Multimodal Talker Localization in Video Conferencing Systems

By

Charn Leung (David) Lo, B. Eng., M. Eng.

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Ottawa-Carleton Institute of Electrical and Computer Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa Ontario, K1S 5B6
Canada
July 5, 2005

Abstract

In a video conferencing environment, it is desirable to isolate the active talker. Traditionally, talker localization is performed acoustically using a beamforming microphone array or videographically using image processing techniques. Since these approaches rely only on the audio or the video data for performing the localization, they are often prone to errors. In this thesis, a new modular multimodal architecture is designed. Data from each localization modality are separated in the beginning, and localizations are performed using each data stream independently. In order to study the effectiveness of this modular multimodal architecture, this thesis combines audio, visual and infrared cues to locate talkers in the video conferencing environment. Special purpose acoustic, video and thermo localizers are developed to perform the localization. Individual results from the localizers are then combined using data fusion techniques to give the final estimation of the talker's location. Two common fusion methods, the summing voter and the Bayesian network, are studied in this thesis. The effectiveness of another two novel fusion methods, the talker occupancy grid assisted summing voter and the talker occupancy grid assisted Bayesian network, are also investigated. A unique algorithm that uses the correlation lags to detect acoustic reflections is also developed in the process of this thesis. Based on the results from experiments and computer simulations, the proposed multimodal localization method outperforms localization methods, in terms of accuracy and robustness, when compared with other single modal methods that rely only on audio, video, or infrared data.

Acknowledgements

I would like to express my gratitude to my thesis co-supervisors Professor Rafik A. Goubran and Professor Richard M. Dansereau for their guidance and advice. I will also like to thank Mitel Networks for their technical and financial support, and Nortel Networks for their financial support.

I am forever grateful of my wife Winnie for her continuous support, encouragement and understanding. I will like to give my heartfelt thanks to God for His unfailing love, His graces, and the chance to have this memorable time throughout the thesis process.

Contents

List of Figures

List of Table

List of Notations and Variables

α[k]	Averaged power profile.
$\overline{\alpha[k]}$	Root-Mean-Square normalized averaged power profile.
Δ	Time delay introduced by differences in arrival time.
Δ_m	Time delay of arrival of the mth microphone.
θ	Incident angle of the plane wave.
δ	Extra distance plane waves needed to travel between two microphones.
a	Planer distances between two microphones on linear microphone array.
Cr	r Chromas
Cb	b Chromas
$D_{(x,y)}$	Pixel value of difference frame of pixel(x,y).
$D_n[k]$	Number of detections in sector n at time k.
e	Evidence
$F(i)_{(x,y)}$	Video frame at time index i.
$G_{(m,n)}$	Talker's occupancy estimates for modality m and sector n.
$Hist(j)_x$	Pixel histogram in the x direction.
$Hist(i)_y$	Pixel histogram in the y direction.
J_s	Total number of detected active sectors.
K_n	Summing voter fusion output for sector n.
$Lag_{(ab,n)}$	Maximum correlation lag of $R_{(ab,n)}$.
N	Total number of microphone in a microphone array.
$P_{(m,n)}$	Correctness probabilities for modality m and sector n.

$\overline{P}_i[k]$	Root-Mean-Square normalized signal power of sector *i*.
P_i[k]	Beamformer output signal power for sector *i*.
$\overline{P_i[k]}$	Root-Mean-Square normalized signal power.
$pixel_{(x,y)}$	Pixel value of pixel (x,y) in YCrCb space.
RMS	Root-Mean-Square.
$R_{(ab,n)}(i)$	Cross-correlation between $\overline{P}_a[k]$ and $\overline{P}_b[k]$.
s	Speed of sound at 20°C and 50% relative humidity, and have the value of 343.99 m/s.
TDOA	Time Delay of Arrival.
Q	Detection quality.
w_m	Beamformer amplitude weight.
Y	Luma
$y_m(t)$	Waveform measured by the m^{th} microphone.
$z(t)$	Beamformer output signal.

Chapter 1 – Introduction

Video conferencing allows two or more people at different locations to communicate with sight and sound. All video conferencing systems have video and audio capabilities. Most systems consist of one or more video cameras which capture the images of the conferencing environment, and one or more microphones which capture the voice activities. Most people have some experience with the low-end systems, like CU-SeeMe and Microsoft Netmeeting. They usually consist of a web camera and a close talking microphone with the traffic going through the public network or the Internet. Most high-end systems, like the systems offered by Mitel Networks, Polycom Incorporated and Andrea Electronics, use arrays of microphones and pan/tilt/zoom camera(s) with the traffic going through a secured network. When compared with more traditional teleconferencing, video conferencing offers the experience of "almost being there". The additional video provides the users with a more natural experience and can better understand the other participants by seeing their unspoken expressions.

Lately, video conferencing is becoming popular, especially among business users. It provides a low cost alternative to traveling. However, the systems that are currently offered commercially leave a lot to be desired.

1.1 Purpose

The purpose of this thesis is to locate participants in video conferencing by combining localization information from different localizers like audio microphone array, video camera and infrared camera. Localization is performed using information acquired from each localizer separately and then combined to form the final estimation of the talker's location.

1.2 Problem Statement

Traditionally, talker localization in video conferencing is done acoustically using a beamforming microphone array [RAY97], [RAY00], [RAY03]. As the microphone array captures the talker's voice and locates him simultaneously, his location is also used to direct a camera capturing his image. An alternative approach to talker localization is to rely on the video information alone to perform localization. Unfortunately, audio or video localizations performed independently are prone to errors. Audio localization is susceptible to acoustic reflections [OMO96], whereas video localization is susceptible to changes in lighting conditions [HSU02], and complex backgrounds. This thesis explores the use of multimodal localization which combines two or more sources giving a more robust localization [STR01].

1.3 Approach of This Thesis

This thesis approaches the multimodal talker localization by separating each localization modality as a separate component. Data streams from the sensing devices are decoupled

early in the beginning. Localizations are then performed on each data stream and individual localization results are then combined using sensor fusion techniques. The approach has the advantage of giving systems that are flexible and can easily be expanded. Each localization modality is developed and tested separately before they are put together as a complete system.

1.4 Thesis Outline

This thesis has twelve chapters in total. Chapter 1 and 2 provide the introduction and the background information. Chapter 3 describes the setup and environments of the experiments and simulations done in the thesis. Chapter 4 describes the architecture of the multimodal talker localization system and its application in video conferencing. Chapter 5 describes how talker localization is done using single modal localization methods. Chapter 6 describes how the multimodal localization architecture can be applied in joint audio-video talker localization. Chapter 7 describes how the talk's occupancy information can be used as weight to improve the overall localization performance. Chapter 8 studies the effect of adding an additional Infrared localizer into the overall multimodal localization architecture. Chapter 9 provides the conclusions and how this thesis can be expanded in the future.

1.5 Summary of Contributions

This section presents a summary of the contributions in this thesis:

1) Proposed a modular architecture for performing multimodal talker localization. This modular design allows flexibilities in the type of sensing device, number of sensing devices, processing method, and fusion method used. This contribution was published in D. Lo, R. A. Goubran, R. M. Dansereau, "Multimodal talker localization in video conferencing environment," in *Proceedings of IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications*, 2004, pp. 195-200.

2) Proposed an algorithm which uses correlation lag to differentiate between an acoustic signal originating from a new talker, and signals caused by acoustic reflection from the original talker. When compared with existing algorithms [BRA99], the proposed algorithm has lower computational requirement because it uses the averaged power of the beamformed signals instead of the raw microphone signals in its calculation. Furthermore, the algorithm can be used with any beamforming technique. This contribution is to be published in D. Lo, R. A. Goubran, and R. M. Dansereau, "Acoustic reflections detection for microphone array applications," accepted to *the 22th IEEE Instrumentation and Measurement Technology Conference, Canada*, 2005. This algorithm was also filed as a patent [LOD03B].

3) Co-invented an algorithm that allows a video conferencing system to find the location of its microphone array. Using this algorithm, microphone arrays and cameras of the conferencing system can be placed arbitrarily in a room or can be moved to a new location during conferencing. The algorithm was filed as a patent [GOU03].

4) Proposed a method, which uses the talker's occupancy grid information to assist a summing voter during the data fusion process, to improve the multimodal talker localization accuracy. The occupancy grid is estimated based on physical characteristics of the sensing devices. These information are then used as weights to bias the summing voter away from unreliable data. This contribution was published in D. Lo, R. A. Goubran, R. M. Dansereau, G. Thompson, D. Schulz, "Robust Joint Audio-Video Localization in Video Conferencing Using Reliability Information," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 4, pp. 1132-1139, August 2004.

5) Proposed a method to improve the robustness of joint audio-video talker localization using the Bayesian network. This method uses the reliability information to stop failed devices from contributing in the fusion process. A simple algorithm was also developed for detecting failures in audio and video devices. When a Bayesian network is used to perform joint audio-video talker localization, the persistent erroneous data stream from a failed device can negatively affect the Bayesian network, resulting in poor localization accuracy. By stopping the failed devices from contributing, better overall localization accuracy is achieved. These contributions were published in D. Lo, R. A. Goubran, R. M. Dansereau, "Robust joint audio-video localization in video conferencing using reliability information II: Bayesian network fusion," accepted to *IEEE Transactions on Instrumentation and Measurement*, August 2005.

6) Developed an algorithm that gauges the correctness of the microphone array's current output using the averaged beam pattern of the microphone array. This algorithm was published as part of [LOD03A].
7) Developed an automatic area-of-interest (AOI) identifying algorithm which isolates regions in a processed video frame in which a talker might be present.
8) Proposed a method to perform multimodal talker localization with Infrared imaging.

In addition to the above contributions, the following tasks were also performed:

1) Developed and constructed a simple motion detection computer routine.
2) Developed and constructed a skin-color detection computer routine.
3) Developed and constructed a simple thermo-graphical detection computer routine.
4) Developed and constructed an automatic white balancing computer routine.
5) Performed joint audio-video talker localization using a simple summing voter as the fusion engine.
6) Performed joint audio-video talker localization using a Bayesian network as the fusion engine.
7) Performed joint audio-video-infrared talker localization using a simple summing voter as the fusion engine.
8) Performed joint audio-video-infrared talker localization using a Bayesian network as the fusion engine.

Chapter 2 Background Review

In a video conferencing environment, it is desirable to isolate the active talker [LOD03A], [LOD04], [WAN98]. Often, the isolation is done by means of audio and/or video localization [LOD03A], [LOD04], [WAN98], [MES02], [WAN00], [ZOT00]. Most commercial systems use a beamforming microphone array to locate the active talker acoustically. Once the talker's location is found, the microphone array sends the talker's direction to the camera. The video camera is then pointed to the talker's direction to capture his/her image. Although less popular, systems that rely on the video to perform localization are becoming more common as video equipment gets cheaper and computers become more powerful. Unfortunately, audio or video localizations alone are prone to errors. For example, audio localization is very susceptible to acoustic reflections [OMO96]; video localization is susceptible to changes in lighting conditions [HSU02] and complex backgrounds. Multimodal localization takes advantage of the complementing nature of multiple sources giving a more robust localization [STR01]. Other researchers have been exploring the use of the multimodal approach [WAN98], [WAN00], [MES02], [ZOT00], [STR01], [WAN99], [WUH02], [TOY00]. This thesis uses the multimodal approach to talker localization.

2.1 Talker Localization Using Audio Method

Traditionally, talker localization in video conferencing was done acoustically using a beamforming microphone array [BRA01]. A microphone array is a collection of two or more microphones distributed in space working collectively as a single device. With a single microphone, the direction of an audio source cannot be determined [JOH93]. However, using two or more microphones, with the help of a beamforming technique, the spatial-temporal relationship can be used to recover directional information about the source [JOH93]. Audio beamforming is a signal processing technique that is used to enhance the audio signal in the incoming direction and at the same time attenuate the signal in all other directions.

2.1.1 Delay-and-Sum Beamforming

Delay-and-Sum is the most commonly used beamforming algorithm in commercially available products. The basic idea behind the delay-and-sum beamforming algorithm is fairly straightforward. Assuming the incoming acoustic signal is a plane wave, the microphone closest to the source will pickup the audio signal first, and then the second closest one and so on. The amount of delay between signals from each microphone is directly proportional to their distance from the source (see Figure 2-1).

In most cases, the geometrical configuration of the microphone array is known and, hence, the amount of delay between signals from each microphone can be calculated. When the correct amounts of delays are applied, the signal in the source direction will be enhanced through constructive interference and the noise in all other directions will be

minimized through destructive interferences. Figure 2-1 shows a linear microphone array with the incident sound wave modeled as plane waves originating from a far field audio source.

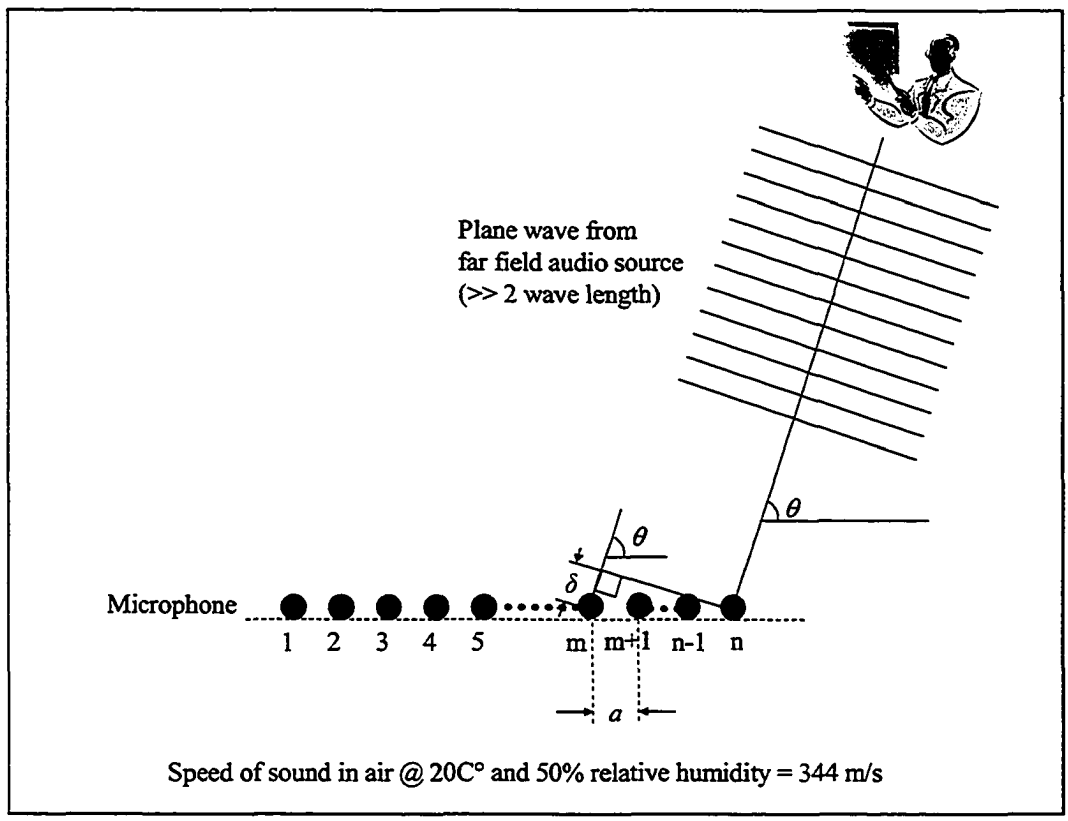

Figure 2-1. Linear microphone array with incident plane wave from a far field source.

For a linear array, assuming the incident angle of the plane wave is θ and using microphone n as reference, the differential distance between the source and microphones m and n can be calculated using

$$\delta = (n-m)\cdot a\cdot\cos(\theta) \qquad (2\text{-}1)$$

where δ is the extra distance the plane waves need to travel to reach microphone m, a is the distance between microphone m and m+1 on the microphone array, θ is the incident angle of the plane wave. Given the speed of sound s is equal to 343.99 m/s at 20°C and

50% relative humidity, the time delay Δ between the audio signals at microphone n and m, also know as the Time Delay of Arrival (TDOA), can be calculated using

$$\Delta = \frac{\delta}{s} \tag{2-2}$$

Now, if delay Δ is applied to the audio signal acquired at microphone m and then summed with the signal acquired at microphone n, signals coming in from the direction of the audio source (θ) is strengthened due to constructive interference and, at the same time, signals from all other directions are weakened due to destructive interferences. This method can be generalized and applied to the rest of the microphones in the array as shown below [JOH93].

$$z(t) = \sum_{m=0}^{N-1} w_m y_m (t - \Delta_m) \tag{2-3}$$

where $z(t)$ is the final beamformed signal, $y_m(t)$ is the waveform measured by the m^{th} microphone, N is the total number of microphones in the array, Δ_m is the TDOA of the m^{th} microphone, and w_m is an amplitude weight applied to the output of the m^{th} microphone. The amplitude weight w_m is also known as the array's shading or taper. It can be used to reduce the array's sensitivity in an undesirable direction (i.e., the sidelobes).

2.2 Talker Localization Using Video Methods

Besides using a beamforming microphone array, a talker can also be located using video methods. As the talker's image is captured using an imaging device like a camera, the talker's image in the scene can be separated from the background with image processing

techniques. Two of the techniques used in this thesis are motion detection and color detection.

2.2.1 Motion Detection

Motion is a useful cue for locating a talker [COL99], [FER01] in a video scene. A talker rarely stays perfectly stationary, especially when he speaks [WAN99]; therefore, motion detection is a useful method for locating a talker in video localization. The most commonly used method in detecting motion is frame subtraction [CUC03]. Motion causes changes in the video frames. Assuming the camera is stationary and the talker is the only moving object in the video scene, motion in the video scene can be isolated by subtracting two consecutive video frames pixel by pixel. In order to avoid detection error due to small changes in lighting condition, binary thresholding is often applied to the video images after frame subtraction is performed. For simplicity, frame subtraction is often done in grayscale. Figure 2-2 demonstrates the method.

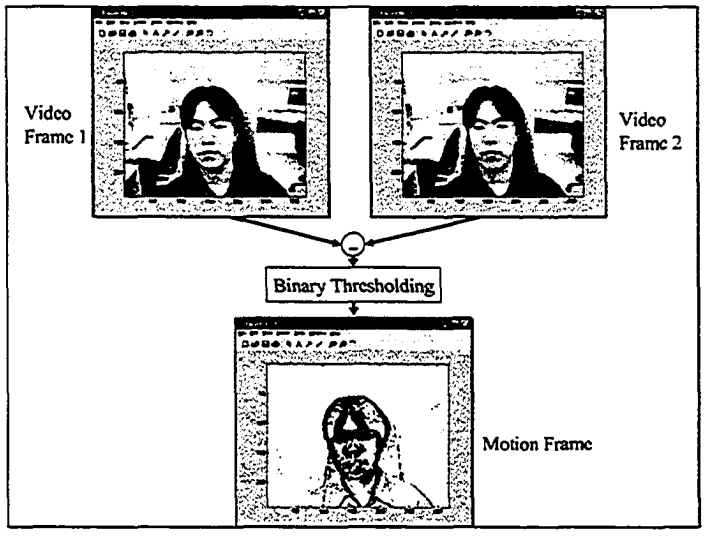

Figure 2-2. A motion detection example using frame subtraction technique.

2.2.2 Color Detection

Color detection works by isolating an object of known color from a video scene. When used in talker localization, color detection is used to isolate skin-color objects in the video scene. Although imaging devices often capture color images in 24-bit Red, Green, Blue (RGB) format, color spaces using RGB like the red-green space [BER01] are not the best for detecting faces [SAB98], [TER00]. Research has found that color analysis done in luma-chroma space [TER00], such as the YCrCb space [POY96], concentrates the skin-color pixels in a tight range [DEC00]. Therefore, a luma-chroma space is well suited for detecting skin color objects. For this reason, video frames acquired in 24-bit RGB format are first transformed into the CCIR601-4 YCrCb color space [POY96] using

$$\begin{bmatrix} Y \\ Cr \\ Cb \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \frac{1}{256} \begin{bmatrix} 65.738 & 129.057 & 25.064 \\ -37.945 & -74.494 & 112.439 \\ 112.439 & -94.154 & -18.285 \end{bmatrix} \begin{bmatrix} Red \\ Green \\ Blue \end{bmatrix} \tag{2-4}$$

where Y is the luma coefficient, Cr and Cb are the complement chromas, Red, Green and Blue are pixels in red, green, and blue color respectively. Except in low light conditions, the luma coefficient can be approximated as a constant for a given video scene. Human skin color pixels plotted in the Cr-Cb space done by Hsu *et al.* [HSU02], Figure 2-3, shows the pixels tightly grouped together as an oval patch.

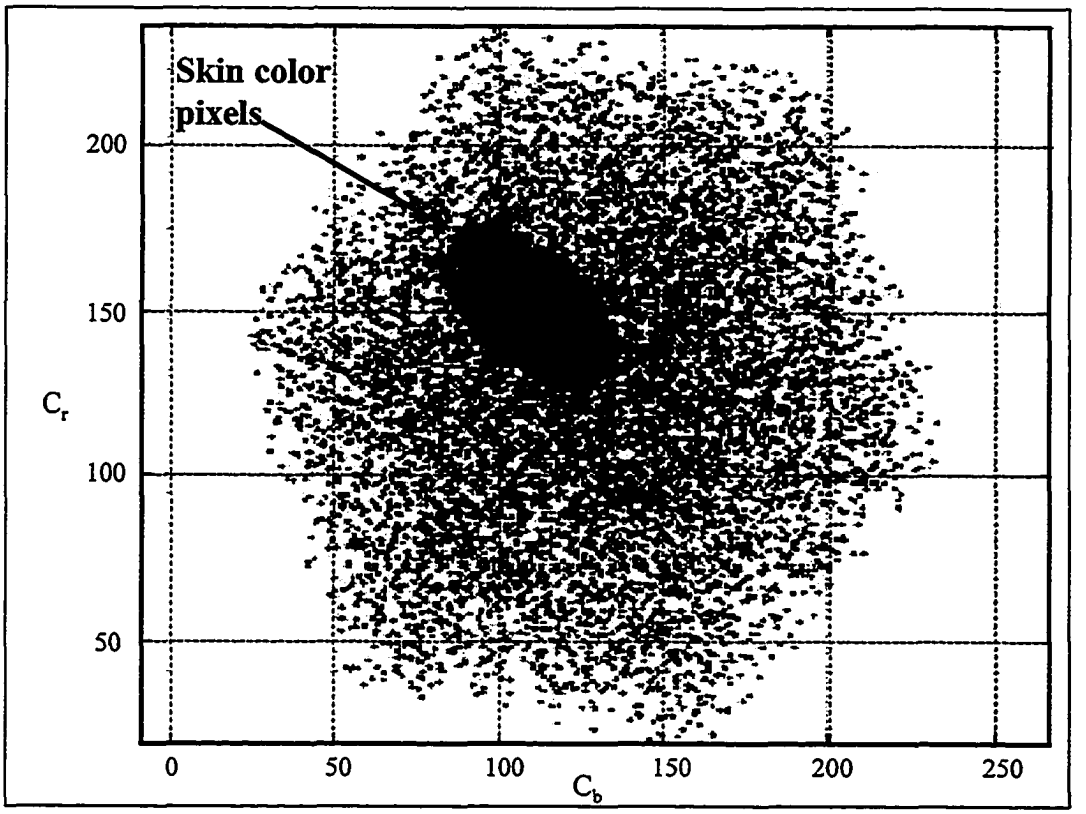

Figure 2-3. Human skin-color pixel plotted in Cr-Cb color space [HSU02].

2.3 Talker Localization Using Infrared Imaging Method

Infrared imaging detects body heat radiated by the talkers [HOL00]. Most infrared cameras map the measured temperatures to grayscale values and then output the image using false color [HOL00]. Currently, no reference can be found on using infrared imaging for the purpose of locating talker in video conferencing application. However, infrared imaging has been used successfully for face tracking [SEK02], navigation in robotics [PET96][SEU94][YEU94], and security surveillance [DAV97][HOL00].

Infrared imaging (IR) complements the video method very well. Video cameras rely on visible light. When the level of visible light is low, most video cameras fail to register any images causing the motion and skin-color localization methods to fail. Consequently, the system loses the ability to locate the talker visually. Low light situations are common in presentations were the lights are dimmed when computer and overhead projectors are being used. However, IR cameras respond to heat and are not affected by lighting conditions. The IR camera will still allow the system to locate the talker visually regardless of the lighting conditions. Since IR imaging maps temperatures to grayscale values, simple binary thresholding works very well as a detection method. The details of the detection method is described in Chapter 5.3

2.4 Talker Localization Using Multimodal Methods

Audio localization using beamforming generally works flawlessly when the talker speaks directly toward the microphone array and no other interfering sounds exist. However, when the talker directs his speech to another direction, say to talk to a colleague sitting on his right, the microphone array often fails to locate the talker correctly due to acoustic reflections [OMO96]. Video localization does not suffer from acoustic reflections, but does fail when the lighting conditions in the scene change drastically [HSU02], or when other people enter and leave the video scene in the background. Since using only audio or only video for localization is prone to failure, researchers are now exploring multimodal approaches by combing audio and video localization methods [LOD03A], [LOD04B],

[WAN98], [MES02], [WAN00], [HSU02], [WUH02], [TOY00], [BRA01]. Joint audio-video localization takes advantage of the complementary nature of the two methods, giving a more robust localization [WHU02]. For example, Wang *et al.* take the approach of cascading audio and video localization methods [WAN98], [WAN99], [WAN00]. Other researchers take the data fusion approach. They have explored the use of a modified Kalman filter [STR01], particle filters [ZOT00], as well as Dempster-Shafer [WUH02] and other statistical methods [TOY00]. While most of these approaches work well, they are primarily using only the audio and video methods, and they often require extensive statistical properties of the localizers [STR01].

Chapter 3 Experimental Setup and Simulation Environment

The results of this thesis were done using experiments and computer simulations. In this chapter, a brief explanation of the setup for the various experiments, and the experimental and simulation environments used in this thesis are given.

3.1 Experimental Environment

All experiments in this thesis are conducted in an anechoic chamber or in reverberant rooms.

3.1.1 Anechoic Chamber

The anechoic chamber is located in the Loeb Building, Carleton University. It measures 3 m x 5 m in size. The floor surface is provided by a suspended metal mesh platform. All wall surfaces are lined with acoustic foam wedges, and the chamber is built to block out all external sound, vibrations, and electro-magnetic interferences.

3.1.2 Reverberant Rooms

Two different reverberant rooms were used to conduct the experiments. The first room is a typical well controlled conference room measuring 3.8 m wide x 5.4 m long x 3 m high lined with drywall, vinyl floor tiles, suspended acoustic tiled ceiling, and uniform lighting. The second room is much larger and is very reverberant. All wall surfaces are made out of concrete, and there are multiple acoustic reflective surfaces like whiteboards. There are no soft surfaces in the room, and the lighting is less uniform. It measures 6.6 m wide x 9.7 m long x 4 m high. The conferencing environment in this room is relatively challenging, and is used as a "stress test".

3.2 Experimental Setup

3.2.1 Equipment for Audio Data Acquisition and Playback

In this thesis, audio localization is done using a beamforming microphone array. The microphone array used is the prototype of a commercially available circular array made by Mitel Networks Corporation (5310 IP Conferencing Unit). The array has a circular housing with six microphones embedded in it, and its diameter is 0.11 m. Since the proprietary voice activation detection algorithm used by Mitel Networks in their 5310 IP conferencing unit works in the 1 k – 1.3 kHz range, audio signals are band-pass filtered to 1k – 1.3 kHz, amplified, and then sampled at 8 kHz. Source localization was done using a Bittware Research System DSP board (BTPC-4062-2) equipped with a 40 MHz ADI DSP processor running a delay-and-sum beamforming algorithm [JOH93]. One or two reproducible audio sources are generated using a loudspeaker (Tannoy Saturn S8LR)

playing back prerecorded voices with the same loudness. Three voice recordings, two male voices and one female voice, were used. The loudspeakers are driven using an amplifier system (Rotel RC-850 pre-amplifier and RB-850 power amplifier). A 1.6 m wide x 6 m high whiteboard is used as an acoustically reflective surface. Acoustically absorbing surfaces is created by applying acoustic foam panels (RPG Diffusor System ProFoam) to the walls. Audio data are processed after the experiments with custom programs that I wrote running under Matlab (version 6.5).

3.2.2 Equipment for Video Data Acquisition

Video data is captured using a Canon VC-C4 camera and 24 bit USB frame grabbers. Two different grabbers are used in this thesis. One is made by Belkin model F5U208. It uses software codecs, and is capable of digitizing video at the rate of 15 frames per second at the resolution of 320 x 240 pixels using a 2.4 GHz Pentium 4 computer. The other one is made by Plextor model ConvertX PX-M402U. It uses a hardware codec, and is capable of digitizing video at the rate of 30 frames per second at the resolution of 720 x 480 pixels. Videos are captured using frame capturing software (FlyCap version 2.5.2). All unused frames are edited out using video editing software (VirtualDub version 1.5.10) to reduce wasted processing time later when these frames are analyzed for motion, skin-color, and thermo contents. Video analysis is done in Matlab (version 6.5) using custom programs that I wrote.

3.2.3 Equipment for Infrared Data Acquisition

Infrared data are acquired using an FLIR Systems ThermoVision 160C infrared camera. The NTSC output from the infrared camera was captured using a Belkin F5U208 USB frame grabber at a rate of 15 frames per second using 320 x 240 pixels resolution. The camera is capable of detecting temperatures from -40°C to 120°C with 0.08°C of resolution. During operation, the camera uses an internal reference to perform periodic calibration to maintain its measurement accuracy. Similar to the video data acquisition, all unused infrared image frames are edited out using video editing software (VirtualDub version 1.5.10) and video analysis is done in Matlab (version 6.5) using custom programs that I wrote.

3.3 Simulation Environment

Computer simulations were also used in this thesis to study the effectiveness of the proposed reflection detection method under different parameters and various conditions. All simulations are done using Matlab (version 6.5). In order to keep the simulations simple, straightforward environmental acoustics are used. The size of the room is assumed to be large enough that the walls will not cause any reflections. An acoustically reflective surface is assumed to cause only the primary reflection (i.e., first order reflection). Also, all audio sources are assumed to be far field.

3.4 Data Fusion Software

Two software packages are used to perform the Bayesian network data fusion. For single iteration and fast prototyping, the Microsoft Research's Belief network Authoring and Evaluation Tool Box (MSBNx version 1.4.2) is used. It has the advantage of having a easy to use graphical user interface and minimum setup time. However, it lacks the ability to perform batch processing. Therefore, each complete evaluation requires multiple runs to cover all the permutations. Batch processing is done using the Bayes Net Toolbox (BNT) for Matlab (version 5, GNU Library GPL). BNT has the advantage of being very flexible but it requires considerable amount of programming efforts.

3.5 Types of Data and File Formats

Two types of data, audio and video, were stored during experiments. Audio data was acquired as the microphone array outputs, and detected active sector. The delay-and-sum beamformer algorithm used by the microphone array is capable of detecting audio signals originating from 12 different directions (sectors). The beamforming algorithms produced 12 beamformed signals and the outputs of the array are the power of the beamformed signals. An audio sector '*i*' is considered active if the power signal exceeds a predetermined threshold. The audio data file is binary in format. Each line of data contains the 12 power signals follow by the currently detected active sector.

Video data are captured by the USB frame grabbers as 24-bit RGB video frames. The Belkin grabber uses the Microsoft MPEG-4 version 2 software coder to compress the video frames in real-time, and then store as an audio video interleave (AVI) file. Video

data captured by the Plextor grabber is compressed using its internal hardware codec in DivX format which is a variant of MPEG-4. For compatibility purposes, files coded using DivX are re-coded using the Microsoft MPEG-4 version 2 software codec and then stored as AVI files.

Chapter 4 Architecture of the Multimodal Talker Localization System

Today's multi-sensor systems are becoming more complex with increasing number of sensors, different types of sensors and increasing complexity of the sensor. The large amount and the complexity of the raw data these sensors generate often make them difficult to combine. In the recent years, the area of data fusion has gain research interests in the multi-sensor applications because it provides a systematic approach to combine and extract useful information from the raw data. This chapter presents a high level architectural view on how data fusion can be used in multimodal multi-sensor systems. Specifically, two fusion methods, the simple summing voter and a Bayesian network, and their improved variants developed in this thesis will be studied. Applications and implementation of these fusion based multimodal systems will be given in the later chapters.

4.1 Multimodal Sensor Fusion

Often, a multimodal multi-sensor system is favored over a single sensor system. By adding more and(or) different type of sensors, the overall system's accuracy and robustness is improved. For example, the system's temporal and spatial coverage can be extended by adding more sensors. Whereas, adding different type of sensors can improve the system's coverage in the measurement space [WAL90]. However, in order to realize

these benefits, the system has to be able to take advantage of the extra information introduced by the extra sensors. Data fusion provides a mean for doing that [WAL90]. It allows information to be systematically combined from multiple sources while refining the states the system is trying to estimate [STE99]. Data fusion has been successfully deployed in the field of robotics [PET96][SEU94][YEU94] and object tracking in a variety of environments [HAL97] [STR01].

Figure 4-1 shows the general architecture of the multimodal sensor fusion system used in this thesis as block diagram. The ***Sensor*** block represents any single sensor modality using either a single sensor or a cluster of sensors. The ***Data Processing*** block processes the raw sensor data. Often, in a multimodal sensor system, fusion happens at the raw data level and the information level [LOD04B]. Therefore, the type of processing performed by the ***Data Processing*** block can range from simple data filtering at the sensor level to complex statistical analyses and features extraction at the information level. The ***Mapping*** block transforms the processed data into a common space in which all processing modules can refer to; for example, a common coordinate system or a common measuring unit. The ***Data Fusion and Decision*** block is responsible to perform the actual data fusion and contains the decision logic for the final output.

The architecture shown in Figure 4-1 is designed to be modular in nature. Data stream from each ***Sensor*** is kept separated at the beginning. Each *Sensor Module* represents a different sensor modality, and has its own associated ***Sensor***, ***Data Processing*** and ***Mapping*** blocks. All mapped data streams are combined at the last stage by the ***Data***

Fusion and Decision block to form the final estimate of the state which the system is trying to approximate. The modular nature of the architecture has the advantages of allowing high degree of flexibility. The type of sensing device, number of sensing devices, the processing method, and the fusion method used can easily be changed without affecting the rest of the system. The architecture can also accommodate an additional *Sensor Module* by simply duplicating the functional blocks and then just plugging it into the system. In this thesis, several examples of how additional *Sensor Modules* are implemented and how they can be incorporated into the system are given in Chapter 6, 7 and 8. Since the modular design decouples the processing required by each sensor modality and the data fusion computation, this architecture is well suited for performing multi-processor computing and distributed computing. The decoupling also allows slower sensing devices to be used without blocking the computation of the ***Final Estimated States***.

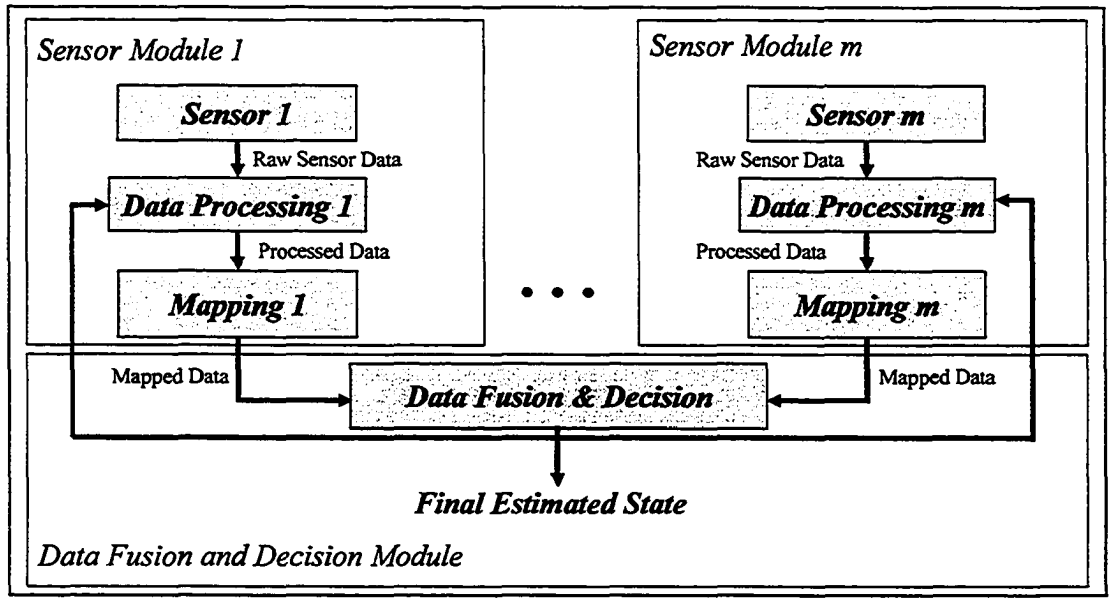

Figure 4-1. General modular multimodal sensor fusion architecture.

