Active Speaker Localisation and Tracking using Audio and Video

A dissertation submitted to the University of Dublin
for the degree of Doctor of Philosophy

Damien Kelly
Trinity College Dublin, March 2010

Signal Processing and Media Applications
Department of Electronic and Electrical Engineering
Trinity College Dublin

To my family,
-YNWA.

Abstract

This thesis is concerned with the problem of tracking active speakers using audio and video data. Particular focus is placed on the task of tracking the current active speaker in a lecture room environment using multiple cameras and multiple microphones. A database of lecture recordings corresponding to this scenario from the European Integrated Project, Computers in the Human Interaction Loop (CHIL) is used to support work presented throughout the thesis.

Within the lecture room environment, the problem of extracting reliable audio and video based features for detecting people is explored. In the audio domain, the use of time-delay estimates from multiple pairs of microphones is examined for localising active speakers. Fundamental limitations on localisation accuracy using this approach are also discussed. In the video domain, background modelling, face detection and skin colour detection are considered as candidate features for the visual detection of people. A new skin colour model is introduced which models for the non-linear dependence of skin-tone on luminance and is shown be effective in modelling skin colour under low illumination. Following the evaluation of audio and video features for locating people, a review of existing techniques which aim to fuse audio and video information for tracking is presented.

This thesis makes two critical analyses in relation to the joint audio-video based localisation problem. The first analysis examines the expected accuracy of audio-based localisation using multiple microphones and video-based localisation using multiple cameras. Within this analysis, the theory of estimating localisation uncertainty for both localisation techniques is unified under a common framework. Using this, single modality localisation is compared to localisation through different audio-video based fusion strategies. The different fusion strategies evaluated are the fusion of audio and video location estimates in the positional domain, the audio domain and the video domain. For each of these fusion methods, it is found that little is to be gained in terms of localisation accuracy through a fusion based approach, in comparison to the accuracy of single modality video-based localisation.

The second critical analysis made in this thesis evaluates the localisation performance of a configuration of microphone arrays in a lecture room. Theoretical bounds on the accuracy of time-delay estimates obtained using the microphone arrays are employed in this analysis. A theoretical bound on the accuracy of localising an active speaker is developed to include important aspects influencing localisation performance such as, the signal-to-reverberant ratio and the speaker and microphone directionality characteristics. It is argued that the sub-optimal placement of microphone arrays results in poor localisation performance. The configuration of microphone arrays in the CHIL lecture room is taken as an example case to demonstrate this point. A novel algorithm is proposed which uses simulated annealing to automatically determine the optimal placement of microphone arrays to minimise localisation uncertainty. The effectiveness and practicality of the proposed algorithm is shown for a number of example

scenarios which consider minimising localisation uncertainty over user defined presenter and audience areas.

Finally, in this thesis a new joint audio-video based algorithm is introduced for tracking the current active speaker in a multi-camera and multi-microphone lecture recording. This algorithm differs from traditional approaches to combining audio and video for tracking in that audio and video based location estimates are not fused in a statistical sense. Instead speakers are detected using video data alone and this information is used to guide an audio-based localisation system which tracks the active speaker. The previous analyses examining localisation accuracy and microphone array performance are used to motivate this approach. The algorithm is proposed as part of a complete system entitled Voxel-based Viterbi Active Speaker Tracking (V-VAST) which reduces a multi-view recording of a lecture to a composite single view presentation consisting of a user defined main view and an automatically inserted view of the current active speaker. The algorithm operates offline and is proposed as a post-production tool for creating a suitable lecture video for presentation over the Internet in eLearning applications or for the purpose of archiving. The tracking accuracy and performance of V-VAST in obtaining a suitable view of the current active speaker is demonstrated on the lecture recordings of the CHIL database.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree that the Library may lend or copy this thesis upon request.

Signed,

Damien Kelly

March 26, 2010.

Acknowledgments

Firstly, I would like to express sincere gratitude to my supervisor Prof. Frank Boland for his help, encouragement, patience and guidance throughout the course of my PhD. His encouragement, in particular, led me towards the path of postgraduate research and for this I am truly grateful. An enormous amount of thanks is also due to Prof. Anil Kokaram, for many inspiring discussions, ideas and invaluable advice. Dr. François Pitié and Dr. Dave Corrigan were also extremely generous with their time and help over the last number of years and I thank them hugely for that.

Thanks must also go to both the staff and post-graduate students of the Electronic & Electrical Engineering Department in Trinity College for their help and also for creating such a friendly and lively working atmosphere. I would like to thank, Bernadette, Conor, Robbie and Nora for their help throughout the years.

A special thanks goes to the members of the SIGMEDIA group and other members of the lab, especially; Dr. Naomi Harte, Gavin, Dan, Deirdre, John, Darren, Gary, Ricardo, Cedric, Craig, Claire, the two Andrews, Liam, Kangyu, Mohamed and Steven. To past members of the lab; Angela, Denis, Deepti and Daire, I thank you also. It has been a privilege to work in the company of such helpful, friendly and engaging people.

As a postgraduate research student I have benefited from the support of the Irish Research Council for Science, Engineering and Technology (IRCSET) and also Science Foundation Ireland (SFI). Many thanks to these organisations for their support.

I count myself extremely lucky to have been gifted with such a wonderful, supportive and loving family. I have been looking forward to this opportunity to express enormous thanks for all their help, advice, support and encouragement. I am truly grateful for everything they have done for me.

Contents

List of Acronyms

CC Cross-Correlation

GCC Generalised Cross Correlation

PHAT Phase Transform

DOA Direction Of Arrival

RT_{60} Reverberation Time

SRP Steered Response Power

TDE Time Delay Estimate

SA Simulated Annealing

ML Maximum Likelihood

MAP Maximum *a posteriori*

SNR Signal to Noise Ratio

CHIL Computers in the Human Interaction Loop

ELRA European Language Resource Association

KF Kalman Filter

PF Particle Filter

EKF Extended Kalman Filter

IEKF Iterated Extended Kalman Filter

UKF Unscented Kalman Filter

MSE Mean Square Error

UT Unscented Transform

MAE Mean Absolute Error

CRLB Cramér-Rao Lower Bound

CPE Correlator Performance Estimate

SRR Signal-to-Reverberant Ratio

GCC-PHAT Generalized Cross-Correlation with Phase Transform

SRP-PHAT Steer Response Power - Phase Transform

SA Simulated Annealing

CP Constant Position

CV Constant Velocity

CA Constant Acceleration

HCI Human Computer Interaction

SIS Sequential Importance Sampling

STFT Short Time Fourier Transform

LTI Linear Time Invariant

PCA Principal Component Analysis

V-VAST Voxel-based Viterbi Active Speaker Tracking

GCF Global Coherence Field

GMM Gaussian Mixture Model

List of Figures

List of Tables

1 Introduction

Continuing advancements in technologies for transmitting multimedia over the Internet, opens a gateway for universities to expand their campus beyond its physical limits and to start offering educational content online. Recently, universities have begun to try and avail themselves of this new opportunity, which is changing the landscape of traditional education. New methods of teaching through this medium are evolving to meet the growing demands of students who desire greater flexibility in education. This has seen the emergence of web-based course content focused towards supplementing normal face-to-face lectures or facilitating distance education and on-demand learning. This represents a new avenue in education popularly known as eLearning.

Within eLearning, two predominant categories exist: synchronous and asynchronous. Synchronous eLearning refers to the live presentation of educational content such as the streaming of multimedia to a remote user. The aim is to provide the user with an equivalent learning experience to that of the in-class participant such as, for instance, the opportunity to interact and ask questions. For many universities the ultimate goal in their eLearning ambitions is to enable students to attend lectures remotely using the Internet in a synchronous manner. However, few have the infrastructure to support this and network technologies have still not reached sufficient bandwidth capabilities to communicate all the necessary audio, video and additional multimedia.

In contrast to synchronous eLearning, asynchronous eLearning is concerned with the presentation of educational resources in an offline manner where the material is not viewed live, but on-demand. Recording lectures and providing them online for asynchronous viewing is a more realistic option for universities at present in comparison to the synchronous alternative,

since it is possible now through existing technology. Presently, many universities offer recorded lectures online through partnerships with Apple iTunes U [11] or Google's youtube EDU [67]. In addition to this, some universities even maintain their own video-on-demand servers supplying recorded lectures such as, Princeton University's WebMedia [149], University Cambridge's CamTv [28], Massachusetts Institute of Technology's OpenCourseWare [127] and University California Berkley's WebCast [185]. As an asset to students for learning, the ability to access recordings of lectures is highly valued since it augments the learning experience and can improve student performance [81, 158]. It is not surprising therefore, that the demand for such asynchronous online lecture content is ever increasing [142].

Recording lectures and making them available online requires a significant amount of effort and commitment on a university's behalf. Commercially available systems exist such as, Panopto's CourseCast [145], Sonic Foundary's MediaSuite [173], Echo360's EchoSystem [49] and Tegrity Campus [181] which enable lectures to be captured automatically without the need for any technical expertise. Some universities have even developed their own systems for lecture capture such as the eClass system used by Georgia Institute of Technology [81]. In general, these systems require the presenter to remain within the field of view of a fixed camera restricting their movements to a small area. Student opinions have shown that restricting the presenter's movements can significantly reduce the perceived classroom experience [130]. One commercially available system which places less constraints on a presenter's movements is Autoauditorium [112]. This is a purely vision based system for tracking the presenter over a large area. The system however is specifically designed for tracking the presenter and does not sufficiently address the task of tracking audience interactions. This is a current limitation since interaction from an audience is a significant element of many lecture presentations. The ultimate goal in the production of video lectures for viewing offline, is to convey exactly that of the in-classroom experience. In-class participants have the freedom to visually follow conversational interactions. If this aspect of a lecture is not conveyed in a recording, the offline viewer's learning experience is reduced in comparison to that of the in-class participant. Furthermore, it is the opinion of professional video producers that lecture recordings which visually capture both the audience and the presenter offer a more enjoyable viewing experience than recordings which simply capture the actions of the presenter [150].

In most cases, the only alternative for universities to adequately capture all the necessary elements of a lecture without restricting the presenter's movements, is to employ a manual camera operator. Professional lecture capturing is expensive and for most universities it is an unrealistic option. Usually, an amateur camera operator is responsible for capturing the lecture which can often result in poor camera operation resulting in poor lecture recordings. This is unacceptable, since the quality of the recorded lecture, in particular, how well the presenter is framed, influences how well it is accepted by students [119]. In addition to this, the use of a human camera operator can hinder the learning experience of in-class participants where given the option, some means of automated capture is preferred since it is less distracting [122].

What is needed therefore, is an intelligent unobtrusive system for automatically capturing all aspects of a lecture. By unobtrusive it is meant that the system does not influence how the lecture should be conducted in order for the capture to be successful. The lecture should be able to proceed under normal circumstances and not be affected in any way by the capturing system. Clearly an important component of such a system is the ability to track the position of the current active speaker. This is important because it is normally the case that the focus of communication in a lecture is at the position of the current active speaker.

In recent times, more intelligent automated lecture capturing systems have begun to emerge for automatically capturing lectures such as Microsoft's lecture capture system [27]. This system represents one of the most sophisticated automated lecture capturing systems in use. This particular tracking algorithm relies on data from two cameras and a microphone array. One camera is focused on the presenter and the visual data from this camera is used specifically for the tasks of tracking the presenter and appropriately capturing the presenter's actions. The second camera which the system employs is assigned to the task of visually capturing the audience members. Individual audience members are located when they ask questions by using the audio data captured at the microphone array. Only the audio data is used to locate audience members who are speaking and this information is used to direct the audience camera to the person who is speaking. As will be highlighted in this thesis, accurate and reliable audio-based localisation can be notoriously difficult to achieve in lecture room environments. A system which relies on audio data only to locate speakers is likely to be unreliable. The designers of the Microsoft lecture capturing system acknowledge this and highlight this point as a significant limiting factor in their system's ability to accurately and consistently capture questions from audience members [27].

Recently however, researchers have begun to examine combining audio and video based measurements for tracking. The basic idea in this approach is that both audio and video can be used to complement each other when applied to tracking and improved reliability and accuracy can be achieved. This thesis is concerned with the combined use of audio and video for tracking. More specifically, this thesis is concerned with using both audio and video for tracking the current active speaker in a lecture room environment. The motivation for this work is the development of automated systems for capturing lectures and creating lecture recordings suitable for presentation over the Internet in eLearning applications for the purpose of archiving.

1.1 CHIL Database

The scenario which is considered is that of tracking the current active speaker in a lecture which has been passively recorded using multiple cameras and multiple microphones. A suitable database of lecture recordings made under these conditions called the Computers in the Human Interaction Loop (CHIL) 2005 evaluation database is considered throughout this thesis. These recordings were made during the CHIL project [30], which was an Integrated Project (IP 506909) jointly coordinated by *Universität Karlsruhe* and the *Fraunhofer Institute IITB* under

the European Commission's Sixth Framework Programme. It began in 2004 with the objective to "*create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves*".

The main focus in this project was towards the office and lecture room scenario. Over the course of the project until it ended in 2007 several evaluations were conducted to examine various technologies in a broad range of tasks including audio-based and video-based tracking; joint audio-video based tracking; speech and person recognition; gesture recognition; automated transcription; activity detection and audio-visual Scene Analysis. These evaluations began with the first evaluation campaign in 2004 and each year after until 2007. For each evaluation a specific evaluation package of annotated audio and video recordings complete with specific metrics and ground truth for each concerned task were created. These packages have recently become available through the European Language Resources Association [50].

Relevant to the lecture scenario of interest to this work is the 2005 evaluation package. Within this package is a set of multi-channel audio and multi-camera video recordings of five seminars made at *ISL, Universität Karlsruhe* in November 2004. The multi-channel audio data consists of various different $44.1kHz$ recordings including a 64-channel microphone array; 4 T-shaped microphone arrays of 4 microphones each; 5 single table-top microphones and a close-talking microphone. The video data consists of $15fps$ recordings from 4 cameras positioned in each of the four corners of the seminar room. A sample of the four camera views of a seminar recording and an illustration of the sensor layout within the CHIL lecture room is shown in figure 1.1.

1.2 Thesis Outline

The following outlines the structure of the thesis,

Chapter 2: Joint Audio-visual Active Speaker Tracking

A review of audio and visual features useful for tracking active speakers is presented in chapter 2. In the audio domain, features arising from the capture of a speech source signal by multiple microphones are described. It is examined how these can be used in localising a speech source. The challenges faced in a lecture room environment such as room reverberations and fundamental limitations in achieving accurate and reliable localisation are discussed.

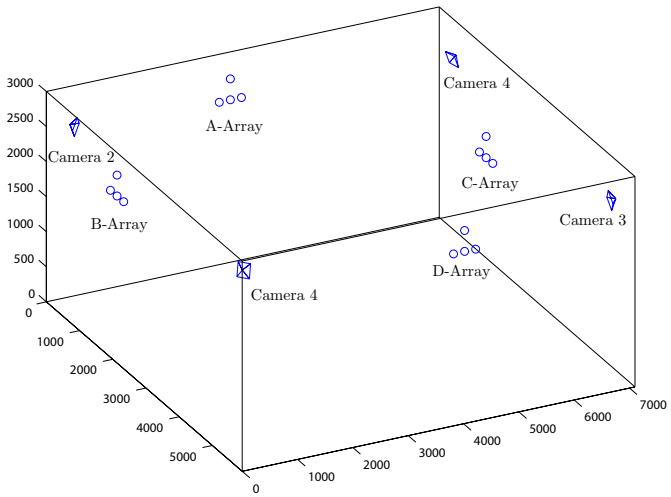
In relation to the video data, some suitable simple features for detecting active speakers in a lecture room environment are presented. In particular, skin colour modelling is examined for use in detecting faces. A new skin colour model is introduced which demonstrates improved skin detection by modelling the non-linear dependence of skin tone on illumination.

(a) *Sample frame from camera 1*

(b) *Sample frame from camera 2*

(c) *Sample frame from camera 3*

(d) *Sample frame from camera 4*



(e) *The setup of the four cameras and four T-shaped microphone arrays of the CHIL room.*

Figure 1.1: *Sample frame from the sequence seminar_2004_11_11_C_segment1 of the CHIL 2005 evaluation package and also an illustration of the layout of the 4 T-shaped microphone arrays and 4 cameras within the room.*

Chapter 3: Audio and Video Features for Active Speaker Localisation

In this chapter, a Bayesian framework for active speaker tracking using audio data is established. Audio-based tracking filters are reviewed in reference to the presented probabilistic framework. Consideration is given to persistent problems in tracking such as motion modelling and the challenge in choosing suitable models for motion. It is shown how the presented tracking framework can be easily extended to include video information. Existing joint audio-video based active speaker tracking techniques are explored within this framework and common strategies for the fusion of audio and video information are described.

Chapter 4: Analysis of Audio-Visual Source Localisation Accuracy

This chapter compares the performance of audio-based localisation using multiple microphones and video-based localisation using multiple cameras in a typical lecture room. Also examined is the accuracy of localisation through the fusion of the estimates from both modalities. In the evaluation, the task examined is that of localising an audio-visual source along a 3D track.

Within this chapter, the theory of uncertainty propagation for estimating the localisation accuracy through multiple cameras and multiple microphones is unified under a common framework. It is through the use of this theory that the comparative analysis of localisation accuracy is made. It is shown that audio contributes little in terms of accuracy when fused with video for localisation.

Chapter 5: Optimal Microphone Placement

This chapter explores the effect of microphone array positions on the expected accuracy of audio-based localisation. A simulated annealing based algorithm is introduced for automatically optimising the positions of the microphone arrays. The analysis is presented from a theoretical viewpoint. The work draws on existing mathematical theory defining lower bounds on localisation accuracy in a reverberant environment. These bounds are further developed to include important aspects which influence localisation performance such as the relative angles between the speakers and the microphones. The CHIL lecture room is examined for a hypothetical lecture scenario and the expected localisation accuracy for the given microphone array setup is examined. Using this example, the usefulness of the proposed algorithm for optimising the microphone array positions is shown.

Chapter 6: Voxel-based Viterbi Active Speaker Tracking V-VAST

This chapter presents a new audio-video based algorithm for tracking the current active speaker in a multi-view recording of a lecture. The algorithm is proposed as part of a system called Voxel-based Viterbi Active Speaker Tracking (V-VAST) which creates a composite single view video sequence of a multi-view lecture recording. The composite view video sequences created, consists

of a user defined main view and an automatically segmented view of the current active speaker. V-VAST operates off-line as a post production step applied to multi-view lecture recordings. In visually segmenting the active speaker from the multiple views available, V-VAST aims to extract the *best* view of the active speaker which is defined as the view in which their face is most visible. The algorithm is extensively tested on multi-view lecture recordings from the CHIL 2005 evaluation database demonstrating consistently accurate and reliable tracking performance.

Chapter 7: Discussion & Conclusion

The final chapter of the thesis provides a summary of the main results and discusses their relevance and significance. Future work and future directions for audio-visual active speaker tracking are also suggested to the reader.

1.3 Contributions

The contributions offered by the work in this thesis are summarised in the following.

- A new skin colour model is presented which models for the non-linear dependence of skin-tone on luminance.
- The analysis of localisation uncertainty in multi-camera and multi-microphone systems is unified under a single framework
- A simulated annealing algorithm is introduced for determining the optimal positions of multiple microphone arrays for minimising localisation uncertainty.
- A new audio-video based active speaker tracking algorithm called V-VAST is proposed for generating a composite view video sequence from a multi-view lecture recording.

1.3.1 Publications

Some of the above works and others arising from this thesis have appeared in the following publications.

[39] D. Kelly, F. Boland. Motion Model Selection in Tracking Humans. *Irish Signals and Systems Confonerence (ISSC)*, pages 363-368, 2006.

[40] D. Kelly, F. Pitié, A. Kokaram, F. Boland. A Comparative Error Analysis of Audio-Visual Source Localization. In *Workshop on Multi-camera, Multi-modal Sensor Fusion Algorithms and Applications M^2SFA^2) in conjunction with 10th European Conference on Computer Vision (ECCV)*, 2008.

[41] D. Kelly, F. Boland, Optimal Microphone Placement for Active Speaker Localization. In *8th IMA International Conference on Mathematics in Signal Processing*, Cirencester, England, UK, Dec. 16th-18th 2008.

2

Audio and Video Features for Active Speaker Localisation

In this chapter, a review is presented of different audio and video based features which can be used for localising active speakers. The first section of this chapter examines audio-based features. In the audio domain, the specific use of multiple microphones for localising active speakers is examined. Numerous features arise from the capture of a speech source by multiple spatially separated microphones, which can be used to locate the source. An overview of these features is presented and how they can be used in localising active speakers is described. Also discussed, are aspects of the lecture room environment such as its acoustic properties which fundamentally limit the accuracy and usefulness of these localisation methods.

The second section of this chapter is dedicated to the analysis of visual features for detecting people using cameras. The multi-camera localisation problem is considered. In addition to the difficulties presented by the multi-camera scenario, the challenges of detecting people reliably in lecture rooms are explored. A new skin colour model is introduced for use in detecting face regions within a scene which is employed later in the V-VAST tracking algorithm introduced in chapter 6.

This chapter also acts as background to chapter 3 which analyses the combination of audio and video features for tracking. It is not the aim of presented material in this chapter to be exhaustive. Instead, it focuses on introducing the localisation problem, basic terminology and feature extraction techniques which will be referred to in later chapters.

2.1 Audio Features

Humans like many animals in nature posses the innate ability to locate sound sources in three-dimensional space. This is achieved through binaural hearing whereby spatial information is extracted from the sound source received at the two ears. This task is performed with relative ease, often in very noisy environments and even in the absence of sight. In many applications; particularly those which aim to interact with or serve people, the ultimate goal is to fully replicate the binaural hearing ability observed in nature. Research efforts over several decades have been dedicated to examining how such spatial information can be extracted from audio signals using multiple microphones.

Up until now this has not been achieved, as many aspects of how humans achieve binaural localisation remain unknown. What is presently known however, is that in locating acoustic sources, humans rely on at least two spatial cues. These cues arise due to the spatial separation of the ears causing sounds to be received at the ears, at different points in time. This results in a relative time delay between the two received sounds known as the interaural time difference. The second cue which humans use which is also due to the spatial separation of the ears, is the relative difference in the intensity of the sound received by each ear. This cue is known as the interaural intensity difference or interaural level difference.

Using two spatially separated microphones, one can aim to localise an acoustic source by extracting similar spatial cues from a source signal received by the microphones. The microphones can be used to indicate the sound intensity observed at their positions in a room. Both this cue and the relative time delay between the received signals can be used to indicate the direction of the source to the microphones. Indeed, this analysis is not restricted to just two microphones but multiple pairs of microphones and arrays of microphones can be used in achieving the localisation task.

In the following sections, the use of multiple microphones for audio-based localisation is described. It is examined how the time-delays between multiple microphones and the differences in sound intensity observed by multiple microphones can be used to localise acoustic sources. The presentation will focus in parts on the general problem of localising acoustic sources. Of course, in this thesis the acoustic source will correspond to an active speaker. Where describing an active speaker simply as an acoustic source is too general, specific problems relating to active speaker localisation will be addressed.

Room Acoustics

When a person speaks, their vocal tract vibrates, which introduces a sound wave into the surrounding medium. The rate at which this vibrating energy is converted into sound energy is defined by the *sound power* P ($Watts$). We are intuitively familiar with the concept of *loudness* in relation to sound sources. Loudness however is subjective, based on perception and is not directly measurable by microphones. Although not a measurable quantity, what is perceived

as loudness is related to the sound power P produced by the source. A $10W$ sound source for instance, will produce a sound which will be perceived as louder than a $1W$ sound source.

Sound Intensity

The perceived loudness of a sound is not only dependent on P but also the distance to the source. A sound source that is heard close to its source appears louder to a listener than the same sound heard from far away. The measurable quantity which relates loudness to the distance from the source is the *sound intensity*. The intensity of a sound at a particular point within a room is determined by the sound power P and also by effective area over which P is dispensed. If an omni-directional source is assumed; that is, one which propagates sound waves equally in every direction, then P is dispersed over a spherical region. The area of this spherical region is equal to $4\pi r^2$ where r is the distance of the sound source to the point where the sound intensity is to be measured. It is important to note that sound intensity is a vector quantity. It is common for sound intensity to be measured perpendicular to the effective spherical area of P. This convention is also maintained in this thesis when referring to sound intensity and is denoted as the *one-sided* (one-direction) sound intensity. By this, the sound intensity I_d at a distance r from a source is defined as [73, Chapter 5], [4, Chapter 6],

$$I_d = \frac{P}{4\pi r^2} \tag{2.1}$$

This states that the observed sound intensity is inversely related to the distance to the source (r) squared. This is known as the inverse square law of sound propagation.

2.1.1 Propagation of Sound in Rooms

Once a sound is introduced into a room it propagates as a wave, which in a normal room will travel at a speed of $343ms^{-1}$. As a wave, it is subject to all forms of wave distorting phenomena such as refraction, diffraction, interference and reflection. The actual room environment; its contents and structure, will dictate whether all or only some of these distortions are observed. Such factors will also dictate the extent of these distortions. In the lecture room environment concerned in this thesis, the presence of people, desks, chairs and structures such as walls mean that all the mentioned wave distortions are likely to be observed. As a result, the sound observed by a microphone at any point in the room will be different to that of the emitted sound.

Reverberation

Of the mentioned environmental distortions of a sound wave, the most detrimental to the localisation task is that due to reflections. These reflections represent echoed sound waves which build up over time and slowly decay once the source sound stops emitting. This phenomenon is known as reverberation. The rate of decay of the reverberant sound is dependent on how well

or poorly surfaces within a room can absorb sound energy. Highly reflective surfaces within a room lead to a highly reverberant environment. The consequence of multiple reflected sound waves corresponding to the source is that it is difficult to discern the true sound wave from that of its reflections. Effectively, reverberation results in multiple "virtual" source locations from which the true source localisation must be determined.

A Model of Sound Propagation

It is common practice to apply a systems based approach to model the distorting effects of a room on an emitted sound wave. In this approach, the room is modelled as a system with an input corresponding to the sound source and an output corresponding to the distorted sound wave. In relation to the acquisition of a sound wave by a microphone, we will consider the output of the system model as that observed by a microphone placed within the room.

A complete system model should reflect all of the relevant factors which influence the propagation of a sound wave from the source to the microphone. In addition to the distortion enforced on the signal by the acoustic room environment, other issues affect the signal received by the microphone. A speaker does not typically adhere to the omnidirectional sound source model [89, 93, 189]. In general, the intensity of the *direct path* (i.e. path between the source and the microphone) will be dependent on the angular orientation of the speaker to the microphone. For example, the observed sound intensity in front of a speaker will be greater than that observed behind the speaker. The property of the source which defines this is the *source directionality* characteristic of the speaker. In a similar manner, the direct path intensity is further attenuated due to the reception directionality characteristics of the microphone.

To simplify a system based analysis of sound propagation, the room is typically considered to be a Linear Time Invariant (LTI) system. For the cases of interest in this thesis such an assumption is likely to always be violated. This is so, because the acoustic conditions in a classroom or seminar room are always changing due to the movement of people and possibly objects within the room. Any invariant assumption on a speaker's directionality characteristics will also be violated, since people continuously move their head while speaking. Over a suitably short duration however this assumption can be applied with reasonable confidence.

By modelling a room as a LTI system, the propagation of a sound wave from a position $\mathbf{x} = [x, y, z]$ to its acquisition by a microphone positioned at $\mathbf{m} = [m_x, m_y, m_z]$ can be completely described by a room impulse response $h(t)$. Since the acoustic conditions of a room vary considerable over the space of the room, the impulse response $h(t)$ is highly dependent on both $\mathbf{x}$ and $\mathbf{m}$. Using a LTI room model, the source signal $s(t)$ as received by a microphone at position $\mathbf{m}$ can be approximated as,

$$x(t) = h(t) * s(t) + n(t) \tag{2.2}$$

where $n(t)$ is the additive noise observed on the microphone signal. This noise term $n(t)$ is included to account for microphone channel noise and any ambient noise of the room such as

air-conditioning fans, paper shuffling etc. This noise is generally assumed to be uncorrelated with the source signal $s(t)$. To facilitate later analysis, the microphone signal $x(t)$ can be equally represented in the frequency domain by,

$$X(\omega) = H(\omega)S(\omega) + N(\omega) \tag{2.3}$$

where $X(\omega)$, $H(\omega)$, $S(\omega)$ and $N(\omega)$ are the frequency domains representations of $x(t)$, $h(t)$, $s(t)$ and $n(t)$ respectively.

In this thesis, a simplification of the impulse response is assumed. It is considered as consisting of a direct path component $h_d(t)$ and a reverberant component $h_r(t)$ such that,

$$h(t) = h_d(t) + h_r(t). \tag{2.4}$$

The received microphone signal then becomes,

$$x(t) = [h_d(t) + h_r(t)] * s(t) + n(t). \tag{2.5}$$

This assumption is in common use in the literature [42, 177].

Characterising Room Reverberation

Since reverberation can greatly affect audio-based source localisation it is useful to derive some insight as to the expected contribution of the direct path and reverberant components in the microphone signals. We can make this analysis by placing some simplifying assumptions on the acoustic conditions of a typical room. The first assumption is that the sound in a room propagates in all directions with equal magnitude and equal probability such that the net sound intensity at a point is zero. Recall that sound intensity is a vector. Therefore under this first assumption the sound intensity is effectively balanced equally in every direction such that the net sound intensity is zero. This implies a *diffuse sound field* assumption on the room. The second assumption made is that the total sound energy content of a room is conserved and is equal to the energy due to the sound source minus that absorbed by the reflecting surfaces.

The most popular microphones in use today measure *sound pressure*, which under the above assumptions can be defined as $p \propto \sqrt{I}$. It is also worth noting that the consequence of this definition also means that $p \propto \frac{1}{r}$; establishing the inverse distance law for sound pressure. Examining the sound intensity due to reverberation and the direct path component can be used to determine their contribution to the microphone signals. The sound intensity due to the direct path has already been defined in equation 2.1. Under the diffuse sound field assumption, the one-sided sound intensity due to the reverberant field is determined as [4, Chapter 6]

$$I_r = \frac{4P}{R_c}, \tag{2.6}$$

where $R_c = \frac{\alpha A}{(1-\alpha)}$ is know as the room constant, A is the total surface area of the room's walls and $0 \leq \alpha \leq 1$ is the mean absorption coefficient over this area. Here, the room constant R_c can be thought of as a measure of the room's ability to absorb sound. An important observation in relation to equation 2.6 is that for constant P, the one-sided sound intensity is constant and independent of the distance r to the source.

Signal-to-Reverberant Ratio

Another useful measure used to estimate the "amount" of reverberation present in a microphone signal is the Signal-to-Reverberant Ratio (SRR) ratio. This is simply the ratio of I_d in equation 2.1 to I_r in equation 2.6 which gives,

$$SRR = \frac{R_c}{16\pi r^2} \tag{2.7}$$

$$= \frac{A\alpha}{16\pi r^2(1-\alpha)}. \tag{2.8}$$

This indicates that under the diffuse sound field assumption, the distance to the source r is the only varying factor in the observed SRR.

Critical Distance

Since audio-based localisation performance increases if the reverberant content of the microphone signals can be reduced, it is useful to know at what distance from the source r does $I_d = I_r$ or equivalently SRR=1. The distance d_c at which this occurs is known as the *critical distance*. By equating equation 2.1 to equation 2.6 and solving for r, the critical distance is obtained as [4, Chapter 6],

$$d_c = \sqrt{\frac{R_c}{16\pi}}. \tag{2.9}$$

If it is desired to obtain a signal such that the direct path component is most dominant in comparison to the reverberant content, then the microphone must be placed at a distance less than d_c to the source.

Reverberation Time

The Reverberation Time (RT_{60}) is the most commonly quoted measure in characterising the acoustic conditions of a room. Under the assumption of an exponential decay of reverberant sound energy, it defines the time taken for the reverberant sound energy of the room to decrease by $60dB$. Originally determined through empirical analysis by Sabine, the RT_{60} is defined as [4, Chapter 6] [73, Chapter 5],

$$RT_{60} = 0.161\frac{V}{A\alpha}, \tag{2.10}$$

where V denotes the volume of the room. As a measure to characterise a room as reverberant or not, the RT_{60} is effective. However, it is not as useful as the SRR in estimating the reverberant content of a microphone signal since it does not consider the direct path component or the distance to the source r.

2.1.2 Microphone Array Signal Processing

To introduce the signal model which will be maintained in describing microphone array techniques, consider again the model defined in equation 2.5. At this point, the nature of the direct path sound propagation to a microphone indexed by m can be introduced as,

$$h_{d_m}(t) = \frac{a}{r_m}\delta(t - \tau_m) \tag{2.11}$$

where $\frac{a}{r_m}$ is the attenuation factor (recall $p \propto \frac{1}{r}$) and a is a scalar constant dependent on the propagation medium and measurement units employed. The time value $\tau_m = \frac{r_m}{c}$ is the source-to-microphone propagation time. The model of equation 2.5 can now be redefined as,

$$x_m(t) = [\frac{a}{r_m}\delta(t - \tau_m) + h_{r_m}(t)] * s(t) + n_m(t) \tag{2.12}$$

$$= \frac{a}{r_m}s(t - \tau_m) + h_{r_m}(t) * s(t) + n_m(t). \tag{2.13}$$

It is common practice in microphone array signal processing to define the observed microphone signals relative to some reference microphone, such as for instance, the microphone indexed by $m = 0$. Adopting this convention, the observed signal at the mth microphone can be defined as $y_m(t) = x_m(t + \tau_0)$ where τ_0 is the propagation time to the 0th microphone. The observed signal $y_m(t)$ then becomes,

$$y_m(t) = \frac{a}{r_m}s(t + \tau_0 - \tau_m) + h_{r_m}(t + \tau_0) * s(t + \tau_0) + n_m(t + \tau_0). \tag{2.14}$$

To simplify notation, the reverberant signal component is absorbed into the noise component $n(t)$ to define a new noise component as,

$$n_{m_m}(t) = h_{r_m}(t + \tau_0) * s(t + \tau_0) + n_m(t + \tau_0), \tag{2.15}$$

such that the complete signal model is finally expressed as,

$$y_m(t) = \frac{a}{r_m}s(t - \tau_{0m}) + n_{m_m}(t) \tag{2.16}$$

where $\tau_{0m} = \tau_m - \tau_0$ is the relative time delay between the 0th and mth microphones. It can be seen from this that the microphone signals can be represented as the direct path component with a noise term corresponding to the reverberations and any ambient noise. It is clear that

the microphone signals contain scaled and shifted versions of the original source signal $s(t)$, with the time shifts under the established model being defined relative to the 0th microphone.

