UNIVERSITY OF CALIFORNIA, SAN DIEGO

Audio Visual Information Fusion for Human Activity Analysis

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Signal and Image Processing)

by

Shankar Thagadur Shivappa

Committee in charge:

Professor Bhaskar D. Rao, Co-Chair
Professor Mohan M. Trivedi, Co-Chair
Professor Virginia De Sa
Professor Yoav Freund
Professor Nuno Vasconcelos

2010

The dissertation of Shankar Thagadur Shivappa is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California, San Diego

2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank my family, my friends and my advisors for their guidance, love and support that has helped me complete my dissertation. I have been very fortunate to have amazing teachers and mentors at every stage of my life who have been a great inspiration to me. I thank them all for instilling in me the ability and confidence to continue my studies up to the doctoral level.

I thank my advisors, Prof. Bhaskar Rao and Prof. Mohan Trivedi, for their guidance and support. They have nurtured my interests and led me through several challenging endeavors in the course of my graduate studies. Their deep insight into the scientific as well as human aspects of research has been extremely beneficial to me. I thank them for supporting my research interests and providing me ample opportunities to independently explore and develop my ideas.

I thank my committee members, Prof. Nuno Vasconcelos, Prof. Virginia de Sa and Prof. Yoav Freund for their time, expertise, and constructive comments, which have enriched my research.

I thank my colleagues and friends in the CVRR lab and DSP lab for their continuous support. Much of my research was facilitated by the earlier work done at these labs and the infrastructure and knowledge-base that was put in place by earlier lab members. I thank Dr. Rajesh Hegde, Dr. Sangho Park, Dr. Tarak Gandhi, Dr. Chandra Murthy, Dr. Joel McCall, Dr. David Wipf, Dr. Ethan Duni, Dr. Kohsia Huang, Dr. Junwen Wu, Dr. Wenyi Zhang, Dr. Stephen Krotosky, Dr. Shinko Cheng, Dr. Erik Murphy-Chutorian, Dr. Yogananda Isukapalli, Dr. Brendan Morris, Anup Doshi, Cuong Tran, Ashish Tawari, Sayanan Sivaraman, Siva Subramaniam, Yuzhe Jin and Ali Masnadi-Shirazi for being my mentors, colleagues and ever-willing test-subjects for my various experiments.

I thank our sponsors, University of California Discovery Program, National Science Foundation (RESCUE project), CALIT2 (Smartspaces lab) and the ECE Graduate Fellowship Program. I had the wonderful opportunity to teach and mentor graduate and undergraduate students as a teaching assistant. I thank Prof. Paul Siegel, Prof. Ken Zeger, Prof. Tara Javidi and my advisors for providing me this valuable opportunity. I would like to thank Eldgridge Alcantara for being a

great mentor from whom I learned my many teaching skills. I thank the ECE departmental staff, who behind the scenes have made everything go smoothly for me. Specifically, I thank Karol Previte, Rosemary Le, Gennie Miranda, Megan Scott, Bernadette Villaluz, MLissa Michelson, Robert Rome, Carrie Weber, Adolfo Juarez, Rachael Pope and Shana Slebioda for the personal help they have provided on so many occasions.

The text of Chapter 2, in full, is a reprint of the material as it appears in: Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments: A survey", Proceedings of the IEEE, October 2010. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 3, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "An Iterative Decoding Algorithm for Fusion of Multimodal Information," EURASIP Journal on Advances in Signal Processing, Special Issue on Human-Activity Analysis in Multimedia Data, Feb-2008. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 3, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Multimodal Information Fusion using the Iterative Decoding Algorithm and its Application to Audio-visual Speech Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, Las Vegas, Apr-2008. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 4, in full, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Audio Visual Fusion and Tracking With Multilevel Iterative Decoding: Framework and Experimental Evaluation", IEEE Journal of Selected Topics in Signal Processing, Special issue on Speech Processing for Natural Interaction with Intelligent Environments, July 2010. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 5, in part, is a reprint of the material as it appears in:

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Role of Head Pose Estimation in Speech Acquisition From Distant Microphones," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, Taiwan, Apr-2009. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 5, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Hierarchical audio visual information fusion in intelligent meeting rooms," Manuscript submitted to IEEE Transactions on Multimedia for review. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 6, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Hierarchical audio visual information fusion in intelligent meeting rooms," Manuscript submitted to IEEE Transactions on Multimedia for review. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 6, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments: A survey", Proceedings of the IEEE, October 2010. The dissertation author was the primary investigator and author of this paper.

VITA

2003	B. Tech. in Electrical Engineering, Indian Institute of Technology, Madras, India
2004	B. Tech. in Electrical Engineering - Communication Systems, Indian Institute of Technology, Madras, India
2007-2010	Graduate Teaching Assistant, University of California, San Diego
2010	Ph. D. in Electrical and Computer Engineering - Signal and Image Processing, University of California, San Diego

PUBLICATIONS

Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments: A survey", *Proceedings of the IEEE, October 2010.*

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Hierarchical audio visual information fusion in intelligent meeting rooms," *In Submission: IEEE Transactions on Multimedia, 2010.*

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Audio Visual Fusion and Tracking With Multilevel Iterative Decoding: Framework and Experimental Evaluation", *IEEE Journal of Selected Topics in Signal Processing, Special issue on Speech Processing for Natural Interaction with Intelligent Environments, July 2010.*

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "An Iterative Decoding Algorithm for Fusion of Multimodal Information," *EURASIP Journal on Advances in Signal Processing, Special Issue on Human-Activity Analysis in Multimedia Data, Feb-2008.*

Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms", *IEEE CVPR Joint Workshop for Visual and Contextual Learning and Visual Scene Understanding, Jun-2009.*

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Role of Head Pose Estimation in Speech Acquisition From Distant Microphones," *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, Taiwan, Apr-2009.*

Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Person Tracking With Audio-visual Cues Using the Iterative Decoding Framework," *IEEE International Conference on Advanced Video and Signal based Surveillance, AVSS 2008, Santa Fe, Sep-2008* [**Best Paper**].

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Multimodal Information Fusion using the Iterative Decoding Algorithm and its Application to Audio-visual Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, Las Vegas, Apr-2008.*

ABSTRACT OF THE DISSERTATION

Audio Visual Information Fusion for Human Activity Analysis

by

Shankar Thagadur Shivappa

Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)

University of California, San Diego, 2010

Professor Bhaskar D. Rao, Co-Chair
Professor Mohan M. Trivedi, Co-Chair

Human activity analysis in unconstrained environments using far-field sensors is a challenging task. The fusion of audio and visual cues enables us to build robust and efficient human activity analysis systems. Traditional fusion schemes including feature-level, classifier-level and decision-level fusion have been explored in task-specific contexts to provide robustness to sensor and environmental noise. However, human activity analysis involves the extraction of information from audio and visual cues at multiple levels of semantic abstraction. This naturally leads to a hierarchical fusion framework.

In this dissertation, the limitations of existing fusion schemes are explored

and new algorithms are developed to address some of these limitations. The iterative decoding algorithm (IDA) fuses the audio and video modalities at the decision level but unlike other schemes, it uses an iterative strategy to infer the joint likelihood of the hidden states from the unimodal likelihoods. The iterative decoding is advantageous to joint modeling and other decision level fusion schemes in terms of ease of training of the models and the performance under low SNR scenarios. The extension of the IDA to more complex tasks, such as audio-visual person tracking and meeting scene analysis, leads to hierarchical fusion frameworks. The multi-level iterative decoding framework for audio-visual person tracking (MID-AVT) uses the iterative decoding framework for tracking multiple subjects using both audio and visual cues from multiple cameras and microphone arrays. The local sensor-level tracks are fused using the IDA to obtain globally consistent tracks. The MID-AVT framework is robust to sensor calibration errors and requires only a rough calibration step to learn the correspondences between different sensors. The location specific speaker modeling (LSSM) framework for audio-visual meeting scene analysis augments the tracking information with speaker recognition information. Speaker recognition using far-field microphones is a challenging task. The LSSM framework addresses this issue by using the speaker's location information to select the corresponding location specific speaker recognition model. In practice, training such contextual models requires intensive labeling of audio-visual datasets. Semi-supervised techniques for model learning and sensor calibration are presented in this dissertation to address this issue. A particular case, learning the LSSM models using face recognition information, is explored in detail and found to perform well in practice. The overall contribution of this dissertation is the exploration of various aspects of hierarchical fusion in audio-visual human activity analysis and the extensive analysis of these hierarchical fusion frameworks on real world audio-visual testbeds.

Chapter 1

Audio visual human activity analysis

1.1 Introduction

Human activity analysis is the process of inferring and interpreting the identity, actions and intent of human subjects. Human activity analysis systems fall under two main categories.

- Interactive systems such as natural human computer interfaces (HCI), interactive gaming interfaces and interactive robots where the subject(s) cooperate and interact with an intelligent system that responds to their actions and intents.
- Passive observation systems such as surveillance systems, archival and retrieval systems and automatic transcription systems where the intelligent system observes the human subjects .

In either case, the focus is on using sensors to collect information about human subjects and analyzing this sensory data using a suitable mathematical framework to extract relevant information. Historically, research in the field of human activity analysis has drawn considerable inspiration from ability of the human brain to perform such an analysis in a natural and seamless manner. Since

human perception is multimodal in nature, with speech and vision being the primary senses, significant research effort has been focussed on developing intelligent systems with audio and video interfaces[19]. Humans are the ultimate intelligent systems equipped with multimodal sensors and the capability to seamlessly process, analyze, learn and respond to multimodal cues. Humans beings seem to learn the cross-modal correspondences early on and use that along with other techniques to combine the multimodal information at various levels of abstraction. This seems to be the ideal approach to sensory information fusion as exemplified by the success of hierarchical modeling schemes. However significant progress is necessary before computers can begin to process multimodal information at the level of humans. The models and algorithms used in intelligent systems need not be motivated by human information processing alone. However, human cognition can provide valuable insight into the what and how of intelligent systems.

Human activity and interaction is inherently multimodal. Vision and hearing are the primary senses used by humans to comprehend the complex world as well as to communicate with each other. Several psychological studies have outlined the fusion of audio and visual information by humans for performing particular tasks. A classic example is that of lip reading. Another example is that of audio source localization. These studies provide the basis for intelligent system researchers to incorporate either audio or visual or both the modalities in order to accomplish a particular task. It is necessary while designing such systems to evaluate the benefits and costs associated with using both audio and visual sensory modalities as opposed to using just one of them.

1.2 Benefits of audio-visual fusion

The traditional interfaces such as keyboard, mouse and even close-talking microphones are considered too restrictive to facilitate natural interaction between humans and computers. Research efforts have been focussed on developing non-intrusive sensors such as cameras and far field microphones so that humans can communicate through natural means like conversational speech and gestures, with-

out feeling encumbered by the presence of sensors. In other words, the computer has to fade into the background, allowing the users of the intelligent systems to conduct their activities in a natural manner. This necessitates the use of multi-modal, especially audio-visual systems. Audio-visual systems are not restricted to human computer interfaces (HCI) alone. In several applications such as meeting archival and retrieval and human behavioral studies, audio-visual fusion can be applied as a post processing step. The techniques discussed in this thesis are also applicable in this context and not restricted to real-time interfaces.

Though different sensors might carry redundant information as suggested in the previous paragraph, these sensors are rarely equal, in the sense, they carry complementary information too, making it advantageous to use certain sensors over others for certain tasks. This is clearly demonstrated in the case of speech and gesture analysis for HCI applications, where the information carried through gestures complements the information presented through speech. Utilizing both these cues leads to a system that can understand the user more completely than using just one of the modalities.

As a consequence of committing to non-intrusive and natural interfaces, the audio and visual sensors are usually deployed in unconstrained environments and operated in a far-field configuration. In such a setting, background noise and environmental factors significantly affect the performance of the systems. A significant benefit of using multimodal sensors is the robustness to environment and sensor noise that can be achieved through careful integration of information from different types of sensors. This is particularly true in cases where a particular human activity can be deduced from two or more different sensory cues, like for example, audio and lip movements in the case of human speech. Many other tasks like person tracking, head pose estimation, affective state analysis also exhibit significant overlap in the information conveyed over multiple modalities, especially audio and video.

Audio-visual information fusion is not restricted to fusing cues from two sensor streams. Multiple sensors are used in practice in the form of microphone arrays and camera networks. Accurate calibration of such a sensor network is a

difficult task and audio-visual fusion can be used to solve this problem[90][30][72].

Systems designed to analyze multiple human subjects have to cope with yet another complexity. In multi-sensor multi-subject analysis systems, audio-visual cues can be integrated at multiple semantic levels. This leads to hierarchical fusion strategies. However, more complex fusion frameworks require a more elaborate training procedure with additional demands on the size and nature of training datasets. Since it is expensive to collect such labeled datasets, there is significant advantage to using audio-visual systems that can adapt in a semi-supervised manner. Using cross-modal correspondences, it is possible to use the cues from one modality to generate labeled training data for another task. This enables the audio-visual systems to evolve over time and adapt to changing scene and sensor configurations.

1.3 Application domains

As discussed in the previous section, the selection of audio-visual fusion strategies is specific to the scene and sensor configuration. In this section we will briefly outline a few practical application domains.

The most extensively researched domain is that of meeting scenes. The challenge here is to use far-field cameras and microphones to analyze the human activity in a meeting scene, which typically has multiple subjects. A practical example of meeting analysis can be seen in [49]. Far-field sensors are necessitated for developing an unobtrusive system. Tasks such as person tracking, speech recognition, speech enhancement and person identification are performed using audio and visual cues.

Natural human computer interfaces are another domain where audio-visual fusion is critical. Here again, far-field sensors are used, however, the subject is usually co-operative and frequently adapts to the system.

Health smart homes and assisted living for people with disabilities is yet another area where audio-visual systems are needed [32]. This includes passive surveillance of the scene for detecting certain events of interest such as an individual

losing consciousness/mobility as well as active interaction with the subjects.

Intelligent vehicles have advanced significantly and include several driver assistance technologies [100][98]. Such driver assistance systems and the interaction with the car's infotainment system could benefit significantly by the use of both audio and visual cues[96]. Speech recognition, person identification, affect analysis are tasks of interest in this context.

Several psychoanalytical studies involve the segmentation and labeling of audio-visual recording of subjects. Using audio-visual fusion framework to develop segmentation algorithms has a great potential in making this process more efficient and affordable.

In figure 1.1, we present audio-visual testbeds involving meeting scenes, natural HCI and intelligent vehicles. Though audio-visual fusion is not commonly employed in the real-world applications at present, there is a lot of potential that needs to be explored and these testbeds are a first step in that direction.

1.4 Challenges in audio-visual information fusion

- Synchronization of audio and video sensors is a primary challenge in developing audio-visual systems for human activity analysis. In some tasks, such as speech recognition, synchronization of audio and video frames is critical. In other tasks, such as person tracking, the synchronization requirements can be relaxed. In any case, cameras and microphones need to capture the signals from the scene in a time synchronous manner. This challenge is acute in systems that require multiple audio-visual sensors.

- Multiple cameras and multiple microphones are needed to design systems that can observe human activity in an unconstrained scene. Cameras and microphones in such a system need to be calibrated with respect to each other and with respect to the world co-ordinates. This poses a significant challenge, especially in practical deployment of systems.

- Suitable mathematical models are needed to infer human activity from noisy sensory observations of the scene. Most of the research in computer vision

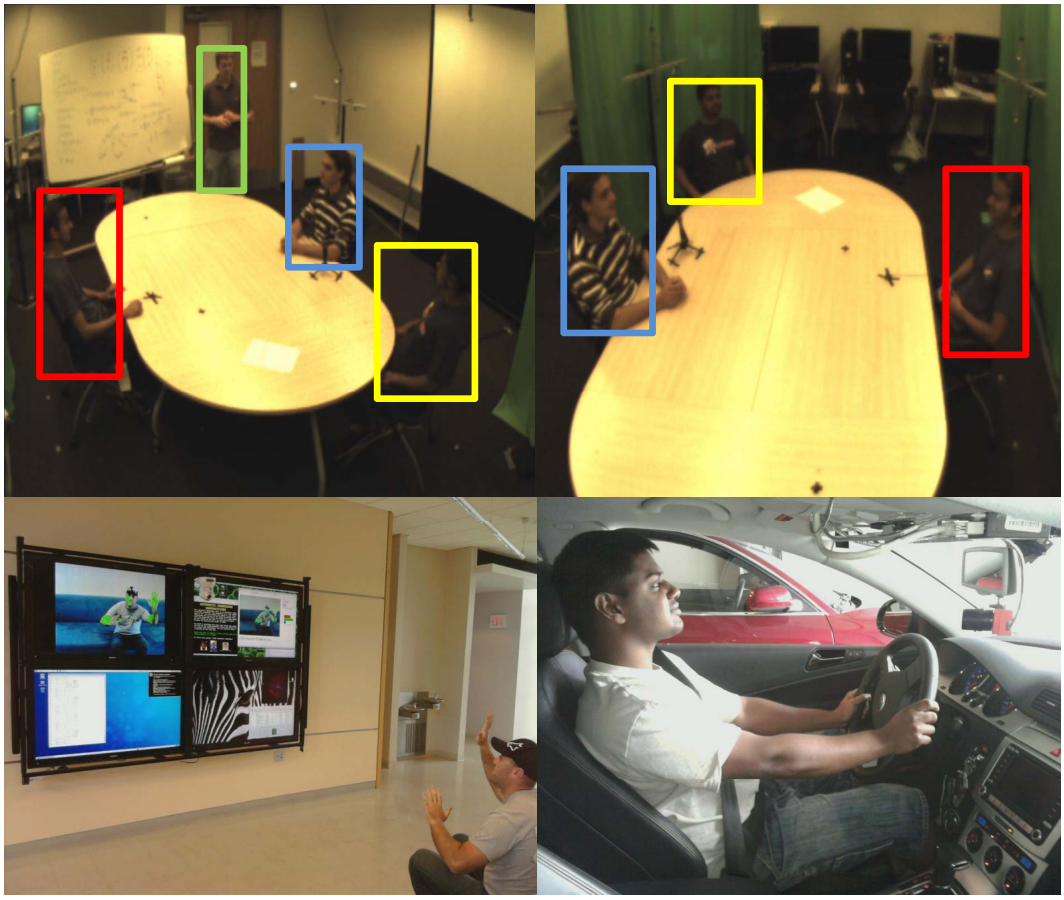

Figure 1.1: Audio-visual testbeds involving a meeting scene, natural HCI and an intelligent vehicle. These are some of the examples where the benefits of audio-visual fusion are being demonstrated in real-world situations.

and speech processing has evolved independently. Audio-visual fusion requires the integration of audio and visual modeling techniques and in many cases new mathematical models have to be developed for this purpose.

- Standard audio-visual datasets are necessary for developing fusion strategies, training the audio-visual models and evaluating their performance. Such datasets are very difficult to collect, even harder to annotate and consequently not readily available. Also, given the broad scope of human activity analysis, datasets collected for one specific task are not well suited for other tasks.

1.5 Research contributions

Cameras and microphones are ubiquitous and the current challenge is no longer the cost of deploying these sensors. Computational resources necessary to process the multiple data streams might be a limitation in some applications such as in mobile devices. However, the ongoing trends indicate that computational power will not be a bottleneck either. In order to be successful, an audio-visual system needs to use effective fusion strategies. As outlined in the later chapters, several audio-visual tasks, their corresponding suitable feature sets and fusion strategies have been explored by the research community. However, the selection of a suitable fusion strategy for a particular task at hand is non-trivial and requires domain expertise. This is a significant hurdle in the widespread deployment of audio-visual systems. Future research needs to address this challenge by developing adaptive and context based fusion strategies. Online learning and automatic sensor calibration strategies will play a major role in the next generation of audio-visual systems.

In this thesis, the challenges outlined in Section 1.4 are addressed in detail. Audio-visual testbeds and the corresponding techniques for synchronized capture of audio and video signals are described in detail. The challenges and techniques in calibrating such sensor networks are presented. Fusion schemes of varying levels of complexity for audio-visual fusion in a wide range of tasks are developed.

- Iterative decoding algorithm (IDA) for the fusion of temporal streams of audio and visual information.
- Multi-level iterative decoding algorithm (MID) for hierarchical fusion of audio and visual cues in more complex tasks.
- Contextual modeling framework for fusion of audio and visual information at different levels of semantic abstraction.
- A semi-supervised learning framework for learning the contextual models.

The Iterative decoding algorithm (IDA) for the fusion of temporal streams of audio and visual information is based on the principle of iterative decoding used

in turbo codes. The IDA fuses the audio and video modalities at the decision level but unlike other decision level fusion schemes, it uses an iterative scheme to infer the joint likelihood of the hidden states from the unimodal likelihoods obtained from the audio and video models. The IDA is advantageous to joint modeling and other decision level fusion schemes in terms of ease of training of models and performance under low SNR scenarios. However, the utility of the iterative decoding algorithm is limited by the fact that the multiple observation streams need to correspond to the same underlying hidden state sequence for effective inference. In practice however, audio-visual human activity analysis includes the observation of multiple subjects using multiple sensors. In such a situation, further steps are required to associate data with the respective source before iterative decoding based inference can be applied.

The multi-level iterative decoding framework for audio-visual person tracking (MID-AVT) scheme uses the iterative decoding framework for tracking multiple subjects using both audio and visual cues from multiple cameras and microphone arrays. The MID-AVT framework extends the iterative decoding algorithm by including a data association step to select appropriate track hypotheses from different sensor views. The performance of the MID-AVT tracker is similar to the popular particle filter based audio-visual tracker. However there are distinct advantages to the MID-AVT framework. It is modular and hence easy to expand to more number of cameras and microphone arrays or any other sensors that can localize persons. It is also applicable to sensors with overlapping and non-overlapping field of 'view'. Since the placement of the sensors is assumed to be arbitrary but fixed, only a rough calibration scheme is necessary to establish the correspondence between sensors. Moreover, the performance of the MID-AVT tracker is robust to small errors in sensor calibration.

However, tracking human subjects is only the first step in analyzing human activity in an intelligent space. The situational awareness needed in an intelligent space is developed by fusing information at multiple levels of semantic abstraction. When audio-visual fusion is explored in the context of such co-performed tasks, not only is an hierarchical integration of audio and video cues necessary, but it is also

beneficial to the performance of the individual tasks because the output of one kind of human activity analysis task contains valuable information for another such task and by interconnecting them, a robust system results. The location specific speaker modeling (LSSM) framework for audio-visual meeting scene analysis augments the tracking information with speaker recognition information. Speaker recognition using far-field microphones is a challenging task. The LSSM framework addresses this issue by using the speaker's location information to select the corresponding location specific speaker recognition model.

Training the contextual models requires extensive amounts of densely labeled training data. A framework to train the contextual models using minimum amount of supervision will make the hierarchical fusion frameworks more applicable in practice. A particular case, learning the LSSM models using face recognition information, is explored in detail and found to perform well in practice.

1.6 Thesis outline

In Chapter 2 we present a review of existing audio-visual fusion schemes. In Chapter 3 we develop the iterative decoding algorithm (IDA) - a probabilistic fusion framework for fusing information from time sequences of audio and video observations based on the theory of turbo codes. The utility of IDA is demonstrated on speech segmentation as well as speech recognition tasks. In Chapter 4 we extend the iterative decoding to solve the problem of person tracking by including a data association framework. This results in the multi-level iterative decoding based audio-visual tracking (MID-AVT) framework. In Chapter 5 we discuss more elaborate hierarchical fusion schemes and explore the utility of contextual fusion schemes in building robust audio-visual meeting analysis systems. Specifically the location specific speaker modeling framework and the role of head pose estimation in speech acquisition from far field microphones is explored. The utility of these fusion schemes on a real world problem of meeting scene analysis is analyzed and evaluated on an extensive set of real world meeting recordings. The details of the real world testbeds are presented in Appendix A. In Chapter 6, we present a semi-

supervised learning scheme to train the location specific speaker models. Finally we present the concluding remarks and future directions in Chapter 7.

Chapter 2

Audio-visual information fusion schemes: A survey

The varied application domains of multimodal human activity analysis systems have always presented a challenge to the systematic understanding of their information fusion models and algorithms. The traditional approach to information fusion schemes classifies them based on early, late and intermediate fusion strategies and describes their associated merits. Achieving robustness to environmental and sensor noise is the traditional motivation for audio-visual information fusion. This category includes the major part of the multimodal fusion strategies studied so far. The most widely accepted notion of sensory information fusion applies to these systems. Those tasks which involve redundant cues in multiple modalities due to the nature of the human activity, fall under this group. Audio-visual speech recognition is the classic example of such a task. It is also one of the earliest areas to generate considerable research interest in multimodal information fusion techniques. In earlier literature[78][47], fusion strategies have been classified as follows -

- **Signal enhancement and sensor level fusion strategies.**
- **Feature level fusion strategies.**
- **Classifier level fusion strategies.**

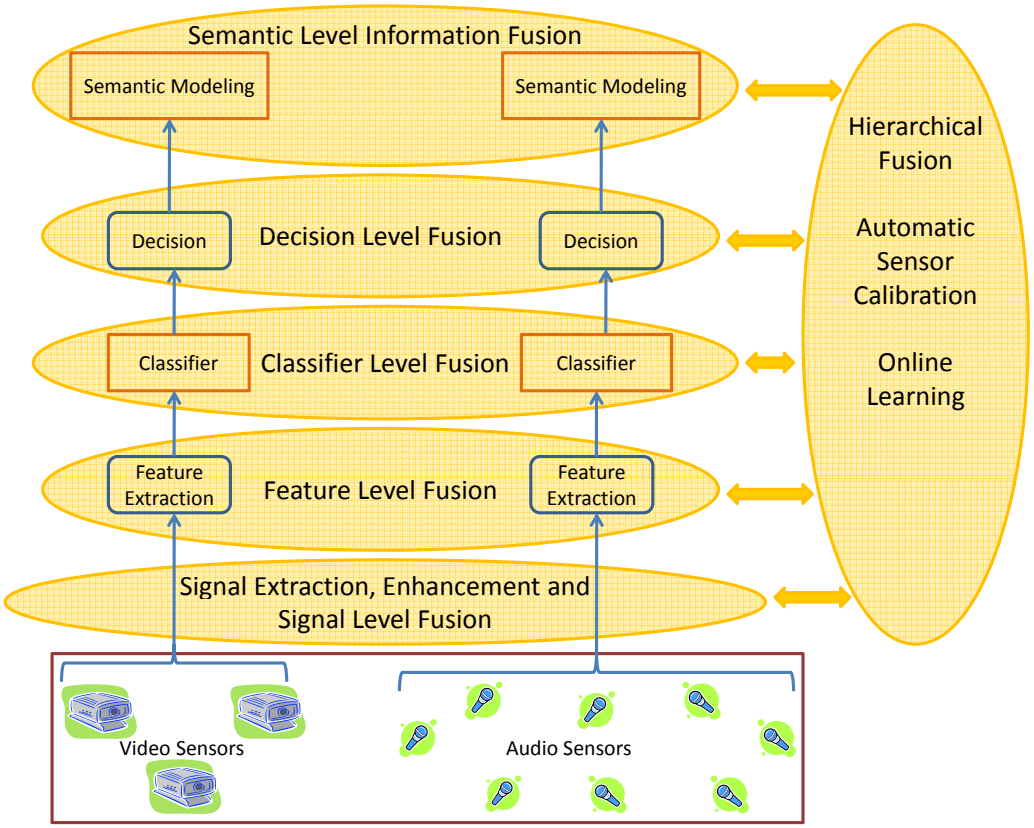

Figure 2.1: Information fusion at various levels of signal abstraction is depicted here.

- **Decision level fusion strategies.**
- **Semantic level fusion strategies.**

2.1 Signal enhancement and sensor level fusion strategies

This includes signal enhancement techniques such as beamforming using microphone arrays. It is conceivable that video information could be useful in the beamforming process as in [89][87][57]. Also, camera networks could benefit from the source localization and pan-tilt-zoom cameras might be able to capture better

images of the scene. However such schemes are rarely described in isolation and are usually part of hierarchical fusion approaches (Section 2.7).

2.2 Feature level fusion strategies

Cognitive scientists refer to this as the early fusion strategy. This is also referred to as the data to decision fusion scheme in literature[99][84]. Some tasks such as automatic speech recognition, person tracking, affect analysis etc produce cues in multiple modalities in a temporally correlated manner. Note that an up-sampling or a down-sampling stage is sometimes necessary in order to align the streams to each other. A representative example is the case of audio signals and lip movements carrying the information about the spoken word in the audio and visual modality respectively. In these cases, an early fusion strategy is feasible. In this case, one concatenates the feature vectors from the multiple modalities to obtain a combined feature vector which is then used for the classification task. Figure 2.2 is a schematic representation of typical feature fusion schemes.

This early fusion has the advantage that it can provide better discriminatory ability for the classifier by exploiting the covariations between the audio and video features[99]. However, the larger dimensionality of the combined feature vector presents challenges for the classifier design. In order to overcome this, standard dimensionality reduction techniques such as DCT, PCA, LDA and QDA are applied. LDA and QDA based systems are know to out-perform PCA based systems in classification tasks. However, in the presence of limited training data, PCA is more stable than LDA[58]. The optimal dimensionality reduction technique also depends on the nature of the classifier used. Theoretically, kernel based classifiers like SVMs do not require an explicit dimensionality reduction step. However, most multimodal systems adopting an early fusion strategy are based on HMM based classifiers and do benefit from dimensionality reduction.

As an example, [70] presents an elaborate scheme for early fusion of audio-visual information for speech recognition which includes both early and late fusion. The early fusion consists of the DCT and multiple PCA steps to reduce the di-

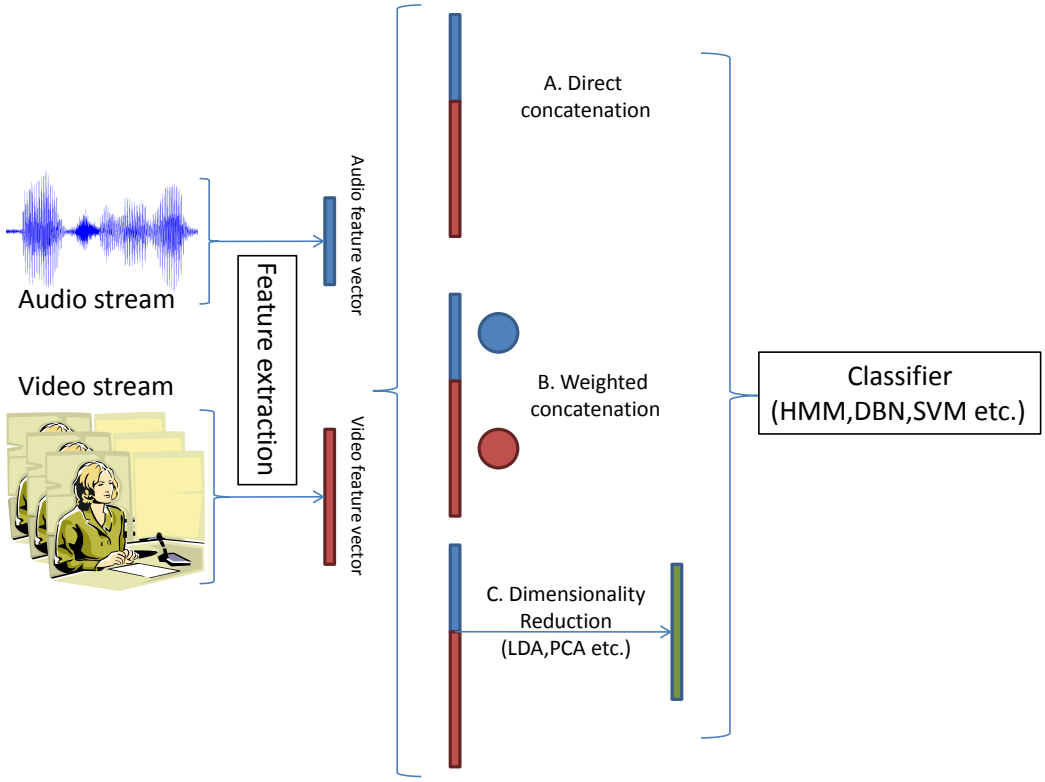

Figure 2.2: Signal and information flow in feature level fusion strategy

mensionality of the audio-visual feature vector. Early fusion strategy with a HMM based classifier is also explored in [59] for the purpose of analyzing group actions in meetings. 39 features including 18 audio features and 21 visual features are concatenated and used to recognize group actions in meeting recordings. This early fusion scheme is second in performance only to an intermediate fusion strategy using asynchronous HMMs (10% vs 9.2% error rates), revealing that the simple early fusion strategy is quite effective if used in the right task. Another example of early fusion for audio-visual tracking can be seen in [72]. Here the microphone arrays and cameras are treated as generalized directional sensors and treated equivalently. [36] proposes an iterated extended Kalman filter (IEKF) for audio-visual source tracking by concatenating audio and visual features.

The early fusion technique has the advantage of being the simplest to implement and is suitable for those applications which require very fast processing of

cues. However, it cannot be applied to most tasks where strictly temporally synchronized cues are not present. Also, the feature concatenation performs poorly when the reliability of the different modalities during the training phase differ from the actual operation phase.

