is gratefully acknowledged.

7. References

Adler, J. (1996). Virtual Audio: Three-Dimensional Audio in Virtual Environments. Technical Report T96-03, Swedish Institute of Computer Science.

Baker, S. and Nayar, S. (1999). A Theory of Single Viewpoint Catadioptric Image Formation. *Int. J. Comp. Vis.*, 35: 1-22.

Basu, A. and Southwell, D. (1995). Omni-Directional Sensors for Pipe Inspection. *IEEE Trans. Syst. Man Cybern*., 25: 3107-3112.

Boult, T. (2000). Dove: Dolphin omnidirectional video equipment. Proc. IASTED Int. Conf. On Robotics and Automation. Honolulu, Hawaii.

Boult, T., Michaels, R., Gao, P., Lewis, C., Yin, W. and Eckmann, M. (2001). Into the Woods: Visual Surveillance of Non-Cooperative and Camouflaged Targets in Complex Outdoor Settings. *Proc .of the IEEE*. Oct. 2001.

Brandstein, M., Adcock, M. and Silverman, H. (1995). A Practical Time-Delay Estimator for Localizing Sound Sources with a Microphone Array. *Computer Speech and Language,* 9: 153-169.

Chai, D. and Ngan, K. (1999). Face Segmentation Using Skin-Color Map in Videophone Applications. *IEEE Trans. Circuits Syst. Video Technol.,* 9: 551-564.

Compernolle, D., and Gerven, S. (1994). Beamforming with Microphone Arrays. Technical Report MI2-SPCH-94-6, ESAT, K. U. Luven, Belgium.

Crowley, J. and Berard, F. (1997). Multi-Modal Tracking of Faces for Video Communication. Proc. IEEE CVPR., pp. 640-647, Puerto Rico.

Davis, A. (1999). *Integrated Collaboration: Driving Business Efficiency into the Next Millennium*. Forward Concepts report.

De Mori, R. (1998). *Spoken Dialogues with Computers*, Academic Press Limited, London, UK.

Fiala, M. and A. Basu, A. (2002). Robot Navigation Using Panoramic Landmark Tracking. Proc. VI 2002. Calgary, Alberta, Canada.

Forsyth, D. (1990). A Novel Algorithm for Color Constancy. *Int. J. Comp. Vis.*, 5: 5-36.

Freed, L. (2000). Microsoft NetMeeting 3.0. *PC Magazine*.

Griebel, S. and Brandstein, M. (2001). Microphone Array Source Localization Using Realizable Delay Vectors. IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics. New Paltz, NY., USA. 71-74

Guentchev, K. (1997). Learning Based Three Dimensional Sound Localization Using a Compact Non-Coplanar Array of Microphones. M.Sc. Thesis, Department of Computer Science, Michigan State University, MI.

Gutchess, D., Jain, A. and Cheng, S. (2000). Automatic Surveillance Using Omni-Directional and Active Cameras. Proc. Asian Conf. Comp. Vis. Taipei.

Herpers, R., Derpanis, K., Topalovic, D. and Tsotsos, J. (1999). Detection and Tracking of Faces in Real environments. Proc. Int. Workshop on Recognition Analysis and Tracking of Faces in Real-Time Systems, 96-104, Korfu, Greece.

Hioki, W. (1990). *Telecommunications*. Prentice Hall. Englewood Cliffs, NJ.

Howell, A. and Buxton, H. (2000) Face Detection and Attentional Frames for Visually Mediated Interaction. Proc. Human Motion 2000. Austin, TX., USA. 143-148.

Huang, K. and Trivedi, S. (2001). NOVA: Networked Omnivision Arrays for Intelligent Environment. Proc SPIE Conf. on Applications and Science of Neural Networks, Fuzzy Systems and Evolutionary Computation IV, 4479-27.

Johnson, D. and Dudgeon, D. (1993). *Array Signal Processing: Concepts and Techniques*. Prentice Hall.