4.1.1 Multimodal Sensor Fusion Using Simple Summing Voter

A summing voter is one of the simplest ways to perform data fusion [LOD03A]. In this thesis, data fusion is used to combine information rather than the raw data from the sensors. A statistical measurement called correctness probabilities $P_{(m,n)}$ is developed to estimate the trustworthiness of the current detection of a particular sensory modality. $P_{(m,n)}$ is computed based on the statistical measurement of how often a state was detected in the past, and is computed for every sensor output. Mathematically, it is computed as the ratio of the number of times the current state is detected over the total number of detections done during a window of time $[i\text{-}td, i]$ (i.e., the histogram or the relative frequency of the localization).

$$P_{(m,n)}[i] = \frac{\sum_{k=i-td}^{i} D_n[k]}{\sum_{s=1}^{N} \sum_{k=i-td}^{i} D_s[k]} \tag{4-1}$$

where $D_n[k]$ is the number of detected state n at time k, td is the width of the window of time to look back to from the current data point, N is the possible states of the detection, and m is sensor modality.

Once all the $P_{(m,n)}$ are computed, a summing voter is chosen to fuse the results using

$$K_n = \sum_m P_{(m,n)} \tag{4-2}$$

where m is the sensor modality, n is the currently detected state, K_n is the fused result for state n, and $P_{(m,n)}$ is the probability of modality m detecting state n as active. Summing

voters have the advantage of being simple, have low computational requirements, and can be easily adapted to incorporate modifications. Majority rule is used as the decision logic and the state with the highest K_n is output as the final estimation.

4.1.2 Multimodal Sensor Fusion Using Bayesian Network

Two basic fusion methods are used in this thesis. Besides the summing voter introduced in the previous section, Bayesian network is another popular fusion method [LOD04A] that is used in this thesis. Consider a Bayesian network over universe U with observed evidence e expressed as the probability

$$P(U,e) = \prod_{A \in U} P(A \mid pa(A)) \cdot \prod_{i} e_i \qquad (4\text{-}3)$$

where $P(U,e)$ is the joint probability of U and e, and $pa(A)$ is the parent set of A.

In this thesis, a Bayesian network is used to fuse high level information like extracted features. Assuming the ***Data Processing*** block of sensor modality m extracts n features from the raw data: $F_{m1}, \ldots, F_{mn}$, the inference modal of the Bayesian network fusion can be represented as direct acyclic graph (DAG) [PEA88] as shown in Figure 4-2. The nodes in the DAG represent the variables, both observed and unobserved, in the universe U, and the lines between the nodes represent probabilistic dependencies as conditional probabilities. The arrows represent the direction of information flow. Therefore, DAGs used in these thesis adopt the convention of putting the lowest level of information at the top, and informational fusion starts from the top down with the final fused result at the bottom of the DAG.

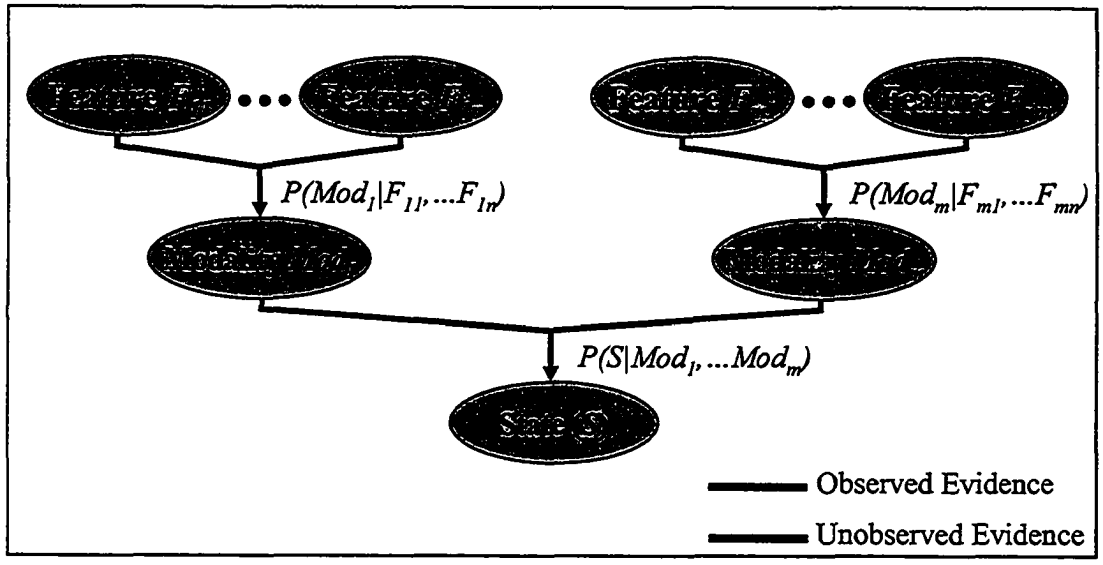

Figure 4-2. Inference model for Bayesian network multimodal sensor fusion.

The extracted features of each sensor modality are treated as observed evidences to support the modality's output in estimating state (S). With the observed evidences, equation (4-2) can be applied onto the inference modal shown in Figure 4-2 and the State node (S) can be found using

$$\begin{aligned} P(S,e) &= P(F_{11},...,F_{mn},Mod_1,...,Mod_m) \\ &= \prod_{i=1}^{m} P(Mod_i \mid F_{i1},...,F_{in}) \bullet P(S \mid Mod_1,...,Mod_m) \end{aligned} \tag{4-4}$$

where S is the state the system is estimating, e is the evidences, F_{mn} is the extracted feature of modality m with the currently value of S as n.

There is more than one way to compute $P(S,e)$. In this study, bucket elimination [PEA88] is used. Bucket elimination marginalizes one non-observed variable at a time and has it replaced with the simplified result, transversing the nodes in the inference model one at a

time. Only the non-observed variables are needed to be marginalized [PEA88]. Before the inference model can be used, each node is populated with its *a priori* knowledge.

4.2 Multimodal Sensor Fusion with Weights

The architecture shown in Figure 4-1 assumes each sensor module contributes equally in the fusion process. However, if the confidence level of one of the sensors is known to be lower then the others, less emphasis should be put on the data stream from this particular sensor. In this thesis, this is accomplished by adding weights to the architecture. Figure 4-3 shows the modified fusion architecture. The architectural components are the same as Figure 4-1 with the exception of the added ***Weight*** block. The ***Weight*** block provides a mechanism to control how much each sensor contributes in the fusion process. By modulating the value of these weights, the system can be dynamically adjusted to adapt changes in the environment [LOD03A], and accounted for failed sensors [LOD04A].

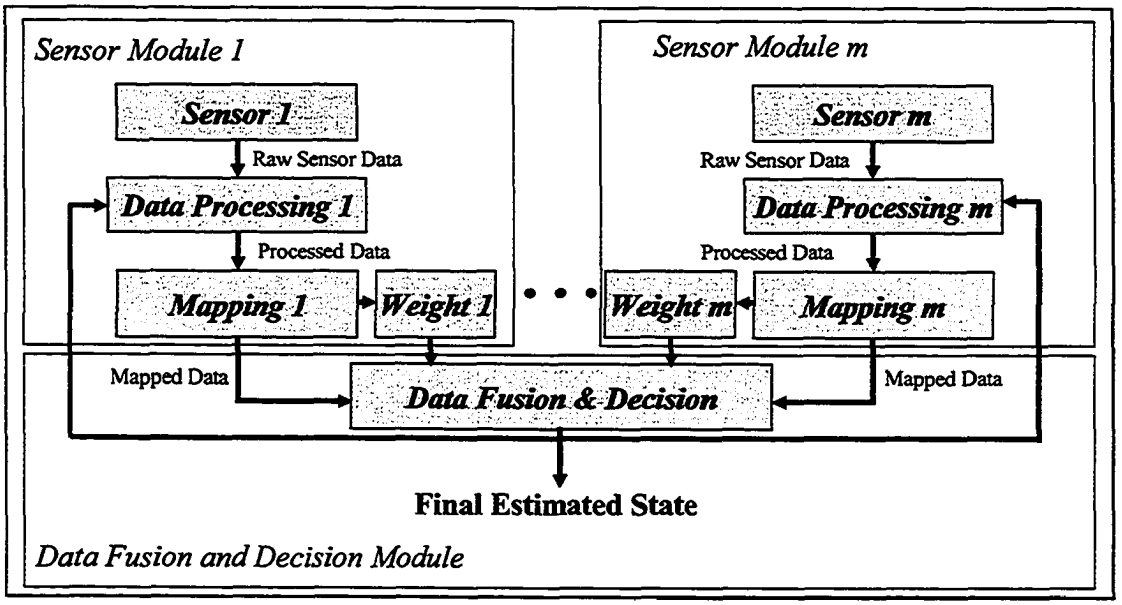

Figure 4-3. General architecture for multimodal data fusion with weights.

4.2.1 Multimodal Sensor Fusion Using Simple Summing Voter with Weights

In order to incorporate the weights into the summing voter, equation (4-2) is modified

$$K_n = \sum_m P_{(m,n)} W_{(m,n)} \tag{4-5}$$

where K_n is the fused result for the final estimated state detecting n as active, and $P_{(m,n)}$ is the probability of modality m detecting state n as active, $W_{(m,n)}$ is the weight corresponding to modality m detecting state n as active.

4.2.2 Multimodal Sensor Fusion Using Bayesian Network with Weights

Similar to the modification done to the summing voter in the previous section, the inference model of the Bayesian network is modified to include weights as well. Each of feature is modified by its corresponding weight, and a new variable *Weighted Feature WF* is used to represent the result. Figure 4-4 shows the modified inference model, and the corresponding fusion equation becomes

$$\begin{aligned} P(S,e) &= P(F_{11},\ldots,F_{mn},W_{11},\ldots,W_{mn},Mod_1,\ldots,Mod_m,S) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{n}P(F_{ij})\cdot\prod_{i=1}^{m}\prod_{j=1}^{n}P(W_{ij})\cdot\prod_{i=1}^{m}\prod_{j=1}^{n}P(WF_{ij}\mid W_{ij},F_{ij})\cdot \\ &\quad \prod_{i=1}^{m}P(Mod_i\mid WF_{i1},\ldots,WF_{in})\cdot P(S\mid Mod_1,\ldots,Mod_m) \end{aligned} \tag{4-6}$$

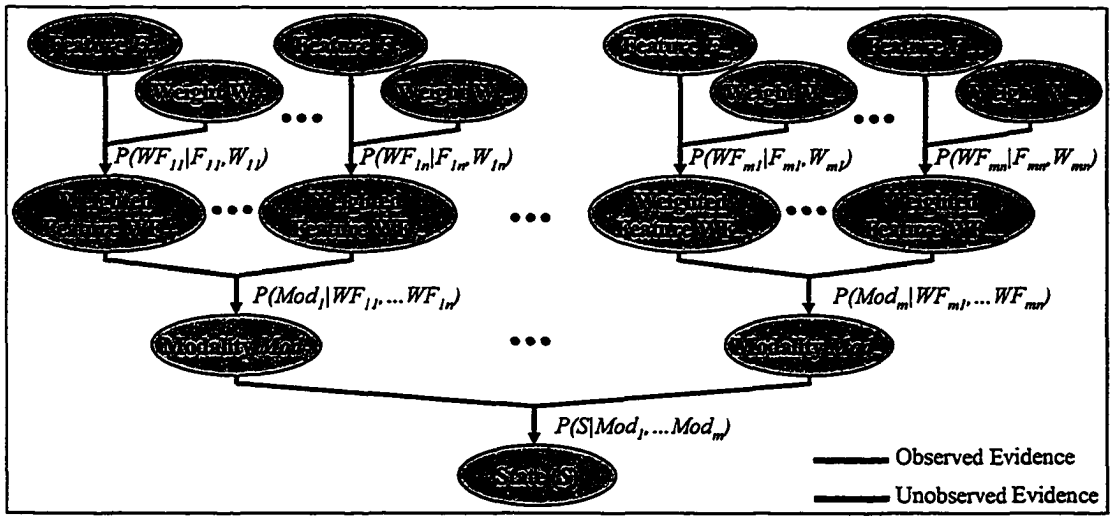

Figure 4-4. Inference model for Bayesian network multimodal sensor fusion with weights.

4.3 Multimodal Talker Localization Architecture

The general multimodal data fusion architecture shown in Figure 4-3 can be applied to a wide range of applications like surveillance [DAV97], and robotics [PET96] [SEU94] [YEU94]. As an exploration platform and case study, the application of this architecture in video conferencing for the purpose of talker localization is studied in this thesis. Figure 4-5 shows the architecture of the general multimodal talker localization. Detectable features of the talker, like speech, movements and body heat, trigger events that can be sensed by different localization modalities. The detection of these features maps well into the ***Analysis*** block in the general architecture. The ***State*** being estimated in this thesis is the talker's location.

There are several approaches to perform multimodal talker localization in video conferencing. Some researchers approach it by cascading different localizers [WAN98] [WAN99] while others might use one modality as the primary localization method with additional modalities as a means of confirmation [FIA04]. The disadvantages of these approaches are that they are essentially "hard-wired" with limited flexibility or are still relying heavily on a particular localization method.

Taking advantage of the modular architecture, data streams are decoupled early in the beginning. Each stream feeds a different *Localization Module*. Each *Localization Module* is responsible for one localization modality, and a purpose specific localizer is used to

localize the talker. Localization results from each *Localization Module* are then combined using the *Data Fusion and Decision Module* to form the final estimate of the talker's location.

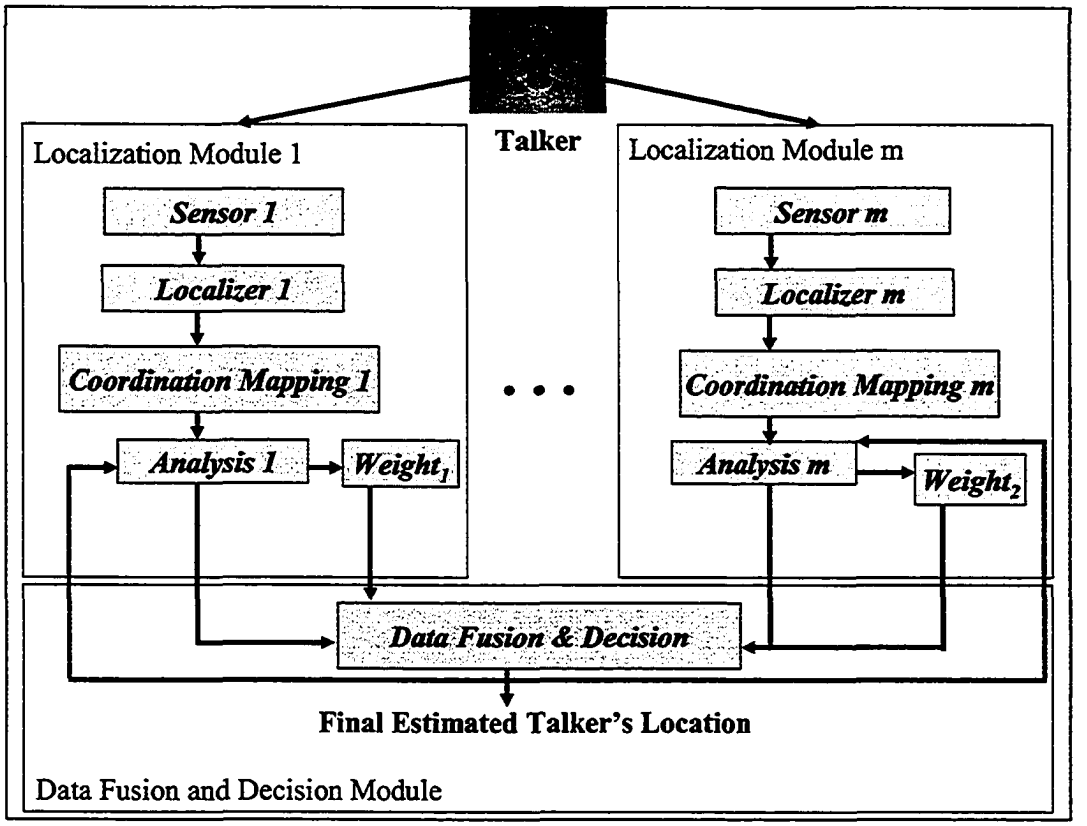

Figure 4-5. Architecture of the general multimodal talker localization system.

Each *Localization Module* contains a *Sensor* block which is any device that can sense the presence of the talker. The *Localizer* block is responsible for performing the localization. Since different sensors and localizers can have different coordinate systems, the *Coordinate Mapping* block is needed to transform the localizer's output into a common coordinate system that is used by the *Analysis* block, and the *Data Fusion and Decision*

modules. Although localization results from different localizers can be contradicting at times, the *Data Fusion and Decision* module is responsible for drawing the best out of the available results from the individual localizers, and makes a collective output based on a predefined decision rule. The *Analysis* block performs statistical analysis on the localization data. Based on the results of the *Analysis* block, the *Weight* block provides an optional bias so that the data fusion engine can put different weights on the result from a specific localizer.

The multimodal nature of the architecture allows the system to use multiple localization modalities. The following chapter explorers how this multimodal talker localization architecture can be used in video conferencing applications. The modular nature of the architecture has the advantage of allowing flexibility in the type of sensing device, number of sensing devices, the processing method, and the fusion method used. Any one of these components can be changed without affecting the rest of the system. Also, the degree of influence from each localization modality in the final result is not fixed but controlled by a weight as outlined as the *Weight* block in the block diagram. Consequently, how much the system relies on a particular modality can be dynamically adjusted according to the localization quality of each modality. This advantage will be explored further in Chapter 7. Since data streams are decoupled in the beginning and the architecture is a modular design, the system can easily be expanded by adding new sensors and localizers as a drop-in by simply duplicating the functional blocks in the localization module. Chapter 8 shows how an additional infrared (IR) localization module can be added and its impact to the overall localization performance is studied.

4.3.1 *Joint Audio-Video-Infrared Talker Localization in Video Conferencing Applications*

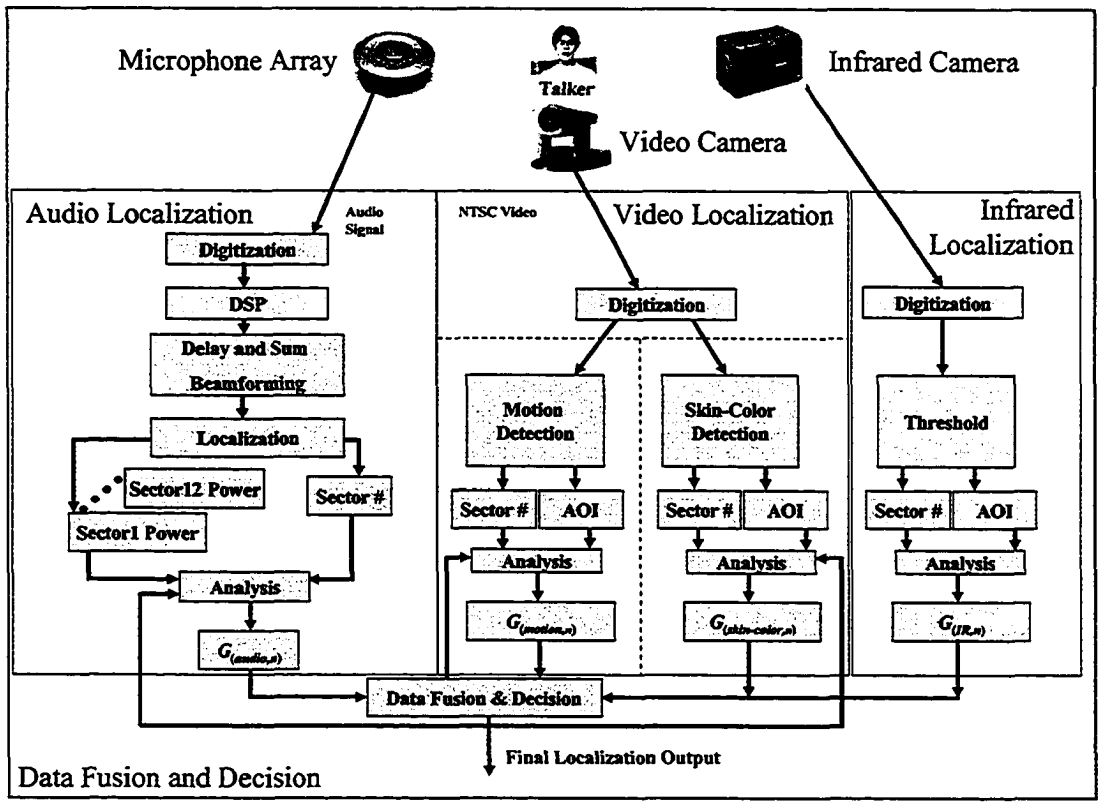

Figure 4-6. System block diagram for a video conferencing application using multimodal talker localization system.

The system block diagram shown in Figure 4-6 is a specific example of how the block diagram shown in Figure 4-5 can be realized. It is intended as an exploration platform for studying how the modular multimodal data fusion architecture shown in Figure 4-3 can be used in particular applications like video applications. In this specific example, three different types of single-modal localization methods are used: audio, video, and IR. Taking advantage of the modularity of the system, each localization modality is developed, tested and composed separately before they are put together as a complete

multimodal system. In this thesis, each localization modality is tested as a standalone single modal localization method first. The multimodal localization system is tested in stages using just the audio and video localization modality, followed by the infrared localization as the third localization modality. The performance of the joint audio-video localization method is first studied and compared, and then the effect of adding the additional infrared localization modality is studied.

Speech, actions and body heat from the talker trigger audio, video and thermo events that can be captured by the microphone array, video camera, and IR camera. The system keeps separate the audio, video, and IR video signals at the initial stages. The audio, video, and IR data are digitized, processed, and then localization is performed on each input modality separately using specialized localizers such as the audio beamforming, motion detection, skin-color detection, and thermo detection localizers. The results from the localizers are statistically analyzed. Details of the analysis are covered in Chapter 5. The *Weight* block in the general architecture, Figure 4-3, is implemented as the probability occupancy grid for the talker's location $G_{(m,n)}$. The probability occupancy grid describes the probabilistic estimates of the talker's occupancy state [SEU94] for localization modality m and localization sector n. Details of $G_{(m,n)}$ are given in Chapter 7. The results from the audio, video, and thermo localization are then combined using data fusion methods like the summing voter and Bayesian networks. Based on the "fused" localization results, the decision logic decides the final estimation of the talker's location. Details of the two different fusion methods used in this thesis will be discussed in Chapter 6, 7, and 8.

In the audio localization modality, a beamforming audio localizer is used to sense the talker's voice activities. In the video localization modality, two different localizers are used. The motion detection localizer is used to sense the talker's motion, and the skin-color detection localizer is used to sense large skin surfaces of the talker, like face and hands. In the IR localization modality, a thermo detection localizer is used to sense the talker's body heat. The details of these single modal localizers will be discussed in Chapter 5. The details of joint audio-video multimodal localization will be discussed in Chapter 6 and 7, and multimodal localization using audio, video and IR methods will be discussed in Chapter 8. Based on experimental results [LOD03A], [LOD03B], [LOD04A], [LOD04B], multimodal talker localization provides the advantages of better localization accuracy and robustness when compared with single modal localization methods [LOD03A], [LOD03B], [LOD04A], [LOD04B].

Chapter 5 Single Modal Talker Localization

In this chapter, we look at how talker localization is done traditionally using various single modal methods like audio, video and infrared. The working principle and the implementation of each of these methods are studied. In the Chapters 6, 7, and 8, these single modal localization methods are combined using the fusion architecture shown in Figure 4-5 to perform multimodal talker localization.

5.1 Talker Localization Using Audio Information

In this thesis, a beamforming microphone array is used as the audio localizer. This section takes a more detailed look at how audio data acquired by the microphone array is used to localize the talker.

5.1.1 Circular Microphone Array

A microphone array can be configured linearly or in other geometrical shapes, a circle for example. The microphone array used in this thesis is circular in shape, and has six microphones embedded in it. Figure 5-1 shows a circular microphone array with a far field source and sound waves incident at angle θ.

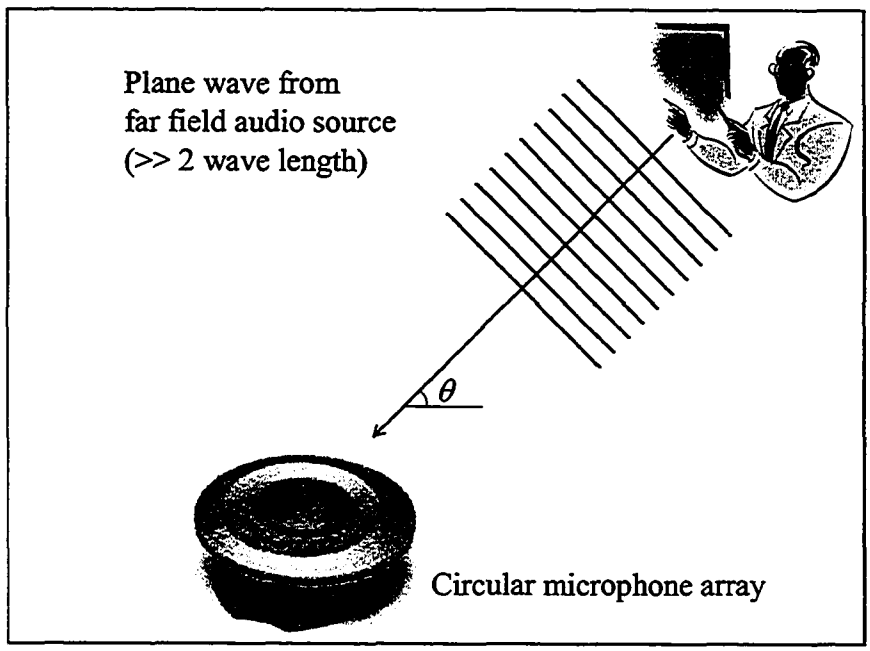

Figure 5-1. Circular microphone array with plane wave coming in from a far field source.

5.1.2 Beamforming and Localization Output of the Microphone Array

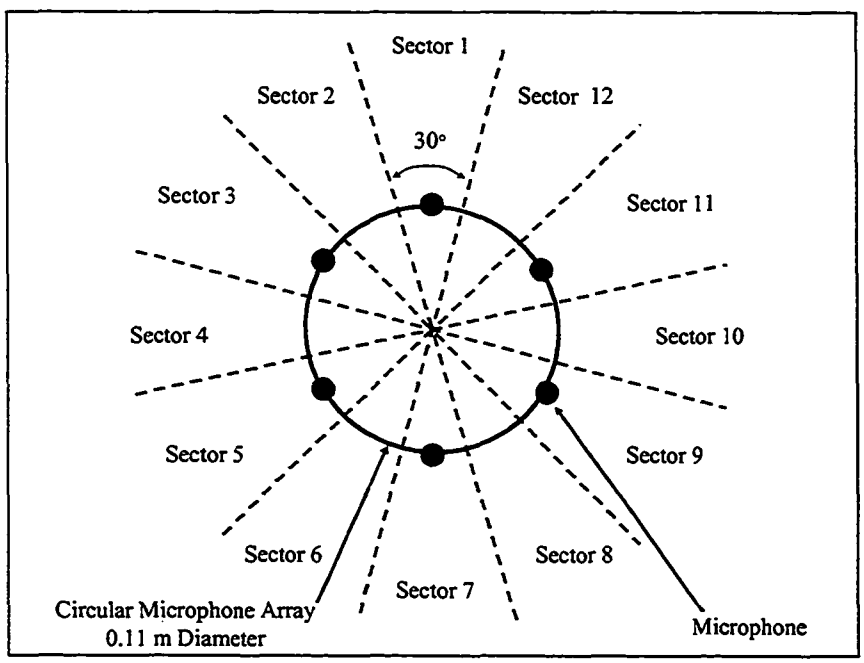

Figure 5-2. Localization sectors of the microphone array.

Although, the microphone array only has six microphones, through delay-and-sum beamforming, the microphone array is capable of segmenting the space around it into 12 sectors with each sector spanning 30° as shown in Figure 5-2. These sectors are labeled from 1 to 12. Throughout this thesis, localization results will be reported in terms of sector numbers.

The delay-and-sum beamforming algorithm used in this thesis is actually capable of giving higher spatial resolution than 12 sectors. However, due to the current design of the microphone array, each main beam is about 40° wide [MIT02]. There is 3 dB drop from the center of the main lobe to the edge of main lobe limiting the spatial resolution [MIT02]. Engineers at Mitel Networks decided to set the spatial resolution of the microphone to 12 sectors giving a good compromise between spatial resolution and detection error rate [MIT02].

The beamforming algorithm combines the signals from the various microphones to enhance the audio signal originating from a desired location and attenuate the audio signals originating from all other locations. Given a microphone array with any number of microphones and beamforming algorithm capable of detecting audio signals originating from N different directions (sectors), the N beamforming algorithms have N output signals $B_1[t]$, $B_2[t]$, ..., $B_N[t]$. The delay-and-sum algorithm is a commonly used beamforming technique [JOH93] and it is chosen for this thesis because of its simplicity. The required delays are calculated based on the physical layout of the microphone array with the assumption that the source is in the far field. The outputs of the microphone

array are the windowed power signals $P_i(t)$, $i=1,\ldots,N$, for each sector which are calculated over the time window $[t-\Delta_t, t]$ from the beamformed signals $B_i(t)$ using

$$P_i(t)=\frac{1}{\Delta_t}\int_{t-\Delta_t}^{t} B_i^2(t)\,dt \qquad i\in[1,N] \tag{5-1}$$

where Δ_t is the width of the time window. Using the sampling notation $B_i[n]\triangleq B_i(nf_s)$, where $n\in\mathbb{Z}$ and f_s is the sampling frequency, then for the numeric implementation we used

$$P_i^{(k)}=\frac{1}{M}\sum_{n=0}^{M-1} B_i^2[n+(k-1)D_B] \qquad i\in[1,N] \tag{5-2}$$

where $M=\lceil \Delta_t f_s \rceil$ samples is the width of the window and D_B controls the spacing of the windows in $B_i[n]$ for the k^{th} window forming $P_i^{(k)}$. Notice that if $D_B=M$, then the windows extracted from $B_i[n]$ are non-overlapping. In this particular study $D_B=M$, $\Delta_t=1$ ms, $f_s=8000$ Hz, $N=12$ sectors, and $M=\lceil(0.001)(8000)\rceil=8$ samples.

Once the windowed power signals $P_i^{(k)}$ are computed for the k^{th} window, then a decision must be made as to which sector $i=i_{active}$ is *active* (*i.e.*, enough sound to be identified as voice activities). The approach used in this study is to set a predetermined power threshold T_{active} and take the sector with the maximum power that is greater than this threshold such that

$$i_{active}=\underset{i}{\operatorname{argmax}}\left\{P_i^{(k)}\right\}>T_{active}. \tag{5-3}$$

In this thesis, we assume that the background noise power is small compared to the speech signals. If the noise levels are higher, then extensions to this approach would require a more in depth analysis of the signal characteristics to distinguish speech from background noise. Also note that if multiple talkers are speaking, then the one with the

greatest windowed power is considered active. Note that this approach to determining the active sector can produce undesirable results when steering a camera since it may erroneously swap back and forth between a talker and an acoustic reflection. The next section looks at how to avoid this problem when considering a switch to a new active sector.

5.1.3 Normalized Cross-Correlation and Correlation Lag

One of the analysis for the audio localization that is done in the Analysis Block in Figure 4-5 and Figure 4-6 is cross-correlation. It is used to avoid having acoustic reflections erroneously displace the active sector. A novel method is developed in this thesis to distinguish between reflections and a second talker. The key to the approach is that the cross-correlation of the speech with the reflection should have a higher maximum than for the two talker scenario. For our approach, we use the windowed power signal $P_i^{(k)}$ in the cross-correlations instead of the beamformed signals $B_i[n]$. Since signal power can vary, before the cross-correlation is computed the windowed power signals $P_i^{(k)}$ are normalized [LOD05]. This normalization is done by first calculating the windowed root-mean-square (wRMS) on the last W values of $P_i^{(k)}$ as follows

$$\text{wRMS}_k=\sqrt{\frac{1}{W}\sum_{n=0}^{W-1}\left[P_i^{(k-n)}\right]^2}. \tag{5-4}$$

Then $P_i^{(k)}$ is normalized by the windowed RMS to obtain

$$\overline{P}_i^{(k)}=\left(\frac{1}{\sqrt{W}}\right)\frac{P_i^{(k)}}{\text{wRMS}_k}. \tag{5-5}$$

Note that since $P_i^{(k)} \geq 0$, then the following holds for $\overline{P}_i^{(k)}$: (i) $\inf \overline{P}_i^{(k)} = 0$, (ii) $\sup \overline{P}_i^{(k)} = 1$, and (iii) $\overline{P}_i^{(k)}$ is undefined if all W of the $P_i^{(k)}$ are zero (*i.e.*, $\overline{P}_i^{(k)} = 0/0$).

5.1.4 Acoustic Reflection Detection

In hands-free conferencing systems, acoustic reflections from the walls can cause reverberations that deteriorate the audio quality. This phenomenon is commonly known as the barrel effect [RAD00]. Microphone arrays and beamforming techniques can be used to remedy this problem by capturing the sound originating from a desired direction and attenuating the sounds originating from all other directions [BED94]. However, simple microphone arrays cannot discriminate between the case of a single talker with a strong acoustic reflection, Figure 5-3(a), and the case of two different talkers, Figure 5-3(b) [LOD03B][LOD05]. This problem is more serious when the microphone array is part of a video conferencing system where the array passes the talker's direction to a steerable camera [LOD03A], [LOD04]. If the microphone array mistakenly interprets a reflection as a second talker, the camera could point to the wall, post, or column that caused the reflection. This scenario is common when a talker speaks toward another participant or the white board during the conference instead of speaking toward the microphone array, which results in a reflected audio signal stronger then the direct path causing the array to localize incorrectly [LOD03B][LOD05].