2.3 Classifier level fusion strategies

Cognitive scientists refer to this as the intermediate fusion strategy. This is typically encountered in cases where HMMs (and their hierarchical counterparts) and Dynamic Bayesian networks are used to model individual streams. In such cases, the information can be fused within the classifier, but after processing the feature vectors separately. Thus a composite classifier is generated to process the individual data streams. The intermediate fusion strategy is an attempt to avoid the limitations of both early and late fusion strategies . Unlike early fusion, fusion at the classifier level does allow the weighted combination of different modalities based on their reliability[33]. These weighted combinations however are taken on each frame, allowing for a much finer combination of cues than in late fusion. Such fusion schemes are widely used in audio-visual speech recognition systems. Figure 2.3 is a schematic representation of typical intermediate fusion schemes.

Asynchrony between the different streams can be modeled to some extent. This is critical in cases such as audio-visual speech recognition where the audio and video asynchrony is of the order of 100ms whereas the frame duration is typically 25ms[38][78]. Different degrees of asynchrony are allowed at the cost of complexity and speed. The multistream HMM[27][62] assumes perfect synchrony between the different streams. On the other extreme is the model that allows complete asynchrony between the streams. This is however infeasible due to the exponential increase in the number of state combinations possible due to the asynchrony. An intermediate solution is given by the product HMM [103] or the coupled HMM [68]. In case of audio-visual speech recognition, this corresponds to imposing phone synchrony as opposed to the frame synchrony of the multistream HMM.

The coupled hidden Markov model and the multistream hidden Markov

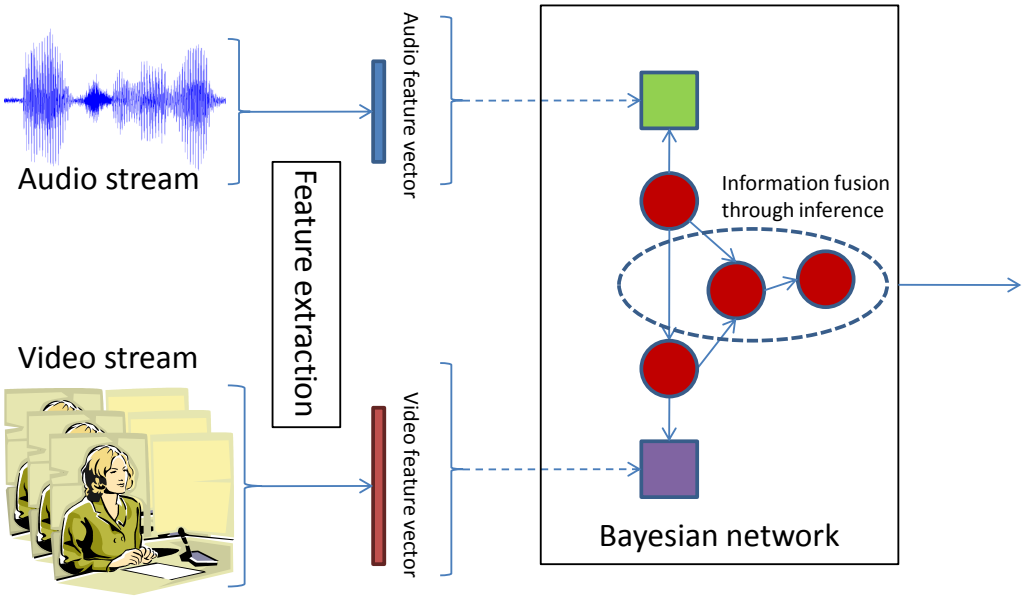

Figure 2.3: Signal and information flow in classifier level fusion strategy

model have been used to improve the performance of audio-visual speech recognition [27, 68, 86]. These schemes have also been applied in other areas of research such as biometrics[29], audio-visual head pose estimation using the particle filter framework[16], audio-visual person tracking[95][88][71][6], audio-visual aggression detection[110]. However, in the real-world situations, the reliability of the different streams varies with time. For example, the video channel in audio visual speech recognition might be completely unreliable if the speaker covers the mouth with the hand or turns away from the camera[85]. In this case, it is useful to be able to estimate the reliability of each channel continuously and weight them accordingly. Stream weight estimation and its adaptive counterparts have been presented in literature[78][40]. The iterative decoding algorithm [85] solves this problem by using techniques burrowed from turbo codes [9]. The iterative decoding algorithm has been applied to the problem of audio visual speech recognition[86] on the GRID audio-visual speech corpus[21] and to the problem of person tracking using the audio-visual cues in [88].

2.4 Decision level fusion strategies

Late or decision level fusion involves the combination of probability scores or likelihood values obtained from separate uni-modal classifiers to come up with a combined decision. In cases where strictly temporally synchronized cues are absent, late integration is still feasible. Typically late fusion involves using independent classifiers, one for each modality and combining the likelihood scores based on some reliability based weighting scheme. The training and decoding these unimodal models scales linearly in the number of streams which makes these schemes particularly attractive. The reliability of the streams is typically used by exponentially weighting the probability scores from individual streams before taking their product. Such a combination scheme with appropriate weighting scheme has been used for audio-visual speech recognition[27]. However, in case of audio-visual speech recognition, the late fusion strategy has been shown to be inferior to the intermediate fusion strategy discussed in the previous section[27]. Figure 2.4 is a schematic representation of typical decision fusion schemes.

The weighting scheme used in late fusion draws upon the work in combination theory to estimate the best weighting factors based on the training data. This is however a limitation when there is a mismatch between the training database and the actual operation. As with the intermediate fusion strategy, decision fusion allows for separate weighting of the different streams based on the reliability. However the fusion is not at the level of frames but at a higher levels. For example, in the audio-visual speech recognition context, the decision level fusion could take place at the utterance level. Decision level fusion allows maximum flexibility in the choice of individual classifiers. [56] explores the use of decision level fusion for audio-visual person identification. The lack of state correspondences in the text independent person ID task imposes the late fusion strategy in this case. The authors also acknowledge the importance of optimal weighting in the decision fusion. [92] is another audio-visual person identification system based on decision level fusion. [112] describes an audio-visual affect recognition which uses decision level fusion to combine facial expressions and prosodic cues for affective state recognition.

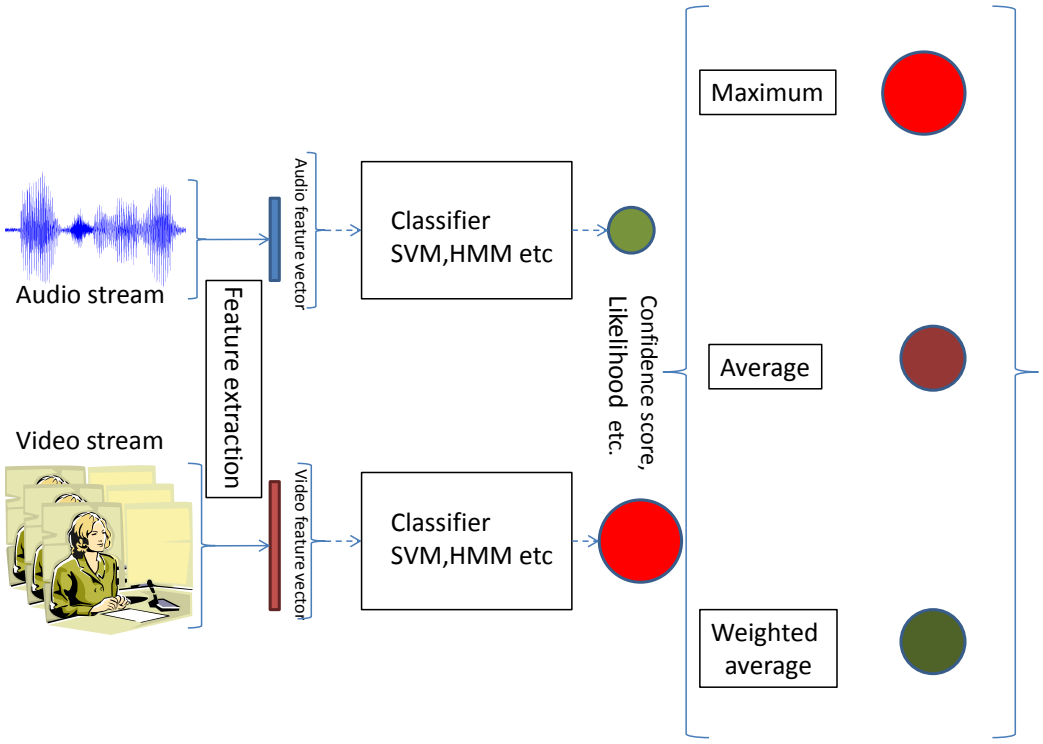

Figure 2.4: Signal and information flow in decision level fusion strategy

2.5 Hybrid fusion strategies

A combination of the above mentioned fusion strategies is also reported in literature. In [78] a combination of feature level fusion with decision level fusion is used in the context of an audio-visual speech recognition task. The audio and visual feature are combined early on through a discriminatory feature selection process and the discriminatory features are used again as one of the streams in a multi-stream based decision fusion technique. There is no theoretical basis for such a scheme, however in practice, it is shown to improve the recognition accuracy. Canonical correlation analysis (CCA) is a statistical approach that combines linear dimensionality reduction and fusion by computing linear projections that are maximally correlated. It is a combination of early and late fusion strategies. [83] applied CCA to a open-set speaker identification problem. More recently, a spectral diffusion framework has been proposed to provide a uniform embedding of data for multisensory fusion [52].

2.6 Semantic level fusion strategies

It is conceivable that higher level information can be merged after the semantic interpretation of the sensory information. This is beyond the scope of our survey because usually such fusion schemes will involve other modalities like text, webpages and other such sources of information that are amenable to semantic interpretation.

2.7 Hierarchical fusion strategies

Traditional fusion schemes, as described so far, have mostly focussed on task based fusion schemes. Audio-visual fusion for speech recognition, person tracking, person identification, emotion recognition have been explored and are also areas of active research. However, in practice, several such tasks have to be co-performed to provide the situational awareness that is required by an effective intelligent system. For example, an intelligent robot will be expected to simultaneously perform the tasks of speaker localization, speech recognition, speaker identification and emotion recognition in order to provide a wholesome communication experience. Similarly, a meeting scene analysis system requires the tracking of human subjects, person identification, speaker localization and speech recognition to automatically analyze meeting scenes.

One of the early research studies in observing human activities in an instrumented room is described in [101]. A graphical summary of the human activity was generated. The audio and visual information was used in identifying the current speaker based on a rule based decision fusion. [11] and [39] describe another meeting room analysis system which also fuses audio-visual stream for person identification, in addition to using the audio for automatic transcription and archival purposes. [79] investigates speech, gaze and gesture cues for high level segmentation of a discourse into topical segments based on a psycholinguistic model. In [73], the authors propose a hierarchial HMM framework for modeling human activity. More recent hierarchical fusion strategies include [113][74][23][8]. In [8], the authors develop a probabilistic integration framework for fusion of audio visual cues

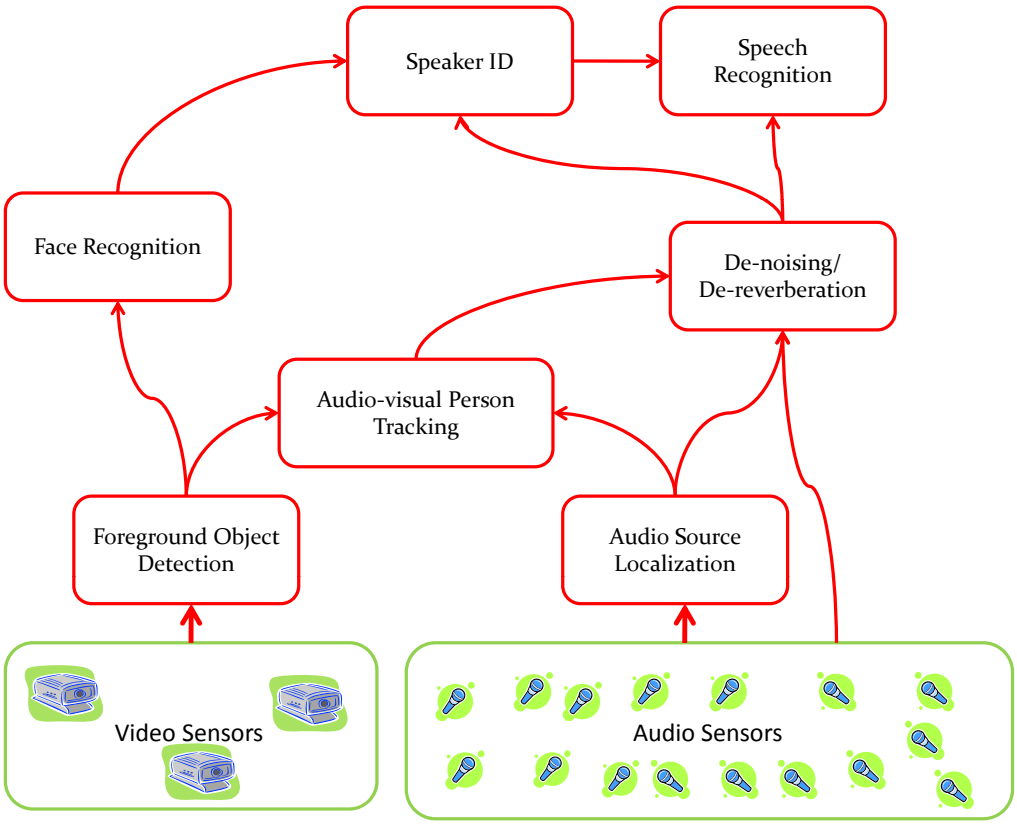

Figure 2.5: Flowchart summarizing the exchange of audio and visual cues at multiple levels of semantic abstraction in a meeting analysis system.

at the track and identity levels. This is an example of fusion at multiple levels of abstraction. Similarly, in [87], the utility of head pose estimation and tracking for speech recognition from distant microphones is explored.

When audio-visual fusion is explored in the context of such co-performed tasks, not only is an hierarchical integration of audio and video cues necessary, but it is also beneficial to the performance of the individual tasks because the output of one kind of human activity analysis task contains valuable information for another such task and by interconnecting them, a robust system results. The interconnected blocks in a hierarchical fusion framework is illustrated in Figure 2.5.

In hierarchical fusion schemes, audio-visual information fusion can involve the following scenarios

- **Reduce the search space in classification tasks** - Audio or visual cues can be used to restrict the search space for classification tasks using the corresponding other modality. In case of parametric statistical model based classification, this can be achieved by having specially trained models for different contexts and switch between these models using the audio or visual cues. Typically, the statistical models are easier to train and have better performance for individual contexts. The task of audio-visual fusion in this case is to robustly identify the particular context based on complimentary cues. In other cases where classification models are based on minimum distance or maximum likelihood, the audio and visual cues can be used to restrict the set of possibilities over which the minimum/maximum is evaluated.

- **Semi-supervised/unsupervised learning of classification models** - The audio and visual cues can be used to select a training set for training statistical models for the classification tasks. This is particularly necessary for the contextual modeling stated above, to be successful. The audio-visual fusion reduces the effort to label the training sets and leads to a semi-supervised or in certain cases, unsupervised learning algorithms that can automatically update the contextual models. The challenge lies in identifying the cues based on their ease and robustness of detection and the minimum supervision needed in labeling the data.

- **Calibration of sensors** - In cases where multiple cameras and microphones are used to collect audio and visual cues, the calibration of the sensors with respect to each other and with respect to the room co-ordinates is an important issue. One solution is to develop algorithms that work in the sensor co-ordinates. Another approach is to develop sensor calibration techniques that use the cameras and microphones together to calibrate each other.

- **Traditional fusion strategies (as describes in earlier sections)** - The most commonly encountered examples of audio visual fusion in literature are cases where the audio and visual modalities carry complimentary information from the same underlying process as in the case of acoustic waveforms and lip

movements conveying information about the underlying speech segment. The fusion challenge is to develop inference algorithms to decipher the underlying process based on the audio and visual observations. This is achieved by fusing the cues at the feature, classifier or decision levels.

In the following chapters we present several fusion frameworks illustrating these fusion paradigms on real world human activity analysis tasks.

2.8 Acknowledgments

The text of Chapter 2, in full, is a reprint of the material as it appears in: Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments: A survey", Proceedings of the IEEE, October 2010. The dissertation author was the primary investigator and author of this paper.

Chapter 3

A Probabilistic framework for audio-visual information fusion

3.1 Introduction

Fusion of information from different streams is a big challenge in multimodal systems. So far, there has not been any standard fusion technique that has been widely accepted in the published literature. Graphical models have been widely discussed as the most suitable candidates for modeling and fusion information in multimodal systems[47]. Since human activity is usually a temporal sequence of events and activities, dynamic Bayesian networks and hidden Markov models are commonly employed to model human activity.

Information fusion can occur at various levels of a multimodal system. A sensor level fusion of video signals from a normal and an infrared camera is used for stereo analysis in [55]. At a higher level is the feature level fusion. The audio and visual features used together in the ASR system built at the John Hopkins University, 2000 workshop [69] is a good example of feature level fusion. Fusion at higher levels of abstraction(decision level) have also been proposed. Graphical models have been frequently used for this task[46]. Fusion at the sensor level is appropriate when the modalities to be fused are similar. At the feature level, fusing more disparate sources becomes possible. At the decision level, all the

information is represented in the form of probabilities and hence it is possible to fuse information from a wide variety of sensors. In this chapter we develop a general fusion algorithm at the decision level.

In this chapter, we develop a fusion technique in the Hidden Markov model (HMM) framework. HMMs are a class of Graphical models that have been used traditionally in speech recognition and human activity analysis[73].

3.2 Advantages of the iterative decoding scheme

A good fusion scheme should have lower error rates than those obtained from the unimodal models. Both the joint modeling framework and the iterative decoding framework have this property. Multimodal training data is hard to obtain. Iterative decoding overcomes this problem by utilizing models trained on unimodal data. Building joint models on the other hand requires significantly greater amounts of multimodal data than training unimodal models due to the increase in dimensionality or complexity of the joint model or both. Working with unimodal models also makes it possible to use a well-learned model in one modality to segment and generate training data for the other modalities, thus overcoming the problem of lack of training data to a great extent.

In many applications like ASR, well-trained unimodal models might already be available. Iterative decoding utilizes such models directly. Thus, extending the already existing unimodal systems to multimodal ones is easier. Another common scheme used to integrate unimodal HMMs is the product HMM [43]. Simulations show that the product rule performs as well as the joint model. But the product rule has the added disadvantage that it assumes a one-one correspondence between the hidden states of the two modalities. The generalized multimodal version of the iterative decoding algorithm in section 3.4.2, relaxes this requirement. Moreover, the iterative decoding algorithm performs better than the joint model and the product HMM in the presence of background noise, even in cases where there is a one-one correspondence between the two modalities.

In noisy environments, the frames affected by noise in different modalities

are at best non-overlapping and at worst independent. The joint models are not able to separate out the noisy modalities from the clean ones. Because of this reason, the iterative decoding algorithm outperforms the joint model at low SNR. In the case of other decision level fusion algorithms like the multistream HMMs [111] and reliability weighted summation rule [28], one has to estimate the quality(SNR) of the individual modalities to obtain good performance. Iterative decoding does not need such apriori information. This is a very significant advantage of the iterative decoding scheme because the quality of the modalities is in general time-varying. For example, if the speaker keeps turning away from the camera, video features are very unreliable for speech segmentation. The exponential weighting scheme of multistream HMMs requires real time monitoring of the quality of the modalities which in itself is a very complex problem.

3.3 Turbo codes and the iterative decoding algorithm

Turbo codes are a class of convolutional codes that perform close to the Shannon limit of channel capacity. The seminal paper by Berrou et al.[9] introduced the concept of iterative decoding to the field of channel coding. Turbo codes achieve their high performance by using two simple codes, working in parallel to achieve the performance of single complex code. The iterative decoding scheme is a method to combine the decisions from the two decoders at the receiver and achieve high performance. In other words, two simple codes working in parallel perform as well as a highly complex code which in practice cannot be used due to complexity issues.

An analogy can be drawn between the redundant information of the two channels of a turbo code and the redundant information in the multiple modalities of a multimodal system. Based on this a modified version of the iterative decoding algorithm to extract and fuse the information from parallel streams of multimodal data can be developed.

Consider a multimodal system to recognize certain patterns of activity in an

intelligent space[102]. It consists of multimodal sensors at the fundamental level. From the signals captured by these sensors, one can extract feature vectors that encapsulate the information contained in the signals in finite dimensions. Once the features are selected, one can model the activity to be recognized, statistically. For an activity that involves temporal variation, Hidden Markov models(HMM) are a popular modeling framework[73].

3.3.1 Hidden Markov Models

Let $\lambda = (A, \pi, B)$ represent the parameters of a HMM with N hidden states, that models a particular activity. The decoding problem is to estimate the optimal state sequence $Q_1^T = \{q_1, q_2 \dots q_T\}$ of the HMM based on the sequence of observations $O_1^T = \{o_1, o_2 \dots o_T\}$.

The Maximum aposteriori probability state sequence is provided by the BCJR algorithm[3]. The MAP estimate for the hidden state at time t is given by $\hat{q}_t = \arg\max P(q_t, O_1^T)$. The BCJR algorithm computes this using the forward and backward recursions.

Define,

$$\begin{aligned}
\lambda_t(m) &= P(q_t = m, O_1^T) \\
\alpha_t(m) &= P(q_t = m, O_1^t) \\
\beta_t(m) &= P(O_{t+1}^T | q_t = m) \\
\gamma_t(m', m) &= P(q_t = m, o_t | q_{t-1} = m'), m = 1, 2 \dots N, m' = 1, 2 \dots N
\end{aligned}$$

Then establish the recursions,

$$\begin{aligned}
\alpha_t(m) &= \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \\
\beta_t(m) &= \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m') \\
\lambda_t(m) &= \alpha_t(m) \cdot \beta_t(m)
\end{aligned}$$

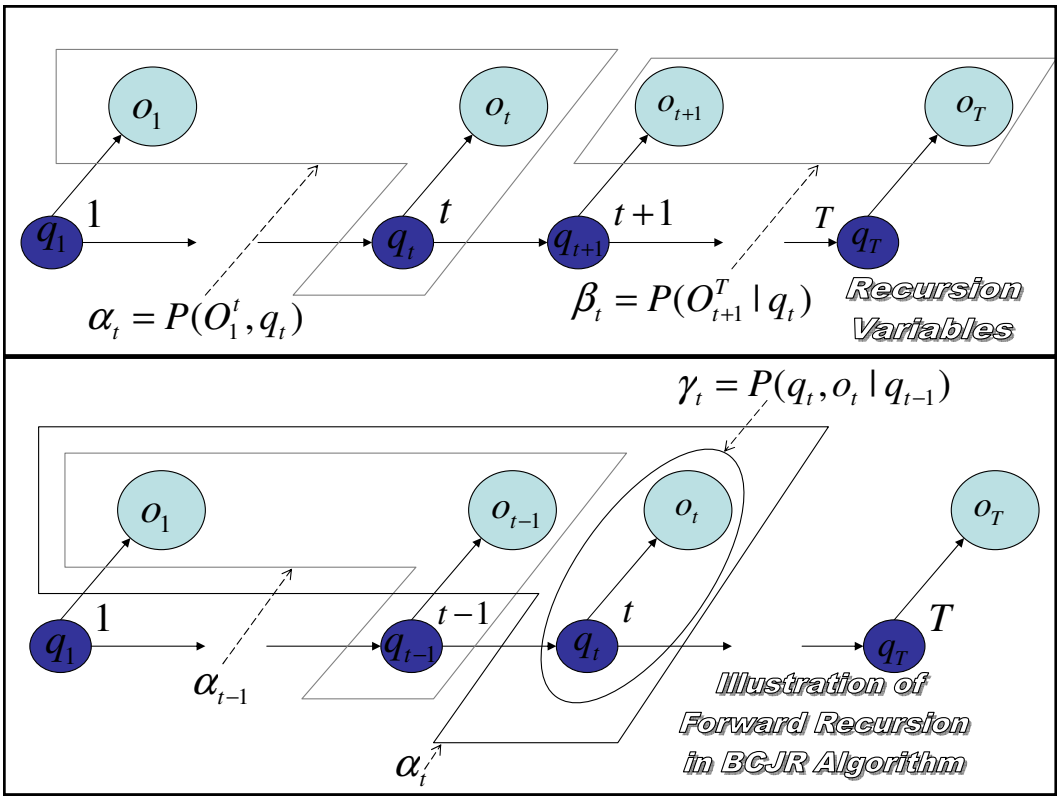

Figure 3.1: Illustrating the forward recursion of the BCJR algorithm

These enable us to solve for the MAP state sequence given appropriate initial conditions for $\alpha_1(m)$ and $\beta_T(m)$.

3.3.2 Multimodal scenario

For the sake of clarity, consider a bimodal system. There are observations O_1^T from one modality and observations $\Theta_1^T = \{\theta_1, \theta_2 \ldots \theta_T\}$ from the other modality. The MAP solution in this case would be $\hat{q}_t = \arg\max P(q_t, O_1^T, \Theta_1^T)$. In order to apply the BCJR algorithm to this case, concatenate the observations(feature level fusion) and train a new HMM in the joint feature space. Instead of building a joint model, one can develop an iterative decoding algorithm that allows us to approach the performance of the joint model by iteratively exchanging information between the simpler models and updating their posterior probabilities.

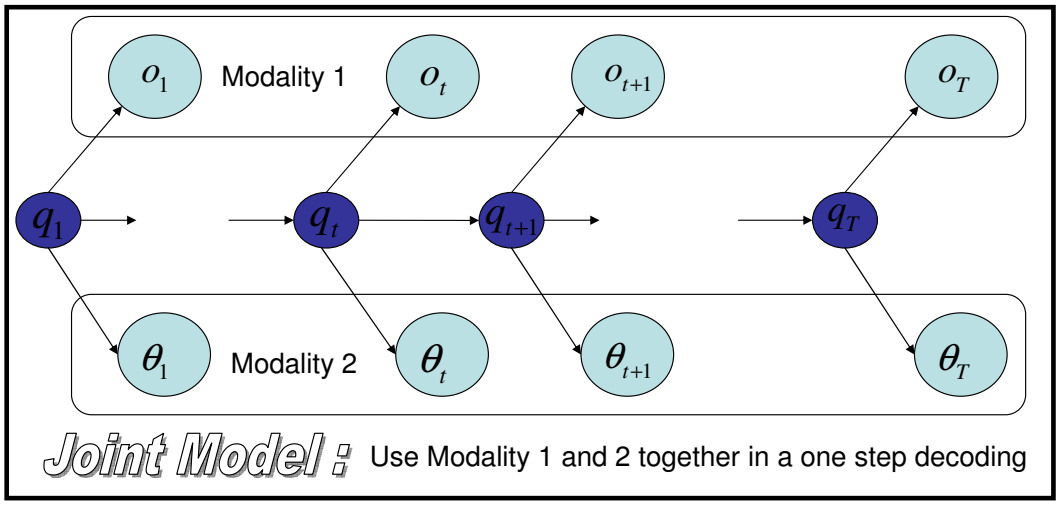

Figure 3.2: Joint Model for a bimodal scenario

3.4 Iterative Decoding Algorithm

This is a direct application of the turbo decoding algorithm[9]. In this section, it is assumed that the hidden states in the two modalities have a one-one correspondence. This requirement is relaxed in the generalized solution presented in the next section.

In the first iteration of the iterative algorithm, decode the hidden states of the HMM using the observations from the first modality, O_1^T. The aposteriori probabilities, $\lambda_t^{(1)}(m) = P(q_t = m, O_1^T)$ are obtained.

In the second iteration, these aposteriori probabilities, $\lambda_t^{(1)}(m)$ are utilized as extrinsic information in decoding the hidden states from the observations of the second modality Θ_1^T. Thus the aposteriori probabilities in the second stage of decoding are given by $\lambda_t^{(2)}(m) = P(q_t = m, \Theta_1^T, {Z^{(1)}}_1^T)$ where $Z_t^{(1)} = \lambda_t^{(1)}$ is the extrinsic information from the first iteration. In order to evaluate $\lambda_t^{(2)}$, the BCJR algorithm is modified as follows.

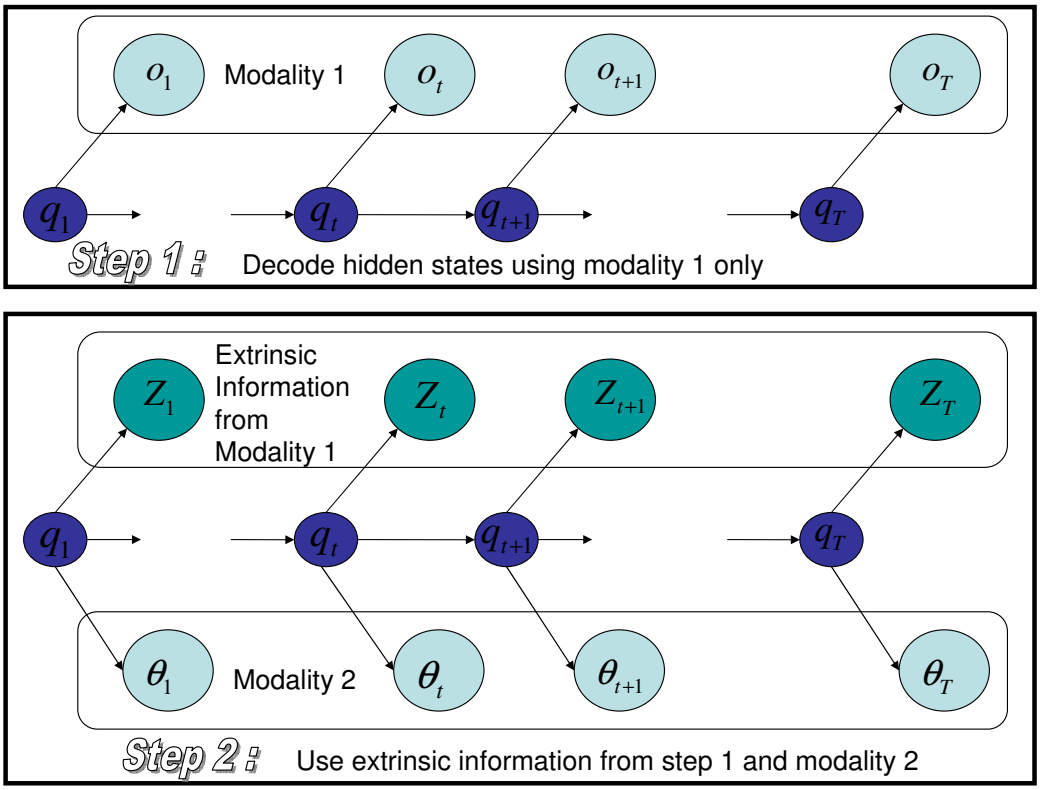

Figure 3.3: First two steps of the iterative decoding algorithm

3.4.1 Modified BCJR algorithm for incorporating the extrinsic information

$$\begin{aligned}
\lambda_t^{(2)}(m) &= P(q_t = m, \Theta_1^T, {Z^{(1)}}_1^T) \\
\alpha_t^{(2)}(m) &= P(q_t = m, \Theta_1^t, {Z^{(1)}}_1^t) \\
\beta_t^{(2)}(m) &= P(\Theta_{t+1}^T, {Z^{(1)}}_{t+1}^T | q_t = m) \\
\gamma_t^{(2)}(m', m) &= P(q_t = m, \theta_t, Z_t^{(1)} | q_{t-1} = m')
\end{aligned}$$

Then the recursions do not change, except for the computation of $\gamma_t^{(2)}(m', m)$.

Since the extrinsic information is independent of the observations from the second modality, $\gamma_t^{(2)}(m', m) = P(q_t = m | q_{t-1} = m') \cdot P(\theta_t | q_t = m) \cdot P(Z_t^{(1)} | q_t = m)$

Here $Z_t^{(1)} = [z_{1t}^{(1)} z_{2t}^{(1)} \dots z_{Nt}^{(1)}]'$ is a vector of probability values. A histogram

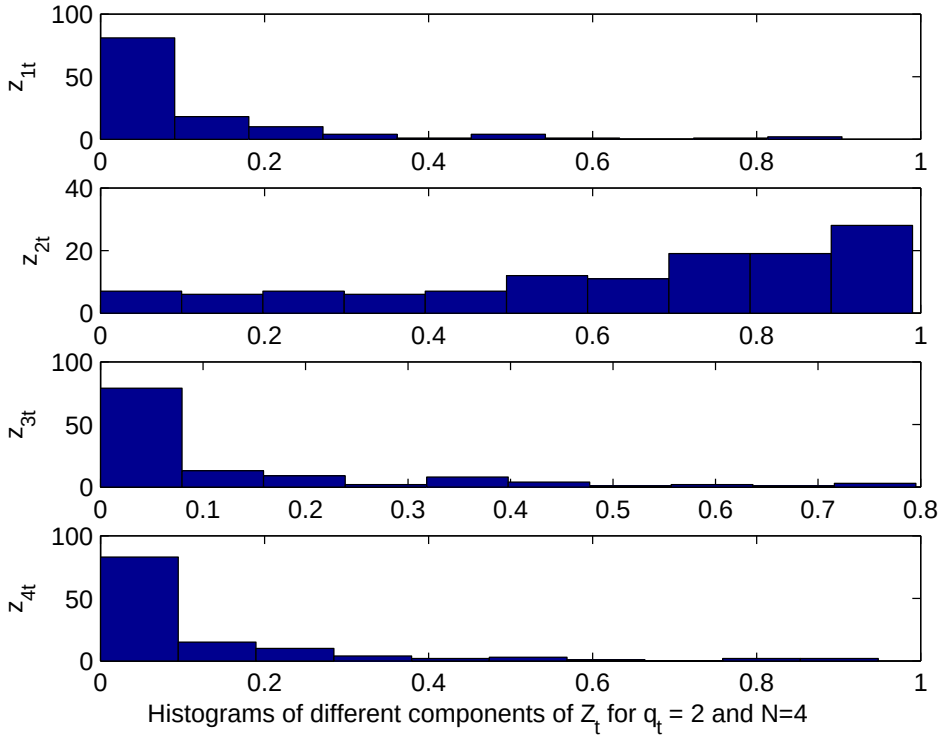

Figure 3.4: A histogram of each component of Z_t for $q_t = 2$ in a $N = 4$ state HMM synthetic problem

of each component of $Z_t^{(1)}$ for $q_t = 2$ in a $N = 4$ state HMM synthetic problem is shown in figure 3.4. From the histogram, one can see that a simple parametric probability model for $P(Z_t^{(1)}|q_t = m)$ is obtained as

$$P(Z_t^{(1)}|q_t = m) = f(1 - z_{mt}^{(1)}; \rho) \cdot \prod_{i \neq m} f(z_{it}^{(1)}; \rho)$$

where,

$$f(x; \rho) = \begin{cases} \frac{1}{\rho} e^{-x/\rho} & , \ x \geq 0, \\ 0 & , \ x < 0. \end{cases}$$

is an exponential distribution with rate parameter $\frac{1}{\rho}$. Other distributions like the beta distribution could also be used. The exponential distribution is chosen due to its simplicity.