Jones, M. and Rehg, J. (1998). Statistical Color Models with Applications to Skin Detection. Technical Report CRL 98/11, Compaq Computer Corp. Cambridge, MA.

Lu, W. and Y. Tan. (2001). A Color Histogram based People Tracking System. Proc. 2001 ISCAS. Sydney Australia. 2: 137-140.

Mehtre, B., Kankanhalli, M., Marasimhalu, A. and Man, G. (1995). Color Matching for Image Retrieval. *Pattern Recogn. Lett*., 16: 325-331.

Peri, V. and Nayar, S. (1997). Generation of Perspective and Panoramic Video from Omnidirectional Video. Proc. DARPA Image Understanding Workshop, 243-245.

Phung, S., Bouzerdoum, A., and Chai, D. A novel skin color model in YCBCR color space and its application to human face detection. Proc. Int. Conf. Img. Proc. 2002. 289-292.

Rabinkin, D. (1994). Digital Hardware and Control for a Beam-Forming Microphone Array. M.Sc. Thesis, Department of Electrical Engineering, Rutgers University, New Brunswick, NJ.

Raja, Y., McKenna, J. and Gong, S. (1998). Segmentation and Tracking Using Color Mixture Models. Proc. Third Asian Conf. Comput. Vis.

Renomeron, J. (1997). *Spatially Selective Sound Capture for Teleconferencing Systems*, M.Sc. Thesis, Department of Electrical and Computer Engineering, Michigan State University, New Brunswick, NJ, USA.

Srisuk, S. and Kurutach, W. A New Robust Face Detection in Color Images. Proc. Fifth IEEE Int. Conf. Automatic Face and Gesture Recognition. Washington, DC., USA. 291-296

Stiefelhagen, R., Yang, J and Waibel, A. (2002). Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues. IEEE Trans. Neural Networks. 13: 928-937

Swain, M. and Ballard, D. (1991). Color Indexing. *Int. J. Comp. Vis.*, 7: 11-32.

Wilson, K. Checka, N, Demirdjian D and T. Darrell. (2001). Audio-Video Array Source Separation for Perceptual User Interfaces. Proc. PUI 2001. Orlando, FL., USA

Yasushi, Y. (1999). Omni-directional Sensing and its Applications. *IEEE Trans. Inf. and Syst*., E82-83.

Yong, R., Gupta, A. and Cadiz, J. (2001). Viewing Meetings Captured by an Omni-Directional Camera, *ACM Trans. Comput.-Hum. Interact*., 450-457.

Zheng, J. and Tsuji, S. (1992). Representation for Route Recognition by a Mobile Robot. Proc. IEEE Int. Conf. Comp. Vis., 55-76.

Zotkin, D., Duraiswami, R., Philomin, V. and Davis, L. (2000). Smart VideoConferencing. Proc. Int. Conf. Mult. Expo. 3107-3112, New York, NY, USA.

Tables

Parameter	**Value**
Sampling Frequency	44100Hz
Sampling Resolution	16 bits
Band Pass Filter Range	200-40000Hz
Filter Coefficients	128
Dif_{thresh}	0.10
V_{thresh}	Computed Dynamically

Table I. Audio system parameters.

Threshold	**Value**
Skin Pixel Classification (Δ)	0.91
Min Pixels per histogram bin (δ)	50
Size Filter	340 pixels

Table II. Video system thresholds.

Scenario	**Description**
1	Subject s_2 speaking only
2	No subjects speaking
3	Subject s_2 and s_4 speaking concurrently.
4	No subjects speaking
5	Subjects s_3 and s_4 speaking concurrently.
6	Subject s_4 speaking only.

Table III. Different scenarios in experiment one.