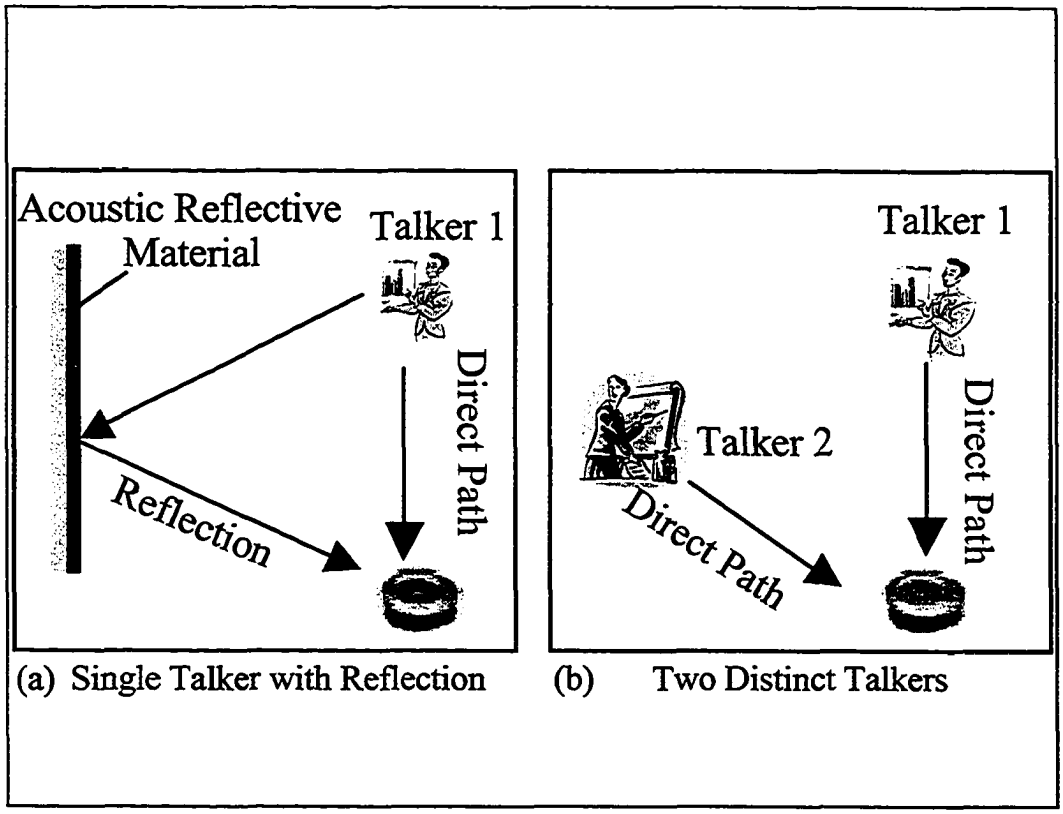

Figure 5-3. Two scenarios; (a) single talker with strong reflection, and (b) two distinct talkers. A simple microphone array detects two talkers, and cannot distinguish the difference between these scenarios.

Most audio systems are designed such that the talkers have to be as close to the microphones as possible. This way, the direct path signal is significantly stronger than the reflections. If the talk's direction is known, directional microphones can be used. In other applications where the location of the desired audio signal is not known or is dynamic, a microphone array equipped with a beamforming algorithm is usually used to locate and track the desired audio signal. This reflection detection method deals with the case when the location of the desired talker is unknown. A number of different approaches like post filtering [MAR98], general cross-correlation [BED94], and subspace tracking [AFF96] have been proposed to solve this problem. Other approaches include the use of near-field

beamforming techniques [RAY03], [ZHE03] to restrict the detection capability of the microphone array to a given distance from the array, thus reducing the magnitude of the acoustic reflections. In the past, other researchers have considered using cross correlation of signals originating from pairs of microphones [BED94], [BRD99].

A portion of this thesis investigates the effectiveness of a new method for discerning whether an audio localization detection originated from an acoustic reflection or a new talker. The proposed acoustic reflection detection method is different from what has been done by other researchers. It considers the average power of the beamformer's output instead of the raw microphone signals; therefore, it is not restricted to a specific beamforming technique and is able to achieve a significant reduction in computational complexity. The novelty of the method is to use correlation lag to discriminate between the case of a single talker with acoustic reflections and the case of multiple talkers regardless of their power levels and how reverberant the environment is. The algorithm is simple and can be performed in real time.

When the microphone array detects a new voice activity, the wRMS power signal $\overline{P}_X^{(k)}=\overline{P}_{i(n)}^{(k)}$ of the current active sector and $\overline{P}_Y^{(k)}=\overline{P}_{i(n-1)}^{(k)}$ of the previous active sector are cross-correlated using equation (5-6) so that it can discern whether the active sector switch is due to a reflection or some other uncorrelated signal such as from another talker.

$$R_{\overline{P}_X\overline{P}_Y}(l)=\sum_k \overline{P}_X^{(k+l)}\overline{P}_Y^{(k)} \tag{5-6}$$

What is of interest is the lag that maximizes this cross-correlation. We define the maximum correlation lag l_{max} as the undirected lag between the two wRMS power signals at which the maximum cross-correlation occurs which can be expressed as follows

$$l_{max} = \left| \arg\max_{l} R_{P_x P_y}(l) \right|. \tag{5-7}$$

In experiments, l_{max} is used as an indicator of an acoustic reflection versus an uncorrelated second talker.

Since the maximum cross-correlation result between the two RMS normalized series is N, the cross-correlation result is first divided by N so that it ranges from zero to one with 0 corresponding to no correlation and 1 corresponding to 100% normalized correlation. Maximum correlation lag is defined as the lag between the two signals at which the maximum cross-correlation occurs. The detection decision is made based on the value of the cross-correlation and its lag

$$\begin{cases} \max(R_{(ab,n)}) > T_{cor} \\ \left| Lag_{(ab,n)} \right| > T_{lag} \end{cases} \Rightarrow \text{Reflection} = \text{True} \tag{5-8}$$

where $R_{(ab,n)}$ is the cross-correlation between the signal power of the two sectors that are being checked, $|Lag_{(ab,n)}|$ is the corresponding magnitude of the maximum correlation lag, T_{cor} and T_{lag} are threshold values for cross-correlation and the maximum correlation lag.

5.1.5 Model for Delay-and-Sum Beamforming Microphone Array

In order to study the acoustic reflection detection method outlined in Section 5.1.4 using computer simulation, a mathematical model for the delay-and-sum beamforming

microphone array is derived. A delay-and-sum beamformer operates on the principle that the proper amount of delays, which corresponds to the TDOA of the individual microphone's signals, are applied to the microphone signals and then summed. The signal in the principle direction is enhanced and signals in other directions are attenuated. Therefore, the amount of TDOAs corresponding to each microphone on the circular array has to be computed. The array was modeled as six microphones evenly placed on the circumference of a circle with 0.11 m diameter, Figure 5-4, using microphone 1 as reference. The TDOA for the other microphones can be computed using trigonometry and then multiplied by the speed of sound in air. For example, the TDOA of microphone 5, as shown in Figure 5-4, can be computed as

$$\begin{aligned} &\varphi_5 = 180 - 60 - \theta = 120 - \theta \\ &\cos\varphi_5 = \frac{\delta_{15}}{c_{15}} \\ &c_{15} = \sqrt{3}\frac{\Phi}{2} \\ &\therefore \delta_{15} = \frac{\sqrt{3}}{2}\Phi\cos(120-\theta)\bullet S \end{aligned} \tag{5-9}$$

Similarly, the TDOA of the rest of the microphones can be computed as

$$\begin{aligned} &\delta_{11} = 0 \\ &\delta_{12} = \frac{\Phi}{2}\cos(\theta-30)\bullet S \\ &\delta_{13} = \frac{\sqrt{3}}{2}\Phi\cos(\theta-60)\bullet S \\ &\delta_{14} = \Phi\cos(90-\theta)\bullet S \\ &\delta_{15} = \frac{\sqrt{3}}{2}\Phi\cos(120-\theta)\bullet S \\ &\delta_{16} = \frac{\Phi}{2}\cos(150-\theta)\bullet S \end{aligned} \tag{5-10}$$

Where δ_{ij} is the TDOA for microphone j using microphone i as reference, S is speed of sound in air, and θ is the principle direction (i.e., the main lobe) of the beamformer.

Once the proper values of TDOA have been computed, the model shown in Figure 5-5 is used to compute the beamformer's output. Using θ equal to 0°, 30° 60°, 90°, 120°, 150°, 180°, 210°, 240°, 270°, 300°, and 330°, the space around the microphone array is partitioned into 12 sectors, matching the sectors that are used in the actual microphone array as mentioned in Section 5.2.

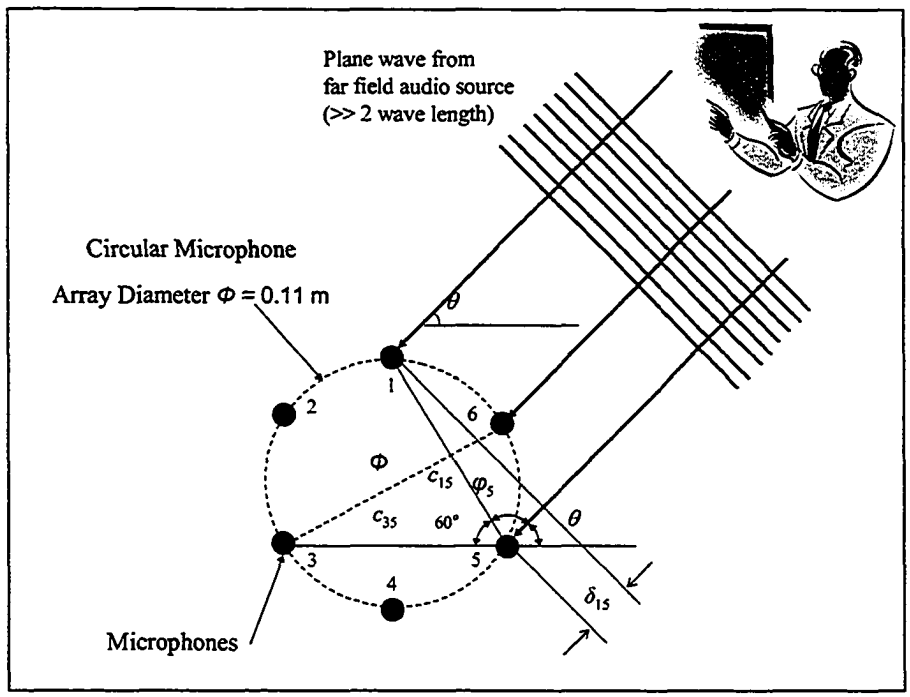

Figure 5-4. Triangulation of the time delay of arrival (TDOA) for a circular microphone array.

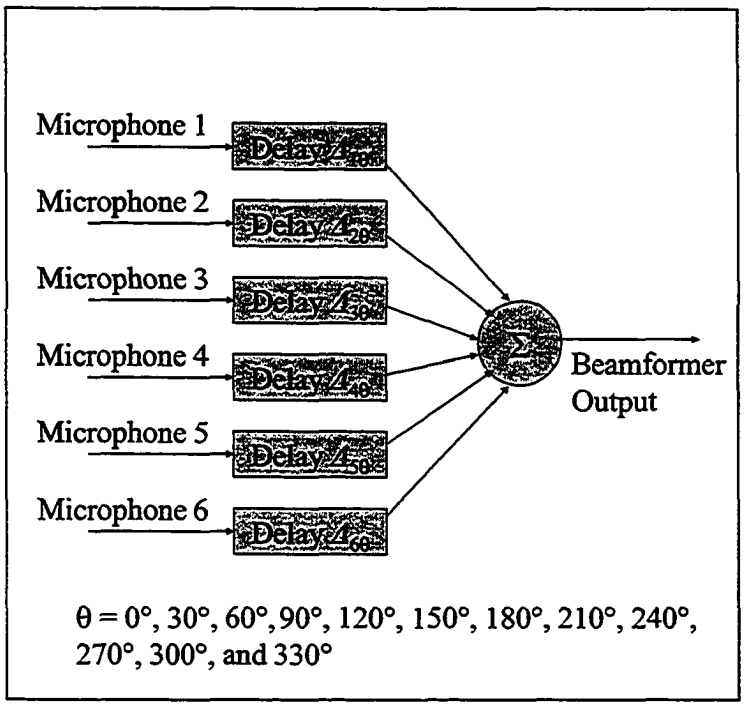

Figure 5-5. Block diagram of the delay-and-sum beamformer.

5.1.6 Model for Human Voice

In the computer simulations, human voice is modeled as Gaussian white noise because of its simplicity and un-biasing nature. The beamforming microphone array used in this thesis band-pass filtered the audio signal to 1 kHz – 1.3 kHz. In order to match the bandwidth of the actual microphone array used as close as possible, a linear phase FIR filter is used to band-pass filter the Gaussian white noise signal to 1 kHz – 1.3 kHz as well.

5.1.7 Averaged Beam Pattern

The averaged beam pattern is another novel method developed in this thesis that is used to gauge how good an audio localization result is. As mentioned before, the delay-and-sum beamformer has the effect of reinforcing the signal in the source direction while attenuating signals from all other directions. With this property in mind, when the output powers are averaged over a time window t_{avg} for each sector and plotted against sector number in polar coordinates, the averaged beam pattern of the beamformer can be observed. It is expected that the power in the active sectors will be higher, whereas the power in the neighboring sectors will be lower. This distinct pattern will be referred to as the averaged beam pattern in this thesis. For example, Figure 5-6 shows the averaged beam pattern where sector 6 is the active sector and t_{avg} equals to 66 ms. The main lobe is centered at sector 6 with 12 dB gain, and 11 dB gain in the two neighboring sectors. There is also a small side lobe with 7 dB gain centered at sector 12 which is directly opposite of sector 6. The small side lobe is typical in a delay-and-sum beamformer [JOH93].

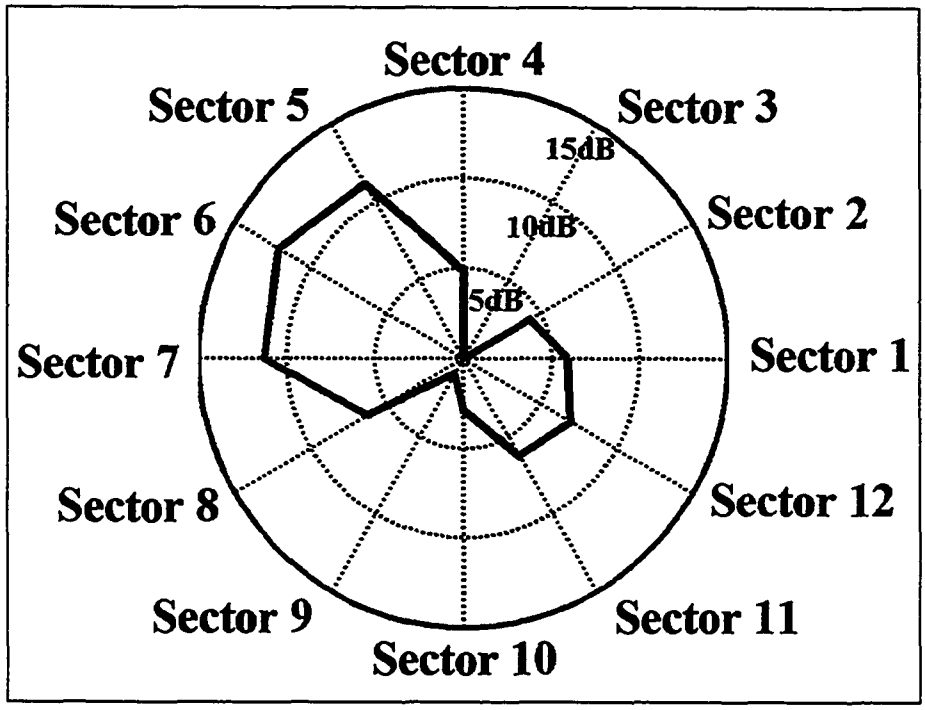

Figure 5-6. Average beam pattern.

5.1.8 Anechoic Experiments

5.1.8.1 One Active Sector At A Time

In order to study and establish the baseline behavior of the beamforming microphone array used in this thesis, a reproducible audio source was presented to the microphone array one sector at a time inside the anechoic chamber. As shown in Figure 5-7, the microphone array was placed in the middle of the anechoic chamber and a loudspeaker was placed at 1 m away from it. An audio file which includes male and female voices speaking in English, Chinese and Japanese was then played back with the same loudness, one sector at a time. Acoustic data were sampled and localization was done using the delay-and-sum beamforming.

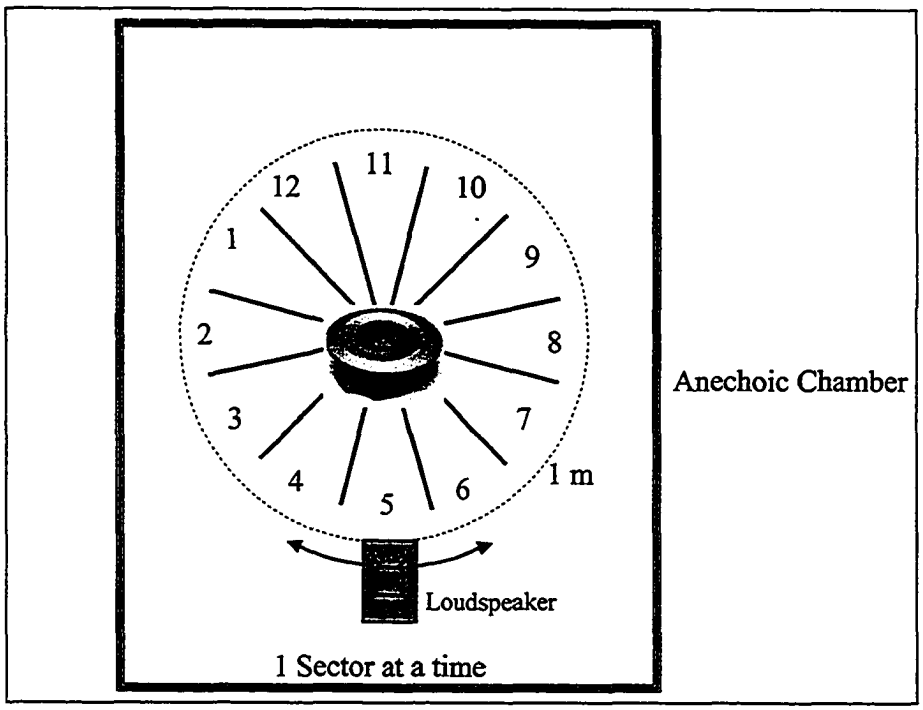

Figure 5-7. Experimental setup for anechoic one sector at a time experiment.

5.1.8.2 Reflection / Multiple Talker Detection in an Anechoic Chamber

In this experiment, the scenarios where a single talker with strong acoustic reflection, and two distinct talkers having a conversation (Figure 5-3) are recreated. The microphone array was placed in the middle of the chamber. A loudspeaker and an acoustic reflective panel were placed at 1 m from the microphone array as shown in Figure 5-8(a). Pre-recorded human speech was played through the loudspeaker. The microphone array captured the voice activities from the direct path and its reflection path at different sectors. The experiment was then repeated with the reflective panel removed, Figure 5-8(b), and then again with the reflective panel replaced by another loudspeaker playing a different speech recording, Figure 5-8(c). All data were recorded and digitized. The

cross-correlation of the two active sectors was computed using the beamformer's power signals. A 250 ms data window was used in the computation and the maximum correlation lags were recorded.

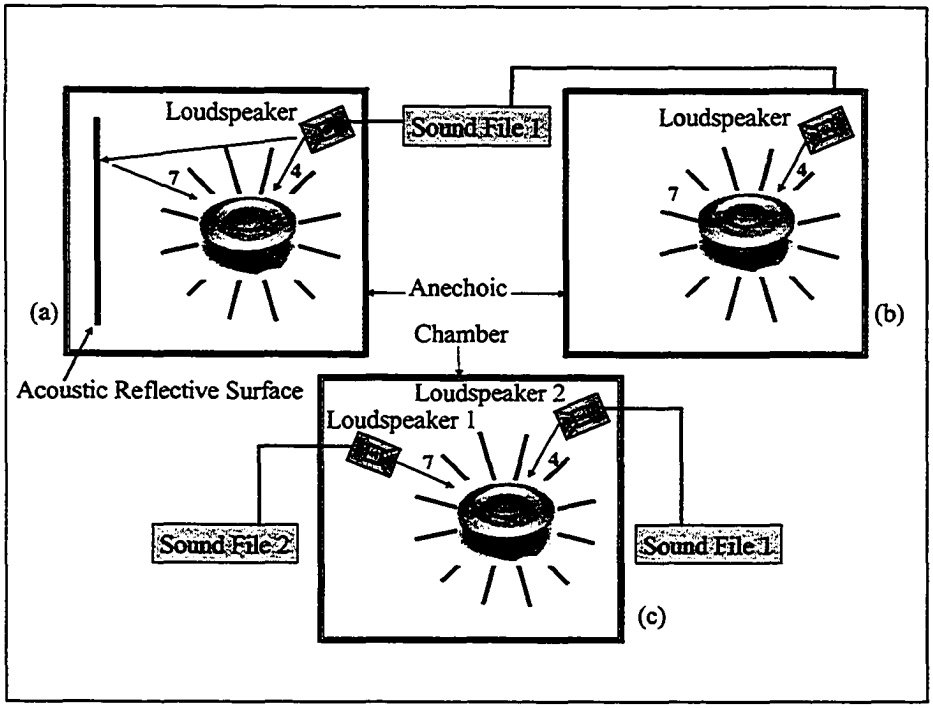

Figure 5-8. Experimental setup for anechoic reflection / multiple talker detection. (a) single talker with strong reflection, (b) single talker with no reflection, and (c) multiple talkers.

5.1.9 Reverberant Room Experiments

5.1.9.1 Reflection / Multiple Talker Detection in a Reverberant Room

The reverberant experiment was conducted in a 5 m x 7 m conference room lined with drywall. The experimental setup is shown in Figure 5-9 and it was similar to the anechoic experimental setup. An acoustic reflective panel and a loudspeaker were placed at two different sectors of the microphone array, Figure 5-9(a). Pre-recorded human speech was

played back on the loudspeaker. Voice activities from the direct path and the reflection path were captured by the microphone array. The experiment was then repeated with the reflective panel replaced by another loudspeaker. In order to remove all reflections, acoustic foam panels were placed behind the second loudspeaker, Figure 5-9(b). Again, a 250 ms data window was used to compute the cross-correlation of the two active sectors, and the maximum correlation lags were recorded.

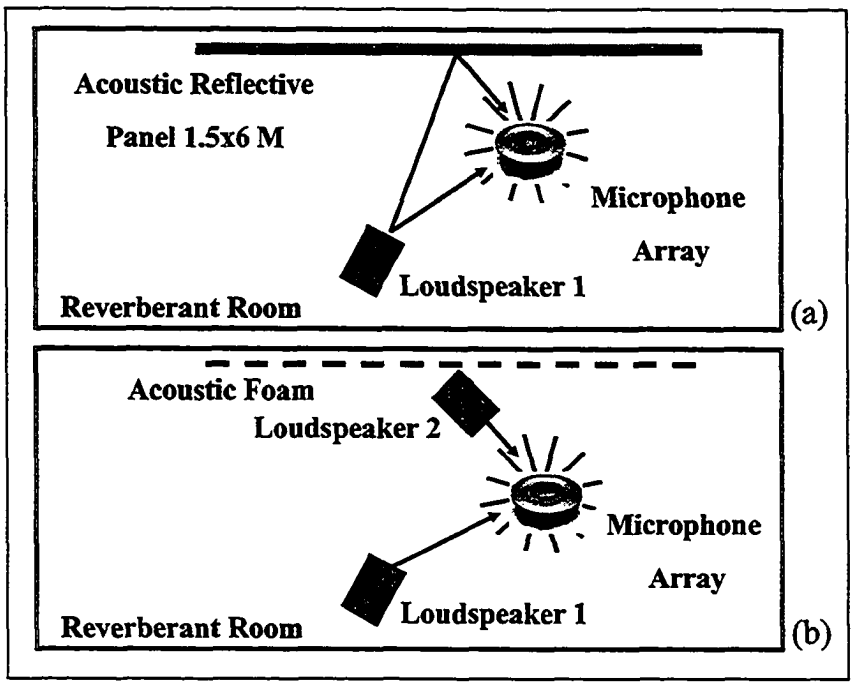

Figure 5-9. Experimental setup for the reverberant reflection / multiple talker experiment. (a) single talker with strong reflection, and (b) multiple talkers.

5.1.10 Computer Simulations

5.1.10.1 Reflection / Multiple Talker Detection using Computer Simulation

Figure 5-10 shows the simulation scenario. In the simulation, the delay-and-sum beamforming microphone array was modeled as what is outlined in Section 5.5. Human speech used in the simulation was modeled as band-limited Gaussian white noise as outlined in Section 5.6. Two simulations were performed. In the first simulation, it was assumed that an acoustic reflective panel was causing reflections, and as a result, an acoustic image was formed. The first order reflection image was used in the calculation. The simulation was repeated assuming there were no reflections and the reflective image was replaced by another uncorrelated band-limited white noise source to simulate the presence of a second distinct talker. The beamformer's output signal powers of the two active sectors were then cross-correlated and the values of the maximum correlation lags were recorded.

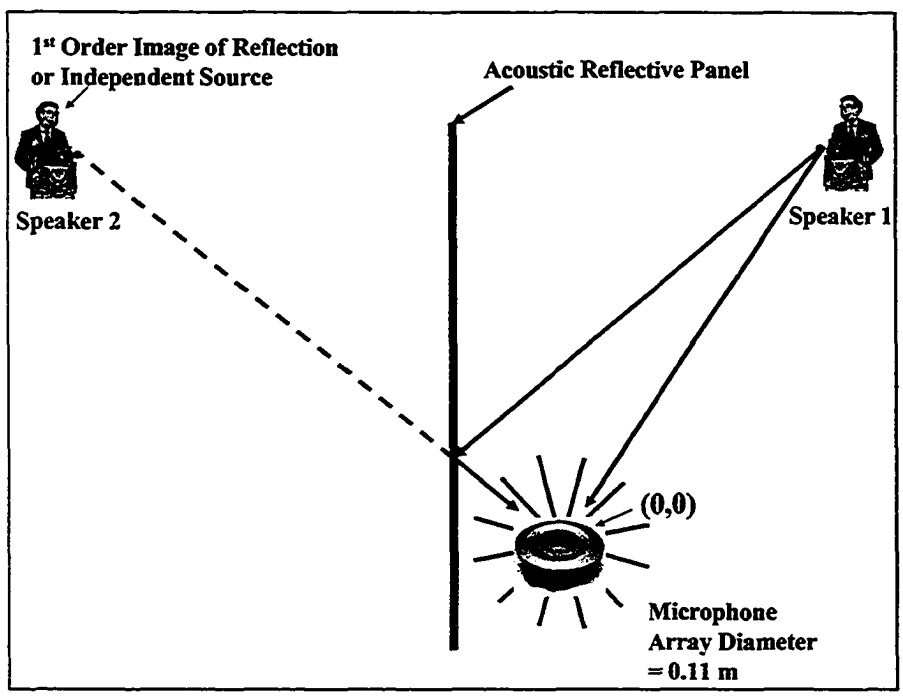

Figure 5-10. Scenario for the reflection / talker computer simulation.

5.1.11 Results and Discussions

Two groups of experiments and computer simulations were done in order to study the effectiveness of the proposed reflection detection method. The first group was conducted in an anechoic chamber and the second group was conducted in a reverberant room. The anechoic chamber provides an idealized environment by eliminating any artifacts due to the room's acoustic properties. The reverberant room provides a most realistic environment to study how the reflection detection method performs in the real word. Figure 5-11, 5-12 and 5-13 show the results of the experiments and simulations. In each of the figures, the upper panel labeled (a) corresponds to the single talker with strong reflection scenario, Figure 5-3(a). The lower panel (b) corresponds to the scenario having two distinct talkers, Figure 5-3(b). Within each panel are two plots. The upper plot shows the maximum normalized cross-correlation and the lower plot shows the maximum correlation lag. The x-axis is time in seconds.

Figure 5-11 shows a typical result for the anechoic experiments. Figure 5-12 shows a typical result for the reverberant experiments. In order to further study how the proposed method behaves under different constraints and with different parameters, the fundamentals, for example beamforming, reflections and speeches, have been extracted, simplified and implemented as a computer simulation. A typical result is shown in Figure 5-13. In all three figures, Figure 5-11, 5-12 and 5-13, note that the single talker scenario consistently exhibits a greater correlation lag. The lag represents the delay between the path of the direct and the reflected signals. In the context of echo detection, the sign of

the correlation lag does not bear useful information. It is only the magnitude that is of significance. The sign of the correlation lag will depend on which sector is being used as the reference in cross-correlation calculation. The occasional jumps in the correlation lag are due to the positive correlation of the background noises when the talkers are not speaking. In Figure 5-11(a), from 1.9s – 7.5s, the mean l_{max} is 43.7 ± 4, whereas the mean l_{max} in Figure 5-11(b) is 3 ± 0. In Figure 5-12(a), from 0.9s – 6.1s, the mean l_{max} is 55 ± 22, whereas the mean l_{max} in Figure 5-12(b) is 7 ± 23. In Figure 5-13(a), from 0.05s – 0.89s , the mean l_{max} is 7 ± 0, whereas the mean l_{max} in Figure 5-13(b) is 0 ± 0. However, the correlation values are very close when the single talker scenario and the two distinct talkers scenario are compared. For the anechoic experiments, the mean of the maximum correlation is 0.74 ± 0.05 for Figure 5-11(a), and 0.83 ± 0.08 for Figure 5-11(b). For the reverberant experiments, the mean of the maximum correlation is 0.51 ± 0.14 for Figure 7-12(a), and 0.52 ± 0.23 for Figure 5-12(b). For the computer simulation, the mean of the maximum correlation is 0.95 ± 0 for Figure 5-13(a), and 0.85 ± 0 for Figure 5-13(b). Comparing the two talkers scenario, 5-11(b), 5-12(b), and 5-13(b), with the single talker with reflection scenario, 5-11(a), 5-12(a), and 5-13(a), although the value of the correlation may still be large due to the side lobes of the beamforming algorithm, the values of the correlation lag were consistently small for two talkers and consistently larger for single talker with reflection.

The results demonstrate that using cross-correlation alone as an indicator of acoustic reflections is not sufficient for robustly detecting the presence of a strong echo. However,

with the addition of maximum correlation lag l_{max}, reflections can be distinguished reliably from other distinct talkers regardless of the reverberation in the environment.

In this section, a novel method that uses the normalized maximum correlation lag has been investigated as a means for detecting reflections in microphone array applications. The presented method uses the beamformer's power signals in its calculation. Because a power signal has less data than the original audio signals, its cross-correlation task poses a lower computational load. Also, the amount of lag is directly proportional to delay in the reflection path of the voice signal. Therefore, the proposed method can also be used to estimate the size of the conference room in order to select the appropriate beamforming algorithm (e.g. near-field parameter).

Voice is one of the primary methods of communication in conferencing. Beamforming microphone arrays provide an effective means in capturing the talker's voice. However, microphone arrays can be sensitive to reflections and other acoustic properties of the room. In the next chapter, we will look into alternative ways to locate the talker using video means.

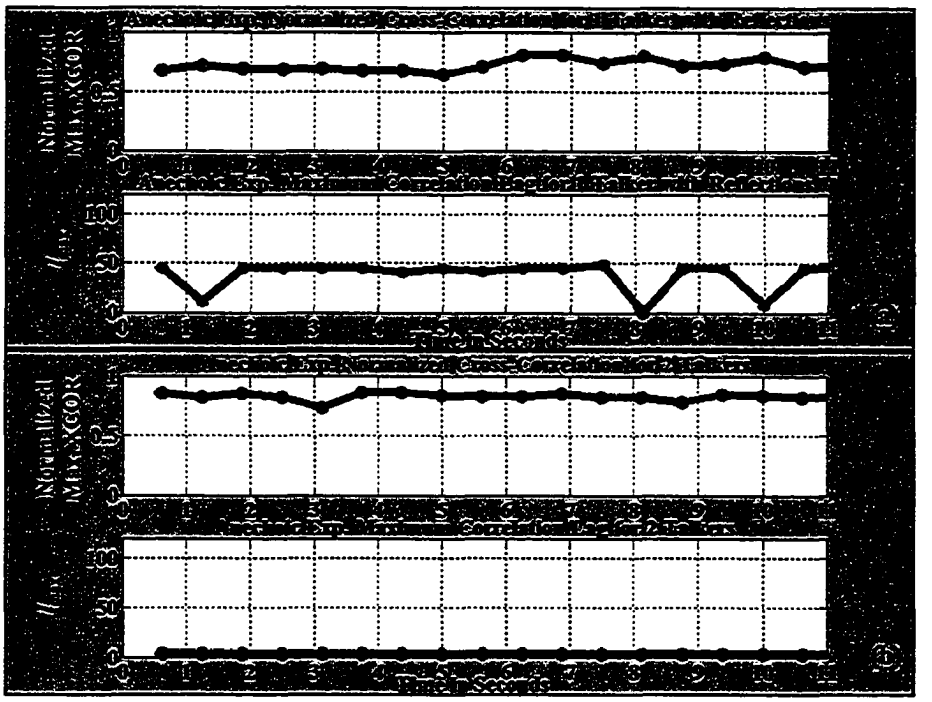

Figure 5-11. Results for anechoic talker / reflection detection experiment; (a) single talker with reflections, and (b) two individual talkers.

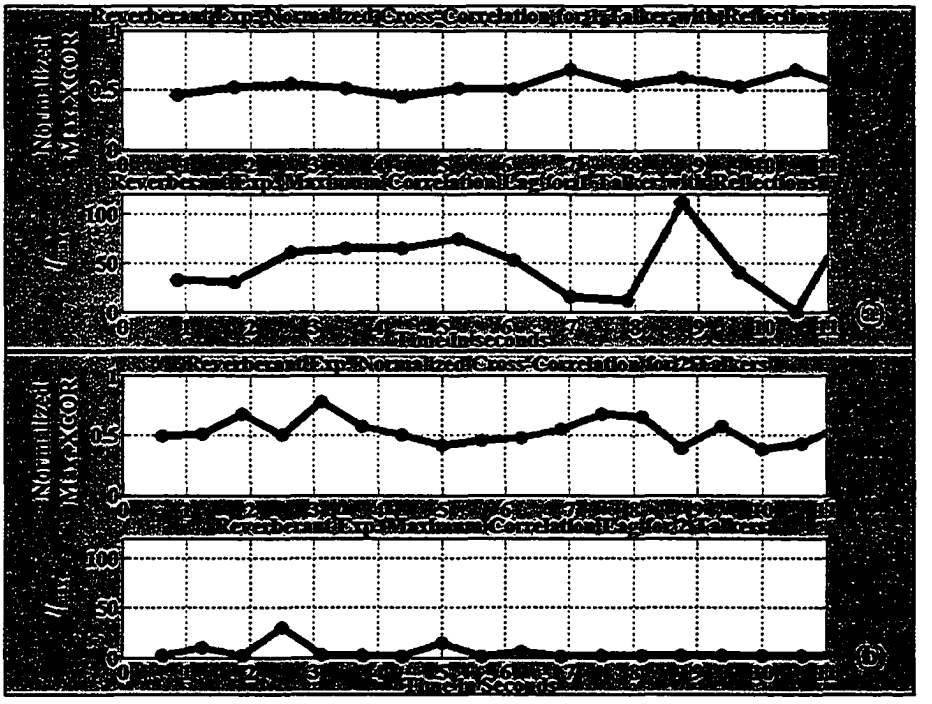

Figure 5-12. Results for reverberant talker / reflection experiment; (a) Single talker with reflections, and (b) Two individual talkers.

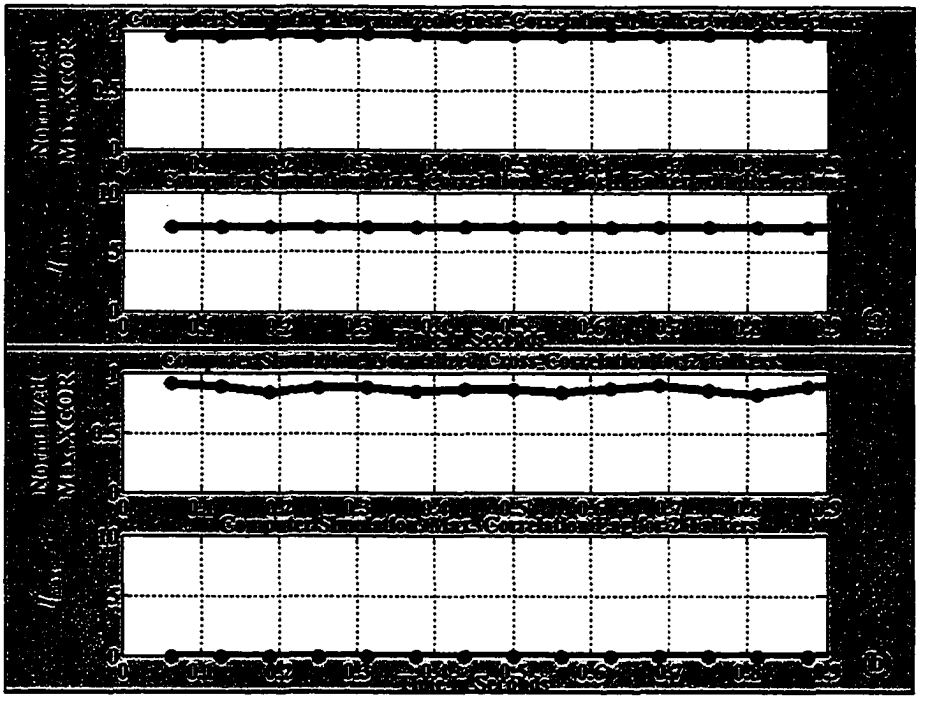

Figure 5-13. Computer simulation results for the reflection / talker detection method; (a) single talker with reflections, and (b) two individual talkers.