In the third iteration, the extrinsic information to be passed back to decoder 1 is the aposteriori probabilities $\lambda_t^{(2)}(m)$. But part of this information ($\lambda_t^{(1)}(m)$), came from decoder 1 itself. If one were to use $\lambda_t^{(2)}$ as the extrinsic information in the third iteration, it would destroy the independence between the observations from the first modality and the extrinsic information. This difficulty can be overcome by choosing another formulation for the extrinsic information based on the following observation,

$$\begin{aligned}
\lambda_t^{(2)}(m) &= \alpha_t^{(2)}(m) \cdot \beta_t^{(2)}(m) \\
\alpha_t^{(2)}(m) &= \sum_{m'} \alpha_{t-1}^{(2)}(m') \cdot \gamma_t^{(2)}(m', m) \\
\lambda_t^{(2)}(m) &= \sum_{m'} \alpha_{t-1}^{(2)}(m') \cdot \gamma_t^{(2)}(m', m) \cdot \beta_t^{(2)}(m) \\
\lambda_t^{(2)}(m) &= P(Z_t^{(1)}|q_t = m) \sum_{m'} \alpha_{t-1}^{(2)}(m') \cdot P(q_t = m|q_{t-1} = m') \cdot \\
&\quad P(\theta_t|q_t = m) \cdot \beta_t^{(2)}(m) \\
\lambda_t^{(2)}(m) &= P(Z_t^{(1)}|q_t = m) \cdot Y_t^{(2)}
\end{aligned}$$

Note that $Y_t^{(2)}$ does not depend on $Z_t^{(1)}$ and is hence uncorrelated with o_t . This argument follows the same principles used in Turbo coding literature [9]. Hence, $Y_t^{(2)}$ is normalized to sum to 1 and the normalized vector is considered to be the extrinsic information passed on to decoder 1 in the third iteration.

The normalized extrinsic information $Z_t^{(2)}(m) = \frac{\lambda_t^{(2)}(m)/P(Z_t^{(1)}|q_t=m)}{\sum_{m'} \lambda_t^{(2)}(m')/P(Z_t^{(1)}|q_t=m')}$ is passed back to decoder 1.

The iterations are continued till the state sequences converge in both the modalities or a fixed number of iterations are reached.

3.4.2 General multimodal problem

In the previous section, it was assumed that the hidden states in the two modalities of a multimodal system are the same. In this section, this restriction is relaxed to allow the hidden states in the individual modalities to just have a known prior co-occurrence probability. In particular, if q_t and r_t represent the

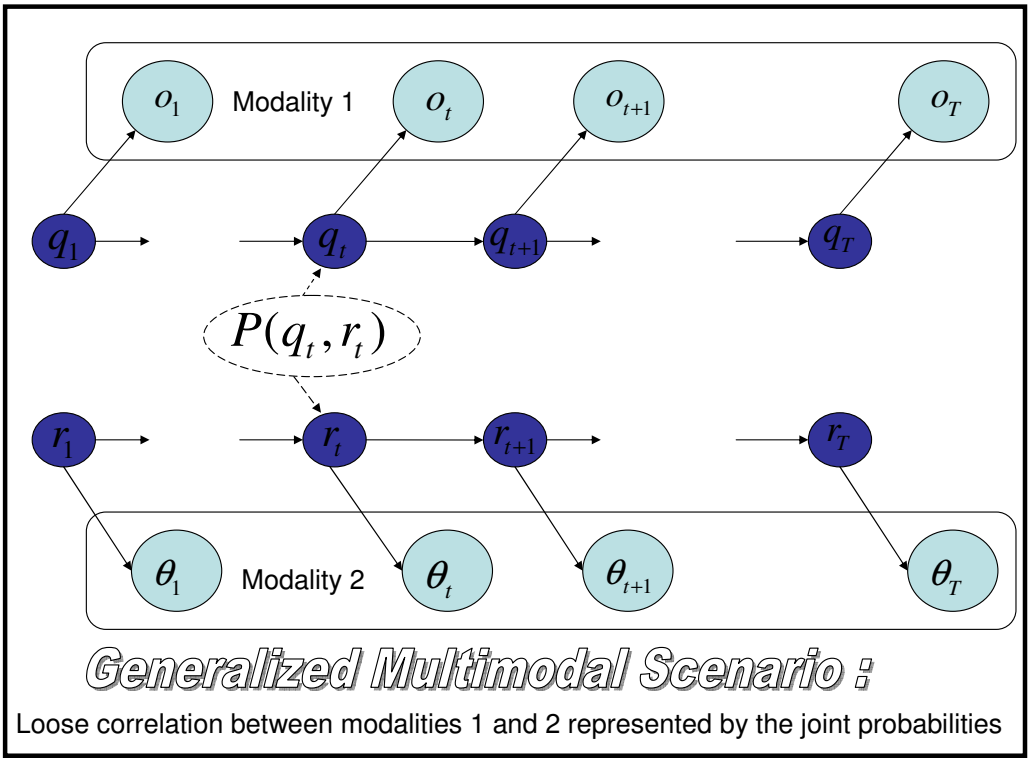

Figure 3.5: A more generalized bimodal problem

hidden states in modality 1 and 2 at time t, then the joint probability distribution $P(q_t = m, r_t = m')$ is assumed to be stationary and known.

This corresponds to the case where there is a loose but definite interaction between the two modalities as seen very clearly in the case of phonemes and visemes, in audio-visual speech recognition. There in no one-one correspondence between visemes and phonemes. But the occurrence of one phoneme corresponds to the occurrence of a few specific visemes and vice-versa.

3.4.3 Iterative decoding algorithm in the general case

This is an extension of the iterative decoding algorithm as presented in the turbo coding scenario. At the jth iteration in the modified BCJR algorithm, in the computation of $\gamma_t^{(j)}(m', m) = P(q_t = m, \theta_t, Z_t^{(j-1)} | q_{t-1} = m')$, one needs to

compute

$$\begin{aligned}\gamma_t^{(j)}(m',m) &= P(r_t = m, \theta_t, Z_t^{(j-1)} | r_{t-1} = m') \\ \gamma_t^{(j)}(m',m) &= P(r_t = m | r_{t-1} = m') \cdot P(\theta_t | r_t = m) \cdot P(Z_t^{(j-1)} | r_t = m)\end{aligned}$$

$$\begin{aligned}\gamma_t^{(j)}(m',m) = P(r_t = m | r_{t-1} = m') \cdot P(\theta_t | r_t = m) \\ \cdot \sum_n \{P(Z_t^{(j-1)} | q_t = n) P(q_t = n | r_t = m)\}\end{aligned}$$

which can be computed from the joint probability distribution $P(q_t = m, r_t = m')$. The rest of the iterative algorithm remains the same as before.

3.5 Experimental evaluation of the iterative decoding algorithm

3.5.1 Performance evaluation on a synthetic dataset

In this section the results of applying the iterative decoding algorithm to a synthetic problem are presented. A synthetic problem is better suited to isolate the performance characteristics of the iterative decoding algorithm from the complexities of real world data, which are dealt with in section 3.5.2.

Observations are generated from an HMM with 4 states and whose observation densities are 4 dimensional Gaussian distributions. A joint model is constructed by concatenating the feature vectors. The goal of the experiment is to decode the state sequence from the observations and compare it with the true state sequence in order to obtain the error rates. The experiment is repeated several times and the average error rates are obtained.

In the first case, the joint model with 8 dimensions and 4 states is used to generate the state and observation sequences. The joint model is used to decode the state sequence from the observations. Next, the observations are assumed to be generated by two modalities with 4 dimensions each. The product rule [43] is an

alternative modeling strategy to the joint model. But in the simulations, it is found that its error rates are the same as that of the joint model. Hence the joint model is considered to be the baseline. The iterative decoding algorithm described in section 3.4 is applied to decode the state sequence and compared to the true state sequence. The results are shown in figure 3.6. The iterative decoding algorithm converges to the baseline performance and it reduces the error rate by almost 50% compared to the unimodal case (iteration 1). Figure 3.6 also shows the standard deviation of error from which it can be seen that the performance is indeed close to the baseline performance. Since the two modalities have similar unimodal error rates, the error dynamics of the iterative algorithm are independent of the starting modality.

In the second example, the observations are generated from two independent HMMS such that the state sequence follows a known joint distribution. The generalized iterative decoding algorithm described in section 3.4.2 is then applied to decode the hidden states. The results are shown in figure 3.7. In this case there is no baseline experiment for comparison as the two streams are only loosely coupled but the general trend in average error rate with each iteration is similar to the case shown in figure 3.6.

In the presence of noise, the iterative algorithm outperforms the joint model as shown in figure 3.8. Based on the standard deviation of error, a standard t-test reveals that the difference between the joint model and the iterative decoding algorithm is statistically significant after the third iteration. In this case additive white Gaussian noise is added to the features of one of the modalities. No apriori information about the noise statistics is assumed to be available. Note that in this case, the individual modalities have varying noise levels and hence the convergence of the iterative algorithm is dependent on the starting modality. But in both the cases, the iterative algorithm converges to the same performance after the third iteration. This illustrates the advantage of iterative decoding over joint modeling as mentioned in section 3.2.

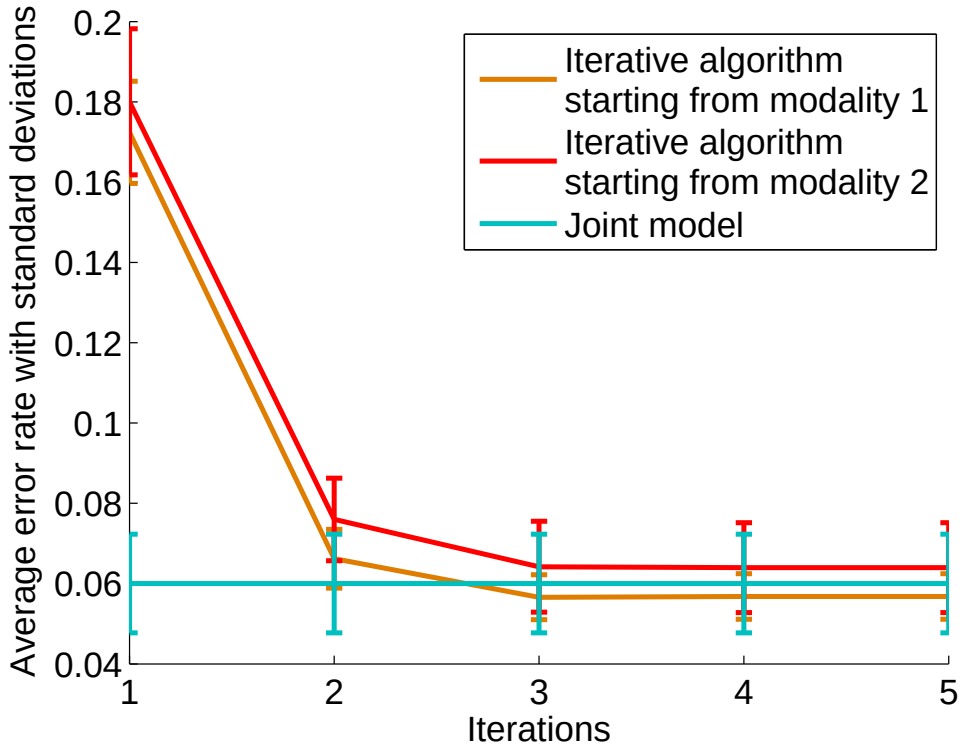

Figure 3.6: Error rate at different iterations for a 4 state HMM problem with one-one correspondence between the two modalities. Note the convergence of the error rate to that of the joint model.

3.5.2 Speech segmentation experiment

In order to evaluate the performance of the iterative decoding algorithm on a real world problem, a simplified version of the meeting room conversation is considered, with one speaker. The goal of the experiment is to segment the speech data into speech and silence parts. The traditional approach to the problem is to use the energy in the speech signal as a feature and maintain an adaptive threshold for the energy of the background noise. This is not accurate in the presence of non stationary background noise like overlapping speech from multiple speakers. In this experiment, the audio and visual modalities to build a multimodal speech segmentation system , that is robust to background noise and performs better than the audio only model or the joint model.

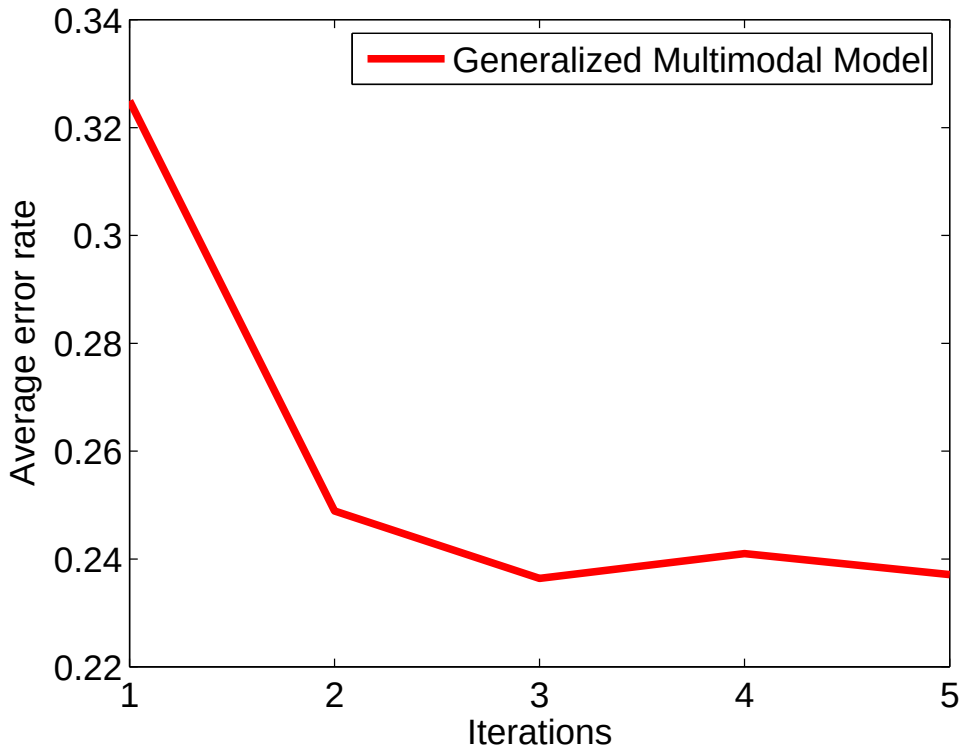

Figure 3.7: Error rate at different iterations for a generalized multimodal problem. Note that the performance follows the same trend as in the previous case.

Data collection

The audio-visual testbed used for this experiment is described in detail in Appendix A.2. 4 minutes of audio-visual data was collected from 20 different speakers. This included 12 different head poses and 2 different backgrounds as shown in figure 3.9 We used 1 minute of data from each speaker, that is a total of 20 minutes of audio visual data to estimate the HMM model parameters. The remaining 3 minutes from each speaker were included in the testing set. That is, a total of 60 minutes of testing data was used.

Feature extraction

Each time-step corresponds to one frame of the video signal. The cameras capture video at 15 fps. The energy of the microphone signal in time window corresponding to each frame is the audio feature. The face of the speaker is detected

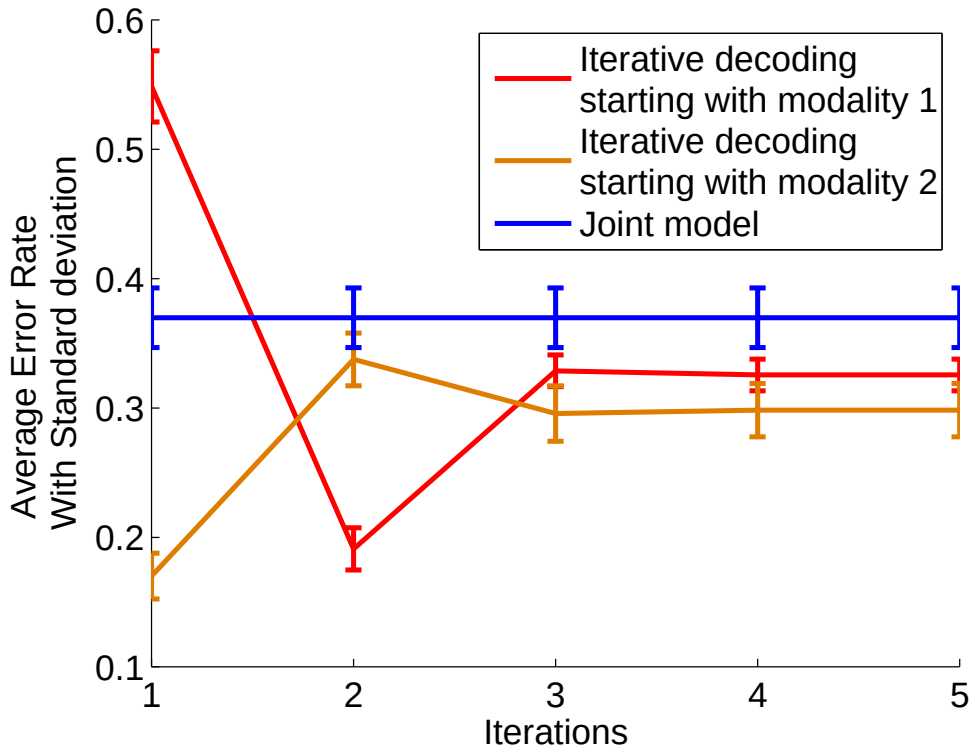

Figure 3.8: Error rate at different iterations in the case of noisy modalities. Note that the iterative algorithm performs better than the joint model at low SNR.

Figure 3.9: Different head poses and backgrounds for one subject out of 20 subjects in the dataset

and tracked using the Viola-Jones face detector [105]. Figure 3.10 shows some sample frames from the face detector output for different subjects. The mouth

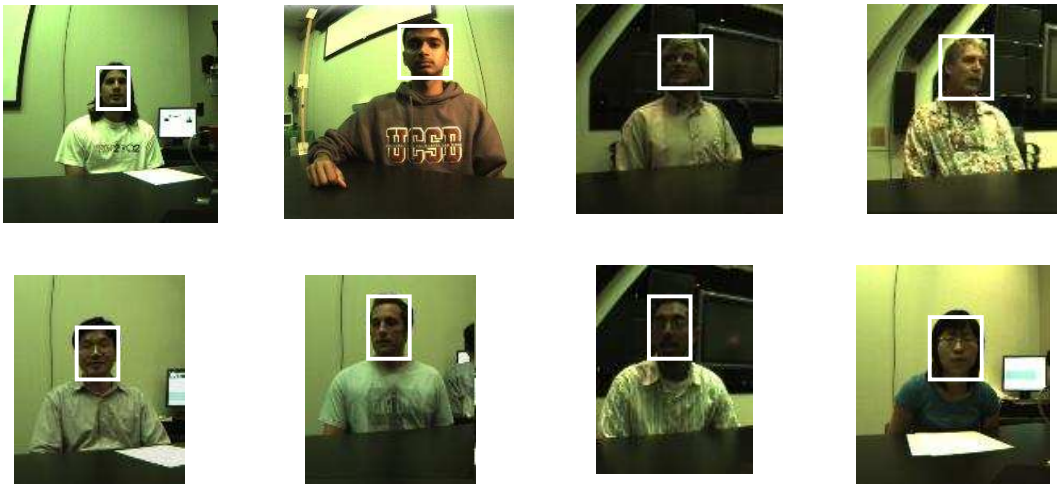

Figure 3.10: Face detection using the Viola-Jones face detector with various subjects.

Figure 3.11: Some snapshots of the lip region during a typical utterance. Observe the variations in pose and facial characteristics of the three different subjects, which limits the performance of a video-only system.

region is considered to be the lower half of the face. The motion in the mouth region is estimated by subtracting the mouth region pixels from consecutive frames and summing the absolute value of these differences. This sum is the video feature vector. Thus a smooth and stable face tracker is essential for accurate video feature extraction. Figure 3.11 shows the different positions of the lips during a typical utterance.

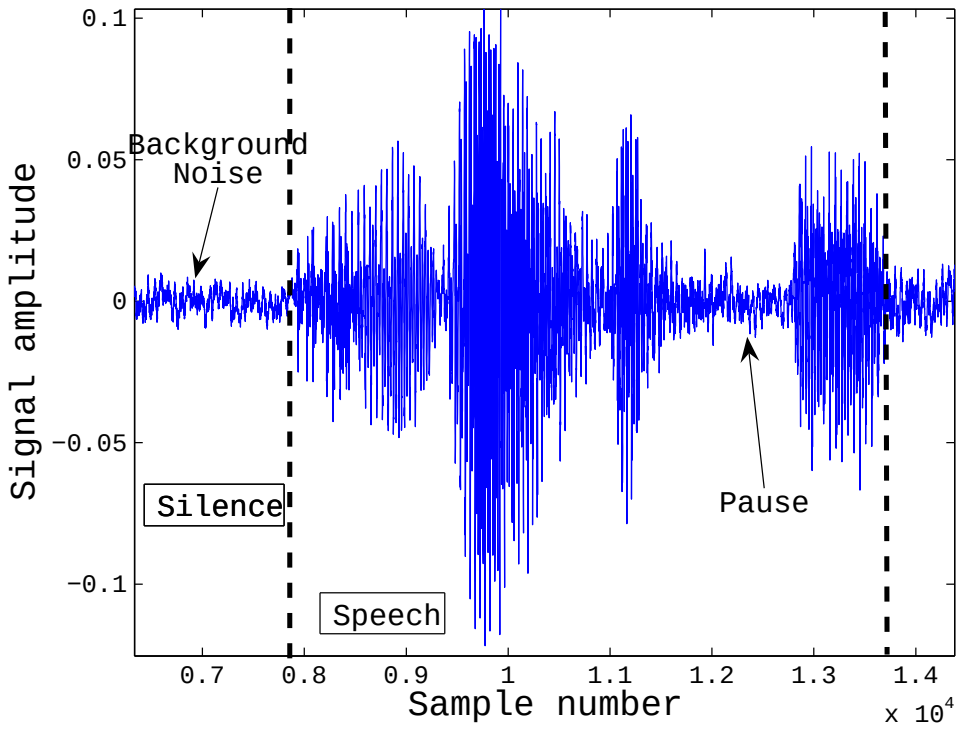

Figure 3.12: Audio waveform of speech in background noise. The short pauses between words which can be confused by an audio-only system for background noise will be detected as speech by the video modality, based on the lip movement.

Model training and results

HMMs in the audio and video domains are trained using labeled speech and silence parts of speech data. The joint model by is also trained by concatenating the features. The results of the experiment on a typical noisy segment of speech is shown in figure 3.14. The ground truth is shown in figure 3.13. From the numerical results in figure 3.15, one can see that by the third iteration, the iterative decoding algorithm performs slightly better than the joint model. This improvement however, is not statistically significant because the background noise in the audio and video domains is not so severe. Though building the joint model is straightforward in this case, it is not so easy in more complex situations, as explained in the introductory sections. Thus the iterative algorithm appears to be a good fusion framework in the multimodal scenario.

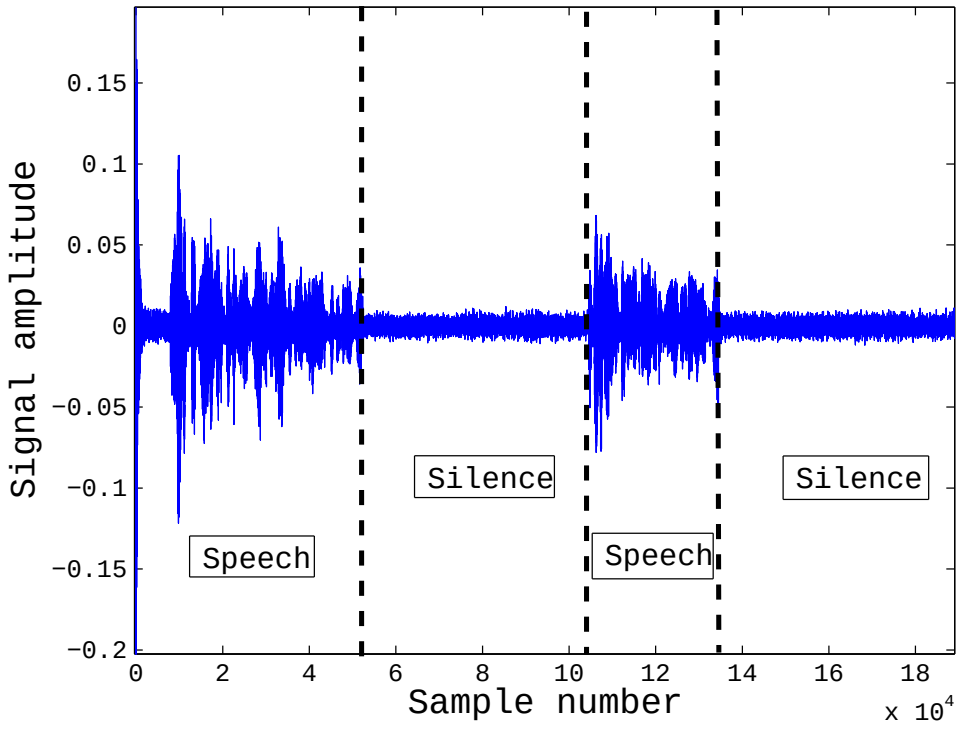

Figure 3.13: Audio waveform from a typical utterance in background noise. The speech and silence parts are hand labeled to be used as ground truth.

3.5.3 Audio visual speech recognition experiment

Database, feature extraction and modeling

A more involved application of the iterative decoding algorithm is in the automatic speech recognition (ASR) domain. Audio-visual fusion by way of fusion of lip reading and acoustic features for ASR has received considerable attention in the research community. The GRID audio-visual speech corpus [21] is a recently collected audio-visual dataset for the evaluation of audio-visual ASR systems (AVSR). The results correspond to a speaker dependent AVSR system. The GRID corpus is a 51 word small vocabulary speech corpus of six word long sentences. 1000 sentences are uttered by each speaker.

900 utterances are used to train the HMMs and the rest are used in the test set. Each word is modeled by a three state HMM with a Gaussian mixture

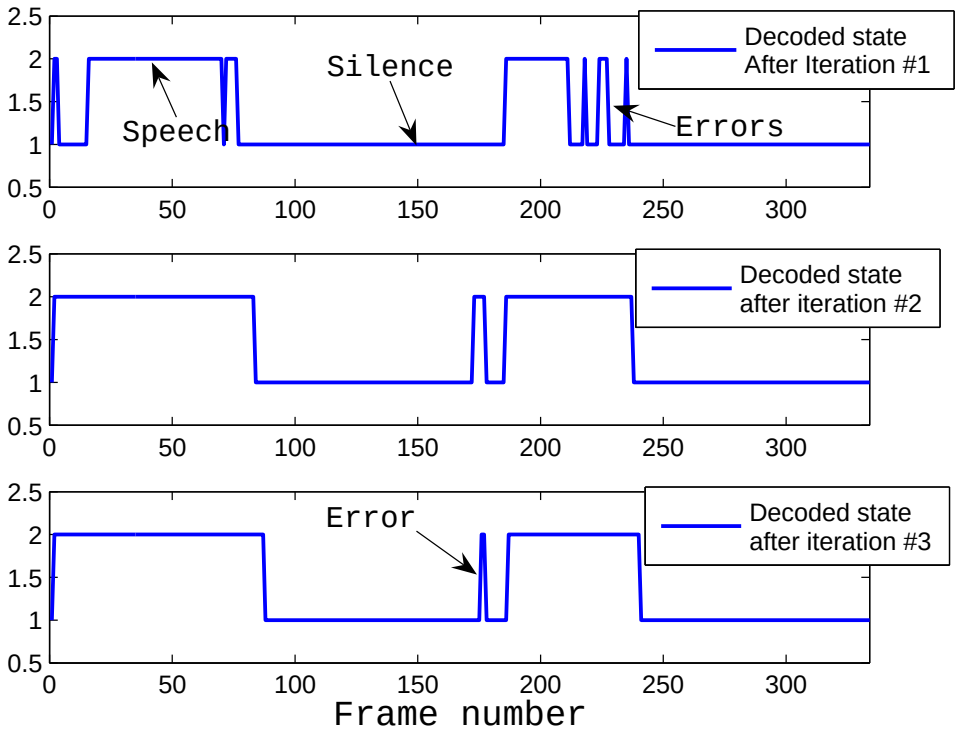

Figure 3.14: The decoded states of the HMM after each iteration. Note the errors in the first iteration being corrected in the subsequent iterations.

model(GMM) observation density. There are 10 components in each GMM with diagonal covariance matrices. The audio feature vectors are the 13 MFCC coefficients computed on 20ms windows of audio signal with a 10ms overlap. The video rate is 25 frames per second. This corresponds to one video frame for every 4 audio frames. The video features are hence upsampled to match the audio and video frame rates. In order to extract the video features, the face of the speaker is detected and tracked using the Viola-Jones face detector[105]. The current frame is subtracted from the previous frame to estimate the motion in the mouth region of the face. The first 16 coefficients of the 2D-DCT of the mouth region motion map are used as the components of the video feature vector .

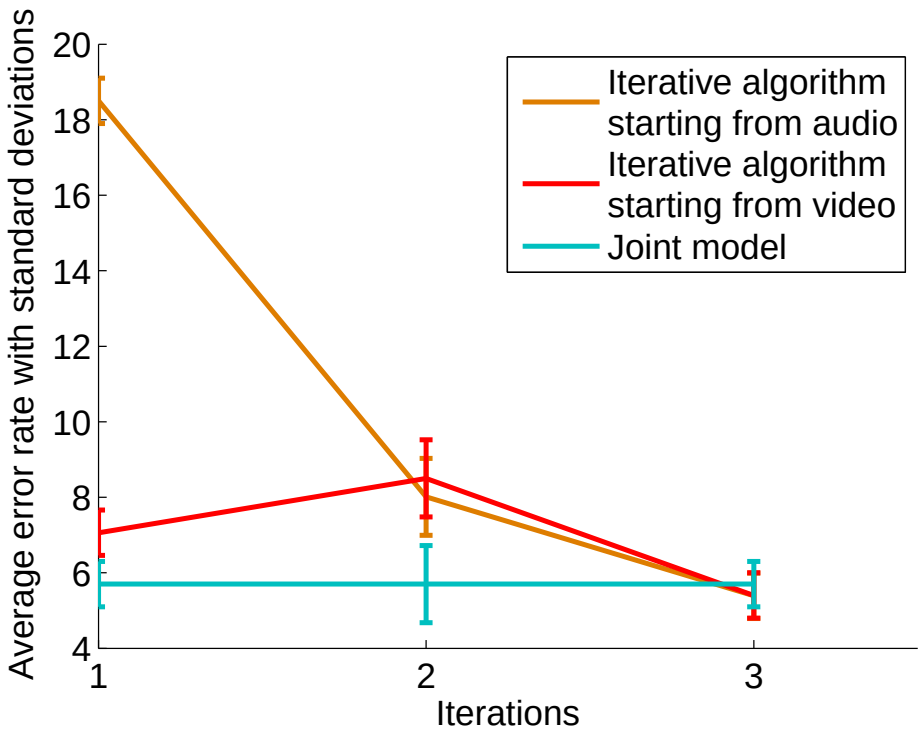

Figure 3.15: Results showing the error rates for the iterative decoding scheme for the speech segmentation problem.