Figures

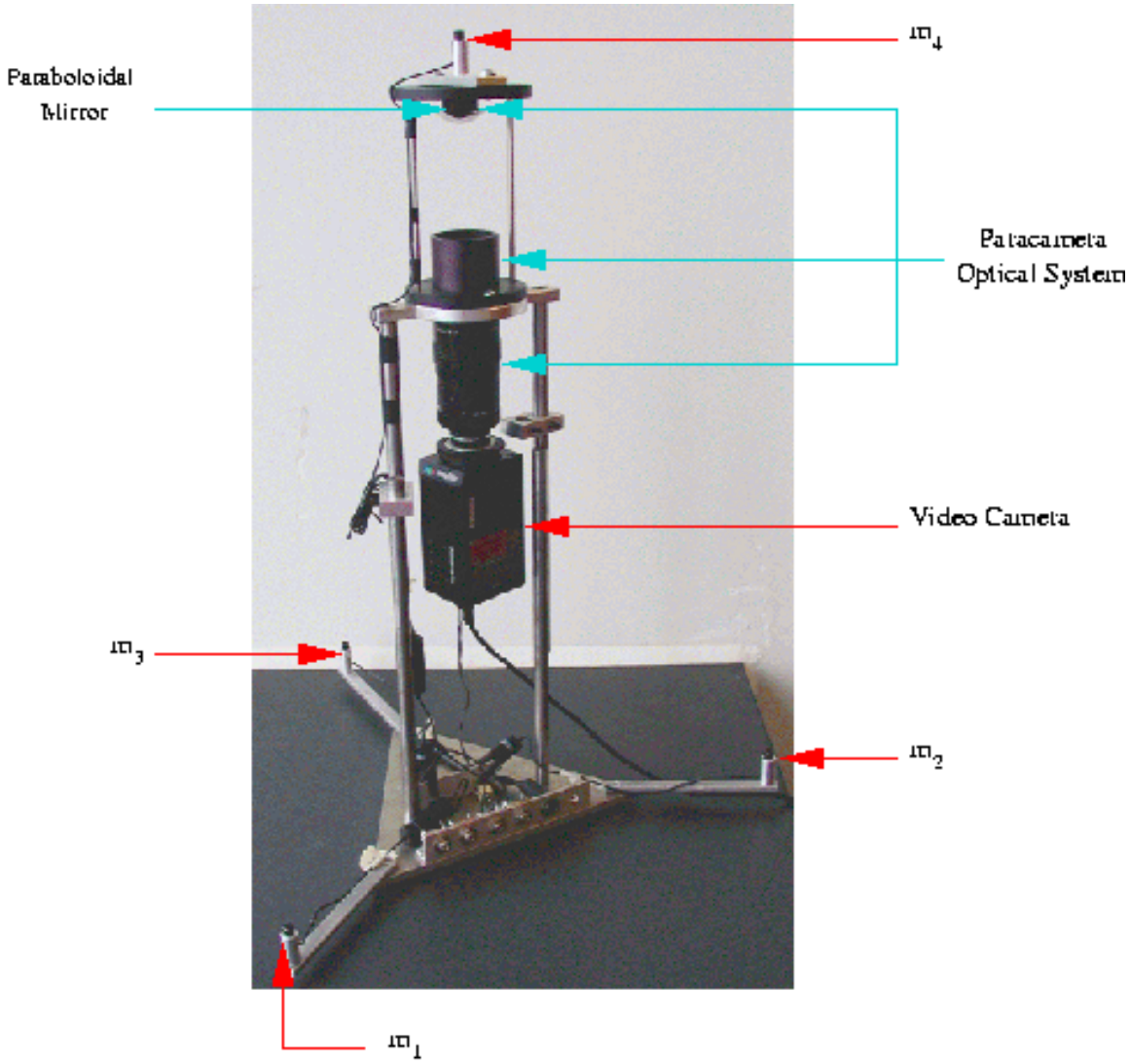
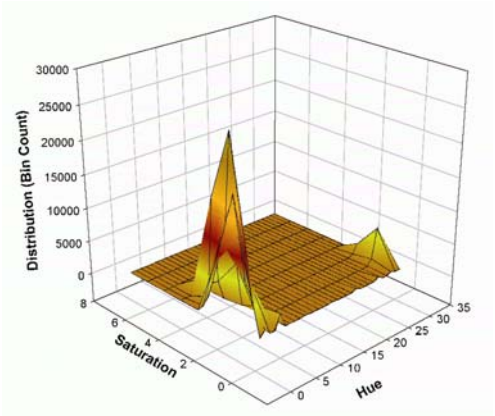


Figure 1. The Eyes 'n Ears Sensor.