5.2 Talker Localization Using Video Information

In the previous chapter, we discussed how to locate the talker using audio information alone. Instead of relying on the audio data, some localization systems take a different approach and rely only on the video information to locate the talker [BIR98], [COL99], [FER01], [HSU02], [SAB98]. In this chapter, we will discuss the advantages, the limitations, and the implementation of some of these methods.

In this thesis, two video localizers are used. One is based on motion detection, which identifies movements, and the other on skin-color detection, which identifies objects with skin-like color. All image processing techniques used in this thesis adopted the

coordinate system where the upper left hand corner is (1,1), the x-direction is along the width of the image and the y-direction is along the height of the image as shown in Figure 5-14. Similar to the audio talker localization, the video data is separated from the audio data in the beginning and video localization is done using video data alone with fusion being introduced later in the thesis.

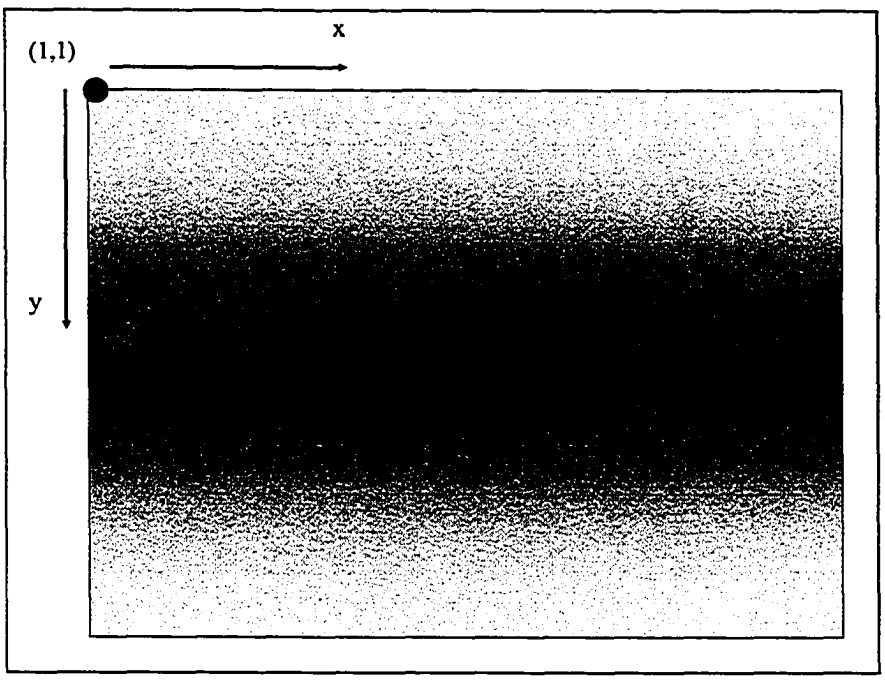

Figure 5-14. Image coordinate system used in this thesis.

5.2.1 Automatic Area-of-Interest (AOI) Identification

When processing a video frame, images often need to be segmented so that the objects of interest can be isolated. Isolating the AOIs helps eliminate unnecessary computations during image processing. There is more than one way to identify AOIs in an image. These methods vary widely and they are still an active area of research in the image

processing field. Two of the more popular methods are defining the AOIs manually on the first video frame and then let the system track the AOIs in the subsequent video frames [LOC94], and defining the AOIs automatically using a human visual system (HVS) model [AGA03] which attempts to mimic how humans direct their focus of attention to areas of high visual interest [OSB98]. In a commercial conferencing system, user intervention decreases the usability of the system, therefore, defining the AOIs manually is not desirable. HVS models allow defining the AOIs automatically. However, HVS models are complex and have high computational requirements, making real-time implementations difficult. In this thesis, a new automatic Area-of-Interest (AOI) identification algorithm is developed to identify patches of pixels where their content within can potentially be of interest. The proposed algorithm uses pixel histogram to find the boundaries of the AOIs. It has the advantage of low computational requirement and operates automatically. The following outlines the steps taken to detect the AOIs.

After a video frame has been processed by a chosen image processing method, such as motion detection and color detection, the video frame is then binary thresholded. The resulting image has black background, pixel value of zero, with the detected objects as white foreground, pixel value of 255. A line-by-line pixel count is performed along the x and then y direction using

$$\begin{aligned} Hist(j)_x &= \sum_{i=1}^{width} V(i,j) \quad ; j = 1...height \\ Hist(i)_y &= \sum_{j=1}^{height} V(i,j) \quad ; i = 1...width \end{aligned} \tag{5-11}$$

where $Hist(j)_x$ is the histogram generated by counting along the x direction, $Hist(i)_y$ is the histogram generated by counting along the y direction, $V(i,j)$ is the video image with video pixel at location (i,j), and *width* and *height* is the corresponding width and height of the video image.

The order of the direction of the pixel counting is not important and will yield the same results. The results of the pixel counting are histograms, one for the x direction and one for the y direction. The histograms are then binary thresholded again. The uniqueness of this automatic AOI identification algorithm is the use of the pixel histograms to find the boundaries of the AOIs. A low to high transition in the histogram signifies the beginning of an AOI and a high to low transition signifies the ending of an AOI. By combining the beginnings and the endings in the x and y directions, a bounding box which encloses large cluster pixels can be formed. These boxes are defined as the AOIs. The total number of the foreground pixels within an AOI has to pass a threshold value before it will be accepted as a valid AOI. Figure 5-15 demonstrates the process of the detection of AOIs using motion detection as an example. Nevertheless, the method is equally applicable to any other methods in which the result can be binary thresholded.

The proposed automatic AOI identifying algorithm works well for images with objects that are tightly clustered after image segmentation was performed. However, when an object is spread sparsely over the image, the algorithm has a tendency to use multiple AOIs instead of a single AOI to enclose the object. If that is the case, different

segmentation techniques like the continuous contour edge detection [IAN96] method can be used to remedy to problem.

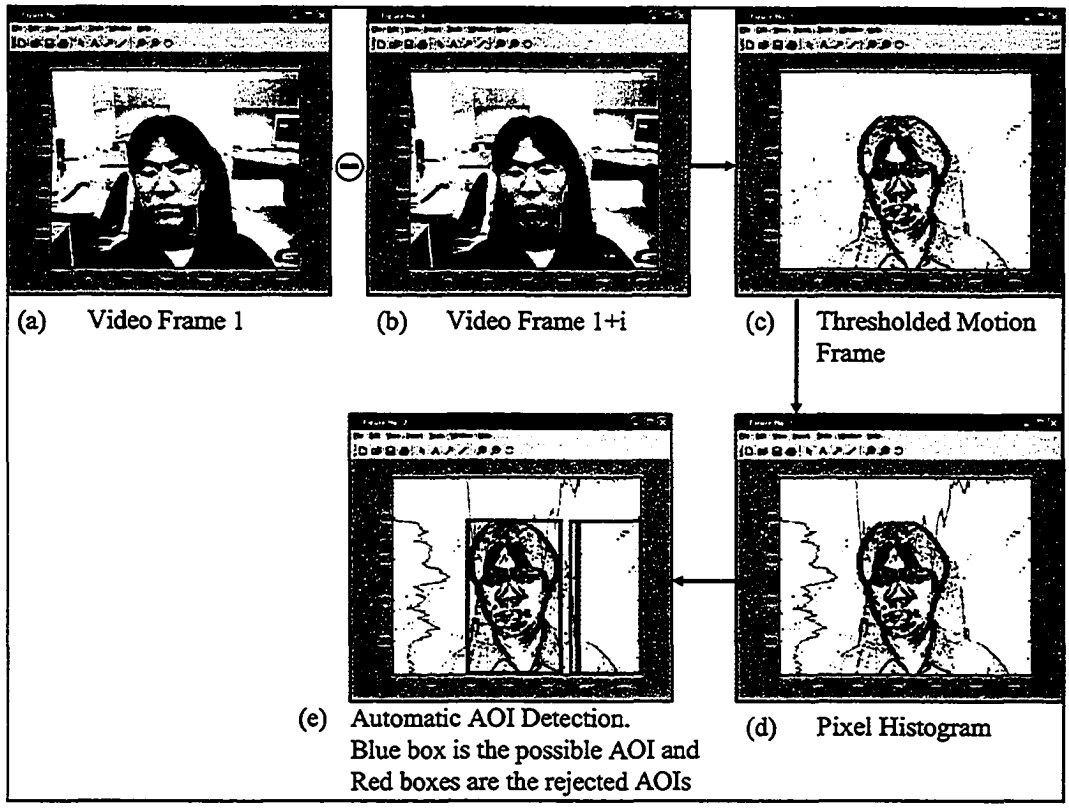

Figure 5-15. Automatic AOI detection example. Blue box in (e) (dark color for black and white printout) is the detected AOI and the red boxes (light grey for black and white printout) are the rejected AOIs.

5.2.2 Motion Detection

The motion detection localizer identifies movements of the talker. Like identifying AOIs, there is more than one way in motion detection, and these methods vary widely as well. Frame subtraction [PIN99][TOY00] and optical flow [MEI99] are two of the well established methods used in performing motion detection. Frame subtraction operates based on the assumption that if a camera is perfectly stationary, changes between

subsequent image frames can only be caused by the motions of objects. Optical flow operates based on the assumption that if an object is moving with constant velocity, the path and the direction (i.e., motion vector) of the image pixels reveals information about the object's distance from the camera [ARC95]. Frame subtraction has the advantage of being simple and easy to implement. However, it requires the camera to be perfectly stationary, and has low accuracy in measuring motions that are moving directly towards or away from the camera. Optical flow has the advantage of allowing the camera to move, and can measure motions in all directions. Since the experimental setup in this thesis used a fix camera, background subtraction is chosen to perform motion detection because of its simplicity.

To perform background subtraction, 24-bit RGB color video frames are first converted to 8 bit grayscale video frames according to ITU-R BT.709 standard using

$$\text{Grey} = 0.2125 * \text{Red} + 0.7154 * \text{Green} + 0.0721 * \text{Blue} \qquad (5\text{-}12)$$

The i^{th} grayscale frame is then subtracted from the $(i+k)^{th}$ grayscale frame, where k is used to control the time interval between frame comparisons which will, in part, depend on the video frame rate. The resulting difference frame is pixels that have changed.

$$D_{(x,y)} = F(i)_{(x,y)} - F(i+k)_{(x,y)} \quad ; x = 1...width; y = 1...hight \qquad (5\text{-}13)$$

where $D_{(x,y)}$ is the resulting difference frame, $F(i)_{(x,y)}$ is video frame at time index i, $F(i+k)_{(x,y)}$ is video frame at time index $(i+k)$, (x,y) is the location of the pixel, and *width* and *height* are the respective width and height of the video image. Again, binary thresholding is applied to the resulting difference frame. The AOI is identified using the

automatic AOI identification algorithm mentioned in the previous section. The AOI provides a bounding box where motion has occurred.

When more than one object is moving in the video scene, multiple AOIs will be detected. Motion detection can only identify where motions happen but it cannot distinguish whether the source of the motions is a talker or something else. Therefore, using just motion detection alone is not sufficient to localize the talker accurately.

5.2.3 Skin-Color Detection

Using color images, skin-color detection can be used to identify objects with skin-like color [HSU02][TER00]. Color can be represented using different color spaces. The most popular color space used by digital storage is the Red-Green-Blue (RGB) color space. Although variations in the RGB space like the R-G space has the advantage of reducing the sensitivity of segmentation to the changes in amount of light [TER00], and is often used in skin-color detection, RGB color space spreads skin-color pixels over a large range making the detection difficult [TER00]. It has been shown that color analysis done in luma-chroma space, such as the YCrCb, concentrates the skin-color pixels in a tight range [TER00] as shown in Figure 5-16. Therefore, a luma-chroma space is well-suited for detecting skin color pixel.

Hsu *et al.* [HSU02] used 137 images from nine subjects in the Heinrich-Hertz-Institute image database to define the skin tone cluster in the YCrCb space. They use a non-linear

model to compensate the luminance in low light and then fitted an ellipse to the skin tone cluster [HSU02]. To reduce computational complexity, this thesis uses a system of eight linear equations, Equation (5-14), to enclose the skin tone cluster and the fitted ellipse as shown in Figure 5-16.

$$
\begin{aligned}
&\text{Line 1: } Cr \geq -0.702 \bullet Cb + 209.945; \\
&\text{Line 2: } Cr \leq -2.12 \bullet Cb + 420.184; \\
&\text{Line 3: } Cr \geq -5.984 \bullet Cb + 669.481; \\
&\text{Line 4: } Cr \geq 2.012 \bullet Cb - 115.329; \\
&\text{Line 5: } Cr \leq 0.7389 \bullet Cb + 99.171; \\
&\text{Line 6: } Cr \geq -0.0333 \bullet Cb + 138.122; \\
&\text{Line 7: } Cr \leq -0.142 \bullet Cb + 183.425; \\
&\text{Line 8: } Cr \leq -0.681 \bullet Cb + 241.713;
\end{aligned}
\qquad (5\text{-}14)
$$

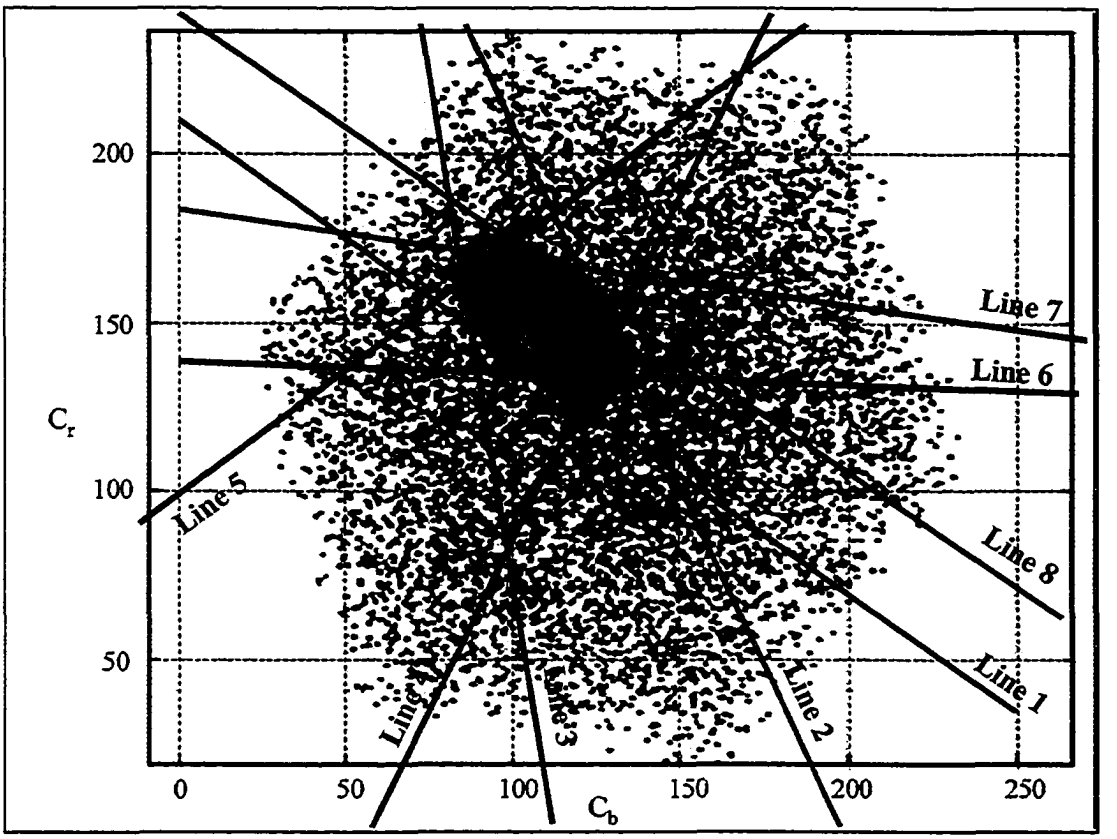

Figure 5-16. A system of 8 equations enclosing the skin pixel area in Cr-Cb color domain (after [HSU02]).

To perform skin-color detection, the 24-bit RGB color video frame is first transformed into the CCIR601-4 YCrCb color space [POY96] using (2-4). Equation (5-14) is then used to determine the thresholding values for detecting skin-color pixels. Each pixel is checked using

$$\begin{cases} \text{if}(Cr_{(x,y)} \geq TestCr1) \,\&\, (Cr_{(x,y)} \leq TestCr2) \,\&\, (Cr_{(x,y)} \geq TestCr3) \,\&\, (Cr_{(x,y)} \geq TestCr4) \,\&\, (Cr_{(x,y)} \leq TestCr5) \\ \&(Cr_{(x,y)} \geq TestCr6) \,\&\, (Cr_{(x,y)} \leq TestCr7) \,\&\, (Cr_{(x,y)} \leq TestCr8) = True \Rightarrow pixel_{(x,y)} = 1 \\ \text{else } pixel_{(x,y)} = 0 \end{cases} \quad (5\text{-}15)$$

where $Cr_{(x,y)}$ is the r-chrominance value of the pixel located at (x,y), $pixel_{(x,y)}$ is the resulting skin-color mask, *TestCr1* is the r-chrominance value computed using the equation of Line 1 in (5-14), *TestCr2* is using the equation of Line 2 in (5-14) and so on. The resulting image is a skin-color pixel mask (Figure 5-17(b)). To clean up the mask, morphological closing is applied to emphasize large groups of pixels such as faces (Figure 5-17(c)). Often, small holes exist within a large group of pixels, such as the eyes and the mouth. Therefore, a flood fill operation is used to fill out any small holes in it (Figure 5-17(d)). Morphological opening is then used to eliminate any small cluster of pixels and background noises (Figure 5-17(e)). The resulting mask represents large skin-color objects such as faces. The mask is then multiplied with the original image giving only objects with skin-color (Figure 5-17(f)). AOIs are then identified using the automatic AOI identifier (Figure 5-17(g)). Figure 5-17 demonstrates the process.

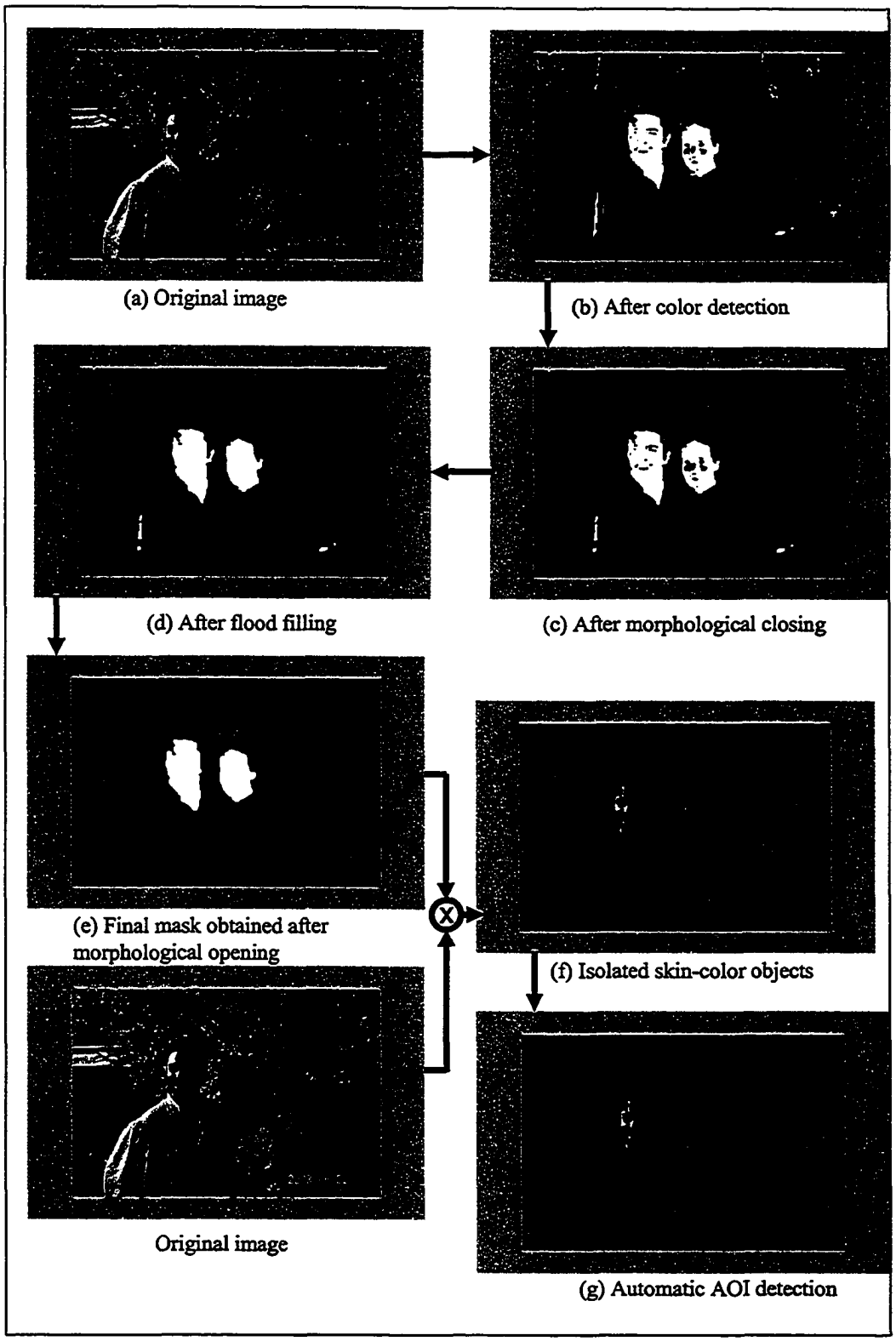

Figure 5-17. An example of skin-color detection.

The outlined skin-color detection method is robust and works well in most well illuminated images of the upper body and full body that occupied big portion of the frame. However, localization methods based on skin-color can only identify the human faces in the image; it has no ability to distinguish who is the active talker. Furthermore, if there are pictures or posters with human faces in the video scene, the skin-color detection method will treat them just as another potential human talker.

5.2.4 Automatic White Balancing

Different lighting conditions can cause the camera to bias the color of an object [HSU02]. This can affect the accuracy of the skin-color detection algorithm. Often, white balancing is used to rectify the problem. White balancing in this thesis is done using a method similar to that of Hsu [HSU02]. Most video scenes contain some white color pixels [HSU02]. White color has the characteristic of a very high luminance value. Therefore, the top 3% of pixels with luminance value larger than a threshold are treated as white. The red, green and blue values of these pixels are then adjusted to pure white accordingly (pixel value of (255,255,255)). These adjustments are then applied to the rest of the pixels in the video frame.

5.2.5 Camera's Field of View to Active Sector Mapping

In a video conferencing system, the camera and the microphone array use a different frame of reference to localize objects. The camera usually uses the room or a fixed point

in the room as the reference point [LOC94]. A video localizer, that uses the camera, will then report the relative position of an object with respect to the reference point. On the other hand, the microphone array uses itself as a point of reference and reports all acoustic localizations relative to the locations of the microphone array. Therefore, localization results from each of the localizer will need to be mapped to a common frame of reference and that is the function of the *Mapping* block in the general multimodal sensor fusion architecture shown in Figure 4-1. If the video and the microphone array are co-located, using two different frames of reference cause little problem since the relative location between the two is fixed and is known. Locations from one reference frame can be mapped to the other one or vice versa. For example, the simplest way to co-locate the camera and the microphone array used in this thesis will be by setting up the camera in such a way that the principle axis for the camera's pan is aligned with the center of the microphone array, so that every 30° of the camera's pan is mapped to a microphone array sector.

In this thesis, the camera and the microphone array are placed separately in the conference room. Since the spatial resolution of the microphone array is less than the camera, the field of view of the camera is mapped to the regions defined by the microphone array sectors as shown in Figure 5-18 so that the locations of the active talker are reported using a common frame of reference. The sectors in which the AOIs fall on are identified as the active sectors. A calibration run performed before the start of the experiment provides the locations of these sectors.

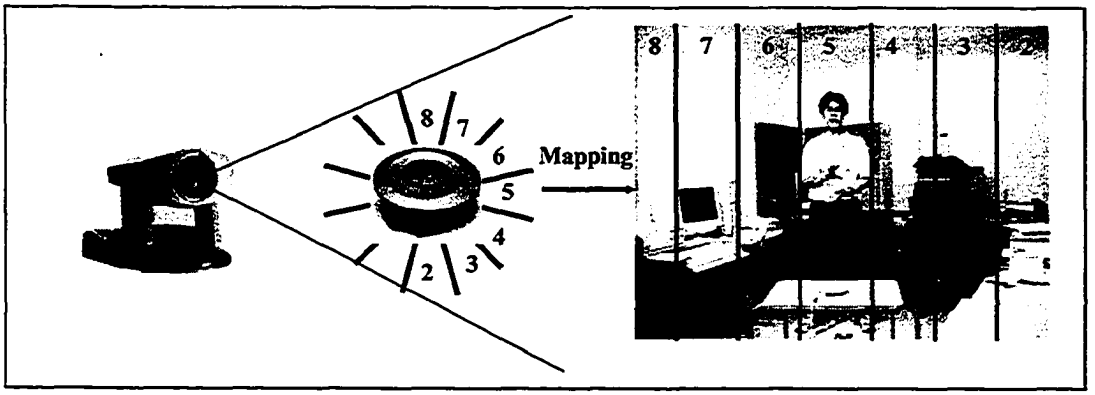

Figure 5-18. Camera's field of view to microphone array active sector mapping.

Video localization methods do not suffer from acoustic reflections, but they do fail when the lighting conditions in the scene change drastically [HSU02], when there is a complex background, and when other people enter and leave the video scene.

Localization methods that use only audio or only video rely on one type of data, and are prone to failure. Multimodal localization takes advantage of informational diversity from multiple sources, giving a more robust localization [STR01]. The following two chapters discuss various multimodal approaches used in this thesis.

5.2.6 Microphone array and camera locations self-discovery procedure, and camera self-calibration procedures

5.2.6.1 Discovery Procedures

The purpose of the self-discovery process is to locate the microphone arrays and cameras which are placed arbitrary in the conferencing environment. The system can include one or more cameras detecting one or more microphone arrays. The discovery process relies on finding markers that are placed on the microphone arrays and cameras in advance. These markers can be anything that is visible to the camera and it can be a single marker or more then one. Each marker will have its own signature and the signature is known in advance. For example, LEDs (or IR-LED) can be placed on the microphone arrays as the markers and the signature will be a special flashing sequence or simply LEDs in different colors. With the known marker signature, the camera will be instructed to do a search for that specific marker signature in the conferencing environment. The marker can be, for example the IP address of the device. Once the marker is located, the camera can "zoom-in" to carry out further confirmation. The confirmation can be in the form of identifying a different marker's signature, which the markers were instructed to send, or just simply identifies some known physical features of the device.

In order to implement the above outlined discovery procedures, a microphone array is fitted with an infrared LED that pulsates at a frequency *Fp* for time interval *Tp* carrying specific information, for example the array's internet protocol (IP) address. The LED is placed in an asymmetric location on the array. The camera detects the presence of the microphone array by searching for the infrared LED on the microphone array. The LED shows up as bright white light in CCD camera's image. The following procedure can be used to perform the task:

1. Set the value of *Fp* and *Tp*. The detection speed is determined by the frequency *Fp*. However, increasing *Fp* increases the video processing computational requirement. Typical values of *Fp* range from 5 to 15 Hz and *Tp* in the range of 20-50% duty cycle.
2. The frame capture card is setup to capture video at frame rate twice of *Fp*.
3. First, make sure nothing is moving in the background. Note that it is crucial to make sure there is no motion in the field of view of the camera during this step. The camera is then instructed to do a stepping panning search for the pulsating light with minimum of two frames captured at each step.
4. Because the frame capture card is capturing the video at twice the speed of *Fp*, if the pulsating LED is indeed in the field of view of the camera, a lit LED will be seen in one of the video frames but not both. Frame subtraction can then be used to detect the presence of LED by subtracting one frame from the other. The resulting frame will be the pixels representing the location of the LED.
5. Using the pan and tilt action of the camera, the LED is centered in the field of view. The camera is then zoomed-in.
6. Step 4 and 5 are repeated iteratively until satisfactory optical resolutions of the markers are obtained. The identity of the microphone array is confirmed by the marker's unique signature.

5.2.6.2 Camera Calibration Procedures

The purpose of the self-calibration process is to relate, regulate and standardize dimensions and locations in the conferencing environment to the video system. On each device there is a set of calibration markers with known geometry and dimension. These markers can be the same set used in the self-discovery process, a completely different set or combination of both. The camera will then be calibrated using these markers.

To calibrate, the camera is instructed to zoom-in to the device identified by the self-discovery process. The intrinsic physical parameters of the camera: focal length, principal point, skew coefficient, and the lens distortions will be found by auto-calibration means. This can be done by instructing the camera to observe the markers, capture the image, and identify where the markers are. The camera is then to pan and tilt slightly and then the observation process will be repeated. The intrinsic parameters of the camera can then be calculated. If the camera has the ability to auto-focus, the focal length of the camera can change depending on the video content. To compensate for these changes, the current setting of the camera lens can be read off from the camera and then mapped to the focal length calculated from the intrinsic parameters. When the camera changes its focus, the new value of the lens setting can be read off from the camera and then back calculates the correct focal length. Using a single camera can result in the loss of depth perception. Fortunately, because we know the dimension or the markers, the depth info can be recovered by calibrating the image size of the markers (in pixel) to the actual size of them.

Figure 5-19 outlines a typical scenario with the microphone array on a tabletop with the camera on an elevation and is tilted towards the microphone. During the calibration, the distance and the orientation of the array will be found using image processing techniques. The pan angle, the tilt angle, zoom factor and focal length will be read from the camera control. The following procedure is an example how it can be done.

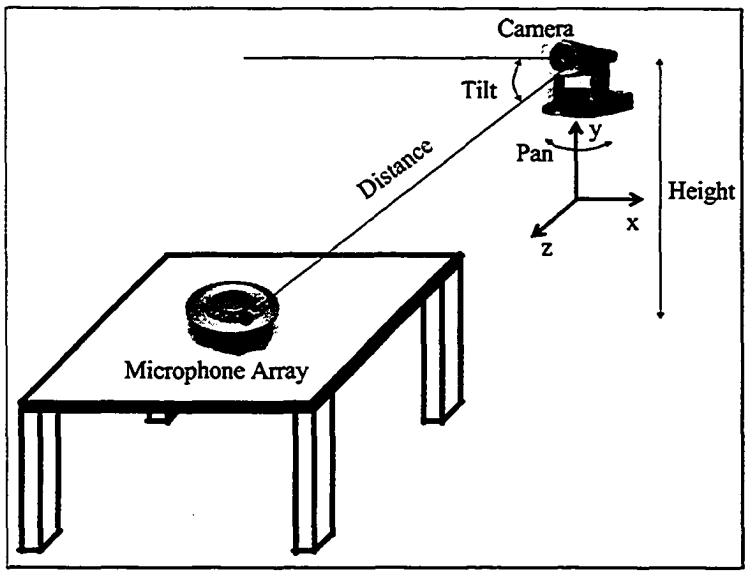

Figure 5-19. Typical Scenario for Microphone Array and Camera Setup.

1. Turn the captured color video frame into a grayscale video frame using equation (5-12).
2. In order to determine the edge of the circular microphone array, spatial domain edge detection methods like the Sobel gradient filter [LOW91] or the equivalent frequency domain high-pass filter [LOW91] can be used.
3. Objects closer to the camera will appear to be larger and objects further away from the camera will appear to be smaller. This size to distance relationship is an intrinsic property of the video system and depends only on the setting of the zoom

factor, focal length of the camera and the lens used. A calibration can be done beforehand by presenting objects of know size to the video system at known distances with different focal lengths. The object size-to-distance relationship is assumed to be linear. A three way look-up table can be created for the size/distance/focal length.

4. Edge detection gives the outline of the microphone array. Since the camera is viewing the microphone array at an angle, the circular outline will appear elliptical. The long axis of the ellipse is always perpendicular to the viewing axis of the camera; therefore, this length is also the true length of the array's diameter. With the known size of the array, the look-up table can be used to recover the distance of the array from the camera.
5. The camera will then be instructed to pan, tilt, and zoom-in until the long axis of the array is centered and occupying the field of view as much as possible. A new frame will be captured and then step 1 – 4 will be repeated.
6. Since the microphone array reports voice activities in sectors with respect to itself, it is essential to find the orientation of the array as well. There is a LED located off center on the top of the microphone array. The same method used in steps 1-4 in Discovery Procedure can be used to find the LED. Based on where the LED is biased towards, the orientation of the array can be established.

5.3 Talker Localization Using Infrared Information

Infrared imaging (IR) complements the video camera used in the current setup very well. Video cameras rely heavily on visible light. When the level of visible light is low, most

video cameras fail to register any images causing the motion and skin-color detection to fail [HSU02]. Consequently, the system looses the ability to locate the talker visually. Low light situations are common in presentations were the lights are dimmed when computer and overhead projectors are being used. However, IR sensors like IR cameras respond to heat and are not affected by the lighting condition. The IR camera will still allow the system to locate the talker visually regardless of the lighting condition in the room. Since most IR cameras map temperatures as grayscale and then assigning pseudo-colors to the grayscale, simple binary thresholding works very well as a detection method. This chapter outlines how IR imaging can be used as a talker localization modality.

IR cameras and sensors have been used successfully in multi-sensor applications for the purpose of navigation and object tracking in the area of robotics [PET96] [YEU94] and surveillance [DAV97]. However, the deployment of IR camera in commercial videoconferencing systems is limited. With the falling price of IR cameras, adding IR cameras in conferencing systems is becoming more feasible. Part of this thesis investigates how IR cameras can be added to a video conferencing system using the modular multimodal localization architecture shown in Figure 4-1. In this chapter, we will discuss how IR imaging can be used as a standalone thermo-graphic localization modality. In Chapter 8, we will discuss how IR imaging can be used in conjunction with other localization modalities, such as audio and video localizers used in this thesis, to locate talkers in video conferencing applications.

5.3.1 Thermo-graphical Detection in IR Images

Heat sources in the video conferencing environment like human bodies radiate more infrared energy than the background. When this infrared radiation is capture by an IR camera, it can be seen as thermo-graphic images. The heat sources will standout from a relatively uniform background. A typical IR image of talkers is shown in Figure 5-20. The two talkers in the foreground are clearly distinguishable from the uniform background. These properties make a thresholding method a good candidate for detecting human bodies thermo-graphically in IR images.

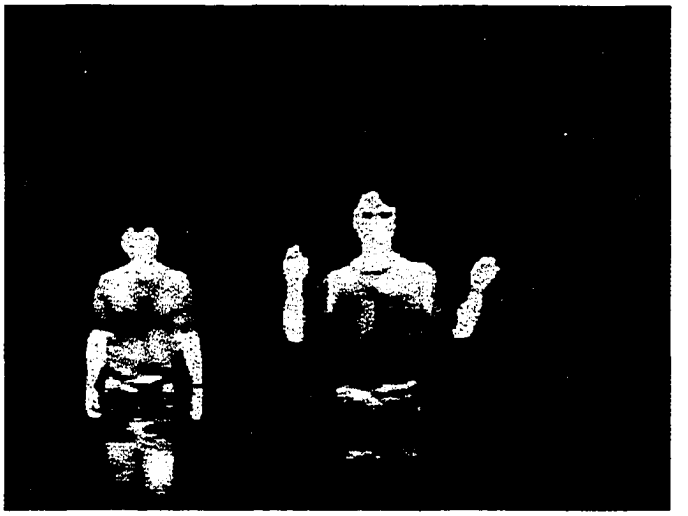

Figure 5-20. Example of infrared image.