Results

In the noiseless environment the audio-only speech recognizer has a state error rate of 12%. The state error rate is measured by comparing the decoded hidden state sequence with the transcriptions. The state error rate is a better estimate of the efficacy of the algorithm than the word error rate as the fusion of information takes place at the state level. The video-only speech recognizer has a state error rate of 27%. The iterative decoding algorithm converges to an error rate of 13% after the third iteration. The audio modality is then corrupted with white noise so the SNR is now reduced to 5dB. The error rate of the audio-only speech recognizer is now 40%. But the iterative decoding algorithm converges to an error rate of 25% after the third iteration. The results are summarized in Figure 3.16.

Note that the error rates presented here are highly dependent on the choice of the audio and video features. Using better video features would naturally lead

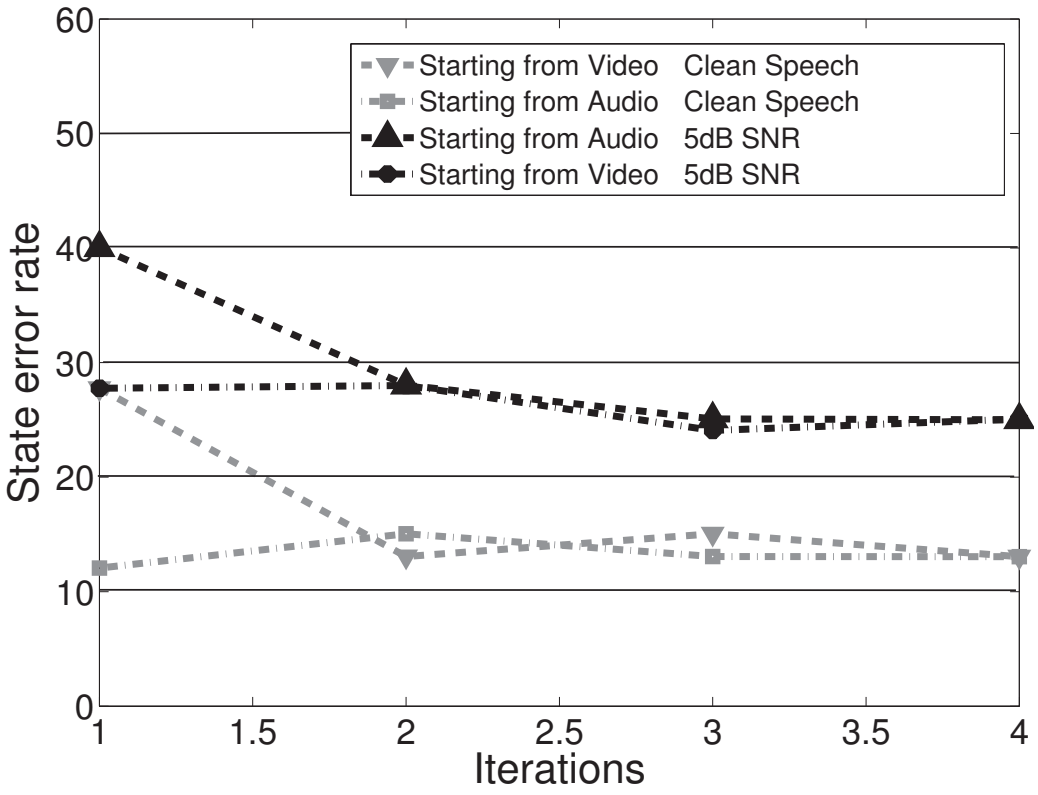

Figure 3.16: State error rates for an audio-visual speech recognition task on the GRID speech corpus using the proposed scheme. After 3 iterations, the error rate of the iterative decoding algorithm converges close to the error rate of the best modality.

to a better performance in the video-only speech recognizer and hence the iterative decoding framework would perform better in the presence of audio noise.

3.6 Concluding remarks on the iterative decoding algorithm

In this chapter a general information fusion framework based on the principle of iterative decoding used in turbo codes has been presented. The iterative decoding algorithm has been extended to the generalized multimodal case and its performance has been evaluated on synthetic experiments as well as on the real world tasks of speech segmentation and speech recognition. The iterative decoding

is advantageous to joint modeling and other decision level fusion schemes in terms of ease of training of models and performance under low SNR scenarios. However, the utility of the iterative decoding algorithm as presented here is limited by the fact that the multiple observation streams need to correspond to the same underlying hidden state sequence for effective inference. In practice however, audio-visual human activity analysis includes the observation of multiple subjects using multiple sensors. In such a situation, further steps are required to associate data with the respective source before iterative decoding based inference can be applied. In the next chapter, a hierarchical framework is presented that uses iterative decoding and data association to track multiple persons using audio and visual cues.

3.7 Acknowledgments

The text of Chapter 3, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "An Iterative Decoding Algorithm for Fusion of Multimodal Information," EURASIP Journal on Advances in Signal Processing, Special Issue on Human-Activity Analysis in Multimedia Data, Feb-2008. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 3, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Multimodal Information Fusion using the Iterative Decoding Algorithm and its Application to Audio-visual Speech Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, Las Vegas, Apr-2008. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Multilevel iterative decoding based audio-visual person tracking (MID-AVT) framework

4.1 Introduction

In the previous chapter, the iterative decoding algorithm was described for the fusion of audio-visual information at the classifier level. In this chapter, a hierarchical fusion framework is presented which combines the iterative decoding algorithm with a data association scheme. This enables the benefits of the iterative decoding scheme to be applied to the case where multiple human subjects are present in the scene which leads to multiple streams of audio and visual observations. The multilevel iterative decoding based audio-visual person tracking (MID-AVT) framework presented in this chapter is an embodiment of such a hierarchical fusion framework.

4.2 Person tracking using audio-visual cues

Robust person tracking is the first step in facilitating detection and analysis of human activity in a monitored space. It is also an integral component of

intelligent spaces, for facilitating seamless interaction between humans and computers. Tracking humans using audio-visual cues can provide robustness to background noise and visual clutter. Tracking based on visual sensors has been widely researched[106]. Microphone array based trackers that track sound sources have also been studied by some researchers[13]. In this chapter the iterative decoding algorithm is developed further to formulate a general fusion framework for multimodal person tracking and applied to track people in an indoor environment with multiple cameras and microphone arrays. Extensive experimental evaluation of the framework is also presented. The evaluation is carefully designed to bring forth the true strengths of the framework and its weaknesses. In section 4.2.1, a survey of related research and the comparative advantage of the MID-AVT framework is presented. In section 4.3, the mathematical formulation of the hidden Markov model based MID-AVT framework is developed. In section 4.4, the laboratory testbed with multiple cameras and microphone arrays which was used for extensive experimentation and evaluation studies is outlined.

4.2.1 Existing audio-visual person tracking schemes

In this section a brief survey of related research activities in the field of multimodal person tracking is presented. Also, the motivation behind using iterative decoding framework to solve the tracking problem is clearly outlined.

Person tracking has been a computer vision problem that received considerable attention[106]. An good review of multi-camera trackers can be found in [45]. Audio source localization is also a well researched field [41][25]. Localizing and tracking individuals using audio-visual information has recently received much attention.

Early effort in tracking speakers using both audio and video cues involved camera epipolar constraints and audio cross correlation. In [77] one camera and two microphones were used and a single person was tracked. Spatial probability maps were used in [75] to track a single speaker using two cameras and three microphones. [104] used a particle filter to track one subject using one camera two microphones. [67] used auditory epipolar geometry and face localization to track

multiple people in the camera view using four microphones. Bayesian network based feature concatenation scheme was explored in [5] using one camera and two microphones. Audio-visual synchrony and correlation have been exploited to locate speakers in [22][31][42]. These early efforts were constrained by the number of sensors used (usually one or two cameras and two to three microphones) and the scene complexity (usually one speaker was tracked).

Subsequent researchers have used Bayesian networks with the particle filtering based inference technique in audio-visual tracking [104] [114] [76] [5] [20] [17] [71] [36] [35] [6]. Approximate inference in the dynamic Bayesian network framework, necessitated by the complexity and non-Gaussianity of the joint models, is performed by the use of particle filters [35],[17]. In the recent past, the CLEAR 2006 and CLEAR 2007 evaluation workshops [95][94] have been a significant research effort in evaluating audio-visual person tracking in meeting and lecture scenes. A wide variety of frameworks were developed and evaluated in these workshops on datasets collected under the initiative of the European CHIL (Computers in Human Interaction Loop) consortium. Among the techniques presented in CLEAR 2006 and CLEAR 2007, [6][1][51] are the closest matching schemes to the MID-AVT framework.

[51] describes an audio-visual 3-D person tracker that uses face detectors as the visual front-end and fuses detections from multiple views to obtain the 3-D location of the person's head. If a speaker is active, the audio localization results are matched to the closest video track and continued to be tracked. If there is no match with the video tracks, the audio track is tracked separately. The results indicate that though the video face detection yields consistent results, the fusion of audio localization information does not perform well. In fact with the addition of audio information the results are worse than the video-only results.

[1] describes an elaborate 3-D voxel based video tracker augmented by the audio localization information. Views from multiple cameras are combined to construct a 3-D voxel representation of the subjects and this 3-D object is then tracked over time. One problem with such an approach is that it relies heavily on the calibration of the cameras to obtain the 3-D co-ordinates of object pixels. This

sensitivity is a recurring feature in other schemes too. Another shortcoming in [1] is that the audio localization information is associated to the video detections using data association techniques. Details of the data association technique used are not provided and one can assume that proximity based data association is one possible solution. This could lead to many false detections because the audio detections are quite noisy. The results do indicate that the audio-visual tracker performs only as well as the video-only tracker. In section 4.4.1, this is explored in more detail.

[6] presents a state-space based fusion strategy for associating audio localization information with the video tracks. 3-D tracks are maintained using a particle filter based tracker. If audio detections are close to video tracks, they are associated with each other. If not, new tracks are created to explain the audio detections till a matching video track is found. The 3-D video tracker described here has the same sensitivity to camera calibration mentioned above. In addition, a separate particle filter is used for each person and hence an estimate of the number of people in the scene is necessary. Also, even when the number of subjects is known accurately, if some subjects are not detected, the tracker tends to initialize false tracks to explain the given number of subjects.

[35] presents an interesting particle filtering framework which incorporates the audio and visual detections into the particle filtering framework. However the tracking framework presented in [35] does not correspond to a 3-D tracker. The camera views are stitched to obtained a panoramic view of the room in which subjects are tracked. An advantage of this system is that the cameras need not be accurately calibrated. However this setup places restrictions on the positions that the subjects can occupy and is difficult to generalize to new scenes especially when larger number of people participate in meetings and lectures.

[17] uses particle filtering to fuse audio and video detections. This is the closest approach to the MID-AVT framework. In [17], two overlapping camera views are used along with a microphone array to localize and track subjects. Occlusions are handled by multi-view and audio localizations. However the evaluation is limited to a simple scene and does not give much insight into the strengths and

weaknesses of the framework. Also, the 3-D tracking relies on accurate calibration of the cameras.

4.2.2 Proposed framework - MID-AVT

The MID-AVT framework is an alternative approach to fusion of audio-visual cues based on iterative decoding for tracking multiple people in a space instrumented with multiple sensors - cameras and microphone arrays. One important requirement of this scheme is that the overlapping fields of view should provide robustness to occlusions. Another goal of the MID-AVT framework is to overcome two major disadvantages with some of the existing schemes outlined above, namely sensitivity to accurate sensor calibration and the necessity to know the number of subjects in the scene. The framework is based on a rough calibration step similar to [35] but unlike [35] there is no constraint on the scene complexity and the tracking process actually incorporates multiple overlapping views which allows for successful tracking through occlusions in some views. Unlike [17], the MID-AVT framework is robust to sensor calibration errors. In Section 4.4.1, the performance of the MID-AVT framework is compared with that of the particle filter framework suggested in [17] on the same dataset. Also the robustness to sensor calibration is demonstrated.

The MID-AVT framework is based on iterative decoding. The iterative decoding scheme as described in the previous chapter is not applicable to tracking as we need to solve the data association problem[4] before using iterative decoding. In the next section a hidden Markov model (HMM) based tracking framework is presented which specifies the tracking problem in a hierarchical manner, allowing the local sensors (camera/microphone array) to maintain track hypotheses and the global tracker to fuse the local tracks from various sensors to generate a robust estimate using iterative decoding. The same framework is also applicable to situations where multiple sensors are used to monitor disjoint spaces. In this case, one cannot expect robustness to sensor limitations as one would in the overlapping-field-of-view case.

The calibration of multimodal sensors is an important issue in tracking. In

the MID-AVT framework, the system only requires a rough calibration step. After this initial calibration, the system can continue tracking even if the sensors are disturbed because we are tracking in the sensor co-ordinate system and not in the 3-D world co-ordinate system. If the calibration is accurate, one can, in addition, infer the 3-D co-ordinates of the subjects. This 3-D location information is not necessary for the tracking algorithm to work. The experimental evaluation results support this claim. This is an advantage over the particle filter based tracking schemes because the particle filters are tracking in the 3-D world co-ordinates and a mismatch in calibration of the sensors is not tolerable.

Also, the MID-AVT framework does not need to know the number of people in the scene. Every individual who presents a signature on any of the sensors is detected and tracked. This is yet another advantage over [35] which assumes that the number of subjects in the scene is known and [6] which assumes that the maximum number of subjects to be tracked is three.

In summary, the MID-AVT framework has several distinct advantages. It is modular and hence easy to expand to more number of cameras and microphone arrays or any other sensors that can localize persons. It is also applicable to sensors with overlapping and non-overlapping field of 'view'. Since the placement of the sensors is assumed to be arbitrary but fixed, only a rough calibration scheme is necessary to establish the correspondence between sensors. The unimodal models considered in this framework are simple and intuitive.

4.3 Computational framework and algorithms

The goal of the MID-AVT framework is to provide a framework for tracking multiple targets(people) in a space instrumented with multiple cameras and microphone arrays. Each sensor detects the subjects in its field of view and maintains an exhaustive list of possible track hypotheses. For example, if one tracked object occludes another, it involves two tracks converging and they may diverge again at a later stage. However, when two tracks converge and diverge there are four possible track hypotheses as shown in the first column of Figure 4.3. Human

motion in indoor environment is highly non-linear and hence at the sensor level there is not enough information to reject the false hypotheses. Once the information from other sensors is also available, a composite tracker can evaluate the likelihood of each hypothesis, incorporating the information from the other sensor and the hypotheses with high likelihood are selected and tracked in the subsequent time frames. This process is graphically depicted in the second and third columns of Figure 4.3. Here, there are two distinct tracks in sensor 2, because there in no occlusion in its view. When the 4 hypotheses from sensor 1 are evaluated with the two distinct tracks from sensor 2, only two hypotheses survive with high likelihood. These surviving tracks are tracked in subsequent time frames. This hypothesis selection process described in Figure 4.3 is intuitive and the iterative decoding algorithm provides a statistical framework to implement it. In Figure 4.3 the flowchart of the MID-AVT framework is provided.

4.3.1 Feature extraction for the cameras

The video features are obtained from a simple foreground object detection scheme. The foreground pixels in a frame are detected by background subtraction. They are then fused into reliable blobs by morphological operations. A bounding rectangle is fit to each distinct blob. The pixel co-ordinates of the center of the ith rectangle, (x_{it}, y_{it}) and the area of the rectangle $[a_{it}]$ are the components of the observation vector $o_{it}^T = [x_{it} y_{it} a_{it}]$. For every frame at time t, for the jth camera, a list of the M_j detected foreground objects $o_{it}^j, 1 \leq i \leq M_j$ is generated.

4.3.2 Feature extraction for the microphone arrays

The audio features are based on the time delay of arrival(TDOA) estimates between pairs of microphones in an array to estimate the location of the sound source. The generalized cross correlation based phase transform (GCC-PHAT) framework [54][12] is used to locate sound sources if present. This technique has been the preferred method of TDOA estimation in established literature[35][72] as it has shown to be robust to reverberations. For simplicity, the TDOA estimates are

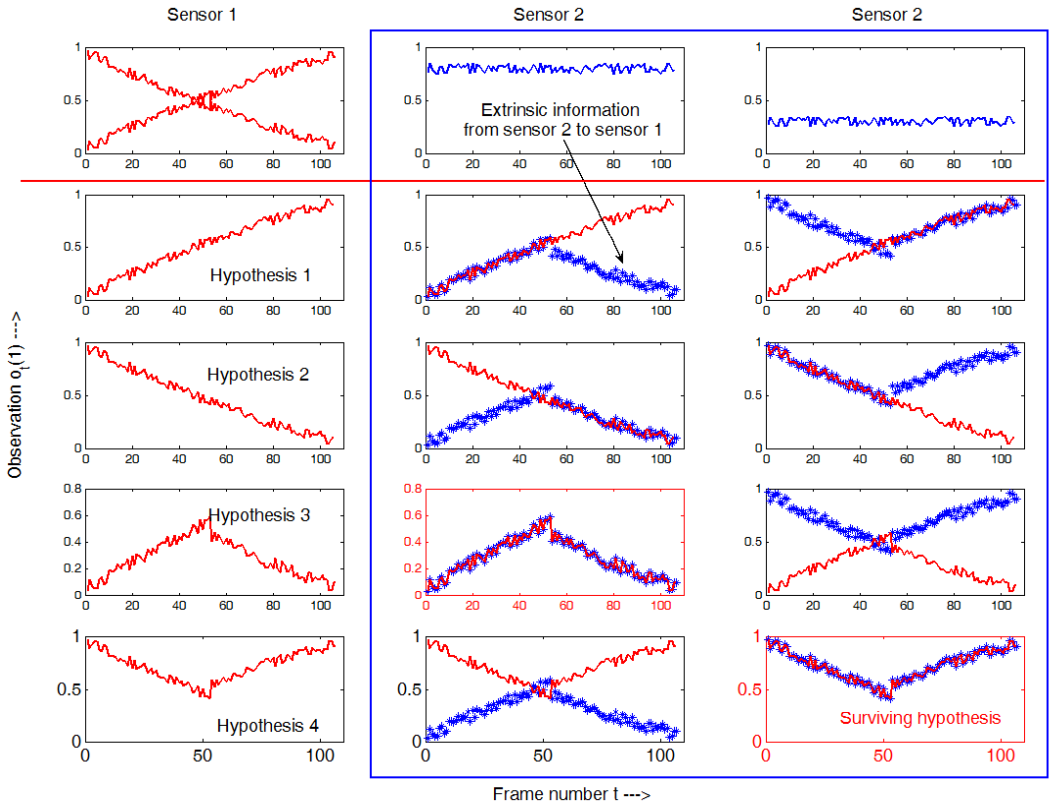

Figure 4.1: The disambiguation of confusable hypotheses using the iterative decoding scheme is illustrated here. The first graph shows the tracks as seen in one of the sensors. The next four images in the first column present the possible hypotheses that are plausible according to the first sensor alone. The second and third columns have two tracks in the field of view of sensor 2. Note that both the second and third column correspond to the same sensor. The extrinsic information that these tracks provide sensor 1 are shown in the next eight images, superimposed with the four hypotheses from sensor 1. The two surviving hypotheses are marked in red.

computed on time windows of audio samples corresponding to the interval between the camera frames. A vector of TDOA values between each microphone i and the reference microphone r is given by $\vec{\tau} = (\tau_{1r}, \tau_{2r} \ldots \tau_{mr})$. The TDOA estimates form the observation vector $o_{1,t}$ corresponding to the microphone array. Thus a microphone network behaves like a 3-d localizer similar to a camera. Note that the use of the SRP-PHAT technique [25] would allow the detection of multiple sound sources simultaneously. There would be M_t detected sources at each time instance

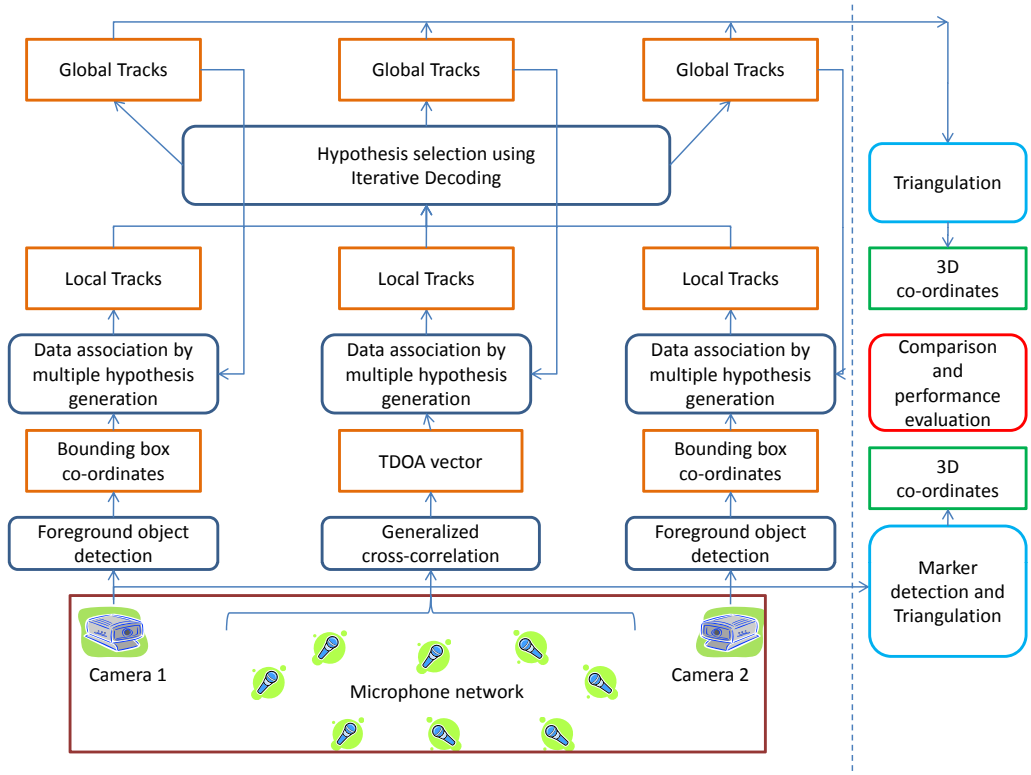

Figure 4.2: The MID-AVT framework involving the local and global track hierarchies along with the groundtruth estimation procedure.

and a list of observations, $o_{i,t}, 1 \leq i \leq M_t$. In the current chapter, it is assumed that there is only one sound source active at any particular time. This audio setup however differs from [72] in the arrangement of microphones. Traditional microphone arrays (linear/planar/spherical) have only angular resolution because the total span of the array is small compared to the source location. Large aperture microphone arrays have much wider total span and provide better resolution in the TDOA space. However such large aperture microphone arrays also require larger audio frames to accurately estimate the TDOA.

After the observations are extracted for each frame of audio, the cameras and the microphone arrays are treated equivalently as in [72]. In the next section the camera or the microphone array in general are referred to as a sensor.

4.3.3 Multiple hypotheses generation - local tracking

The object detection module associated with each sensor detects the foreground objects (or sound sources) in each frame. In the presence of multiple objects of interest, all distinguishable objects are detected by each camera. False positive errors could occur in the presence of background noise or clutter. False negative errors could occur due to occlusions. The tracking framework will address both these issues.

Consider frames from time $t = 1 \dots T$. Start with a list of features of detected objects $o_{i,t}, 1 \leq i \leq M_t$ at time t, where M_t is the number of detected objects at time t. In the current setup, $o_{i,t}$ are image coordinates of the detected objects. More elaborate features such as size, color can also be added under the same framework. To start with let the initial track value for track j be $l_{j,0}$. At each time step, the tracks are updated according to the rule $l_{j,t} = \{o_{i,t} | d(l_{j,t-1}, o_{i,t}) \leq r\}$, where $d(x, y)$ is the Euclidean distance between x and y. If more than one observation lies within Euclidean distance r from $l_{j,t-1}$, the old track is split to account for each such observation. If no observation lies within radius r, we assign the past value $l_{j,t-1}$ to the track. This corresponds to occlusions or the object leaving the field of 'view' of the sensor. This is a very simple data association framework and would result in a lot of false positives, as it maintains tracks corresponding to all the possibilities in case of any occlusions or merging and diverging of tracks. Only those possibilities are discarded where the data association can be completed without ambiguity based on nearest neighbors. In the next step, using the information from other tracks, the hypotheses that are unlikely under a probabilistic joint model are rejected.

4.3.4 Multiple hypotheses selection and filtering- global tracking

Consider the set of all hypotheses $h^k = l_j | 1 \leq j \leq N_k$ from sensor k which has N_k hypotheses. In the global tracking step, all possible combinations of these hypotheses, one from each sensor are considered. There are $\prod_k N_k$ such combina-

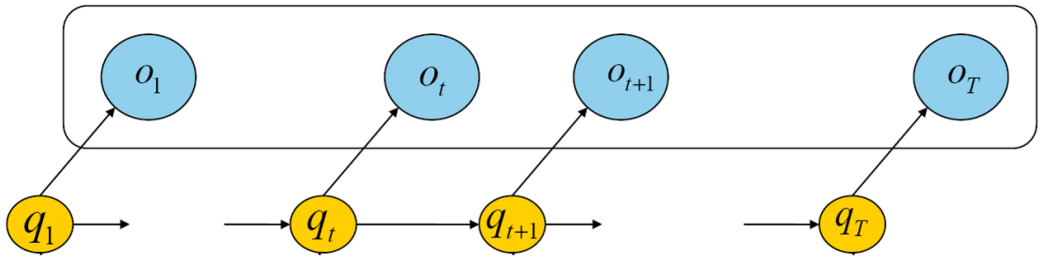

Figure 4.3: The HMM for smoothing the observations from sensor k. Note that the hidden states $q+t$ are described in the same feature space as the observations o_t and hence they are referred to as the temporally smoothed observations.

tions. The likelihood of each combination $C = (l^1_{j_1}, l^2_{j_2} \ldots l^N_{j_N})$ under the iterative decoding framework with HMM λ_k for sensor k is evaluated. Spurious tracks have a low likelihood and are discarded. The remaining tracks are then passed down to the local trackers to use as initial tracks for the next time window.

Since the MID-AVT framework is based on specific audio and video features and also includes an extra data association step, the iterative decoding algorithm is presented again in the next section as opposed to borrowing the same notation from the previous chapter.

4.3.5 Iterative decoding algorithm

Consider a hidden Markov model Λ_k for sensor k with N hidden states (see Figure 4). For clarity of presentation, the sensor index k is dropped. Λ has a parametric transition density. The hidden state q_t corresponds to the true location of the object at time t in the same feature space as the observation vectors of sensor k. Thus the hidden states are, in a Bayesian sense, the temporally smoothed observations. The conditional distribution of the observation o_t when the hidden state is q_t is assumed to be Gaussian. Now, the decoding problem is to estimate the optimal state sequence $Q^T_1 = \{q_1, q_2 \ldots q_T\}$ of the HMM based on the sequence of observations $O^T_1 = \{o_1, o_2 \ldots o_T\}$.

The Maximum aposteriori probability (MAP) state at time t is calculated using the BCJR (Bahl Cocke Jelinek and Raviv) algorithm[3] which is also re-

ferred to as the forward-backward sum-product algorithm in the graphical models community. Note that any other inference technique can also be used. The MAP estimate for the hidden state at time t is given by $\hat{q}_t = \arg\max P(q_t, O_1^T)$. The BCJR algorithm computes this using the forward and backward recursions.

The forward recursion variable $\alpha_t(m)$, the backward recursion variable $\beta_t(m)$, the joint likelihood of the hidden state and the observation sequence $\lambda_t(m)$ and the recursion variable $\gamma_t(m', m)$ are defined as follows,

$$\lambda_t(m) = P(q_t = m, O_1^T) \tag{4.1}$$

$$\alpha_t(m) = P(q_t = m, O_1^t) \tag{4.2}$$

$$\beta_t(m) = P(O_{t+1}^T | q_t = m) \tag{4.3}$$

$$\gamma_t(m', m) = P(q_t = m, o_t | q_{t-1} = m') \tag{4.4}$$

where, $m = 1, 2 \ldots N, m' = 1, 2 \ldots N$

Then establish the recursions,

$$\alpha_t(m) = \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \tag{4.5}$$

$$\beta_t(m) = \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m') \tag{4.6}$$

$$\lambda_t(m) = \alpha_t(m) \cdot \beta_t(m) \tag{4.7}$$

At the first sensor HMM, the hidden states are decoded using the observations from the first sensor. The following aposteriori probabilities are obtained, $\lambda_t^{(1)}(m) = P(q_t = m, O_1^T)$.

In the second sensor HMM, these aposteriori probabilities, $\lambda_t^{(1)}(m)$ are utilized as extrinsic information in decoding the hidden states from the observations of the second sensor. Thus the aposteriori probabilities in the second stage of decoding are given by $\lambda_t^{(2)}(m) = P(q_t = m, O_1^T, {Z^{(1)}}_1^T)$ where $Z_t^{(1)} = \lambda_t^{(1)}$ is the extrinsic information from the first sensor.

$$\lambda_t^{(2)}(m) = P(q_t = m, O_1^T, {Z^{(1)}}_1^T) \tag{4.8}$$

$$\alpha_t^{(2)}(m) = P(q_t = m, O_1^t, {Z^{(1)}}_1^t) \tag{4.9}$$

$$\beta_t^{(2)}(m) = P(O_{t+1}^T, {Z^{(1)}}_{t+1}^T | q_t = m) \tag{4.10}$$

$$\gamma_t^{(2)}(m', m) = P(q_t = m, o_t, Z_t^{(1)} | q_{t-1} = m') \tag{4.11}$$

In order to distinguish the hidden states of sensor 1 from those of sensor 2 at time t, denote them as $q_{1,t}$ and $q_{2,t}$ respectively. Similarly the observations are denoted by $o_{1,t}$ and $o_{2,t}$. Then the recursions do not change, except for the computation of $\gamma_t^{(2)}(m', m)$.Since the extrinsic information is independent of the observations from the second modality,

$$\begin{aligned}\gamma_t^{(2)}(m', m) =& P(q_{2,t} = m, o_{2,t}, Z_t^{(1)} | q_{2,t-1} = m') \\ \gamma_t^{(2)}(m', m) =& P(q_{2,t} = m | q_{2,t-1} = m') \\ & \cdot P(o_{2,t} | q_{2,t} = m) \cdot P(Z_t^{(1)} | q_{2,t} = m)\end{aligned}$$

$$\begin{aligned}\gamma_t^{(2)}(m', m) = & P(q_{2,t} = m | q_{2,t-1} = m') \cdot P(o_{2,t} | q_{2,t} = m) \\ & \cdot \sum_n \{P(Z_t^{(1)} | q_{1,t} = n) P(q_{1,t} = n | q_{2,t} = m)\}\end{aligned}$$

where $q_{2,t}$ and $o_{2,t}$ correspond to the hidden state and observation at time t for modality 2.

Assuming that $P(Z_t^{(1)} | q_{1,t}) = 1$ if $q_{1,t} = \arg\max_n Z_{t,n}^{(1)}$ and 0 otherwise, where $Z_{t,n}^{(1)}$ is the nth component of the vector $Z_t^{(1)}$, which corresponds to a hard decision rule, the missing piece in the framework is only the evaluation of $P(q_{1,t} = n | q_{2,t} = m)$. In section 4.3.6, the process of sensor calibration is described by which this distribution is estimated.

Alternatively, one can visualize the iterative decoding as follows. Consider the HMM based tracker for each sensor k. The observation model for this HMM which defines the conditional distribution of the observation $o_{k,t}$ when the hidden state is $q_{k,t}$ is assumed to be Gaussian. Now, the iterative decoding algorithm involves incorporating extrinsic information from sensor $k-1$ while decoding the hidden states of sensor k. In order to do so, augment the observation model to include the extrinsic information as well. The extrinsic information from sensor $k-1$ is denoted by $Z_t^{(k-1)}$. The augmented observation model is now represented as

$$P(o_{k,t}, Z_t^{(k-1)}|q_{k,t}) = P(o_t|q_t) \cdot P(Z_t^{(k-1)}|q_{2,t})$$
$$P(o_{k,t}, Z_t^{(k-1)}|q_{k,t}) = P(o_t|q_t) \cdot P(Z_t^{(k-1)}|q_{1,t}) \cdot P(q_{1,t}|q_{2,t})$$

Assuming that $P(Z_t^{(k-1)}|q_{k-1,t}) = 1$ if $q_{k-1,t} = \arg\max_n Z_{t,n}^{(k-1)}$ and 0 otherwise, where $Z_{t,n}^{(k-1)}$ is the nth component of the vector $Z_t^{(k-1)}$, which corresponds to a hard decision rule, one is now left with the evaluation of $P(q_{k-1,t} = n|q_{k,t} = m)$. In section 4.3.6, we describe the process of sensor calibration is described by which this distribution is estimated.

We proceed to sensor $k+1$ with the extrinsic information $Z^{(k)}$ from sensor k. We proceed likewise till we decode the hidden states of the last sensor from the extrinsic information of the previous sensor. In the next iteration, we use the extrinsic information of the last sensor to decode the hidden states of the first sensor. Then the second iteration proceeds as the first, with updated state sequences. Finally we threshold the overall log-likelihood of the track combinations to select the surviving tracks in each sensor 'view'.