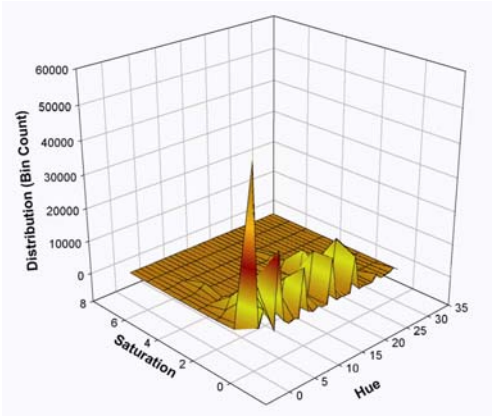
(a) Raw Paracamera image

(b) Unwrapped image

Figure 2. Paracamera Images. A panoramic view can be easily generated by un-warping the hemispherical view obtained by the Paracamera. Perspective Images of any size may then be extracted from the panoramic.

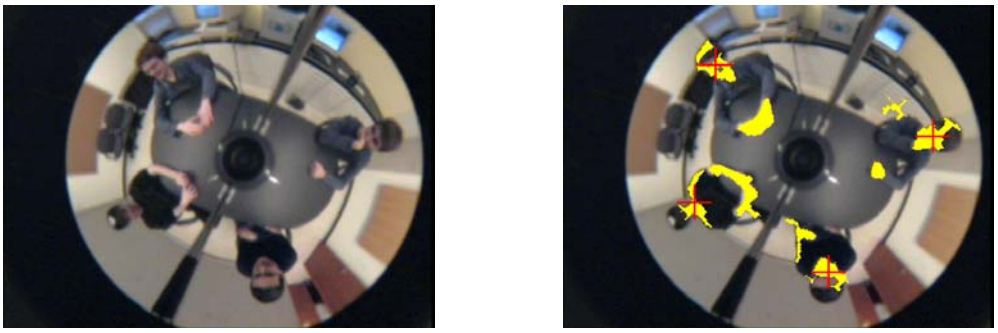


(a) Skin color histogram



(b) Non-skin color histogram

Figure 3. Hue-saturation histogram models for skin (a) and non-skin (b).



(a) Input image (b) Identified faces

Figure 4. Faces identified in the input omnidirectional image. Pixels identified as skin are colored yellow, and the centers of the identified faces are indicated with a red cross.