Frequency Domain Representation

It is useful at this point to introduce a frequency domain representation for a source signal received by an array of microphones which is referred to in later analysis. In the frequency domain equation 2.16 becomes,

$$Y_m(\omega) = \frac{a}{r_m} \exp^{-j\omega\tau_{0m}} S(\omega) + N_{m_m}(\omega). \tag{2.17}$$

A vector of the signals observed at an array of M microphones can be defined as,

$$\mathbf{Y}(\omega) = [Y_0(\omega), .., Y_m(\omega), .., Y_M(\omega)]^T \tag{2.18}$$

$$\mathbf{N}_m = [N_{m_0}(\omega), .., N_{m_m}(\omega), .., N_{m_M}(\omega)]^T, \tag{2.19}$$

to give a vector of observed microphone signals as,

$$\mathbf{Y}(\omega) = S(\omega)[\frac{a}{r_0}, .., \frac{a}{r_m} \exp -j\omega\tau_{0m}, .., \frac{a}{r_M} \exp -j\omega\tau_{0(M)}]^T + \mathbf{N}_m \tag{2.20}$$

$$\mathbf{Y}(\omega) = S(\omega)\mathbf{D}(\omega) + \mathbf{N}_m. \tag{2.21}$$

The vector $\mathbf{D}(\omega)$ is known as the steering vector or propagation vector. Usually the scaling factors in the steering vector are chosen such that its norm is unity.

Spatio-Spectral Correlation Matrix

Again, to facilitate later discussion in this thesis, a commonly used measurement in array based applications, known as the spatio-spectral correlation matrix is defined as,

$$\mathbf{R}_{YY}(\omega) = E[\mathbf{Y}(\omega)\mathbf{Y}^H(\omega)] \tag{2.22}$$

$$= |S(\omega)|^2\mathbf{D}(\omega)\mathbf{D}(\omega)^H + \mathbf{R}_{N_m N_m}(\omega) \tag{2.23}$$

where $\mathbf{R}_{N_m N_m}$ is the noise spectral matrix and H denotes the conjugate transpose. The use of the spatio-spectral matrix is later re-examined in chapter 3 in the review of joint audio-video based tracking methods.

Signal Coherence

An aspect of microphone array signals which will be seen to greatly influence the accuracy of localisation is that of signal coherence. Two signals are said to be coherent if they are delayed and attenuated versions of each other [72]. In a noiseless anechoic environment therefore, signals

received by multiple spatially separated microphones are perfectly coherent. The existence of noise and reverberation act in reducing signal coherence. It will seen in later analysis that signal coherence places a lower limit on the accuracy of localisation. It is therefore an important similarity measure between microphone signals used for localisation and is defined as [20],

$$\gamma_{y_m y_n} = \frac{G_{y_m y_n}(\omega)}{\sqrt{G_{y_m y_n}(\omega) G_{y_m y_n}(\omega)}} \tag{2.24}$$

where $G_{y_m y_n}(\omega)$ is the cross-power spectrum of $y_m(t)$ and $y_n(t)$. A common measure of coherence is the magnitude coherence squared $|\gamma_{y_m y_n}|^2$ which is bounded by zero and unity [56].

2.1.3 Time-delay based Localisation

Time-delay based localisation using multiple microphones represents an *indirect* approach to the localisation problem. Firstly, the relative time-delays between the microphone signals of equation 2.16 are estimated. Secondly, these time-delays are then used to infer the sound source position.

In order for Time Delay Estimate (TDE)-based localisation to be possible it is necessary to define the manner by which a $3D$ source position $\mathbf{x} = [x, y, z]^T$ relates to an observed time delay. Since the speed of sound c is known and given two microphone positions $\mathbf{m}_m$ and $\mathbf{m}_n$, the expected time-delay observed by the pair of microphones due to a source at $\mathbf{x}$ can be approximated as,

$$\tau_{mn} = \frac{f_s}{c}(|\mathbf{m}_m - \mathbf{x}| - |\mathbf{m}_n - \mathbf{x}|) \tag{2.25}$$

where f_s is the sampling frequency. Including the sampling frequency in the definition means that time-delays are quoted in units of audio samples rather than seconds. The $\mathbb{R}^3 \rightarrow \mathbb{R}$ mapping defined by equation 2.25 is referred to as the *time-delay measurement* function in this thesis. An illustration of the time-delay τ_{mn} arising from two microphones is illustrated in figure 2.1a.

For a given time-delay τ_{mn}, the source position $\mathbf{x}$ is constrained in $3D$ space to a hyperboloid of two sheets with focal points corresponding to the microphone positions $\mathbf{m}_m$ and $\mathbf{m}_n$. Several techniques have been proposed for estimating the source position as the approximate intersection of multiple hyperboloids corresponding to TDEs at multiple pairs of microphones [199]. Such approaches for TDE-based localisation can be complex, since finding the intersection is a nonlinear problem [192].

Furthermore, the hyperboloid defined by a TDE at a pair of microphones is sensitive to slight changes in the estimated time-delay. As a result, the position estimate derived from the intersection of hyperboloids is normally equally sensitive to slight estimation errors. To address this issue, the localisation problem can be formulated as the solution to a set of intersecting spheres where the spheres are centred at the microphone locations [70]. This is popularly known in the literature as the *spherical intersection* method for source localisation.

A formulation of the localisation problem similar to that of spherical intersection solves for

the source position by considering it as residing at the centre of concentric spheres defined by the source and microphone positions [7]. This concept is illustrated in figure 2.2a for the $2D$ case of concentric circles, however, it is easily formulated for application to the $3D$ case of concentric spheres.

It is possible also, to approach the localisation task by simplifying the wave front model. When the distance between the source and the microphones is large relative to the distance between the microphones, the source is said to be in the *far-field*. In cases where this condition can be ascertained, the sound wave impinging on a pair of microphones can be assumed to be *planar*. In this case, a directional angle to the source relative to the microphones can be approximated using a TDE. This is shown in figure 2.1b where under the far field assumption, the DOA of the source sound wave is approximated as

$$\theta_{nm} = \cos^{-1}\left(\frac{\tau_{mn}c}{f_s d_{mn}}\right), \tag{2.26}$$

where d_{mn} denotes the distance between the microphones. This is referred to in this thesis as the *DOA measurement function*.

Localisation using the DOA method in the absence of TDE uncertainty is straightforward. The source location is simply determined as the intersection of bearing lines defined through knowledge of the microphone positions $\mathbf{m}_m$ and $\mathbf{m}_n$ and the DOA angles to a set of microphone pairs. In the presence of TDE uncertainty however, these bearing lines are unlikely to intersect at a single point. To account for such a situation, the closest intersecting points of all bearing line pair combinations can be determined and the source position estimate obtained as a weighted average of these points [113]. This localization strategy for the $2D$ case is illustrated in figure 2.2b. Extending this method to the $3D$ localisation problem is straightforward.

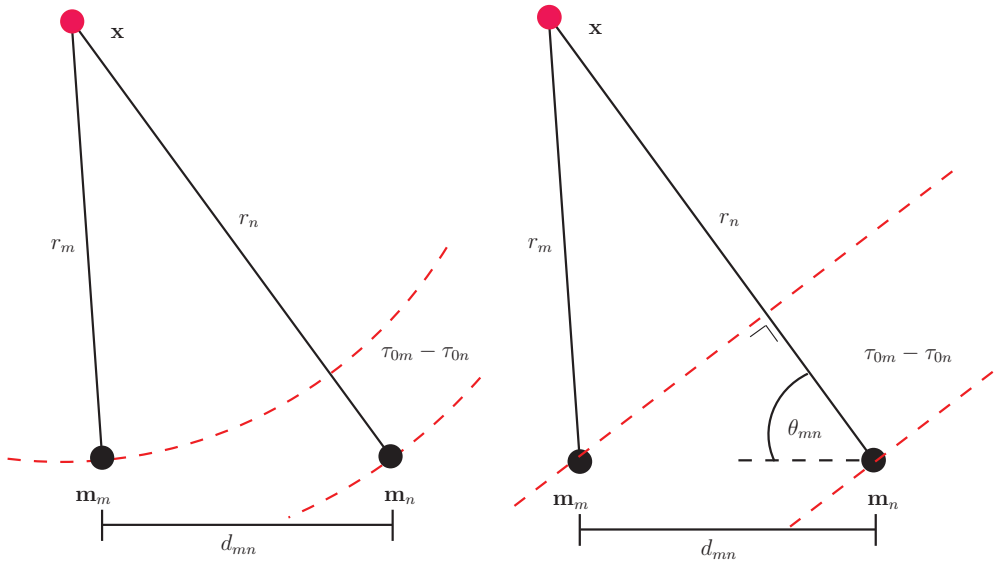
2.1.3.1 Time Delay Estimation

In order to employ TDE-based localisation it is necessary to estimate the relative time delay between multiple microphone pairs by some means. This thesis considers cross-correlation based time-delay estimation techniques which are the most straightforward and by far the most commonly used in speaker tracking applications. The reader is referred to [85] and references therein for a review of existing time-delay estimation techniques.

In an attempt to reduce the effects of reverberation and noise on the source signal received at the microphones, it is normal practice to apply a pre-filter to the microphone outputs. If a pre-filter $h_{f_m}(t)$ is applied to the microphone signal $y_m(t)$ then the processed microphone output is given by,

$$z_m(t) = h_{f_m}(t) * y_m(t). \tag{2.27}$$

Under this assumed signal model, the cross-correlation of $z_m(t)$ and $z_n(t)$ can be used to estimate the relative time delay τ_{mn} between the two microphone signals. The cross-correlation function



(a) *Time-delay $\tau_{mn} = \tau_{0m} - \tau_{0n}$ between microphones $\mathbf{m}_m$ and $\mathbf{m}_n$ corresponding to a spherical wave-front model.*

(b) *DOA θ_{mn} between microphones $\mathbf{m}_m$ and $\mathbf{m}_n$ corresponding to the source position $\mathbf{x}$. The DOA is defined by the time-delay $\tau_{mn} = \tau_{0m} - \tau_{0n}$ under the planar wave-front model.*

Figure 2.1: *Illustration of the relative time delay and Direction Of Arrival (DOA) at a pair of microphones.*

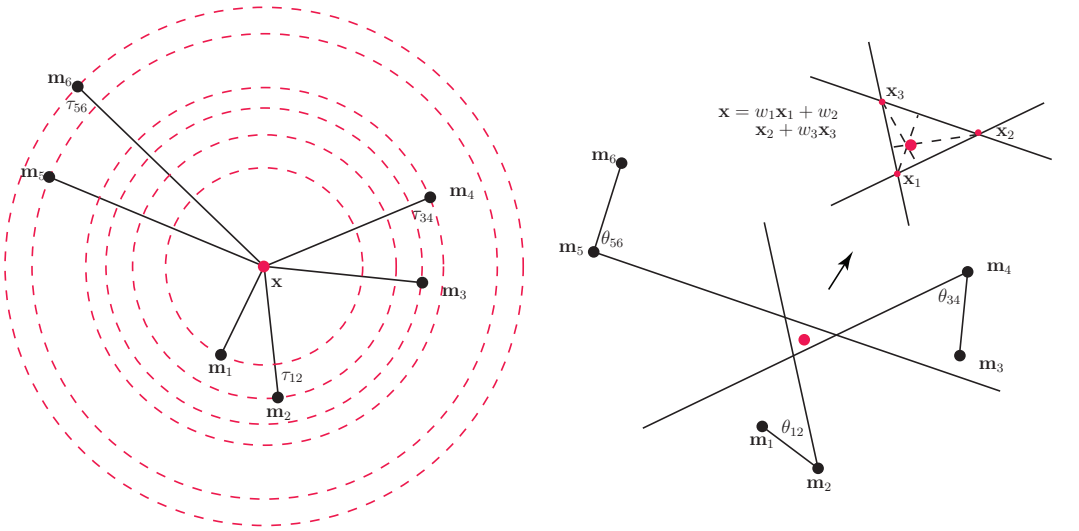
is defined as,

$$R_{z_m z_n}(\tau) = E[z_m(t) z_n(t - \tau)] \tag{2.28}$$

where $E[\cdot]$ denotes the expectation operator. The cross-correlation function attains a maximum at the relative time-shift between $z_m(t)$ and $z_n(t)$ where the correlation between the two signals is greatest. The time-delay can therefore be estimated as,

$$\hat{\tau}_{mn} = \arg\max_{\tau} R_{z_m z_n}(\tau). \tag{2.29}$$

For two microphone signals z_m and z_n which are uncorrupted by noise or reverberation, the maximum of equation 2.29 is expected to occur at the true time delay $\tau_{mn} = \tau_{0m} - \tau_{0n}$. In real environments however, since reverberation can act to create multiple "virtual" acoustic sources within a room, multiple peaks can be observed in the cross-correlation function. It can happen that, when the distorting effects of reverberation are significant, the maximum peak may not correspond to the true source position [12]. In such circumstances, the TDE obtained through equation 2.29 will be erroneous. It is common in the time-delay estimation literature to refer to such TDEs as *anomalous*. This terminology will be used throughout this thesis however it should be noted that this is equivalent to the term *outlier* used in the general estimation literature.



(a) *The source is located at the centre of a set of concentric circles where a single microphone is located on the circumference of a each circle [7].*

(b) *When the bearing lines to multiple microphone pairs do not intersect at a unique point, multiple possible source locations are obtained. In this example, the bearing lines defined by the angles θ_{12}, θ_{34}, θ_{56} intersect at locations $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$. When this occurs, a weighted sum of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ can be used to estimate the true location i.e. $\mathbf{x} = w_1\mathbf{x}_1 + w_2\mathbf{x}_2 + w_3\mathbf{x}_3$. The weights w_1, w_2 and w_3 are determined based on the probabilities of the TDEs used to estimate the DOA bearings θ_{12}, θ_{34}, θ_{56} [125].*

Figure 2.2: *Time-delay and DOA localisation techniques.*

Generalised Cross-Correlation

Using the Wiener-Khintchine theorem, the cross correlation function of equation 2.28 can also be represented as,

$$R_{z_m z_n}(\tau) = \mathcal{F}^{-1}\{E[Z_m(\omega)Z_n^*(\omega)]\} \tag{2.30}$$

where $\mathcal{F}^{-1}\{\cdot\}$ is the inverse Fourier transform and the asterisk $*$ denotes the complex conjugate. In terms of the applied pre-filters as in equation 2.27, the cross-correlation function is,

$$R_{z_m z_n}(\tau) = \int_{-\infty}^{\infty} H_{f_m}(\omega)H_{f_n}^*(\omega)G_{y_m y_n}(\omega)e^{j\omega\tau}d\omega \tag{2.31}$$

$$= \int_{-\infty}^{\infty} \Psi(\omega)G_{y_m y_n}(\omega)e^{j\omega\tau}d\omega \tag{2.32}$$

where $G_{y_m y_n}(\omega)$ is the cross-power spectrum of the unfiltered microphone signals $y_m(t)$ and $y_n(t)$ and $\Psi(\omega) = H_{f_m}(\omega)H_{f_n}^*(\omega)$ is the combined frequency domain weighting equivalent of the time domain prefilters $h_{f_m}(t)$ and $h_{f_n}(t)$. This is the Generalised Cross Correlation (GCC) formulation of the time-delay estimation problem. It is common to define the frequency domain

weighting functions $\Psi(\omega)$; also known as processors, rather than time domain pre-filters. For the case where $\Psi(\omega) = 1$, this corresponds directly to cross-correlation without pre-filtering.

Various frequency domain weighting functions exist for use with GCC, such as the Maximum Likelihood (ML) weighting [20],

$$\Psi_{ML}(\omega) = \frac{1}{|G_{y_m y_n}(\omega)|} \frac{|\gamma_{mn}|^2}{[1 - |\gamma_{mn}|^2]}. \tag{2.33}$$

As can be seen from this, the ML processor acts to accentuate the signals at frequencies where the signal coherence is highest. This model is built on an anechoic (free space) sound propagation model with uncorrelated noise on the received signals. Therefore, the effects of a reverberant environment on ML time-delay estimation, tends to reduce its accuracy and reliability considerably [12]. As a result it is not generally suitable in its basic form for speaker localisation in real room environments.

A more suitable processor which has found extensive use in TDE-based speaker localisation is the Phase Transform (PHAT) weighting [20],

$$\Psi_{PHAT}(\omega) = \frac{1}{|G_{y_m y_n}(\omega)|}. \tag{2.34}$$

The effect of this filter is to assign an equal weighting to the signals at each frequency. As a result, the correlation function obtained is determined using the phase information of the signals only. Since each frequency band is equally weighted however, errors are accentuated in frequency bands where the signal power is low [20]. Despite this, there is a strong theoretical basis [26, 177] supporting the use of the PHAT weighting above other processors in reverberant environments which is supported by empirical analysis [86]. Although the PHAT weighting is useful to counteract the effects of reverberation, it is not without its limitations and shows poor performance in low reverberant and low noise environments [88, Chapter 5].

As previously stated, the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) weighting does not attempt to account for the presence of noise. To address this Wang et al. [75] proposed a modified form of the PHAT processor to be applied as,

$$\Psi^{MOD}_{PHAT}(\omega) = \frac{1}{\gamma|G_{y_m y_n}(\omega)| + (1-\gamma)|N_m(\omega)|^2} \tag{2.35}$$

where $0 \leq \gamma \leq 1$ is a weighting which is equivalent in its definition to the SRR. One difficulty in GCC weighting functions defined for background noise is that an estimate of the noise must be made. Typically, an estimate of background noise is made during periods when the source is not active.

Additional frequency domain processors have been proposed specifically for speech source localisation. Brandstein et al. [113] define a pitch-based frequency domain weighting processor. The premise in their approach is that distinctly periodic portions of the received signal spectrum

indicate signal content which is largely unaffected by reverberation or noise. Their weighting aims to increase the contribution of such signal content in the correlation estimate and decrease that which deviates from this assumption. They employ a harmonic speech model and measure the deviation of the received microphone signals from this model to weight the signal content. The proposed weighting is defined as,

$$\Psi_s = \frac{(1 - max[e_{m,i}, e_{n,i}])^\beta}{|G_{y_m y_n}(\omega)|} \tag{2.36}$$

where $e_{m,i}$ and $e_{n,i}$ are the normalised errors between the ith harmonic of the microphone signals m and n respectively, in relation to that defined by the harmonic speech model. The variable β is a heuristically determined parameter and in the range $\beta = [1, 2]$. Comparing, equation 2.34 and equation 2.36 it can be seen that Ψ_s is effectively a weighted version of the PHAT processor, therefore it can observe similarly poor performance when the Signal to Noise Ratio (SNR) is low.

2.1.3.2 Fundamental Limitations on Time Delay Estimation

There is a fundamental limitation on the performance of time-delay estimation. The effect of the source signal properties which influence this performance limit can be characterised by the Cramér-Rao Lower Bound (CRLB). This performance measure defines a lower limit on the achievable variance of any unbiased TDE [6] and defines the uncertainty on the TDE locally about the true time-delay.

Assuming that the signal and noise spectra are constant for $-2\pi B \leq \omega \leq 2\pi B$ where B is the signal bandwidth, then the CRLB on the variance of a time-delay estimate is given by [6,56],

$$\sigma^2_{CRLB} = \left[2T \int_0^{2\pi B} \omega^2 \left(\frac{SNR^2}{1 + 2SNR}\right) d\omega\right]^{-1} \tag{2.37}$$

$$= \frac{3}{8\pi^2} \frac{(1 + 2SNR)}{SNR^2} \frac{1}{B^3 T}. \tag{2.38}$$

It is clear that the assumption of constant signal power is unrealistic in speech applications. In the absence of a more specific treatment of the speech localisation problem in the available literature however, some insight into the expected performance of time-delay estimation can still be gained under this assumption.

The key observations in equation 2.38 are that the minimal achievable variance is affected by the SNR, signal bandwidth B and the length of the observation window T. In particular,

these are related to the variance of a TDE in the following manner,

$$\sigma^2_{CRLB} \propto \frac{1}{SNR^2} \tag{2.39a}$$

$$\sigma^2_{CRLB} \propto \frac{1}{B^3} \tag{2.39b}$$

$$\sigma^2_{CRLB} \propto \frac{1}{T} \tag{2.39c}$$

Therefore, in designing a time-delay estimator, increasing any of these quantities improves the accuracy of TDEs. In speech applications it is clear that the bandwidth B is not under the control of the designer and is typically $3kHz$ ($400Hz - 3400kHz$). The SNR however, generally decreases as the source-to-microphone distances increases. Therefore, ensuring that the microphones are close to the source can increase the SNR and improve TDE accuracy. Also, increasing the time analysis window T will act to improve the accuracy of TDEs. It is often the case however that T is dictated by the required measurement update rate. In tracking, typically the highest update rate is desired restricting T to small values. A high update rate is therefore employed at the cost of reduced TDE accuracy.

The CRLB, does not completely reflect the true nature of time-delay estimation performance as it incorporates noise analysis only and does not consider the effects of reverberation. As the SNR decreases, a thresholding effect is observed whereby the accuracy of a TDE diverges from that estimated by the CRLB. This is a result of the wrong peak in the cross-correlation function of equation 2.29 being selected as that corresponding to the true time-delay estimate. This type of error arising due to anomalous TDEs is known as a "large" error in time-delay estimation problems and is not modelled by the CRLB. The CRLB only describes uncertainty *locally* about the true time-delay which is often referred to as a "small" error in the estimation problem. Considerable efforts have been made to theoretically model the performance of time-delay estimation in the presence of both large and small errors. Chapter 5 continues this discussion where these theoretical models are used in examining the localisation performance of a configuration of microphone arrays in a lecture room.

Improving the Reliability of Time Delay Estimates (TDEs)

Since reverberation can have such a significant effect on the accuracy and performance of TDE-based localisation, it can be useful to establish some measure of TDE reliability.

The most basic reliability measure is based on the energy of the microphone signals. Typically, high signal energy will indicate the presence of a speech source. Only determining TDE over frames of significant energy can therefore yield more accurate estimates. More sophisticated approaches aim to directly classify the signal and speech or non-speech by learning the characteristic features of speech signals [44].

Additional strategies for improving the reliability of TDEs have been proposed based on the *precedence effect*. The precedence effect is also know as the *law of the first wave-front* and is a

psyhcoacoustical phenomenon observed in humans where localisation is only performed on the direct sound wave. Employing the precedence effect in time-delay estimation requires firstly determining the onset of the direct sound wave in the microphone signals. Once this onset is determined, time-delay estimation is only applied over a short period at the detected onset [33]. By only considering a short period about the onset, time-delay estimation is not performed on the later arriving reverberations over which TDEs are less reliable. Additional techniques have examined modelling the precedence effect for localisation by learning the association of the spectral content of the microphone signals to that of localisation precision [108].

The value of the maximum peak in the GCC function and the ratio of the first and second largest peak can also be used to estimate TDE reliability. Typically, the value of the maximum peak has a direct relation to the likelihood of it corresponding to a true sound source. Furthermore, the ratio of this peak value to that of the second largest peaks can also indicate its significance and therefore is reliability. Analysis of both these reliability criteria, has been shown to yield equivalent performance to that obtained by modelling the precedence effect [33].

Determining reliable TDEs can also be considered as an outlier estimation problem through some temporal based statistical analysis. Given a number of short temporal windows it is possible to build a histogram of the TDEs from which the true TDE can be determined as in [139]. Such an approach can be particularly useful in speech applications where the source signal is not always continuous but possibly intermittent.

2.1.4 Localisation through Sound Intensity Differences

Given the definition of sound intensity in 2.1, it can be seen that it contains information relating to the distance to the source r. If P is known, the distance to the sound source from a particular point could be determined simply by measuring the sound intensity at that point. In speaker localisation problems, P is never known and not easily estimated. Furthermore, the inverse square law is a simple model of sound propagation and is not always adhered to, particularly if the omni-directional source model is violated.

In applications where the assumption of omni-directional sound propagation is valid, then a measure of the energy of the microphone signals can be used to localise an acoustic source. Consider an estimate of the energy of the mth microphone signal in equation 2.16 obtained as,

$$E_m = \frac{a}{r_i^2}\int_0^T s_i^2(t-\tau_{0m})dt + \int_0^T n_{m_m}^2(t)dt \tag{2.40}$$

$$= E_s + \int_0^T n_{m_m}^2(t)dt, \tag{2.41}$$

where E_s is the source signal energy. It is assumed that the time analysis window T is sufficiently large such that the time-delay τ_{0m} does not affect the signal energy estimates. Assuming for a pair of microphones indexed by m and n, that $\int_0^T (n_{m_m}^2(t) - n_{m_n}^2(t))dt$ is a zero mean random

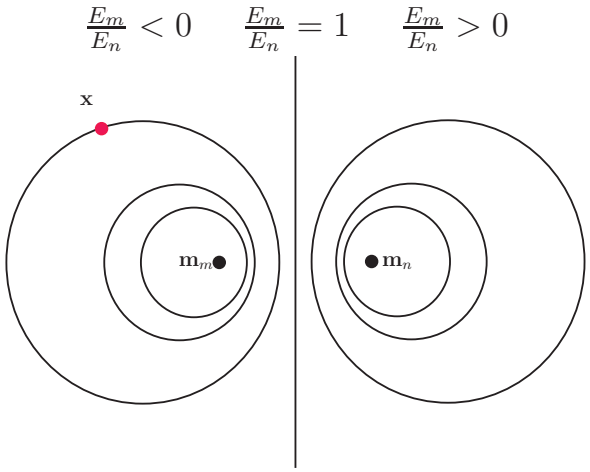


Figure 2.3: *Localisation using interaural level differences.*

variable, the following relationship between microphone signal energies can be defined as [172],

$$\frac{E_m}{E_n} = \frac{r_n^2}{r_m^2} \tag{2.42}$$

This relationship constrains the position of the sound source to a surface in $3D$ space as illustrated in figure 2.3. The intersection of the $3D$ surfaces defined by multiple pairs of microphones using this relation can then be used in localising the position of the source.

The successful use of interaural level differences for localisation requires the direct measurement of the source signal energy using multiple microphones. If the microphone signals are distorted due to reverberation or noise, this is generally not possible. As a result, techniques which use interaural level differences for localisation have found little use in real room environments. Furthermore, the assumption of an omnidirectional source is generally too restricting for the speaker localisation task.

2.1.5 Steered Response Power based Localisation

Steered Response Power based localisation methods search a number of hypothesised speaker locations for the position corresponding to the maximum received power at the microphones. For each location under examination, a steering vector is defined to steer an array of microphones to that position. A measure of the total received power from that position can then be used to clarify as to whether or not the position corresponds to an active speaker.

Given the observed vector of signals in equation 2.21, we can aim to recover the source signal $S(w)$ through determining the steering vector $\mathbf{D}(\omega)$ and apply it to the observation vector $\mathbf{Y}(\omega)$

as,

$$\mathbf{D}^H(\omega)\mathbf{Y}(\omega) = \mathbf{D}^H(\omega)S(\omega)\mathbf{D}(\omega) + \mathbf{D}^H(\omega)\mathbf{N}_m(\omega) \tag{2.43}$$

$$= S(\omega) + \mathbf{D}^H(\omega)\mathbf{N}_m(\omega) \tag{2.44}$$

The operation on the observation vector $\mathbf{Y}(\omega)$ in equation 2.43 can be thought of as time aligning the microphone output signals and determining their weighted sum. This is known as the *delay and sum* beamformer. The success of the delay-and-sum beamformer relies on the source signals to sum coherently and the noise and reverberant signal components to sum incoherently.

Since the steering vector is dependent on the source position $\mathbf{x}$ the beamformer can be steered to any hypothesised position. Using the audio-based measurement function of equation 2.25 the expected time delay corresponding to a hypothesised source position $\mathbf{x}$ can be determined. These expected time-delays can then be used to define a steering vector $\hat{\mathbf{D}}(\omega, \mathbf{x})$. Through this it is possible to examine the total signal power received from the direction of $\mathbf{x}$ as,

$$P_{SRP}(\mathbf{x}) \int_{-\infty}^{\infty} |\hat{\mathbf{D}}^H(\omega, \mathbf{x})\mathbf{Y}(\omega)|^2 d\omega. \tag{2.45}$$

This can be used for localisation where multiple hypothesised locations are examined in terms of their steered response power and the location at which this yields a maximum is taken to be the source position. Such localisation methods are regarded as Steered Response Power (SRP) based localisation techniques. They determine an estimate of the source localisation as,

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \int_{-\infty}^{\infty} |\hat{\mathbf{D}}^H(\omega, \mathbf{x})\mathbf{Y}(\omega)|^2 d\omega. \tag{2.46}$$

One criticism of Steered Response Power (SRP)-based localisation techniques is that if the localisation search space is large then these techniques can be computationally demanding.

Steer Response Power - Phase Transform (SRP-PHAT)

The SRP for a hypothesised location can be equivalently represented in terms of the GCC functions arising from every pair of microphones. It can be shown that the relation in equation 2.46 is equivalently defined as [88, Chapter 6],

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \left[\sum_{m=0}^{2M} \sum_{n=0}^{2M} R_{z_m z_n}(\tau_{mn}) \right]. \tag{2.47}$$

where τ_{mn} is the expected time delay at microphones m and n due to the source at position $\mathbf{x}$. When the GCC function is obtained using the PHAT processor of equation 2.34, then this process of localisation is known both as the Steer Response Power - Phase Transform (SRP-PHAT) [87] and the Global Coherence Field (GCF) [123]. These localisation techniques simply

require evaluating the GCC-PHAT function for each microphone pair at the theoretical time delay corresponding to $\mathbf{x}$ and then summing the result over all microphone pairs. Since SRP-PHAT based localisation techniques employ the PHAT processor, they benefit form its robust performance in reverberant environments.

2.2 Video Features

In its simplest form, the problem of detecting a person in a video sequence reduces to a binary classification problem whereby image pixels are determined as either foreground corresponding to the person, or background corresponding to everything else in the scene. In the same way the use of multiple microphones for audio-based localisation is inspired by the binaural localisation ability of humans, video-based person detection techniques also aim to extract features which humans use in visually locating people.

Many low-level visual features are in common use for detecting people in video sequences such as, the motion observed in a scene, colour information and edge details. Such low level visual information alone however, is often not sufficient for the accurate detection of people. Models of high level visual features such as faces can be developed for detecting people using low-level visual features.

In this section, the extraction of suitable video-based features for the detection of people are briefly examined. In particular, the multi-camera environment and the challenges which it presents to the detection problem are analysed. Background modelling for foreground detection and existing techniques for face detection are introduced. In relation to face detection, focus is directed on the problem of modelling skin colour. A new model of skin colour is introduced which aims to adequately account for the non-linear dependence of skin tone on illumination.

Also examined in this section, is the relationship between a $3D$ point and its projection onto the image plane of a camera. This introduces the measurement function for cameras which will be used in later chapters of this thesis that consider the problem of localisation using multiple cameras.

Challenges in the Lecture/Seminar Room Environment

The lecture room represents a challenging visual tracking environment. Many issues exist which make the detection of people difficult, such as, varying illumination; visual clutter and the presence of multiple people. In relation to the multi-view tracking problem considered in this thesis, three significant problems are highlighted.

- **Varying Illumination:** Both the temporal and spatial variation of illumination can be problematic to many visual trackers. Although temporal variation in illumination can be observed in typical lecture room environments, it does not represent the most difficult problem since lighting conditions often remain consistent. However, there can be

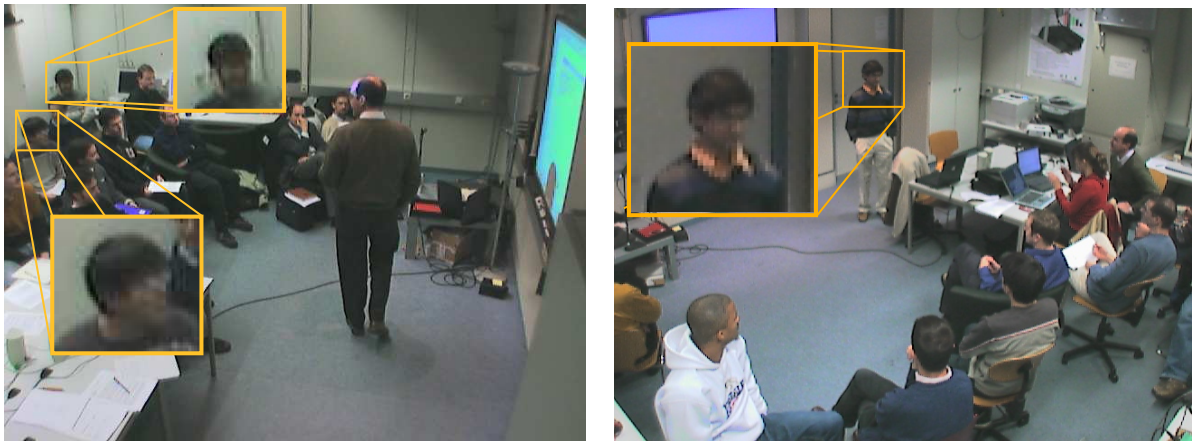
significant variation in illumination spatially. This situation generally arises naturally in lecture room environments, whereby lighting is deliberately focused on a presenter. Also, it is often the case that lighting is lowered over an audience area so as to improve the visual contrast of projector displays. Spatially varying illumination can be problematic in the detection of people since it distorts edge-based visual features. Additionally, edges not arising from features of interest are introduced at the gradients between regions under different illumination.

Shadows arising due to people or objects within the scene can also contribute to the spatial variation in illumination. Shadows make the problem of detecting people within a scene challenging since it is often difficult to discern a shadow from a person. Not only do shadows typically maintain the silhouetted shape of a person, but they also undergo the same motion as that of the person creating the shadow. Moving people create shadows which can also introduce local temporal variations in illumination within the room. Problems due to varying illumination can be accentuated in the multi-view problem since each view can observe the same position under different illumination due to their different viewpoints.

- **Static Foreground:** It is likely to be the case that the presenter will continuously move throughout a presentation. In this work however it is desired to not only detect a presenter but every person within the room, since every person is a potential active speaker. The difficulty in this is that whereas motion may provide a simple cue for detecting the position of a moving presenter, it is not guaranteed to be as effective in detecting other people in the scene such as audience members. This is because audience members typically undergo little motion and are effectively static. Therefore, they represent *static foreground* regions which must be accounted for and other features besides motion must be used for detection.