4.3.6 Sensor Calibration

The camera and microphone locations are assumed to be arbitrary but fixed. Hence only a rough calibration step is needed to establish a relationship between the state space of different sensors. In the iterative decoding algorithm presented in section 4.3.5, the missing piece of the framework is the problem of estimating $P(q_{1,t} = n|q_{2,t} = m)$ for sensor pair $(1, 2)$. There are efficient ways of learning and storing this distribution by using decision trees, piecewise linear approximations and kernel based density estimation techniques [24]. In this chapter a simple kernel density estimation scheme is used to estimate the conditional distribution $P(q_{1,t} = n|q_{2,t} = m)$, by first estimating the joint distribution $P(q_{1,t} = n, q_{2,t} = m)$ from a set of training points collected during the calibration step. In order to collect training points, an initial calibration step where a single person carrying a sound source walks around the space monitored by the sensors is required. Tracking is

now trivial as there is only one object. The observations from several frames are used to estimate the joint distribution $P(q_{1,t} = n, q_{2,t} = m)$ using a Gaussian kernel of appropriate bandwidth for smoothing.

During the initial calibration phase, a person carrying a sound source walks around the room. From the audio signals, the TDOA vector corresponding to the sound source is computed and from the video frames, the (x, y) pixel co-ordinate of the foreground object is obtained. Note that the calibration step establishes correspondences between sensors in the sensor co-ordinate system. The calibration of the cameras and microphone arrays to the world co-ordinate system is not required for the MID-AVT framework. However this is required to measure the ground truth for evaluating the accuracy of the tracker and to compare it with other tracking schemes.

4.4 Experimental evaluation

The MID-AVT framework is evaluated on two different datasets MID-AVT-UCSD-1 and MID-AVT-UCSD-2 collected in the Audio-visual testbed at the Smartspaces laboratory at CALIT2, UCSD. The sensor and scene configuration as well the nature of these datasets is described in Appendix A.3. Also, the ground truth related to the actual location of the human subjects is also collected for both the datasets.

4.4.1 Evaluation Results

The MID-AVT framework and the particle filter based tracker from [17] were compared on the MID-AVT-UCSD-1 dataset. For the HMMs in the MID-AVT framework, 500 hidden states per sensor were used and 100 particles for each subject were used in the particle filter. All four cameras and the four cross shaped microphone arrays were used as sensors. Neither algorithm was implemented in real-time, however the iterative decoding algorithm was 2.5 times slower than the particle filtering approach. Moreover, the iterative decoding was carried out on blocks of length 5 seconds and hence there is a minimum delay of 5 seconds in

generating the global tracks. However there are applications such as automatic meeting summarization where such a delay is tolerable.

The tracker is evaluated by counting the number of frames a subject is tracked correctly (tracker output matches ground truth location by 500 mm). The MID-AVT scheme had an average accuracy of 76% on the MID-AVT-UCSD-1 dataset while tracking all the subjects in the meeting scene. The errors were mostly missed detections involving subjects who blended in with the background due to dark clothing and remained silent for most of the meetings. In Figure 4.4.1 the different views of one of the meeting scenes is shown. Note that one of the subjects is completely missing in the tracker output. Also, the active speaker was tracked using the audio detections alone and associating this detection with the corresponding global track during the course of the meetings and it was found that the active speaker was accurately found in 85% of total frames. In Figure 4.4.1 snapshots from the global tracking process as seen from one of the camera views is shown. Note that the active-speaker tracking tracks the different active speakers as they take turns in the conversation. However in the meeting scenes there is only one dominant speaker and hence the audio observations do not improve the localization accuracy of the tracker. Also, there is not much movement of the seated participants which is not a very challenging tracking scenario. The average root mean-squared error of the speaker location was 21cm. The particle filter based tracker was evaluated and was found to perform with an accuracy of 74%. There is no appreciable difference between the performance of the two trackers.

MID-AVT-UCSD-2 dataset involves a more challenging tracking scenario. In Figure 4.4.1 a snapshot of a scene from this dataset is shown. In Table 4.3, the fraction of times the global tracker successfully resolves the ambiguity during occlusions and noisy detections based on the information from the other sensors is presented. In Figure 4.4.1, one of the tracks from a clip and the associated groundtruth is shown. The root mean squared error between the track and the ground truth is 11cm. Again, the performance of the particle filter scheme is very similar to that of MID-AVT framework.

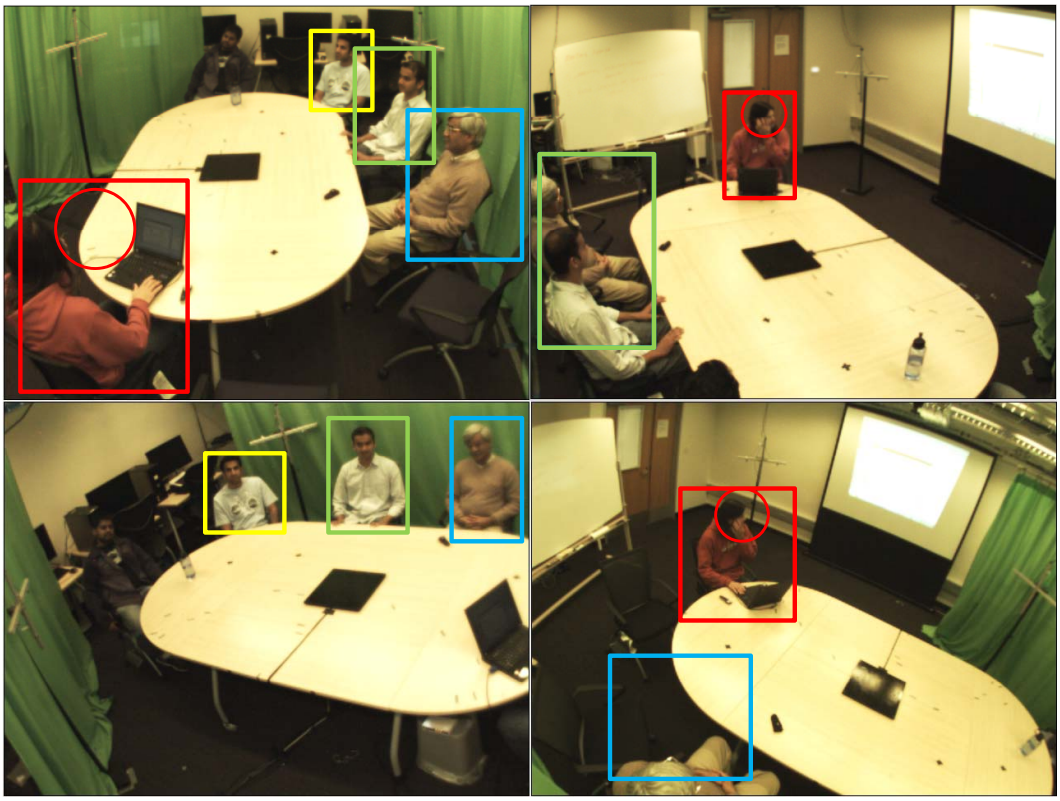

Figure 4.4: A snapshot showing the different views from the tracker. Note that at the moment the snapshot was taken, one subject was missed by the tracker due to lack of contrast with the background. He also remained silent during the meeting and was not picked up by the audio localizer either.

4.4.2 Sensitivity to sensor calibration

In order to demonstrate the robustness of the iterative decoding scheme, the calibration mismatch is simulted by applying a small fixed random rotation transformation to each camera view. This corresponds approximately to the case where the camera calibration is inaccurate. In this new configuration the experiments were repeated on the MID-AVT-UCSD-2 dataset and the results are presented in Table 4.4. Five random rotation transformations (and in each case different cameras were perturbed by different angles) were applied to the videos. Each rotation was selected randomly to lie between -10^o and 10^o around the camera axis. The average results are shown in Table 4.4. The performance of the particle filter tracker degrades considerably while the proposed framework maintains the

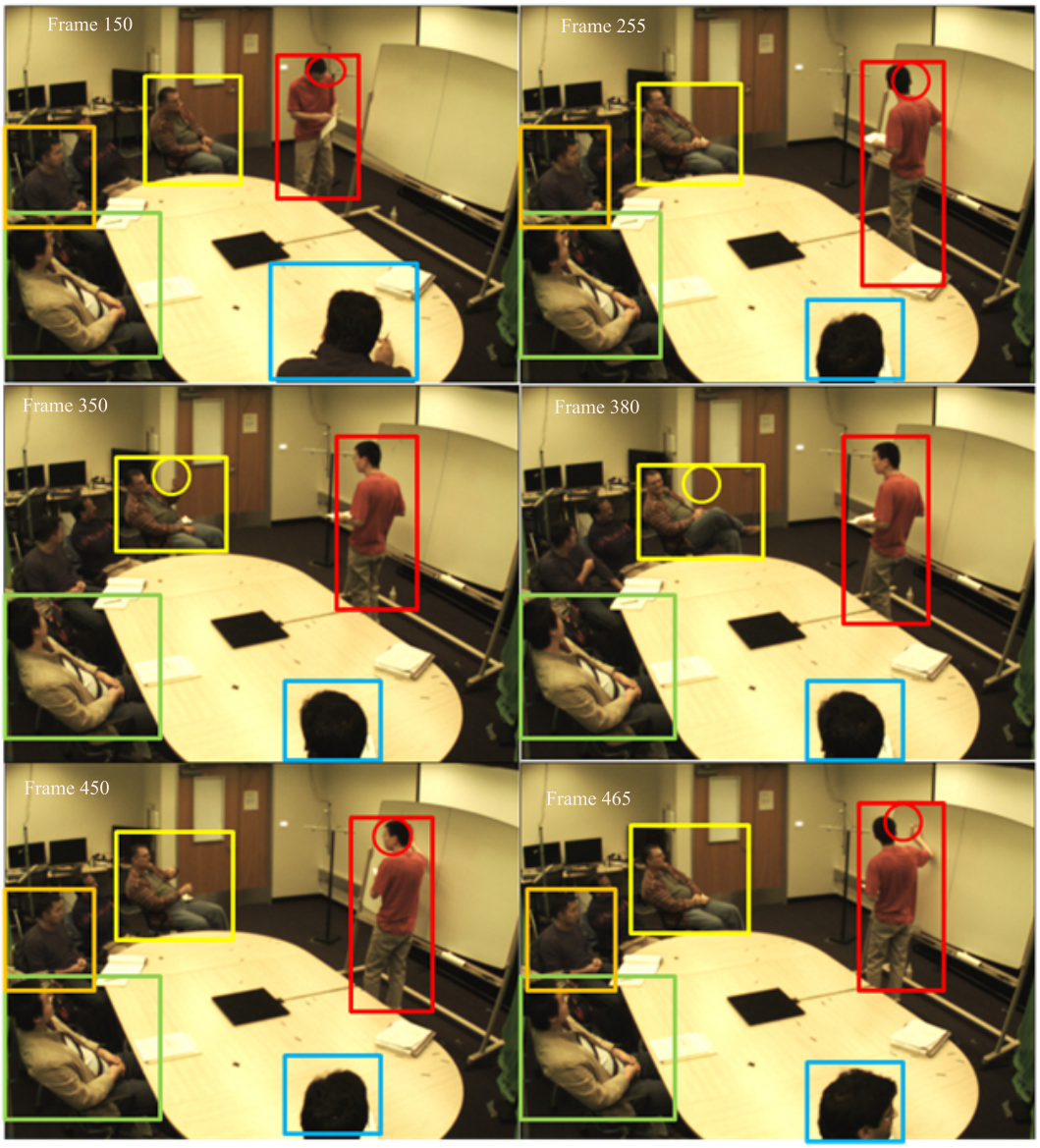

Figure 4.5: Different snapshots during a meeting illustrate the active speaker tracking that highlights the current active speaker by drawing a circle around the head of the associated track.

tracking accuracy. The particle filter maintains the tracks in the 3-D co-ordinates and hence the mismatched calibration affects the tracking process. However, in the proposed MID-AVT framework, local tracking occurs in the image co-ordinates and is robust to calibration errors. Note that here only a small perturbation to the sensor configuration has been applied.

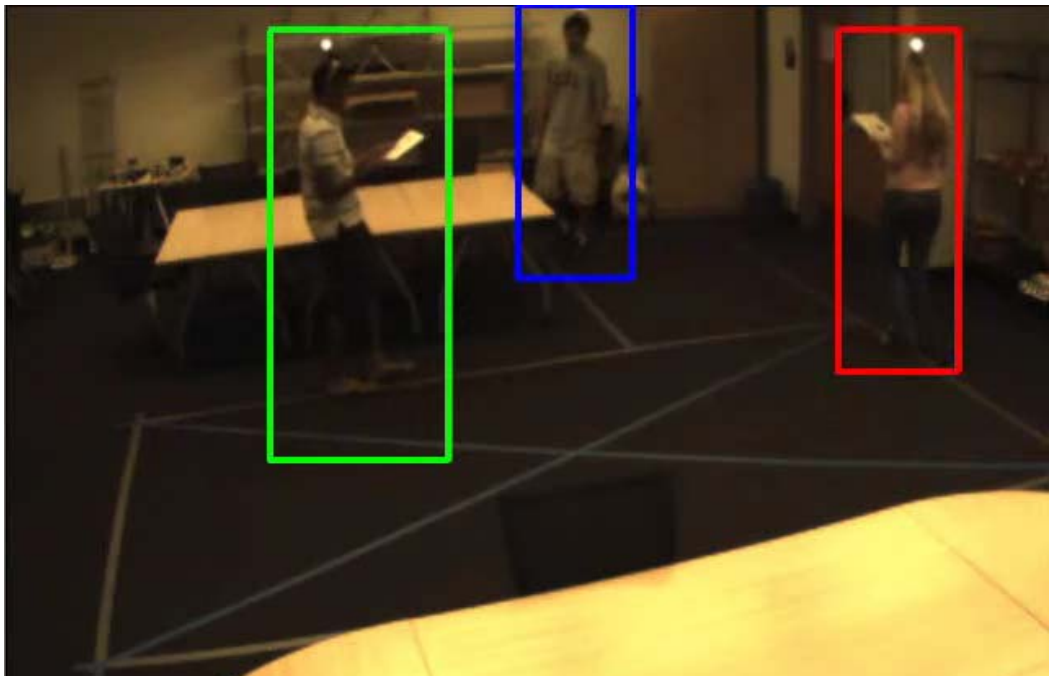

Figure 4.6: A snapshot from the tracking process on data set 2 which involves subjects moving continuously and hence resulting a lot of occlusions, with tracks merging and diverging in camera views.

4.5 Concluding remarks on the MID-AVT framework

In this chapter the MID-AVT framework is developed and evaluated for tracking multiple subjects in a space instrumented with multiple cameras and microphone arrays. The MID-AVT framework extends the iterative decoding algorithm by including a data association step to select appropriate track hypotheses from different sensor views. The performance of the MID-AVT tracker is similar to the popular particle filter based audio-visual tracker. However there are distinct advantages to the MID-AVT framework. It is modular and hence easy to expand to more number of cameras and microphone arrays or any other sensors that can localize persons. It is also applicable to sensors with overlapping and non-overlapping field of 'view'. Since the placement of the sensors is assumed to be arbitrary but fixed, only a rough calibration scheme is necessary to establish

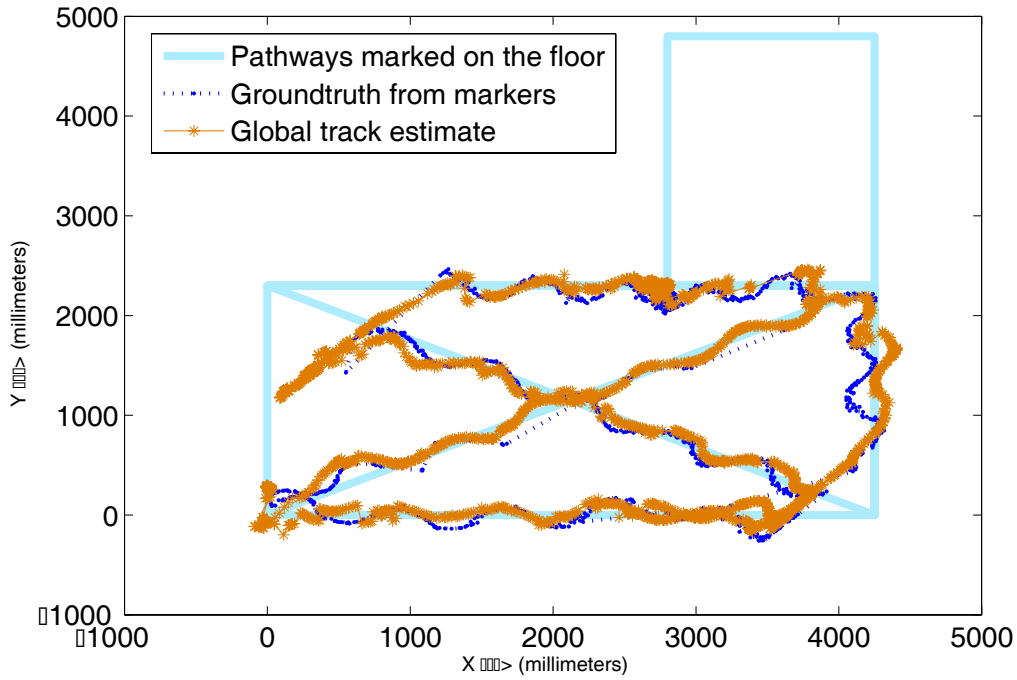

Figure 4.7: A track and its associated ground truth in world co-ordinates.

the correspondence between sensors. Moreover, the performance of the MID-AVT tracker is robust to small errors in sensor calibration. However, tracking human subjects is only the first step in analyzing human activity in an intelligent space. The situational awareness needed in an intelligent space is developed by fusing information at multiple levels of semantic abstraction. In the next chapter, an advanced hierarchical framework is presented to extract different kinds of information apart from tracking to facilitate the development of situational awareness in intelligent spaces. Audio-visual fusion is explored in the context of such a hierarchical framework.

4.6 Acknowledgments

The text of Chapter 4, in full, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, “Audio Visual

Fusion and Tracking With Multilevel Iterative Decoding: Framework and Experimental Evaluation", IEEE Journal of Selected Topics in Signal Processing, Special issue on Speech Processing for Natural Interaction with Intelligent Environments, July 2010. The dissertation author was the primary investigator and author of this paper.

Table 4.1: Summary of fusion strategies in audio-visual person localization and tracking

Fusion strategy for audio-visual person localization and tracking	Sensors	Scene complexity	Model	Publication and Year
Proximity based speaker association	1,2	S	Camera epipolar geometry and audio cross-correlation	Pingali et. al. [77] 1999
SNR based weighted average of SPMs	2,3	S	Spatial probability maps	Aarabi [75] 2001
Feature concatenation without weighting	1,2	S	Probabilistic tracking with particle filters	Vermaak et. al. [104] 2001
Proximity based association of audio and visual events	1,4	M	Auditory epipolar geometry and face localization	Nakadai et. al. [67] 2001
Product rule	2,14	M	Probabilistic tracking with particle filters	Zotkin et. al. [114] 2002
Importance sampling and product rule	2,14	M	Probabilistic tracking with particle filters	Gatica-Perez et. al. [76] 2003
Speaker detection using audio	1,3	M	Skin tone based face detection in omni-camera	Kapralos et. al. [50] 2003
Feature concatenation without weighting	1,2	M	Bayesian network	Beal et. al. [5] 2003

Table 4.2: Summary of fusion strategies in audio-visual person localization and tracking (contd.)

Fusion strategy for audio-visual person localization and tracking	Sensors	Scene complexity	Model	Publication and Year
Weighted addition of proposal distributions from each sensor	5,2	M	Probabilistic tracking with particle filters	Chen and Rui [20] 2004
Product rule	2,16	M	Probabilistic tracking with particle filters	Checka et. al. [17] 2004
Sequential state update using audio and video	4,16	M	Iterated extended Kalman filter	Gehrig et. al. [36] 2005
Feature concatenation without weighting	2,14	M	Markov Chain Monte Carlo particle filter	Gatica-Perez et. al. [35] 2007
Feature concatenation without weighting	4,14	M	Particle filter	Bernardin et. al. [6] 2007
Finite state machine for appropriate weighting	1,14	M	Particle filter	Bernardin et. al. [6] 2007
Iterative decoding algorithm	2, 8	M	Hidden Markov Model	Shivappa et. al. [90] 2010

Table 4.3: Results from MID-AVT-UCSD-2 - percentage of occlusions that are correctly resolved by the MID-AVT framework in comparison with the Particle filtering based tracker. Note that the performance of the two schemes is very similar.

		1 camera	Mics.	1 camera and mics.	2 cameras	2 cameras and mics.
MID-AVT framework						
	1 subject	95%	76%	95%	98%	98%
	2 subjects	53%	42%	68%	85%	87%
	3 subjects	38%	40%	65%	80%	83%
	4 subjects	**33%**	**34%**	**55%**	**69%**	**73%**
Particle filter based tracker						
	4 subjects	**30%**	**20%**	**35%**	**69%**	**74%**

Table 4.4: Results from MID-AVT-UCSD-2 dataset (4 subject case) when a random rotation transformation is applied to the camera views - percentage of occlusions that are correctly resolved by the tracker is shown in the table. This demonstrates that the MID-AVT framework is robust to small camera calibration errors.

	1 camera	Mics.	1 camera and mics.	2 cameras	2 cameras and mics.
MID-AVT framework	**33%**	**34%**	**55%**	**68%**	**70%**
Particle filter based tracker	**20%**	**20%**	**35%**	**49%**	**54%**

Chapter 5

Hierarchical frameworks for audio-visual information fusion in meeting scenes

5.1 Introduction

Audio visual fusion has been recognized as a critical component in the design of intelligent systems. Traditional fusion schemes have focussed on feature, classifier and decision level fusion[91]. More recently, hybrid as well as hierarchical fusion schemes have been explored to extend the benefits of fusion to more complex tasks such as meeting scene analysis, smart health homes and intelligent automobiles. Traditional fusion schemes have mostly focussed on task based fusion schemes. Audio-visual fusion for speech recognition, person tracking, person identification, emotion recognition have been explored and are also areas of active research. However, in practice, several such tasks have to be co-performed to provide the situational awareness that is required by an effective intelligent system. For example, an intelligent robot will be expected to simultaneously perform the tasks of speaker localization, speech recognition, speaker identification and emotion recognition in order to provide a wholesome communication experience. Similarly, a meeting scene analysis system requires the tracking of human subjects,

person identification, speaker localization and speech recognition to automatically analyze meeting scenes.

It can be shown that when audio-visual fusion is explored in the context of such co-performed tasks, not only is an hierarchical integration of audio and video cues necessary, but it is also beneficial to the performance of the individual tasks because the output of one kind of human activity analysis task contains valuable information for another such task and by interconnecting them, a robust system results. In this chapter we focus our attention on a meeting scene analysis system and present the results of our research in hierarchical fusion schemes in this context. A meeting scene involves complex interaction between multiple human subjects in an environment that feels natural to the participants. Hence the complete understanding of a meeting scene involves information at multiple levels of semantic abstraction. Hence a hierarchical fusion strategy is very relevant in this context. A hierarchical scheme is presented to infer the structure and dynamics of a meeting scene. Several competing methods for fusion are compared. Immediate application areas of such a meeting analysis system include meeting archival and teleconferencing. The robust methods to analyze and interpret meeting scenes that are developed in this chapter lead to efficient archival and retrieval of information hidden in hours of audio-visual recordings of meeting scenes. In a broader sense, we also believe that the fusion schemes developed in the context of meeting scenes can help design such hierarchical fusion schemes in other application domains such as intelligent vehicles, smart homes and natural human-computer interaction. In Figure 5.1, the interconnected blocks in the hierarchical fusion framework for meeting scene analysis are illustrated. Though the specific details will be brought out in later sections, it can seen from Figure 5.1 that a task such as person tracking can be accomplished by fusing audio and video cues whereas a task such as speaker identification can be assisted by face recognition which in turn can use information from the person tracking block. Thus, audio-visual fusion can occur at different levels and hence the name hierarchical fusion.

In Section 5.2, we present a brief summary of research in the field of audio-visual analysis of meetings and hierarchical fusions schemes. The advantages of

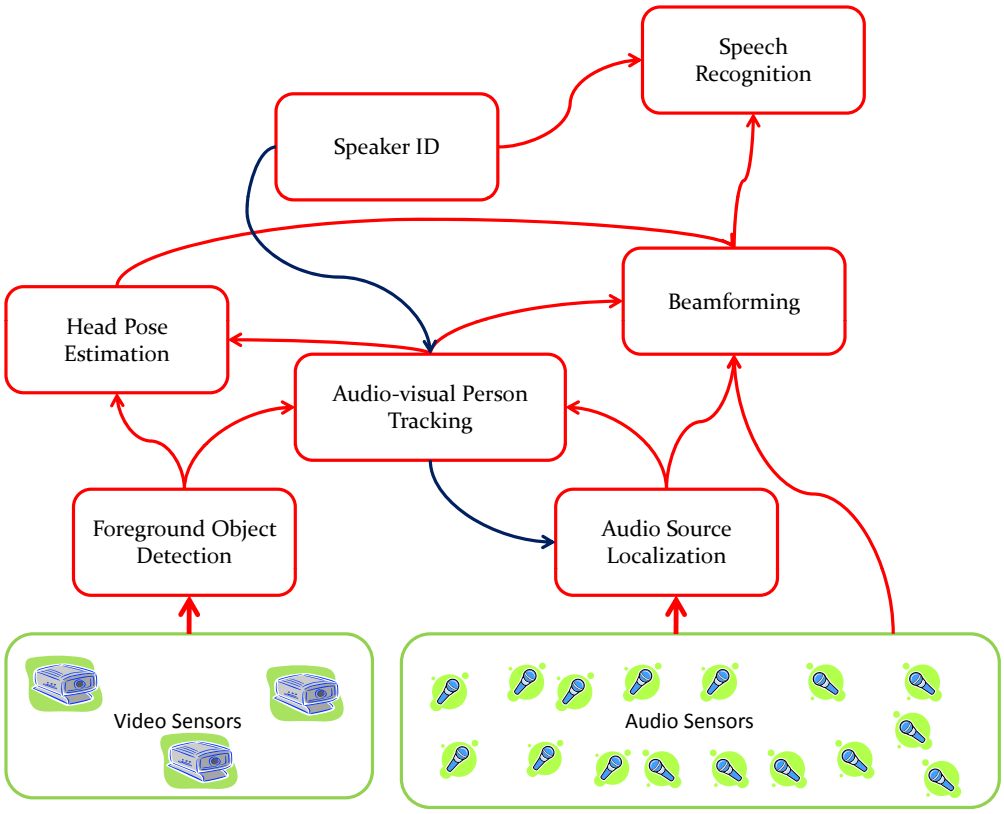

Figure 5.1: Flowchart summarizing the exchange of audio and visual cues at multiple levels of semantic abstraction in our meeting analysis system.

hierarchical fusion schemes are examined and various possible fusion paradigms are analyzed in Section 5.3. In Sections 5.9.1, 5.5.1 and 5.6, several individual components that make up the meeting analysis system are described. In each of these sections, we motivate the utility of the respective task, the challenges and existing algorithms. We then describe how contextual constraints as well as the fusion of information from other tasks can simplify and improve these algorithms and compare the performance before and after fusion. The meeting room and sensor configurations in the testbed in Appendix A. Though the contextual constraints and fusion strategies explored in this chapter and the next chapter are in the domain of meeting scene analysis, possible extensions to other domains and applications where such a hierarchical fusion might be beneficial are presented in Section 6.6.

5.2 Literature review in meeting scene analysis

In this section we will present a very brief overview of existing work in the area of hierarchical schemes for human activity analysis as well as recent research in meeting analysis systems.

One of the early research studies in observing human activities in an instrumented room is described in [101]. A graphical summary of the human activity is generated. The audio and visual information is used in identifying the current speaker based on a rule based decision fusion. [11] and [39] describe another meeting room analysis system which also fuses audio-visual stream for person identification, in addition to using the audio for automatic transcription and archival purposes. [79] investigates speech, gaze and gesture cues for high level segmentation of a discourse into topical segments based on a psycholinguistic model. In [73], the authors propose a hierarchial HMM framework for modeling human activity. More recent hierarchical fusion strategies include [113][74][23][8]. In [8], the authors develop a probabilistic integration framework for fusion of audio visual cues at the track and identity levels. This is an example of fusion at multiple levels of abstraction. Similarly, in [87], the utility of head pose estimation and tracking for speech recognition from distant microphones is explored. In [97] the role of contextual fusion in emotion recognition is explored.

Recently, there has been a lot of interest in developing smart meeting room technologies. [109] is a recent survey of techniques and existing challenges in the domain of smart meeting rooms. Also, a number of multimodal meeting rooms equipped with multimodal sensors have been established by various research groups and consortiums. Annotated audio-visual corpora have been collected and standard evaluations have been organized to compare existing frameworks on specific tasks. Though our work in this chapter does not closely align with the existing frameworks, it is necessary to view its practical implications in comparison to existing schemes for meeting analysis. A recent effort in collecting and organizing multimodal corpora is presented in [53]. Recent evaluations of meeting scene analysis systems include the CLEAR 2006 evaluation [93] and CLEAR 2007 evaluation [94].

In our previous study, we presented the MID-AVT framework for tracking persons using audio and visual cues [90] which included extensive evaluation in real world meetings. However, a person tracking framework is only the first step in analyzing a meeting scene. More cues such as person identity, active speaker location and clean speech for recognition need to be extracted. The main challenge is to perform this cue extraction in a robust manner using far field sensors under unconstrained conditions in a typical real-world meeting scene. In [82], the authors present a multimodal fusion approach for speaker localization and segmentation. This corresponds to a classifier level fusion of audio-visual cues. In contrast, the hierarchical fusion framework presented here is shown to address this challenge using a different approach - using the information from different tasks to assist other tasks, it is possible to design simple yet effective algorithms at each step of the hierarchical framework. Further simplification can be achieved by utilizing the domain knowledge or the typical scene configuration. This is also illustrated in the current chapter. In order for the fusion strategy to be practical, it needs to be applicable in different sensor and scene configurations as well as adapt to changes in the same. To this end we present a semi supervised learning scheme which allows the classifications models to be updated in an online manner, with minimum supervision, needed only when training models for new subjects.

5.3 Hierarchical fusion schemes

In general, audio-visual information fusion can involve the following scenarios (as illustrated with examples in Figure 5.2.)

- **Reduce the search space in classification tasks** - Audio or visual cues can be used to restrict the search space for classification tasks using the corresponding other modality. In case of parametric statistical model based classification, this can be achieved by having specially trained models for different contexts and switch between these models using the audio or visual cues. Typically, the statistical models are easier to train and have better performance for individual contexts. The task of audio-visual fusion in this case

is to robustly identify the particular context based on complimentary cues. We provide two specific examples of such contextual modeling in figure 5.2. In other cases where classification models are based on minimum distance or maximum likelihood, the audio and visual cues can be used to restrict the set of possibilities over which the minimum/maximum is evaluated. In Figure 5.2, a speaker localization framework based on video based search space reduction is given as an example of such a fusion strategy.

- **Semi-supervised/unsupervised Learning of classification models** - The audio and visual cues can be used to select a training set for training statistical models for the classification tasks. This is particularly necessary for the contextual modeling stated above, to be successful. The audio-visual fusion reduces the effort to label the training sets and leads to a semi-supervised or in certain cases, unsupervised learning algorithms that can automatically update the contextual models. The challenge lies in identifying the cues based on their ease and robustness of detection and the minimum supervision needed in labeling the data. In Figure 5.2, we provide a specific example of using face recognition to train speaker recognition models.

- **Calibration of sensors** - In cases where multiple cameras and microphones are used to collect audio and visual cues, the calibration of the sensors with respect to each other and with respect to the room co-ordinates is an important issue. One solution is to develop algorithms that work in the sensor co-ordinates. Another approach is to develop sensor calibration techniques that use the cameras and microphones together to calibrate each other.

- **Traditional fusion strategies** - The most commonly encountered examples of audio visual fusion in literature are cases where the audio and visual modalities carry complimentary information from the same underlying process as in the case of acoustic waveforms and lip movements conveying information about the underlying speech segment. The fusion challenge is to develop inference algorithms to decipher the underlying process based on the audio and visual observations. This is achieved by fusing the cues at the

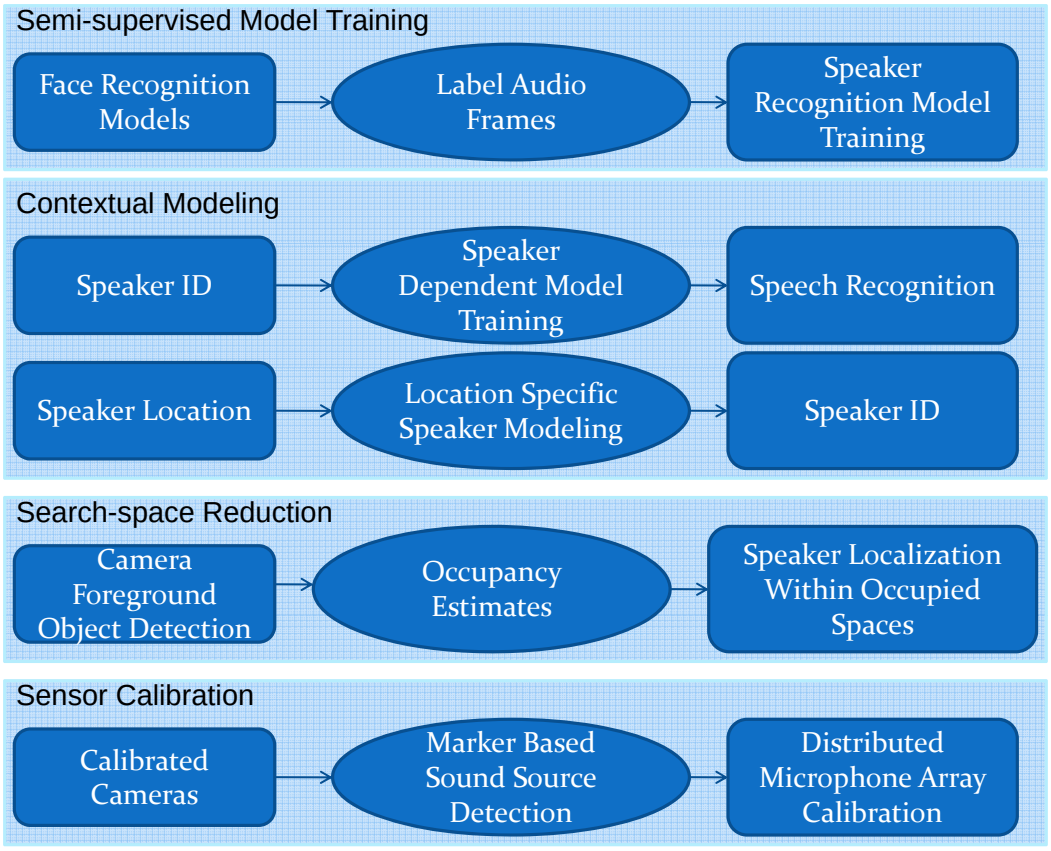

Figure 5.2: Different fusion paradigms apart from the traditional audio-visual fusion scenario are presented here with examples.

feature, classifier or decision levels.