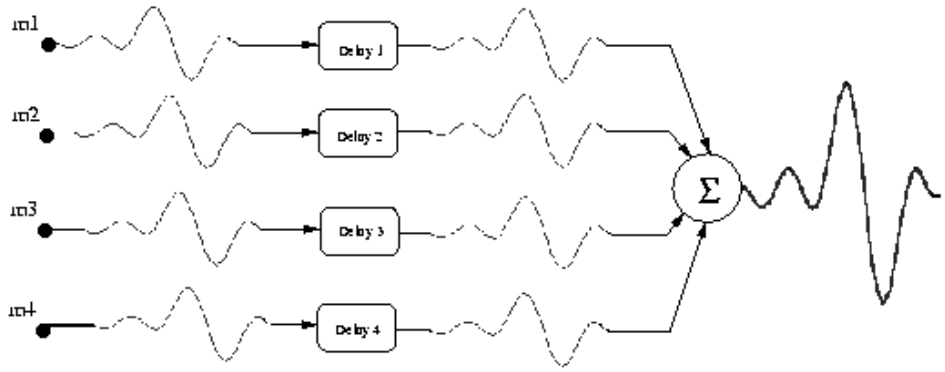
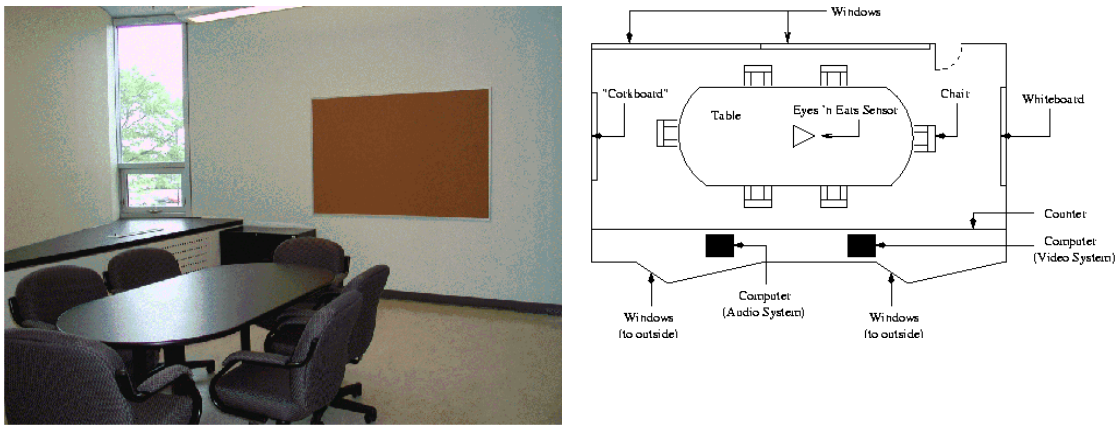
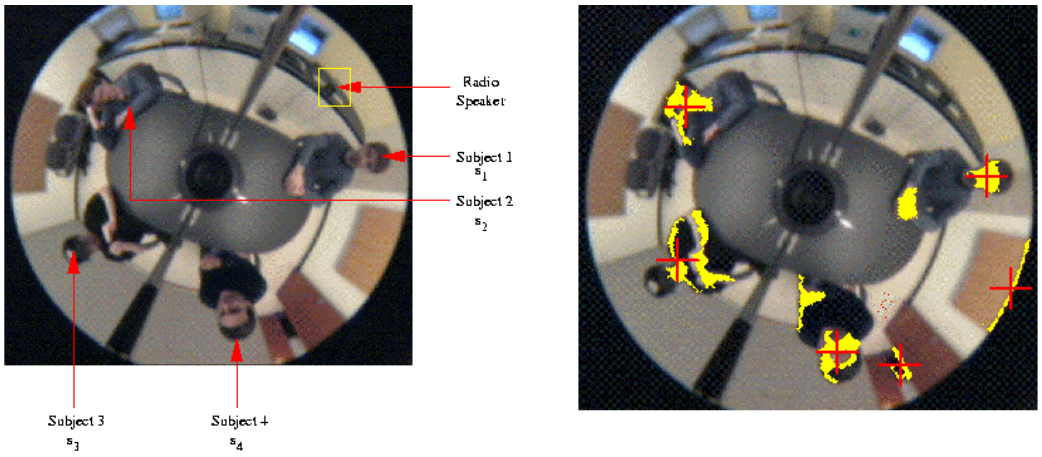


Figure 5. Forming the Beamformed Signal. Summing the appropriately delayed signal from each microphone will lead to a beamformed signal with maximized energy output.



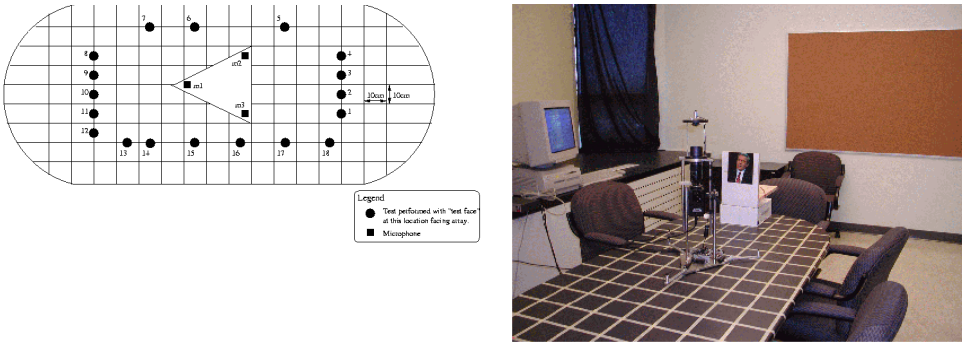
(a) Actual meeting room (b) Schematic diagram of the room

Figure 6. Experimental setup.