- **Low Resolution:** It is likely in a multi-view scenario where a relatively small number of cameras is used that people will be positioned far away from some of the cameras. As a result, they will only occupy small regions within the image plane resulting in a low resolution capture of the person. This does not favour the extraction of high-level visual features such as faces, since in low resolution, the edge details of facial features are effectively lost. This is worsened in cases where the captured data is also compressed. An example of a challenging scenario for the detection of faces which typically occurs in the considered database of CHIL lectures is illustrated in figure 2.4. From this it is seen that in some cases, little detail of facial features remain due to low resolution capture and compression.

  Since there are multiple views, there may be certain exceptions where the person is captured at a higher resolution by another camera. This however, is not always guaranteed. Furthermore, at least two camera views are required to infer the $3D$ position of detected people. Observing a high resolution capture of people in two views is unlikely under the



(a) (b)

Figure 2.4: *Example of difficult face detection problems in the lecture recordings of the CHIL database. The low resolution capture of people who are not close to the camera combined with JPEG compression of the video frames act in removing considerable face detail making the task of face detection challenging.*

given multi-camera scenario. Therefore, the extraction of high-level visual features are likely to fail or at least be unreliable for detecting all visible faces.

- **Dissimilar Views:** The problem of dissimilar views is specific to the multi-view problem and corresponds to the situation where a single object is observed at different orientations due to the different camera viewpoints. Therefore, the reliable detection of a person in multiple views must address the problem of a person appearing differently in each view. This has significant implications on the face detection problem since a face observed by multiple cameras is captured at different poses. This is discussed further later in this chapter, where the face detection task is described.

2.2.1 Background Modelling

The most common approach to the problem of foreground detection is to employ background models based on the observed frames. Pixels which deviate significantly from the assumed model are simply regarded as foreground within the scene. The use of a threshold is common in discerning such a deviation where the foreground is determined as satisfying the following,

$$|\mathbf{I}_t - \hat{\mathbf{B}}_t| > b_t, \tag{2.48}$$

where $\mathbf{I}_t$ denotes the pixel intensities of the current frame, $\hat{\mathbf{B}}_t$ are the estimated background pixel intensities and b_t is the chosen threshold value. This approach is known as background subtraction and a recent review of existing techniques may be found in [147].

The background model $\hat{\mathbf{B}}_t$ and threshold b_t can be preset to the specific tracking task. It is reasonable to assume that an unoccupied view of a lecture room can be obtained for each camera. Subtracting this background model from that of an occupied scene could be used to detect newly introduced people and objects. The determination of a time varying background model $\hat{\mathbf{B}}_t$ and also time varying threshold b_t which are updated periodically however, is more desirable. The periodical update of the background model and threshold is preferred since they can be adjusted to account for changes in the background, varying illumination and also to compensate for any camera noise. Without such adaptation, errors in the background model would accumulate over time.

It is important in this approach however that adaptation is applied to background pixels only. This is necessary so as to avoid the background model adapting to the detected foreground. The consequence of updating a background model at foreground locations is that stationary foreground regions eventually become part of the background. This would be particularly problematic in relation to static foreground regions undergoing little motion, such as the audience. The use of an appropriate binary mask which differentiates between foreground and background is often applied at the update stage so as to avoid updating on foreground pixels and is applied through,

$$\hat{\mathbf{B}}_t = F\hat{\mathbf{B}}_{t-1} + (F-1)\hat{\mathbf{B}}_t \tag{2.49}$$

where $F = 1$ indicates foreground and $F = 0$ indicates background. This translates to only updating the background model $\hat{\mathbf{B}}_t$ given that the pixel corresponds to that of background. Since the binary mask F is derived from $\hat{\mathbf{B}}_t$ by equation 2.48, errors in the background model $\hat{\mathbf{B}}_t$ are often reciprocated through the binary mask.

Temporal Median Filter

A temporal median filter can be used to model the background pixels. In this approach the background is assumed to be accurately described as the median of its previous n pixel values. Thus, the model of the background at time t is determined as,

$$\hat{\mathbf{B}}_t = median\left[\mathbf{I}_{t-(n+1)}, \mathbf{I}_{t-(n+2)}, ..., \mathbf{I}_t\right]. \tag{2.50}$$

One disadvantage of this method is that a buffer of the previous n frames must be stored in order to evaluate the median at time t. Typical applications only consider evaluating the median over some subset of the previous n frames [153].

Running Gaussian Average

Another proposed technique for modelling background pixels is to use a running Gaussian average [25]. This approach proposes to model each pixel location independently. The pixel value at time t is assumed to fit a Gaussian probability density function (pdf) where the estimated

background pixel is assumed to be $\hat{\mathbf{B}}_t = \mu_t$ where μ_t is the mean of the pdf and is determined by,

$$\mu_t = \alpha_t \mathbf{I}_t + (1 - \alpha_t)\mu_{t-1}, \tag{2.51}$$

where α_t is an appropriate weight between the current frame $\mathbf{I}_t$ and the previous mean μ_{t-1} with α_t determined emprically. The variance σ_t^2 may be preset or determined in a similar manner to that of equation 2.51. Commonly, the threshold is set to $b_t = 1.96\sigma_t^2$ defining the 95 percentile confidence interval for the background pixel value.

The use of a single Gaussian distribution in modelling the intensity of a background pixel is most useful in situations where the observed pixel intensity adheres to a uni-modal distribution. The uni-modal nature of a Gaussian distribution is inappropriate for modelling distributions of a multi-modal nature. Many factors can cause background pixel intensities to exhibit multi-modal behaviour. Commonly observed scenarios include occlusions and sudden changes in illumination. A single Gaussian statistical model does not fully capture multi-modal behaviours such as these. A more appropriate Gaussian-based approach to modelling multi-modal pixel intensities is to use a Gaussian mixture model [22].

Eigenbackground

An eigenbackground approach to modelling the background determines a suitable model of the background through principal component analysis (PCA) [134]. A training dataset of n frames is used during a learning phase where the mean background μ_0 is determined from the n frames together with its covariance matrix $\Sigma_{\mathbf{B}}$. Eigenvalue decomposition of the covariance matrix through $L = \Phi\Sigma_{\mathbf{B}}\Phi^T$ results in Φ, a matrix of the eigenvectors of $\Sigma_{\mathbf{B}}$ and L, a diagonal matrix of their corresponding eigenvalues. Classification of foreground proceeds as follows. Given the current frame $\mathbf{I}_t$, it is first projected into the eigenvector space through $\mathbf{I}'_t = \Phi_m^T(\mathbf{I}_t - \mu_0)$ where Φ_m is the matrix of the m most significant eigenvectors of Φ. The inverse projection is then determined by $\hat{\mathbf{B}}_t = \Phi_m\mathbf{I}_t + \mu_0$. Finally, foreground is detected using the thresholding approach shown in equation 2.48. The premise in this approach is that static background regions are well described by such an eigenspace model whereas moving regions within the frame are not. The consequence of this is that moving objects do not appear in the background model.

Temporal Differencing

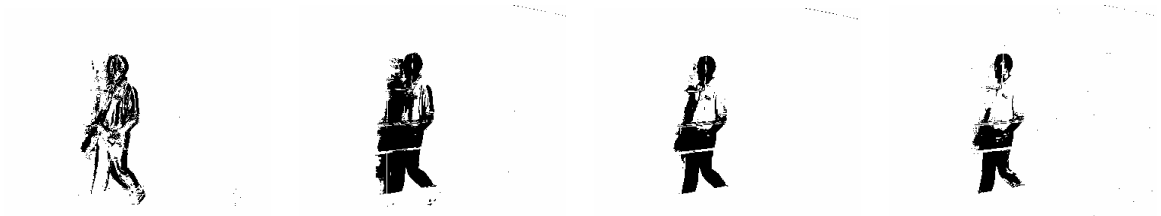
Temporal differencing can be applied to frames within a frame sequence to determine moving people. This approach assumes that a person is continuously moving and also that the motion between frames is sufficient to indicate the foreground region corresponding to the person. This approach is equivalent to assuming the estimated background $\hat{\mathbf{B}}_t$ is simply the previous frame i.e. $\hat{\mathbf{B}}_t = \mathbf{I}_{t-1}$. The threshold b_t of equation 2.48 in this case corresponds to the temporal difference between frames. The only assumption which this technique makes regarding the background is that its pixel values are temporally stationary. A temporal differencing approach using three

(a) Sample Frame

(b) Temporal Median Filter Background Estimate

(c) Running Gaussian Average Background Estimate

(d) Eigenbackground Estimate (n=5)



(e) Temporal Differencing $b_t = 15$

(f) Temporal Median $b_t = 30$

(g) Running Gaussian Average $b_t = 15$

(h) Eigenbackground $n = 5$ $b_t = 76$

Figure 2.5: *Example of the performance of various background estimation techniques on a video sequence from the AV16.3 database [59]. Shown in this figure is the estimated background of the sample frame in (a) using; (b) a temporal median filter where the median is taken from the previous* 100 *frames subsampled by a factor of* 5 *(i.e.* $n = 20$*); (c) a running Gaussian average and (e) the eigenbackground approach* ($n = 5$)*. The detected foreground using background subtraction using temporal differencing, the temporal median filter, the running Gaussian average and eigenbackground techniques are illustrated in (e), (f), (g) and (h) respectively. These examples highlight some of the challenges in foreground detection through background modelling such as the presence of shadows. This is particularly evident in the foreground detected through the temporal median filter in (f). Additional problems occur where the foreground is a similar colour to that of the background which results in large unconnected regions in the detected foreground. This is most evident in the foreground detected by the running Gaussian average and eigenbackground in (g) and (h) respectively. Also, the detected foreground using temporal differencing in (e) is seen to result in large unconnected foreground regions.*

successive frames for foreground detection is presented in [196].

An example of the described techniques for background estimation is presented in figure 2.5. From this it is seen that each of the described techniques are effective for background estimation. In the example some of the difficulties which can arise in the detection of people using background subtraction are described.

2.2.2 Face Detection

Faces are high level visual features which are obvious candidates for localising people. Due to the importance of face detection in many vision-based applications such as surveillance, Human Computer Interaction (HCI), facial expression analysis, visual recognition and people tracking,

it has received considerable attention in the signal processing research domain. In the following sections, existing methods for detecting faces in video are examined. This analysis is not aimed toward an exhaustive or comprehensive review. Instead, existing vision-based methods for face detection are examined for their suitability for locating and tracking speakers in the lecture room environment considered. For recent and comprehensive reviews of face detection, the reader is referred to those by Yang et al. [110] and Hjemlås et al. [47].

Several decades of visual face detection research, has seen the evolution of five main categories of face detectors. These classes correspond to *knowledge-based*; *template matching*; *feature invariant*; and *appearance-based* techniques [110]. These are briefly discussed in the following.

Knowledge-based Methods

Knowledge-based techniques for face detection are based on simple detection rules which encode *a priori* knowledge of face characteristics and visual features. The *a priori* information used, generally consists of simple descriptive features representing human intuition as to what constitutes a face. Typical descriptive information can relate to the expected number of facial features such as eyes, nose and mouth; their relative positions on a face; in addition to their symmetry. Generally, existing knowledge-based methods for face detection apply a top-down detection strategy where a likely face location is first determined, then local features are extracted and evaluated against the established detection rules. Typically, such techniques analyse only the gray-scale intensity characteristics of faces at different image resolutions [21]. The basic premise of such approaches is that the macroscopic features of faces are adequately captured at very low resolution. Face regions in low resolution images often simply appear as uniform intensity regions. This feature can be used to indicate likely face positions. The detection process then implements a coarse to fine detection scheme usually employing pyramid images. The initialised face regions are used to inform the detection of higher level facial features, such as the eyebrows/eyes, nostrils/nose and mouth at higher resolutions.

The challenge in rule-based face detection is to define a set of rules which are flexible enough to be equally applicable to non-frontal faces, varying facial expressions and to be robust in the presence of beards/mustaches as well as glasses. This is not straightforward since the assumption of symmetry or the expected number of visible facial features changes under head rotations. For instance, the number of visible eyes varies between frontal and profile face views. In addition to this, sufficiently high resolution is required to detect such local facial features. As a result, the successful application of rule-based techniques to face detection is largely constrained to problems with frontal faces captured at sufficiently high resolution enabling the detection of high-level facial features. Furthermore, due to the specific use of image intensities only, illumination changes both spatially and temporally can distort the grey-scale facial features making them undetectable.

Template Matching

Template matching is one of the simplest and earliest techniques for locating known patterns in images. As a result it is also representative of some of the earliest approaches to the face detection problem. Using a predefined face template, such techniques aim to identify face regions in an image by direct comparison with the defined template. Elliptical templates are popularly employed and compared to detected edges or silhouettes within an image for head detection. Usually the comparison between an image region and a template is made through some correlation based similarity measure.

Early template matching systems for face detection considered adaptable shape-based templates built on curves defining a face's outline [186]. Such primitive shape models however are never likely to be sufficient to distinguish faces from other objects in complex backgrounds such as observed in a typical lecture room. The use of multiple shaped-based templates for individual facial features such as eyes, nose and mouth can be combined with the outline shape of a face to define more complex face models [78]. Such detailed models have greater potential to discriminate against false-positives which may arise in complex backgrounds containing face-like shaped objects. Once again however such detailed face models require the detection of high-level facial features which is only possible if the face is captured at sufficiently high resolution.

A criticism of template based approaches is that they can not effectively account for variations in pose, scale or shape. This is so, since it is difficult to incorporate such variability into a single template. Although the use of templates are still important in the visual detection of faces, current approaches use more informed strategies for defining face templates. Most current template-based face detectors use information from sets of training face and non-face images to develop a template. The aim in these approaches is to define a template which is supported by this example data in some statistical sense. Although still employing template-based analysis, theses techniques are classed among appearance-based methods.

Appearance-based Methods

Appearance-based techniques attempt to learn an appropriate face model or template using sets of face and non-face example data. Generally, these techniques employ statistical based analysis and machine learning to achieve this.

The appearance-based approach of eigenfaces aims to determine a suitable model of a face using Principal Component Analysis (PCA) [126]. A set of training face images are used during a learning phase to attain a set of representative features which characterises the variation between face images. The method of analysis is identical to that applied in the case of eigenbackground modelling as in section 2.2.1 the only difference being that the PCA is applied to example faces in comparison to example background data.

More complex appearance-based face detectors built on machine learning techniques such as neural networks [74] and support vector machines [48] have been employed to detect faces by

learning to recognise representative visual face features. In general, eigenfaces, neural networks, or support-vectors machines are trained for detecting fixed-pose (usually frontal) faces and in general can only handle fixed-pose (usually frontal) face detection. They can be applied to the multi-pose face detection problem through the use of multiple trained detectors. For instance, a neural-network can be trained on example data of faces at different poses and applied in turn to an image for detecting multi-pose faces. In contrast, the pose estimation problem can be considered separately to the face detection task. The approach of [68, Chapter 5] employs a neural network trained on various face pose examples for pose estimation. This information is then used in selecting the neural network trained on detecting faces at the estimated pose. A disadvantage of such an approach is that in general, a considerable amount of training is required to achieve accurate detection rates at a sufficient number of different poses.

By far the most popular and extensively used appearance-based face detector is that proposed by Viola and Jones [143]. They define a robust real-time face detector which uses a cascade of Haar-like feature based classifiers. In their approach they use a set of horizontal, vertical and diagonal filters in extracting image features at different scales. An optimal cascade of classifiers based on extracted image features is then learned from example grey-scale face and non-face images. The resulting face detector operates by identifying likely face regions in an image using simple and efficient classifiers. More complex classifiers are applied only to these regions for verification as face locations. Applying the classifiers in a cascade of low to high complexity enables non-face regions to be quickly rejected allowing high level classification to be concentrated on likely face regions only. The result of this is that classification is easily achieved in real-time. Similar to the previously described appearance-based face detection methods, applying the Viola and Jones algorithm to the problem of multi-pose face detection requires training an individual detector for each considered pose [144].

Feature Invariant Techniques

Feature invariant approaches examine the bottom-up face detection problem where firstly local facial features such as eyes, nose and mouth are detected. Secondly, the geometry, structure and relative positions of these features are used to infer the presence of a face. In general, the detection of local facial features requires some model of the feature to be defined. For instance, edge detail or corners can be used to build an appropriate model. This can be difficult in many cases since such features are likely to be distorted due to image noise. Previous discussion has highlighted current challenges in detecting faces in images such as varying pose, low resolution and illumination changes. Feature-based face detection aims to extract visual facial features which are invariant under such conditions and use these in locating faces.

The approach in [104] examines six facial features corresponding to the eyebrows, eyes, nose and mouth. Such features at low resolution appear as dark blobs in contrast to the light background of the face. Once detected, the edge structure locally about these features is examined

to build a feature vector of details such as edge length and edge intensity. The invariance of the technique to illumination changes relies on the assumed invariance of edge detection under such conditions. The extracted features are statisically compared with the learned distribution of feature vectors obtained from training data. In an effort to incorporate a level of invariance to pose, combinations of four features among the six concerned are examined to account for the possible occlusion of facial features.

The success of face detection techniques which rely on the extraction of local facial features requires that edge information and gradient information is present in the image. Low resolution and illumination changes will dictate whether this is possible. In low resolution images much of this detail is not retained as was seen in the examples presented in figure 2.4. Colour information when available however, is relatively unaffected by resolution. It can therefore be used for locating likely face regions when high-level facial features cannot be resolved. As a result, the development of skin colour models for detection of skin regions has attracted considerable attention among researchers for face detection. In this thesis, a skin colour model is considered for detecting likely face regions. In the following section the task of skin colour modelling is described and a new skin colour model is introduced.

2.2.3 Skin Colour Modelling

The most common approach to the modelling of skin colour is to transform the *RGB* space to some chrominance colour space so as to decouple skin colour into independent chrominance and luminance components. Usually the luminance information is discarded and only the distribution of skin colour is modelled in chromatic colour spaces using a single Gaussian [111] or mixture-of-Gaussians [165]; the latter being the preferred approach [179]. The motivation in this is that a skin colour model can be developed which is invariant to changes in illumination. Further motivation arises due to the observation that the distribution of different skin colour types occupy a dense compact region within chromatic colour spaces [102]. Therefore, in chromatic colour spaces the variation between the colour of different skin types is reduced which makes it easier to form a general skin colour model which encompasses the various different skin types.

A considerable amount of effort has been dedicated to the analysis of the most appropriate color space for use in the skin colour modelling task. The argument for and against the use of different colour spaces generally focuses on how well regions of skin colour are separated from regions corresponding to non-skin colours. Improvement in the separation of the two colour classes, for example, has been attributed to the use of the normalised r-g and HSV colour spaces in comparison to the *RGB* colour space [79]. There is significant evidence to suggest that chromatic colour spaces do improve skin colour detection in comparison to detection methods using the *RGB* colour space [79, 187].

The significance of the chosen colour space in this observation however has been discounted [2, 115]. In this thesis, the view of Albiol et al. [2] is maintained in that a transformation to a

different colour space does not increase the separation between skin and non-skin colour classes. It is viewed that, it is simply the case that existing statistical modelling techniques such as the single Gaussian and mixture-of-Gaussians are more appropriate for modelling skin in chromatic colour spaces than in the *RGB* colour space. Therefore, it is not the separation between the two classes which is increased through a transformation to chromatic colour spaces, but the ease by which the two classes can be separated in the transformed colour space using standard statistical models. This is not suggesting that the results of improved colour detection in chromatic colour spaces are incorrect, only a reinterpretation of the results in that they can not be regarded in isolation to their employed skin colour model. This is a subtle distinction. It is an important one however, as it proposes that skin and non-skin colour classes can be as equally discriminated in the *RGB* colour space. It simply suggests that a more complex model of the distribution of skin colour is required to achieve this when the *RGB* colour space is used. In this section, skin colour is modelled in the *RGB* colour space.

One criticism of existing chrominance colour spaces is that they do not adequately account for the non-linear dependence of skin colour on the luminance component [156]. This nonlinear dependence can be observed in figure 2.6 where a sample of $\approx$ 5 million skin colour pixels captured under varying illumination in the *RGB* colour space is shown. The sample data in this figure is taken from *PICS, The Psychological Image Collection at Stirling* [146].

Hsu et al. [156] proposed a colour transformation in the *YCbCr* colour space to remove the non-linear dependence of skin colour on luminance. They considered the chroma *Cb* and *Cr* as functions of luminance Y and fitted piecewise linear boundaries to the skin colour cluster in the *YCbCr* colour space. In the following section a much simpler technique for modelling the non-linear dependence of skin tone on luminance than that proposed by Hsu et al. [156] is defined. The proposed skin colour model is shown to adequately capture the variation of skin tone under varying illumination. Unlike Hsu et al. this model is defined directly in the *RGB* colour space and requires the estimation of fewer model parameters.

Estimating the Nonlinear Relation between RGB Skin Colour and Luminance

It is attempted in this section to fit a $3D$ curve to the sample *RGB* skin colour data in figure 2.6 so as to establish the non-linear relation of *RGB* skin colour to that of luminance. The luminance component of colour in *RGB* space is commonly defined as,

$$Y = 0.3R + 0.59G + 0.11B. \tag{2.52}$$

A $3D$ curve in RGB space can be parameterised in terms of the luminance component Y as,

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \frac{1}{3} \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \begin{bmatrix} Y^3 \\ Y^2 \\ Y \end{bmatrix} \tag{2.53}$$

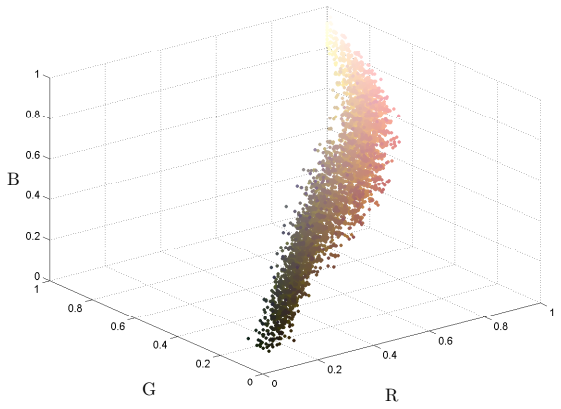


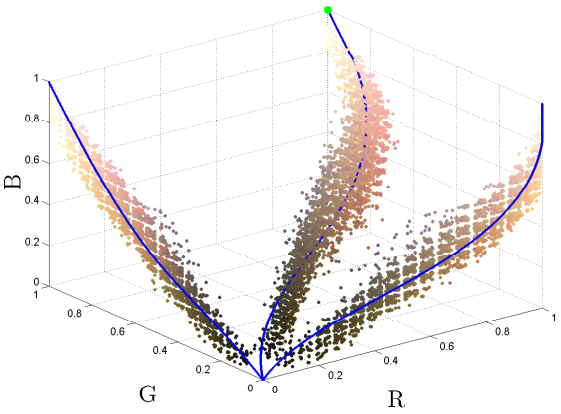
Figure 2.6: *Sample of skin colour pixels in RGB space captured under varying illumination obtained from the* PICS, The Psychological Image Collection at Stirling *[146].*

where the following constraints $\sum_j \frac{1}{3} q_{ij} = 1, \forall i$ are enforced on the coefficients to ensure that the curve endpoints are constrained to $[0,0,0]^T$ and $[1,1,1]^T$. The need to constrain the endpoints of the curve to $[0,0,0]^T$ and $[1,1,1]^T$ is to ensure that the 3D curve spans the full luminance range $[0,1]$ i.e. to ensure that $Y = 0$ for $R = G = B = 0$ and $Y = 1$ for $R = G = B = 1$. It is important to note that order of the polynomial in Y as defined by equation 2.53 was chosen in this analysis purely on the basis that it provided the best visual fitting to the sample RGB skin colour data.

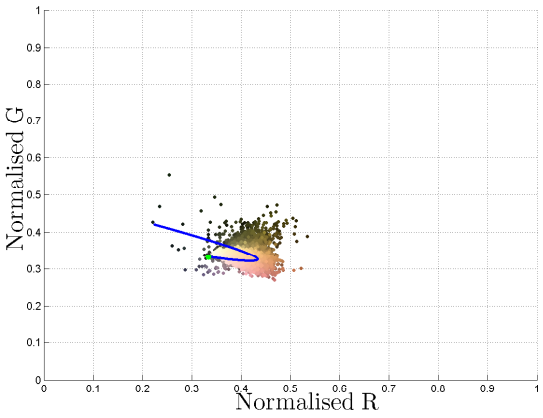
The result of fitting the curve of equation 2.53 to the sample RGB data through minimising the sum of the squared error between the sample RGB data and the parameterised curve is shown in figure 2.7a. Points along this curve correspond to a least squares estimate of skin colour in RGB space for varying luminance. Also seen in figure 2.7 is the estimated $3D$ curve in various different chrominance colour spaces such as the normalised R-G colour space in figure 2.7b, the $yCbCr$ colour space in figure 2.7d and the HSV colour space in figure 2.7c. It can be seen from these that although the sample skin colour is distributed over compact and dense regions in each, the nonlinear variation of skin colour with luminance is still evident.

Modelling the Skin Colour by Two Polynomial Projections in *RGB* Colour Space

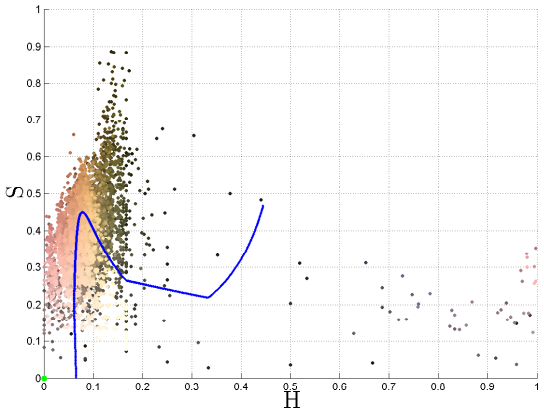
Although the estimated $3D$ curve obtained in the previous section could be used directly as a model of skin colour, it was found that the constraint that the end-points of the curve reside at $[0,0,0]^T$ and $[1,1,1]^T$, does not accurately reflect that observed in the sample skin data. Estimating both the curve coefficients and the end-points simultaneously is not straightforward. Although the end-points could be manually initialised this is undesirable. Instead, an alternative method is used which effectively estimates the best $3D$ curve in RGB space by two $2D$ projections. This is achieved by projecting the sample data of figure 2.6 onto the RG and RB planes and then fitting a polynomial to each of these to model the non-linear correlation between



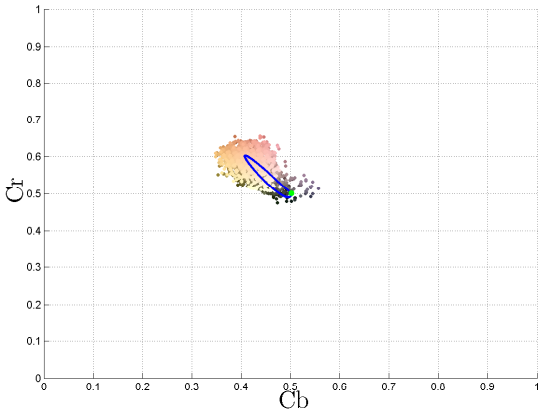
(a) *The estimated* 3D *curve (blue) of equation 2.53 with coefficients* $[q_{11}, q_{12}, q_{13}, q_{21}, q_{22}, q_{23}, q_{31}, q_{32}, q_{33}] = [-4.87, 5.98, 1.89, 1.75, -2.31, 3.56, 3.90, -3.91, 3.01]$ *fitted to the sample* RGB *skin colour data. Also show are the projections of the* 3D *plot onto the* RB *and* GB *planes.*



(b) *Sample skin colour data in normalised* R-G *colour space including the transformed* 3D *curve.*



(c) *Sample skin colour data in* HSV *colour space including the transformed* 3D *curve.*



(d) *Sample skin colour data in* yCbCr *colour space including the transformed* 3D *curve.*

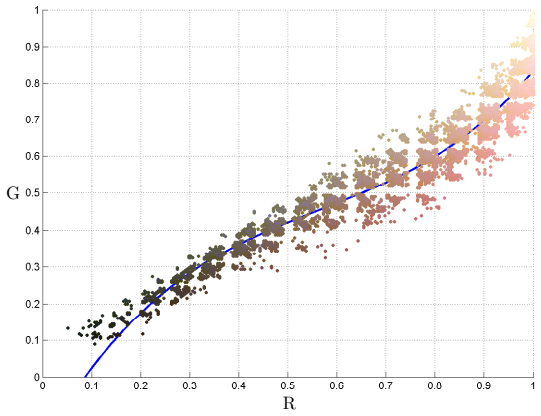
Figure 2.7: *Non-linear dependence of skin tone on luminance in different colour spaces*

the channels. More formally, two polynomials $f_G(R)$ and $f_B(R)$ are defined as,

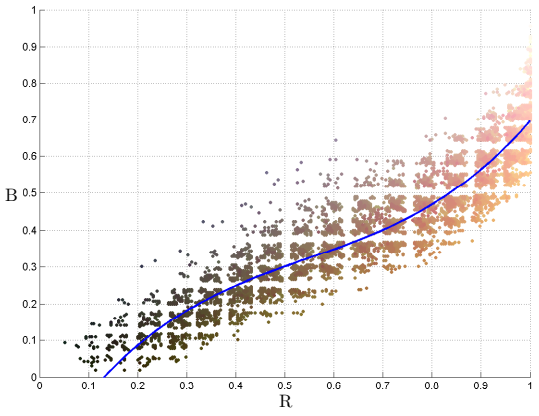
$$G = f_G(R) = a_n R^n + a_{n-1} R^{n-1} + ... + a_1 R + a_0, \tag{2.54a}$$

$$B = f_B(R) = b_n R^n + b_{n-1} R^{n-1} + ... + b_1 R + b_0. \tag{2.54b}$$

which model the G and B colour components of a skin pixel in relation to its R component. For both $f_G(R)$ and $f_B(R)$ a polynomial of order three with coefficients $a = [1.77, -2.92, 2.07, -0.23]$ and $b = [1.90, -3.23, 2.34, -0.18]$ was found to best fit the sample skin data. These cubic polynomials are shown in figure 2.8a and figure 2.8b.



(a) *RG plane and the cubic polynomial* $f_G(R)$ *(blue) with coefficients* $a = [1.90, -3.23, 2.34, -0.18]$.



(b) *RB plane and the cucbic polynomial* $f_B(R)$ *(blue) with coefficients* $b = [1.77, -2.92, 2.07, -0.23]$.

Figure 2.8: *Estimating the correlation between the R, G and B components of skin colour by two polynomials.*

The classification of a pixel $\mathbf{p}$ in an image $\mathbf{I}$ where $\mathbf{I}(\mathbf{p}) = [R, G, B]$ as skin or non-skin proceeds in a deterministic manner based on two conditions. The first condition is that the polynomial relations of equation 2.54 are satisfied as follows,

$$C_1(\mathbf{p}) = (|G - f_G(R)| < t_0) \ \cap \ (|B - f_B(R)| < t_1) \tag{2.55}$$

where typical values for the thresholds t_0 and t_1 are 0.06 and 0.08 respectively.

In addition to this, a condition on the dominance of the R component in skin colour is incorporated into the detection process. The usefulness of this colour cue has been previously reported [83, 180] and is defined by,

$$C_2(\mathbf{p}) = (R/G > \beta) \ \cap \ (R/B > \beta). \tag{2.56}$$

where $\beta > 1$. A typical value for β is 1.1.

Overall, the conditions defined in equation 2.55 and equation 2.56 define a deterministic pixel wise approximation of skin colour regions. Therefore, a skin colour mask for an image $\mathbf{I}$ may be defined at each pixel location $\mathbf{p}$ as,

$$\mathbf{S}(\mathbf{p}) = \begin{cases} 1 & C_1 = true, \ C_2 = true \\ 0 & \text{Otherwise.} \end{cases} \tag{2.57}$$

In the following analysis, the sample frame in figure 2.9a is used to compare the proposed skin colour detection algorithm against two existing techniques. The two existing techniques examined are the skin detection algorithm of Hsu et al. [156] and that proposed by Jones et al. [120]. These methods are chosen for comparison since they both aim to accurately detect

skin colour under varying illumination. The following discussion aims to highlight the main advantages of the proposed skin colour detection technique above these other approaches.

Figure 2.9a shows an example frame of a lecture recording from the CHIL database and the $3D$ scatter plot of the RGB values for this frame are shown in figure 2.9b. This sample frame is used in the following discussion to describe a typical scenario where the proposed algorithm results in improved skin colour detection beyond that of the existing detection techniques of Hsu et al. and Jones et al.

The method of Hsu et al. [156] as previously described, utilises the $YCbCr$ colour space and defines a model specifying a volume in this colour space which represents skin colour. This volume incorporates the distortion of skin colour due to changes in illumination. The success of Hsu's method for modelling skin colour under varying illumination relies on a colour balancing pre-process. The colour balancing procedure proceeds as follows. Firstly, a set of pixels $\mathbf{q}$ representing the top 5% of the luminance range is identified. Secondly, if the number of pixels in $\mathbf{q}$ is greater than 100 then a colour balancing operation is applied. The colour balancing operation linearly scales the colour value of each pixel in the image such that the average grey-scale value of the pixels in $\mathbf{q}$ corresponds to white in the employed colour space. This colour balancing operation is a key component in the success of the skin colour detection method of Hsu et al. This reliance on a colour balancing operation has been highlighted as a significant limiting factor in Hsu's algorithm. In particular it has been shown that colour balancing can actually reduce the ability of Hsu's algorithm to accurately detect skin regions [15]. One of the advantages of the proposed skin colour detection algorithm is that it does not rely on any pre-processing or colour correcting operations. It is therefore not limited in this regard when compared to Hsu's approach.

Jones et al. [120] approach the problem of modelling skin colour under varying illumination in a different manner. They aim to model the statistical distribution of skin colour and non-skin colours in the RGB colour space using Gaussian mixture models. The statistical colour models are trained on approximately 3 million images acquired from the world-wide-web (WWW). Since they approach skin colour modelling in the RGB colour space they do not make any assumption on the dependence of skin colour on luminance. As a result, their statistical skin colour model is flexible enough to capture any dependence of skin colour on luminance, including any non-linear relation. By using a Gaussian Mixture Model (GMM) to model skin colour, the method of Jones et al. effectively identifies clusters of skin colour which exist in the training data. One possible criticism of their approach is that they do not restrict the GMM in any way to ensure that the statistical model represents skin colour spanning the full luminance range. A possible failure arising from this is that if the training data only represents skin colour captured at discrete levels of illumination then the modes of the GMM will tend to only occupy these regions. Essentially, if the training data does not contain significant samples captured at low levels of illumination then they will not be adequately represented in the statistical model. Furthermore, the fitting of a GMM requires the pre-definition of a specific number of clusters. This can result in skin colour

only being adequately modelled at discrete levels of illumination. The proposed skin detection technique however does not have such a restriction and models skin colour continuously over the full range of illumination. Also, since the proposed model is deterministic it avoids problems where insufficient training data is available to statistically model skin colour at all possible levels of illumination.