5.4 Speaker identification using location specific speaker models (LSSM)

In the next few sections, we describe the various blocks of our meeting analysis system and present specific cases of the fusion paradigms described above.

5.5 Person Tracking and Speaker Localization

Tracking the subjects of the meeting is the fundamental step in a meeting analysis system [7][45]. In earlier systems, person tracking has been mostly explored in the context of video surveillance systems. In our previous work, [91], we have addressed the issue of tracking multiple subjects using cues from microphone arrays and cameras. The MID-AVT framework presented in [91] is a probabilistic framework and its performance is compared with the other audio-visual person tracking schemes. The MID-AVT framework has been shown to be robust to sensor calibration errors which is a major issue in practical systems.

Typically, in a meeting scene, there is minimal movement of the subjects and varying lighting conditions during presentations which affects the robustness of the background subtraction schemes. Also, during the course of a meeting, the focus is on the active speaker and thus it is necessary to robustly track the current active speaker. Also, in a meetings scene, we have certain constraints that can help us simplify and improve the performance of the person tracking block. Specifically, we can use planar microphone arrays positioned on the table as the participants are seated around the table. However such a planar microphone array provides angular resolution of the speakers and this restricts their applicability to scenes where all the participants are seated around the table and no participant is obstructed by another participant such that they are at the same angular position relative to the center of the microphone array on the table. The problem of speaker localization can be further simplified by discretizing the search space as the participants are likely to occupy certain specific locations around the table. These locations can be learned from earlier meeting recordings as in [63] or can be assigned at the beginning and refined as more meetings take place as in our framework presented here.

5.5.1 Discrete speaker location based active speaker localization

We assume that the participants of a meeting are likely to occupy certain locations, usually where the chairs are placed, around a table. We start with an exhaustive list of possible locations $\mathbf{L} = \{l_1, l_2, \ldots l_M\}$. A further simplification is to select a best view camera for each location and denote the bounding box for each location l_j in the camera co-ordinates as $V(l_j)$. We shall also denote the microphone array's time difference of arrival (TDOA) vector for location l_j as $T(l_j)$. A simple approach to localizing the active speaker is to estimate the current TDOA vector and compute the nearest location based on Euclidean distance in the TDOA space. The current active speaker location is thus given by $l_t = \arg\min_j |T(t) - T(l_j)|^2$. An inherent assumption in this case is that more than one speaker is not active at a particular time. However, if a robust algorithm can be developed for locating multiple simultaneous speakers, it can be used in our framework to remove this restriction. Since the main focus of this thesis is to explore the nuances of audio-visual fusion, we have not explored more complex algorithms for localizing multiple speakers.

5.5.2 Visual cues to reduce the speaker localization search space

Audio based speaker localization can be further improved by using video cues to restrict the search space. It is a relatively simple task in the video domain to detect the presence or absence of a foreground at a specific location. We use a simple background subtraction followed by thresholds to detect the presence of a human subject at a particular location. For this purpose, the bounding box for each location $B(l_j)$ from the best view camera is used and the percentage of foreground pixels within the bounding box $F(l_j)$ are estimated and the particular location is used in the search for the active speaker only if this percentage exceeds a threshold. $l_t = \arg\min_{j|F(l_j)>t} |T(t) - T(l_j)|^2$. The overall algorithm is summarized in Figure 5.3.

Table 5.1: Comparative performance of active speaker localization with and without visual fusion on our meeting dataset.

Number of Subjects Present	Number of Possible Speaker Locations Searched	Accuracy without visual cues	Accuracy with visual cue fusion
3	5	85%	97%
3	6	78%	96%
3	7	70%	96%
4	5	85%	94%
4	6	82%	94%
4	7	73%	94%

In Figure 5.4 we illustrate the advantage of using the visual cues in reducing the search space in a sample meeting scene with three speakers. We observe that by restricting the nearest speaker location search to the set of locations that are actually occupied by subjects, instead of an exhaustive search over all the possible locations around the table, the localization accuracy improves by 21% on our dataset consisting of a typical 5 minute meeting clip involving three subjects. We evaluated the visual cue fusion for the speaker localization task on our real meeting dataset collected in our audio visual testbed described in Appendix A. The dataset consists of 10 meetings each of 5 minute duration. There are 3 to 4 subjects in the meetings and the results are shown in Table 5.1. The improvement in performance is very critical in our current set up because the active speaker location is a fundamental cue that is used in further tasks and any loss of performance here is propagated to the higher semantic levels. This is an example of audio-visual fusion where easily extractable visual cues are used to reduce the search space and increase the robustness of an audio processing task.

As we shall see in the next few sections, active speaker location is a fundamental piece of information for the other tasks in the meeting analysis framework and any error at this step is likely to be propagated to the higher levels of analysis.

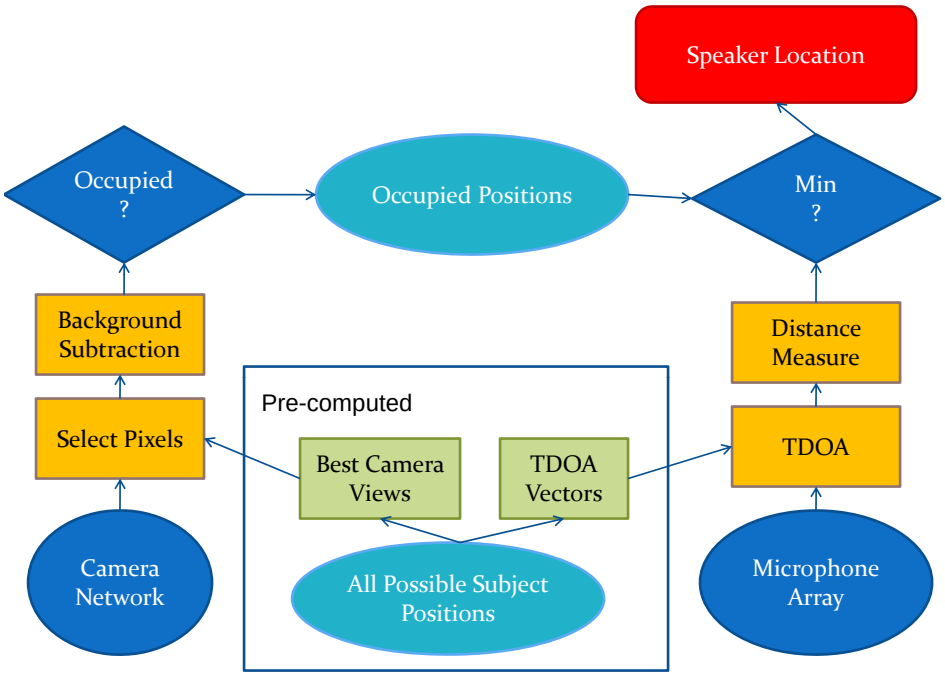

Figure 5.3: Flowchart summarizing the fusion of visual cues in the speaker localization task.

By utilizing robust and simple visual cues we have succeeded in obtaining a very low error rate in the speaker location estimates which significantly improves the overall performance of our meeting analysis systems.

5.6 Speaker Identification

In the previous section, we described an audio-visual fusion approach to detect the active speaker in a meeting scene. In this section we will focus our attention on the next important step in a meeting analysis system which is the identification of the active speaker. The main challenge is to recognize the current speaker using far-field microphones. The alternative is to have all the meeting participants wear lapel microphones which contradicts the non-intrusive nature of

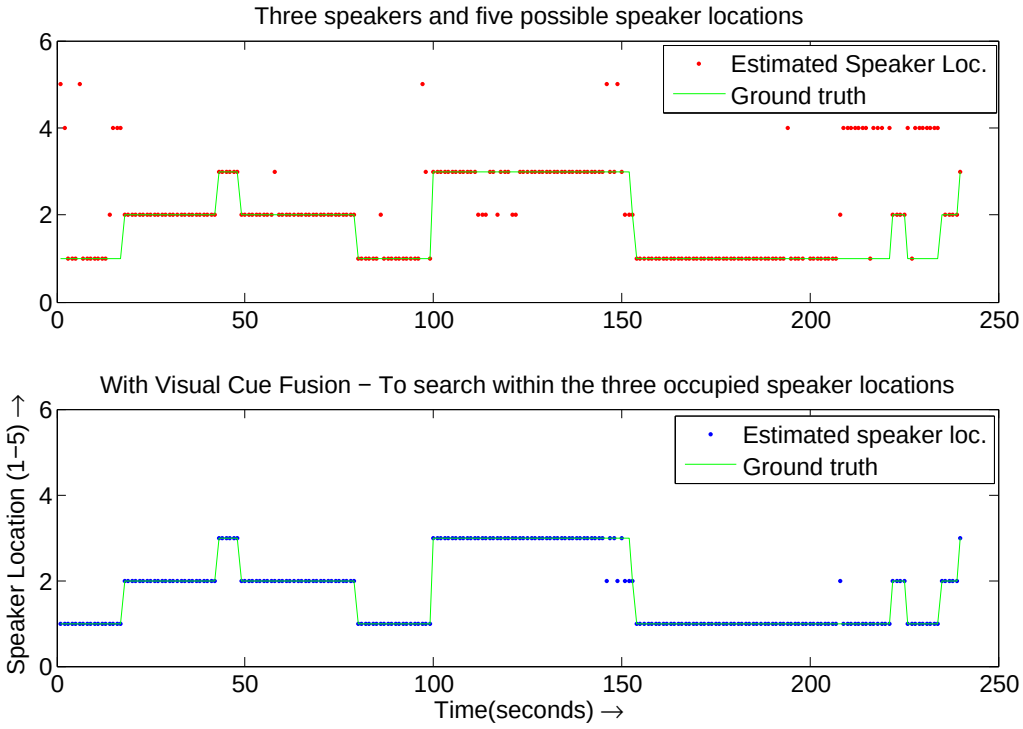

Figure 5.4: A typical meeting segment with three subjects where the search over all five possible subject locations leads to more errors than when the search is limited to the occupied spots based on visual cues.

the intelligent space design. We will first explore a standard speaker identification schemes based on Gaussian mixture models (GMMs).

5.6.1 Gaussian Mixture Models

In this section we describe a Gaussian mixture model (GMM) based text-independent speaker recognizer for a single microphone[81]. The speech signal from the microphone is windowed into frames of approximately $25ms$ duration with a $10ms$ overlap between windows. A energy threshold is used to detect speech frames and discard silence frames. 12 Mel-spaced cepstral coefficients (MFCC) are extracted for each speech frame and these constitute the feature vector for the GMM classifier. At frame index t, let the MFCC feature vector be denoted by x_t.

The GMM defines the likelihood for x_t as

$$p(x_t|\lambda) = \sum_{m=1}^{M} w_m N(x_t|\mu_m, \Sigma_m)$$

where w_m are the mixture weights and $N(x_t|\mu_m, \Sigma_m)$ is a multivariate Gaussian density function with mean μ_m and co-variance matrix Σ_m. The Expectation maximization algorithm is used to choose the GMM parameters for each speaker based on maximum likelihood (ML). Note that to train the GMM, the training observations $X = \{x_1, x_2 \dots x_\tau\}$ are used frame-wise. The GMM does not model the inter-frame dependence that is characteristic of speech waveforms. Some existing work suggest to augment the feature vector with Δ and $\Delta - \Delta$ (velocity and acceleration) components to better model the inter-frame dependence. We do not consider this issue here. However, the inclusion of these features should if any, increase the performance of our proposed scheme.

For a matched training-testing speaker location in our experimental setup, the average recognition accuracy was 94%. When the test sequence was from a new location, the accuracy dropped to 62%. This rapid degradation in recognition accuracy can be attributed to the varying channels between the test and the training sets. In a reverberant room, the far-field microphone receives the speech signal $s(n)$ directly from the speaker along with considerable reflections from the walls and other surfaces in the room. These can be encapsulated in the form of a transmission channel $h(n)$ between the speaker and the microphone. Hence the received signal at the microphone is $y(n) = s(n) * h(n)$. Expressed in the frequency domain for each frame at time t, one can represent this as $Y_t(e^{j\omega}) = S_t(e^{j\omega}).H(e^{j\omega})$. Note that the channel is a function of the speaker location and hence the degradation in performance for mismatched training and testing speaker locations.

Two main approaches have been explored in existing literature to address this issue which we will review now.

5.6.2 Cepstral Mean Subtraction

The MFCC feature vector computation involves taking the log of the magnitude of the frequency spectrum of each frame of speech which can be represented

as $log(|Y_t(e^{j\omega})|) = log(|S_t(e^{j\omega})|) + log(|H(e^{j\omega})|)$. Note that an average of Y_t with respect to time t yields an estimate of $log(|H(e^{j\omega}|)$ + some long term characteristic of speech signal which can be discarded to retain only the relevant information in $log(|S_t(e^{j\omega})|)$. This is the basis of Cepstral mean subtraction (CMS) technique. One issue in using (CMS) directly in a reverberant environment is that the CMS cannot account for the channel response that is longer than the $25ms$ frame. This is usually the case in typical medium sized meeting rooms. [48] proposes a spectrum subtraction technique to undo the effects of the long channel response by treating the long tail of the channel response akin to noise that needs to be incorporated in the model. However even such a modeling does not give us the performance close to a matched situation. In our experiments, we performed CMS over $10s$ time window and the results are summarized in Table 5.2.

Thus we see that the CMS technique provides improvement over the mismatched speaker location scenario (75% instead of 62%) but the accuracy is not comparable to that of the matched situation. Also, CMS results in slightly worse performance even in the matched scenario.

5.6.3 Microphone Array Beamforming

Another approach to deal with the reverberation is to enhance the quality of speech using a beamformer[37][60]. Here multiple microphones are used to enhance the quality of the speech signal. A beamformer works by enhancing the signal from the location of interest while suppressing signals from other locations. We use the speaker localization results from section 5.5.1 in a delay and sum beamformer with a six channel microphone array. The output of beamformer is used to train and test the speaker recognition system. The results are summarized in Table 5.2. Though the beamformer improves the performance in the presence of background noise, it is not so effective in dealing with reverberation.

5.6.4 Location Specific Speaker Modeling

We see that while using techniques such as CMS provides us some improvement, even using a microphone array instead of a single microphone does not provide us performance close to the matched speaker location condition for training and testing. We propose a novel approach to use different models trained for specific speaker locations along with a microphone array based speaker location estimate to choose the appropriate model for speaker recognition. The general flow of the scheme is explained in Figure 5.6. We train speaker models for specific locations $j = 1, 2, \ldots M$ in the room. Each location corresponds to a specific seat around the table or a presentation position as shown in Figure 5.5. In a typical meeting room there are a few such positions which makes this training feasible. Only one of the microphones from the microphone array is chosen to train the speaker models. Though it makes intuitive sense to use the microphone closest to the speaker location, in our testbed, the circular microphone array is located at the center of the table and there is no clear advantage in choosing one microphone over the other. A separate GMM λ_{ij} is trained for each speaker i for each location j in the meeting space. Even though a single microphone is used for training the models, we do need the entire microphone array during runtime to locate the current speaker as described in Section 5.5.1. Based on the speaker's location, the appropriate model is chosen for the speaker recognition task which provides us with the speaker ID.

At time t, if the speaker location is estimated to be location $j(t)$, then the models $\lambda_{ij(t)}$ are used to estimate the likelihood of the MFCC vectors. This location specific speaker modeling (LSSM) technique provides an accuracy of 92% on our meeting dataset as shown in Table 5.2.

A further improvement in accuracy of speaker recognition is possible by joint detection based on multiple observations corresponding to the same location. Specifically if $Y(t) = \{y_1, y_2 \ldots y_T\}$ correspond to the MFCC vectors for the audio frames around time t which correspond to the same location $j(t)$, then $logP(Y(t)|\lambda_{ij(t)}) = \sum_{k=1}^{T} logP(y_k|\lambda_{ij(t)})$ and the speaker identity $id(t)$ is the ML estimate $id(t) = \arg\max_i logP(Y|\lambda_{ij(t)})$. This corresponds to a maximum likeli-

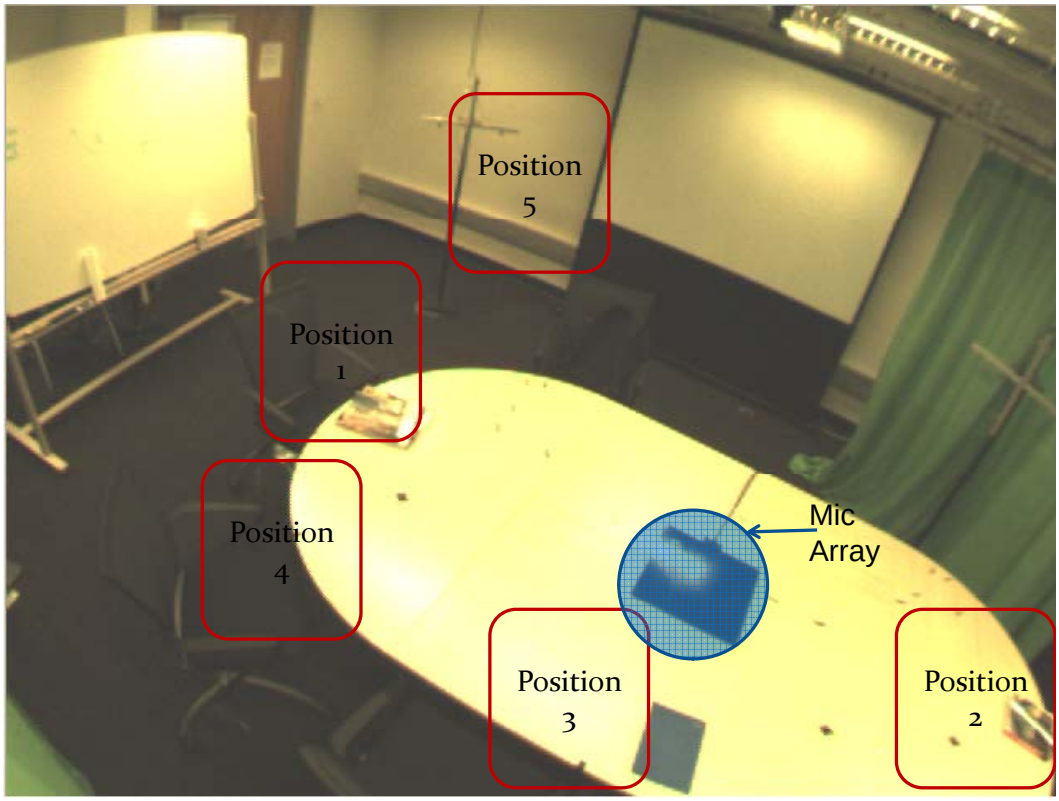

Figure 5.5: The Meeting room and the sensor configurations showing the possible speaker locations around the table.

hood based decision rule at the conversation side level as opposed to the frame level which further improves the performance of the speaker recognition system. The improvement is based on the length of the conversation segments.

The LSSM technique would be impractical in a real world setting as it requires the training of the speaker model for each possible location around the meeting table. However, in the next chapter we develop a semi-supervised speaker model training technique which can automatically update and train the models starting from audio-visual meeting recordings.

In the following sections of this chapter, yet another hierarchical fusion framework is explored that utilizes the video from the cameras to estimate the head pose of the meeting participants and use this contextual information to select the appropriate beamformer taps for reconstruction of speech from far-field microphone

Table 5.2: Comparative performance of the LSSM, CMS and beamforming techniques for matched and mismatched speaker locations

Approach	Matched	MisMatched
Baseline	**94%**	**62%**
CMS	87%	75%
Beamforming	78%	78%
LSSM	**92%**	**92%**

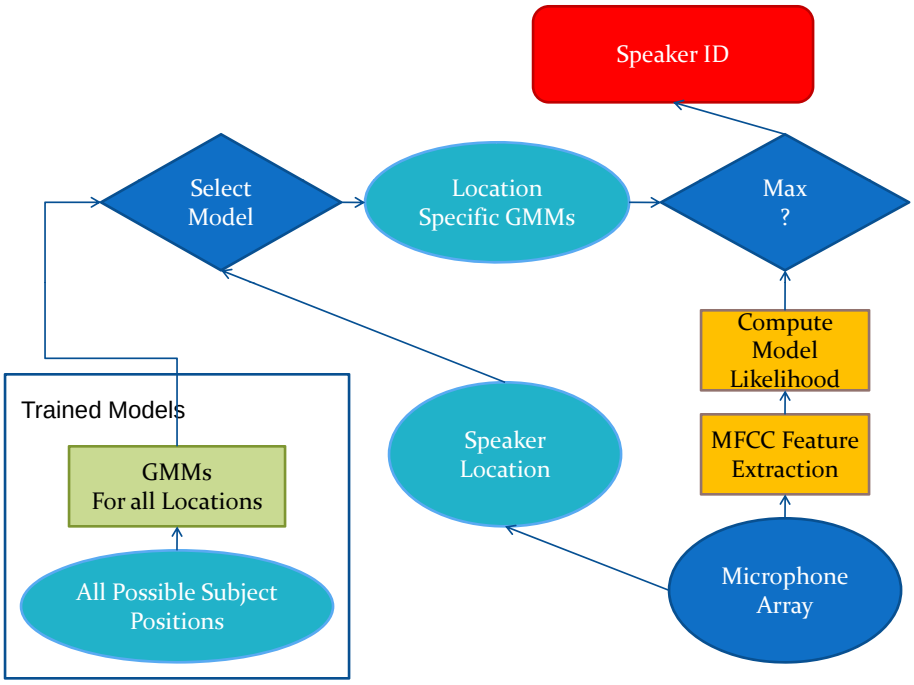

Figure 5.6: Flowchart summarizing the Location specific speaker modeling which involves the fusion of the speaker location cue to select the appropriate model for speaker recognition.

for robust speech and speaker recognition.

5.7 Role of head pose in speech and speaker recognition systems

Speech acquisition from distant microphones in a reverberant environment is a challenging task [41][86]. The signal at the distant microphone is distorted due to echoes and techniques based on SNR measurements cannot be employed to select the best set of microphones. Existing approaches to acquiring clean speech from distant microphones include microphone array based beam-forming techniques. In such systems, recent research has focussed on augmenting the microphone array based system with information from video cameras which are used to track the speakers and provide accurate location information. The sensitivity of the speech acquisition systems to location errors has been studied in [57]. In the next few sections we explore the sensitivity of distant speech acquisition systems to the orientation of the speaker's head in addition to speaker location. Speech recognition accuracy can be significantly improved by using the correct beamforming parameters for the particular location and the orientation of the speaker's head. The orientation of the speaker's head can be estimated using both the audio and video modalities. An audio-visual head pose estimation system is presented in [16]. A detailed survey of video-only head pose estimation can be found in [66]. In our system we use the head orientation estimates and location estimates from the video modality, to improve the quality of speech enhancement by the microphone array. We adopt a delay, filter and sum strategy and report the improvement in the speech recognition accuracy on a large vocabulary speaker dependent speech recognition task.

5.7.1 Room Acoustic Transfer Function

Let microphone i be at location (x_i, y_i, z_i) and the head of the speaker be centered at (x_s, y_s, z_s) and oriented in the direction (ϕ_s, θ_s) in the polar co-ordinates relative to the original co-ordinates. The location and directivity of the microphones is assumed to fixed and we do not model changes in those parameters. Let us assume that the source signal $s(t)$, measured using a close talking microphone,

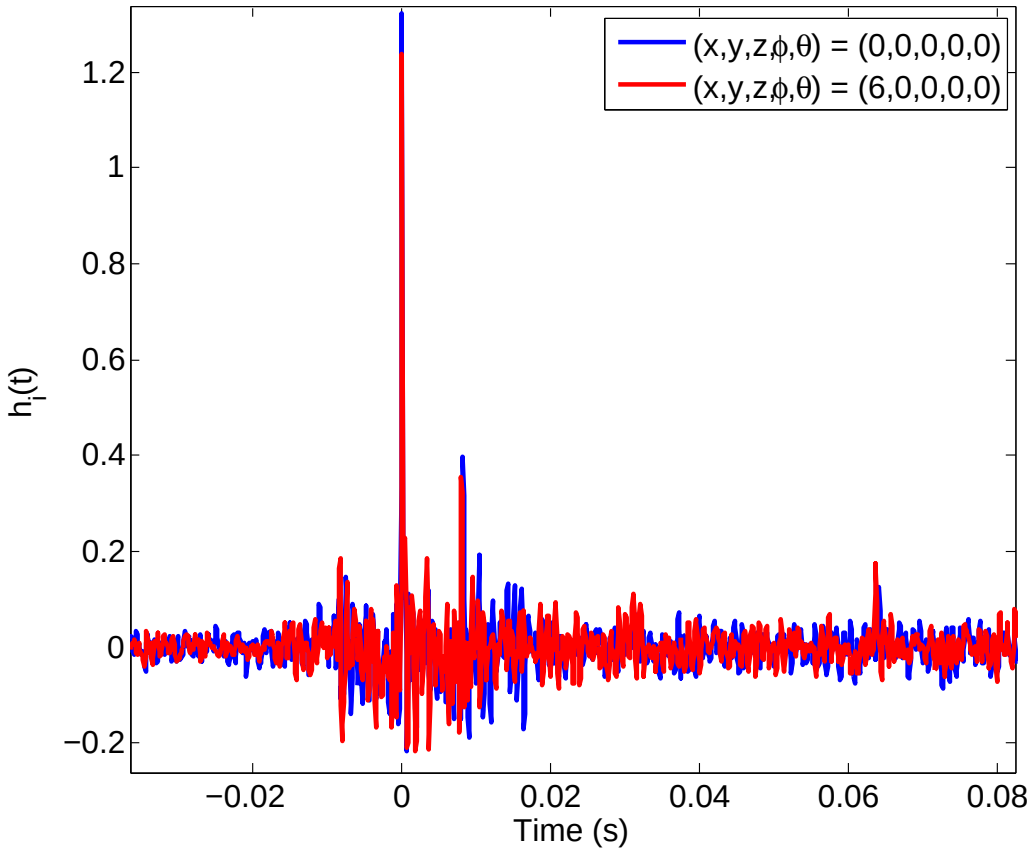

Figure 5.7: Room acoustic impulse response, for two source locations 6" apart. The impulse response is estimated by assuming that $s(t)$, measured using a close talking microphone, is the input and $h_i(t)$, the signal received at microphone i, is the output of the channel. The same measurements are then repeated for another location of the speaker, 6" away from the first.

encounters a channel whose impulse response is h_i. The transfer function corresponding to this channel will be referred to as the room acoustic transfer function. If we represent the signal received at microphone i by $y_i(t)$, then, $y_i(t) = s(t) * h_i$. In Figure 5.7, we see an example of $h_i(t)$ for two different source locations.

We claim, h_i depends on $x_i, y_i, z_i, x_s, y_s, z_s, \phi_s$ and θ_s. Since the microphone is assumed to be fixed, we could reduce the dependence to x_s, y_s, z_s, ϕ_s and θ_s. It is easy to see why this is indeed the case. The location of the speaker relative to the microphone and the room determines the relative delay and amplitude of the

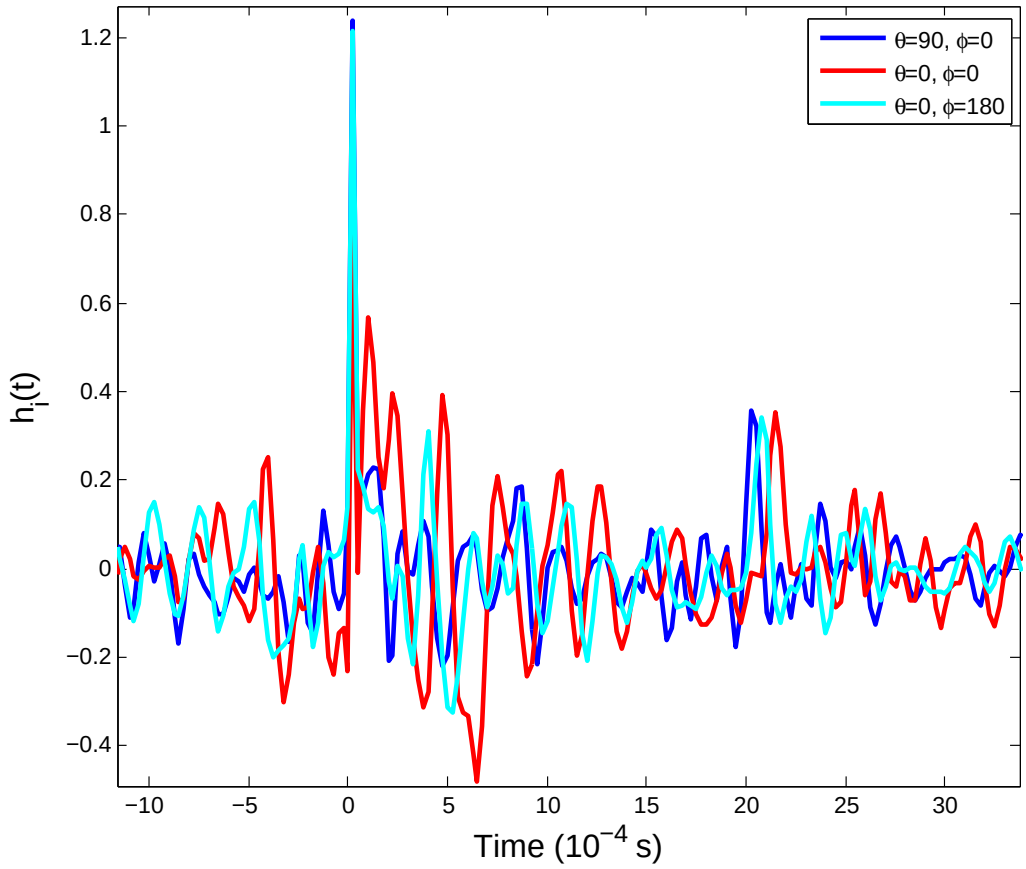

Figure 5.8: Room acoustic impulse response, for same location but three different speaker head orientations, estimated as in Figure 5.7. Note that the impulses responses are very different, indicating the sensitivity to head pose.

reflections from the walls and other surfaces in the room, contributing to the tail of the impulse function. The human vocal tract acts as a directed source. This is especially true for frequencies greater than 4kHz. In [16], the head radiation pattern is discussed in detail. From the directional nature of head radiation pattern one can deduce the dependence of the room acoustic transfer function on ϕ_s. In Figure 5.8, the dependence of h_i on the orientation of the speaker's head confirms this deduction. In the next Section we present a framework to utilize the head pose information for effective speech acquisition from distant microphones.

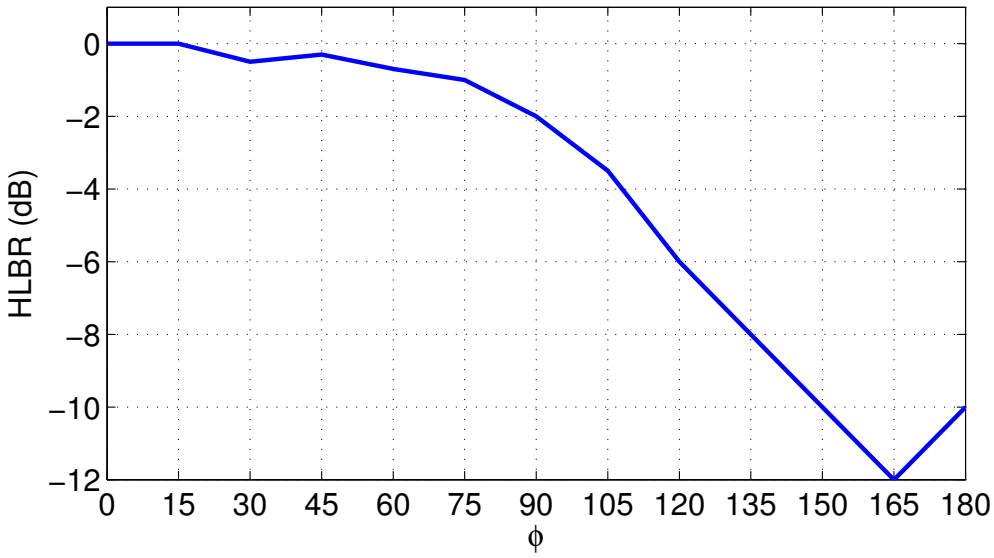

Figure 5.9: The ratio of energy in the high($>$ 4kHz) and low frequency bands($<$ 200Hz) vs the angle ϕ around the speaker's head, as shown in Figure 5.11. We can see that the human vocal tract is highly directional for frequencies above 4kHz, with little or no attenuation in the front of the mouth, but experiencing a 10dB attenuation at the rear of the head.