As mention earlier, most IR cameras map measured temperatures to a range of grayscale values and then apply pseudo colors to the image for easy viewing. Therefore, only the grayscale values or the luminance values carry useful information. In this thesis, a simple binary thresholding method is used for detecting human bodies. To thermo-graphically

detect an object in the IR images, a calibration image is captured with the temperature calibration bar added to the image as shown in Figure 5-21. This calibration is generated by the internal electronics of the camera showing how the measured temperature is mapped.

Figure 5-21. Calibration infrared image with a calibration bar on the right.

Using this calibration bar, a pixel value corresponding to a predefined threshold temperature is found. In this thesis, a threshold temperature of 25 - 28 °C is used. The threshold temperature value is chosen based on the assumption the conferencing environment is an air conditioned indoor room. Therefore, most non-heat generating background will be around 18 – 22 ° C. Using the pixel value histogram, the talker will

be shown as peaks in the high value range and the background will be shown as large peaks in the low value range as shown in Figure 5-22. With this distinction, a threshold value which is above the background and below the talker can be found. In most cases, the temperature of a lightly clothed human body will measure in the range of 25 – 30 ° C in the experiments conducted in this thesis.

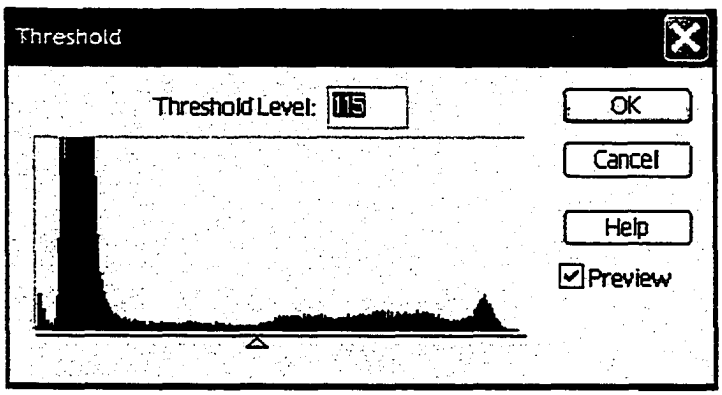

Figure 5-22. Pixel histogram of IR image.

Before thresholding can be applied, image pixels are first converted from color to greyscale using equation (5-12). Pixels with values above the threshold are set to black (0) and pixels with values less than the threshold are set to white (255). Figure 5-23 shows an example of a thermo-graphically detected image using the binary thresholding method just mentioned.

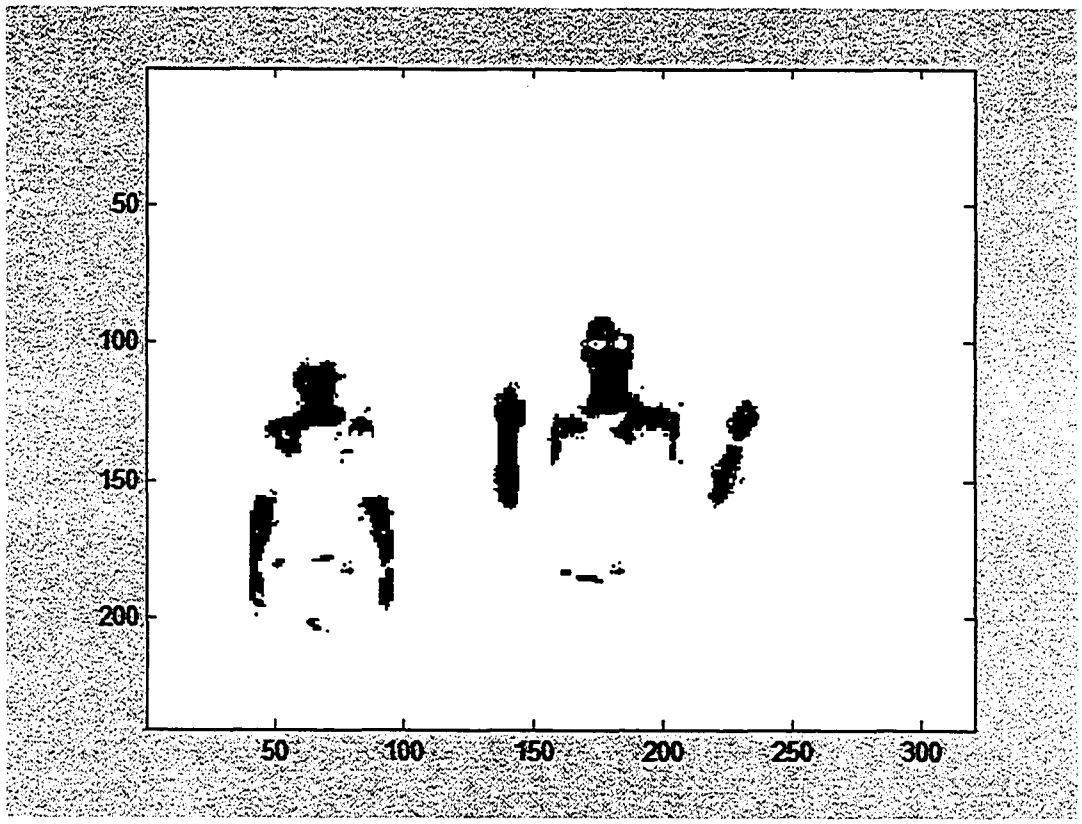

Figure 5-23. Thermo-graphically detected IR image with 27 °C binary thresholding.

5.3.2 AOI Identification and Camera's Field of View to Active Sector Mapping

The IR camera used in this thesis output IR images in the NTSC format. Therefore, from an imaging processing point of view, the output of an IR camera is very much like just any other video camera and can be processed the same way. Some of the software and algorithms developed for the video localization like the automatic AOI identification routine (Chapter 5.2.1) and the camera's field of view to active sector mapping method (Chapter 5.2.5) can also be applied to the IR images without any modification. To find the AOIs, the automatic AOI identification routine is applied to the thermo-graphically

detected image, like the one shown in Figure 5-20. Figure 5-24 shows an example of the identified AOIs. Figure 5-25 shows an example of how the camera's field of view is mapped to detection sectors.

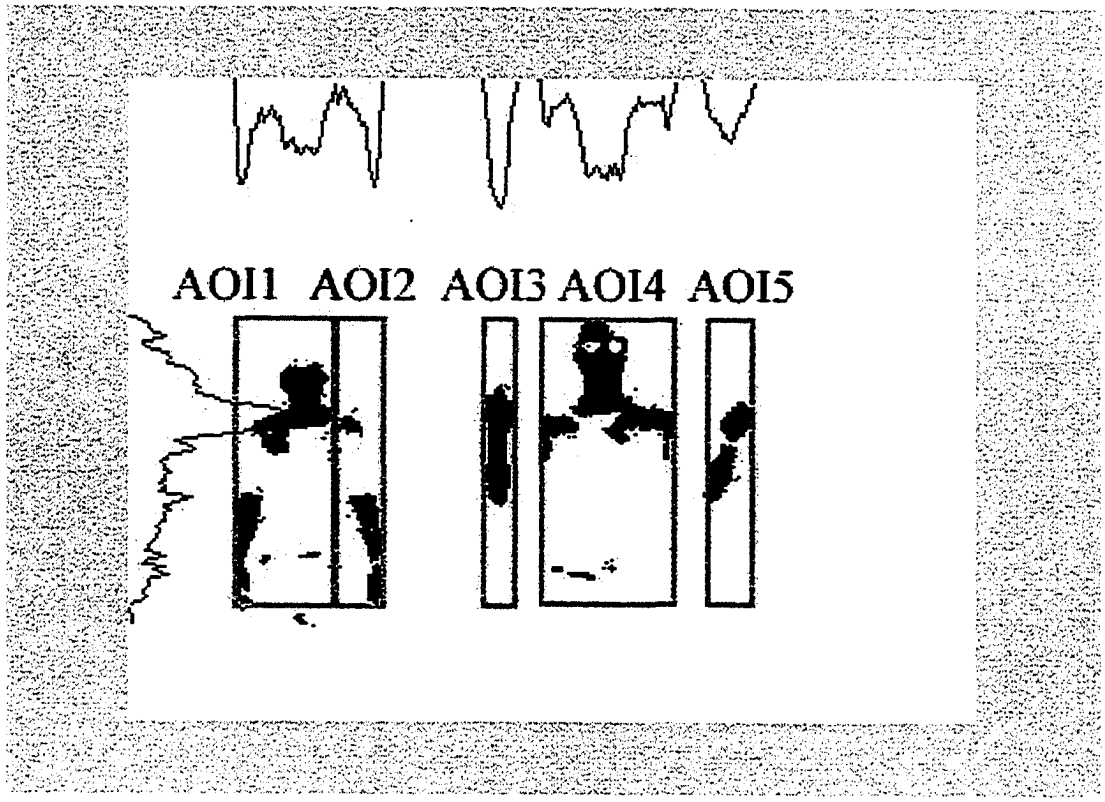

Figure 5-24. Thermo-graphically detected objects with AOIs.

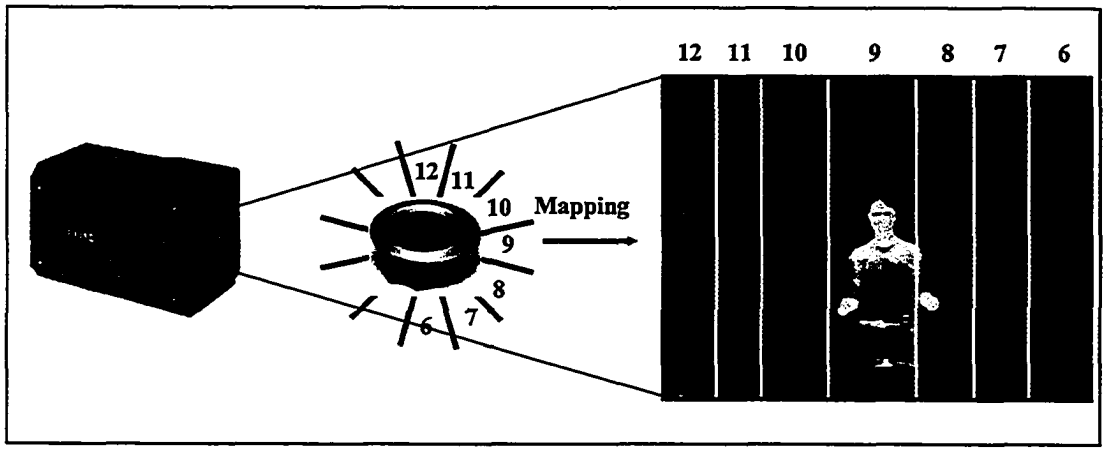

Figure 5-25. IR camera's field of view to detection sectors mapping.

The thermo-graphic detection method outlined above works well for video conferencing conducted in a well controlled indoor room. It has the advantage of being simple and fast. Unlike video localization, IR is not affected by lighting conditions nor will it be confused by a poster on the wall. However, the outlined thermo-graphic detection method does not distinguish an active talker from other participants in the background or other heat sources like a furnace's air register or warm computer monitor. In Chapter 8, we will see how this simple thermo-graphic localization method can be combined with other localization methods to form a more sophisticated localization system.

Chapter 6 Multimodal Talker Localization Using Joint Audio-Video Information

Using only audio or only video for localization is prone to failure [LOD03A], [LOD04]. Joint audio-video localization takes advantage of the complementing nature of the two sources, giving a more robust localization [BRA01]. Using the general multimodal talker localization architecture shown in Figure 4-6, a joint audio-video talker localization system can be designed as shown in Figure 6-1. The system consists of an audio localization module, a video localization module, and a data fusion module and decision module. A modification has been added to the video localization module. Instead of having only one localizer, a second localizer is added. In this particular design, motion detection and a skin-color detection localizer are used in the video localization. The advantage of adding a second localizer is that the performance gain is similar to adding a whole new video localization module without incurring the cost of an extra camera and the effort of synchronizing multiple video sources. However, since both video localizers are fed off from the same camera signal, when the camera malfunctions, both localizers will fail simultaneously.

6.1 Joint Audio-Video Talker Localization System

The block diagram of the joint audio-video talker localization system is given in Figure 6-1. Inheriting the properties of the general architecture, the audio and video data are decoupled in the beginning and localizations are done with each modality separately. In

order to have a common coordinate system that both the audio and video localizers can refer to, the space at which the talker moves is segmented into 12 sectors and then mapped to the microphone array sectors as done in Chapter 5.2.5. Analysis is done to estimate the correctness probability ($P_{(m,n)}$) of the current output for each localizer, where m is the localization method, which can be audio beamforming, motion, or skin-color detection, and n is the active sector. For each sector, the correctness probabilities from the audio and video localizers are combined using data fusion techniques.

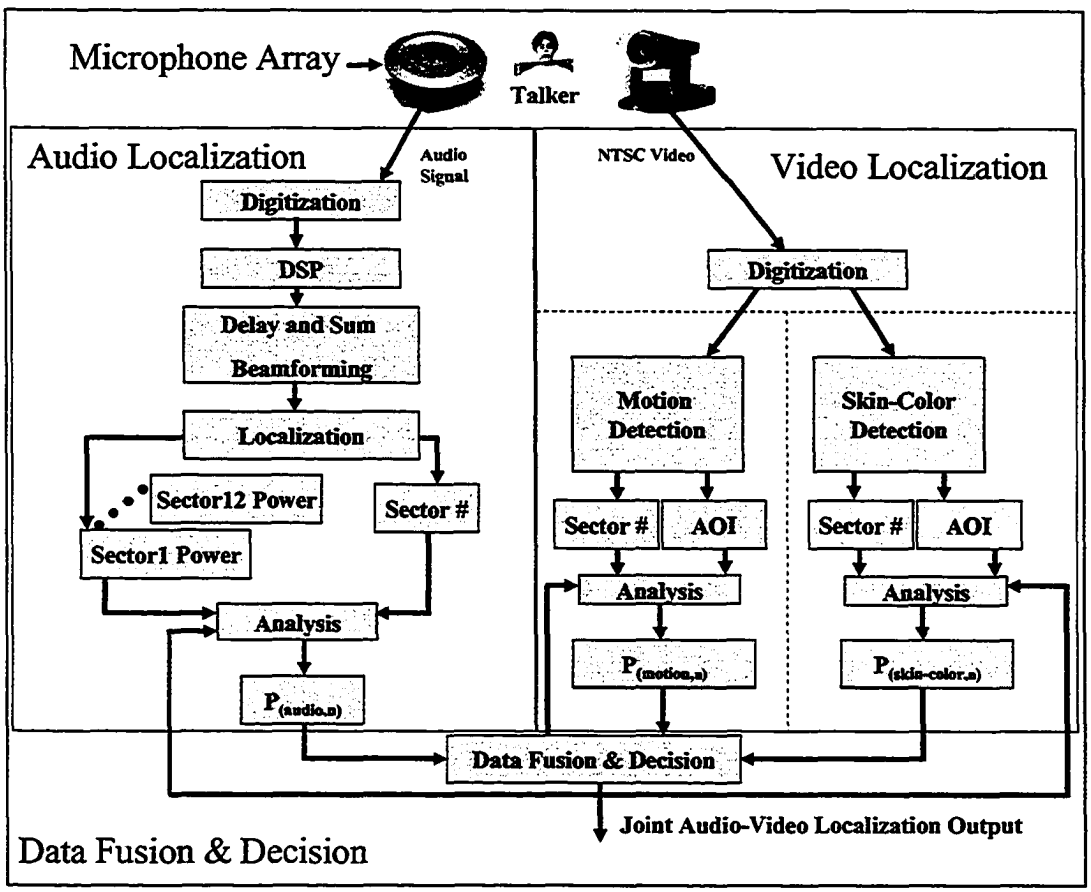

Figure 6-1. Joint audio-video talker localization system block diagram.

6.2 Joint Audio-Video Talker Localization Using Summing Voter Fusion

As outlined in section 5.1, audio localization is done using a delay-and-sum beamforming circular microphone array. Video localization is done using a motion detection localizer and a skin-color detection localizer. The motion detection localizer identifies movements of the talker while the skin-color detection localizer identifies objects with skin-like color. In order to estimate the trustworthiness of the current detections, the correctness probabilities $P_{(m,n)}$ is calculated using Equation (4-1). For convenience, Equation (4-1) is repeated here;

$$P_{(m,n)}[i] = \frac{\sum_{k=i-td}^{i} D_n[k]}{\sum_{s=1}^{N} \sum_{k=i-td}^{i} D_s[k]} \tag{6-1}$$

where $D_n[k]$ is the number of detections in sector n at time k and td is the width of the window of time to look back to from the current data point, N is the number of sectors, and m is the detection method which can be the audio beamforming, motion, or skin-color detection. For example, if td is 20 samples, N is 12, and the motion detection localizer identifies sector j as active 5 times in the last 20 detections, using Equation (6-1), $P_{(m=\text{motion}, n=j)} = \frac{5}{12 \cdot 20} = 0.021$.

Once all the $P_{(m,n)}$ are computed, a summing voter is chosen to fuse the different localization results by applying Equation (4-2),

$$K_n = \sum_m P_{(m,n)} = P_{(audio,n)} + P_{(motion,n)} + P_{(skin-color,n)} \tag{6-2}$$

where m is the localization method, n is the currently detected active sector, K_n is the fused result for sector n, and $P_{(m,n)}$ is the probability of method m detecting sector n as active. Summing voters have the advantage of being simple, have low computational requirements, and can be easily adapted to incorporate modifications. Since three detection methods are used, m can be audio beamforming, motion detection, or skin-color detection. Majority rule is used as the decision logic and the sector with the highest K_n is taken as the fused localization output.

6.3 Experiment — Joint Audio-Video Talker Localization Using a Summing Voter

Recordings were done in a 3.8 m x 5.4 m x 3 m reverberant room. The microphone array was placed on a 1.5 m x 3 m conference table. The video camera was placed two meters away from the microphone array. Acoustic data were sampled and source localization was done using delay-and-sum beamforming. The video data were digitized at 320x240 pixels, 15 frames per second. Audio/video synchronization was done by hand clapping, which both the microphone array and the camera registered. The experiment was done with the talker standing stationary at one meter in front of the microphone array, giving a presentation. Audio and video localization disturbances were introduced during the experiment. Audio disturbance was introduced as the talker directed his voice toward the sidewalls causing strong acoustic reflections. Although motion detection and skin-color detection do not suffer from acoustic reflections, they too can fail in locating the talker. Motion detection is especially susceptible to other motion in the background. Similarly,

skin-color detection is especially susceptible to complex background scenes and changes in lighting conditions [HSU02]. Subsequently, two forms of video disturbances were introduced into the experiment: i) another person sat and performed tasks in the background, and ii) the lights were dimmed. Since the range of the talker's motion was limited, a fixed camera was used to simplify the experiment. The camera was set up so that the field-of-view would capture sectors 4 -10, as shown in Figure 6-2. Before the experiments were started, initialization runs were done to calibrate the mapping required for relating the locations of the sectors and the camera's field of view.

Six different scenarios, as outlined in Table 6-1, are considered in the experiments. These 6 scenarios represent different permutations of the audio and video localization disturbances the localization system can potentially face. In order to compare the performance between different sensor fusion techniques, the total localization error rate for scenarios (1) – (6) and the error rate for just scenario (6) are computed. The error rate for scenario (6) is especially of interest because it represents the tracking error rate, and it reflects the dynamic behavior of the system. The error rate for a given time frame t_{start} to t_{end} is computed as

$$Error\ Rate = \frac{\sum_{i=t_{start}}^{t_{end}} t_{error}(i)}{(t_{end} - t_{start})} \qquad t_{end} > t_{start} \tag{6-3}$$

where t_{error} is the duration of the localization error.

Table 6-1. Scenarios for audio and video disturbances.

	Scenario
(1)	Talker giving a presentation in front of the camera and directing his voice toward the microphone array.
(2)	Talker giving a presentation in front of the camera and directing his voice toward sidewalls causing strong acoustic reflections.
(3)	Scenario (2) + Another person sitting in the background at sector 10 facing the camera.
(4)	Scenario (1) + Lights were dimmed.
(5)	Scenario (2) + another person, facing the camera, conducting tasks in the background at sector 10.
(6)	Talker walk around the microphone array in the following sequence: sector 5, sector 4, sector 5, sector 6, sector 7, sector 8

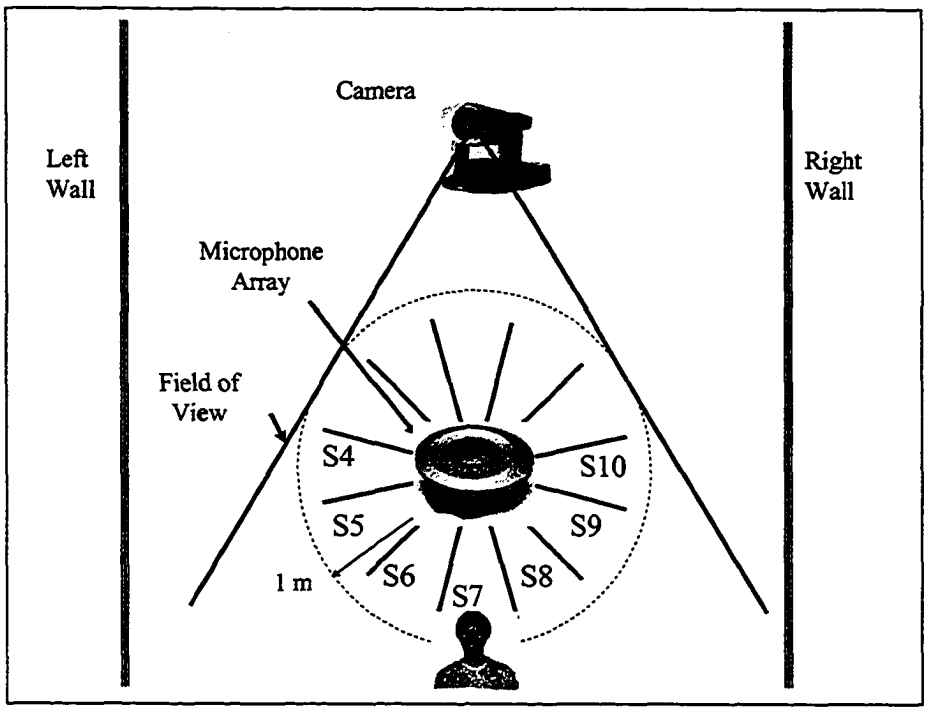

Figure 6-2. Experimental setup for joint audio-video talker localization experiments.

6.3.1 Experimental Results

Figure 6-3 shows the localization results for the six scenarios. The results are concatenated in time and presented in Figure 6-3 with scenario (1) represented by plots from 0 s – 2.9 s, scenario (2) from 2.91 s – 5.84 s, scenario (3) from 5.85 s – 8.86 s, scenario (4) from 8.87 s – 11.94 s, scenario (5) from 11.95 s – 15.18 s, and scenario (6) from 15.19 s – 25.00 s, respectively. These scenarios are also marked as (1), (2), (3), (4), (5), and (6) in Figure 6-3. During scenarios (1) – (5), the talker is always standing at sector 7 with the presence of different disturbances causing different localizers to locate the talker incorrectly. In scenario (6), the talker walked around the microphone array starting from sector 5 toward sector 4, then back to sectors 5, 6, 7, and then 8.

The three plots on the left show the localization output of each localizer when it is used as a stand-alone device. Figure 6-3(a) shows the localization output of the motion detection localizer. Figure 6-3(b) shows the localization output of the skin-color detection localizer. Figure 6-3(c) shows the localization output of the microphone array. Figure 6-3(d) shows the fused localization result using a simple summing voter. For reference, Figure 6-3(e) shows the hypothetical localization output for an ideal localizer which located the talker at sector 7 for scenario (1) – (5), and the walk around sequence of sector 5-4-5-6-7-8 during scenario (6). Since both the motion detection and the skin-color detection localizers are capable of detecting multiple objects, more than one sector can be identified as active simultaneously. In most cases, the motion and skin-color detection do not missed the correct sector completely. Instead, the localization errors will be shown as multiple active sectors. Audio localization errors simply show as discontinuities in the plots. The microphone array localizer and the fused localization outputted only one active sector and therefore, localization errors would show as jumps in the plots.

In scenario (1), the talker was standing in sector 7 and spoke directly towards the microphone array. This would be the "normal" operating scenario. Although the microphone array has a short lived localization error, the summing voter fusion methods reject the error and provide perfect output for scenario (1). In scenario (2), when the talker speaks towards the walls, the acoustic reflections confuse the microphone array and cause it to localize incorrectly. This scenario is very common during conferencing when the talker turns his head to address a question from a participant. Again, the fusion

method combines the localization results from the two video localizers and successfully removes all the audio localization errors to give a perfect output. In scenario (3), the talker speaks toward the side wall while another person, who is facing the camera, was sitting at sector 10 in the background. The motion detection and the skin-color detection recognized the person in the background. The simple summing voter fusion method, Figure 6-3(d), suffers some localization errors. In scenario (4), the lights were dimmed so that only the outline of objects was visible. While the skin-color detection fails, the motion detection still detects the motion of the talker's outline with only a few localization errors. The fusion methods are not affected by the skin-color detection failure and outputted the correct sector number. In scenario (5), heavy acoustic reflections were combined with another person conducting tasks in the background at sector 10, providing a challenging situation. The fusion method suffers the same localization errors but it performed considerably well. In scenario (6), the talker walked around the microphone array from sector 5 toward sector 4, then back to sectors 5, 6, 7, and then 8. The localization error rate for scenario (6) (i.e. tracking error rate) is 23.9% while the total error rate is 12.6%. It can be seen in Figure 6-3(a) that motion detection was less selective and shows strong multiple detections as the talker crossed between sectors; whereas, skin-color detection given in Figure 6-3(b) was more specific and can locate the talker better when he was moving from one sector to the next. The discrepancy can be understood as motion detection identified the motion of the talker's face as well as his body movements, while the results from skin-color detection follows only the face of the talker. The fusion method removes some localization errors but not all.

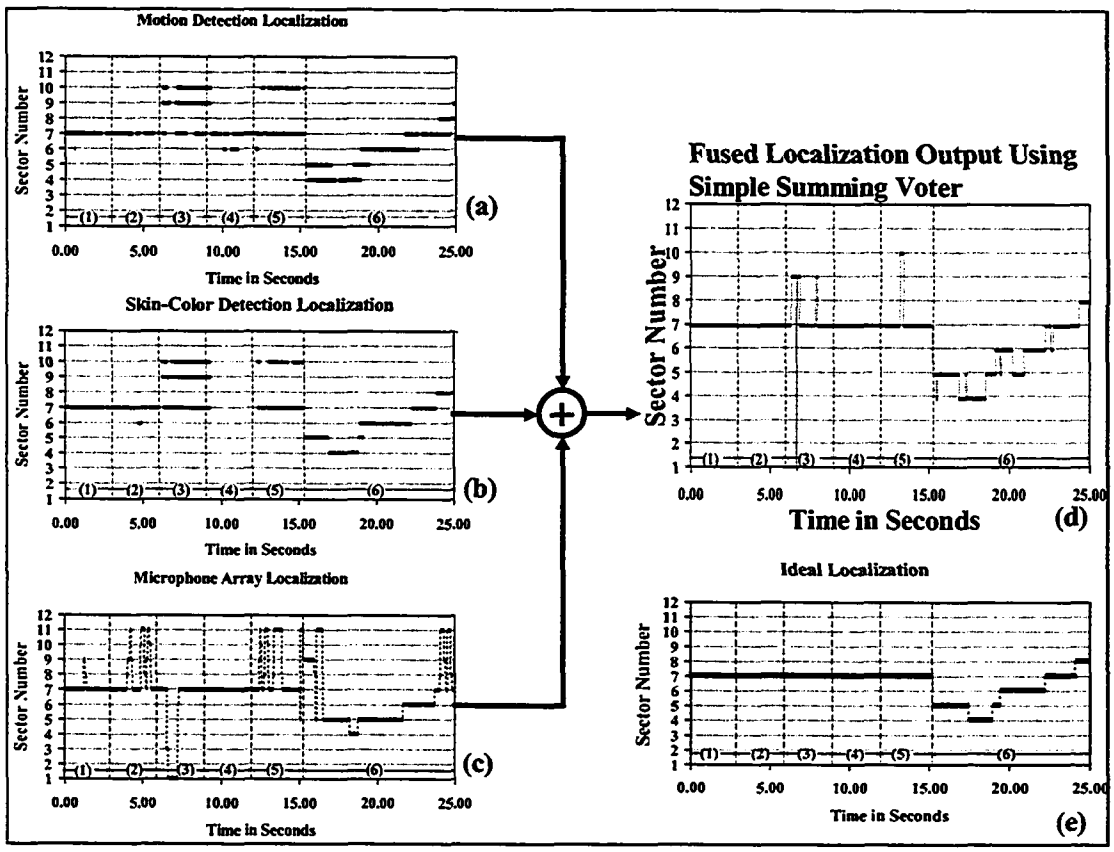

Figure 6-3. Fused localization result, (a) localization from motion detection localizer, (b) localization from skin-color detection localizer, (c) localization from microphone array localizer, (d) fused localization output using simple summing voter, and (e) localization from an ideal localizer.

6.4 Joint Audio-Video Localization Using Bayesian Network Fusion

Bayesian networks are often used in data fusion [PAV00][TOY00][ZOT01]. Performing joint audio-video localization using a Bayesian network is very similar to what is done using the summing voter. The same methods are used to perform audio and video localizations. The main difference is the use of a Bayesian network as a fusion engine instead of the summing voter. Although the summing voter fusion engine improves the overall localization performance, it does not take into account the unique characteristics of each localizer like the averaged beam pattern of the microphone array as shown in

Chapter 5.1.7. The Bayesian network allows the inclusion of these unique characteristics as part of the fusion process and therefore further improves the overall localization performance.

By applying Equation (4-3), a Bayesian inference model can be constructed.

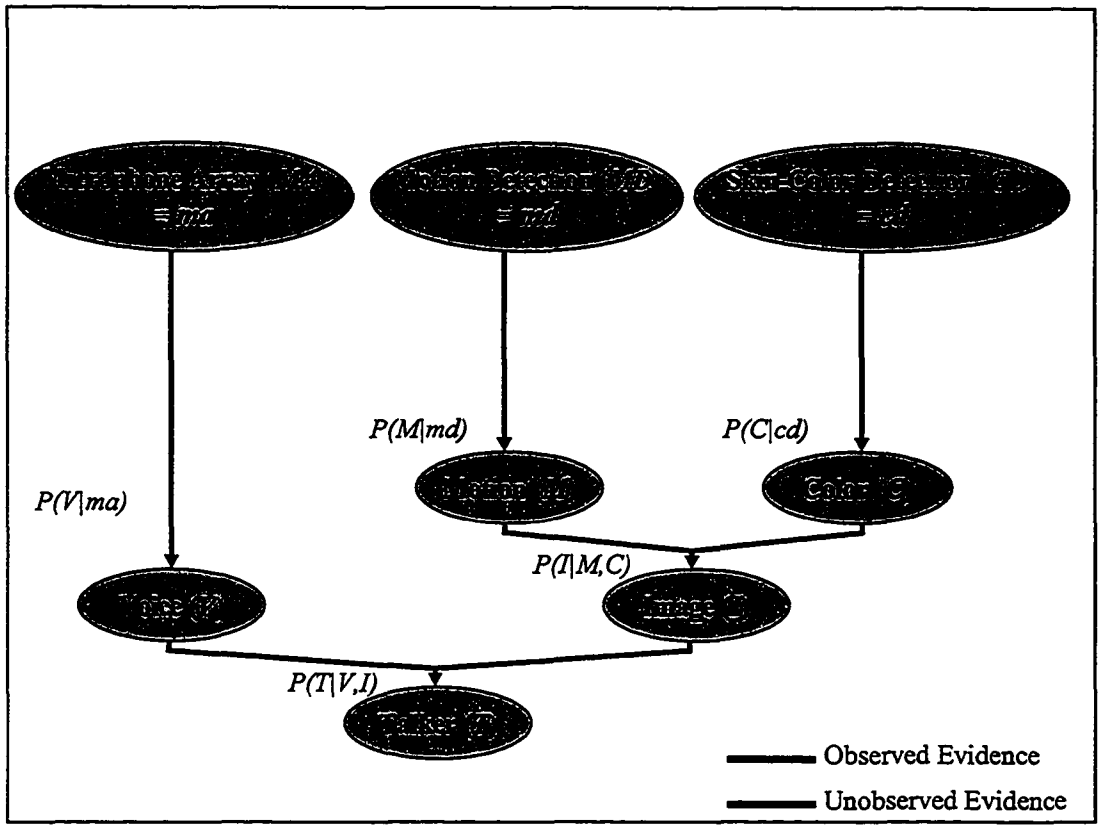

Figure 6-4. Bayesian inference model for performing data fusion on the localization results.

Figure 6-4 shows the Bayesian inference model for performing data fusion on the localization results. The observed evidence *e* is the localization outputs from the microphone array, the motion detection localizer, and skin-color detection localizer as well as their corresponding reliability estimates. The localizers are represented by the nodes Microphone Array (*MA*), Motion Detection (*MD*), Skin-color Detection (*CD*).

Motion (M), Color (C), Voice (V) and Image (I) are the unobserved random variables. The Talker node (T) is the talker's location which we are trying to find. Each arrow represents a conditional probability. The values of the observed evidence are represented by $MA=ma$, $MD=md$, and $CD=cd$, respectively. With the observed evidence, Equation (4-4) can be applied onto the inference model and the Talker node (T) can be found using

$$P(T \mid V,I) = P(ma) \cdot P(md) \cdot P(cd) \cdot P(M \mid md) \cdot P(C \mid cd) \cdot P(I \mid M,C) \cdot P(V \mid ma) \qquad (6\text{-}4)$$

As mentioned in Chapter 4, bucket elimination [PEA88] is used to marginalize the variables. The value of MA, MD, and CD are observed. Only the non-observed variables M, C, V and I need to be marginalized following [PEA88]. Note that the order in which $P(T,e)$ is marginalized does not affect the end result. The *a priori* is obtained using an initialization run before the start of experiments and from other standalone experiments like the one active sector at a time experiment shown in Section 5.1.8.1, which measures the baseline behaviour of the microphone array.

6.5 Experiments — Joint Audio-Video Talker Localization Using a Bayesian Network

The experimental setup for this section is similar to what is used in Section 6.3. The same audio and video disturbances as listed in Table 6-1 are used in this set of experiments. Before the experiments were started, initialization runs were done to populate the inference model with *a priori* knowledge, as well as to calibrate the mapping required

for relating the locations of the sectors and the camera's field of view. In order to perform the initialization, the talker makes a 20 s presentation at a fix distance in front of the microphone array and the camera one sector at a time. After the talker has finished the 20 s presentation at one sector, he walks to the next sector slowly while talking. Two runs are made for each initialization. The transition of the audio and the video data going from one sector to the next is used to determine the calibration required to map the audio sectors to the camera's field of view. Using equation 6-1, the data from the initialization runs are used to calculate the conditional probabilities for all the unobserved evidence, $P(M|md)$, $P(C|cd)$, $P(I|M,C)$ and $P(V|ma)$, needed in equation (6-4). Since the observed evidence are obtained from the localizers' output directly, the probabilities of the observed evidence, $P(ma)$, $P(md)$ and $P(cd)$, equal to 1.

6.5.1 Experimental Results

Figure 6-5 shows the fusion results using the Bayesian network. Figure 6-5(a) shows the localization output using only the microphone array. Figure 6-5(b) shows the localization output using only the motion detection. Figure 6-5(c) shows the localization output using only skin-color detection. These three cases represent the localization performance if only one of the localizer is used. In each figure, the plots on the right are the fused localization results. Figure 6-5(d) show the final fused results.

When comparing the results using Bayesian network fusion, Figure 6-5(d), with the results using simple summing fusion, Figure 6-3(d), Bayesian network fusion is marginally worse in the overall talker localization performance. More localization errors were found in scenarios (2), (3), and (6).