Figure 2.10 presents a comparison between the proposed skin colour detection technique and that proposed by Hsu et al. [156] and Jones et al. [120]. The skin colour mask estimated with the proposed skin colour detection technique on the sample frame of figure 2.9a is presented in figure 2.10a. Figure 2.10b shows the results of skin detection on the same frame using the skin detection algorithm as proposed by Hsu et al [156] and figure 2.10c shows the result of skin detection using the technique of Jones et al. [120]. In each case the skin masks have been *cleaned* using equivalent open, close and hole filling morphological operations. From these figures it can be seen that both the proposed skin detection technique and the method of Hsu outperform the technique of Jones et al. for detecting skin colour under low illumination.

Overall, the proposed skin detection technique can be seen to perform similarly to the method of Hsu. However, Hsu's method appears to perform poorly at classifying non-skin regions under high illumination. This can be seen in figure 2.10b where significant wall regions about the presenter in the scene are incorrectly classified as skin.

The need to adequately model skin colour at low levels of illumination can be further supported by a direct comparison between the proposed detection technique and that of Jones et al. The estimated skin colour pixels of the sample frame using the proposed method in RGB space are shown in figure 2.10d. Shown in figure 2.10e is the region of the skin colour pixels estimated by the new approach which are classified as non-skin by the method of Jones et al. This demonstrates that the proposed detection method correctly classifies a greater number of skin colour pixels under low illumination than that of Jones et al. Motivated by these results, the presented skin colour detector is employed later in chapter 6 in an active speaker tracking application.

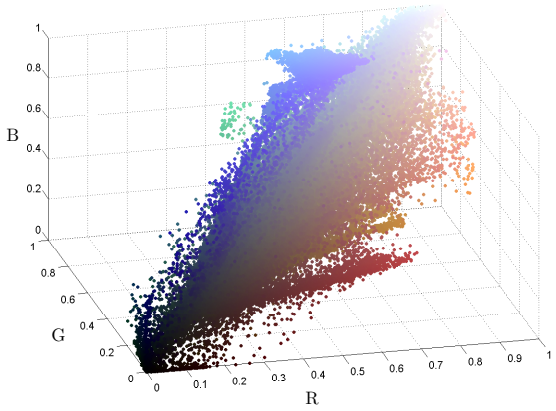
2.2.4 Camera Measurement Function

This section introduces the mathematical formulation of the projection of $3D$ points onto the image plane which will serve as analysis in later chapters. Extensive treatment of the presented material and also additional details which are not covered here can be found in the textbooks of O. Faugeras [138] and Hartley et al. [154].

2.2.4.1 The Pinhole Camera Model

The pinhole camera model is a simple model which describes image formation by central projection. By central projection in this context, it is meant that a point $\mathbf{x} \in \mathbb{R}^3$ in space is mapped onto a point $\mathbf{p} \in \mathbb{R}^2$ in the image plane such that $\mathbf{p}$ and $\mathbf{x}$ form a straight line through a point

(a) *Sample frame from the CHIL database.*



(b) *Scatter plot of sample frame (a) in the* RGB *colour space.*

Figure 2.9: *Sample frame from the CHIL database together with a scatter plot of the frame's pixels in the RGB colour space*

$\mathbf{C} \in \mathbb{R}^3$. An illustration of this relation is presented in figure 2.11. In this model the point $\mathbf{C}$ is known as the *camera centre* and defines the point through which all object to image rays are projected. The perpendicular distance between the image plane and the *camera centre* is known as the *focal length* and is denoted f. A basic pinhole camera model is therefore completely described by both $\mathbf{C}$ and f.

In essence, what the pinhole camera model represents is a set of well defined geometrical relationships in $3D$ space between object points and their corresponding imaged points on the image plane. Using this, a $\mathbb{R}^3 \mapsto \mathbb{R}^2$ central projection mapping can be defined which relates an object point $\mathbf{x}$ to its corresponding imaged point $\mathbf{p}$. This mapping can be defined as follows.

Consider the point $\mathbf{p} = [p_x, p_y]^T$ obtained by the projection of a $3D$ point $\mathbf{x} = [x, y, z]^T$ onto the image plane using the pinhole camera model defined by $\{\mathbf{C}, f\}$. This corresponds to the scenario as depicted in figure 2.11. Using the relation of similar triangles it can be seen that $\frac{p_x}{f} = \frac{x}{z}$ and $\frac{p_y}{f} = \frac{y}{z}$. Through this, the image point $\mathbf{p}$ is determined as,

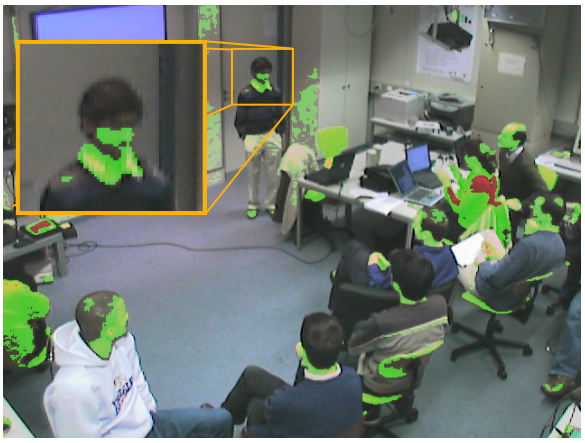
$$\begin{bmatrix} p_x \\ p_y \end{bmatrix} = \begin{bmatrix} f\frac{x}{z} \\ f\frac{y}{z} \end{bmatrix}. \tag{2.58}$$

The $\mathbb{R}^3 \mapsto \mathbb{R}^2$ central projection mapping for an arbitrary point $\mathbf{x} = [x, y, z]$ therefore is defined as,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} f\frac{x}{z} \\ f\frac{y}{z} \end{bmatrix}. \tag{2.59}$$

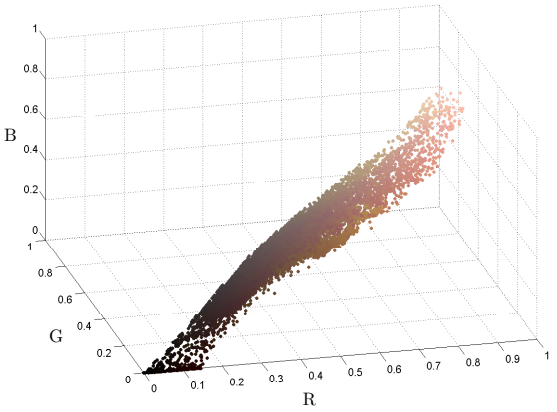
Until now we have examined the projection of $3D$ points to their $2D$ images in Euclidean

(a) *Extracted skin regions (Proposed Method)*

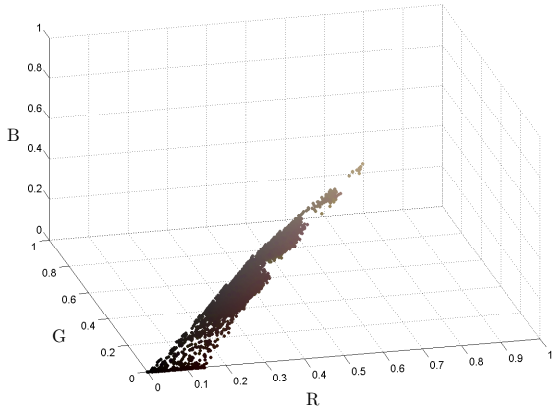


(b) *Extracted skin regions (Hsu et al. [156])*

(c) *Extracted skin regions (Jones et al. [120])*



(d) *Detected skin pixels using the proposed skin colour model which accounts for the non-linear dependence of skin tone on luminance.*



(e) *Colour region in the RGB colour space classified as skin by the proposed method but as non-skin by the method of Jones et al. [120].*

Figure 2.10: *Comparison of the proposed skin colour detection technique to that of Hsu et al. [156] and Jones et al. [120]*

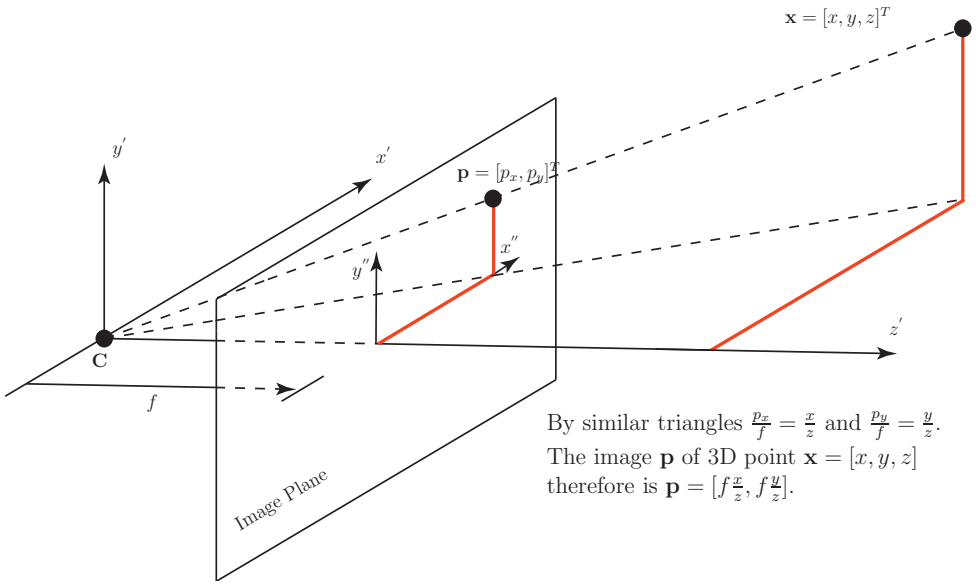


Figure 2.11: *The pinhole camera model defined by the focal length f and the camera centre* **C**. *The projection of a 3D point onto the image plane can be defined using the relation of similar triangles.*

space. There are concepts in multi-view geometry however which Euclidean geometry can not adequately describe. For instance, consider the case of an image point as $z \to 0$, i.e. $[f\frac{x}{0}, f\frac{y}{0}]$. This is an infinite point which is not defined in Euclidean space. Situations such as this motivate the need for a different geometric space in which such points can be represented. The approach in multi-view geometry is to extend the image plane from a Euclidean space $\mathbb{R}^2$ to a projective space $\mathbb{P}^2$ which enables the representation of such infinite points.

2.2.4.2 Projective Space and Homogeneous Coordinates

To introduce the concept of the projective space $\mathbb{P}^2$ we can consider again the scenario illustrated in figure 2.11, but instead consider the image point **p** as the point of intersection of the $3D$ line from **C** to **x** with that of the image plane. In examining this problem it can be assumed without loss of generality that the camera centre lies at the origin i.e. $\mathbf{C} = [0, 0, 0]^T$. In addition to this, the z axis is redefined as $w = \frac{z}{f}$. This simply corresponds to a scaled z axis and enforces that the image plane irrespective of f is always the plane defined by $w = 1$.

An intuitive way of thinking of the projective space $\mathbb{P}^2$ is that of consisting of the set of rays in $\mathbb{R}^3$ which project through the camera centre. Consider now, the parameterized form of the $3D$ line through the image point **p** as [69, pp. 158-159],

$$\begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} = \begin{bmatrix} kp_x' \\ kp_y' \\ kw \end{bmatrix} \tag{2.60}$$

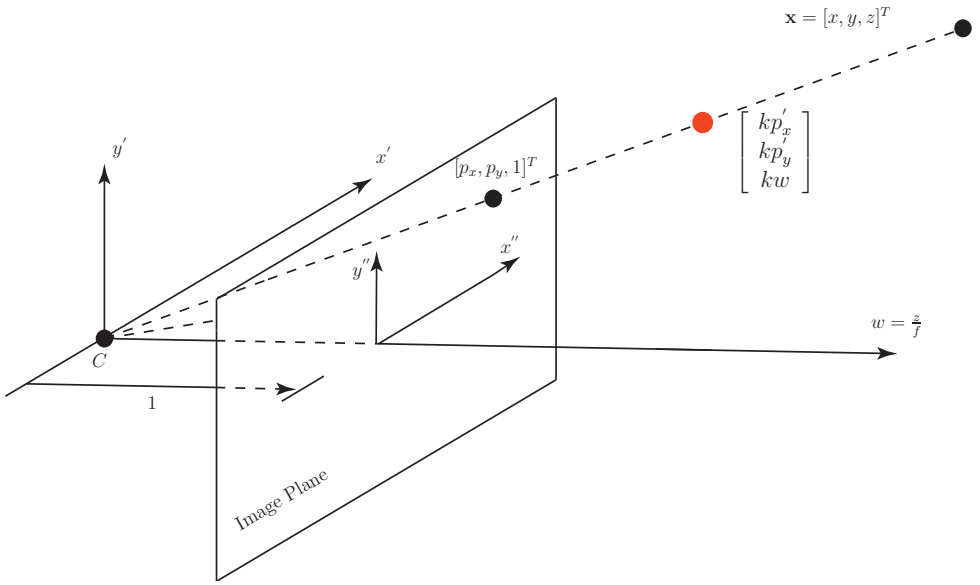


Figure 2.12: *Parameterized form of the 3D line through the point on the image plane* $[p_x, p_y, 1]$ *and the camera centre* $\mathbf{C}$*. Each* $[p'_x, p'_y, w]$ *for* $k \neq 0$ *defines a unique 3D line through the* $\mathbf{C}$*.*

which is illustrated in figure 2.12. For $k \neq 0$ the point $[kp'_x, kp'_y, kw]$ in equation 2.60 defines a unique $3D$ ray through $\mathbf{C}$. In this way, the point $[kp'_x, kp'_y, kw]$ can be seen to represent a unique point in $\mathbb{P}^2$. Solving for $k = \frac{1}{w}$, an equivalent representation of the $3D$ coordinate of the image point is defined in terms of w as,

$$\begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} = \frac{1}{w} \begin{bmatrix} p'_x \\ p'_y \\ w \end{bmatrix}. \tag{2.61}$$

The significance of equation 2.61 is that an image point can be represented by an ordered triple of real numbers $[p'_x, p'_y, w]$ in $\mathbb{P}^2$ knowing that the point on the image plane can be recovered by dividing by w i.e. $\mathbf{p} = [\frac{p_x}{w}, \frac{p_y}{w}]$. The point $[p'_x, p'_y, w]$ is said to be the *homogeneous coordinate* of the point $[p_x, p_y]$. In particular, given equation 2.60 any point $[kp'_x, kp'_y, kw]$ for $k \neq 0$ is equally a *homogeneous coordinate* of $[p'_x, p'_y]$. It is from this that the term homogeneous arises. From this it can be seen that infinite points in Euclidean space simply correspond to the points in $\mathbb{P}^2$ with *homogeneous coordinate* where $w = 0$. At this point, the notation $\tilde{\mathbf{p}}$ is introduced to refer specifically to the homogeneous representation of the a non-homogeneous point $\mathbf{p}$.

2.2.4.3 The Calibration Matrix

The benefit of using homogeneous coordinates is that the $\mathbb{R}^3 \mapsto \mathbb{R}^2$ mapping as defined in equation 2.59 which is non-linear in z can be defined as a linear $\mathbb{R}^3 \mapsto \mathbb{P}^2$ mapping. Using

homogeneous coordinates the mapping in 2.59 is redefined as,

$$\begin{bmatrix} p_x^{'} \\ p_y^{'} \\ w \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{2.62}$$

$$= \mathbf{Kx}. \tag{2.63}$$

The matrix $\mathbf{K}$ is known as the calibration matrix. In this formulation $\mathbf{K}$ represents the simplest case of a pinhole camera which assumes that the image coordinates are equally scaled in both the x and y axes. In general, this is not the case and the number of pixels per unit area m_x on the x axis is often different to the number of pixels per unit area m_y on the y axis. In addition to this, the origin of pixel measurements is often not the centre of the image plane $[c_x, c_y]$ but at a point $[x_0, y_0]$, known as the *principal point*. To account for this, image measurements are translated with respect to the *principal point* and scaled appropriately by the factors m_x and m_y on each axis. This is defined in the calibration matrix by,

$$\mathbf{K} = [m_x, m_y, 1] \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.64}$$

$$= \begin{bmatrix} m_x f & 0 & m_x c_x \\ 0 & m_y f & m_y c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.65}$$

$$= \begin{bmatrix} \alpha_x & 0 & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.66}$$

In the most general case, the calibration matrix is defined as,

$$\mathbf{K} = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.67}$$

where the *skew* parameter s is introduced to account for the case of non-rectangular pixels.

2.2.4.4 The Camera Projection Matrix

In examining the pinhole camera model it was assumed that the camera centre corresponded to the origin. Typically, $3D$ points will exist in some other *world coordinate* space and not that of the camera's coordinate system. This is not an issue since $3D$ points can easily be translated and rotated through into the coordinate space of the camera before projection onto the image

plane. The image point therefore becomes

$$\hat{\mathbf{p}} = \mathbf{KR}(\mathbf{x} - \mathbf{C}), \tag{2.68}$$

where $\mathbf{R}$ is a $3D$ rotation matrix representing the rotation between the world coordinate space and the coordinate space of the camera. A problem does arise however in that the $\mathbb{R}^3 \mapsto \mathbb{P}^2$ mapping of equation 2.68 is now non-linear. Similar to the manner in which image points can be represented by homogeneous coordinates, $\mathbf{x}$ can be defined in $\mathbb{P}^3$ with homogeneous coordinates $\tilde{\mathbf{x}} = [x, y, z, 1]$. Using this representation the non-linear $\mathbb{R}^3 \mapsto \mathbb{P}^2$ mapping of equation 2.68 can be redefined as a linear $\mathbb{P}^3 \mapsto \mathbb{P}^2$ mapping through

$$\tilde{\mathbf{p}} = \mathbf{KR}[\mathbf{I}_{3\times3}| - \mathbf{C}]\tilde{\mathbf{x}}, \tag{2.69}$$

where $\mathbf{I}_{3\times3}$ is a 3×3 identity matrix. The camera centre $\mathbf{C}$, rotation matrix $\mathbf{R}$ and the calibration matrix $\mathbf{K}$ can be encompassed in a single 3×4 matrix $\mathbf{P}$ called the *camera projection matrix* such that

$$\tilde{\mathbf{p}} = \mathbf{P}\tilde{\mathbf{x}}. \tag{2.70}$$

Given a point $\mathbf{x}$ and camera matrix $\mathbf{P}$ therefore, the homogeneous coordinate of the image point $\mathbf{p}$ is first obtained by equation 2.70 as,

$$\begin{bmatrix} p_x^{'} \\ p_y^{'} \\ w \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{2.71}$$

where a_{uv} is the (u, v) entry in $\mathbf{P}$. Secondly the non-homogeneous coorindate of the image point $\tilde{\mathbf{p}}$ is obtained by dividing by w as in 2.61 to give,

$$\begin{bmatrix} \frac{p_x^{'}}{w} \\ \frac{p_y^{'}}{w} \end{bmatrix} = \begin{bmatrix} \frac{a_{11}x+a_{12}y+a_{13}z+a_{14}}{a_{31}x+a_{32}y+a_{33}z+a_{34}} \\ \\ \frac{a_{21}x+a_{22}y+a_{23}z+a_{24}}{a_{31}x+a_{32}y+a_{33}z+a_{34}} \end{bmatrix}. \tag{2.72}$$

This is referred to as the camera measurement function in this thesis and its formulation in equation 2.72 is later employed in the analysis of localisation accuracy presented in chapter 4.

2.3 Final Comments

This chapter has introduced the various theory in the extraction of audio and video based features for localising active speakers. The use of audio information obtained from multiple microphones for use in localisation was described. It was seen that both time-delays estimates

at multiple microphones pairs, energy measurements and steered-response power can all be used to localise active speakers. The general issues relating to the acoustic environment which audio-based localisation techniques must contend with such as reverberation were discussed. Also, measures to characterise the level of reverberation in a room were presented.

Those techniques which employ time-delay estimation were seen to be fundamentally limited in terms of their accuracy by the source signal bandwidth, time-analysis window and SNR. It was also seen that two components of error exist in the time-delay estimation problem such as "small" errors which describe uncertainty locally about the true time-delay and "large" errors corresponding to anomalous TDEs. Introduced in this section was the time-delay measurement function defining the relationship between a speech source position to that of a time-delay observed at a pair of microphones. Also, the DOA measurement function was defined, establishing the relationship between a given speech source position and the DOA at a pair of microphones.

In addition to the overview of multi-channel audio-based features, various techniques for the detection of people in video sequences were considered, such as, background modelling and face detection. Feature-based face detection through skin colour modelling was proposed as the most applicable to the multi-view scenario where faces are observed at various poses and often at low resolution. A new skin colour model was introduced which was seen to demonstrate the ability to detect skin colour under low illumination. Finally, the camera measurement function was introduced describing how a *3D* point is projected onto the image place of a camera.

3

Joint Audio-visual Active Speaker Tracking

Combining audio and video features for tracking active speakers is a relatively new concept in the signal processing research community. It is clear that under ideal conditions it is possible to do this using microphone array techniques alone. Early approaches, attempted to track active speakers using the multi-channel audio-based features presented in section 2.1. As was detailed in the presentation, the necessary conditions for this to be possible are never realised. In the early development of audio-based localisation techniques, issues such as the reverberation phenomenon coupled with the challenge of discerning a speech source from among noise sources were quickly identified as significant barriers. Although considerable advances have been made over several decades [72], these issues still remain significant barriers. Researchers continue in their efforts to improve audio-based localisation performance and advancements are continuously being reported. Regardless of this, it is unlikely that solely audio-based active speaker tracking systems will ever meet a satisfactory level of performance for use in everyday applications. Even in favourable acoustic environments, their performance is inherently limited. This is the case because a person can only be located using such techniques when actively speaking.

Consequently, a strong inclination now exists in the research community towards the use of video with audio for tracking active speakers. This offers the potential to greatly improve tracking performance since it is not affected by reverberation which can cause audio-based localisation to fail and can be used to estimate a person's position when not actively speaking. Many basic visual features exist such as those discussed in section 2.2 that can be used to affirm likely speaker positions within a scene. However, video alone can not fully address the problem of detecting speaker activity. Visual information such as lip movements can indicate likely speech

activity, but the dominant information relating to such is contained within the audio data. At a very basic level, video can be used to indicate likely speaker positions and guide an audio-based tracking system for determining speech activity.

At a more sophisticated level, if every potential speaker could be detected using video data, the active speaker tracking task reduces to that of determining which of the known speakers in the scene is currently active. Determining the position of an active speaker from a small set of hypothesised positions is a much more approachable task than blind localisation using audio information alone. Nishiguchi et al. [170] use this approach in localising speaking students in a lecture room scenario. A set of known seated student positions are continuously monitored for occupancy through motion detection using temporal differencing on the video data. A histogram type analysis of TDEs built over occupied speakers positions is then used to evaluate which student is speaking.

Similar systems have been proposed for guiding beamforming based localisation techniques. As was previously stated in section 2.1 both the SRP-PHAT and the use of a delay-and-sum beamformer can be computationally expensive to implement if the space to be searched for an active speaker is large. Reducing the search space to a small number of hypothesised locations using video data makes them much more tractable. This strategy is a popular approach in HCI and desktop video-conferencing systems. The work of Bub et al. [183] presents a beamformer which is steered towards a speaker located using skin colour, motion and face-shape based visual cues. Similar video based methods for steering beamformers are described in [13, 14, 116]. Also, the use of face detection is made in [80] to reduce the SRP-PHAT search space in speaker localisation which in turn is used to steer a beamformer.

Unfortunately, video-based tracking is not without its own challenges. Illumination changes, occlusions and low lighting conditions are just some of the problems which currently make the accurate detection of people in video difficult. It is the nature of both audio and video as sensing modalities however that they tend to fail independently. Therefore, in the same way that video-based localisation can be used to compensate for the limitations of audio-based tracking; audio-based tracking can be used to compensate video.

Audio can also facilitate video localisation in cases where a video-based estimate is not possible. This can occur where the speaker is not in the camera's field of view. The alternative solution in a purely video-based approach would be to complete an exhaustive visual search of the tracking space which can be computationally prohibitive. Audio guided systems for visual speaker localisation have been proposed for automated video-conferencing applications [23, 24, 61, 62] and HCI systems [19]. In these speaker tracking systems audio-based localisation is used to define a rough estimate of a speaker's position which is then refined using video data.

Audio-guided and video-guided speaker localisation systems however, do not fully utilise the complementary nature of both the audio and video modalities. In particular, they do not consider how audio and video should be combined when location estimates from both modalities are available. If the tracking problem can be modelled with accuracy, then fusing multiple

measurements from different sensors can result in an overall improved positional estimate [188]. Whether such strategies can be applied using audio and video sensors for tracking people is a question which has recently interested the signal processing research community. In this chapter the tracking techniques which facilitate the incorporation of both audio and video for tracking active speakers are explored.

Aiming to cover all the relevant video-based and audio-based tracking literature relating to active speaker tracking would be impractical. It is felt that existing literature provides sufficient treatment in relation to video-based tracking [10, 90, 92]. For this reason, this chapter introduces the active speaker tracking problem initially from an audio-based perspective and useful tracking techniques for this problem are presented. The joint audio-video based problem is examined afterwards by considering how audio-based tracking methods can be extended to include visual information. In this regard, the review follows the trend observed in the research community in recent years towards active speaker tracking; from purely audio-based tracking, to a joint audio-visual perspective.

3.1 Bayesian State Sequence Estimation

Many speaker localisation and tracking tasks can be stated as problems of estimating a hidden state $\mathbf{x}_k$ over some time duration $k = 0, .., K$ based on a set of observations $\mathbf{y}_{0:K} = \{\mathbf{y}_0, .., \mathbf{y}_K\}$. In the context of the work described in this thesis, we can consider $\mathbf{x}_k \in \mathbb{R}^p$ as the hidden state of an active speaker, such as for instance, the speaker's position in *3D* space ($p = 3$). However, the general definition of $\mathbf{x}_k$ as a p-dimensional vector is considered to be applicable to cases where $\mathbf{x}_k$ may relate to other hidden states of interest such as perhaps, the time-delay relating to an active speaker in the audio-domain ($p = 1$).

Applying a Bayesian analysis to the state estimation problem, $\mathbf{x}_k$ can be considered as a random variable. The task can then be defined as that of estimating the sequence of states $\mathbf{x}_{0:K}$ for times $k = 0, .., K$ from the sequence of observations $\mathbf{y}_{0:K}$. From a Bayesian perspective, this translates to the problem of determining the *posterior* distribution $p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K})$. Using Bayes' rule the *posterior* distribution is defined as [5],

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K}) = \frac{p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K})p(\mathbf{x}_{0:K})}{\int p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K})p(\mathbf{x}_{0:K})d\mathbf{x}_{0:K}}. \tag{3.1}$$

Since the *posterior* distribution is completely defined by both $p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K})$ and $p(\mathbf{x}_{0:K})$ the relation in equation 3.1 can be re-written,

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K}) \propto p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K})p(\mathbf{x}_{0:K}). \tag{3.2}$$

The presented analysis refers to two specific scenarios in relation to the state estimation problem. The first is the *online* estimation problem where the *posterior* of equation 3.1 is

estimated incrementally at each time-step k based on the available observations $\mathbf{y}_{0:k}$. This problem can also be referred to as a *filtering* problem where estimating $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ is of interest. The second scenario corresponds to the *off-line* case where $\mathbf{x}_k$ is determined at each time step k using the complete set of observations $\mathbf{y}_{0:K}$. This is often referred to as the *smoothing* estimation problem which addresses the issue of determining the distribution $p(\mathbf{x}_k|\mathbf{y}_{0:K})$.

Since both $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ and $p(\mathbf{x}_k|\mathbf{y}_{0:K})$ are marginals of the *posterior* distribution in equation 3.1, in simple estimation problems they can be determined by integrating over the nuisance states. In most practical estimation problems, this requires complex high dimensional integrals making their estimation intractable [5]. Therefore, practical applications of Bayesian techniques to state estimation problems often require *prior* assumptions and approximate models of both the state $\mathbf{x}_k$ and measurements $\mathbf{y}_k$ so as to restrict the form of $p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K})$.

To this end, in state sequence estimation problems it is common to assume a Markovian representation of $\mathbf{x}_k$ and model the evolution of the state as,

$$\mathbf{x}_k = \mathbf{T}_k(\mathbf{x}_{k-1}, .., \mathbf{x}_{k-r}, \mathbf{v}_k). \tag{3.3}$$

where the function $\mathbf{T}_k$ is known as the *process model* and $\mathbf{v}_k$ defines the process noise reflecting uncertainty in the model. In addition to this, it is common to define the relationship between the state $\mathbf{x}_k$ and its observation $\mathbf{y}_k$. This can be defined through a *measurement model* as,

$$\mathbf{y}_k = \mathbf{H}_k(\mathbf{x}_k, \mathbf{w}_k) \tag{3.4}$$

where $\mathbf{H}_k$ defines the *measurement function* and $\mathbf{w}_k$ is the noise observed on the observations.

3.1.1 Recursive Bayesian Filter

The recursive Bayesian filter addresses the *online* estimation problem and enables a recursive determination of $p(\mathbf{x}_k|\mathbf{y}_{0:k})$. This filter arises by modelling $\mathbf{x}_k$ as a first order Markov process. This corresponds to cases where the process model of equation 3.3 is defined for $r = 1$. Under this assumption, equation 3.3 and equation 3.4 define two important relations in the definition of the recursive Bayesian filter. These are,

$$p(\mathbf{x}_k|\mathbf{x}_{0:k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}), \tag{3.5}$$

which asserts the dependence of $\mathbf{x}_k$ on $\mathbf{x}_{k-1}$ only and

$$p(\mathbf{y}_k|\mathbf{x}_{0:k}) = p(\mathbf{y}_k|\mathbf{x}_k), \tag{3.6}$$

which defines the current observation $\mathbf{y}_k$ as dependent only on the current state $\mathbf{x}_k$. The probability density function (*pdf*) $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is known as the *state transition pdf* and defines the evolution of the state from time-step $k-1$ to k. In the specific case of tracking the posi-

tional state of an active speaker, the state transition density defines the dynamical model of the speaker's motion. The *pdf* $p(\mathbf{y}_k|\mathbf{x}_k)$ defines a probabilistic model of the measurement process and is known as the measurement *likelihood function*. Given this, $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ can be determined recursively at each time instance in a two-step *predict* and *update* operation given by [54],

$$\text{PREDICT}: \quad p(\mathbf{x}_k|\mathbf{y}_{0:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{0:k-1})d\mathbf{x}_{k-1} \tag{3.7a}$$

$$\text{UPDATE}: \quad p(\mathbf{x}_k|\mathbf{y}_{0:k}) \propto p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{0:k-1}). \tag{3.7b}$$

Again, in most cases where the dimensionality or degrees of freedom of $\mathbf{x}_k$ is large, the integral of in the prediction step is intractable. There are however certain cases where the recursive Bayesian filter can be directly applied such as the case where the state is constrained to a Gaussian linear state space and both the process noise $\mathbf{v}_k$ and measurement noise $\mathbf{w}_k$ are uncorrelated and drawn from Gaussian distributions. Under these assumptions, the posterior $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ is Gaussian for all times k and the state can be estimated using the Kalman Filter (KF) [194, Chapter 5]. The KF is described in the following section however for a more detailed treatment the reader is referred to the tutorial by Welch et al. [65].

3.1.2 The Kalman Filter (KF)

The Kalman Filter (KF) can be applied to estimating $\mathbf{x}_k$ in cases where both $\mathbf{v}_k$ and $\mathbf{w}_k$ are zero-mean uncorrelated Gaussian random processes and the process model $\mathbf{T}_k$ in equation 3.3 and measurement model $\mathbf{H}_k$ in equation 3.4 are linear functions. In this case $\mathbf{x}_k$ and $\mathbf{y}_k$ are defined by,

$$\mathbf{x}_k = \mathbf{A}_k\mathbf{x}_{k-1} + \mathbf{v}_k \tag{3.8a}$$

$$\mathbf{y}_k = \mathbf{G}_k\mathbf{x}_k + \mathbf{w}_k \tag{3.8b}$$

where $\mathbf{A}_k$ is the state transition matrix and $\mathbf{G}_k$ is the measurement matrix. Under these assumptions the predict and update recursion of equation 3.7 becomes,

$$\text{PREDICT:} \quad p(\mathbf{x}_k|\mathbf{y}_{0:k-1}) = \mathcal{N}(\mathbf{x}_k; \hat{\mathbf{x}}_{k|k-1}, \mathbf{P}_{k|k-1}) \tag{3.9a}$$

$$\text{UPDATE:} \quad p(\mathbf{x}_k|\mathbf{y}_{0:k}) = \mathcal{N}(\mathbf{x}_k; \hat{\mathbf{x}}_{k|k}, \mathbf{P}_{k|k}) \tag{3.9b}$$

where $\hat{\mathbf{x}}_{k|k-1}$ and $\mathbf{P}_{k|k-1}$ are the *a priori* state estimate and state error covariance respectively and $\hat{\mathbf{x}}_{k|k}$ and $\mathbf{P}_{k|k}$ are their corresponding *a posteriori* estimates. The notation $\mathcal{N}(x; \mu, \sigma^2)$ refers to a normal distribution over x with mean μ and covariance σ^2. The KF is heavily reliant on the fact that the state estimate $\hat{\mathbf{x}}_{k-1|k-1}$ and error covariance matrix $\mathbf{P}_{k-1|k-1}$ are linearly transformable [164]. That is, $\hat{\mathbf{x}}_{k-1|k-1}$ can be propagated through the linear process model by $\mathbf{A}_k\hat{\mathbf{x}}_{k-1|k-1}$ and the error covariance $\mathbf{P}_{k-1|k-1}$ by $\mathbf{A}_k\mathbf{P}_{k-1|k-1}\mathbf{A}_k^T$. Through this the prediction

equations are defined as [194],

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}_k \hat{\mathbf{x}}_{k-1|k-1} \tag{3.10a}$$

$$\mathbf{P}_{k|k-1} = \mathbf{A}_k \mathbf{P}_{k-1|k-1} \mathbf{A}_k^T + \mathbf{Q}_k \tag{3.10b}$$

where $\mathbf{Q}_k$ is the covariance of the process noise $\mathbf{v}_k$. The complete algorithm defines the set of update equations as [194],

$$\boldsymbol{\nu}_k = \mathbf{y}_k - \mathbf{G}_k \hat{\mathbf{x}}_{k|k-1} \tag{3.11a}$$

$$\mathbf{S}_k = \mathbf{G}_k \mathbf{P}_{k|k-1} \mathbf{G}_k^T + \mathbf{R}_k \tag{3.11b}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{G}_k^T \mathbf{S}_k^{-1}. \tag{3.11c}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \boldsymbol{\nu}_k \tag{3.11d}$$

$$\mathbf{P}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{G}_k) \mathbf{P}_{k|k-1} \tag{3.11e}$$

where $\boldsymbol{\nu}_k$ is known as the innovation sequence which is the difference between the observation and its prediction, $\mathbf{S}_k$ is the covariance of the innovation sequence, $\mathbf{K}_k$ is the Kalman gain and $\mathbf{R}_k$ is the covariance of the measurement noise $\mathbf{w}_k$.