5.8 Speech acquisition from distant microphones

Current research on speech acquisition using distant microphones implicitly or explicitly model speaker location. Speaker location is implicitly used in aligning the signals from different microphones with one another [13]. In more advanced schemes, in addition to the proper alignment of the signals using the appropriate delay, location specific parameters are used in beamforming [57]. However, to date, no research has included the orientation of the speaker's head in the beamforming techniques. This is mainly due to the difficulty of estimating the orientation of the head. Using video, however, we can estimate the head pose of the speaker[66] and use this information in acquiring clean speech from distant microphones. This can be done in one of the following ways,

- Use specific microphone array beamformer coefficients for the current location and orientation of the speaker.

- Use a subset of the microphones for acquiring the speech by selecting those microphones that have a strong direct path from the speaker.
- Use the best microphone for the present speaker location and orientation.

Note that the later options are specific instances of the earlier ones. However, they are also easier to implement in practise. Thus there is a trade-off between generality and convenience. In Section 5.9, we present results that provide practical insights to this trade-off. The other issue that is addressed in Section 5.9 is that of the sensitivity of automatic speech recognition to the orientation of the speaker's head in each of the three situations considered above. This allows us to implement a practical system by training beamformers for particular orientations of the speaker's head. In more specific instances, such as meeting rooms, the participants tend to face each other while speaking and this would allow the training of beamformers for these particular cases. These cases are also explored in Section 5.9. Also note that energy/SNR based selection of the 'best' microphones does not convey the same information as a microphone that has a dominant direct path and register clearer signals from the speaker.

5.9 Computational Framework and Algorithms

In Figure 5.10, we present the framework of our proposed scheme. The configuration of the sensors and the layout of the room are shown in 5.11

5.9.1 Audio-visual person tracking

The localization of speaker is based on our earlier work. We refer the reader to [88] for details. The audio localization includes the time difference of arrival (TDOA)estimation as a first step and these TDOA estimates are used in the beamformer for aligning the signals from different microphones.

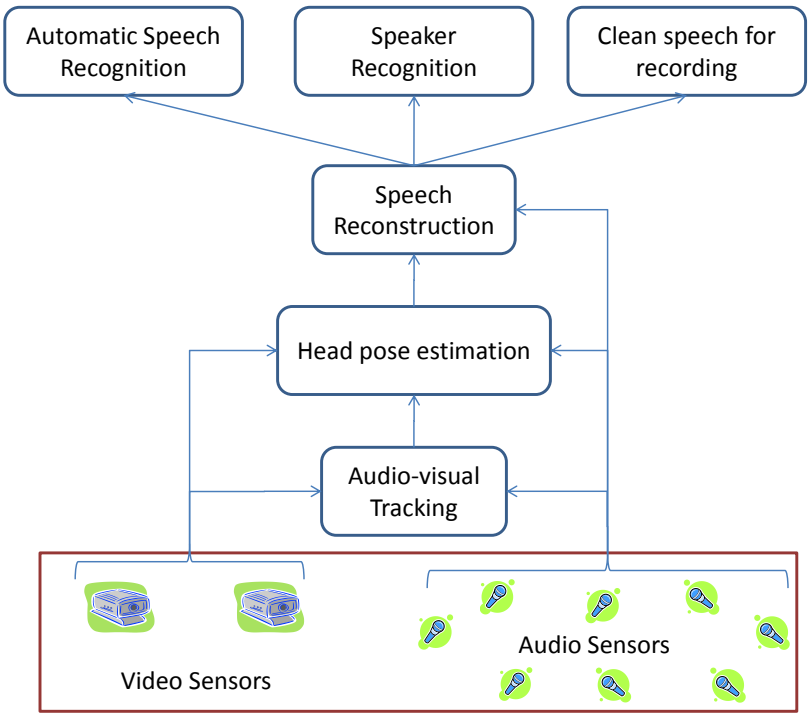

Figure 5.10: The overall system flowchart that uses the head pose information to select the appropriate beamformer taps.

5.9.2 Audio-visual head pose estimation

In section 5.1, we discussed some aspects of audio-visual head pose estimation. Our video head pose estimation algorithm using calibrated video cameras is based on the algorithm discussed in [65]. The audio head pose estimation is not incorporated in the present system, but could be a future addition.

5.9.3 Filter and sum beamformer

In our experiments we use a filter and sum beamformer to reconstruct the speech signal from the distant microphones . The signal $s_i(t)$ from the ith microphone is delayed by the appropriate delay T_i to align all the microphones with one another. During the training phase, they are aligned with a reference microphone $s_r(t)$ that is placed close to the speaker and the filter taps are trained by a stochas-

tic gradient descent algorithm. Note that by explicitly constraining a subset of the filters to have all zero taps, we can select a subset of the microphones. And in the extreme case, select only one of the microphones. These cases correspond to the three options mentioned in Section 5.8

5.9.4 Automatic Speech Recognition

A commercially available speech recognition software, the dragon naturally speaking system is used for recognizing the acquired speech signal. The recognition system is adapted for each speaker separately, using a close talking microphone. This is the same microphone used as the reference microphone in training the beamformer taps (Section 5.9.3). The results correspond to a person dependent large vocabulary continuous speech recognition task based on the standard dictation mode of the speech recognizer.

5.10 Experimental Evaluation

In this Section we describe the experimental setup in the Smartspace lab at UCSD. We present the details of the system used to evaluate the theory presented above. The results presented in Section 5.10.1 are from this setup. Figure 5.11 shows the layout of the room in which the audio-visual system is deployed. There are 2 rectilinear cameras and 8 omnidirectional microphones deployed in the room as shown in Figure 5.11. The cameras and microphones are calibrated with respect to the room co-ordinates. The setup is close to a typical meeting with 4 participants and a presenter. Thus each participant has 4 foci of attention, corresponding to the 4 other participants in the meeting. For each orientation of the speaker, corresponding to the speaker facing one of these foci, we present the following results.

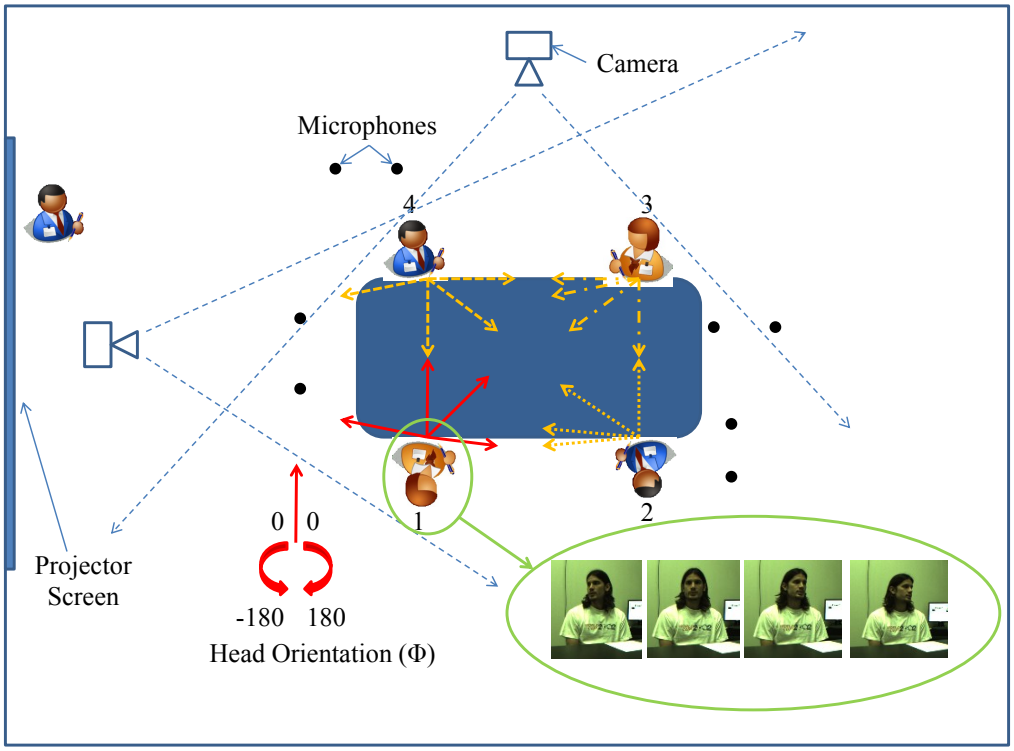

Figure 5.11: Layout of the audio-visual testbed at the Smartspace lab at UCSD.

5.10.1 Results

- Case A: A filter and sum beamformer is trained using all 8 microphones for that particular orientation and speaker location.
- Case B: A filter and sum beamformer is trained using a subset of microphones "in front of" the speaker, for that orientation.
- Case C: The single best microphone, based on speech recognition accuracy, is selected and the signal is directly used for speech recognition.

The baseline results to compare the performance of our scheme are as follows.

- Case D: Close talking microphone.

- Case E: A filter and sum beamformer is trained using all 8 microphones, with the training data including all possible orientations at the given speaker location (orientation agnostic).

- Case F: A filter and sum beamformer is trained using all 8 microphones for a "forward" orientation at the given speaker location.

The results are presented in Table 6.2. From these results, it is clear that by training the beamformer for particular head orientations in any of the three cases A, B, C, one can achieve an improvement over cases E and F.

Table 5.3: Comparisons of speech recognition accuracies for the beamformers described above. Note that the first three cases require the estimation of the head pose of the speaker, the last two cases represent the best one can do in the absence of such information.

Location	Case A	Case B	Case C	**Case D**	Case E	Case F
	With	head	pose	**Baseline**	No head	pose
1	85%	87%	85%	**90%**	77%	78%
2	84%	85%	84%	**91%**	78%	77%
3	81%	82%	81%	**85%**	72%	71%
4	85%	87%	85%	**90%**	77%	78%

In Figure 5.12, we present the results of head orientation mismatch on the speech recognition accuracy. The baseline for comparison is the accuracy of the close-talking microphone. Then there is the beamformer trained for the correct orientation of the speaker along with the beamformer trained for the nominal orientation (angle zero) and used for other orientations of the head. From this we can conclude that using the right head orientation in selecting the beamformer improves the speech recognition accuracy by 10% in some cases.

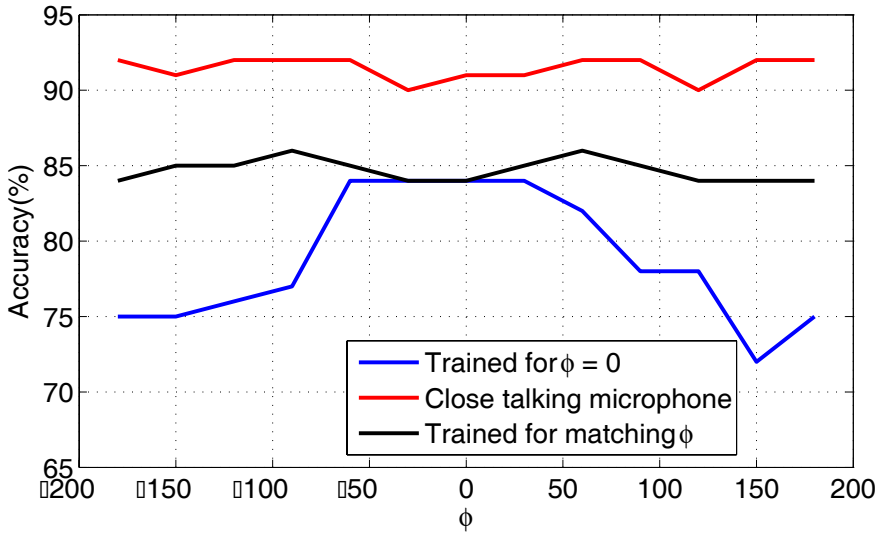

Figure 5.12: Sensitivity of the speech recognition task to head orientation mismatch.

5.11 Concluding remarks

We have presented an audio-visual system to effectively acquire speech signals from far-field microphones in a meeting room scenario and demonstrated the improvement in speech recognition accuracy obtained by training beamformers for particular head pose of the speaker. In the more general problems, where the speakers are not constrained to occupy certain locations and face particular directions as in a meeting room, there are open issues that have to be addressed regarding the practicality of storing and using different beamformers for different positions and speaker head orientations. Future work could explore reducing the constraints in our system and demonstrating the improvement in speech quality, speech recognition and speaker recognition tasks in a general setting.

5.12 Acknowledgments

The text of Chapter 5, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Role of Head Pose Estimation in Speech Acquisition From Distant Microphones," IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, Taiwan, Apr-2009. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 5, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Hierarchical audio visual information fusion in intelligent meeting rooms," Manuscript submitted to IEEE Transactions on Multimedia for review. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Crossmodal learning in hierarchical audio-visual fusion schemes

6.1 Introduction

In the previous chapter, hierarchical fusion frameworks were presented, to extract information in a robust manner from audio and visual cues. Training the contextual models requires extensive amounts of densely labeled training data. A framework to train the contextual models using minimum amount of supervision will make the hierarchical fusion frameworks more applicable in practice. This area of audio-visual information fusion has not been explored in detailed. In the next section, we provide some background and literature review of supervised and unsupervised learning schemes in human activity analysis systems. In the rest of this chapter we develop a framework for learning the location specific speaker models in a semi supervised manner using face recognition information.

6.2 Background

Significant progress has been made in designing systems that fuse audio and visual information to achieve better accuracy and robustness to background noise. A very commonly seen paradigm in fusion schemes is the appropriate weighting of input streams according to their reliability. The assessment of the quality of individual streams is a challenging task in itself and needs further research. Common measures like SNR are useful but there is a necessity for other measures to quantify the reliability of extracted cues from audio and video streams. As an example, consider a speaker whose lips are sometimes partially or fully occluded from the camera due to his changing orientation on an audio-visual speech recognition system. How to enable the system to weight the audio and visual cues appropriately in this case?

The question can be reposed as, how to build a system that can adapt to changing situations? This leads us to the bigger problem of learning. Most of the systems described in this survey learn model parameters in a supervised fashion. This requires a lot of annotated training data and also places the restriction that the conditions during the deployment of the system cannot be significantly different from the training conditions. This highlights the utility of semi-supervised and unsupervised methods for learning parameters. Semi-supervised learning allows one to learn model parameters from a small amount of annotated training data and large amounts of non-annotated training data. Unsupervised learning has also been used in ambient intelligence systems to classify previously unseen activities [14]. Existing research has addressed the problem of learning at several levels. Multimodality has an advantage in unsupervised learning through the presence of cross-modal correspondences.

Recent work in cognitive sciences has led to the design of systems that can learn primitive correspondences across different modalities in a manner similar to the learning experiences of a human child. More specifically, systems that ground language in perceptual cues have been proposed. From the previous sections we can conclude that there are a very large number of strategies for fusion of information in human activity analysis systems. Humans are extremely competent at such

tasks and seem to employ an near-optimal fusion strategy for each situation. However, most approaches to automatically recognize multimodal actions are based on having a annotated training set[107]. To quote the authors,

> ... However, no matter based on feature or semantic fusion, most systems do not have learning ability in the sense that developers need to encode knowledge into some symbolic representations or probabilistic models during the training phase. Once the systems are trained, they are not able to automatically gain additional knowledge even though they are situated in physical environments and can obtain multisensory information. ...

Yu and Ballard[108] present a unified framework to learn perceptually grounded meanings of spoken words without transcriptions. This is the first step towards building a system that can learn for perceptual cues without the necessity to encode the knowledge in some symbolic representation. The perceptually grounded words provide the symbolic representation[107]. The opposite process where visually-guided attention helps in understanding a complex auditory scenes has also been studied in literature [10].

Modeling schemes influence the fusion strategy used and the modeling schemes are themselves are heavily task oriented, as seen in the preference of the speech recognition community in using HMMs and the tracking community in using particle filters. An intelligent system will have to simultaneously perform these tasks in order to perform tasks like a human. A framework to fuse the different systems would have to be developed. Learning such a framework by starting with a certain amount of pre-programmed intelligence, but streamlining the models and adding extra functionalities both by supervised learning and observing multimodal data for cross-modal correspondences in a manner similar to the development of human cognition is a challenge towards which the research community is making advances towards.

Finally, the entire meeting analysis system as a whole as well as the results from extensive evaluation are presented in Section 6.4.

6.3 Face recognition for learning location specific speaker models

In this section we will present a framework for learning the location specific speaker models in a semi supervised manner using face recognition information.

6.3.1 Face recognition using eigenfaces

Person identification using face recognition techniques have been extensively researched. However, in a meeting setting, face recognition is much more computationally expensive compared to speaker recognition. Hence face recognition is not preferable to speaker recognition for realtime meeting analysis systems. Moreover, face recognition in unconstrained scenarios has challenges of its own such as illumination changes, varying head pose as well as background clutter [109][102]. However, face recognition can be still be useful in validating training data for training location specific speaker models. Since the focus of this chapter is to explore the significance of audio-visual fusion, we select a simple face recognition scheme. However, for a practical system, one still needs the face recognition system to work with far-field cameras. For this purpose, we use a face detection and tracking system based on the OpenCV implementation of the Viola and Jones' face detector. The face detection step is followed by a principal component analysis (PCA) based Eigen-face model for extracting the most relevant features for face recognition [44]. We use a k-nearest neighbor classifier in the PCA space to classify the incoming face as being from one of the existing faces in our dataset. On our meeting dataset, this approach provides an average of 59% accuracy on a 15 person training set on a per-frame basis. However, by taking majority decisions on segments of multiple frames, we can significantly improve the accuracy of the face recognition system. Table 6.1 shows the variation of face recognition accuracy for groupings of different number of frames. We can infer from the table that if we can base our decision on 23 frames or more, we can achieve more than 95% accuracy.

Table 6.1: Performance of the Eigenface based face recognition system on a 15 subject dataset.

Number of frames averaged	1	3	11	23
Accuracy	59%	67%	81%	95%

6.3.2 Semi-supervised learning scheme for location specific speaker models

If we assume that our face recognition models are accurate, one can learn the location specific speaker models from meeting recordings in an unsupervised manner with the following algorithm. At each time t for which an audio frame is available, let $j(t)$ be the location of the active speaker as determined in Section 5.5.1. For the entire length of the meeting recording, collect the audio frames that correspond to active speaker location j. These frames are added to the training set for training models for location j. In order to label these frames with the correct speaker ID, we use the face recognition system. The video frames corresponding to the best view for location j are passed to the Viola and Jones face detector and the detected faces are used in the face recognition system. A joint decision is made on the complete set of detected faces, ensuring a high degree of confidence in the recognition result. The output of the face recognition system is the label for the audio frames. After the new frames are added, the location specific speaker models are re-estimated using the EM algorithm. The overall algorithm is summarized in Figure 6.1.

Thus the supervision required in the training of the models is limited to adding new face images for new subjects in the dataset. Once this is completed, the system can continually update and refine the location specific speaker models. Note that when a particular location model is not available for a particular speaker, the first run of the meeting analysis system will be error prone.

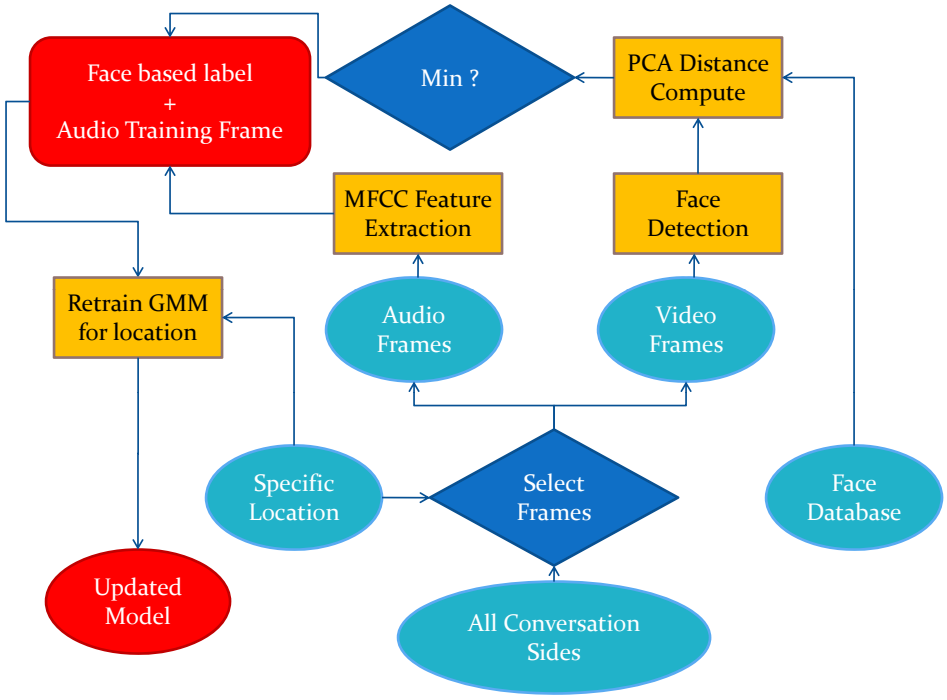

Figure 6.1: Flowchart summarizing the fusion of face recognition results for labeling the audio frames resulting in a semi-supervised approach for the LSSM framework.

6.4 Implementation of the meeting analysis system

In this section we describe the overall setup of the meeting analysis system. The sensors consist of the six element circular microphone array and four rectilinear cameras as described in Section A.4.1. The audio is sampled at 44.1kHz and analyzed in frames that are 25 milliseconds in length and have an overlap of 15 milliseconds. The video frames are captured at 15fps. For the sake of clarity, the results of the meeting analysis are presented at a resolution of one second. Note that this choice of time scale is for convenience and clarity of presentation and can be modified without affecting the results as long as the audio sampling rate and

video frame rates are consistent.

The apriori information that is necessary for the meeting analysis system is the set of all possible seating positions around the table, the best view camera and bounding box corresponding to each of these positions and the expected time difference of arrival (TDOA) vectors for each location. Also, we provide a labeled face dataset on which the face recognizer is built using a k-nearest neighbor classifier in the eigenface space. Also, one of the microphones in the array is selected to construct the speaker recognition models.

The first step in the meeting analysis is to capture the audio and video frames in a synchronous manner. The technique involved in synchronizing the audio and video streams is beyond the scope of our work. Also, the methods for estimating the best view camera, the bounding box and estimated TDOA vector for the different locations are standard techniques used in computer vision [102] and microphone array analysis [90] and we do not analyze them in detail in the current chapter.

- For each video frame, average number of pixels that are different from the background pixels with in the bounding boxes for each location are computed. If this average is greater than a certain threshold, the location is declared as occupied.

- For each audio frame, the TDOA vectors are computed. The nearest video frame is queried for occupied locations. The distances from the estimated TDOA vectors for occupied locations to the current TDOA vector are computed and the location with the minimum distance is declared as the active speaker location.

- The meeting is broken down into conversation sides within which the active speaker location does not change. Each conversation side is of varying length in time.

- The MFCC feature vectors are computed for the audio frames of the microphone selected for the speaker recognition task. For each conversation side

which corresponds to one speaker location by definition, the average likelihood of the MFCC feature vectors under available speaker GMMs for the particular location is computed. The speaker whose GMM that achieves the maximum average likelihood is the speaker for the given conversation side.

- Note that to begin with, there are no speaker models for any locations. In order to train the models, the face recognition results are used to generate a labeled dataset. Again, the conversation sides are considered and the face detector is implemented within the bounding box of the best view camera's frames corresponding to the active speaker location for the conversation side. The detected faces are recognized using the labeled face dataset and a majority decision is taken. In order to achieve a high order of confidence, conversation sides that are shorter than 2 seconds or have less than 24 detected faces are dropped. The MFCC vectors from the selected microphone are labeled with the face recognition output and added to the training set to train the speaker recognition GMMs for the particular location.

- We experimented with the possibility of using the detected faces to improve the face dataset. However, several non-faces are detected as faces by the face detector and hence adding these false positives into face dataset degrades the performance of the face detector. Hence the supervised labeling of the face examples is necessary. However this is considerably simplified compared to the effort involved in collecting and training audio frames for each speaker for each possible location.

- Also, note that the training of the speaker models will automatically adapt the GMMs to any small changes in the speaker seating locations with respect to the original seating locations. One can update the expected TDOA vectors and bounding boxes periodically for the different seating locations.

- The result of the meeting analysis system is the meeting recoding organized in the form of conversation sides each of which has a location and speaker ID associated with it as well as the audio data and the best view video frames associated with it . This is extremely useful to support querying for

Table 6.2: Performance of the meeting analysis system with retraining from the face recognition.

Meeting clip	Cumulative number of minutes analyzed	Number of speaker models	Number of new speakers	accuracy before training	Accuracy after training
1	4	3	3	-	92%
2	8	3	0	91%	90%
3	13	3	0	91%	92%
4	17	5	2	35%	89%
5	20	5	0	87%	89%
6	24	6	1	61%	86%
7	28	6	0	89%	89%
8	33	8	2	32%	86%
9	38	8	0	86%	87%

specific speakers, locations, interactions as well as other criterion. Also, if implemented in an online manner, this system will be very valuable to an intelligent teleconferencing system.

6.5 Evaluation results

We evaluated the meeting analysis and the semi supervised speaker model learning scheme on a sequence of meeting recordings. Our face dataset consisted of 15 subjects. The meeting recordings were held in a natural setting with three to 4 subjects and typically consisted of 4 minute long clips. In Table 6.2 we present the overall results of the meeting analysis and training process. The accuracy is computed by comparing the active speaker recognition results with the actual active speaker on a per second basis.

Note that whenever a new speaker comes into the meeting scene, the recog-

Table 6.3: Performance of the Eigenface based face recognition system on a 15 subject dataset.

Meeting clip	2	3	5	7	9
Number of speaker models	3	3	5	6	8
Accuracy	90%	92%	89%	89%	87%

nition accuracy declines. However, retraining the system using the face recognition results in improved performance. In a steady state, meeting analysis system will perform with the accuracies denoted in Table 6.3.

Finally, we apply our meeting analysis system to the set of recorded meetings and organize the conversation sides based on speaker location and speaker ID. Such an analysis allows to browse through meetings in an intelligent manner, allowing us to query for specific speakers. Also, the audio and video clips for each conversation side can be retrieved upon query in an efficient manner. In Figure 6.2 we illustrate one such meeting clip which has been tagged by our meeting analysis system. Note that the conversation sides that were identified earlier based on speaker location have now been tagged with speaker ID as well.

6.6 Concluding remarks

In the current and the last chapter we have presented an analysis of the hierarchical fusion of audio visual cues for building practical intelligent systems. We have analyzed the different fusion paradigms which are apart from the traditional fusion strategies. We have described how cues can be fused to reduce the search space and also to enable contextual modeling, both resulting in an increased performance and simplified modeling. We have also described a semi-supervised learning scheme that uses cross-modal interaction to generate labeled training datasets, reducing the effort involved in training the contextual models. The current chapter illustrates these principles in the context of a meeting analysis system.

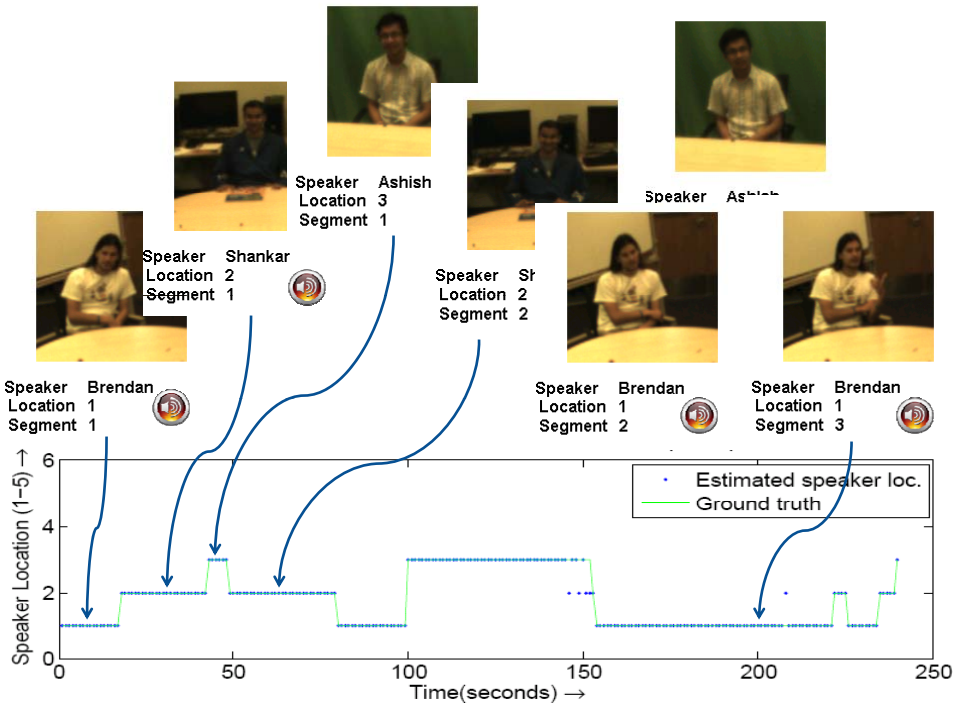

Figure 6.2: The result of analysis of a meeting recording allows us to organize the meeting based on the location of the speaker and speaker identity. This facilitates intelligent archival and browsing of meeting recordings. Note that the audio and video clips corresponding to each conversation side is indexed with the speaker location and speaker ID.

6.7 Acknowledgments

The text of Chapter 6, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Hierarchical audio visual information fusion in intelligent meeting rooms," Manuscript submitted to IEEE Transactions on Multimedia for review. The dissertation author was the primary investigator and author of this paper.

The text of Chapter 6, in part, is a reprint of the material as it appears in: Shankar T. Shivappa, Mohan M. Trivedi, and Bhaskar D. Rao, "Audio-visual Information Fusion In Human Computer Interfaces and Intelligent Environments:

A survey", Proceedings of the IEEE, October 2010. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Concluding remarks and future directions

Audio-visual information fusion is an important area in the design and implementation of intelligent systems that need to analyze and interpret human activity. In this thesis, several key aspects of audio-visual information fusion has been explored in the context of several tasks such as audio-visual speech recognition, person tracking and meeting scene analysis. The benefits of fusion are most evident in the robust and efficient performance of hierarchical fusion schemes. In the area of semi-supervised and unsupervised learning using crossmodal correspondences, we have barely scratched the surface and there is a huge potential for further research.

In practice there are several situations where such a hierarchical fusion model can be very effective. Natural human computer interaction can benefit significantly from a hierarchical fusion framework. Applications in the areas of smart health homes, assisted living technologies and intelligent automobiles are all relevant in this context. The proliferation of smart phones and mobile computing platforms has opened up increasing application areas for audio-visual fusion. The audio and visual information captured on a phone are usually noisy and will need specific fusion framework to achieve the robustness required for most of the applications. The specific interaction between the different audio and visual cues is domain specific and needs to be researched in the individual domains. However,

the fusion paradigms outlined in this thesis will be applicable in these application areas and enable the development of intelligent audio-visual systems.

Appendix A

Audio-visual testbeds

A.1 Review of existing audio-visual meeting corpora, testbeds and evaluations

Audio-visual analysis of human activity in meeting rooms for meeting scene understanding, segmentation, archival and retrieval has received a lot of attention in the recent past. Systematic comparison of the different fusion approaches in meeting scene analysis is extremely challenging due to varying scenarios considered by different groups. Moreover, many of the systems described above (ASR, biometrics, tracking, emotion detection etc.) are used as subsystems of the meeting analysis system.

One of the early research studies in observing human activities in an instrumented room is described in [101]. A graphical summary of the human activity is generated. The audio and visual information is used in identifying the current speaker based on a rule based decision fusion. [11] and [39] describe another meeting room analysis system which also fuses audio-visual stream for person identification, in addition to using the audio for automatic transcription and archival purposes. [79] investigates speech, gaze and gesture cues for high level segmentation of a discourse into topical segments based on a psycholinguistic model.

More recent work in [59] models the action of the group of individuals in a meeting instead of individual actions. HMMs are used to statistically model the

state of the group using audio and video features and the interactions between individuals are inherently accounted for in the model. Using this formulation, meetings are segmented into five categories: Discussions, Monologues, Note-Taking, Presentations and White-Board presentations. Different fusion schemes were evaluated and the early integration strategy performed the best followed closely by the asynchronous HMM. The feature concatenation scheme could suffer from the curse of dimensionality. Intuitively, there is a certain amount of asynchrony between the audio and visual streams in a meeting scene and this hints at the possible inadequacy of using simple HMMs to model the meeting scenes. [113] describes a two layered HMM model to segment the meeting at the individual and group levels respectively. In this case, the asynchronous HMM performs best at the lower level as expected. Dynamic Bayesian networks were explored for suitability in modeling meetings in [26]. An comparison of various modeling techniques is provided in [2].