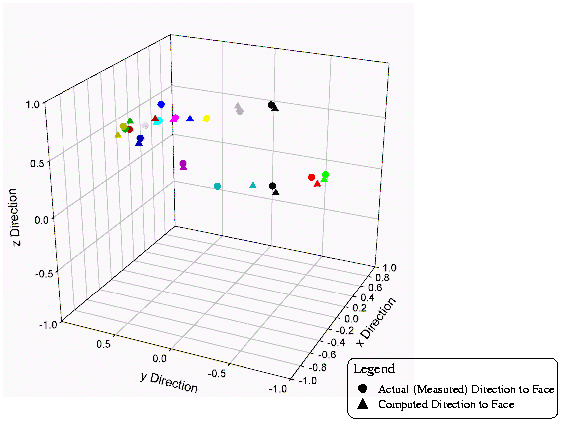
(a) Setup (b) Sample output

Figure 7. Experiment one set-up (a) and sample output (b). Four participants in a typical group meeting, seated around the Eyes 'n Ears sensor. The radio distracter is located counter clockwise of the participant on the right (s_1).

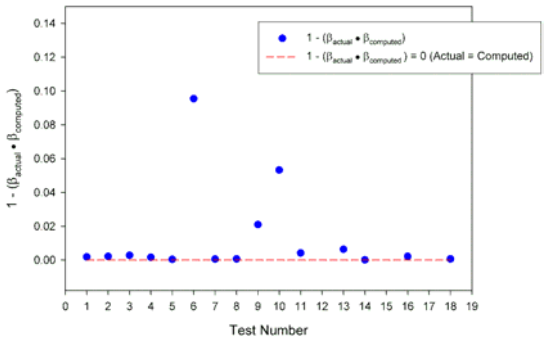


(a) Location of tests (b) Experimental setup

Figure 8. Location of test face during each of the trials in experiment two.



(a) Actual and measured directions



(b) Difference between actual and measured directions

Figure 9. Experimental results. (a) shows the actual and measured directions to the 18 face locations, while (b) shows the difference between the two directions. Each data point corresponds to three repetitions.

Figure Captions

Figure 1. The Eyes ‘n Ears Sensor.

(a) Raw Paracamera image (b) Unwrapped image
Figure 2. Paracamera Images. A panoramic view can be easily generated by un-warping the hemispherical view obtained by the Paracamera. Perspective Images of any size may then be extracted from the panoramic.

(a) Skin color histogram (b) Non-skin color histogram
Figure 3. Hue-saturation histogram models for skin (a) and non-skin (b).

(a) Input image (b) Identified faces
Figure 4. Faces identified in the input omnidirectional image. Pixels identified as skin are colored yellow, and the centers of the identified faces are indicated with a red cross.

Figure 5. Forming the Beamformed Signal. Summing the appropriately delayed signal from each microphone will lead to a beamformed signal with maximized energy output.

(a) Actual meeting room (b) Schematic diagram of the room
Figure 6. Experimental setup

(a) Setup (b) Sample output
Figure 7. Experiment one set-up (a) and sample output (b). Four participants in a typical group meeting, seated around the Eyes 'n Ears sensor. The radio distracter is located counter clockwise of the participant on the right (s_1).

(a) Location of tests (b) Experimental setup
Figure 8. Location of test face during each of the trials in experiment two.

(a) Actual and measured directions
(b) Difference between actual and measured directions
Figure 9. Experimental results. (a) shows the actual and measured directions to the 18 face locations, while (b) shows the difference between the two directions. Each data point corresponds to three repetitions.