The Motion Modelling Problem

There is a prevalent concern in using the KF for positional tracking which warrants much discussion. This relates to that of choosing suitable models for motion. In relation to the Bayesian formulation this translates to defining the state transition probability density $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ to most appropriately model the dynamical evolution of the state $\mathbf{x}_k$.

Many suitable linear motion models for use with the KF exist, such as the Constant Position (CP), Constant Acceleration (CA) and Constant Velocity (CV) motion models as presented in Table 3.1 [190]. When the problem is that of tracking people, it can be particularly difficult to assign the observed motion to any one of these models. In any typical tracking scenario, a person will transgress over many different states of motion. With the KF this is of particular concern since in its standard form it is restricted to linear state evolution models. It is important to know therefore from the set of possible linear motion models that the correct model is being used.

Divergence in the Kalman Filter

Under optimal tracking conditions, the innovations sequence of equation 3.11a can be shown to be white and orthogonal [194, Chapter 5]. As a consequence of this, any modelling inaccuracies propagate in the innovation sequence indicating divergence in KFs. Divergence in the Kalman filter occurs as either true divergence, where the errors become unbounded, or apparent divergence where finite degradations are observed in the filtered state estimates [55]. Inaccurate

motion modelling results in apparent divergence where the state does not evolve according to the assumed state model of equation 3.8a but rather to some true linear process

$$\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{A}}_k \tilde{\mathbf{x}}_k + \tilde{\mathbf{v}}_k \tag{3.12}$$

such that the true measurements become,

$$\tilde{\mathbf{y}}_{k+1} = \mathbf{G}_k \tilde{\mathbf{x}}_k + \tilde{\mathbf{w}}_k. \tag{3.13}$$

It can be seen from equation 3.11a that where equation 3.12 represents the true state model, the innovation sequence does not represent the true innovation. Instead, the innovation in the case of inaccurate motion modelling can be obtained by replacing $\mathbf{y}_{k+1}$ in equation 3.11a by $\tilde{\mathbf{y}}_{k+1}$. The actual innovation therefore is [148],

$$\tilde{\boldsymbol{\nu}}_{k+1} = \tilde{\mathbf{y}}_{k+1} - \mathbf{G}_{k+1} \hat{\mathbf{x}}_{k+1|k}. \tag{3.14}$$

By substituting this expression for the true innovation into equation 3.11d, the true error in the state estimate becomes,

$$\mathbf{e}_{k+1|k+1} = \tilde{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}_{k+1|k+1}, \tag{3.15}$$

which is found, after some manipulation to be [39],

$$\mathbf{e}_{k+1|k+1} = \boldsymbol{\phi}_{k+1} \mathbf{A}_k \mathbf{e}_{k|k} + \boldsymbol{\phi}_{k+1} \Delta\mathbf{A}_k \tilde{\mathbf{x}}_k + \boldsymbol{\phi}_{k+1} \tilde{\mathbf{v}}_k - \mathbf{K}_{k+1} \mathbf{w}_k \tag{3.16}$$

where $\boldsymbol{\phi}_{k+1} = [\mathbf{I} - \mathbf{K}_{k+1}\mathbf{G}_{k+1}]$ and $\Delta\mathbf{A}_k = \tilde{\mathbf{A}}_k - \mathbf{A}_k$. In expanding equation 3.14, it can be seen that all modelling errors manifest in the innovation sequence i.e.

$$\tilde{\boldsymbol{\nu}}_{k+1} = \mathbf{G}_{k+1} \mathbf{A}_k \mathbf{e}_{k|k} + \mathbf{G}_{k+1} \Delta\mathbf{A}_k \tilde{\mathbf{x}}_k + \mathbf{G}_{k+1} \tilde{\mathbf{v}}_k + \mathbf{w}_{k+1}. \tag{3.17}$$

Computer Simulated Tracking Problem

The class of motion models as in table 3.1 are considered in this section. These different models are considered in estimating the evolution of a state vector $\mathbf{x}_k = [x_k, \dot{x}_k, \ddot{x}_k]$ where x_k, $\dot{x}_k$ and $\ddot{x}_k$ denote *1D* position, velocity and acceleration components. In relation to equation 3.17, it can be seen that if $E[\mathbf{w}_k]$ is zero mean then the error due to an inaccurate motion model propagates in the expected value of the innovation sequence through $\Delta\mathbf{A}_k$ and $\mathbf{v}_k$. In this simulated case, $\mathbf{G}_k E[\mathbf{v}_k]$ is the change in acceleration over the time step T. Over periods of constant acceleration therefore, $\mathbf{G}_k E[\mathbf{v}_k] = 0$ and the expected value of the innovation sequence in the simulated case

Model	CP	CV	CA
$\mathbf{A}_k$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & T & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & T & \frac{T}{2}^2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix}$
$\mathbf{G}_k$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Table 3.1: Transition matrices for a constant position model CP, constant velocity model CV and constant acceleration model CA.

becomes,

$$\begin{aligned} E\left[\tilde{\boldsymbol{\nu}}_{k+1}\right] &= \mathbf{G}_{k+1}\mathbf{A}_k E\left[\mathbf{e}_{k|k}\right] + \mathbf{G}_{k+1}\Delta\mathbf{A}_k\tilde{\mathbf{x}}_k + \mathbf{G}_{k+1}E\left[\mathbf{v}_k\right] \\ &= \mathbf{G}_{k+1}[\mathbf{A}_k E[\mathbf{e}_{k|k}] + \Delta\mathbf{A}_k\mathbf{x}_k + E[\mathbf{v}_k]]. \end{aligned} \tag{3.18}$$

The following examines $E\left[\boldsymbol{\nu}_k\right]$ in tracking the simulated motion using the CP, CV and CA models where the actual motion observed is CA.

- Using the CP Motion Model, $\Delta\mathbf{A}_k = \begin{bmatrix} 0 & T & \frac{T}{2}^2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix}$

$$E\left[\tilde{\boldsymbol{\nu}}_{k+1}\right] = E\left[\mathbf{e}_{k|k}\right] + T\dot{x}_k + \frac{T}{2}^2\ddot{x}_k \tag{3.19}$$

- Using the CV Motion Model, $\Delta\mathbf{A}_k = \begin{bmatrix} 0 & 0 & \frac{T}{2}^2 \\ 0 & 0 & T \\ 0 & 0 & 1 \end{bmatrix}$

$$E\left[\tilde{\boldsymbol{\nu}}_{k+1}\right] = E\left[\mathbf{e}_{k|k}\right] + \frac{T}{2}^2\ddot{x}_k \tag{3.20}$$

- Using the CA Motion Model, $\Delta\mathbf{A}_k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$$E\left[\tilde{\boldsymbol{\nu}}_{k+1}\right] = E\left[\mathbf{e}_{k|k}\right] \tag{3.21}$$

It can be seen from equation 3.19 that using a CP motion model over periods of constant acceleration ($\ddot{x}$ constant), the expected value of the innovation sequence increases linearly. Equation 3.20 reveals that a CV motion model results in an offset in the expected value of the in-

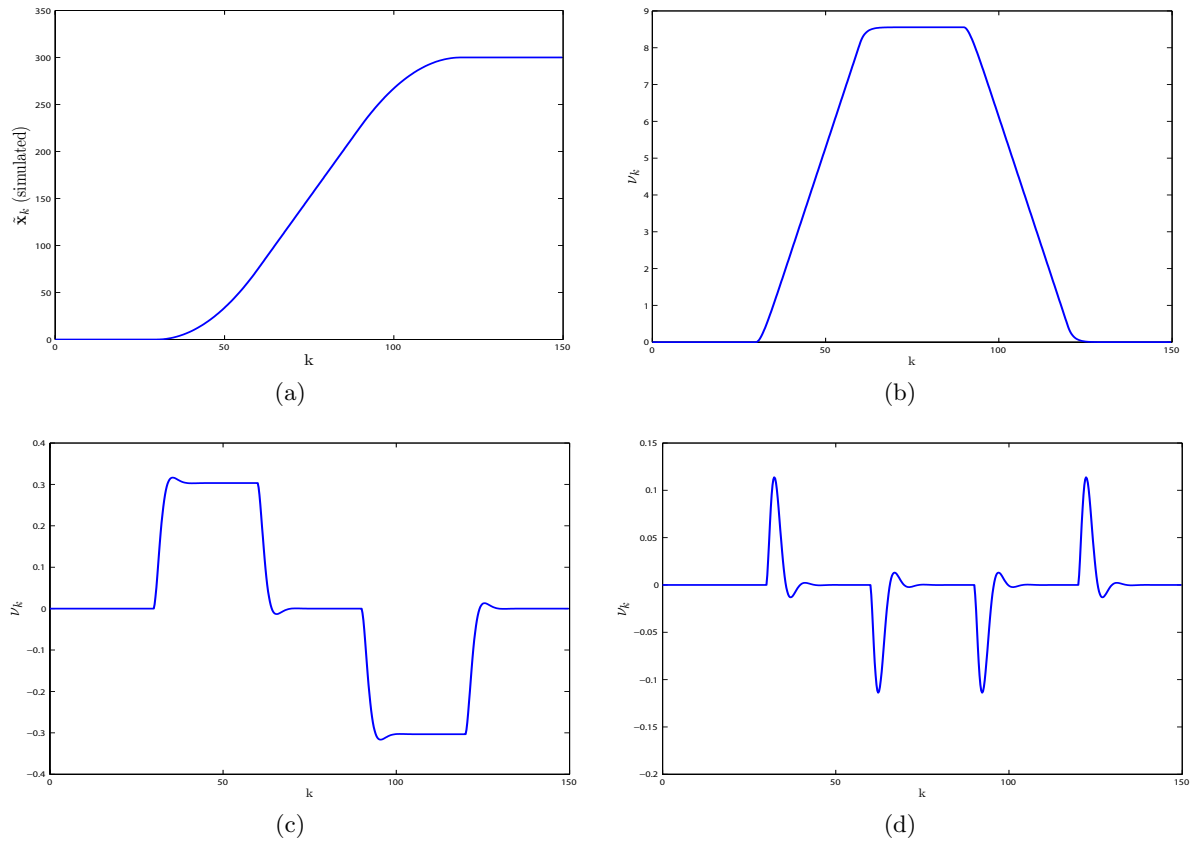


Figure 3.1: *Innovation sequences resulting in tracking the observed motion in (a). The resulting innovation sequence using the CP, CV, CA motion models are shown in (b), (c) and (d) respectively. The simulated motion is that of a person starting at an initial stationary position at* $k = 0$ *and accelerating from* $k = 30$ *to* $k = 60$*, maintaining a constant velocity (5m/s) from* $k = 60$ *to* $k = 90$ *and then decelerating back to a stationary position at* $k = 120$*. (*$T = 0.25s$*).*

novation sequence. As the CA motion model estimates all states of the simulated motion it is the optimal model and it is seen in equation 3.21 to be zero mean. Similarly, over periods of constant velocity ($\dot{x}$ constant) both the CV and CA motion models are zero mean and an offset is observed in $E[\boldsymbol{\nu}_k]$ where the CP model is used. This theoretical analysis is shown to correspond directly to the observed innovation sequence in a simulated tracking example as presented in figure 3.1. In this example the simulated tracking problem is that of a person starting from an initial stationary position to a state of constant velocity and then decelerating back to a stationary position.

It is seen therefore that continually monitoring the statistical properties of the innovation sequence is critical for detecting modelling errors and necessary to ensure accurate performance. This is the basis of many existing methods of divergence detection and control in Kalman filters [51]. The present analysis shows that divergence trends in the innovation sequence can be used in determining the most suitable linear motion model for a set of available linear models.

3.1.3 Kalman Filters for Audio-based Tracking

Since localisation estimates obtained using audio data can be noisy, the KF provides a simple, effective and easily implementable tool for the filtering of the positional estimates. In its standard form it can not be directly used in the case of a nonlinear measurement function as is the case in time-delay estimation and direction-of-arrival tracking. It can be used if the measurements are assumed to be an audio-based positional estimate obtained through some closed form estimate using TDEs or otherwise. In this case it can be used for the spatial filtering of location estimates [34].

The KF does not perform optimally if the assumption of linear motion is violated or if an incorrect motion model is used. A basic analysis of the innovation can be used to monitor performance and provide some insight into choosing the best model. Adaptive motion models are therefore essential to ensure accurate tracking. Extending the KF to the case of adaptive multiple motion models as in [35], is straightforward.

In many cases the assumption of linear motion however is too restrictive. In this case variants of the KF such as the Extended Kalman Filter (EKF), Iterated Extended Kalman Filter (IEKF) and Unscented Kalman Filter (UKF) exist which can address this. Each of these can handle cases where either/both the process function $\mathbf{T}_k$ or measurement function $\mathbf{H}_k$ is non-linear. As was seen in section 2.1.3.1 the time-delay measurement function of equation 2.25 and DOA measurement function of equation 2.26 are non-linear. Variants of the KF which can account for this are of particular interest in audio-based tracking.

Extended Kalman Filter (EKF)

In many cases of a nonlinear process function $\mathbf{T}_k$ or measurement function $\mathbf{H}_k$, a local linearisation of the functions may be sufficient. This is the approach of the EKF. In the case where both of the functions in $\mathbf{T}_k$ and $\mathbf{H}_k$ are nonlinear the extended Kalman filter linearises both functions such that $\nabla\mathbf{T}_k = \frac{\partial \mathbf{T}_k}{\partial \mathbf{x_{k-1}}}|_{\hat{\mathbf{x}}_{k-1|k-1}}$ and $\nabla\mathbf{H}_k = \frac{\partial \mathbf{H}_k}{\partial \mathbf{x}_k}|_{\hat{\mathbf{x}}_{k|k-1}}$ replace $\mathbf{A}_k$ and $\mathbf{G_k}$ respectively in equation 3.11. This equates to a first order linearisation of the functions $\mathbf{T}_k$ at the previous state estimate $\hat{\mathbf{x}}_{k-1|k-1}$ and a first order linearisation of $\mathbf{H}_k$ at the *a priori* state estimate $\hat{\mathbf{x}}_{k-1|k}$. The EKF however is not restricted to a first order linearisation and where a higher order estimate is necessary it may be extended to the second order case [193]. The EKF through linearising the non-linear measurement function has enabled the simplicity of the KF to be applied to the direction of arrival audio-based tracking problem [136].

Iterated Extended Kalman Filter (IEKF)

The Iterated Extended Kalman Filter (IEKF) is a variant of the EKF which is aimed at improving the state estimate by re-iterating the update stage of the EKF. The EKF determines the partial derivative of the measurement function $\mathbf{H}_k = \frac{\partial \mathbf{H}_k}{\partial \mathbf{x}_k}|_{\hat{\mathbf{x}}_{k|k-1}}$ at the *a priori* state estimate $\hat{\mathbf{x}}_{k|k-1}$. After the update step of the EKF however, an improved *posterior* state estimate $\hat{\mathbf{x}}_{k|k}$ is

available. The IEKF re-evaluates the derivative of the measurement function at this new estimate $\hat{\mathbf{x}}_{k|k}$ i.e. $\mathbf{H}_k = \frac{\partial \mathbf{H}_k}{\partial \mathbf{x}_k}|_{\hat{\mathbf{x}}_{k|k}}$ and the update filtering step is repeated. Since at each successive update a *better posterior* state estimate is expected, the IEKF repeats this process for a fixed number of iterations or until no further improvement is observed. This iterative process of the IEKF reduces the error introduced by linearising the measurement function. The extension of the EKF to an IEKF is trivial and has been proposed for TDE-based tracking in preference to the basic EKF [184].

Unscented Kalman Filter (UKF)

The Unscented Kalman Filter (UKF) [167] is also a variant of the EKF which can handle nonlinear process and measurement functions. It achieves this not by linearisation but by propagating the state and error covariance using the *unscented transform*. The basic strategy of the unscented transform is to deterministically define a set of points $\mathcal{X}^i_{k-1|k-1}$, $i = [0, .., 2p]$ known as *sigma points* with associated weights $\mathcal{W}^i_{k-1|k-1}$. These points are chosen to appropriately estimate the state $\hat{\mathbf{x}}_{k-1|k-1}$ and error covariance $\mathbf{P}_{k-1|k-1}$ through,

$$\hat{\mathbf{x}}_{k-1|k-1} = \sum_{i}^{2p} \mathcal{W}^i_{k-1|k-1} \mathcal{X}^i_{k-1|k-1} \tag{3.22a}$$

$$\mathbf{P}_{k-1|k-1} = \sum_{i}^{2p} \mathcal{W}^i_{k-1|k-1} [\mathcal{X}^i_{k-1|k-1} - \hat{\mathbf{x}}_{k-1|k-1}][\mathcal{X}^i_{k-1|k-1} - \hat{\mathbf{x}}_{k-1|k-1}]^T. \tag{3.22b}$$

Once defined, the prediction stage consists of propagating the sigma points through the process function to determine the set of transformed sigma points $\mathcal{X}^i_{k|k-1} = \mathbf{T}_k(\mathcal{X}^i_{k-1|k-1})$. Then by applying equation 3.22 to the set of transformed points, the *a priori* state estimate $\hat{\mathbf{x}}_{k|k-1}$ and error covariance $\mathbf{P}_{k|k-1}$ are determined in the prediction step. The update procedure is the same as that of the standard KF however the unscented transform is also used at this step in propagating $\hat{\mathbf{x}}_{k|k-1}$ and $\mathbf{P}_{k|k-1}$ through the nonlinear measurement function where necessary.

In comparison to the linearisation approach of the EKF, Julier et al. show the unscented transform to be more accurate than linearisation in cases where function is highly non-linear [164]. Additionally, they claim that the unscented transform has significant computational benefits over that of the standard EKF. The application of the UKF to audio-based tracking has been reported in [161] where they compare the use of the UKF to that of the EKF. They concluded in their analysis that the UKF only marginally outperforms the EKF and with almost equivalent computational burden.

3.1.4 The Particle Filter

The Particle Filter (PF) is a sequential Monte Carlo method for estimating the *posterior* $p(\mathbf{x}_k|\mathbf{y}_{0:k})$. This is achieved by assuming the $p(\mathbf{y}_k|\mathbf{x}_k)$ to be sufficiently well approximated

by a finite number N_s of support points $\mathbf{x}_k^i$, $i = 1, ..., N_s$. These points are associated with weights w_k^i such that an approximation of $p(\mathbf{y}_k|\mathbf{x}_k)$ can be obtained through [160],

$$p(\mathbf{x}_k|\mathbf{y}_{0:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{3.23}$$

The weights are determined by the process of importance sampling and are recursively estimated through [160],

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{y}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{y}_{0:k})} \tag{3.24}$$

where $q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{y}_{0:k})$ is known as the importance density or proposal density. Ideally it is desired for the samples $\mathbf{x}_k^i$ to be distributed according to the posterior. Since it is the *posterior* which is being estimated, it is typically not possible to draw samples from it directly. It is the proposal by which the samples $\mathbf{x}_k^i$ are initially drawn and should be chosen to closely approximate $p(\mathbf{x}_k|\mathbf{y}_{0:k})$. A proposal density which poorly approximates the *posterior* will result in a lot of the samples $\mathbf{x}_k^i$ having small weights and not accurately representing the true *posterior* distribution. A scalar measure $\tilde{N}_{eff} = 1/\sum_{i=1}^{N_s}(w_k^i)^2$ known as the *effective* sample size can be used as a measure of the effectiveness of particles in estimating the distribution.

It is common in tracking to assign the proposal distribution as equal to the *prior* $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and form what is known as the bootstrap filter [132]. The bootstrap filter is implemented in the same predict and update process as the previously discussed recursive filter given by,

$$\begin{aligned}
&\textsc{Predict}: && \text{Perturb the particles according to the process model} \\
& && \mathbf{x}_{k|k-1}^i = \mathbf{T}_k(\mathbf{x}_k^i, \mathbf{v}_k). \\
&\textsc{Update}: && \text{Determine the particle likelihoods } p(\mathbf{y}_k|\mathbf{x}_k^i) \text{ and} \\
& && \text{associated weights according to equation 3.24 with} \\
& && q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{y}_{0:k}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}) \text{ i.e. } w_k^i = w_{k-1}^i p(\mathbf{y}_k|\mathbf{x}_k^i).
\end{aligned} \tag{3.25a}$$

The bootstrap filter can suffer from a phenomenon known as degeneracy where after a short number of iterations the effective sample size reduces and the weights of all particles but one is zero. Up until the adaptation of this filter as proposed by Gordon et al. the use of the bootstrap was constrained to few applications. Gordon et al. in their work suggested resampling the particles after the update stage and resetting the particle weights to $\frac{1}{N_s}$. This removes particles which do not contribute (small weights) to the estimate of $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ and results in increasing the effective sample size N_{eff}. It is known as the Sequential Importance Sampling (SIS) particle filter and is the most popular particle filtering method. Although resampling the particles at each time-step can counteract degeneracy, it can lead to a scenario where particles with large weights are resampled repeatedly reducing particle diversity. This is another failure characteristic of the particle filter known as *sample impoverishment*.

One notable problem using the $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ as the proposal density with the SIS particle filter is that it is not conditioned on the current observation $\mathbf{y}_k$. This means that particles can be propagated to regions away from the current observation which can lead to sample impoverishment. The auxiliary SIS particle filter is a variant of the basic algorithm to address this. It aims to sample from a proposal density $q(\mathbf{x}_k^i, i|\mathbf{y}_k)$ where i refers to the index of the particle at the previous time step $k-1$. The proposal distribution ensures that particles close to the current observation are assigned large weights in the resampling step.

Since particle filtering methods attempt to approximate probability densities by a set of samples, no restrictions are placed on the form of posterior $p(\mathbf{x}_k|\mathbf{y}_{0:k})$, likelihood $p(\mathbf{y}_k|\mathbf{x}_k)$ or state transition density $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. Thus, the PF is not restricted to estimate states confined to a linear Gaussian state space as with the standard KF. It can be used therefore in the case of nonlinear motion models and nonlinear measurement functions. In addition to this it is much more straightforward to implemented than the EKF, IEKF or UKF since it only requires maintaining the particles and their associated weights.

The ability of the PF to model complex likelihood functions is especially useful in the audio-based tracking problem. Both in the beamforming and time-delay audio-based tracking problem the likelihood function can have multiple modes. The multiple modes corresponding to multiple likely speaker positions can arise due to the presence of multiple speakers or because of the reverberation phenomenon. Enabling the modelling of such through complex likelihood functions makes the particle filter a powerful technique for tracking in the presence of reverberation [45]. A variety of likelihood functions have been proposed using the GCC for TDE-based tracking using an Auxiliary PF [100], and also for beamforming [32]. Further discussion in relation to these is delayed at this point since they are examined in the later analysis of the fusion problem.

3.1.5 Grid-based Approximation

An alternate means to particle filtering is to approximate the *posterior* through grid-based methods. Intuitively, this corresponds to the case where $\mathbf{x}_k^i$ in equation 3.23 does not define a set of random particles but rather a grid of points in the state space. What distinguishes this technique from that of particle filtering is that the grid points are defined deterministically.

The complexity of the *posterior* distribution generally dictates the resolution of the most effective approximating grid. Typically, if the *posterior* is complex with many multiple modes a high grid-resolution is required. As such, it is often too restrictive for many tracking problems.

In the offline tracking problem, if a grid-based estimate of the *posterior* $p(\mathbf{x}_k|\mathbf{y}_{0:k})$ is available then a MAP estimate

$$\mathbf{x}_{0:K}^{MAP} = \arg\max_{\mathbf{x}_{0:K}} p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K}), \tag{3.26}$$

of the state sequence $\mathbf{x}_{0:K}$ can be obtained using the Viterbi algorithm [57].

It is a consequence of discrete-time sampling the source signal in audio-based tracking that without some form of interpolation, the TDEs are constrained to discrete values. If the concerned

estimation problem defines the state of interest $\mathbf{x}_k$ as the true time-delay in the audio-domain then it can be addressed through grid-based methods.

Tung et al. [178] recognised the inherently discrete nature of TDEs enforced by the discrete time-delay estimation problem. They apply the Viterbi algorithm to obtaining a MAP estimate of the true time delay over a set of observations. Due to the offline manner of this approach it is only useful when offline methods can be used. The Viterbi algorithm however can also be used over windowed observations in a delayed tracking scheme.

The success of the Viterbi algorithm for tracking relies on determining accurate grid-based estimates of the posterior distribution. Later in chapter 6 a joint audio-video based active speaker tracking algorithm is introduced which uses the Viterbi algorithm. Video information is used by the tracking technique to obtain a grid-based estimate of the posterior distribution. Audio-based information is then used to determine the position of the active speaker from the grid-based estimate of the posterior.

3.2 Combining Audio and Video Observations for Tracking

At this point in the review after examining various approaches to audio-based tracking we introduce a sequence of video-based observations $\mathbf{z}_{0:k}$ of the state sequence $\mathbf{x}_{0:k}$. We consider therefore the task of estimating the state based on the combined set of audio and video observations $\{\mathbf{y}_{0:k}, \mathbf{z}_{0:k}\}$. If the video-based measurements $\mathbf{z}_{0:k}$ are assumed to be defined by a similar model to that of equation 3.4 then the likelihood function for the joint audio-video tracking problem becomes $p(\mathbf{y}_k, \mathbf{z}_k|\mathbf{x}_k)$.

Much of the existing work in handling the joint audio-video based likelihood function relies on the independent measurement assumption. This assumption states that both the audio and video measurements are independent of one another and only depend on the current state. In terms of the likelihood function this translates to

$$p(\mathbf{y}_k, \mathbf{z}_k|\mathbf{x}_k) = p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{z}_k|\mathbf{x}_k). \tag{3.27}$$

Little evidence arises to argue against such an assumption since the only common dependence between the observations under the assumed observation models is the current state. Therefore if the observation models can be justified, the independent measurement assumption is reasonable.

3.2.1 Simple Average

The most intuitive method of combining multi-modal measurements under the independent measurement assumption is to firstly obtain two individual state estimates $\hat{\mathbf{x}}_k^A$ and $\hat{\mathbf{x}}_k^V$ using the audio and video data respectively. Then, a global state estimate can be obtained as the

weighted average of the two estimates as,

$$\hat{\mathbf{x}}_k = \sum_{n=[A,V]} w_n \hat{\mathbf{x}}_k^n. \tag{3.28}$$

where w_n are weights associated with the single modality estimates. An important question arises in considering such an approach. It is difficult in practice to define the most appropriate weights w_n. Ideally, these should relate in some manner to the reliability and accuracy of the estimates $\hat{\mathbf{x}}_k^n$. In such a fusion scheme unreliable state estimates should be given a zero weighting and state estimates which are reliable should be weighted to reflect their level of accuracy. Ineffective weights can result in a single state estimate skewing the fused estimate.

It is unlikely that practical fusion applications exist where appropriate weighting can be employed through a simple scalar weighting scheme. In general, it is rather ad-hoc in its approach and cannot be regarded as optimal in a statistical sense. Given the lack of statistical information relating to the accuracy and reliability of the state estimates $\hat{\mathbf{x}}_k^n$ however, the simple average approach does provide an easily implementable approach to fusion despite its limitations.

As previously stated though, the simple average approach is only reasonable in cases where the state estimates are deemed reliable and similarly close to the true state. The simple average is employed under these condition by Bernardin et al. [103]. In their proposed audio-video based tracker they consider three states of operation. These correspond to that where; a reliable audio-based estimate exists but no video estimate; a reliable video-based estimate exists but no audio-based estimate and the case where both audio and video-based estimates are reliable. In their system both audio and video based measurements are regarded as reliable if their relative error is less than $0.5m$. Where both audio and video are deemed reliable their system employs the simple averaging strategy to fusing both estimates where each modality is weighted equally. Despite the criticisms which arise in relation to the simple averaging approach, their results show that it can be effective in achieving accurate results. However, its overall performance is critically dependent on the ability to accurately classify the single modality estimates as reliable or unreliable.

3.2.2 Audio-video Fusion using Kalman Filters

If the KF modelling assumptions as outlined in section 3.1.2 can be assumed to apply to the state $\mathbf{x}_k$ and observations $\mathbf{y}_k$ and $\mathbf{z}_k$, then the KF can be used to fuse audio and video based observations. Under the assumption of independent measurements implied by equation 3.27 it provides a convenient structure for the fusion of multi-modal estimates. Also, unlike the simple averaging scheme, the KF enables statistical information to be incorporated into the fusion problem. Uncertainty in either modality can be reflected by increasing the corresponding level of the measurement noise.

A basic application of the KF to the multi-modal measurement estimation problem can be

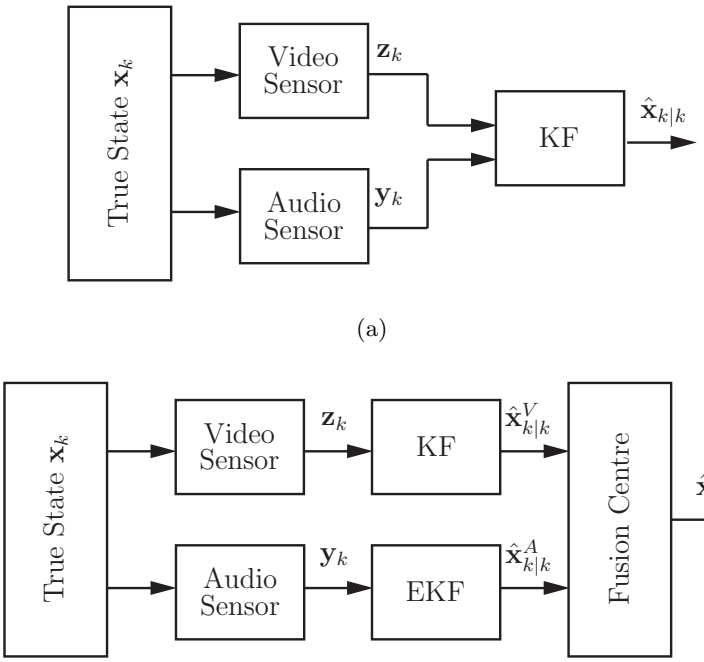


Figure 3.2: *Centralised (a) and Decentralised (b) Joint Kalman Filters (KFs). The centralised KF is not implementable due to the non-linear audio-based measurement function. The decentralised KF enables the incorporation of local KFs. The first local KF determines a video-based posterior state estimate and the second local EKF determines an audio-based posterior state estimate. A fusion centre then fuses both local state estimates into a global state estimate.*

taken through a *centralised* approach. A centralised approach to fusing multiple measurements using the KF is achieved by simply augmenting the observation vector $\mathbf{y}_k$ in equation 3.8b to include the additional video based measurement and define a new measurement vector as $[\mathbf{y}_k, \mathbf{z}_k]^T$. The basic concept of the centralised KF is illustrated in 3.2a.

The use of a centralised KF is not possible in audio-video based tracking problems where the audio and video sensors provide measurements in different coordinate spaces. For example, if the state $\mathbf{x}_k$ is defined as a position in the image plane, then since the video based observations are also made in this space, the measurement function can be defined as a simple linear model. In contrast however, the audio measurements are not made directly in the image plane but possibly as angles of azimuth and elevation represented in polar coordinates relative to the image plane. To address such issues Strobel et al. [136] have proposed the *decentralised* KF.

The decentralised approach implements two KFs locally for each modality to determine two single modality *posterior* state estimates $\hat{\mathbf{x}}^A_{k|k}$ and $\hat{\mathbf{x}}^V_{k|k}$ using the audio and video data respectively. A standard centralised KF is then used to determine a combined *a posteriori* state estimate $\hat{\mathbf{x}}_{k|k}$ from both $\hat{\mathbf{x}}^A_{k|k}$ and $\hat{\mathbf{x}}^V_{k|k}$. This decentralised approach is illustrated in figure 3.2b. So long as the initial conditions of the state and the prior information are the same for each

local KF $\hat{\mathbf{x}}_{k|k}$ and $\mathbf{P}_{k|k}$ are obtained by,

$$\hat{\mathbf{x}}_{k|k} = \mathbf{P}_{k|k}\left(\mathbf{P}^{-1}_{k|k-1}\hat{\mathbf{x}}_{k|k-1} + \sum_{n=[A,V]}\left[(\mathbf{P}^{n}_{k|k})^{-1}\hat{\mathbf{x}}^{n}_{k|k} - (\mathbf{P}^{n}_{k|k-1})^{-1}\hat{\mathbf{x}}^{n}_{k|k-1}\right]\right) \tag{3.29a}$$

$$\mathbf{P}^{-1}_{k|k} = \mathbf{P}^{-1}_{k|k-1} + \sum_{n=[A,V]}\left[(\mathbf{P}^{n}_{k|k})^{-1}\hat{\mathbf{x}}^{n}_{k|k} - (\mathbf{P}^{n}_{k|k-1})^{-1}\hat{\mathbf{x}}^{n}_{k|k-1}\right]. \tag{3.29b}$$

where $\mathbf{P}^{A}$ and $\mathbf{P}^{V}$ are the state error covariances corresponding to the audio and video based local state estimates respectively. Equation 3.29 represents the operation of the fusion centre component in figure 3.2b. Implementation of this fusion strategy therefore simply requires that each local KF communicates their local *a priori* and *a posteriori* estimate of both the state and state error covariances.