A number of multimodal meeting rooms equipped with multimodal sensors have been established by various research groups and consortiums. Annotated audio-visual corpora have been collected and standard evaluations have been organized to compare existing frameworks on specific tasks. Table A.1 lists the details of a few important meeting corpora. Another recent effort in collecting and organizing multimodal corpora is presented in [53].

Recent evaluations of meeting scene analysis systems include the CLEAR 2006 evaluation [93] and CLEAR 2007 evaluation [94].

A.2 UCSD-CVRR audio-visual testbed 1

In this section we describe an experimental testbed that is set up at the Computer Vision and Robotics Research(CVRR) lab at University of California, San Diego. The goal of this exercise is to develop and evaluate human activity analysis algorithms in a meeting room scenario. Figure A.1 shows a detailed view of the sensors deployed.

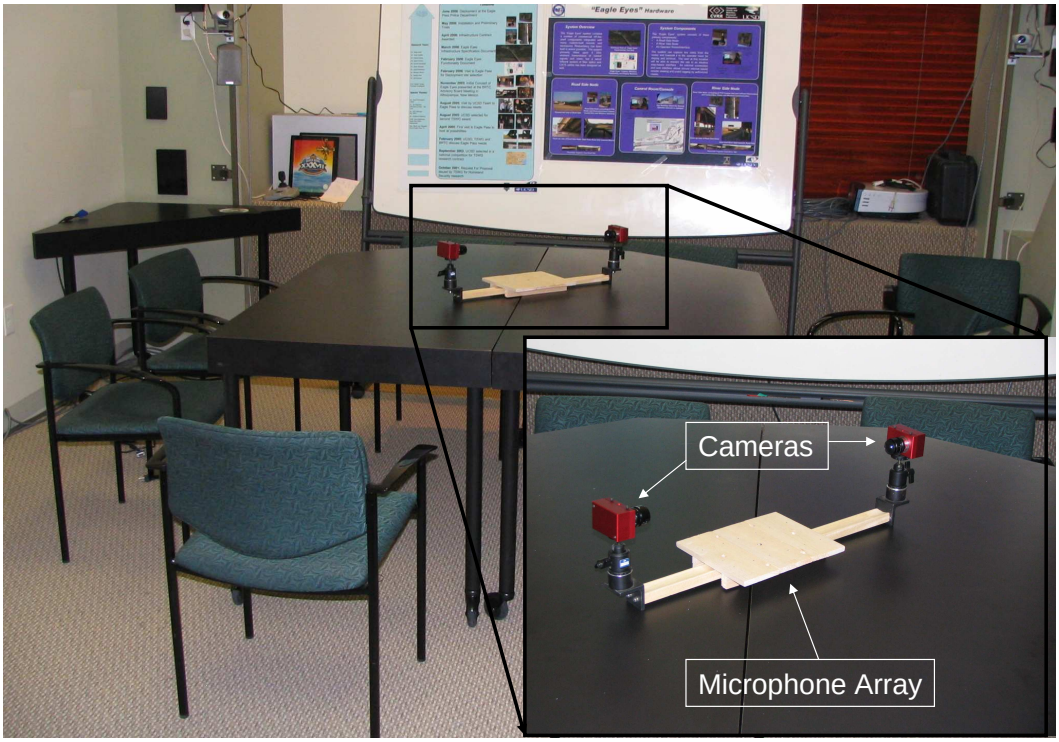

Figure A.1: Testbed and the associated audio and video sensors

A.2.1 Audio sensors

The sensors consist of a microphone array. The audio signals are captured at 16kHz on a Linux workstation using the Advanced Linux Sound Architecture(ALSA) drivers. JACK is a useful audio server that is used here to capture and process multiple channels of audio data in realtime(as required).

A.2.2 Video sensors

We use a synchronized pair of wide angle cameras to capture the majority of the panorama around the table. The cameras are placed off the center of the table in order to increase their field of view as shown in the enlarged portion of figure A.1.

A.2.3 Synchronization

In order to facilitate synchronization, the video capture module generates a short audio pulse after capturing every frame. One of the channels in the microphone array is used to record this audio sequence and synchronize the audio and video frames.

A.3 UCSD-CALIT2 audio-visual testbed 2

A.3.1 Test bed details

In this section we present the details of our laboratory testbed with multiple cameras and microphone arrays. The testbed is located in the Smartspaces lab at CALIT2 in the University of California, San Diego. The testbed is equipped with 24 microphones and 4 cameras. The layout is shown in Figure 6. The cameras have significantly overlapping field of view and different perspectives. The cameras have a resolution of 640x480 pixels and capture frames, synchronously, with each other and the microphones, at 7.5 fps. The audio signal is sampled at 44.1kHz. There are four microphone arrays with 4 microphones each, arranged in the form of a cross with dimensions 40cm x 40 cm. In addition there is a circular array with 6 microphones in the center of the table and two microphones at the end of the table. The cameras have a overlapping field of view with different perspectives.

A.3.2 Ground truth estimation

In order to obtain the ground truth, we use standard chessboard pattern based camera calibration techniques to calibrate the cameras with respect to the world co-ordinates. The microphones are manually located in the camera view and their location is estimated by triangulation. A sound source with a bright source of light is moved around the monitored space. By triangulation, the position of the light is accurately determined at each frame. The positions of the microphones are then optimized to match the TDOA values obtained at each frame with those computed from the sound source co-ordinates. This calibration allows us to obtain

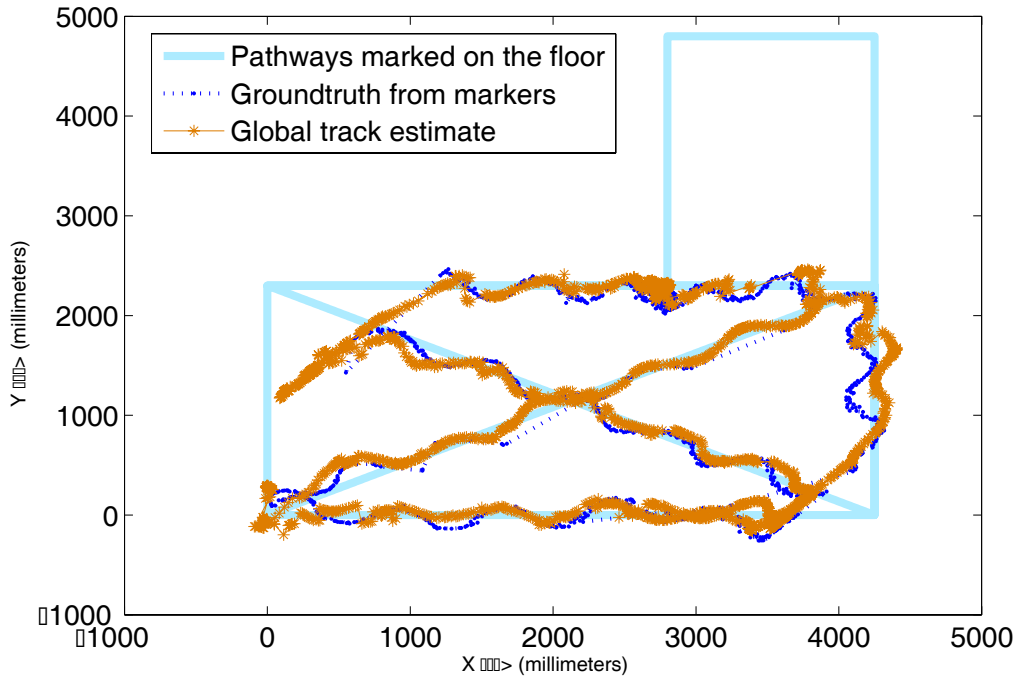

Figure A.2: A track and its associated ground truth in world co-ordinates.

visual-marker based ground-truth estimates for comparison of our results. The location estimate from the triangulation procedure was compared with the actual location measurement. The standard deviation of the error was 2.4 cm on a test set that involved 100 different spots distributed in the room .

A.3.3 Datasets

Meetings among the lab members were recorded for the evaluation of the MID-AVT framework. The meetings consist of 4 to 6 subjects. There are clips where the subjects are either involved in a discussion or one person is giving a presentation. In collecting this dataset (MID-AVT-UCSD-1), we have tried to keep the sensor configurations comparable to the CHIL meeting rooms [64] which were used in the CLEAR 2006 and CLEAR 2007 evaluation workshops. During presentations and meetings, there is not much movement among the participants and usually only one speaker is active at a particular time, which is true for a

majority of the time in many meetings. The individual segments range from 5 minute to 15 minutes in duration. Some meeting segments were annotated by manually marking the position of the subjects' head once every second for a total for a total of 1200 seconds. This corresponds to 9000 frames and these frames were used in our evaluation.

In addition we also have a separate dataset (MID-AVT-UCSD-2) of scenes involving 1-4 subjects that involves a lot more movement of the subjects. This dataset involves multiple subjects who are involved in a continuous conversation with mostly one active speaker at any time, moving around in the room. This dataset has significant number of occlusions and tracks converge and diverge frequently. This dataset has shorter clips ranging from 1 to 5 minutes and the evaluation is presented on a total of 3000 frames which involve about 30 occlusions which were manually detected and marked for evaluation. There are only two cameras and a total of 8 microphones in this dataset.

A.4 UCSD-CALIT2 audio-visual testbed 3

A.4.1 Scene and sensor configuration

In the current section we focus our attention on an intelligent meeting space and explore the hierarchical fusion in the context of meeting analysis system. Our results are based on the analysis of real-world meetings collected in our audio-visual testbed in the Smart spaces lab at CALIT2, UCSD. In this section we describe the physical set up of this meeting room and the sensors deployed in it.

The meeting room testbed is 23' long and 13' wide. A six element circular microphone array is located at the center of the table. There are four rectilinear cameras that cover the meeting scene from different vantage points such that every location has a best viewing camera associated with it.

The meeting analysis system described in its entirety in Section 6.4 is trained and tested on a set of real world meeting scenes collected in the above described testbed.

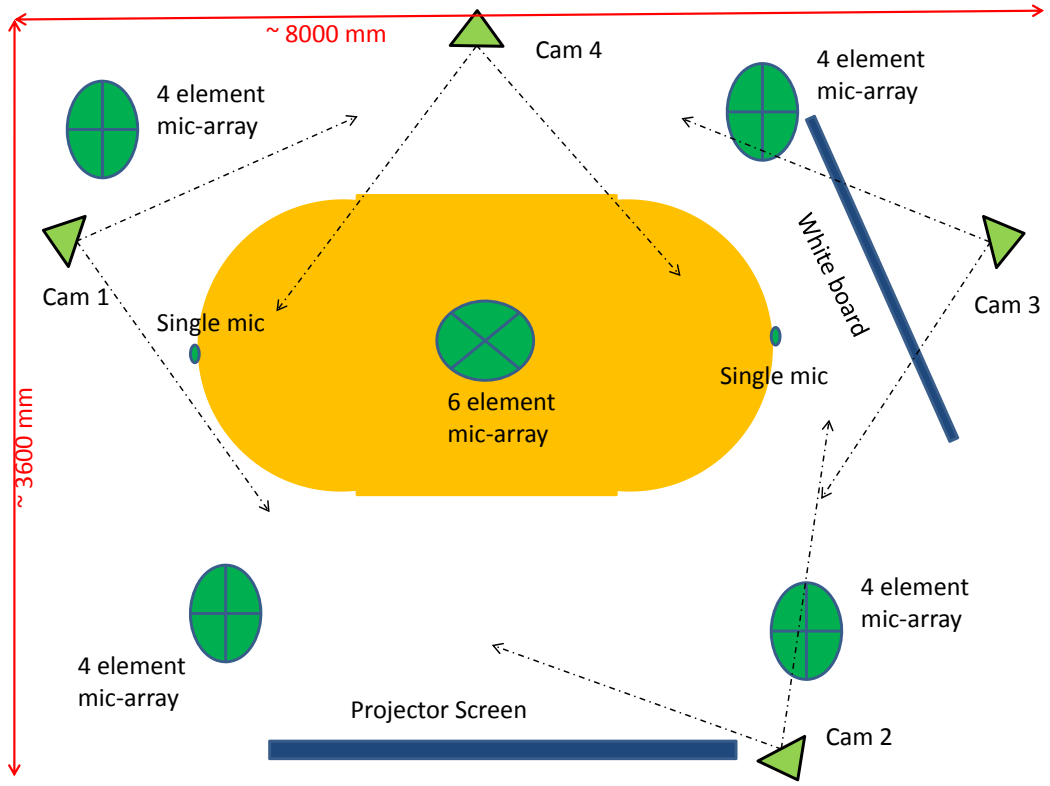

Figure A.3: The configuration of the meeting room for data set 1. The 4 cameras and 24 microphones are shown with their approximate fields of view. The dimensions of the room are approximately 360 cm x 800 cm.

Table A.1: Standard audio-visual meeting scene corpora and their sensor, scene and participant information.

ISL [15] : The Interactive System Labs of CMU, Pittsburgh has collected a database consisting of more than 100 diverse meetings, combined total of 103 hours (4.3 days). Each meeting lasted an average of 60 minutes. The meetings have an average of 6.4 participants. The meetings have been collected since 1999. A meeting in the database is a minimum of three individuals speaking to one another. The results are presented in a maximum of eight mono audio files in WAV format, so-called speaker and recording protocol files containing information about the participants, equipment, environment and scenario, three video tapes, one transcription file of the entire meeting, so-called marker file containing begin and end time stamps for conversation contributions, and a list of the meetings vocabulary. The meeting scenarios include ProjectWork Planning, Military Block Parties, Games, Chatting, and Topic Discussion.

ICSI [61] : International Computer Science Institute, Berkeley, California has collected a 75-meeting corpus with audio and transcripts of natural meetings recorded simultaneously with head-worn and tabletop microphones. The corpus contains 75 meetings of 4 main types and 53 unique speakers. The data totals to over 70 meeting-hours and up to 16 channels for each meeting. The ICSI effort is predominantly an audio scene analysis and meeting transcription effort.

NIST [34] : NIST has constructed a Meeting Data Collection Laboratory (MDCL) to collect corpora to support meeting domain research, development and evaluation. The NIST Smart Data Flow architecture, developed by the NIST Smart Spaces Laboratory, streams and captures all of the sensor data from 200 mics and 5 video cameras on 9 separate data collection systems in a proprietary time-indexed SMD format.

Table A.2: Standard audio-visual meeting scene corpora and their sensor, scene and participant information. (contd.)

The NIST architecture also ensures that all data streams are synchronized (via the Network Time Protocol and NIST atomic clock signal) to within a few milliseconds. The NIST Meeting Room Pilot Corpus consists of 19 meetings/15 hours recorded between 2001 and 2003. In total, the multi-sensor data comes to 266 hours of audio and 77 hours of video.

CHIL [64] : The CHIL (Computers in the Human Interaction Loop) consortium is an European reserach effort with the participation of 15 partner sites from nine countries under the joint coordination of the Fraunhofer - IITB and the Interactive Systems Labs (ISL) of the University of Karlsruhe, Germany. Five smart rooms have been set up as part of the CHIL project, and have been utilized in the data collection efforts. Two types of interaction scenarios constitute the focus of the CHIL corpus: lectures and meetings. The CHIL corpus is accompanied by rich manual annotations of both its audio and visual modalities. In particular, it contains a detailed multi-channel verbatim orthographic transcription of the audio modality that includes speaker turns and identities, acoustic condition information, and name entities for part of the corpus. Furthermore, video labels provide multi-person head location in the 3D space, as well as information about the 2D face bounding box and facial feature locations visible in all camera views. In addition, head-pose information is provided for part of the corpus. Each smart room contains a minimum of 88 microphones that capture both close-talking and far-field acoustic data. There exists at least one 64-channel linear microphone array, namely the Mark III array developed by NIST. The video data is captured by five fixed cameras. Four of them are mounted close to the corners of the room, by the ceiling, with significantly overlapping and wide-angle fields-of-view.

Table A.3: Standard audio-visual meeting scene corpora and their sensor, scene and participant information. (contd.)

VACE [18] : Under this research effort, Air Force Institute of Technology (AFIT) modified a lecture room to collect multimodal, time-synchronized audio, video, and motion data. In the middle of the room, up to 8 participants can sit around a rectangular conference table. 10 camcorders and 9 Vicon MCam2 near-IR cameras, driven by the Vicon V8i Data Station record the video data. For audio, the participants wear Countryman ISOMAX Earset wireless microphones to record their individual sound tracks. Table-mounted wired microphones are used to record the audio of all participants (two to six XLR-3M connector microphones configured for the number of participants and scenario, including two cardioid Shure MX412 D/C microphones and several types of low-profile boundary microphones (two hemispherical polar pattern Crown PZM-6D, one omni-directional Audio Technica AT841a, and one four-channel cardioid Audio Technica AT854R). For the VACE meeting corpus, each participant is recorded with a stereo calibrated camera pair. The Vicon system is used to obtain more accurate tracking results to inform subsequent coding efforts, while also providing ground truth for video-tracking algorithms.

AMI & AMIDA [80] : The AMI and AMIDA projects are EU projects concerned with the recognition and interpretation of multiparty meetings.Three standardized meeting rooms were constructed at IDIAP, TNO and University of Edinburgh. Each room consisted of at least 6 cameras and 12 microphones. The different recording streams are synchronized to a common timeline. The corpus consists of 100 hour annotated corpus of meetings, with speech annotations aligned to the word level. Also, manual annotations of the behavior of the meeting participants are provided at various levels namely dialogue acts, topic segmentation, extractive and abstractive summaries, named entity, gaze direction etc.

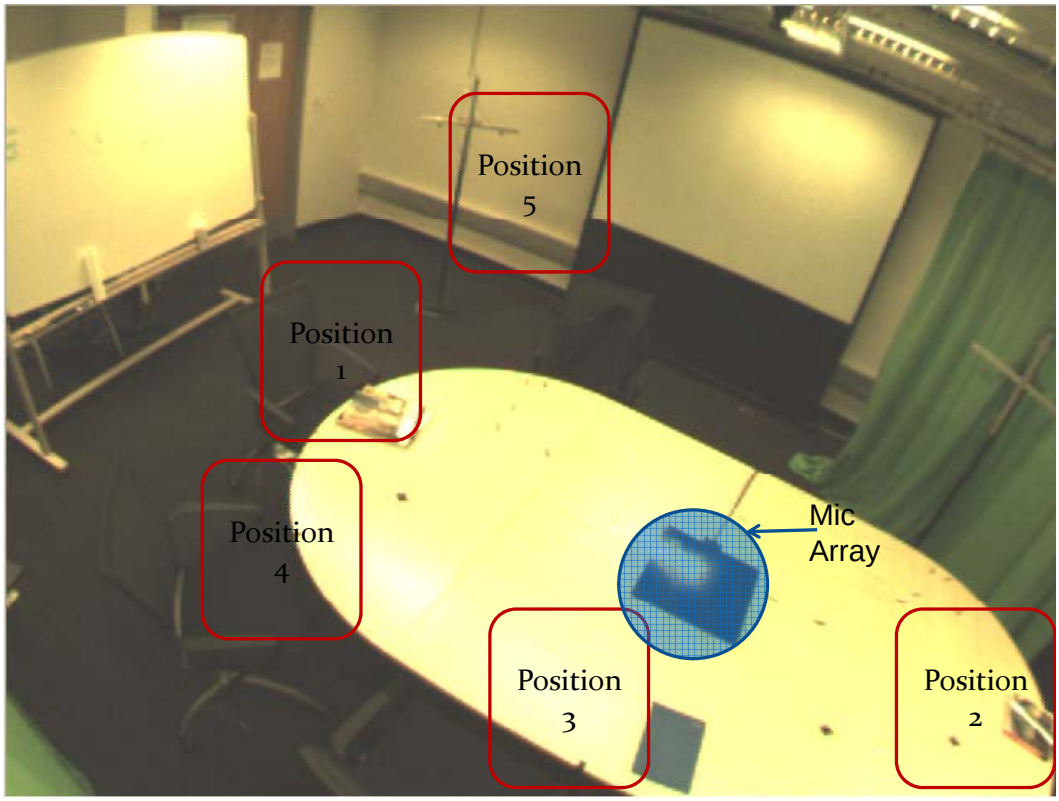

Figure A.4: The Meeting room and the sensor configurations showing the possible speaker locations around the table.

Bibliography

[1] A. Abad, C. Canton-Ferrer, C. Segura, J. L. Landabaso, D. Macho, J. R. Casas, J. Hernando, M. Pardas, and C. Nadeu, "Upc audio, video and multimodal person tracking systems in the clear evaluation campaign," *Proceedings of the First International CLEAR Evaluation Workshop - Multimodal Technologies for Perception of Humans*, 2007.

[2] M. Al-Hames, C. Lenz, S. Reiter, J. Schenk, F. Wallhoff, and G. Rigoll, "Robust multi-modal group action recognition in meetings from disturbed videos with the asynchronous hidden markov model," *IEEE International Conference on Image Processing*, 2007.

[3] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, Mar. 1974.

[4] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. Academic Press, 1988.

[5] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003.

[6] K. Bernardin, T. Gehrig, and R. Stiefelhagen, "Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking," in *CLEAR Evaluation Workshop*, 2007.

[7] K. Bernardin, R. Stiefelhagen, A. Pnevmatikakis, O. Lanz, A. Brutti, J. R. Casas, and G. Potamianos, *Person Tracking*, ser. Human-Computer Interaction Series. Springer London, 2009.

[8] K. Bernardin, R. Stiefelhagen, and A. Waibel, "Probabilistic integration of sparse audio-visual cues for identity tracking," in *Proceeding of the 16th ACM international conference on Multimedia*, 2008.

[9] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding:turbo-codes," in *Proceedings of the IEEE International Conference on Communications*, May 1993.

[10] V. Best, E. J. Ozmeral, and B. G. Shinn-Cunningham, "Visually-guided attention enhances target identification in a complex auditory scene," *Journal of the Association for Research in Otolaryngology*, 2007.

[11] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel, "Multimodal meeting tracker," in *in Proceedings of RIAO2000*, 2000.

[12] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," 1997.

[13] M. Brandstein and D. Ward, *Microphone Arrays*. Springer, 2001.

[14] O. Brdiczka, J. Maisonnasse, P. Reignier, and J. Crowley, "Detecting small group activities from multimodal observations," *Applied Intelligence*, 2007.

[15] S. Burger, V. MacLaren, and H. Yu, "The isl meeting corpus: the impact of meeting type on speech style," in *ICSLP*, 2002.

[16] C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pardàs, and J. Hernando, "Audiovisual head orientation estimation with particle filtering in multisensor scenarios," *EURASIP J. Adv. Signal Process*, vol. 2008.

[17] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[18] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, T. X. Han, J. Tu, Z. Huang, M. Harper, Y. Xiong, D. Mcneill, R. Tuttle, and T. Huang, "Vace multimodal meeting corpus," in *in Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005.

[19] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, Jan 2001.

[20] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proceedings of the IEEE*, 2004.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, Nov 2006.

[22] R. Cutler and L. S. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo (III)*, 2000.

[23] P. Dai and G. Xu, "Context-aware computing for assistive meeting system," in *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, 2008.

[24] S. Dasgupta and Y. Freund, "Random projection trees for vector quantization," *IEEE Transactions on Information Theory*, 2009.

[25] J. H. DiBiase, H. F. Silverman, and M. S. Branstein, "Robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[26] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic bayesian networks," *IEEE Transactions on Multimedia*, Jan. 2007.

[27] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, Sept. 2000.

[28] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Transactions on Multimedia*, Oct. 2005.

[29] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, and H. Abut, "Multimodal person recognition for human-vehicle interaction," *IEEE Multimedia Magazine*, 2006.

[30] E. Ettinger and Y. Freund, "Coordinate-free calibration of an acoustically driven camera pointing system," in *International Conference on Distributed Smart Cameras (ICDSC)*, 2008.

[31] J. W. Fisher, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *NIPS*, 2000.

[32] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A multimodal corpus recorded in a health smart home," in *LREC 2010 workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*.

[33] A. Garg, G. Potamianos, C. Neti, and T. S. Huang, "Frame-dependent multistream reliability indicators for audio-visual speech recognition," in *Proceedings of the International Conference on Multimedia and Expo*, 2003.

[34] J. Garofolo, C. Laprum, M. Michel, V. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus," in *in Proc. of Language Resource and Evaluation Conference.*, 2004.

[35] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech and Language Processing*, 2007.

[36] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, Oct. 2005.

[37] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proceedings of ICSLP*, 1996.

[38] K. W. Grant and S. Greenberg, "Speech intelligibility derived from asynchronous processing of auditory-visual information," 2001.

[39] R. Gross, M. Bett, H. Yue, X. J. Zhu, Y. Pan, J. Yang, and A. Waibel, "Towards a multimodal meeting record," 2000.

[40] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream HMMs in Audio-Visual Speech Recognition," in *10th International Conference on Multimodal Interfaces*, 2008.

[41] T. Gustafsson, B. D. Rao, and M. M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, Nov. 2003.

[42] J. Hershey and J. Movellan, "Audio vision: Using audiovisual synchrony to locate sounds," in *NIPS*, 2000.

[43] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, "Integration of multimodal features for video scene classification based on hmm," in *Proceedings of IEEE Workshop on Multimedia Signal Processing*, 1999.

[44] K. S. Huang and M. M. Trivedi, "Robust real-time detection, tracking, and pose estimation of faces in video streams," in *Proceedings of International Conference on Pattern Recognition*, Jun, 2004.

[45] ——, "Video arrays for real-time tracking of person, head, and face in an intelligent room," *Machine Vision and Applications*, vol. 14, no. 2, pp. 103-111, Jun. 2003.

[46] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," in *Proceedings of the IEEE International Workshop on Human Computer Interaction in conjunction with ICCV*, Oct. 2005.

[47] ——, "Multimodal human-computer interaction: A survey," *Comput. Vis. Image Underst.*, 2007.

[48] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[49] B. Kane, S. Luz, and J. Su, "Capturing multimodal interaction at medical meetings in a hospital setting: Opportunities and challenges," in *LREC 2010 workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality.*

[50] B. Kapralos, M. R. M. Jenkin, and E. Milios, "Audiovisual localization of multiple speakers in a video teleconferencing setting," 2003.

[51] N. Katsarakis, F. Talantzis, A. Pnevmatikakis, and L. Polymenakos, "The ait 3d audio / visual person tracker for clear 2007," *Proceedings of the First International CLEAR Evaluation Workshop - Multimodal Technologies for Perception of Humans*, 2007.

[52] Y. Keller, S. Lafon, R. Coifman, and S. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Transactions on Signal Processing*, 2009.

[53] M. Kipp, J. C. Martin, P. Paggio, and D. Heylen, *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications.* Lecture Notes on Aritificial Intelligence, Springer, 2009.

[54] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustic, Speech and Signal Processing*, 1976.

[55] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Journal of Computer Vision and Image Understanding - Special Issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum*, Dec. 2006.

[56] M. Liu, H. Tang, H. Ning, and T. S. Huang, "Person identification based on multichannel and multimodality fusion," in *CLEAR*, 2006.

[57] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. on Audio, Speech, and Language Processing*, Nov 2007.

[58] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Feb 2001.

[59] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Mar. 2005.

[60] I. Mccowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proceedings of 2001 A Speaker Odyssey: The Speaker Recognition Workshop*, 2001.

[61] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and

C. Wooters, "Meetings about meetings: research at icsi on speech in multiparty conversations," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, April 2003.

[62] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust asr," *Speech Communication*, 2001.

[63] B. Morris and M. M. Trivedi, "An adaptive scene description for activity analysis in surveillance video," in *IEEE International Conference on Pattern Recognition*, 2008.

[64] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, J. R. Casas, J. Turmo, L. Cristoferetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Journal on Language Resources and Evaluation*, Dec 2007.

[65] E. Murphy-Chutorian and M. M. Trivedi, "3d tracking and dynamic analysis of human head movements and attentional targets," *Second ACM/IEEE International Conference on Distributed Smart Cameras*, Sept. 2008.

[66] ——, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[67] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," in *IJCAI*, 2001.

[68] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[69] C. Neti, G. otamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop," in *Proceedings of IEEE Workshop Multimedia Signal Processing*, 2001.

[70] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop," in *Proceedings of IEEE Workshop Multimedia Signal Processing*, 2001.

[71] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proceedings of the 7th international conference on multimodal interfaces*, 2005.

[72] A. ODonovan and R. Duraiswami, "Microphone arrays as generalized cameras for integrated audio visual processing," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[73] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Proceedings of International Conference on Multimodal Interfaces*, Oct. 2002.

[74] N. M. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[75] S. G. Z. P. Aarabi, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, 2001.

[76] D. G. Perez, G. Lathoud, I. McCowan, J. M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," 2003.

[77] G. Pingali, G. Tunali, and I. Carlbom, "Audio-visual tracking for natural interactivity," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, 1999.

[78] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, Sept. 2003.

[79] F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K. E. McCullough, N. Furuyama, and R. Ansari, "Gesture, speech, and gaze cues for discourse segmentation," *IEEE conference on Computer Vision and Pattern Recognition*, 2000.

[80] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The ami and amida projects," in *HSCMA*, April 2008.

[81] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, 1995.

[82] V. Rozgic, K. J. Han, P. G. Georgiou, and S. Narayanan, "Multimodal speaker segmentation in presence of overlapped speech segments," in *Proceedings of Tenth IEEE International Symposium on Multimedia*, 2008.

[83] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audio-visual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Signal Processing*, 2007.

[84] J. L. Schwartz, J. Robert-Ribes, and P. Escudier, "Ten years after summerfield: A taxonomy of models for audio-visual fusion in speech perception," *Hearing by Eye II*, 1998.

[85] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "An iterative decoding algorithm for fusion of multi-modal information," EURASIP *Journal on Advances in Signal Processing - Special Issue on Human-Activity Analysis in Multimedia Data*, 2008.

[86] ——, "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[87] ——, "Role of head pose estimation in speech acquisition from distant microphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[88] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Person tracking with audio-visual cues using the iterative decoding framework," in *5th IEEE International Conference On Advanced Video and Signal Based Surveillance*, 2008 [Best Paper Award].

[89] ——, "Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms," in *IEEE CVPR Workshop: ViSU'09*, 2009.

[90] S. T. Shivapppa, B. D. Rao, and M. M. Trivedi, "Audio visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *IEEE Journal of Special Topics in Signal Processing, Special Issue*, 2010.

[91] S. T. Shivapppa, M. M. Trivedi, and B. D. Rao, "Audio-visual information fusion in human computer interfaces and intelligent environments: A survey," *Proceedings of IEEE*, 2010.

[92] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos, "A decision fusion system across time and classifiers for audio-visual person identification," in *CLEAR*, 2006.

[93] R. Stiefelhagen, K. Bernardin, R. Bowers, J. S. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *CLEAR*.

[94] R. Stiefelhagen, R. Bowers, and J. Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007 (Lecture Notes in Computer Science)*.

[95] R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling multimodal human-robot interaction for the karlsruhe humanoid robot," *IEEE Transactions on Robotics*, Oct. 2007.

[96] K. Takeda, H. Erdogan, J. H. L. Hansen, and H. Abut, *In-Vehicle Corpus and Signal Processing for Driver Behavior (Springer USA 2009).*

[97] A. Tawari and M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on Multimedia, Special Issue on Multimodal Affective Interfaces*, vol. 12, issue 6, October 2010. pp. 502-509.

[98] A. Tawari and M. M. Trivedi, "Contextual framework for speech based emotion recognition in driver assistance system," in *IEEE Intelligent Vehicles Symposium*, 2010.

[99] P. Teissier, J. Robert-Ribes, J. L. Schwartz, and A. Guerin-Dugue, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech and Audio Processing*, Nov 1999.

[100] C. Tran and M. M. Trivedi, "Towards a vision-based system exploring 3d driver posture dynamics for driver assistance: Issues and possibilities," in *IEEE Intelligent Vehicles Symposium*, 2010.

[101] M. M. Trivedi, K. S. Huang, and I. Mikic, "Activity monitoring and summarization for an intelligent meeting room," in *Proceedings of the IEEE International Workshop on Human Motion*, 2000.

[102] ——, "Dynamic context capture and distributed video arrays for intelligent spaces," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 35, Jan. 2005.

[103] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr 1990.

[104] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," *Eighth IEEE International Conference on Computer Vision.*

[105] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[106] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, 2006.

[107] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Trans. Appl. Percept.*, vol. 1, no. 1, 2004.

[108] ——, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomput.*, 2007.

[109] Z. Yu and Y. Nakamura, "Smart meeting systems: A survey of state-of-the-art and open issues," *ACM Computing Surveys*, Feb. Volume 42 , Issue 2 , 2010.

[110] W. Zajdel, J. Krijnders, T. Andringa, and D. Gavrila, "Cassandra: audio-video sensor fusion for aggression detection," in *4th IEEE International Conference On Advanced Video and Signal Based Surveillance*, 2007.

[111] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audiovisual affective expression recognition through multistream fused hmm," *IEEE Transactions on Multimedia*, June 2008.

[112] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, "Audio-visual affect recognition in activation-evaluation space," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, July 2005.

[113] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modeling individual and group actions in meetings: A two-layer hmm framework," *IEEE International Conference on Computer Vision and Pattern Recognition*, June 2004.

[114] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, 2002.