In the inference model, Figure 6-4, the localization results from the motion detection localizer and the skin-color detection localizer are first combined into the Image node (*I*). The results from the Voice node (*V*) and the Image node (*I*) are then combined again to form the final estimate of the talker's location. The model assumes the *I* and the *V* node contribute equally in the fusion process. As an artifact, the final estimates are bias towards the audio localizer. This effect is reflected in the high total error rate of 38.7%, and is particularly evident in scenario (6), Figure 6-5(d), showing 75% of location error rate. The way the Bayesian network fusion localize incorrectly matches very closely with what the microphone array is outputting in Figure 6-5(c) scenario (6). Instead of drawing a conclusion on the performance of the Bayesian network here, we will save the conclusion for the later chapters. In the following two chapters, we will investigate how the occupancy information can be used to improve the accuracy of the Bayesian network fusion, and how adding an additional IR localizer can also impact the overall performance. Their performance will be studied and a conclusion will be given in these chapters when more experimental results are available.

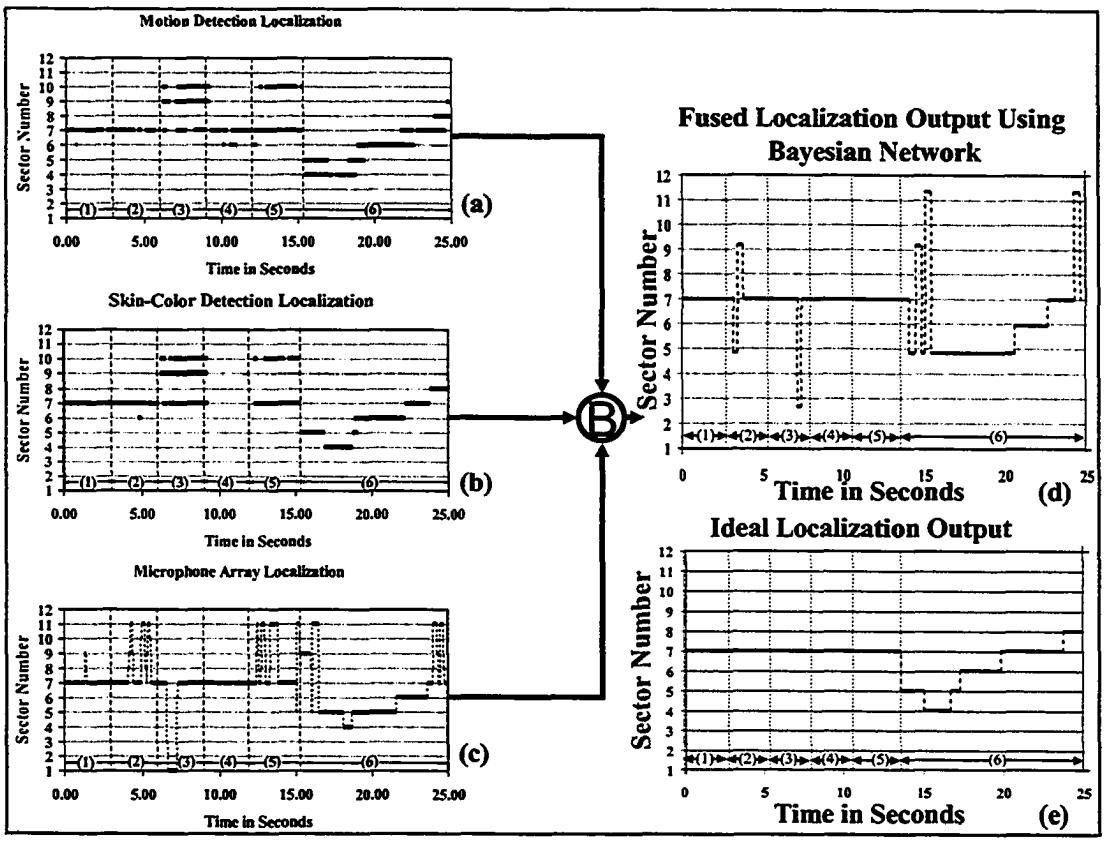

Figure 6-5. Joint audio-video localization results using the Bayesian network.

Chapter 7 Multimodal Talker Localization Using Joint Audio-Video Information and the Occupancy Estimates

In this chapter, we study how weights can be used to improve the overall performance of the multimodal talker localization system.

7.1 Using Occupancy Estimates as Weights

As mentioned in Chapter 4, the general basic fusion architecture shown in Figure 4-1 assumes each localizer contributes to the fusion process equally. However, if the correctness of the output of each localizer can be estimated, a weighting function can be used to bias the less reliable localizers away from the fusion process so that they contribute less in the final fused output. The use of weights in controlling a fusion processor was discussed in Chapter 4.2, and the general architecture for performing multimodal data fusion with weights was given in Figure 4-3.

In this chapter, we investigate the use of the occupancy information as weights. The confidence level of each localizer's output is estimated using the occupancy grid mapping technique. The occupancy grid mapping technique is widely used in robotics especially for the purpose of navigation [ELF89] [PET96] [YEU94]. The technique divides the environment into a discrete grid and assigns each grid location a value related to the

probability that the location is occupied by an obstacle. Initially, the entire grid is assigned with equal value. Sensor readings are then used to modify the grid value to reflect the probability that a specific grid location is being occupied.

One of the novelties of this thesis is adapting the occupancy grid concept and then using it to improve the overall talker localization performance and robustness. In order to compute the occupancy information, the grid is setup so that it coincides with the activation sectors of the microphone array as shown in Figure 5-2. As the localizers locate the talker in the conferencing environment, we estimate the probability of the talker occupying a particular sector. The occupancy estimates are derived based on known physical properties of each individual localizer and the current measurement of that localizer. In order to achieve better localization performance, the occupancy estimates are introduced into the fusion engines as weights to control how the localization results from individual localizers are combined during the data fusion process. Although the summing voter fusion engine used in Section 6.3 improves the overall localization performance, it does not take into account the unique characteristics of each localizer. Occupancy estimates allow the inclusion of these unique characteristics as part of the fusion process and therefore further improves the overall localization performance. As seen in Figure 6-5 (d), a Bayesian network fusion can perform poorly if proper weights are not applied to control how data streams are fused. The occupancy estimates can be used to provide the proper weights. Also, when one or more localizers fail, the persistent erroneous data streams from the failed localizers can negatively affect a statistically based data fusion method, like the Bayesian network, resulting in poor localization

accuracy [CHA02]. The occupancy estimates provide a means to automatically stop the failed devices from contributing in the data fusion process, hence improving the overall robustness of the system. In this chapter, the impact of adding reliability estimates into the fusion engine, and how occupancy estimates can be used to eliminate the failed localizers from the fusion process are investigated.

7.2 Joint Audio-Video Talker Localization using Occupancy Assisted Summing Voter Fusion

This section explores the use of occupancy estimates in combination with the summing voter fusion engine used in the previous chapter. Taking advantage of the Weights block in the generalized system architecture in Figure 4-1, the reliability estimates function as weights to assist the summing voter. It dynamically discriminates erroneous audio and video localization results by biasing the data fusion engine away from these questionable data.

The audio and video localization methods used are identical as in the joint audio-video talker localization using summing voter fusion, Section 6.3, and using Bayesian network fusion, Section 6.4.

7.2.1 Talker's Occupancy Estimates $G_{(m,n)}$ and Correctness Probability $P_{(m,n)}$ for Grid Location n and Localization Modality m

The video conferencing space is divided into a discrete two-dimensional polar grid which coincides with the activation sectors used by microphone array as shown in Figure 5-2. The occupancy estimates for each detection modality are derived and computed based on the specific physical properties of the sensors and their current output. The following section outlines how the occupancy information is estimated and why.

7.2.1.1 Occupancy Estimates for Audio Localization

In a video conferencing environment, acoustic reflections and multiple talkers often confuse the microphone array and cause the microphone array to locate the active talker incorrectly [OMO96]. The occupancy measurement $G_{(audio,n)}$ of the audio localizer is designed to discriminate localization errors from these sources.

The main concept for deriving the occupancy estimates for the audio detection is based on the averaged beam pattern method developed in Section 5.1.7. Because of the use of a delay-and-sum beamformer, a single talker gives a unique beam pattern. If the averaged beam pattern is collected for one active sector at a time in an ideal condition, such as the anechoic chamber, these averaged beam patterns can be normalized and stored as a set of reference patterns. As the averaged beam pattern of the current detection is cross-correlated with the reference pattern, the amount of deviation can be used to estimate the likelihood that the current detected sector is occupied by the talker.

After audio localization is performed, the output is used to calculate its occupancy $G_{(audio,n)}$ and the associated correctness probability $P_{(audio,n)}$. First, the average power of each sector is computed, and normalized by dividing by its root-mean-square (RMS) value using

$$\overline{\alpha[k]} = \frac{\alpha[k]}{RMS}; \qquad RMS = \sqrt{\frac{\sum_{k=1}^{N} \alpha[k]^2}{N}} \quad k \in [1,N] \tag{7-1}$$

where α[k] is the averaged power profile, $\overline{\alpha[k]}$ is the RMS normalized averaged power profile, and N is the number of sectors, which is 12 in this study. The averaged beam pattern, similar to what is shown in Figure 5-6, is then composed.

The RMS normalized beam pattern is then cross-correlated with the corresponding N reference averaged beam patterns. Since the maximum cross-correlation result between the two RMS normalized series is N, the cross-correlation result is first divided by N so that it ranges from zero to one. The maximum value of the scaled cross-correlation result is then used as $G_{(audio,n)}$.

$$\begin{aligned} R_{(\overline{\alpha}\overline{\rho},n)}(i) &= \sum_{k=1}^{N-i} \overline{\alpha}_n[k+i] \cdot \overline{\rho}_n[k] \quad i=1,2,3,...2N\text{-}1;\ k \in [1,N] \\ G_{(audio,n)} &= \frac{1}{N} \max_i (R_{(\overline{\alpha}\overline{\rho},n)}(i)) \end{aligned} \tag{7-2}$$

where $R_{(\overline{\alpha}\overline{\rho},n)}(i)$ is the cross-correlation between $\overline{\alpha}_n[k]$ and $\overline{\rho}_n[k]$, N is the dimension of the profile which is also the total number of sectors in this case. $\overline{\alpha}_n[k]$ is the RMS normalized averaged beam pattern for sector n, and $\overline{\rho}_n[k]$ is the reference averaged beam pattern for sector n.

The reference averaged beam pattern $\overline{\rho}_n[k]$ is generated through controlled experiments done in an anechoic chamber where audio is presented to only one sector at a time as shown in Section 5.1.8.1. The averaged sector power is RMS normalized and subsequently stored as the reference averaged beam pattern. There are a few situations where an averaged beam pattern can deviate from its reference pattern which will result in a low cross-correlation value. For example, the current sector is not the active sector at all; the microphone array is picking up reflections instead of the active talker or multiple talkers are speaking simultaneously.

7.2.1.2 Occupancy Estimates for Video Localization

In this thesis, the occupancy of the motion detection $G_{(motion,n)}$ and the occupancy of the skin-color detection $G_{(skin\text{-}color,n)}$ are derived from the changes of the foreground-to-background ratio (f-b ratio) within an area-of-interest (AOI) [TOY00]. The f-b ratio is computed as the ratio of black pixels (foreground) to white pixels (background) within a binary thresholded AOI. Changes in the f-b ratio should be gradual. Any sudden jumps in the f-b ratio indicate questionable localization result.

Similar to its audio counterpart, video localization results are used to compute the reliabilities and the correctness probabilities for both the motion and skin-color detection. Since the degree of motion should be limited in finite time, large changes of the f-b ratio

suggest questionable localization results. Based on the changes in f-b ratio, a detection quality, Q, is assigned. In order to allow the occupancy calculation to have better stability in a highly dynamic conferencing environment, the detection quality assignments are set up in a way that small fluctuations of the change of f-b ratio do not affect the assigned value of the detection quality significantly. However, when the change in f-b ratio is large, it is heavily penalized. Consequently, an inverted power curve like distribution, power of square root, is used to assign the detection quality to the corresponding changes in f-b ratio. Table 7-1 shows the assignment values used in this thesis. The occupancy of the active sectors is then computed using Equation (7-3) and sectors which are not reported as active are assumed to have equal occupancy and are computed using Equation (7-4)

Table 7-1. Foreground-to-background ratio to detection quality assignment.

Changes In Foreground-to-Background Ratio (%)	Assigned Detection Quality Q
0-9.99	1
10-19.99	$\sqrt{0.9}\approx 0.95$
20-29.99	$\sqrt{0.8}\approx 0.89$
30-39.99	$\sqrt{0.7}\approx 0.84$
40-49.99	$\sqrt{0.6}\approx 0.77$
50-59.99	$\sqrt{0.5}\approx 0.71$
60-69.99	$\sqrt{0.4}\approx 0.63$
70-79.99	$\sqrt{0.3}\approx 0.55$
80-89.99	$\sqrt{0.2}\approx 0.45$
90->100	$\sqrt{0.01}=0.1$

$$G_{(video,n)} \text{ for the active sector} = Q \tag{7-3}$$

$$G_{(video,n)} \text{ for the inactive sector} = \frac{J_S - \sum_{i=\text{active sector}} G_{(video,i)}}{N - J_S} \qquad N < J_S \tag{7-4}$$

where Q is the assigned detection quality based on the mapped grading scale in Table 7-1, J_S is the total number of detected active sectors, N is the number of sectors, and *video* can be either motion or skin-color detection.

7.2.1.3 Correctness Probability for Audio and Video Localization

The probability of correctness for audio localization $P_{(audio,n)}$ for the current output is a statistical measurement of how often the output sector was detected in the past. It is computed using the same method as outlined in Chapter 6.2

$$P_{(m,n)}[i] = \frac{\sum_{k=i-td}^{i} D_n[k]}{\sum_{s=1}^{N} \sum_{k=i-td}^{i} D_s[k]} \tag{7-5}$$

where $D_n[k]$ is the number of detections in sector n at time k and td is the width of the window of time to look back to from the current data point, N is the number of sectors, and m is the detection method which can be the audio beamforming, motion, or skin-color detection for this thesis.

The probability of correctness for motion detection, $P_{(motion,n)}$, and for skin-color detection, $P_{(skin\text{-}color,n)}$, of the current output is computed using Equation (7-5) as well. Motion detection identifies people as well as any periodic movements like fan blades and

monitor flicker. Similarly, skin-color detection identifies people as well as any objects that have skin-like color. Therefore, additional steps are necessary to reduce the chance of detecting artifacts. Periodic motions give few changes over a long period of time; therefore, active sectors with small changes in the f-b ratio over a long period of time are treated as static background objects in the experiment. In this thesis, a simple method is developed to check for periodic or static objects. AOIs containing the objects are first found using the automatic AOI identifying algorithm from Section 5.2.1. A sliding window is then used to compute the sum of the changes of the f-b ratio in each AOIs. Only AOIs with the sum above a predefined threshold are identified as valid AOIs.

7.2.2 Occupancy Assisted Summing Voter Fusion and Final Localization Decision

The summing voter fusion equation, (6-2), is modified to take the occupancy information into account. By using the occupancy estimates as weights, the voter automatically relies more on the localizer that is more trustworthy. The final output for any sector at a given point in time is decided based on the product of its correctness probability $P_{(m,n)}$ and associated occupancy $G_{(m,n)}$, and then summed as follows

$$K_n = \sum G_{(m,n)} P_{(m,n)} = G_{(audio,n)} P_{(audio,n)} + G_{(motion,n)} P_{(motion,n)} + G_{(skin-color,n)} P_{(skin-color,n)} \quad (7\text{-}6)$$

where m is the localization method, n is the currently detected active sector, K_n is the fused result for sector n, $G_{(m,n)}$ is the occupancy for method m detecting sector n as active, and $P_{(m,n)}$ is the probability of method m detecting sector n as active. Since three detection methods are used in this study, m can be audio beamforming, motion detection, or skin-

color detection. The sector that gives the highest K_n is taken as the fused localization output.

7.2.3 Experimental Results

In this chapter, the effect of introducing occupancy estimates into the summing voter is studied. Similar to what has been done with the simple summing voter fusion in Section 6.2, six different scenarios, Table 7-1, with different audio and video localization disturbances are used to study the effect of occupancy estimates on the summing voter fusion. The localization results for the six scenarios are concatenated in time, and shown in Figure 7-1. These scenarios are marked as (1), (2), (3), (4), (5), and (6) in the figure and the time ranges for each scenario are the same as what were listed in section 6.3.1.

The three plots on the left show the localization output of each localizer when it is used as a stand-alone device. While Figure 7-1(d) shows the fused localization result using the occupancy assisted voter, Figure 7-1(a) shows the localization output of the motion detection localizer, Figure 7-1(b) shows the localization output of the skin-color detection localizer, Figure 7-1(c) shows the localization output of the microphone array, and Figure 7-1(e) show the localization output for an ideal localizer. Similar to what is done in Section 6.3, both the motion detection and the skin-color detection localizers are capable of detecting multiple objects, more than one sector can be identified as active simultaneously. However, localization errors for the motion and skin-color detection can show as multiple active sectors as well. Audio localization errors simply show as

discontinuities in the plots. The microphone array localizer and the fused localization output only one active sector and therefore, localization errors would show as jumps in the plots.

In scenario (1), the talker was standing in sector 7 and spoke directly towards the microphone array. This would be the "normal" operating scenario. Although the microphone array has a short lived localization error, both fusion methods reject the error and provide perfect output for scenario (1). In scenario (2), when the talker speaks towards the walls, the acoustic reflections confuse the microphone array and cause it to localize incorrectly. This scenario is very common during conferencing when the talker turns his head to address a question from a participant. Again, both fusion methods combine the localization results from the two video localizers and successfully remove all the audio localization errors to give a perfect output. In scenario (3), the talker speaks toward the side wall while another person, who is facing the camera, was sitting at sector 10 in the background. The motion detection and the skin-color detection recognized the person in the background. Comparing the joint audio-video talker localization results using summing voter fusion, Figure 6-3, with the results from occupancy assisted summing voter, Figure 7-1, the simple summing voter fusion has more localization errors than the occupancy assisted voter. Objects with no or very few changes over time are treated as background objects and hence allow the occupancy assisted voter to achieve a better performance. The performance gain for scenario (3) is due to the occupancy estimates. The occupancy estimates for both the motion and skin-color detection put a heavier weight on sector 7 because it gives the best detection quality Q. In scenario (4),

the lights were dimmed so that only the outline of objects is visible. While the skin-color detection fails, the motion detection still detects the motion of the talker's outline with only a few localization errors. Both fusion methods are not affected by the skin-color detection failure and have perfect outputs. In scenario (5), heavy acoustic reflections were combined with another person conducting tasks in the background at sector 10, providing a challenging situation. Both fusion methods suffer the same localization errors but they performed considerably well. In scenario (6), the talker walked around the microphone array from sector 5 toward sector 4, then back to sectors 5, 6, 7, and then 8. The total localization error rate for the occupancy assisted summing voter fusion method shown in Figure 7-1 is 7.2 % while the tracking error rate is 18.5%. Comparing the results with the simple summing voter method used in 6.3.1, which has total error rate of 12.6% and tracking error rate of 23.9%, the occupancy assisted voter has better overall and tracking localization performance. The occupancy estimates provided additional information to better reject localization errors, and hence gives better localization performance.

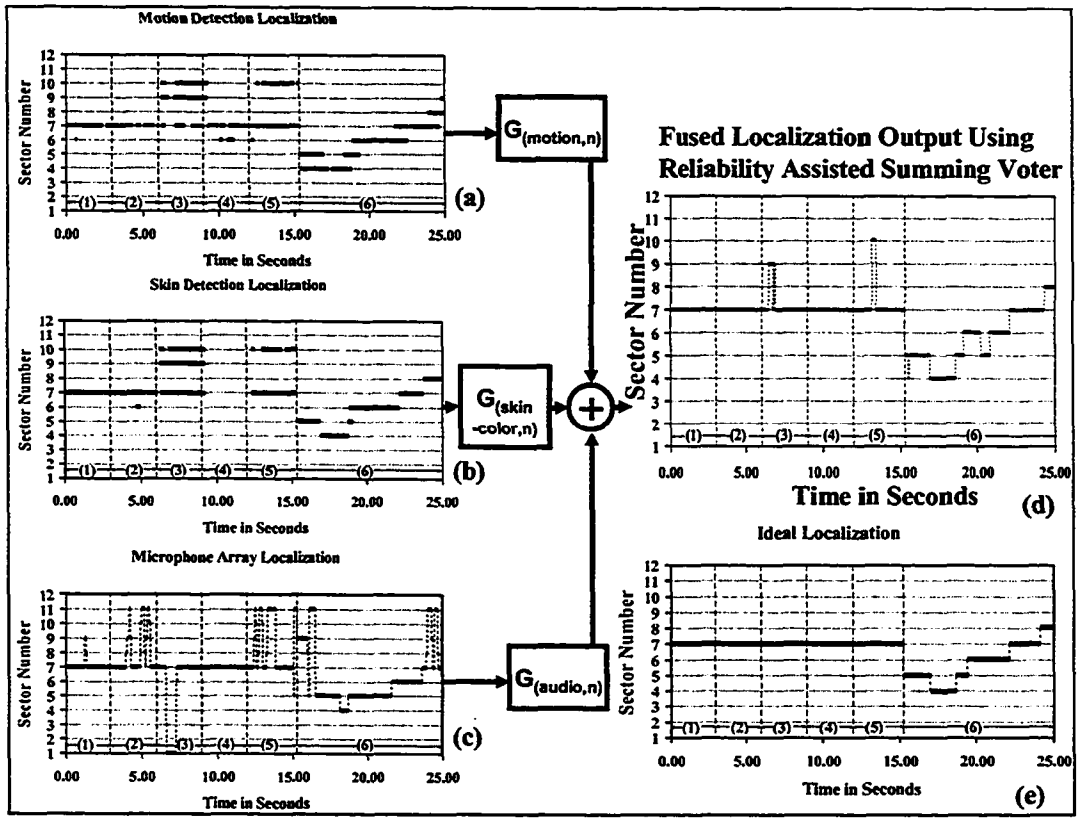

Figure 7-1. Fused localization result with occupancy information, (a) localization from motion detection localizer, (b) localization from skin-color detection localizer, (c) localization from microphone array localizer, (d) fused localization output using occupancy assisted summing voter, and (e) localization from an ideal localizer.

7.3 Joint Audio-Video Talker Localization using Occupancy Assisted Bayesian Network Fusion

Similar to the occupancy assisted summing voter method above, the occupancy assisted Bayesian network fusion takes advantage of the *Weights* block in the generalized system architecture (Figure 4-2) as well. Unlike the occupancy assisted summing voter, the Bayesian network fusion allows the inclusion of the unique characteristics of individual localizers as part of the fusion process, therefore, adding occupancy to the Bayesian decision does not improve the accuracy of the overall localization substantially. Instead,

the occupancy estimates are aimed to improve the robustness of the overall localization. When one or more localizers fail, the persistent erroneous data streams from the failed localizers can negatively affect a statistically based data fusion method, like the Bayesian network, resulting in poor localization accuracy [CHA02]. This section explores the use of the occupancy estimates to automatically stop the failed devices from contributing in the data fusion process, hence improving the overall robustness of the system. The effect of adding occupancy estimates into the Bayesian fusion engine, and using occupancy estimates to stop the failed localizers from contributing in the fusion process are studied. Audio and video localization methods used are identical as the joint audio-video talker localization using Bayesian network fusion as detailed in Chapter 6.4.

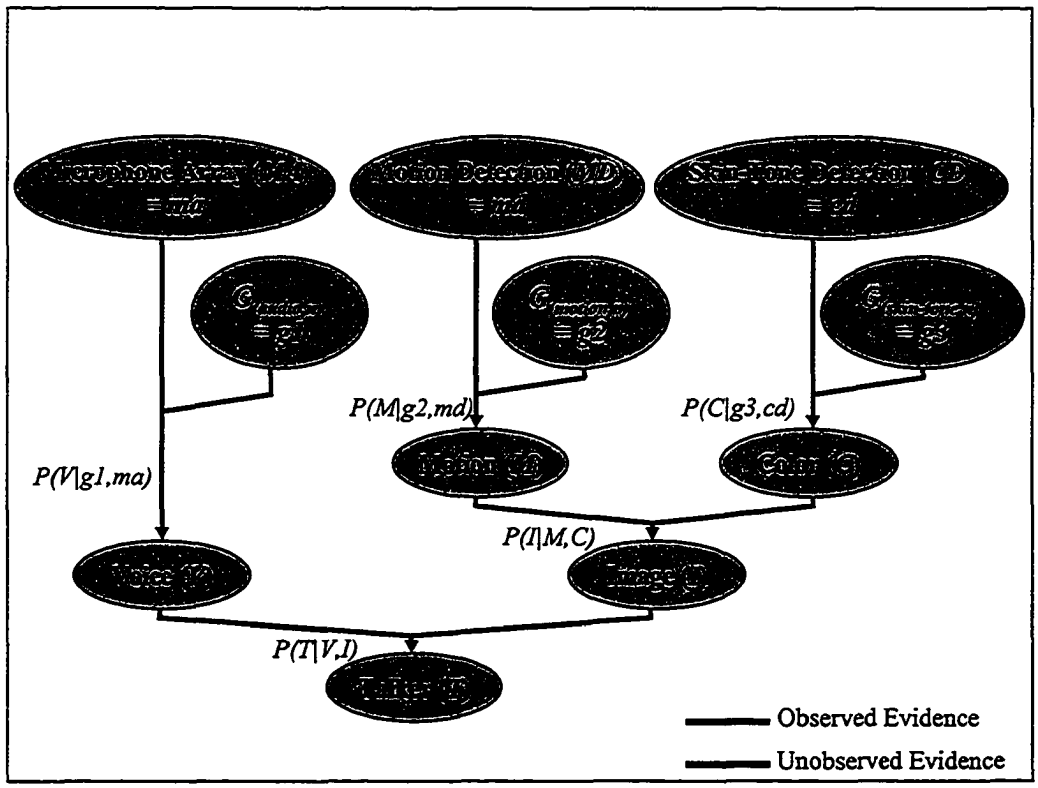

Figure 7-2. Bayesian inference model with occupancy estimates for joint audio-video localization.

7.3.1 *Occupancy Assisted Bayesian Network Fusion and Final Localization Decision*

Since the audio and video localization methods used are the same as the occupancy assisted summing voter method, the occupancy estimates $G_{(m,n)}$ and correctness probabilities $P(m,n)$ are calculated the same way as before using Equation (7-3), (7-4), and (7-5). The Bayesian network, Figure 6-4, and the fusion equation, (6-4), are modified to accommodate the addition of the occupancy information.

Figure 7-2 shows the Bayesian inference model for performing data fusion on the localization results and the occupancy information. The observed evidence e is the localization outputs from the microphone array, the motion detection localizer, and skin-tone detection localizer as well as their corresponding occupancy estimates. The localizers are represented by the nodes Microphone Array (*MA*), Motion Detection (*MD*), Skin-color Detection (*CD*) and their corresponding occupancy estimations are represented by node $G_{(audio,n)}$, $G_{(motion,n)}$, and $G_{(skin\text{-}color,n)}$, respectively. Motion (*M*), Color (*C*), Voice (*V*) and Image (*I*) are the unobserved random variables. The Talker node (*T*) is the talker's location which we are trying to find. Each arrow represents a conditional probability. The values of the observed evidence are represented by $MA=ma$, $MD=md$, $CD=cd$, $G_{(audio,n)}=g1$, $G_{(motion,n)}=g2$ and $G_{(skin\text{-}color,n)}=g3$, respectively. With the observed evidence, (7-3) can be applied onto the inference model and the Talker node (*T*) can be found using

$$\begin{aligned} P(T|V,I) &= P(ma, md, cd, g1, g2, g3, M, C, V, I, T) \\ &= P(ma)\cdot P(md)\cdot P(cd)\cdot P(g1)\cdot P(g2)\cdot P(g3)\cdot P(V|g1,ma)\cdot \\ &\quad P(M|g2,md)\cdot P(C|g3,cd)\cdot P(I|M,C) \end{aligned} \tag{7-7}$$

Bucket elimination [JEA88] is used to marginalize the non-observed variables. The value of *MA*, *MD*, *CD*, $G_{(audio,n)}$, $G_{(motion,n)}$ and $G_{(skin\text{-}color,n)}$ are observed variables and *M*, *C*, *V* and *I* are non-observed variables. Before the inference model can be used, each node is populated with its *a priori* knowledge. The *a priori* knowledge can be obtained during an initialization run before the start of experiments and from other standalone experiments.

7.3.2 Device Failure Detection

When a localizer fails, its output should be considered erroneous. Since the output of a failed localizer is usually persistent, statistically based fusion methods, such as a Bayesian network, often have no choice but to fuse the failed output along with the other outputs causing significant degradation on the overall localization accuracy [CHA02]. For example, when the lights are dimmed to give a presentation, the skin-color video localizers produce erroneous outputs in establishing the AOI. Therefore, it is important to eliminate any contribution from the failed localizers. One of the novelties of this thesis is the use of the occupancy estimates to bias the Bayesian network away from fusing the erroneous data from the failed localizers, hence eliminating their contribution. For motion and skin-color detection, if all the pixels in the video scene are black, the localizers will assume there is a device failure and force the corresponding occupancy estimate to zero as shown in Equation (7-8). Similarly, if there is no detectable audio signal from the microphones within a time window, the microphone array will assume there is a device failure and force its occupancy estimate to zero as shown in (7-9).

$$If\ (\sum_{row=1}^{X}\sum_{column=1}^{Y} P_{(row,column)}) < T_{video} \Rightarrow G_{(video,n)} = 0; \quad n = 1...12 \tag{7-8}$$

$$If \sum_{i=1}^{6} Mic_i < T_{audio} \Rightarrow G_{(audio,n)} = 0; \quad n = 1...12 \tag{7-9}$$

where $P_{(row,column)}$ is a video pixel located at (row, column) in the video image, X and Y are the respective width and height of the video image, T_{video} and T_{audio} are predefined threshold values which are zero in this study, Mic_i is the microphone number of the microphone array, $G_{(video,n)}$ is the occupancy estimates of the motion detection and skin-color detection, and $G_{(audio,n)}$ is the occupancy estimates of the microphone array.

If the failed device is repaired and comes back to life, or conditions improve (i.e., light levels improve for video), the corresponding $G_{(video,n)}$ or $G_{(audio,n)}$ will no longer trigger the device failure detection mechanism and, hence, will not be forced to zero. As a result, the output of the repaired device will be included in the fusion process automatically. How often the device failure check is performed will determine how fast the system responds to a device failure or after the fault has been repaired. Also, as the number of devices used in the system increases, the computational load imposed by the device failure checking will also increase.

7.3.3 Experimental Results

Figure 7-3 shows the localization results for the occupancy assisted Bayesian network fusion engine. Figure 7-3(a) shows the localization output using only the microphone array. Figure 7-3(b) shows the localization output using only the motion detection. Figure 7-3(c) shows the localization output using only skin-color detection. These three cases

represent the localization performance if only one of the localizers is used. In each figure, the plots on right are the fused localization results.

In order to study the effect of device failure, a new set of experiments was conducted using the same experimental. Device failure was simulated by turning off the camera during the experiment 7.1 s after the experiment was started. Figure 7-4(d) and Figure 7-5(d) show the Bayesian network fused results with and without occupancy estimates added in the case of devices failure. Without any video input, both the motion detection (Figure 7-4(b) and Figure 7-5(b)) and the skin-color detection (Figure 7-4(c) and Figure 7-5(c)) malfunctioned, and reported sector 5 as the active sector which is their default AOI location. Without the help of the occupancy estimates, the fused localization output (Figure 7-4(d)) was incorrectly reporting sector 5 as the active sector. However, with the occupancy estimates added, the Bayesian network was biased away from the motion detection and the skin-color detection localizers. The fused localization output (Figure 7-5(d)) was correctly reporting sector 7 as the active sector.

The results show that adding occupancy estimates to the Bayesian network fusion engine significantly improves the accuracy of the overall localization performance with total error rate of 3.2% and tracking error rate of 4.3%, when compared to joint audio-video localization done using the Bayesian network (Figure 6-5(d)) which has total error rate of 38.7% and tracking error rate of 75%, a simple summing voter (Figure 6-3(d)) which has total error rate of 12.6% and tracking error of 23.9%, and the occupancy assisted summing voter (Figure 7-1(d)) which has total error rate of 7.2% and tracking error rate

of 18.5%. When occupancy information is introduced into the fusion process, the occupancy information affects the summing voter, Figure 7-1(d), and the Bayesian network fusion engines, Figure 7-3(d), differently. The improvement of the Bayesian network is much more substantial. This can be understood as the fact that the Bayesian network inference model is chosen with the occupancy estimates added in mind. As mentioned in Chapter 6.5, without the occupancy estimates adjusting how much the system relies on each modality, the Bayesian network will by default rely more on the audio localizer than the other two. In the Bayesian network, Figure 7-2, the inference of the talker's localization is derived from its two immediate parent nodes Voice (V) and Image (I). The V node is directly affected by the microphone array, whereas, I is the combined effect of both the motion detection and the skin-color detection node. Nodes V, I are placed in the network as anchor points for adding the occupancy estimates into the network. Without the occupancy estimates controlling how much each localization modality contributes to the overall fusion, the Bayesian is network putting roughly twice as much weight on the microphone array localizer than the other two video localizers.

In the case of a device failure, its output should be considered erroneous (Figure 7-4). The simulated results show that the occupancy information automatically biases the Bayesian fusion engine away from failed localizers, thus making the overall system less sensitive to device failure resulting in improved overall robustness (Figure 7-5). For this to function correctly, the occupancy estimator or the localizers themselves, has to be capable of detecting device failure. In essence, the occupancy information effectively

controls how the statistical distribution (*a priori* knowledge) places inference on the Bayesian fusion process.

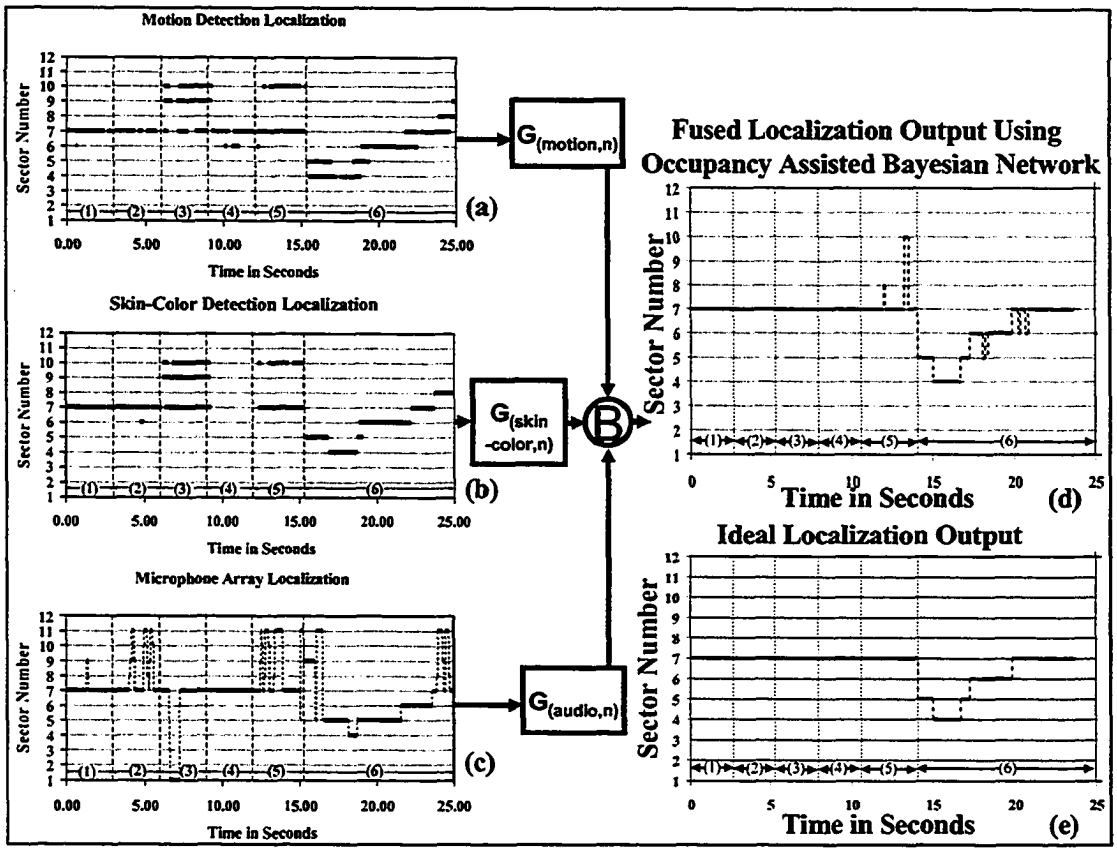

Figure 7-3. Results for joint audio-video localization using Bayesian network fusion with occupancy estimates.