The block structure of the decentralised KF means that it is not restricted to just a single video-based and single audio-based estimate. Its decentralised structure means that it can be used to combine multiple local state estimates from both audio and video sensors and can be extended to include multiple motion models [64]. The only requirement is that the all measurements must be synchronised [135]. This requirement for synchronous measurements is a restriction for speaker tracking applications since audio-based measurements are only available when the speaker is active which is unlikely to be at the same rate as the video measurements. Synchronising the audio and video measurements temporally before the filter update stage is the normal approach in addressing this issue [1].

Besides the benefit of handling combinations of linear and non-linear measurement functions the decentralised KF in general exhibits the additional benefit of faster convergence than a conventional KF [18]. This is the case, due to its ability to process multiple sensor data simultaneously.

Since being popularised by Strobel et al. many other uses for the dencentralised KF have been reported for speaker tracking [1,9,133]. Katsarakis et al. [133] in their work describe the use of the decentralized KF filter for combining *3D* positional measurements from multiple cameras and multiple microphone arrays. Their use of the decentralised structure also incorporates a weighting to each local state estimate before the global state estimate is obtained to reflect the relative confidence in both modalities. The details of suitable weights however is not elaborated upon. A similar implementation of their algorithm is made in [9] as the main tracking component in a multi-modal person identification system. Abad et al. [1] also employ the decentralised KF for speaker tracking to combine location estimates obtained through the SRP-PHAT algorithm in the audio-domain and voxel-based estimates from multiple cameras in the video domain.

Incremental Update based Approach

The decentralized KF requires a state estimate to be determined by local single modality KFs. This requires the state to be completely observable by each modality measurement function. Gehrig et al. [174] have proposed an alternative to the decentralised KF which is based on the incremental update method of Welch et al. [66]. In this work a KF algorithm is described which enables the state estimate to be incrementally updated by *incomplete* observations. For example, a 2D pixel measurement or 1D TDE relating to the *3D* position of an active speaker, each represent incomplete measurements since neither fully observes the three dimensions of the state. Gehrig et al. highlight that such incomplete information can be still used to estimate the state. In their approach the state is fully observed by multiple incomplete observations which are used to estimate the state incrementally. They demonstrate this in their approach and successfully apply this strategy to incrementally estimate the *3D* position of an active speaker using TDEs from multiple microphones and detected faces from multiple cameras. One significant advantage of this method is that it enables asynchronous measurement updates. This is of significant advantage in the case of active speaker tracking since audio-based measurements are not always available at the same time as video-based measurements and only when the speaker is active. Essentially, the incremental update approach enables the state estimate to be refined as the measurements become available.

3.2.3 Audio-video Fusion using Particle Filters

The PF has a distinct advantage over the KF since it does not strictly require the likelihood function to be defined as Gaussian. As a result much more complex likelihood functions can be defined. In turn, this enables more complex low level audio and video features to be incorporated into the tracking problem. This is achieved through the definition of an audio-based likelihood and a video-based likelihood function. Given these definitions fusion can be applied in the Bayesian tracking framework in a straightforward manner. This is presented in the following sections. First however, some examples of audio-based and video-based likelihood functions which exist in the current literature are examined.

Audio-based Likelihood Functions

Consider the case of a set of $m = 1, .., N_{mic}$ microphone pairs from which we wish to define an audio-based likelihood function $p(\mathbf{y}_k|\mathbf{x}_k)$ where the measurements $\mathbf{y}_k$ are a set of TDEs. The expected time-delay at the mth microphone pair for the state $\mathbf{x}_k$ can be obtained using the time-delay measurement function $\tau_m^o = g(\mathbf{x}_k)$ as in equation 2.25. This can be used to define a vector of expected time delays corresponding to $\mathbf{x}_k$ as $\tau^o = [\tau_1^o, .., \tau_{N_{mic}}^o]^T$. A vector $\hat{\tau} = [\hat{\tau}_1, .., \hat{\tau}_{N_{mic}}]^T$ of measured time-delays can be obtained at each microphone pair using the GCC function. The measured time delays could also be determined through a beamforming technique to estimate the time-delay corresponding to the maximum received signal power at the microphones.

Given this, the most simplistic approach to defining an audio-based likelihood function is to assume a Gaussian error distribution on the measured time-delays [43, 197]. Through this the likelihood can be defined as,

$$p(\mathbf{y}_k|\mathbf{x}_k) = \mathcal{N}(\tau^o; \hat{\tau}, \Sigma_{\hat{\tau}}) \tag{3.30}$$

where $\Sigma_{\hat{\tau}}$ is the covariance matrix of the measured time-delays. Although the Gaussian error distribution is in common use [43, 197], it has the limitation of only considering a single TDE at each microphone pair. It does not provide a strategy to model the occurrence of anomalous TDEs.

A more sophisticated audio-based likelihood function can be defined which can incorporate a multiple peak analysis of the GCC function. Through the GCC function of the mth microphone pair a total of N_p peaks $\{\hat{\tau}_m^1, .., \hat{\tau}_m^{N_p}\}$ can be identified corresponding to likely speaker positions. This is advantageous since as revealed in previous discussions, multiple peaks can arise and the most significant peak does not always correspond to the true source position. In addition to this, it can occur that no single peak can be identified as the most likely estimate of the true time-delay. Vermaak et al. [100] propose an audio-based likelihood function which considers multiple TDEs [100, 101]. They define an audio-based likelihood over multiple hypotheses which consider each peak as either corresponding to the true source position or clutter. A likelihood function for the mth microphone pair in their work is proposed as,

$$p_m(\mathbf{y}_k|\mathbf{x}_k) \propto \sum_{j=1}^{N_p} q_j \mathcal{N}(\tau^o; \hat{\tau}^j, \sigma_{\tau^j}^2) + q_0 \tag{3.31}$$

where q_0 is the predefined probability of a TDE relating to clutter, q_j is the probability of the jth peak corresponding to the true speaker position and $\sigma_{\tau^j}^2$ the variance of the TDE at the jth peak. The complete likelihood over all microphone pairs is simply obtained using the independent measurement assumption, as the product of the individual likelihoods at each microphone pair.

It is necessary for the chosen value of q_0 to reflect the reliability of a time-delay estimator in a given room environment. This is difficult since experimental evidence suggests that q_0 varies with the source position [121]. Typically, if a speaker is positioned close to a wall within a room, this can increase the probability of an anomalous estimate indicating a higher value of q_0 than for other regions within the room. Implementing such environment information in defining q_0 is not straightforward. In their experimental analysis they assume all q_j to be equal and evaluate q_0 empirically. Overall this is seen to be sufficient to achieve robust tracking performance in a reverberant environment. Alternative strategies for choosing values for q_j can be through some reliability measure such as those discussed in section 2.1.3.2.

A more straightforward formulation of an audio-based likelihood can be made using the GCC function directly. An audio-based likelihood function formed directly from the GCC is given

as [31, 105, 106, 159],

$$p_m(\mathbf{y}_k|\mathbf{x}_k) \propto \max[R(\tau^o), q_0]^l \tag{3.32}$$

where $R(\tau_0)$ is the GCC function and $l \in \mathbb{R}^+$. The max operation with positive q_0 is to account for cases where the cross-correlation function is negative since such a scenario would not lead to a valid probability distribution. One issue with the definition of equation 3.32 is that due to the presence of multiple sharp peaks in the GCC function, sharp peaks are present in the likelihood function. This is not ideal for particle filter methods which are susceptible to the degeneracy problem under such conditions. The l term in equation 3.32 is used to emphasis significant peaks and reduce the overall contribution of smaller peaks in the GCC function. Effectively, it operates to smooth the likelihood function making it more suitable for recursive particle filtering [31].

The flexibility of the PF means that there is little restriction in defining an audio-based likelihood function and definitions are not confined to those formed using the GCC function. Checka et al. [128] propose an audio-based likelihood function in the frequency domain through modelling the spatio-temporal covariance matrix at a speaker's location. They address the multiple speaker tracking scenario and extend the positional speaker state $\mathbf{x}_k$ to include a binary speech activity bit s. Similar to the definition of the spatio-spectral covariance matrix as presented in section 2.1.2 they define the spatio-spectral covariance matrix at $\mathbf{x}_k$ as,

$$\mathbf{R}_{s_j}(\omega) = s_j \mathbf{D}(\omega)\mathbf{D}(\omega)^H \tag{3.33}$$

where $\mathbf{D}(\omega)$ is the steering vector corresponding to $\mathbf{x}_k$. This model effectively is an estimate of the spatio-spectral received due to the jth speaker with speech activity defined by s_j. Their model of the spatio-spectral covariance matrix for a microphone array given over all speaker positions is established as,

$$\mathbf{R}(\omega) = \mathbf{R}_b(\omega) + \sum_j \mathbf{R}_{s_j} + \lambda \mathbf{I} \tag{3.34}$$

where $\mathbf{R}_b(\omega)$ is an empirically measured spatio-spectral covariance matrix due to background noise sources and λ is a tuning parameter which is also measured empirically. Using their specified model of the microphone array's spatio-spectral matrix, an audio-based likelihood is formed by relating the model $\mathbf{R}(\omega)$ to the measured spatio-spectral matrix.

To apply this, they obtain an N_ω point Short Time Fourier Transform (STFT) for each microphone signal and define a vector $\mathbf{w}(\omega)$ for each of the available microphone pairs. The likelihood is then defined by,

$$p(\mathbf{y}_k|\mathbf{x}_k) = \prod_{n=1}^{N_\omega} \mathcal{N}(\mathbf{w}(n); 0, \mathbf{R}(n)). \tag{3.35}$$

The benefit of the likelihood defined by Checka et al. is that it can incorporate information relating to multiple speakers and also their state of speech activity. Building the spatio-spectral

covariance model however is not straightforward and requires considerable empirical analysis. It therefore lacks the simplicity of the previously discussed likelihood functions simply based on the GCC function.

Video-based Likelihood Functions

With the video based measurement function $f(\cdot)$ as defined in equation 2.72 the *3D* position $\mathbf{x}_k$ of a speaker can be located in the image plane as $\mathbf{p} = f(\mathbf{x}_k)$. Once a hypothesised speaker position $\mathbf{x}_k$ is located in the image plane, the likelihood of that position can be defined in terms of image features such as those described in section 2.2. Typical features include foreground regions obtained through background subtraction, detected skin colour regions, detected faces or edges can all be used. The most simple approach in building a video-based likelihood is to identify an image region corresponding to a potential speaker using high level image features such as a head or face. This analysis can applied to each of $j = 1, .., N_{cam}$ cameras to identify the pixel locations $\hat{\mathbf{p}}_j$ of the speaker in the jth view. A vector of pixel locations can then be set across all views as $\hat{\mathbf{p}} = [\hat{\mathbf{p}}_1, .., \hat{\mathbf{p}}_{N_{cam}}]^T$. In an identical manner to the simple audio-based likelihood, a simple video-based likelihood can be defined assuming a Gaussian error distribution on the pixel measurements $\hat{\mathbf{p}}$ as [43],

$$p(\mathbf{z}_k|\mathbf{x}_k) = \mathcal{N}(\mathbf{p}; \hat{\mathbf{p}}, \Sigma_{\hat{\mathbf{p}}}). \tag{3.36}$$

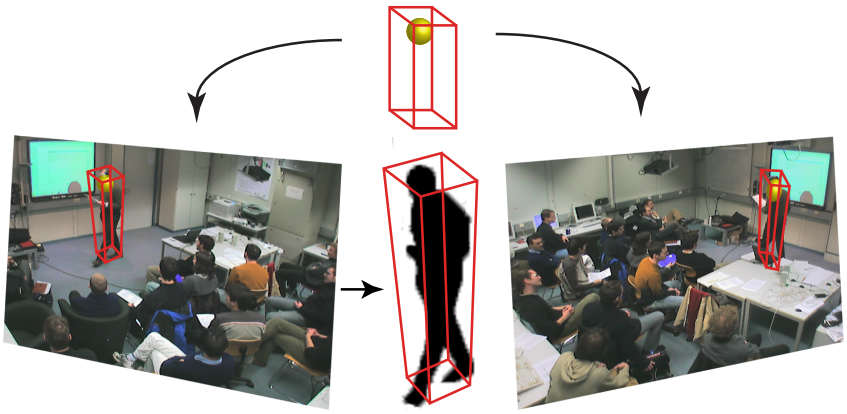
where $\Sigma_{\hat{\mathbf{p}}}$ is the covariance of the pixel-based measurements.

Shape The PF offers the freedom to define video-based likelihood based on multiple image features. In particular hypothesised speaker positions (particles) do not need to be strictly point estimates. Instead, each particle can be associated with a body model such as a cylinder or a more complex model. The projection of this model then using the video-based measurement function $f(\cdot)$, gives this model's representation in the image plane. Using this model, more complex likelihood functions based on the model's correspondence to the image data can be defined. This is the approach of Nickel et al. [106] and used also in [159] and [105]. Their approach performs adaptive background subtraction on multiple camera views in order to determine foreground regions corresponding to a moving speaker. They define a simple cuboid body model in *3D* space and estimate the projected body shape into multiple views. This is applied for each particle in *3D* space. The likelihood of each hypothesised particle is then defined as the total proportion of the projected body shape that is occupied by detected foreground. An illustration of this model for two sample camera views of figure 3.3a and figure 3.3b is presented in figure 3.3c.

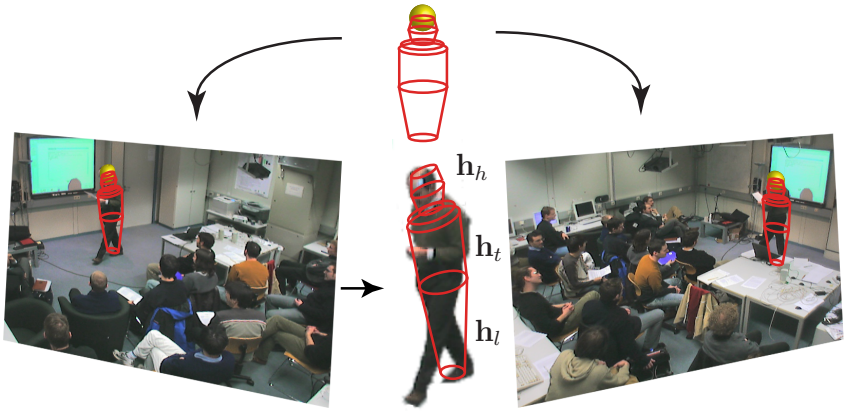
Faces In addition to this foreground based likelihood, Nickel et al. [106] also define a likelihood function to weight particles using a face detector. Once the pixel location of a particle is

(a) *Sample left image*

(b) *Sample right image*



(c) *Simple cuboid body model of [106] in* 3D *space and projected into two views. The video based likelihood is defined to weight particles based on the amount of foreground contained within the bounds of the projected body model in the image plane.*



(d) *Cylinder type model proposed by [151] shown in* 3D *space and projected into two views. A video-based likelihood is formed based on the similarity between the projected body shape in the image plane to that of detected edge regions. Also established in the model are histogram based colour models of the head* $\mathbf{h}_h^m$*, torso* $\mathbf{h}_t^m$ *and legs* $\mathbf{h}_l^m$ *of the body. A colour-based likelihood is defined which weights particles based on the similarity between the histograms* $\mathbf{h}_h^m$*,* $\mathbf{h}_t^m$ $\mathbf{h}_l^m$ *to their respective measurements* $\mathbf{h}_h$*,* $\mathbf{h}_t$ $\mathbf{h}_l$*. The similarity measure used is the Bhattacharyya similarity measure.*

Figure 3.3: *Illustrated example of different body shape models for particle filters. The example frames are taken from the* 2005 *CHIL Evaluation Package [171].*

determined within the view of the camera the Viola and Jones [143] face detection algorithm is applied to a specified region about the projected particle's location. Restricting the search region to a small area about the projected particle location helps to reduce the computational demands of the algorithm. Each particle is then weighted in proportion to the overlapping area of the particle's face region with that of other positively labelled face regions belonging to the other particles.

Contours It is also possible to incorporate contour information into the video-based likelihood. Brunelli et al. [151] utilise a similar body model to that of Nickel et al. but use contour information rather than detected foreground to measure how well it corresponds to the image data. They use a better fitting cylindrical type body model in contrast to the cuboid model that Nickel et al. use. An example of the body model used is shown in figure 3.3d. To incorporate contour information into a likelihood function they compare the correspondence of the projected body shape in the image plane to detected edge regions in the image. The method used for edge detection re-enforces the standard Sobel edge detector with edge information obtained by temporal frame differencing. Through this, the particles are weighted in proportion to the sum of Euclidean distances between the detected edge regions in the image and that of the projected edge structure of the body model in the image plane. The use of contour information is popular in joint audio-video tracking systems and has been used for elliptical head shape [195] and head and shoulder [101] contour models for speaker tracking.

Colour In addition to analysing contours, Brunelli et al. [151] build a likelihood function based on colour analysis. To achieve this they divided the assumed body model into three regions corresponding to the speaker's head, torso and leg regions. Their tracking system implements an automatic acquisition procedure to obtain a colour histogram model for each body region. Using these colour models they define a colour likelihood which weights particles based on the similarity between the histogram models and colour histogram measurements from the projected body region in multiple camera views . The similarity measure which they use in measuring the closeness of the measured histograms to the measured data is the Bhattacharyya distance. The Bhattacharyya distance provides a convenient similarity measure for colour based histograms and a simple means for matching colour histograms and is the basis of colour histogram analysis in many active speaker tracking systems [103, 114, 195, 197].

Combining Measurements from Different Modalities

Given the definition of the audio and video based likelihood functions, the Bayesian formulation of the tracking problem provides a direct approach to fusing the audio and video information. Re-examining the Bayesian tracking problem with the likelihood data defined by equation 3.27,

the posterior $p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{z}_k)$ becomes,

$$p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{z}_k) \propto p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{3.37}$$

Here, the audio and video based information fusion is achieved by *pooling* the likelihoods which is known in the estimation literature as the *independent likelihood pool* [94, Chapter2], [96, Chapter 4]. This is the most dominant method of fusing audio and video information in speaker tracking systems. The majority of the state-of-the art speaker tracking systems use this fusion strategy [36–38, 43, 101, 114, 128, 129, 151, 195, 197].

A number of important characteristics of this strategy are evident. Firstly, in contrast to the previously encountered fusion techniques, the independent likelihood approach does not require the specification of any weighting between the different modalities. This information is completely defined in the likelihood models.

Linear Opinion Pool

An alternative to the independent likelihood model for particle filters is known as the *Linear Opinion Pool* [96, Chapter 4]. Following along similar lines to that of the simple weighted average, this strategy implements a weighted sum of the single modality *posteriors* $p(\mathbf{x}_k|\mathbf{y}_k)$ and $p(\mathbf{x}_k|\mathbf{z}_k)$ as,

$$p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{z}_k) \propto \beta p(\mathbf{x}_k|\mathbf{y}_k) + (\beta - 1)p(\mathbf{x}_k|\mathbf{z}_k) \tag{3.38}$$

where β defines the relative weight between the single modality audio and video based *posterior* estimates. If it is assumed that the same initial conditions and *prior* on the state $\mathbf{x}_k$ is maintained for each single modality estimate, then it can be seen that,

$$p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{z}_k) \propto \left[\beta p(\mathbf{y}_k|\mathbf{x}_k) + (1 - \beta)p(\mathbf{z}_k|\mathbf{x}_k)\right] p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{3.39}$$

Therefore, the linear opinion pool in this case can be implemented by a weighted sum of the single modality likelihood functions. This approach has recently found application in the audio-video tracking literature [105, 106, 140, 141].

Similar to the criticisms of the simple weighted average fusion scheme this method of information fusion suffers from a lack of strong conditions for choosing the weights for the audio $\beta_A = \beta$ and video $\beta_V = (1 - \beta)$ data . As a result, in its application to joint audio-video based tracking, the weights are often restricted to empirically defined measures. In both Nickel et al. [106] and Stiefelhagen et al. [159] a variable weight is defined as $\beta_A = \frac{m_o}{N_{mic}}(0.6)$ where m_0 is the number of TDEs meeting a defined reliability measure. A different strategy is implemented in [105] where the weighting is defined as $\beta_A = (\sigma^2_{A_x} + \sigma^2_{A_y})^{-\frac{1}{2}}$ where $\sigma^2_{A_x}$ and $\sigma^2_{A_y}$ are the variances of localisation in the x and y axes respectively. The video-based weight β_V is obtained in a similar fashion. This is preferable to a fixed empirical measure since it aims to make the best use of the available statistical information in relation to the observed measurement uncertainty.

The fusion of audio and video as proposed by Aarabi et al. [140, 141] although not directly formulated in the manner of equation 3.38, can be considered as a linear opinion pool approach. They also take an empirical strategy in establishing the weights β_i and define the weighting by $\beta_V = (1 - \beta_A)$ with $\beta_A = 0.05$. Both proposed joint audio video based systems stress an affinity for video-based localisation above that of audio which is reflected in the weightings.

3.3 Final Comments

This chapter presented the problem of audio-based active speaker tracking in a Bayesian state estimation framework. Online state estimation techniques such as Kalman filtering, variants of the KF which can handle nonlinearities and the PF were examined. Also, the offline state estimation problem using a grid-based approach was considered.

In the analysis, the problem of motion modelling was introduced in relation to the KF. It was shown that ensuring optimal tracking is difficult because the problem of choosing the most appropriate motion model is difficult. Sub-optimal tracking will persist if poor motion models are used. Such inaccurate *prior* information can significantly affect the utility of multiple measurements in improving tracking accuracy and reliability.

Various key requirements of tracking filters for audio-based tracking were identified such as the need for them to handle the non-linear time-delay and DOA measurement functions. Therefore the EKF, IEKF, UKF and PF were all presented as important tools for audio-based tracking. Of these, the PF was deemed to be most useful since it enables the incorporation of multiple time-delay measurements in the tracking problem.

The latter analysis in the review, focused on the fusion problem and how existing audio-based tracking filters can be extended to include video-information. The decentralised KF and various fusion strategies using PFs were examined. The particle filter was seen to offer considerable flexibility to encompass both low-level audio and video based features for tracking.

Two prominent strategies for fusing audio and video information were identified in relation to PFs. These were the linear opinion pool requiring a scalar weighting for each modality and the independent likelihood pool which did not require such a weighting scheme.

4

Analysis of Audio-Visual Source Localisation Accuracy[1]

Known techniques for fusing audio and video based location estimates were examined in chapter 3. In general, these techniques show improved localisation performance beyond that possible through the use of either audio or video data alone. This result is mainly due to the complementary nature of the audio and video modalities. Since audio and video tend to fail independently, it is reasonable to expect an overall improvement in tracking reliability by combining both modalities.

Reliability however, is not the only aspect of tracking performance which is of interest. The accuracy of localisation is also important. It is necessary at this point to place a distinction between reliability and accuracy in the measure of localisation performance. In this chapter, the term accuracy is used to refer specifically to the measure of how close a location estimate is to the true location. The term reliability is used to refer to the consistency by which accurate localisation is achieved.

There is conflicting evidence in the literature that improved tracking accuracy is achieved through a joint audio-video based approach. Specifically, Strobel et al. report improved tracking reliability, but note no clear improvement in accuracy beyond the best available single-modality positional estimate [136]. Furthermore, literature shows evidence of a fusion-based approach

[1]Results from this chapter have been published in: Damien Kelly, François Pitié, Anil Kokaram and Frank Boland. A Comparative Error Analysis of Audio-Visual Source Localization. *Workshop on Multi-camera Multi-modal Sensor Fusion Algorithms and Applications* (M^2SFA^2) *in conjunction with 10th European Conference on Computer Vision 2008 (ECCV 2008) [41]*

actually under performing single modality tracking [133]. Results such as these, indicate that there are issues relating to the performance of fusion-based systems which are currently not fully understood.

The general approach in evaluating the performance of joint audio-video trackers, is to determine the tracking performance against some ground truth (eg. [128, 133, 136]). Tracking performance however is not only dependent on the accuracy of audio and video-based position estimates but also on the employed motion model. Thus a well chosen motion model can result in undue credit being attributed to a multi-modal approach in improving tracking accuracy.

Furthermore, this measure of performance can only give an indication of the expected accuracy of the resulting fused track. It cannot be used to determine how well a system is performing in relation to its best possible performance with known uncertainties and motion model. Currently, little effort has been made in the literature to evaluate the expected performance of joint audio-video based tracking systems. This means that such systems are being proposed without any theoretical basis on which to measure and evaluate their performance.

It is informative to examine the performance of a joint audio-video based tracking system by examining the localisation accuracy in each domain individually. In this way the contribution of both audio and video in improving localisation accuracy through a fused estimate can be determined. Current literature proposes techniques for predicting the *3D* error associated with localisation using multiple cameras [60] and multiple microphones [17]. The incorporation of such theory in the analysis of joint audio-video based tracking to date has not been adequately investigated. This chapter addresses this issue by unifying the theory of estimating localisation uncertainty through multiple cameras and multiple microphones under a single framework. This is then used to gain some insight into what might degrade the performance of joint audio-video based fusion.

In this work, a framework based on covariance mapping theory is used to estimate localisation uncertainty. This mapping theory is used to determine the *3D* localisation error associated with audio-based localisation using TDEs from multiple microphones and video-based localisation through triangulation using multiple cameras. Given this, a direct comparison is made between the localisation accuracy of both modalities in terms of their ability to provide accurate location estimates of a moving audio-visual source.

Covariance mapping is also used to determine a representation of uncertainty on the time delay estimates in the video domain and similarly to determine a representation for uncertainty on pixel measurements in the audio domain. Effectively, audio and video localisation uncertainty is examined in the positional domain, the audio domain and also the video domain. Maximum likelihood data fusion is applied in these three domains and a resulting *3D* fused localisation estimate is obtained in each case. The effectiveness of these fusion strategies is examined from a theoretical basis and their ability to provide accurate location estimates of a moving source is evaluated. In addition to this, the different fusion strategies are compared to that of a simple audio-video switch-based localisation approach where the location estimate is derived from the

best available single modality location estimates.

In order to make this analysis tractable, the general assumption of Gaussian observation noise is made on the TDEs in the audio domain and also on the pixel measurements obtained from each camera in the video domain.

4.1 Uncertainty Mapping

In this analysis we examine the *3D* position $\mathbf{x} \in \mathbb{R}^3$ of an audio-visual source observed indirectly by pixel measurements $\mathbf{p} = f(\mathbf{x})$ in the video domain and time delay estimates $\boldsymbol{\tau} = g(\mathbf{x})$ in the audio domain. Using these measurements we wish to map their respective covariances $\Sigma_{\mathbf{p}}$ and Σ_{τ} into positional space in order to estimate the associated covariance $\Sigma_{\mathbf{p}}^{V}$ of a video-based location estimate $\mathbf{x}^{V}$ and the covariance $\Sigma_{\boldsymbol{\tau}}^{A}$ of an audio-based location estimate $\mathbf{x}^{A}$. Also of interest in this analysis is the fused ML audio-video based location estimate and its associated covariance. In the positional domain the fused estimate can be obtained as [71],

$$\mathbf{x}^{Pos} = \Sigma_{\mathbf{x}}^{Pos}((\Sigma_{\mathbf{x}}^{A})^{-1}\mathbf{x}^{A} + (\Sigma_{\mathbf{x}}^{V})^{-1}\mathbf{x}^{V}). \tag{4.1}$$

where

$$\Sigma_{\mathbf{x}}^{Pos} = ((\Sigma_{\mathbf{x}}^{A})^{-1} + (\Sigma_{\mathbf{x}}^{V})^{-1})^{-1}, \tag{4.2}$$

is the associated covariance of the estimate.

Fusion in this application can also be considered in two additional domains, the audio domain corresponding to time delays and the video domain corresponding to pixel measurements. In order to examine fusion in these other domains, it is necessary to transform both audio and video measurements and associated uncertainty to equivalent levels of representation. For instance, for fusion in the video domain an equivalent representation of audio-based measurements and uncertainty must be determined in the image plane. Similarly, for fusion in the audio domain, video-based localisation measurements and uncertainty must be transformed to an equivalent representation in the audio domain.

Transforming audio and video measurements to equivalent levels of representation in this context is straightforward and is achieved using the measurement functions $f(\cdot)$ and $g(\cdot)$. For example, the equivalent representation of an audio-based location estimate $\mathbf{x}^{A}$ in the video domain is simply obtained as $\mathbf{p}^{A} = f(\mathbf{x}^{A})$. Similarly, the equivalent representation of a video-based location estimate $\mathbf{x}^{V}$ in the audio domain is obtained as $\boldsymbol{\tau}^{V} = g(\mathbf{x}^{V})$.

In order to fuse both audio and video measurements in different domains however it is not only necessary to transform measurements between domains but also their associated covariances. Therefore, in the video domain, in addition to the covariance of pixel measurements $\Sigma_{\mathbf{p}}$, the covariance $\Sigma_{\mathbf{p}}^{A}$ of $\mathbf{p}^{A}$ needs to be determined. Likewise, in addition to the covariance of time delay estimates $\Sigma_{\boldsymbol{\tau}}$, the covariance $\Sigma_{\boldsymbol{\tau}}^{V}$ of $\boldsymbol{\tau}^{V}$ must be determined in the audio domain. Given this, it is then possible to fuse the audio and video measurements in either the audio or video

domain as in equation 4.2 and equation 4.1.

The resulting covariance of the fused estimates can then be mapped from each domain into positional space. This enables the covariance $\Sigma_{\mathbf{x}}^{Vid}$ of *3D* localisation through fusion in the video domain and the covariance $\Sigma_{\mathbf{x}}^{Aud}$ of *3D* localisation through fusion in the audio domain to be evaluated.

In summary, the following location estimates with their associate covariances are considered,

$\boldsymbol{\tau}, \Sigma_{\boldsymbol{\tau}}$: Audio-based measurement (TDEs).

$\mathbf{p}, \Sigma_{\mathbf{p}}$: Video-based measurement (Pixels).

$\mathbf{x}^A, \Sigma_{\mathbf{x}}^A$: Audio-based *3D* location estimate.

$\mathbf{x}^V, \Sigma_{\mathbf{x}}^V$: Video-based *3D* location estimate.

$\boldsymbol{\tau}^V, \Sigma_{\boldsymbol{\tau}}^V$: Video-based location estimate transformed into the audio domain of time-delays.

$\mathbf{p}^A, \Sigma_{\mathbf{p}}^A$: Audio-based location estimate transformed into the video domain of pixels.

$\mathbf{x}^{Vid}, \Sigma_{\mathbf{x}}^{Vid}$: *3D* location estimate through fusion in the video domain.

$\mathbf{x}^{Aud}, \Sigma_{\mathbf{x}}^{Aud}$: *3D* location estimate through fusion in the audio domain.

$\mathbf{x}^{Pos}, \Sigma_{\mathbf{x}}^{Pos}$: *3D* location estimate through fusion in the positional domain.

4.1.1 Linear Approximation Mapping

The mapping of covariances between the positional domain, audio domain and video domain can be achieved through a first order Taylor series expansion of the audio and video measurements functions and their inverses. Here the process of mapping the covariance $\Sigma_{\mathbf{x}}$ of the source position $\mathbf{x}$ to obtain a corresponding measure of pixel uncertainty $\Sigma_{\mathbf{p}}$ in the video domain is presented. Consider the position $\mathbf{x} \in \mathbb{R}^3$ as a random variable of Gaussian distribution, mean $E[\mathbf{x}]$ and covariance $\Sigma_{\mathbf{x}}$. A pixel-based observation by N_{cam} cameras of this position results in a random vector $\mathbf{p} \in \mathbb{R}^{2\times N_{cam}}$ where $\mathbf{p} = f(\mathbf{x})$. If the measurement function $f(\mathbf{x})$ has a continuous first order derivative then a first order Taylor series expansion of $f(\mathbf{x})$ enables the mean and covariance of $\mathbf{p}$ to be approximated [166]. The mean of $\mathbf{p}$ is approximated by

$$E[\mathbf{p}] \approx f(E[\mathbf{x}]), \tag{4.3}$$

and its covariance $\Sigma_{\mathbf{p}}$ by,

$$\Sigma_{\mathbf{p}} = \frac{\partial f(E[\mathbf{x}])}{\partial \mathbf{x}} \Sigma_{\mathbf{x}} \frac{\partial f(E[\mathbf{x}])}{\partial \mathbf{x}}^T \tag{4.4}$$

where $E[\cdot]$ is used to denote the expectation operator. The transformation of uncertainty in this manner by a first order Taylor series expansion of the measurement function follows directly the

theory of the EKF [194] as in section 3.1.3. This assumption will be later explored and validated in section 4.4.1.

The inverse mapping of (4.4) is difficult to obtain in cases where the inverse measurement function $\mathbf{x} = f^{-1}(\mathbf{p})$ can not be explicitly defined. This occurs in cases where the inverse relation between $\mathbf{p}$ and $\mathbf{x}$ is instead implicitly defined by

$$F(\mathbf{p}, \mathbf{x}) = \mathbf{p} - f(\mathbf{x}) \tag{4.5}$$

and only an estimate $\mathbf{x}^V$ of the position can be determined. In practice the estimate $\mathbf{x}^V$ is usually defined in a least squares sense [169, chapter 14] as that which minimises the criterion function

$$C(\mathbf{p}, \mathbf{x}) = |F(\mathbf{p}, \mathbf{x})|^2. \tag{4.6}$$

Here, the notation $|\cdot|$ is used to denote the Euclidean norm. In this scenario the first order derivative of the inverse measurement function can be approximated using the implicit functions theorem [138, chapter 5]. This is found to be,

$$\frac{\partial f^{-1}(E[\mathbf{p}])}{\partial \mathbf{p}} \approx -\left(\frac{\partial F}{\partial \mathbf{x}}\right)^{\dagger} \frac{\partial F}{\partial \mathbf{p}} \tag{4.7}$$

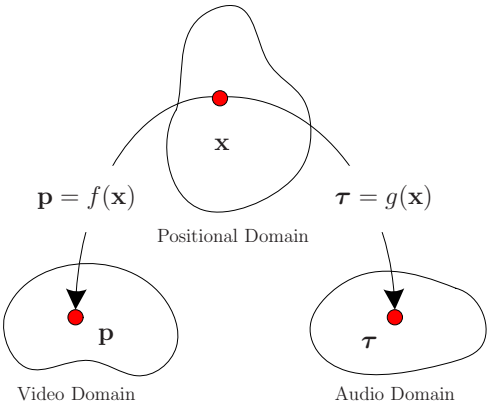
where $\dagger$ is used to denote the pseudo inverse.