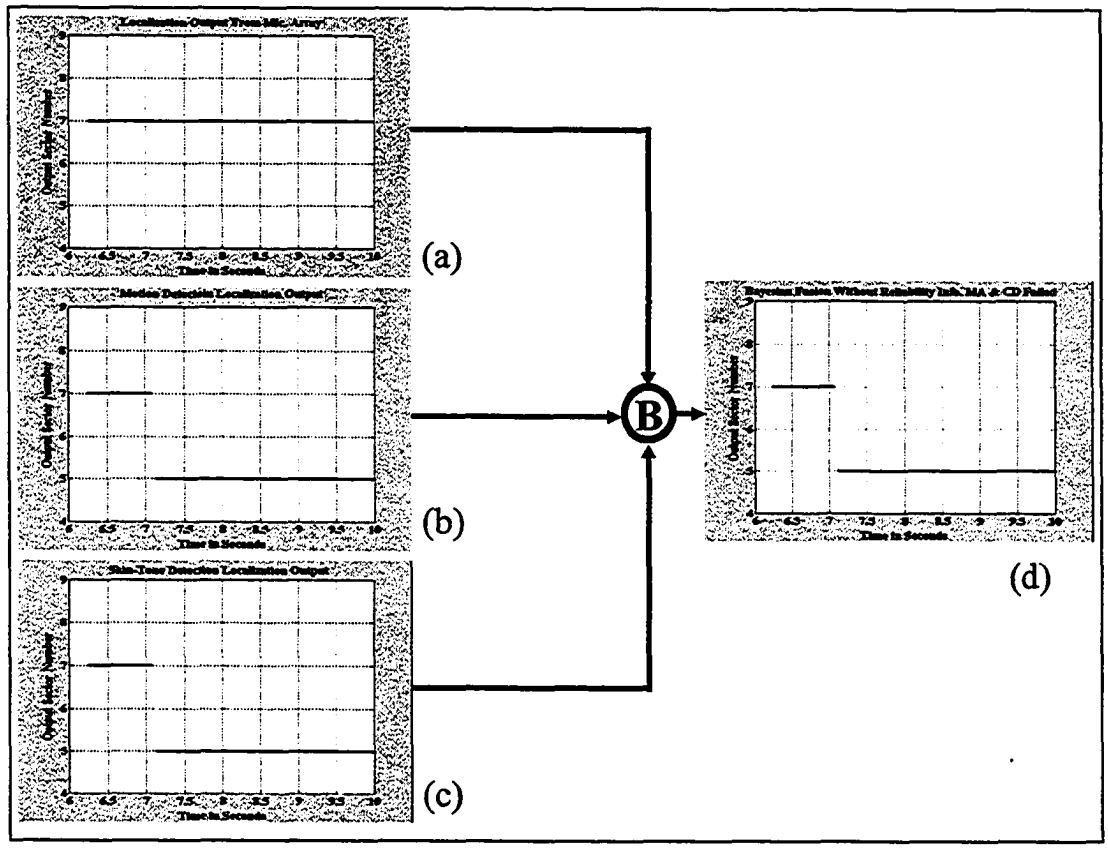

Figure 7-4. Results for joint audio-video localization using Bayesian network fusion estimates in the case of device failure.

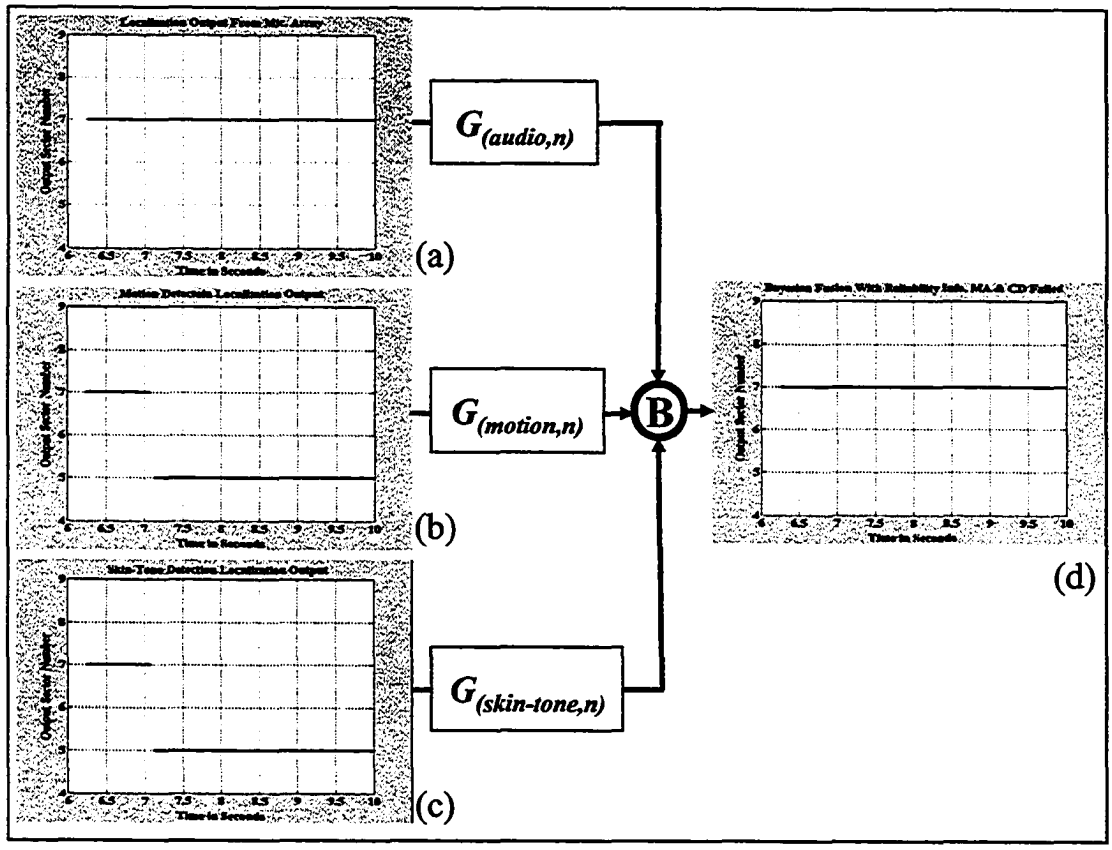

Figure 7-5. Results for joint audio-video localization using Bayesian network fusion with occupancy estimates in the case of device failure.

Chapter 8 Multimodal Talker Localization Using Audio, Video and Infrared Information

In the previous two chapters, we have studied how the modular multimodal localization architecture can be applied in video conferencing using an audio localizer and two video localizers. In this chapter, we add a new IR localization module into the architecture. The impact of the new module to the overall localization performance is studied.

8.1 System Block Diagram of the Joint Audio-Video-IR Talker Localization

Figure 4-6 shows the block diagram of the joint audio-video-IR talker localization architecture. For convenience, Figure 4-6 is repeated in Figure 8-1. It is one example of how the general modular multimodal talker localization architecture shown in Figure 4-3 can be used to construct a system for the video conferencing application.

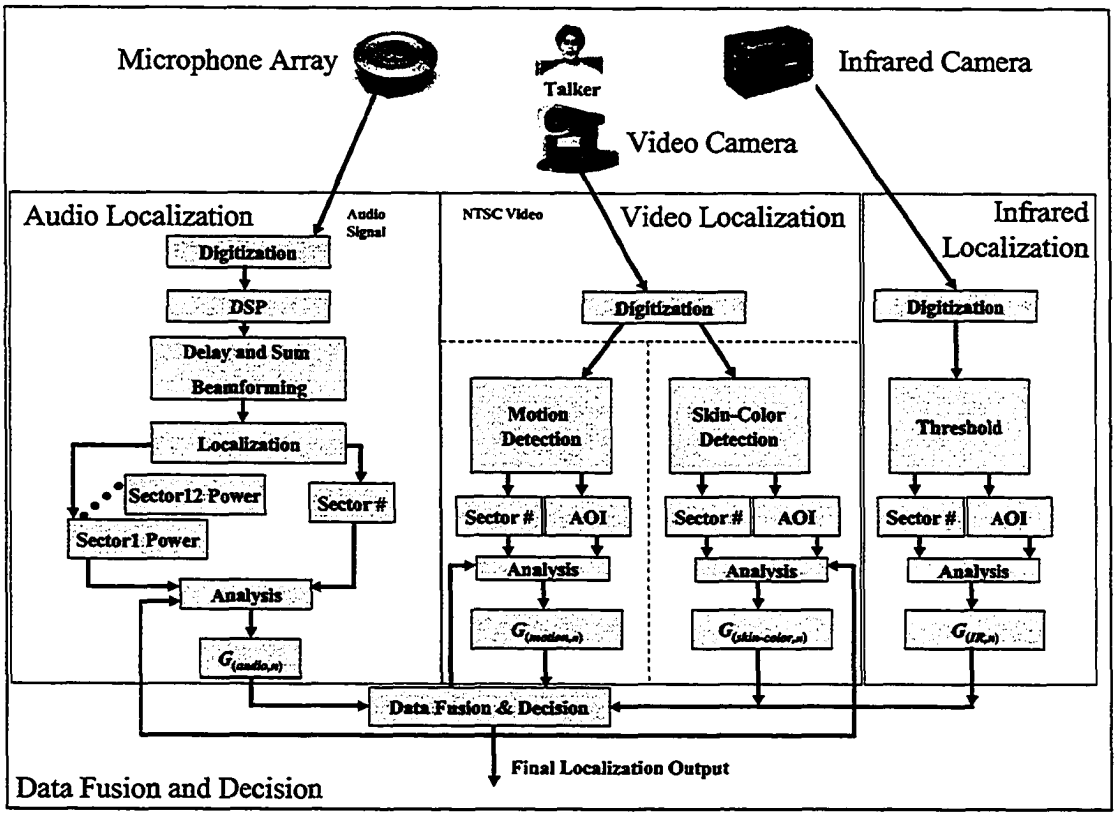

Figure 8-1. System block diagram for a video conferencing application using multimodal talker localization system.

8.2 Occupancy Estimates $G_{(IR,n)}$ and Correctness Probability $P_{(IR,n)}$ for Infrared Localization

How the talker can be detected and localized thermo-graphically is outlined in Chapter 5.3. As mentioned in Chapter 5.3, from an image processing point of view, the IR images can be treated as regular video images. The same method that is used to derive the occupancy estimates and the correctness probabilities for the motion and skin-color localizers can also be used for the IR localizer.

After IR images are binary thresholded and the active AOIs are identified as outline in Chapter 5.3.1 and 5.3.2, the changes of the f-b ratio within an active AOI are used to lookup the corresponding detection quality Q using Table 7-1. The occupancy information $G_{(IR,n)}$ is then computed using Equation (7-3) and (7-4). The correctness probability $P_{(IR,n)}$ for the IR localization modality is computed as the histogram of the localizer's output using Equation (6-1); similar to what is done in the video localizers.

8.3 Joint Audio-Video-IR Multimodal Taker Location

Similar to what is done in the previous two chapters, the simple summing voter, the occupancy assisted Bayesian Network and the occupancy assisted summing voter used as the fusion engines are also evaluated, and their localization performance are evaluated.

8.3.1 Joint Audio-Video-IR Talker Localization using Simple Summing Voter

The fusion equation is obtained by expanding Equation (6-2) to include the new IR localization module.

$$K_n = \sum_m P_{(m,n)} = P_{(audio,n)} + P_{(motion,n)} + P_{(skin-color,n)} + P_{(IR,n)} \qquad (8\text{-}1)$$

8.3.1.1 Experimental Results

In order to study the impact of the newly added IR localization module, a new set of experiments, similar to what is done in Chapter 6 and 7, is conducted. Figure 8-2 shows the experimental setup.

The experiments done in the previous two chapters are conducted in a typical conference room. The environment is relatively well controlled so that no extreme conditions are presented to the localization system. For example, the acoustic tile ceiling helps minimize reverberations, the lighting is reasonability uniform, and the furniture in the room are mostly off-white so that the skin-color detection localizer will not be confused. In order to better evaluate the benefit of adding the IR localization module, the experiments in this chapter are conducted in a more challenging environment. The reverberant room used in these experiments is much larger and is measured 6.6m wide x 9.7 m long x 4 m high. No wall surface in the room is lined with soft materials. There are also multiple acoustic reflective objects in the room which will cause strong reflections with both long and short delays. The lighting is less uniform and the furniture is in skin-tone like color. The microphone array was placed on a skin-color like color conference table. The table was placed off center in the room closer to the right wall. This will cause the acoustic reflections to be stronger in the right side near sector 1, 2 and 12. The video camera was placed two meters away from the microphone array and the IR camera is placed half a meter behind the video camera. The video images were digitized at 720 x 480 pixels, 30 frames per second. The IR images were digitized at 320 x 240 pixels, 15 frames per second. Audio/video/IR synchronization was done by hand clapping which all could register. The experiment was done with the talker standing at one meter in front of the

microphone array giving a presentation. The experiments were conducted over six scenarios as listed in Table 6-1.

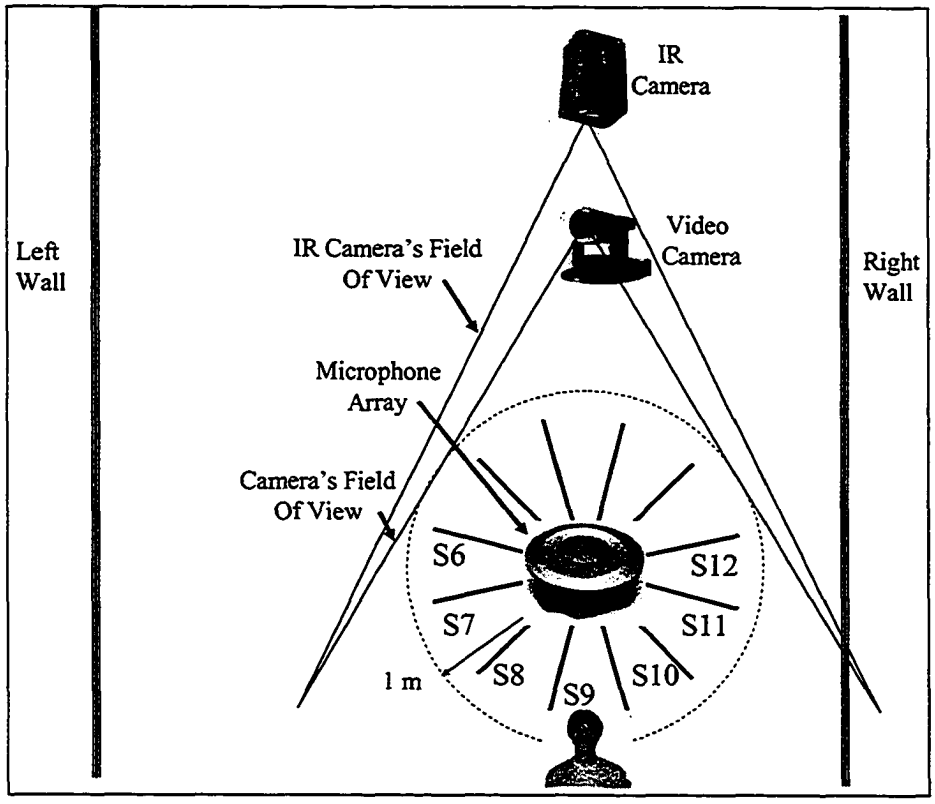

Figure 8-2. Experiment setup for the joint audio-video-IR experiments.

The challenging environment is intended to be a "stress test" to the system. It can be observed that the performance of the microphone array shown in Figure 8-3(d) degraded substantially with large number of localization errors. The performance of the skin-color detection localizer shown in Figure 8-3(b) degraded as well, mostly because of the furniture. The IR localizer, Figure 8-3(a), and the motion detection localizer, Figure 8-3(c), performed reasonably well. Figure 8-3(e) shows the joint audio-video-IR localization result using the simple summing voter. Figure 8-3(f) shows the hypothetical

ideal localization result, and is served as reference. The multimodal localization system performs well in removing the localization errors in scenarios (1), (2) and (3). However, it fails to remove the errors in scenario (4) where the lights are turned off. The room becomes so dark that both the motion detection localizer and the color detection localizer fail completely. Meanwhile, the microphone array is suffering from large numbers of reflections. Only the IR localizer is somewhat reliable. In scenario (6), the talker walks around starting at sector 9, towards the direction of sector 6, stopping at sector 6 briefly and then back towards the direction of sector 11. The IR localizer did marginally better in this scenario, but none of the single modal localizers reflect this motion sequence clearly. The multimodal system extracts most of this motion sequence well with the exception of the brief stopping at sector 6 and at the end where the talker stops at sector 11. There are also some sporadic localization errors, but taking into consideration that none of the single modal localizers are giving a clean detection of this motion sequence, the multimodal localization system is performing well with a total localization error of 30.7% and tracking error of 45.8%.

In order to study the impact of adding the IR localization module, the same set of experimental data is reprocessed with the IR localizer removed. Figure 8-4 shows the fused localization output for the system which uses the simple summing voter fusion engine. The overall localization performance degraded slightly with a few more localization errors especially in scenarios (3) and (6). Besides that, the localization output is basically unchanged.

The benefit of adding the addition IR localizer is not clear based on the experiments conducted in this section. There are a few factors that can affect how well the system is able to take advantage of an additional localizer. The localization performance of the new localizer and the fusion engine used are some of these factors. In the next two sections, we will investigate how different fusion methods can make a difference.

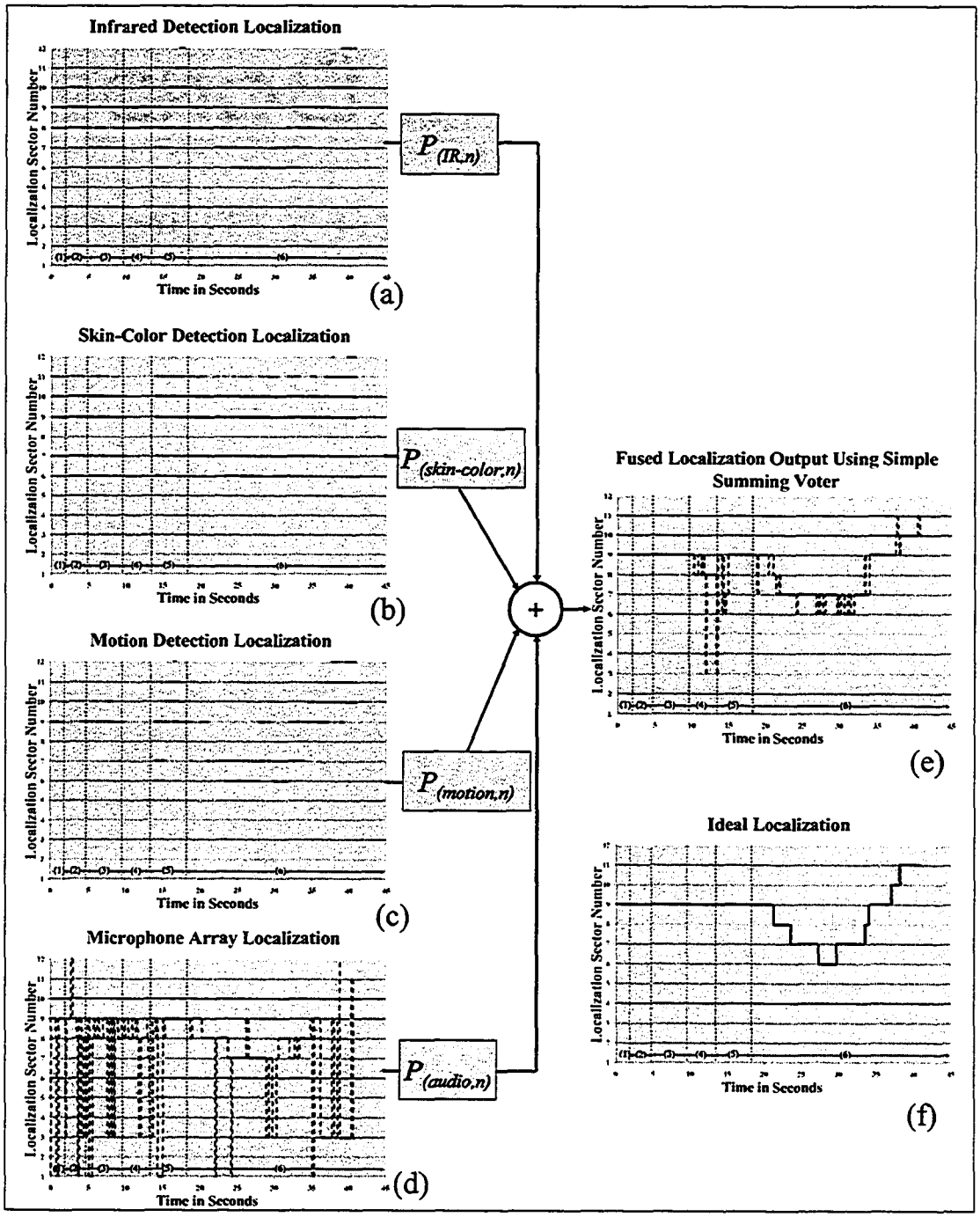

Figure 8-3. Joint audio-video-IR localization results using simple summing voter fusion

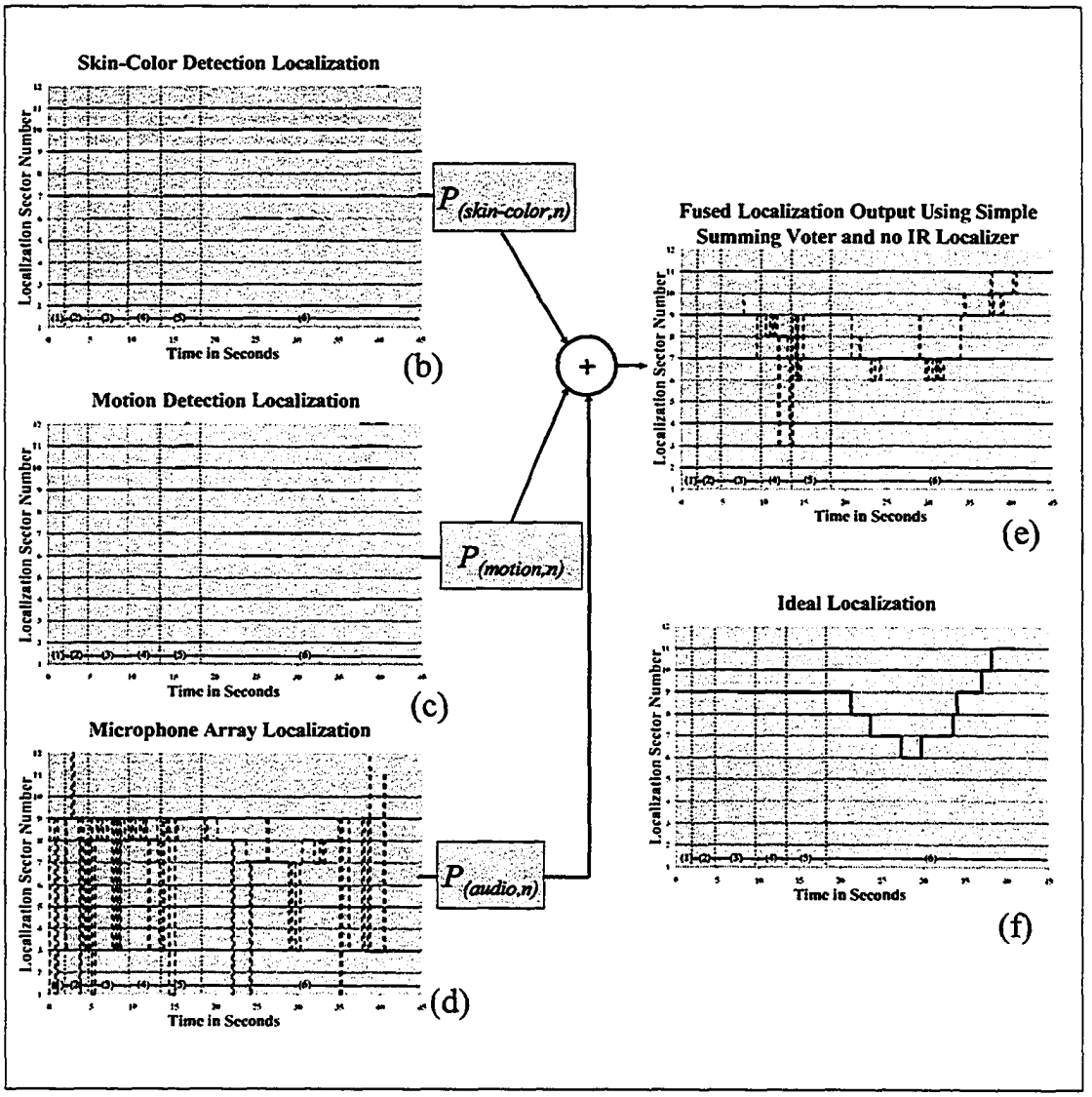

Figure 8-4. Joint audio-video localization results using simple summing voter fusion

8.3.2 Joint Audio-Video-IR Multimodal Talker Location Using Occupancy Assisted Summing Voter

To derive the new fusion equation, Equation (7-6) is expanded to include the IR localizer.

$$K_n = \sum_m G_{(m,n)} P_{(m,n)} = G_{(audio,n)} P_{(audio,n)} + G_{(motion,n)} P_{(motion,n)} + G_{(skin-color,n)} P_{(skin-color,n)} + G_{(IR,n)} P_{(IR,n)} \quad (8\text{-}2)$$

8.3.2.1 Experimental Results

The fused result for the multimodal localization system using the occupancy assisted summing voter is shown in Figure 8-5(e). With the occupancy estimates added into the fusion engine as weights, the overall localization performance does not change much. The localization result is improved in scenario (4) with the two extreme errors removed. However, the localization performance in scenarios (1), (2) and (3) is degraded with a few single value localization errors. When the results of the occupancy assisted summing voter, Figure 8-5(e), are compared with the results of the simple summing voter, Figure 8-3(e), adding the occupancy estimates only provides some benefit in the tracking performance, however, the overall improvement is marginal. The tracking error drops from 45.8% (Figure 8-3(e)) to 35.6% (Figure 8-5(e)), and the total error rate drops from 30.7% (Figure 8-3(e)) to 29.2% (Figure 8-5(e).

In order to study the impact of the IR localization module with the occupancy assisted summing voter, the IR localizer is removed and the fused results are shown in Figure 8-6. When comparing Figure 8-5(e) and Figure 8-6(e), removing the IR localizer degrades the overall localization performance significantly. It is particularly worse in scenario (4). The IR localizer is the best performing localizer among all the single modal localizers. Without the help of the IR localizer, the overall performance suffers.

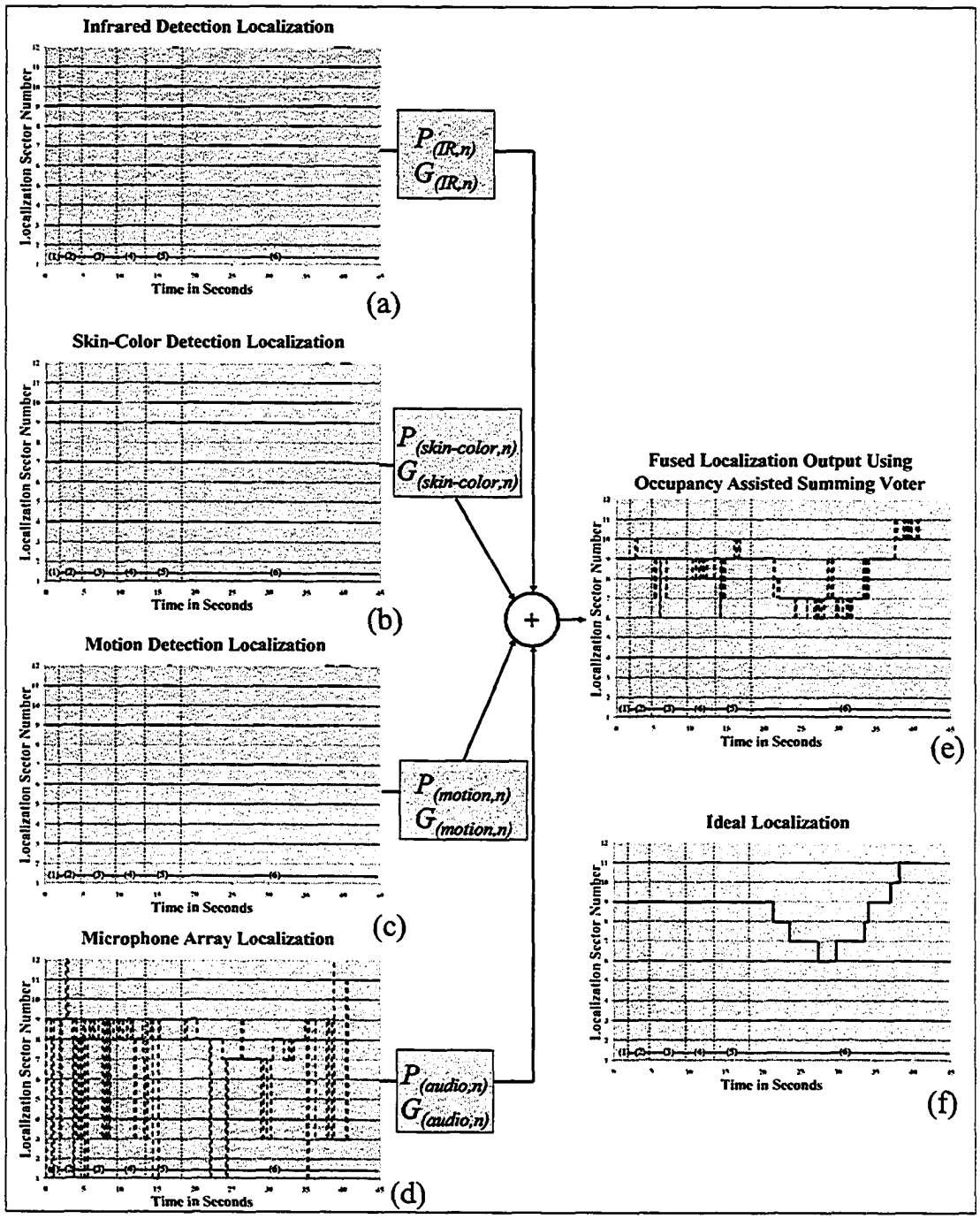

Figure 8-5. Joint audio-video-IR localization results using occupancy assisted summing voter fusion.

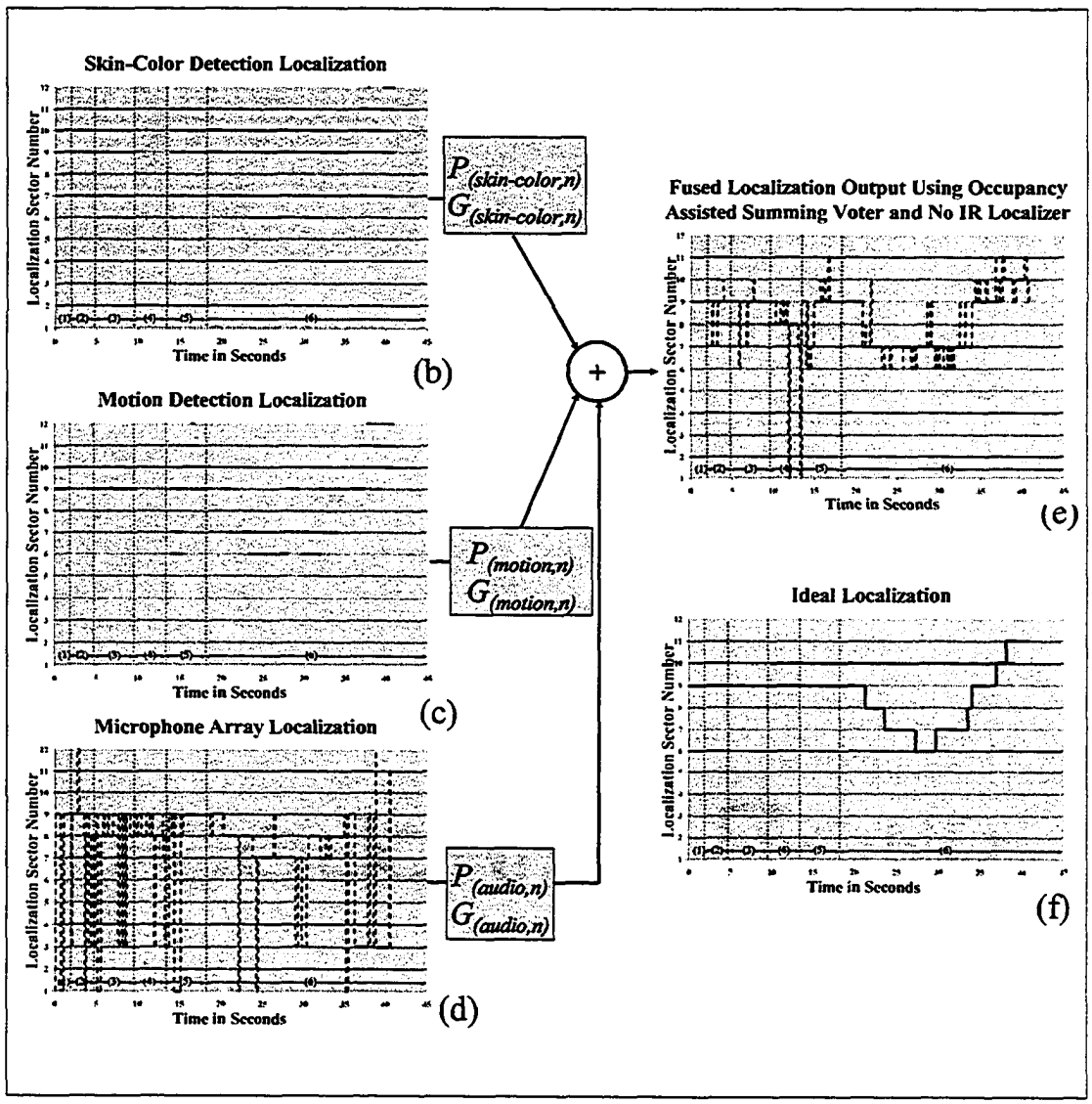

Figure 8-6. Joint audio-video localization results using occupancy assisted summing voter fusion.

8.3.3 Joint Audio-Video-IR Multimodal Taker Location Using Occupancy Assisted Bayesian Network Fusion

The Bayesian network shown in Figure 7-2 and the fusion Equation (7-7) are expanded to include the IR localizer. Figure 8-7 shows the modified Bayesian network.