Using the relations defined in both equation 4.4 and equation 4.7, measurement uncertainty can be mapped between the positional, audio and video domains. The mappings under consideration are illustrated in figure 4.1.

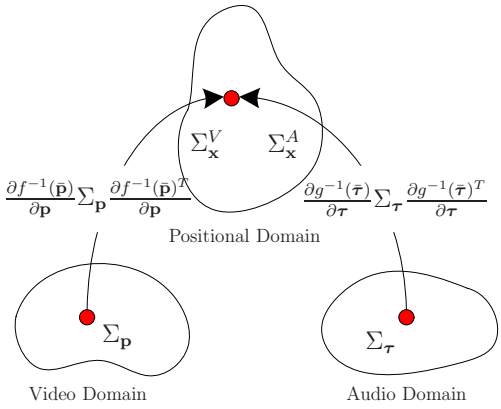
4.2 Audio-based Measurement Function

The audio measurement function as defined in equation 2.25 can be used to define a vector of time delays $\boldsymbol{\tau}$ associated with the *3D* point $\mathbf{x}$. This function is completely described by the positions of the microphones, the sampling frequency and the speed of sound. Let $\mathbf{m}_{ij} = [X_{ij}, Y_{ij}, Z_{ij}]^T$, $j = [1, 2]$ denote the positions of the microphones of the ith microphone pair configuration and $\boldsymbol{\tau} = [\tau_1, .., \tau_i, ..., \tau_{N_{mic}}]^T$ be the vector of time-delay estimates for N_{mic} microphone pairs. For the ith microphone pair, the expected time delay τ_i given the source position $\mathbf{x}$ may be determined by expanding equation 2.25 to obtain,

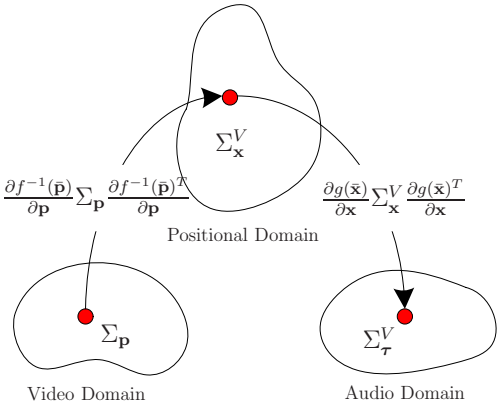
$$\begin{aligned}\tau_i = &\left(\frac{f_s}{c}\right)[((X_{i1} - x)^2 + (Y_{i1} - y)^2 + (Z_{i1} - z)^2)^{\frac{1}{2}} \\ &- ((X_{i2} - x)^2 + (Y_{i2} - y)^2 + (Z_{i2} - z)^2)^{\frac{1}{2}}] \\ &= g_i(\mathbf{x}).\end{aligned} \tag{4.8}$$



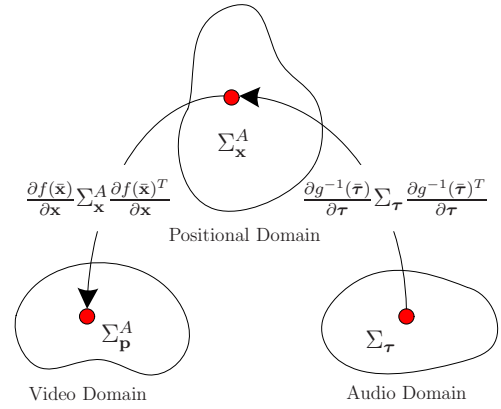
(a) *Audio and video measurement functions.*



(b) *Mapping audio and video uncertainty into* 3D *space. Here* $\Sigma^V_{\mathbf{x}}$ *represents the covariance of a video-based location estimate and* $\Sigma^A_{\mathbf{x}}$ *denotes the covariance of an audio-based location estimate.*



(c) *Mapping video uncertainty into the audio domain. The covariance of a video based* 3D *location estimate in the audio domain is denoted* $\Sigma^V_{\boldsymbol{\tau}}$.



(d) *Mapping audio uncertainty into the video domain. The covariance of an audio based* 3D *location estimate in the video domain is denoted* $\Sigma^A_{\mathbf{p}}$.

Figure 4.1: *The mapping of uncertainty between the audio, video and positional domains through a first order Taylor series expansion of the measurement functions and their inverses. The video measurement function is denoted* $f(\mathbf{x})$ *and the audio measurement function is denoted* $g(\mathbf{x})$. *The corresponding inverse measurement functions are denoted* $f^{-1}(\bar{x})$ *and* $g^{-1}(\bar{\tau})$ *respectively. For clarity in this figure the following notational simplifications are made* $\bar{\mathbf{x}} = E[\mathbf{x}]$, $\bar{\mathbf{p}} = E[\mathbf{p}]$ *and* $\bar{\boldsymbol{\tau}} = E[\boldsymbol{\tau}]$.

In this analysis the sampling frequency used is $f_s = 48kHz$ and the speed of sound is approximated by $c = 343ms^{-1}$. The time delays referred to therefore are in units of audio samples. Using equation 4.8 and equation 4.4, *3D* positional uncertainty can be propagated into the domain of time-delay estimates.

Given equation 4.8 a set of implicit functions can be defined, one for each microphone pair as

$$G_i(\boldsymbol{\tau}_i, \mathbf{x}) = \boldsymbol{\tau}_i - g_i(\mathbf{x}), \tag{4.9}$$

for which $\mathbf{x}^A$ is the *3D* location estimate which minimises the criterion function,

$$C_g(\boldsymbol{\tau}, \mathbf{x}) = \sum_{i}^{N_{mic}} G_i(\boldsymbol{\tau}_i, \mathbf{x})^2. \tag{4.10}$$

Both equation 4.9 and equation 4.7 enable the uncertainty on time delay estimates to be propagated into the *3D* positional domain. In predicting the error region associated with a time-delay based location estimate the relation defined in equation 4.7 has not previously been stated in the literature. As a result, more complex approximate derivations have been proposed [17]. For this reason a complete derivation of equation 4.7 relevant to microphones arrays is presented in appendix A.

4.3 Video-based Measurement Function

The video measurement function relates the *3D* point $\mathbf{x}$ to a vector of pixel measurements $\mathbf{p} = [\mathbf{p}_1, .., \mathbf{p}_i, .., \mathbf{p}_{N_{cam}}]^T$, where $\mathbf{p}_i = [p_{x_i}, p_{y_i}]^T$ is the 2D pixel measurement corresponding to the *ith* camera view. This function is dependent on the camera matrices of the multi-camera views and their associated distortion parameters. Assuming that the distortion characteristics are known and distortion therefore can be corrected, the projection of a *3D* point to the *ith* image plane from equation 2.72, is described by [60],

$$\mathbf{p}_i = \begin{bmatrix} \frac{a^i_{11}x + a^i_{12}y + a^i_{13}z + a^i_{14}}{a^i_{31}x + a^i_{32}y + a^i_{33}z + a^i_{34}} \\ \\ \frac{a^i_{21}x + a^i_{22}y + a^i_{23}z + a^i_{24}}{a^i_{31}x + a^i_{32}y + a^i_{33}z + a^i_{34}} \end{bmatrix} \tag{4.11}$$

$$= f_i(\mathbf{x}) \tag{4.12}$$

where a^i_{uv} is the (u, v) entry in the camera matrix corresponding to the *ith* camera view. This is the non-homogeneous form of the projection of the point $\mathbf{x}$ on to the image plane. The camera matrix's complete form is described in equation 2.72. Using equation 4.11 and equation 4.4, *3D* positional uncertainty can be propagated into the video domain.

Generally, $\mathbf{x}$ is determined as the point $\mathbf{x}^V$ which satisfies the implicit function,

$$F_i(\mathbf{p}_i, \mathbf{x}) = \begin{bmatrix} x(p^i_{31}p_{x_i} - a^i_{11}) + y(a^i_{32}p_{x_i} - a^i_{12}) + z(a^i_{33}p_{x_i} - a^i_{13}) + a^i_{14} \\ x(a^i_{31}p_{y_i} - a^i_{21}) + y(a^i_{32}p_{y_i} - a^i_{22}) + z(a^i_{33}p_{y_i} - a^i_{23}) + a^i_{24} \end{bmatrix} \tag{4.13}$$

such that some criterion function,

$$C_f(\mathbf{p}, \mathbf{x}^V) = \sum_i^N |F_i(\mathbf{p}_i, \mathbf{x}^V)|^2 \tag{4.14}$$

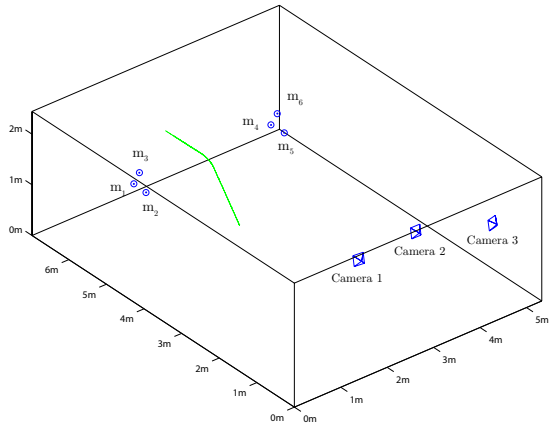
is minimised. Using equation 4.13 and equation 4.7 enables uncertainty on pixel measurements to be propagated into the *3D* positional domain.

4.4 Configuration of Experimental Audio-Video Localisation System

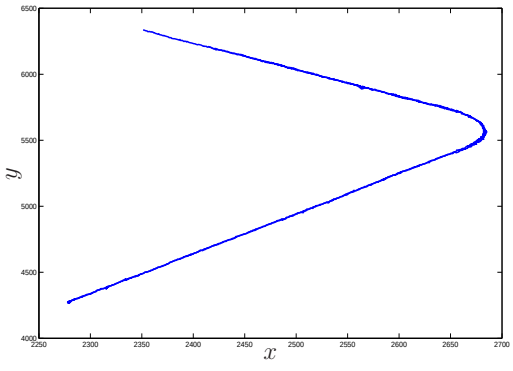
In order to examine *3D* audio and visual localisation accuracy, multiple video cameras and multiple microphones were used to record an audio-visual source moving along a *3D* path. In the analysis, three 720×576 resolution video cameras and six microphones were used. The recordings were conducted in a small lecture room with dimensions $[5.33m, 6.98m, 2.45m]$ and a reverberation time (RT_{60}) of approximately $0.5s$. The six microphones were arranged into two 3-element arrays. The array geometry used was that of a vertical equilateral triangle (dubbed the "triad array" in [88]) with the spacing between the microphones set to $0.34m$. The cameras were positioned within the room so as to create the largest possible *3D* space visible to all cameras. Within this space a track was configured for an audio-visual source to follow. The complete room setup and track of the audio-visual target is illustrated in figure 4.2.

The audio-visual source consisted of a speaker fitted with an LED. In order to maximise the visibility of the target source, the room was darkened and the LED was located in each camera view through intensity thresholding followed by *blob extraction/connected component analysis* [152, chapter 9]. The *centre of mass* of the blob was then taken as the target's pixel location in the frame. The video data was recorded at a frame rate of $25fps$ and a single *3D* location estimate was determined for each frame. The total duration of the recording was 2594 $frames$ or approximately $2min\ 53sec$. The *3D* location estimate was determined by *linear triangulation* using the target's pixel location estimate in each camera view [154, chapter 12]. The estimate obtained satisfied the criterion function defined in equation 4.14. This required calibrating the cameras within the space of the room which is described in the latter parts of this section.

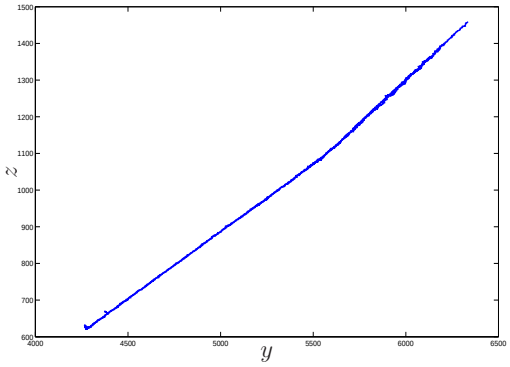
To maximise the accuracy of TDEs at the microphones, Gaussian white noise was output through the target's speaker. Gaussian white noise was chosen on the basis of equation 2.38, where the CRLB on the variance of a TDE is seen to be $\sigma^2_{CRLB} \propto \frac{1}{B^3}$, with B being the source



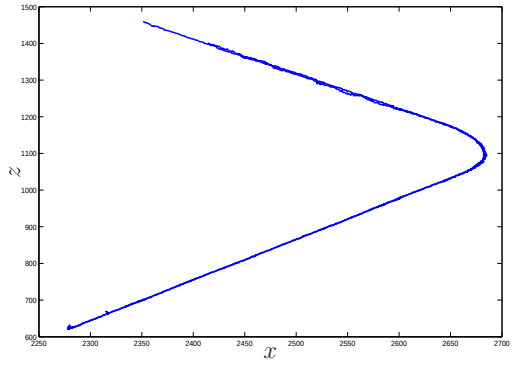
(a) The experimental setup of three cameras and microphones $\{m_1, .., m_6\}$. The track of the audio-visual source is shown in green.



(b) Source path in the xy plane



(c) Source path in the yz plane



(d) Source path in the xz plane

Figure 4.2: *Experimental setup used in the evaluation of audio-visual source localisation. Additional details on the video cameras and microphones, as well as their positions within the room can be found in Appendix B. Also presented in Appendix B is a retrospective note which considers the optimality of the chosen configuration of microphones in the experimental setup.*

signal bandwidth. An audio sampling frequency of $48kHz$ was used and TDEs were obtained using GCC-PHAT [20] from $40ms$ ($25fps$) audio data frames. *Parabolic interpolation* [58] was used in obtaining the TDEs, hence not constraining the estimates to integer values. The audio and video recordings were temporally synchronised using a *clapper board* and the $40ms$ audio frames were centred in relation to corresponding video frames. An audio-based *3D* location estimate was determined at each video frame. Recursive least squares using the Levenberg-Marquardt algorithm and the criterion function of equation 4.10 was used to obtain *3D* location estimates using the TDEs. It should be noted that all possible configurations of microphone pairs in each *triad-array* were used in determining TDEs, however, TDEs corresponding to inter-array microphone pairs were not incorporated into the location estimate.

After the placement of the three video cameras they were fully calibrated within the space

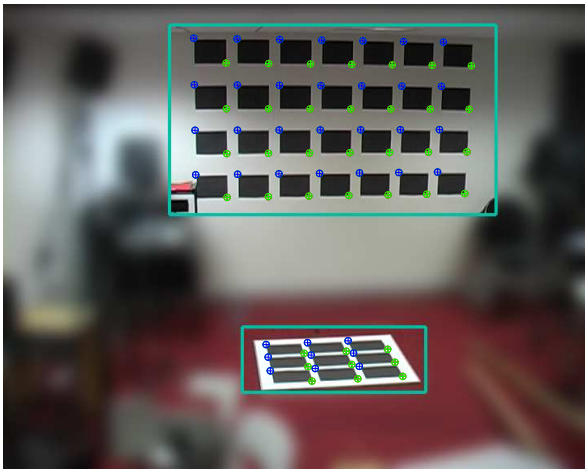
of the room using a set of 74 point correspondences from known *3D* points. The *3D* positions of the points were measured manually using a measuring tape, a square and a level using a single wall as a datum plane. The set of 74 known *3D* point were divided into a *training set* of 37 points and a *test set* also consisting of 37 points. The training set was used to estimate both the intrinsic and extrinsic parameters of each camera and the test set was used to determine the expected accuracy of reconstructed *3D* points. The set of training and test points in each view is shown in figure 4.3 together with the estimated *3D* reconstruction of the points using the three fully calibrated cameras. The overall reconstruction error over the training set was found to be $5mm$ and over the test set it was found to be $13mm$. This level of reconstruction accuracy was found to be comparable to the accuracy of $12mm$ obtained in [109, chapter 4] which utilised a similar number of *3D* calibration points. It is important to emphasise that in the experiment, the set of test points occupied the same two planes as the set of training points. For this reason, the quoted level of localisation accuracy for the set of test points does not truly reflect the expected level of accuracy at locations not lying on two surfaces used for calibration. A more detailed description of the calibration process used is presented is appendix D.

Given both the audio-based and video-based tracking results, further calibration procedures were applied so as to optimise the relative calibration between the audio and video tracking spaces. This was found to be of critical importance so as to ensure no bias existed between audio-based location estimate to that of video-based location estimates. In effect, the relative calibration process determined the positions of the microphones within the coordinate space of the cameras and refined the initial manual measurement of the microphone positions.

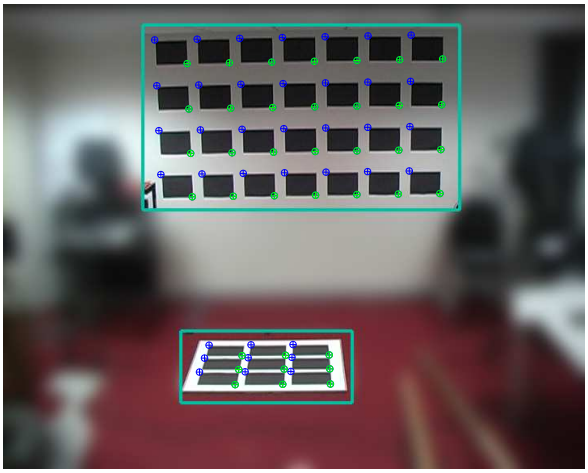
Sufficient reconstruction accuracy was achieved from multi-view visual reconstruction such that any error in localisation relative to audio based localisation was deemed negligible. The visually reconstructed track therefore was taken as the true track's *3D* position. Gaussian noise was added synthetically to pixel measurements to simulate noisy visual localisation. Through this, the covariance $\Sigma_{\mathbf{p}}$ was controlled in the evaluation. In the audio-domain the variance of TDEs at each microphone pair was measured empirically using a *running* variance estimate. The TDEs were assumed to be statistically independent. The covariance Σ_{τ} therefore was formed as a diagonal matrix with the diagonal components being the empirically measured TDE variances.

4.4.1 Validity of First Order Error Propagation

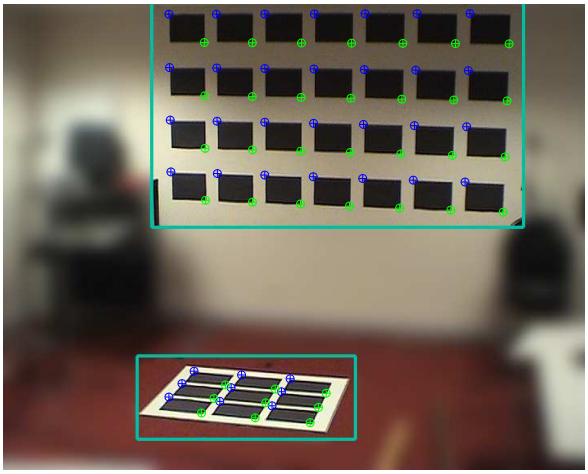
Of particular concern in the application of the error propagation techniques presented in section 4.1.1 is the validity of using a first order Taylor series expansion. In cases where the local linearity assumption is violated, linearisation can introduce errors. To address this concern, the linear mapping techniques described in section 4.1.1 for estimating measurement covariances are compared to that of the Unscented Transform (UT) which can provide a higher order approximation. The violation of a linear approximation therefore should arise in discrepancies between the linear mapping techniques and the UT. In addition to this both estimation techniques are



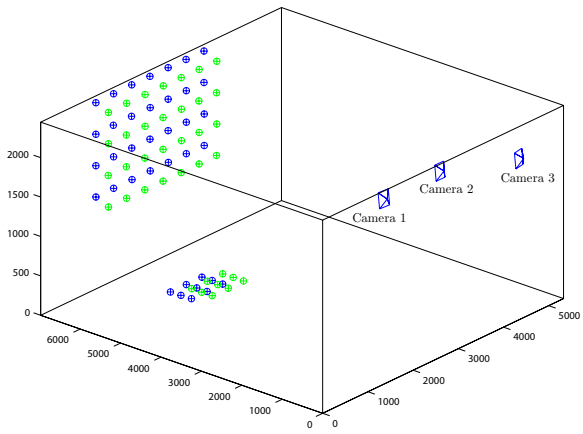
(a) *Camera View 1 with point correspondence* $\mathbf{p}$ *(blue).*



(b) *Camera View 2 with point correspondence* $\mathbf{p}^{'}$ *(blue).*



(c) *Camera View 3 with point correspondence* $\mathbf{p}^{''}$ *(blue).*



(d) *Reconstructed training points (blue) and test points (green) using the three cameras.*

Figure 4.3: *The* 37 *point correspondences* $\{\mathbf{p} \leftrightarrow \mathbf{p}^{'} \leftrightarrow \mathbf{p}^{''}\}$ *(blue) in the three camera views used to calibrate the cameras within the* 3D *space of the room. The point correspondence shown in green are test points used in testing the accuracy of reconstructed* 3D *points. In this figure, the clutter within the room has been blurred so as to highlight the positions of the image point correspondences.*

compared in relation to Monte Carlo based approximation.

4.4.2 Comparison with the Unscented Transform

The Unscented Transform (UT) was briefly introduced in section 3.1.3 as an alternative approach to linearisation for determining the mean and covariance of a random variable that undergoes a nonlinear transformation [167]. It was described in section 3.1.3 how a set of sigma points $\mathcal{X}_i$, $i = 1, .., p$, with associated weights $\mathcal{W}_i$ can be used to adequately approximate the probability

distribution of a random variable. The set of sigma points of the UT are chosen deterministically. In this respect the UT differs from a Monte Carlo approximation where sample points are determined randomly and the true estimate is only obtained asymptotically as the number of samples approaches infinity [118]. The sigma points therefore represent a minimal approximation of the random variable's statistics. The effect of the nonlinear transformation on the random variable is estimated by applying the non-linear transformation to the set of sigma points. The *a posteriori* statistics of the transform random variable are then estimated using the set of transformed sigma points.

In the following the UT is examined in more detail where it is applied to the problem of propagating uncertainty in the audio-video based localisation system. To examine the use of the UT in this regard, we will consider again the problem in section 4.1.1 of estimating the covariance $\Sigma_{\mathbf{p}}$ of a pixel-based observation $\mathbf{p}$ of the *3D* point $\mathbf{x}$, with associated covariance $\Sigma_{\mathbf{x}}$. The set of sigma points are determined through the following [164],

$$\mathcal{X}_0 = E[\mathbf{x}] \tag{4.15a}$$

$$\mathcal{W}_0 = \frac{\kappa}{(p+\kappa)} \tag{4.15b}$$

$$\mathcal{X}_i = E[\mathbf{x}] + (\sqrt{(p+\kappa)\Sigma_{\mathbf{x}}})_i \tag{4.15c}$$

$$\mathcal{W}_i = \frac{1}{2(p+\kappa)} \tag{4.15d}$$

$$\mathcal{X}_{i+p} = E[\mathbf{x}] - (\sqrt{(p+\kappa)\Sigma_{\mathbf{x}}})_i \tag{4.15e}$$

$$\mathcal{W}_{i+p} = \frac{1}{2(p+\kappa)} \tag{4.15f}$$

where $\kappa \in \mathbb{R}$, p is the dimension of $\mathbf{x}$ (which in this example is $p = 3$) and $(\sqrt{(p+\kappa)\Sigma_{\mathbf{x}}})_i$ is the *ith* row of the matrix square root of $(p+\kappa)\Sigma_{\mathbf{x}}$. In this definition, it is assumed that the square root of a symmetric matrix $\mathbf{B}$ is defined as $\mathbf{A}$ such that $\mathbf{B} = \mathbf{A}^T\mathbf{A}$. The set of sigma points as defined in equation 4.15 is a symmetric set of *3D* points which lie on the $\sqrt{(p+\kappa)}$ contour of the covariance $\Sigma_{\mathbf{x}}$. In the case of $\Sigma_{\mathbf{x}}$ being a diagonal covariance matrix, the sigma points lie on the principal axes of the error ellipsoid of $\mathbf{x}$. The sigma points $\mathcal{X}_i$ are then propagated through the video-based measurement function $f(\cdot)$ by,

$$\mathcal{Y}_i = f(\mathcal{X}_i). \tag{4.16}$$

Given the transformed sigma points $\mathcal{Y}_i$ therefore, the expected value is determined using the weights $\mathcal{W}_i$ as,

$$E[\mathcal{Y}_i] = \sum_{i=0}^{2p} \mathcal{W}_i \mathcal{Y}_i \tag{4.17}$$

and associated covariance by,

$$\Sigma_{\mathcal{Y}_i} = \sum_{i=0}^{2p} \mathcal{W}_i \left[\mathcal{Y}_i - E[\mathcal{Y}_i]\right] \left[\mathcal{Y}_i - E[\mathcal{Y}_i]\right]^T . \tag{4.18}$$

In this way an approximation of both $E[\mathbf{p}]$ and $\Sigma_{\mathbf{p}}$ is obtained as,

$$E[\mathbf{p}] \approx E[\mathcal{Y}_i], \tag{4.19}$$

$$\Sigma_{\mathbf{p}} \approx \Sigma_{\mathcal{Y}_i}. \tag{4.20}$$

In the case of $\mathcal{W}_0 = 0$ the UT enables the covariance to be estimated to the same order of accuracy as a first order linearisation approach (i.e. EKF) [166]. For the case of a Gaussian random variable $\mathbf{x}$, it has been shown that in choosing $(p+\kappa) = 3$ the UT determines an estimate correct up to the third order [167, Appendix I].

Shown in figure 4.4a are the results of 1000 Monte Carlo simulations for localisation at 12 positions denoted $\{A, B, .., L\}$ in the room from noisy time-delay estimates. Also shown in this figure are the predicted 95 percentile error regions for each position as determined by linear covariance mapping and the UT. The variance of time-delay estimates in the simulations was set to 1 audio sample in each case.

The first observation in these results is that both linear covariance mapping and UT produce similar uncertainty estimates with no significant difference at any of the 12 positions. This can further be seen in Table 4.1a where the principal components of the covariance estimates for the three different estimation techniques are quoted for positions $\{C, F, I, L\}$. In relation to the Monte Carlo estimate of localisation uncertainty, both techniques underestimate localisation uncertainty at positions with relatively high uncertainty such as at position C and overestimate uncertainty at positions with relatively low localisation uncertainty such as at position I.

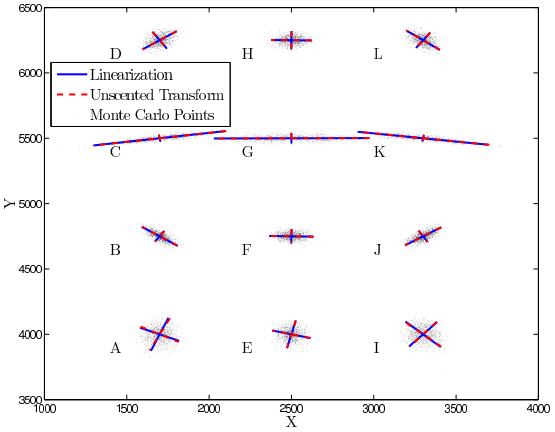
Figure 4.4b presents the above analysis for the case of video-based localisation using noisy pixel measurements. In the simulations the variance of pixel measurements was set to 5 $pixels^2$ in both x and y image axes for each of the 12 positions within the room. Again in this case both the UT and linear covariance mapping result in similar uncertainty estimates. For the positions $\{C, F, I, L\}$, the principal components of the estimates from Table 4.1b are seen to be identical. In relation to the covariance estimates obtained using Monte Carlo simulation however, both the UT and linear covariance mapping tend to result in overestimating localisation uncertainty. To illustrate the relative scale of the associated error for both audio-based and video-based locaisation, the *3D* error ellipsoids are shown in relation to the positions of the sensors within the room in figure 4.4e.

Shown in figure 4.4c are the audio based localisation estimates mapped into the image plane with the predicted 95 percentile error ellipses. In this example, the principal components of the positions $\{C, F, I, L\}$ are shown in Table 4.1c. For this particular case, the UT is seen to provide

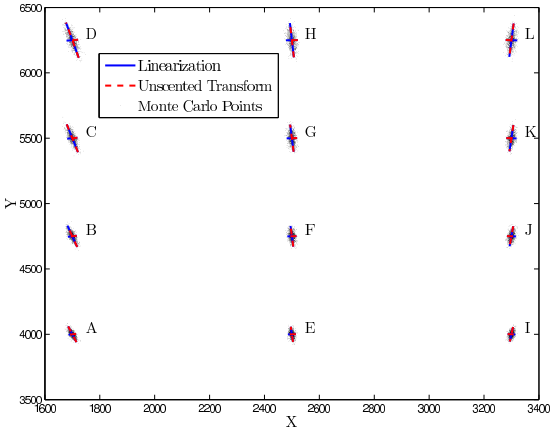
more accurate estimates than the linear covariance mapping approach. This is most significant for position I.

Finally, in figure 4.4d,the result of mapping the video based location estimates to time delays in the audio domain is shown. Also shown are the predicted error bars representing the 95 percentile error regions associated with the mapped time delays. The estimated principal components from Table 4.1d using the UT, linear covariance mapping and Monte Carlo simulation in this case are seen to be identical for the positions $\{C, F, I, L\}$.

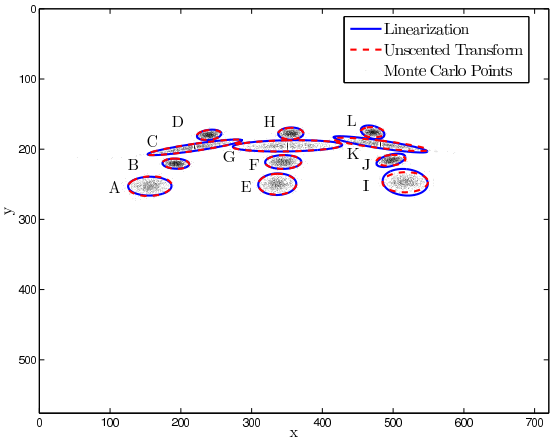
In summary, the simulation shows that linear covariance mapping, performs equally as well as the UT. This suggests that under the assumption of Gaussian noise, first order linearisation is sufficient for the propagation of uncertainty in the considered audio-video based system.



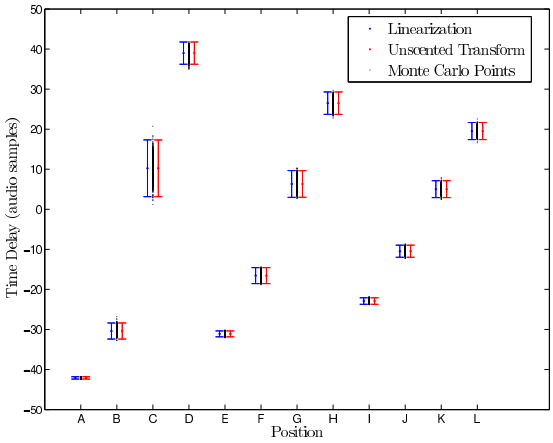
(a) *Audio-based localisation error in the xy-plane.*



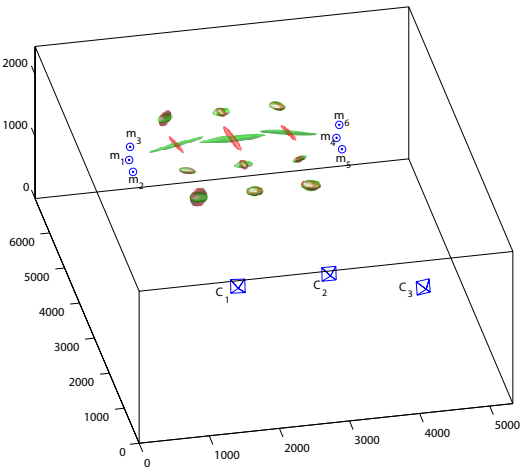
(b) *Video-based localisation error in the xy-plane.*



(c) *Audio-based localisation error mapped into the image plane of camera C_2.*



(d) *Video-based localisation error mapped to time delay domain of microphone pair $\{m_1 - m_2\}$ where the error bars are offset for clarity.*



(e) 3D *Error ellipsoids for audio (green) and video (red) based localisation.*

Point	Position		
	x (mm)	y (mm)	z (mm)
A	1700	4000	1500
B	1700	4750	1500
C	1700	5500	1500
D	1700	6250	1500
E	2500	4000	1500
F	2500	4750	1500
G	2500	5500	1500
H	2500	6250	1500
I	3300	4000	1500
J	3300	4750	1500
K	3300	5500	1500
L	3300	6250	1500

(f) 3D Position of each test point in the room

Figure 4.4: *Validating first order error propagation in the configuration of six microphones $\{m_1, m_2, .., m_6\}$ and three cameras $\{C_1, C_2, C_3\}$ using linear uncertainty mapping, the UT and Monte Carlo simulation. Localisation error is examined for twelve positions denoted $\{A, B, ..L\}$ within the room. The 95 percentile error region is illustrated in each case.*

Pos.	Standard Deviation of the Principal Components of the Covariance Estimates (mm)					
	L		**MC**		**UT**	
C	1^{st}	161.9	1^{st}	170.0	1^{st}	164.1
	2^{nd}	12.2	2^{nd}	12.5	2^{nd}	12.2
	3^{rd}	11.0	3^{rd}	11.4	3^{rd}	11.1
F	1^{st}	53.8	1^{st}	52.9	1^{st}	53.8
	2^{nd}	22.8	2^{nd}	23.3	2^{nd}	20.1
	3^{rd}	19.2	3^{rd}	20.1	3^{rd}	19.2
I	1^{st}	57.9	1^{st}	57.4	1^{st}	58.2
	2^{nd}	50.7	2^{nd}	49.2	2^{nd}	50.8
	3^{rd}	22.7	3^{rd}	22.5	3^{rd}	22.7
L	1^{st}	50.3	1^{st}	52.1	1^{st}	50.4
	2^{nd}	29.7	2^{nd}	30.9	2^{nd}	29.7
	3^{rd}	16.9	3^{rd}	16.4	3^{rd}	16.9

(a) *Audio-based* 3D *localisation error.*

Pos.	Standard Deviation of the Principal Components of the Covariance Estimates (mm)					
	L		**MC**		**UT**	
C	1^{st}	43.7	1^{st}	42.9	1^{st}	43.7
	2^{nd}	7.6	2^{nd}	7.5	2^{nd}	7.6
	3^{rd}	7.4	3^{rd}	7.1	3^{rd}	7.4
F	1^{st}	32.1	1^{st}	30.7	1^{st}	32.1
	2^{nd}	6.7	2^{nd}	6.9	2^{nd}	6.7
	3^{rd}	6.4	3^{rd}	6.3	3^{rd}	6.4
I	1^{st}	22.8	1^{st}	22.4	1^{st}	22.8
	2^{nd}	5.6	2^{nd}	5.7	2^{nd}	5.6
	3^{rd}	5.4	3^{rd}	5.2	3^{rd}	5.4
L	1^{st}	51.6	1^{st}	52.4	1^{st}	51.6
	2^{nd}	8.5	2^{nd}	8.5	2^{nd}	8.5
	3^{rd}	8.0	3^{rd}	7.7	3^{rd}	8.0

(b) *Video-based* 3D *localisation error.*

Pos.	Standard Deviation of the Principal Components of the Covariance Estimates (pixels)					
	L		**MC**		**UT**	
C	1^{st}	26.3	1^{st}	27.5	1^{st}	26.5
	2^{nd}	2.0	2^{nd}	2.0	2^{nd}	2.0
F	1^{st}	9.7	1^{st}	9.6	1^{st}	9.7
	2^{nd}	3.9	2^{nd}	3.9	2^{nd}	3.8
I	1^{st}	12.9	1^{st}	12.3	1^{st}	12.6
	2^{nd}	7.8	2^{nd}	5.3	2^{nd}	5.3
L	1^{st}	7.0	1^{st}	7.1	1^{st}	6.8
	2^{nd}	3.5	2^{nd}	2.8	2^{nd}	2.7

(c) *Audio based* 3D *localisation propagated into the image plane.*

Pos.	Standard Deviation of the Principal Components of the Covariance Estimates (audio samples)					
	L		**MC**		**UT**	
C	1^{st}	2.8	1^{st}	2.8	1^{st}	2.8
F	1^{st}	0.8	1^{st}	0.8	1^{st}	0.8
I	1^{st}	0.3	1^{st}	0.3	1^{st}	0.3
L	1^{st}	0.9	1^{st}	0.9	1^{st}	0.9

(d) *Video-based localisation propagated into audio domain (time-delays).*

Table 4.1: *The standard deviation of the principal components of the covariance estimates for the labelled positions in the simulation example illustrated in figure 4.4. The abbreviations* ***L****,* ***MC*** *and* ***UT*** *are use for Linearisation, Monte Carlo Simulation and Unscented Transform respectively. These terms are used to refer to the technique by which the covariance estimate was obtained. Also,* 1^{st}*,* 2^{nd} *and* 3^{rd} *are used to label the first, second and third principal components of the covariance estimate respectively*

4.5 Comparative Error Analysis and Discussion

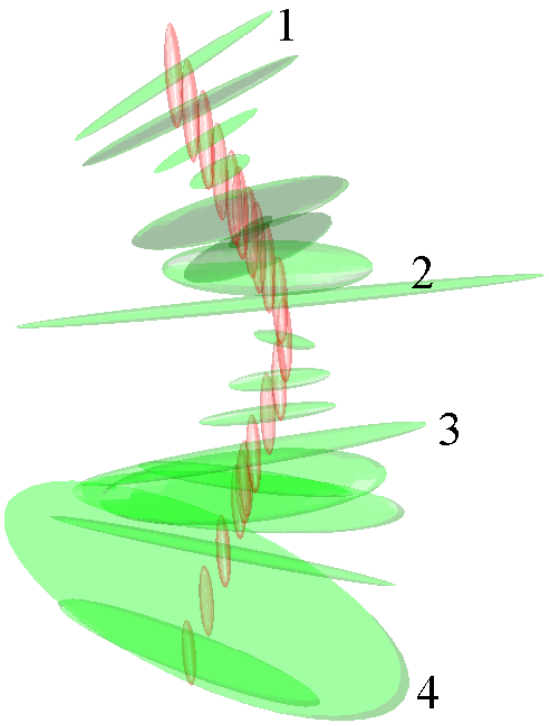
In this section, both audio-based and video-based localisation are examined in determining location estimates for the scenario of a moving audio-visual source as described in section 4.4. Also evaluated is the fused audio-video based estimate using equation 4.1 and equation 4.2 for the different cases of fusion applied in the audio domain, the video domain and the positional domain.