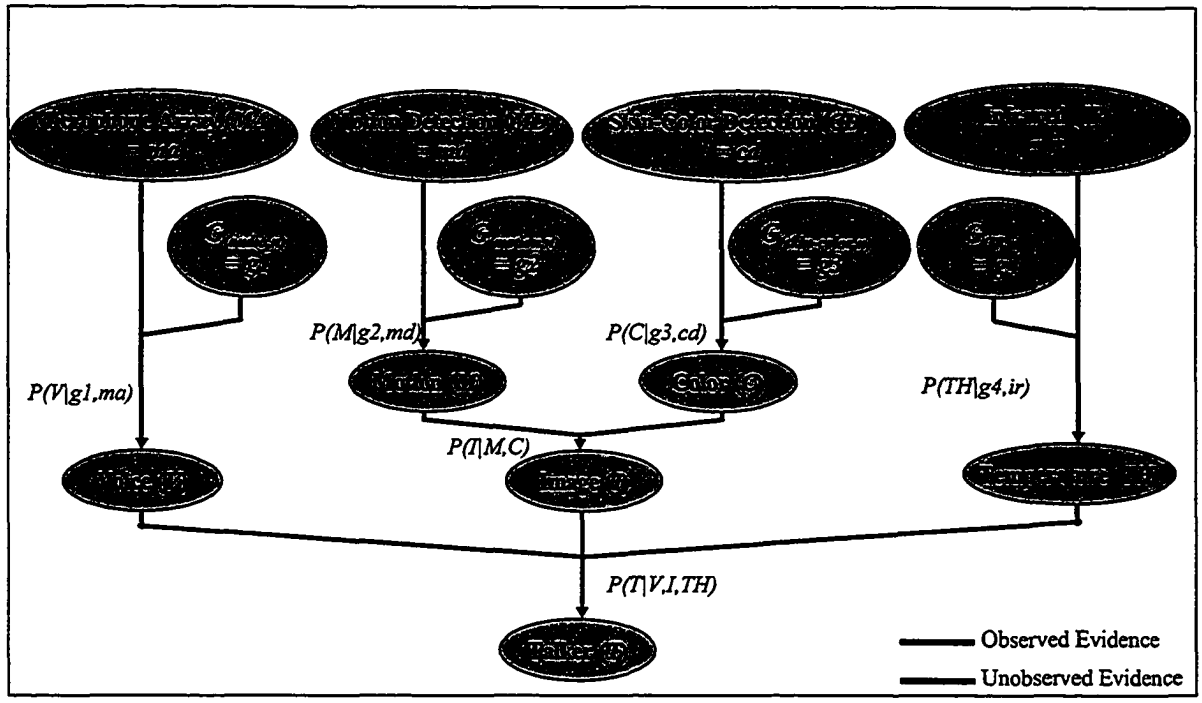

Figure 8-7. Bayesian network for the joint audio-video-IR talker localization system.

The observed evidence e is the localization outputs from the microphone array, the motion detection localizer, and skin-color detection localizer as well as their corresponding occupancy estimates. The localizers are represented by the nodes Microphone Array (*MA*), Motion Detection (*MD*), Skin-Color Detection (*CD*), Infrared Detection (*IR*) and their corresponding occupancy estimations are represented by node

$G_{(audio,n)}$, $G_{(motion,n)}$, $G_{(skin\text{-}color,n)}$, and $G_{(IR,\ n)}$, respectively. Motion (M), Color (C), Voice (V), Image (I) and Temperature (TH) are the unobserved random variables. The Talker node (T) is the talker's location which we are trying to find. Each arrow represents a conditional probability. The values of the observed evidence are represented by $MA = ma$, $MD = md$, $CD = cd$, $IR = ir$, $G_{(audio,n)} = g1$, $G_{(motion,n)} = g2$, $G_{(skin\text{-}color,n)} = g3$, and $G_{(IR,n)} = g4$, respectively. With the observed evidence, (8-3) can be applied onto the inference model and the Talker node (T) can be found using

$$\begin{aligned} P(T|V,I,TH) &= P(ma, md, cd, ir, g1, g2, g3, g4, M, C, V, I, TH, T) \\ &= P(ma)\cdot P(md)\cdot P(cd)\cdot P(ir)\cdot P(g1)\cdot P(g2)\cdot P(g3)\cdot P(g4)\cdot P(V|g1,ma)\cdot \\ &\quad P(M|g2,md)\cdot P(C|g3,cd)\cdot P(TH|g4,ir)\cdot P(I|M,C) \end{aligned} \quad (8\text{-}3)$$

8.3.3.1 Experimental Results

Figure 10-8 shows the fused localization output of the system which used the occupancy assisted Bayesian network as the fusion engine. There is a significant improvement in the overall localization performance when compared with the simple summing voter and the occupancy assisted summing voter systems shown in Figure 8-3 (e) and Figure 8-4 (e). For scenarios (4) and (5), the localization errors which are in the extreme seen in Figure 18-3(e) are removed. There is also improvement in scenario (5). In scenario (6), there is a significant improvement in extracting the talker's motion sequence. Although, there are still some localization errors, the motion sequence is much better defined. The stopping over at sector 6 and the ending at sector 11 can now be seen. At the stop over at sector 6, the fused localization output toggles between sector 6 and 7. This is expected since the localization output of the microphone array is completely wrong while the other three

single modal localizers output both sector 6 and 7 as the active sectors. Depending on the value of the occupancy estimates, the fused output is expected to toggle between the two sectors. The occupancy assisted Bayesian network fusion method shows significant improvement in the localization performance over the other methods studied in this chapter with total error rate of 13.7% and tracking error rate of 13.6%.

Figure 8-9 shows the fused localization results without the IR localizer. Similar to the cause in the occupancy assisted summing voter, without the help of the best performing single modal localizer, the overall localization performance suffers significantly. The degradation is more pronounced in the relatively difficult scenarios (4), (5) and (6).

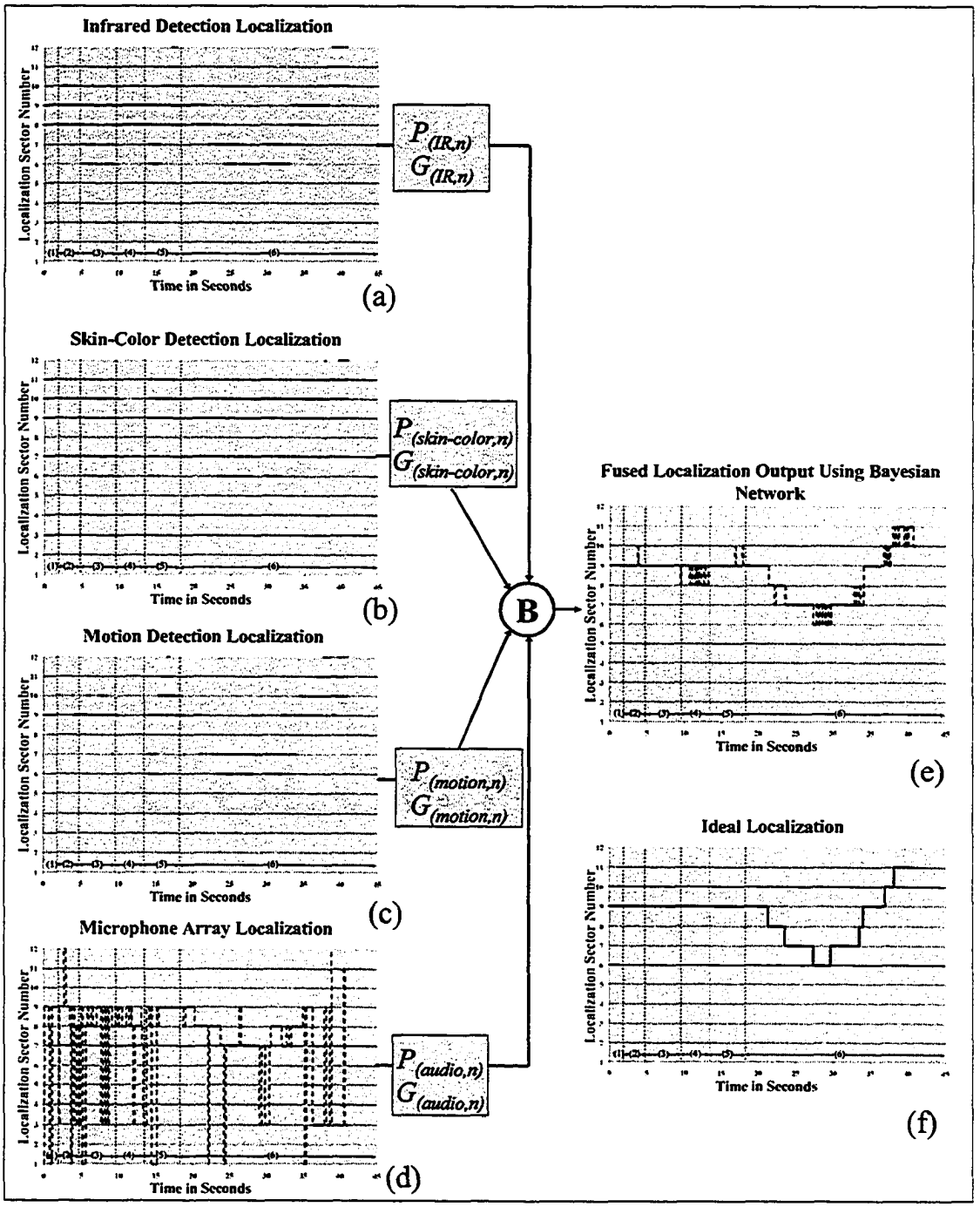

Figure 8-8. Joint audio-video-IR talker localization using occupancy assisted Bayesian network fusion.

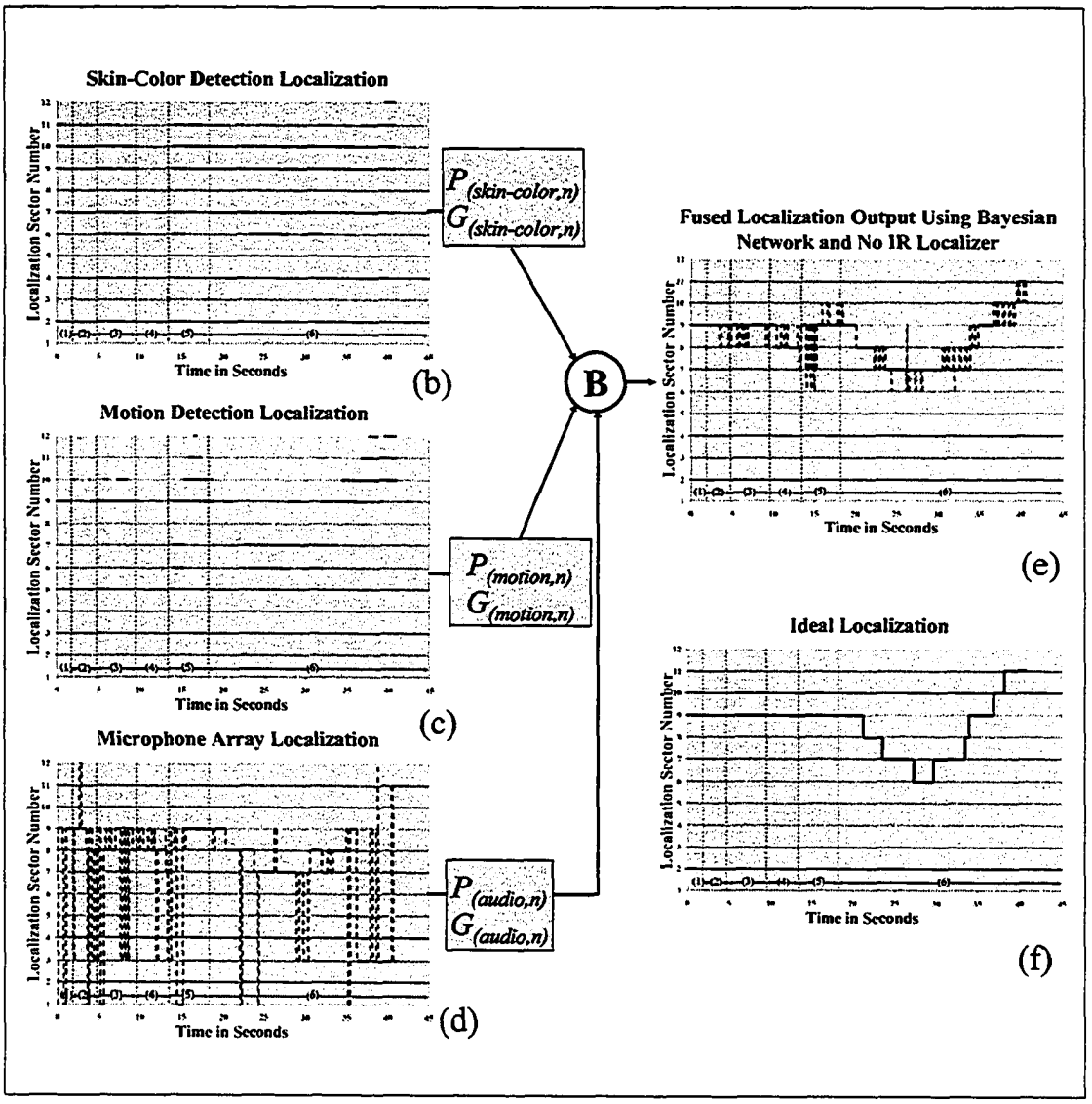

Figure 8-9. Joint audio-video localization using occupancy assisted Bayesian network fusion.

8.4 Conclusions

In this chapter, we demonstrated that the modular multimodal localization architecture is flexible enough to accommodate a new localizer by simply adding a new term in the fusion equation. As an example, we studied the implementation of an IR localization module and its impact on the overall localization performance. Also, how different fusion methods can have different degrees of improvement were investigated.

A challenging environment is used to perform a "stress test" on the multimodal localization system. Some of the single modal localizers suffer from poor localization performance. For example, the microphone array has so many localization errors to the point that it will render the microphone array unusable if it was used as a standalone localizer. The performance of the skin-color localizer degrades substantially as well. The skin-color like furniture confused the localizer causing it to output sector 7, 9, 10 and 11 most of the time. Despite all the performance degradation in the single modal localizers, all three of the multi-modal localizers studied in this chapter perform reasonably well and are significantly better than any one of the single modal localizer. Based on the experimental results, the occupancy assisted Bayesian network is the best performing among all the three fusion methods studied in this chapter.

All the fusion methods studied in this chapter benefit from the additional IR localizer. Again, the occupancy assisted Bayesian network fusion provided the most improvements. This can be understood as the Bayesian network extracts the most information out from

the additional localizer. The additional information is encoded into the Bayesian network through the *a priori* knowledge and the initialization run that is used to populate the conditional probabilities at each node. Through these conditional probabilities, the system collected sensor specific properties which in turn are used during the inference processing used to solve the Bayesian network [PEA88].

Chapter 9 Conclusions and Future Works

9.1 Conclusions

The modular multimodal localization architecture developed in this thesis is very flexible and yet allows individual localizers to be kept relatively simple. With the flexibility of the modular architecture, modality specific refinements can easily be implemented without affecting the operation of other parts in the architecture. However, the trade-off is that a data fusion engine is needed to combine the results from the individual localizers. In this thesis, the deployment of this modular architecture in the area of video conferencing applications is studied as an example.

Video conferencing allows users to communicate more naturally using both sight and sound. Most commercially available video conferencing systems rely only on audio or video to locate the active talker. Acoustic reflections, changes in lighting condition and complex backgrounds often cause audio and video localization failure making localization which uses only one modality unreliable. This thesis investigates the use of multimodal techniques like the joint audio-video and joint audio-video-infrared localization that combines audio, video and infrared information in locating a talker. Audio, video, and infrared localizers often have very different modes of failure, and hence they complement each other very well.

By applying the general modular multimodal localization architecture in video conferencing, the resulting system keeps separate all data streams at the very beginning. Each data stream is then fed into its own purpose specific localizer(s). The localization results from each localizer are then combined using data fusion techniques to generate the final overall localization result. The key is designing a system that can effectively take advantage of the redundancy of information about the talker's location from various modalities. The system has to have the ability to adjust how much it relies on a particular modality when the trustworthiness of a particular modality fluctuates up and down.

In order to study how well and easy the architecture accommodates refinements, a few refinements in audio detection, and in data fusion were studied. A new acoustic reflection detection method which aims to improve the microphone array's ability to perform in the presence of acoustic reflection was studied. Two data fusion methods, the summing voter and Bayesian networks, were studied. The effectiveness of a novel refinement, which uses occupancy estimates, applied to the two fusion methods was also study.

As a refinement in fusion method, this thesis studied how effective it is to improve localization using occupancy information. Using known physical properties of the localizers, the temporal trustworthiness of the localization results are estimated. A summing voter is then used to perform sensor fusion where the occupancy information is used as a weighting function to control the fusion. This method provides the effect of dynamically discriminating unreliable localization results from the audio and the video localizers. The results show that the new method provides additional improvement over fusion methods that do not use the occupancy information, especially in scenarios with

complex background activities or where the talker is walking around in the scene. The results from this study also confirm that when each localizer is used separately as a stand-alone device, they are prone to localization errors. However, when their outputs are fused, the improvements are very noticeable. The addition of the occupancy information as a weighting factor further improves the accuracy of the system. In the case where there are strong acoustic reflections, a pan/tilt/zoom camera is often incorrectly pointed at a wall instead of the talker, or even worse, the camera pans back-and-forth between different reflections rendering the system unusable.

Besides using a summing voter, the effectiveness of a Bayesian network as the fusion engine was also studied. The Bayesian network is constructed with the occupancy estimates. The Bayesian network by itself does not provide much improvement on the overall localization performance. However, with the added occupancy estimates, the improvements are substantial. The total localization error rate drops from 38.7% in Figure 7-1(d) to 3.2% in Figure 7-3(d) with tracking error rates drops from 75% in Figure 7-1(d) to 4.3% in Figure 7-3(d). The improvement results from the Bayesian network incorporating localizer specific information into the fusion engine in the form of *a priori* knowledge. With the help of simple device failure detection, the occupancy estimates provide a means for the fusion engine to handle failed localizers automatically by biasing the fusion engine away from the failed localizers. This method reduces the system sensitivity to device failure, thus improving the overall robustness. A challenging environment is used as a "stress test" to see how the multimodal system performs under more extreme conditions. Although some of the single modal localizers fail almost

completely, the multimodal system performs reasonably well and is able to locate the talker correctly most of the time. In the cases of the occupancy assisted Bayesian network, occupancy assisted summing voter, and with simple summing voter fusion, the total localization error rates are 13.7%, 29.2% and 30.7%, respectively.

In order to study how well the flexible architecture accommodates additional localization modules, an infrared localizer is added to the joint audio-video system. The changes that are needed to implement the additional localizer are minimal and well contained. All the components that were developed previously for the audio and video localizers are not affected. Only an extra term is needed in the fused equation to make the fusion engine take the new localizer into account. All the fusion methods studied in this thesis benefit from the additional infrared localizer. The occupancy assisted Bayesian network is more effective than the other two fusion methods in reaping the benefit of the additional localizer.

9.2 Future Works

The modular multimodal architecture developed is very flexible. Other sensors and localization methods can also be added as drop-in modules to improve the overall performance. This chapter looks at a few that can potentially be added. The examples are just some suggestions, and are not a comprehensive list.

9.2.1 Multiple Cameras

Multiple cameras provide a different viewing angle. Multiple cameras system is particularly useful when the active talker is occluded from the view of one of the cameras. A typical setup for multiple cameras often has one or more cameras viewing the active talker from far field and the other camera(s) zoomed-in with a close-up view on the active talker capturing details of his actions and his facial expressions.

9.2.2 Stereoscopic Camera

Single camera vision systems suffer from a many-to-one mapping problem [DUD73]. Any point in space which is along the same line-of-sight will give the same coordinates in the image plane [DUD73]. Consequently, the depth information is lost [LOC94]. Stereoscopic camera uses a pair of cameras which are placed apart from each other by distance l and viewing from a parallax angle α as shown in Figure 9-1. With a stereoscopic camera system, the three dimensional location of the talker can be found using six calibrated points with known locations [LOC94]. This will allow the camera to realize its full spatial resolution, and provide much higher localization precision.

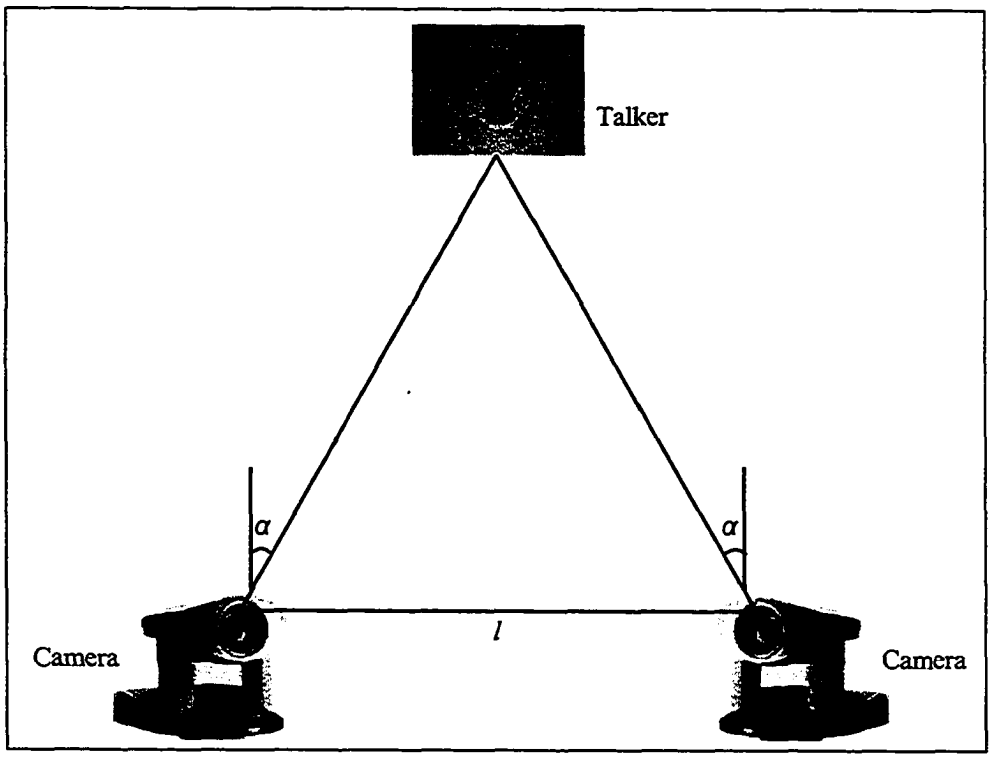

Figure 9-1. Taker localization using stereoscopic camera system.

9.2.3 Multiple Microphone Arrays

Similar to a single camera system, a single microphone array also suffers a many-to-one mapping problem. Although, the microphone array used in this thesis optimizes the detection sensitive at 1 m away from the array, any talker located in the same sector at different distances away from the microphone array will activate the same sector. However, with an additional microphone array placed at a separate location, voice activities from different talkers will activate different sectors on the other microphone array, and therefore can be distinguished. Figure 9-2 demonstrates the scenario. With only microphone array A, voice activities from both talkers will activate sector 11.

However, with the addition of microphone array B, voice activity from talker 1 will activate sector 2 while voice activity from talker 2 will activate sector 12 in microphone array B.

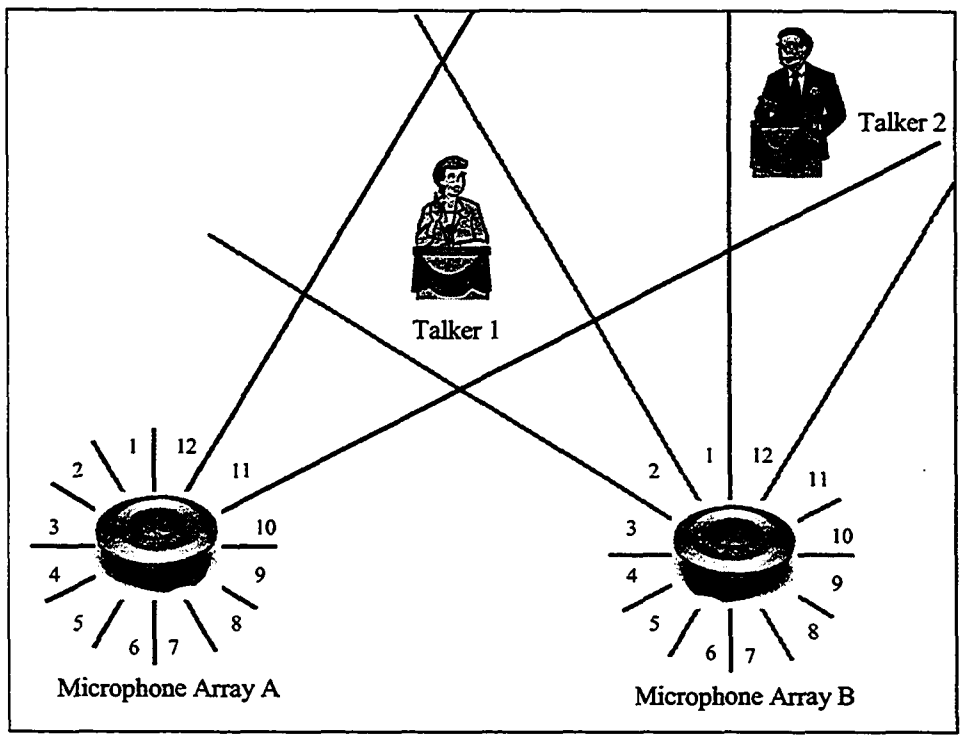

Figure 9-2. Talker localization using two microphone arrays.

9.2.4 Other Applications

Multimodal talker localization provides a more reliable alternative to single modality talker localization. In this thesis we used video conferencing application as a case study in applying the modular multimodal localization architecture and how it can be realized. Each component is kept simple on purpose. However, by joining audio, video, and infrared localization methods, new dynamics are formed. Therefore, there is a need to study how to optimize the architecture for different applications.

Also, how this architecture can be applied is really an open-ended question. With the flexibility and the adaptability nature of the modular multimodal architecture, the architecture and the methods developed can be equally valuable to other applications like surveillance and mobile robotics.

References

[AFF96] S. Affes and Y. Grenier, "A source subspace tracking array of microphones for double talk situations," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, vol. 2, pp. 909 -912.

[AGA03] G. Agarwal, A. Anbu, and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed MPEG domain," in *Proceedings of International Conference on Multimedia and Expo*, 2003, vol. 2, pp. 133-136.

[ARC95] C. Archibald and P. Kwok, *Research in Computer and Robot Vision*, Singapore: World Scientific Publishing Co., 1995.

[BAC97] H. Bacakoglu and M. Kamel. "A three-step camera calibration method," *IEEE Transactions on Instrumentation and Measurement*, vol. 46, no.5, pp. 1165-1172, October 1997.

[BED94] S. Bedard, B. Champagne, and A. Stephenne, "Effects of room reverberation on time-delay estimation performance," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1994, no. 2, pp. 261-264.

[BER01] L. M. Bergasa, M. Mazo, A. Gardel, M. A Sotelo, and L. Boquete, "Unsupervised and adaptive Gaussian skin-color model," *Image and Vision Computing*, vol. 18, pp. 987-1003, Sept. 2001.

[BIR98] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 1998, pp. 232-237.

[BRA99] M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2914-2919, 1999.

[BRA01] M. Brandstein and D. Ward, *Microphone Arrays. Signal Processing Techniques and Applications*. New York: Springer-Verlag, 2001, pp. 3-16.

[CHA02] T. Chaodhury, J. M. Rehg, V. Pavlovic, and A. Pentland, "Boosted learning in dynamic Bayesian networks for multimodal detection", in *Proceedings of the Fifth International Conference on Information Fusion*, 2002, vol. 1, pp. 550 – 556.

[COL99] A. Colmenarez, B. Frey, and T. Huang, "Detection and tracking of faces and facial features," in *IEEE International Conference on Image Processing*, Oct. 1999, pp. 657-661.

[CUC03] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337-1342, Oct. 2003.

[DAN03] R. M. Dansereau, C. Li, and R. A. Goubran, "Lip feature extraction using motion, color, and edge information," in *Proceedings of The 2nd IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications*, 2003, pp. 1-6.

[DAV97] A. L. Davis, "An integrated solution for effective video alarm verification," in *Proceedings of IEEE International Conference on Security Technology*, 1997, pp. 154-157.

[DEC00] D. DeCarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Internal Journal of Computer Vision*, vol. 38, no. 2, pp. 99-127, July 2000.

[DUD73] R. O. Duda, *Pattern Classification & Scene Analysis*. New York: John Wiley & Sons Inc., pp. 379-404.

[DUD01] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. USA: John Wiley & Sons, Inc., 2001, pp. 20-64.

[ELF89] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46-57, June 1989.

[FER01] R. Féraud, O. J. Bernier, J.-E. Viallet, and M. Collobert, "A fast and accurate face detection based on neural network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 42-53, Jan. 2001.

[FIA04] M. Fiala, D. Green, and G. Roth, "A panoramic video and acoustic beamforming sensor for videoconferencing," in *Proceedings of IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications*, October 2004, pp. 47-52.

[GOU03] R. Goubran, D. C-L Lo, M. Nasr, D. Schulz, and G. Thompson, "Self-discovery method," British Patent application number 0330253.6, December 31, 2003, patent pending.

[HOL00] G. C. Holst, *Common Sense Approach to Thermal Imaging*, Washington: The International Society for Optical Engineering, 2000, pp. 46-138.

[HSU02] R. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706, May 2002.

[IAN96] G. Iannizzotto and L. Vita, "A fast, accurate method to segment and retrieve object contours in real images," in *Proceedings of International Conference on Image Processing*, 1996, vol. 1, pp. 841-843.

[JOH93] D. Johnson and D. Dudgeon, *Array Signal Processing*. New Jersey: Prentice Hall, 1993, pp. 112-113.

[LIC03] Chengliang Li, R. M. Dansereau, and R. A. Goubran, "Acoustic speech to lip feature mapping for multimedia applications," in *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, 2003, vol. 2, pp. 829-832.

[LOC94] C. L. Lo, *Biaxial Strain Study Of Porcine Aortic Valve Using Stereographic Technique*, Master of Engineering Thesis, Canada: University of Western Ontario, 1994, pp. 14-21.

[LOD03A] D. Lo, R. A. Goubran, R. M. Dansereau, G. Thompson, and D. Schulz, "Robust joint audio-video localization in video conferencing using reliability information," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 4, pp. 1132-1139, August 2004.

[LOD03B] D. C-L Lo, R. A. Goubran, D. Schulz, and G. Thompson, "Detecting acoustic echoes using microphone arrays," British Patent application number 0324536.2, October 21, 2003, patent pending.

[LOD04A] D. Lo, R. A. Goubran, and R. M. Dansereau, "Robust joint audio-video localization in video conferencing using reliability information II: Bayesian network fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 4, pp. 1541-1547, August 2005.

[LOD04B] D. Lo, R. A. Goubran, and R. M. Dansereau, "Multimodal talker localization in video conferencing environment," in *Proceedings of IEEE International Workshop on Haptic, Audio Visual Environments and Their Applications*, Canada, 2004, pp. 195-200.

[LOD05] D. Lo, R. A. Goubran, and R. M. Dansereau, "Acoustic reflections detection for microphone array applications," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, May 2005, vol. 2, pp. 1139-1143.

[LOW91] A. Low, *Introductory Computer Vision and Image Processing.* United Kingdom: McGraw-Hill Book Company, 1991, pp. 84-180.

[MAR98] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240-259, May 1998.

[MES02] C. Messom, S. Demidenko, and K. Subramaniam, and G. Gupta, "Size/position identification in real-time image processing using run length encoding," in *Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference*, 2002, vol. 2, pp. 1055-1059.

[MEI99] T. Meier and K. N. Ngan, "Video segmentation for content-based coding," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 9, pp. 1190-1203, December 1999.

[MIT02] Mitel Network, *Source Localizer Design,* Mitel Network internal document number 1P00, Canada: Mitel Network, 2002.

[OMO96] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, vol. 2, pp. 921–924.

[OSB98] W. Osberger and A. J. Maeder, " Automatic identification of perceptually important regions in an image", in *Proceedings of International Conference on Pattern Recognition*, 1998, vol. 1, pp. 701-704.

[PAV00] V. Pavlovic, A. Garg, J.M. Rehg, and T.S. Huang, "Multimodal speaker detection using error feedback dynamic Bayesian networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2000, vol. 2, pp. 34 – 41.

[PEA88] J, Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, California: Morgan Kaufmann Publishers, 1988, pp. 150-197.

[PET96] E. M. Petriu, D. Ionescu, D. C. Petriu, F. C. A. Groen, H. Spoelder, S. K. Yeung, S. Elgazzar, and L. Korba, "Multisensor system for mobile robot navigation," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, 1996, vol. 1, pp. 388 - 392.

[PIN99] G. Pingali, G. Tunali, and I. Carlbom, "Audio-Visual tracking for natural interactivity," in *Proceedings of the seventh ACM international conference on Multimedia*, 1999, pp. 373 – 382.

[POY96] C. A. Poynton, *A technical introduction to digital video*. New York: J. Wiley, 1996, pp. 176-177.

[RAD00] B. D. Radlovic and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 728-737, November 2000.

[RAY97] J. G. Ryan and R. A. Goubran, "Optimum near-field response for microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 19-22, 1997.

[RAY00] J. G. Ryan and R. A. Goubran, "Array optimization applied in the near field of a microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 173-176, March 2000.

[RAY03] J. G. Ryan and R. A. Goubran, "Application of near-field optimum microphone arrays to handsfree mobile telephony," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 2, pp. 390-400, March 2003.

[SAB98] E. Saber and A. M. Tekalp, "Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions," *Pattern Recognition Letters*, vol. 19, pp. 669-680, 1998.

[SEU94] S. K. Yeung, W. S. McMath, E. M. Petriu, N. Trif, and C. Gal, "Teleoperator-aided multi-sensor data fusion for mobile robot navigation," in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 1994, pp. 470-476.

[STR01] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, vol. 18, no.1, pp. 22–31, January 2001.

[STE99] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," in *Proceedings of SPIE AreoSense*, Orlando, USA, 1999, pp. 430-441.

[SEK02] A. S. Sekmen, M. Wilkes, and K. Kawamura, "An application of passive human-robot interaction: human tracking based on attention distraction," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 32, No. 2, pp. 248-259, March 2002.

[TER00] J. C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images," in *Proceedings of IEEE International Conference on Face and Gesture Recognition*, 2000, pp. 54-61.

[TOY00] K. Toyama and E. Horvitz, "Bayesian modality fusion: probabilistic integration of multiple vision algorithms for head tracking," in *Proceedings of Fourth Asian Conference on Computer Vision*, January 2000.

[WAL90] E. Waltz and J. Llinas. *Multisensor Data Fusion.* Artech House: Norwood, MA, 1990.

[WAN98] C. Wang and M. Brandstein, "A hybrid real-time face tracking system," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, vol. 6, pp. 3737-3740.

[WAN99] C. Wang and M. Brandstein, "Multi-source face tracking with audio and visual data," in *IEEE 3rd Workshop on Multimedia Signal Processing*, 1999, pp. 169-174.

[WAN00] C. Wang, S. Griebel, and M. Brandstein, "Robust automatic video-conferencing with multiple cameras and microphones," in *IEEE International Conference on Multimedia and Expo*, 2000, vol. 3, pp. 1585-1588.

[WUH02] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor fusion using Dempster-Shafer theory," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, 2002, vol. 1, pp. 7-12.

[YEU94] S. K. Yeung, W. S. McMath, E. M. Petriu, N. Trif, C. Gal, "Teleoperator-aided multi-sensor data fusion for mobile robot navigation," in *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 1994, pp. 470 – 476.

[ZHE00] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Coherent interference suppression with an adaptive array using spatial affine projection algorithm," *IEEE 52nd Vehicular Technology Conference*, 2000, vol. 1, pp. 105-109.

[ZHE00B] Y. R. Zheng and R. A. Goubran, "Adaptive beamforming using affine projection algorithms," in *Proceedings of 5th International Conference on Signal Processing*, 2000, vol. 3, pp. 1929-1932.

[ZHE03] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "A nested sensor array focusing on near field targets," in *Proceedings of IEEE Sensors Conference*, 2003, vol. 2, pp. 843-848.

[ZHE04] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Experimental Evaluation of a Nested Microphone Array With Adaptive Noise Cancellers," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 3, pp. 777-786, June 2004.

[ZOT00] D. Zotkin, R. Duraiswami, L. Davis, and I. Haritaoglu, "An audio-video front-end for multimedia applications," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2000, vol. 2, pp. 786-791.

[ZOT01] D. Zotkin, R. Duraiswami, and L.S. Davis, "Multimodal 3-D tracking and event detection via the particle filter," *in Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp 20 – 27.