Using the covariance mapping techniques presented in section 4.1.1 and given the covariance of time delay estimates, the error associated with audio-based localisation can be determined. Similarly, given the covariance of pixel based measurements, the error associated with video-based localisation can be determined. Through this, for the given audio-visual localisation system, the uncertainty of the various localisation approaches can be evaluated.

Shown in figure 4.5a are the predicted 95 percentile error ellipsoids for points along the path of the audio-visual source. The error ellipsoids are determined for the case of pixel measurements with variance equal to 5 $pixels^2$ in both the x and y image axes respectively and TDEs with variances determined empirically by a running variance estimate. From this, it can be seen that in each case of both audio-based and video-based localisation, the error associated with a location estimate is non-uniform in space. Also, the orientation of the error regions is dependent on the configuration of the sensors. Since the error ellipsoids are not uniform in space, direct comparison of audio-based and video-based localisation uncertainty cannot be made. The trace, denoted $tr(\cdot)$, of the covariance matrix defining the error ellipsoids is therefore chosen as a performance measure for the accuracy of localisation.

Presented in figure 4.5b is a table quoting this measure of accuracy for both audio-based and video-based location estimates at the numbered points along the track as shown in figure 4.5a. Also included in the table are the expected values of *3D* localisation accuracy for the three different fusion strategies. From this, it can be seen that, as expected, the accuracy of the fused audio-video based location estimate is greater than either audio-based or video-based localisation alone. It is clear however that the fused estimate shows a greater improvement on audio-based localisation than that of video-based localisation. This highlights the greater contribution of video measurements in the fused estimates in comparison to that of audio measurements. Furthermore it is seen that, of the fusion strategies examined, the best localisation estimate is obtained through fusion in the positional domain with the worst localisation accuracy being observed for fusion in the video domain. In particular, in the case of fusion in the video domain, there is little improvement in accuracy beyond video-based localisation. This suggests that in the image plane the contribution of audio data in improving localisation accuracy is small.

Again, using the trace of the localisation covariance matrix as a measure of accuracy, audio-based localisation is compared to simulated video-based localisation in figure 4.6a. In this figure, the percentage of frames where audio-based localisation was found to be more accurate than video-based localisation for varying pixel measurement noise is shown. From this it is seen



(a) *xy* plane view of the 95 percentile error ellipsoids over the track length.

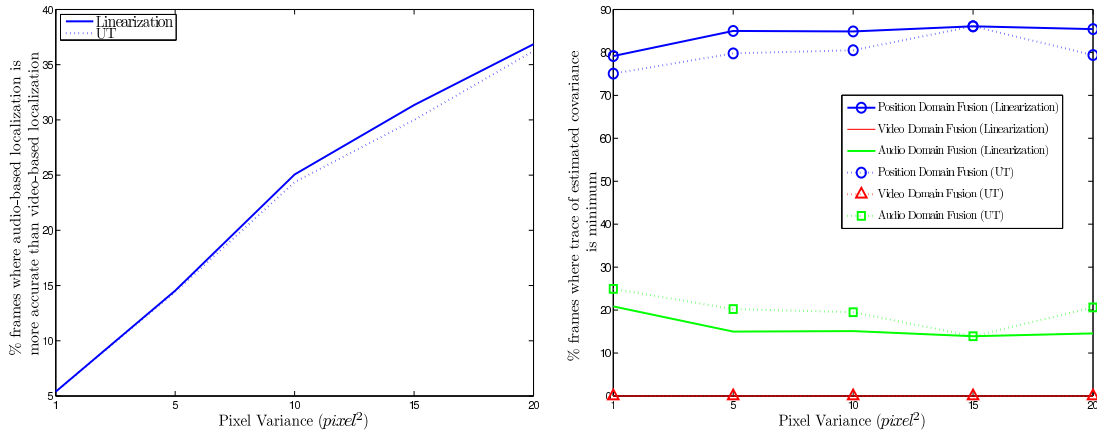
Location	1	2	3	4
$tr(\Sigma_{\mathbf{x}}^{A})$ (cm^2)	178	717	328	639
$tr(\Sigma_{\mathbf{x}}^{V})$ (cm^2)	29	21	17	14
$tr(\Sigma_{\mathbf{x}}^{Aud})$ (cm^2)	2	17	13	9
$tr(\Sigma_{\mathbf{x}}^{Vid})$ (cm^2)	18	20	16	13
$tr(\Sigma_{\mathbf{x}}^{Pos})$ (cm^2)	2	1	2	7

(b)

Figure 4.5: *The 95 percentile error ellipsoids for audio-based localisation (red) and video-based localisation (green) over the track length are shown in (a). Shown in (b) is the trace of the covariance matrices corresponding to the numbered error ellipsoids in (a).*

that even with a pixel measurement noise of 20 $pixel^2$ the percentage of frames where audio localisation is more accurate over the track duration is less than 40%. Although this is the case, for pixel measurement noise with variance 1 $pixel^2$, audio-based localisation was found to be more accurate for 5% of the track duration. This reveals significant variation in audio-based localisation accuracy over a typical track duration.

Figure 4.6b further examines the predicted performance of the different fusion strategies in localising the target for the case of varying pixel uncertainty. Presented in this figure is the percentage of frames over the track duration where the minimum localisation accuracy is achieved by each fusion strategy. This analysis is based on the predicted covariances associated with the obtained location estimates. The results are shown for both cases of estimating the covariances using linearisation and the UT. Figure 4.6b indicates therefore, that fusion in the positional domain provides the most accurate location estimate regardless of the pixel measurement uncertainty. The figure also indicates that, irrespective of the variance of pixel measurements, at no point along the track does fusion in the video domain provide the most accurate location estimate. One contributing factor, in the result that positional domain fusion is best, is the configuration of the sensors. Examining figure 4.5a again, it can be seen that for audio-based



(a) *Percentage of frames where audio based localisation was found to be more accurate than video-based localization for pixel measurements with increasing variance.*

(b) *Percentage of frames where the trace of the estimated covariance is minimum for each fusion strategy.*

Figure 4.6: *Comparison of audio-visual source localisation and various fusion based localisation strategies.*

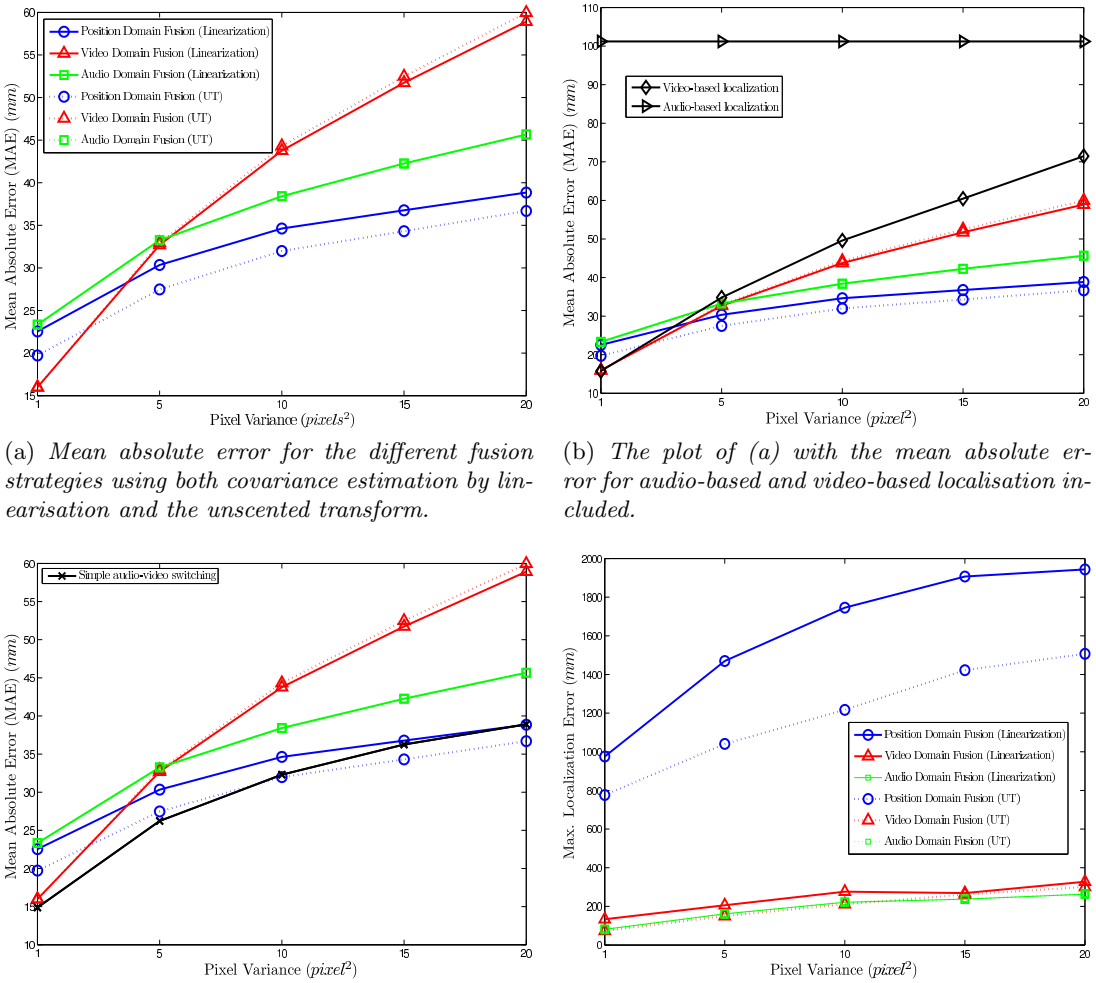
localisation, uncertainty is greatest along the x-axis and least along the y-axis whereas in the case of video-based uncertainty the opposite is true.

Until now the analysis has been confined to examining localisation accuracy based on estimates of the covariance associated with the different location estimates. In figure 4.7 the accuracy of the actual location estimates are examined against the ground truth track data. In each of the presented figures, the results are shown for the fused location estimates obtained using covariance estimates by linearisation and the UT.

In figure 4.7a the Mean Absolute Error (MAE) [2] of localisation over the track duration is shown for the different fusion strategies. It can be seen from this that the results correspond to the previous theoretical analysis and the best localisation accuracy is obtained through fusion in the positional domain with the least accurate fusion strategy being fusion in the video domain. This result however, is only observed for pixel measurement uncertainty greater than 5 $pixels^2$.

Figure 4.7b shows the MAE of audio-based and video-based localisation in relation to the results of the various fusion-based approaches. This figure shows that even with pixel measurement uncertainty of 20 $pixels^2$ a fused estimate greatly improves upon the results of audio-based localisation and to a lesser extent, video-based localisation. One unexpected result occurs for pixel measurement uncertainty below 5 $pixels^2$ where the MAE for the fusion approaches are seen to be greater than that of video-based localisation. The reason for the transition at the value of 5 $pixel^5$ is attributed to a possible bias still existing between audio-based and video-based location estimates despite the relative calibration of the tracking spaces as described in

[2] [191] highlights the ambiguous use of the term "absolute error" in relation to vectors and recommends instead "Euclidean error". In this thesis the term "absolute error" is adhered to, however, to avoid misinterpretation the term "absolute error" when referred to is equivalent to the "Euclidean error".



(a) *Mean absolute error for the different fusion strategies using both covariance estimation by linearisation and the unscented transform.*

(b) *The plot of (a) with the mean absolute error for audio-based and video-based localisation included.*

(c) *The plot of (a) with the mean absolute error for the simple audio-video based switching strategy included.*

(d) *The maximum error over the track duration for increasing pixel measurement uncertainty.*

Figure 4.7: *Comparison of audio-based, video-based and fusion-based localisation performance.*

section 4.4. This highlights the need for more accurate means of relative calibration between audio and video sensors in the implementation of audio-visual fusion systems. In the absence of more accurate calibration techniques than those employed in this analysis, at a level of pixel measurement uncertainty below 5 $pixels^2$, fusing audio-based measurements with video-based measurements is not worthwhile.

The use of audio-video based fusion is seen to improve upon the accuracy of single modality localisation. This however does not give any indication as to whether the fused estimate improves localisation accuracy beyond the best available sequence of single modality location estimates. For instance, consider the hypothetical scenario where the best single modality location estimate is known at each frame and that a simple switch strategy is employed to choose the most accurate estimate. Such a strategy determines the optimum sequence of single modality location estimates

over the track duration. Since the true track location is known in this analysis, the simple audio-video based switching strategy can be examined. This can be used to evaluate fusion-based localisation approaches since they should outperform the simple switching strategy. The MAE of the simple switch-based joint audio-video approach in localising the moving source is shown in figure 4.7c. For comparison, the MAE for the various fusion strategies is also included. From this it can be seen that the accuracy of the simple switching-based approach is only achieved by fusion in the positional domain at relatively large values of pixel uncertainty in the region of 15 $pixels^2$ to 20 $pixels^2$.

An interesting observation is made when the different fusion strategies are examined in relation to the maximum observed error over the track duration. From figure 4.7d it is seen that for all values of pixel measurement uncertainty, the largest observed error occurs in the localisation results corresponding to positional domain fusion and is significantly larger than that observed for audio domain fusion or video domain fusion. This result is in contrast to the previous observation that the best overall localisation accuracy over the track duration is obtained through positional domain fusion. This suggests that large localisation errors occur in the positional domain fusion approach, but the frequency of these errors is low enough such that the overall performance is not greatly affected.

The reason for this result can be deduced from a closer examination of the x, y and z localisation results in figure 4.8 where the various localisation approaches are examined. This figure includes; figure 4.8a showing audio-based localisation; figure 4.8b showing video-based localisation; figure 4.8c showing localisation resulting from fusion in the positional domain; figure 4.8d showing localisation resulting from fusion in the video domain and figure 4.8e showing localisation resulting from fusion in the audio domain.

In examining figure 4.8a it is seen that large errors are observed over the track duration for audio-based localisation. These large errors are due to errors on the TDEs which are propagated from the audio measurement domain into the *3D* location estimates. Even small errors in TDEs can introduce significant localisation errors. In a similar manner, noisy pixels measurements can introduce significant localisation errors when propagated from the video measurement domain into *3D* positional space. This can be seen from the noisy video-based localisation estimates of figure 4.8b.

When audio and video based location estimates are fused in the positional domain, large errors in either of the two location estimates will result in a large error in the overall fused estimate. From figure 4.8a it can be seen that there are many instances over the track duration where audio-based localisation accuracy is low. This means that when a poor video-based location estimate occurs, there is a high probability that it will be fused with a poor audio-based location estimate. When such situations arise, the error in the resulting fused location estimate is large. This phenomenon is evident in the large localisation errors seen in the results of positional domain fusion as shown figure 4.8c.

To apply fusion in the video domain the audio-based location estimates are simply propagated

into the image plane of the cameras. This projection does not reduced the errors in the audio-based location estimates. As a result, some large errors are also apparent in the localisation results of video domain fusion as can be seen in figure 4.8d.

It is interesting to note that the localisation results are significantly smoother when fusion is applied in the audio domain as can be seen in figure 4.8e. The reason for this observation is that when fusion is applied in the audio domain, the more reliable video measurements act to reduce the large errors on the TDEs before the *3D* location estimate is determined. Reducing the error on each individual TDE significantly reduces the total error observed when all TDEs are combined into a *3D* location estimate. In the other fusion approaches considered, such as positional domain fusion and video domain fusion, the errors arising from the audio data are seen to be most dominant. The results presented here suggest that applying fusion in the audio domain has the greatest effect on reducing the localisation errors arising from uncertainty on the TDEs. This means that fusion in the audio-domain results in smoother localisation over the track duration than either of the two other fusion approaches considered. However, when the uncertainty on the TDEs is low, fusion in the audio-domain results in less accurate localisation when compare to localisation through positional domain fusion or fusion in the video-domain.

4.6 Final Comments

The use of error propagation was presented in this paper as a useful means of evaluating the performance of a joint audio-video based localisation system. Through comparative studies with the UT and Monte Carlo simulation, first order error propagation was shown to adequately map audio and video measurement uncertainty across the positional, video and audio tracking domains. In the comparison of audio-based and video-based localisation accuracy, video-based localisation was found to outperform that of audio-based localisation in terms of accuracy and consistency. Maximum likelihood fusion of location estimates in the audio domain, video domain and positional domain was examined. In general, the appropriateness of each fusion strategy is dependent on the configuration of the sensors. For the examined configuration of sensors in this analysis, the best *3D* localisation accuracy was found to be achieved where fusion is applied in the positional domain. Audio was found to contribute little in terms of localisation accuracy when fusion is applied in the video domain.

In addition to this, the performance of the three different ML fusion-based techniques was found to be poor when compared to the simple audio-video based switching localisation approach. In particular the use of fusion was only found to approach the localisation performance of the simple switch-based approach for relatively large pixel measurement uncertainty.

In contrast to the best overall tracking performance being obtained through fusion in the positional domain, the largest observed error was also found to occur through this fusion-based approach. Fusing video measurements with audio measurements in the audio domain was found to reduce the uncertainty on the TDEs. Applying fusion at the level of the TDEs resulted in a

smoother sequence of location estimates but they were found to be less accurate.

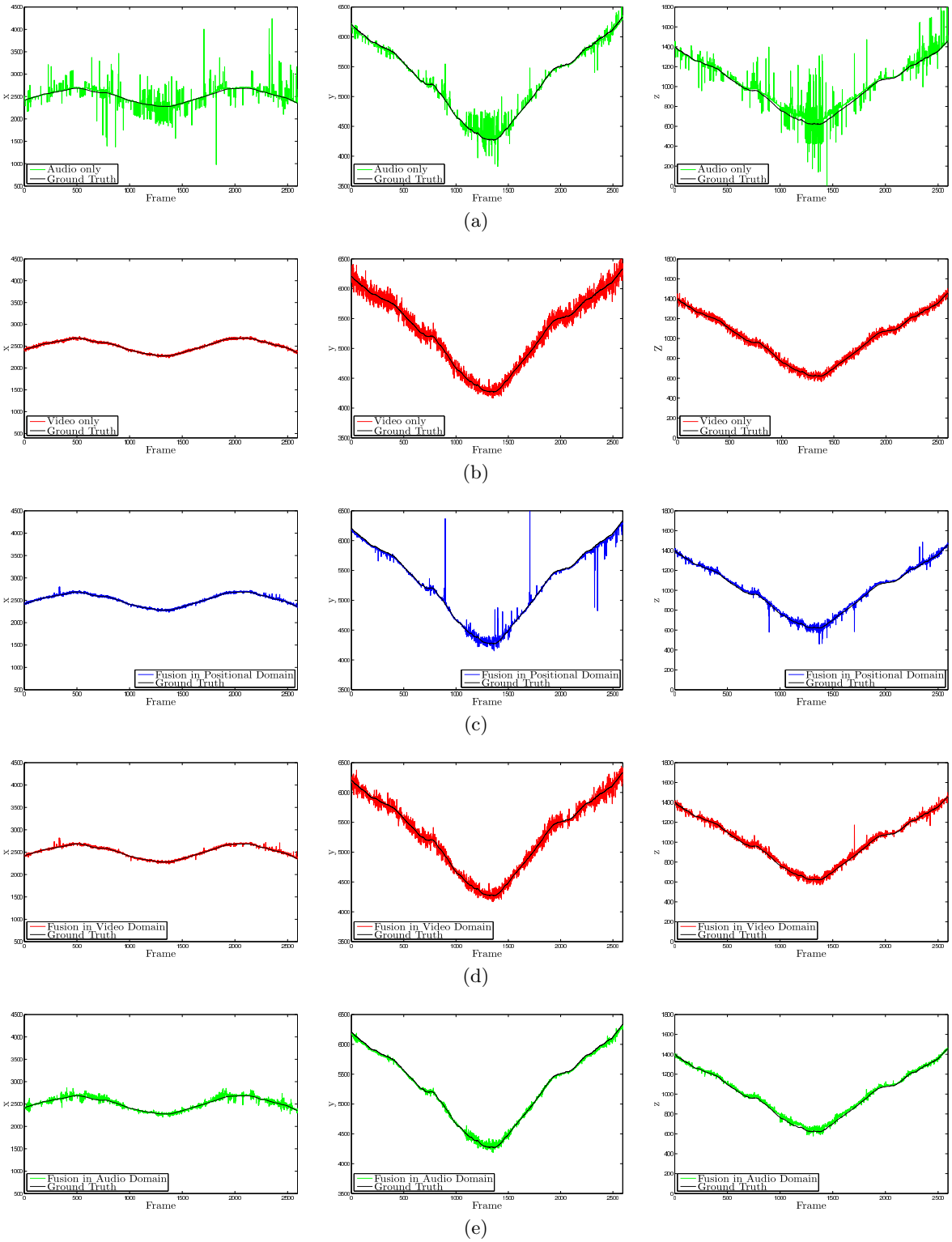


Figure 4.8: *The x, y and z localisation results of (a) audio-based tracking, (b) video-based tracking, (c) fusion in the positional domain, (d) fusion in the video domain and (e) fusion in the audio domain.*

5

Optimal Microphone Placement[1]

There is a need in joint audio-video based systems for a practical means for determining the optimal placement of microphone arrays within a room for Time Delay Estimate (TDE) based localisation. In the context of audio-based localisation this is important since it improves the accuracy of location estimates. In relation to joint audio-video based systems such as that evaluated in chapter 4 however, the need to optimise the placement of arrays becomes more pronounced. This is seen from examining the fused location estimate of equation 4.1 where reducing the audio-based localisation uncertainty can be seen to improve the contribution of the audio modality in the joint estimate.

Earlier approaches to optimal sensor placement considered the performance of sensors which are constrained to specific geometries such as linear sensor arrays [29, 117] and also circular arrays [162]. Now such array geometries are in common usage and a new problem has recently arisen to determine where multiple such arrays should be placed within a room to minimise localisation uncertainty.

In the sensor placement literature, some relevent work exists such as that by Abel [98] and more recently by Zhang [76] and Yang et al. [16]. However, each of these works only consider the single source case. This does not address the problem at hand since the most useful applications of microphone arrays is to the multiple source case.

Literature which focuses on optimal sensor placement for the multi-source case does exist

[1] Results from this chapter have been published in: Damien Kelly, Frank Boland. Optimal Microphone Placement for Active Speaker Localization in *8th IMA International Conference on Mathematics in Signal Processing. Dec. 16th-18th, 2008, Cirencester, UK. [40]*

[46, 95, 137]. None of these treatments of the problem incorporate satisfactory models of TDE uncertainty where the errors on the TDEs are assumed to be equal and independent of the sensor positions. This assumption neglects the fact that the positions of the sensors in relation to the source affect aspects such as the SNR which has a direct influence on the accuracy of the TDEs [6]. Furthermore, the specific problem of speaker localisation is not addressed in these works.

To the best knowledge of the author, the work of Brandstein et al. [17] is the only explicit treatment of the topic of optimal microphone array placement for speaker localisation. They derive the localisation error for a given array geometry and model the variance of TDEs as a function of the distance from the speaker to the microphones and also of the speaker/microphone directionality characteristics. Given this definition of localisation error, they determine the optimal microphone array configuration in a conference room setting from a small set of possible configurations. No algorithm for automatically determining an optimal configuration however is given.

In this chapter it is aimed to extend on the work in [17] to determine the optimal placement of microphone arrays within a room so as to minimise the localisation uncertainty over an audience area. Drawing from the work of Gustafsson et al. [175] a more appropriate model of TDE uncertainty in a reverberant room is employed and a simulated annealing algorithm is proposed for automatically determining an optimal configuration of microphone arrays.

Problem Statement

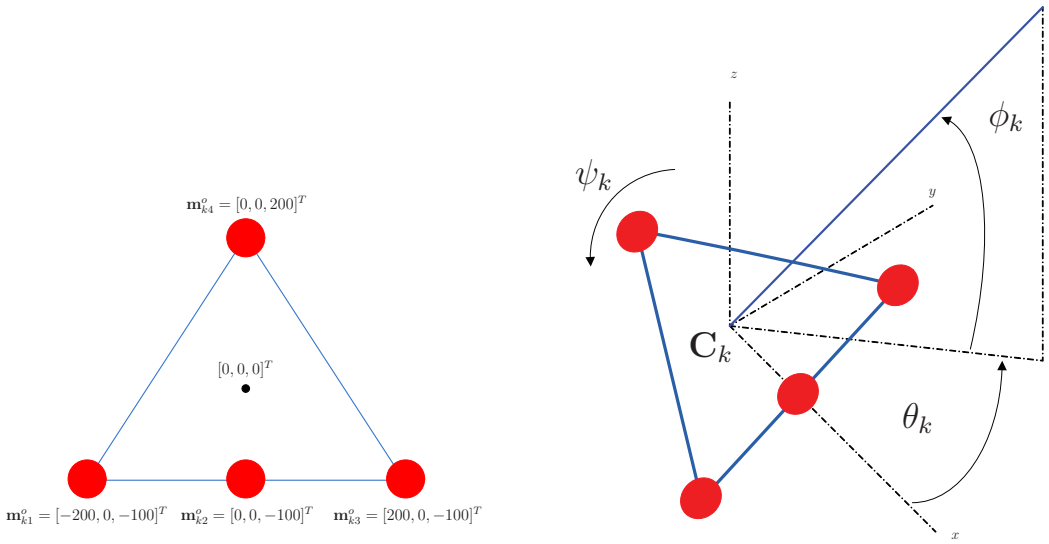
For a defined set of speaker positions $\mathbf{x} = \{\mathbf{x}_1, .., \mathbf{x}_i, .., \mathbf{x}_{N_s}\}$ within a room, the task is to determine the optimal placement of $k = 1, ..., K$ microphone arrays such that the localisation error at each speaker position is minimised. Each microphone array consists of a known but unrestricted array geometry and number of microphones. This information is assumed to be encompassed in a set of microphone positions

$$\mathbf{M}_k^o = \{\mathbf{m}_{k1}^o, .., \mathbf{m}_{kj}^o, ..\mathbf{m}_{kN_{mic}}^o\} \tag{5.1}$$

centred about the origin, where $\mathbf{m}_{kj}^o = [X_{kj}, Y_{kj}, Z_{kj}]^T$ is the position of the jth microphone and N_{mic} is the total number of microphones in the array. An example of the form of $\mathbf{m}_k^o$ for an inverted T-shaped microphone array is presented in figure 5.1a. With this knowledge defining the relative positions of microphones within the array, the placement of the array within the room can be completely described by

$$\mathcal{M}_k = \{\mathbf{C}_k, \mathbf{O}_k\}, \tag{5.2}$$

where $\mathbf{C}_k$ is the array's centre and $\mathbf{O}_k = [\theta_k, \phi_k, \psi_k]$ is a vector of orientation angles. The orientation angles θ_k, ϕ_k and ψ_k denote angles of pan, tilt and roll respectively. An illustrated



(a) (b)

Figure 5.1: *An example of an inverted T-shaped microphone array defined by* $\mathbf{m}^o_k = \{[-200, 0, -100]^T, [0, 0, -100]^T, [200, 0, -100]^T, [0, 0, 200]^T\}$ *is shown in (a). In (b) the parameterisation of microphone positions in the array by a centre position* $\mathbf{C}_k$ *and orientation vector* $\mathbf{O}_k = [\theta_k, \phi_k, \psi_k]$ *is illustrated.*

example of this parameterisation for the inverted T-shaped microphone array of figure 5.1a is shown in figure 5.1b. For any value of the parameters $\mathbf{C}_k$ and $\mathbf{O}_k$, the positions of the microphones of the kth array are defined by,

$$\mathbf{m}_{kj} = \mathbf{R}_\theta \mathbf{R}_\phi \mathbf{R}_\psi \mathbf{m}^o_{kj} + \mathbf{C}_k, \tag{5.3}$$

where

$$\mathbf{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5.4}$$

$$\mathbf{R}_\phi = \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix}, \tag{5.5}$$

$$\mathbf{R}_\psi = \begin{bmatrix} 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \\ 1 & 0 & 0 \end{bmatrix} \tag{5.6}$$

are rotation matrices corresponding to the orientation vector $\mathbf{O}_k$.

The localisation problem for a speech source at position $\mathbf{x}_i$ requires determining TDEs from

a speech signal received by the complete set of microphone arrays

$$\mathcal{M} = \{\mathcal{M}_1, .., \mathcal{M}_k, .., \mathcal{M}_K\}. \tag{5.7}$$

A single array $\mathcal{M}_k$ yields a vector of TDEs

$$\boldsymbol{\tau}_{ik} = [\tau_{i1}, .., \tau_{iN_\tau^k}], \tag{5.8}$$

where N_k denotes the total number of microphone pair combinations comprising $\mathcal{M}_k$. For instance, the inverted T-shaped microphone array described in figure 5.1a yields $N_k = \binom{N_{mic}}{2} = \binom{4}{2} = 6$. Therefore, the total number of microphone pairs for an array configuration $\mathcal{M}$ is,

$$N_p = \sum_{k=1}^{K} N_k. \tag{5.9}$$

From the N_p pairs of microphones an $N_p \times 1$ vector of time-delays,

$$\boldsymbol{\tau}_i = [\boldsymbol{\tau}_{i1}, .., \boldsymbol{\tau}_{ij}, ..\boldsymbol{\tau}_{iN_p}]^T \tag{5.10}$$

is obtained from which the speaker position is to be estimated.

It should be noted that the location estimate is assumed to be determined using TDEs arising from inner-array microphone pairs only. The formulation of the problem does not consider TDEs arising from inter-array microphone pair configurations. The problem is constrained in this manner since in general, localisation systems utilising correlation-based time-delay estimation and distributed microphone arrays, follow this practice. This arises due to the dependence of correlation-based time-delay estimation on signal coherence which defines the Cramér-Rao Lower Bound (CRLB) on the variance of the TDEs [20, 56]. The concept of signal coherence defining the minimal achievable variance of a TDE was introduced in section 2.1.3.2 where it was seen that improving signal coherency improves the performance of time-delay estimation. As a result, array geometries employed in speaker localisation commonly ensure that the spacing between microphones is small since this is known to improve signal coherency [157]. Thus, only using inner-array microphone pair combinations, which are known to be closely spaced, improves both the likelihood of high signal coherence and the likelihood of obtaining accurate and reliable TDEs.

Given this formulation of the problem, the approach described in this chapter for optimising a given microphone array configuration $\mathcal{M}$ can be summarised in the following:

- A model of the uncertainty of TDEs $\boldsymbol{\tau}_i$ is defined as a function of the speaker position $\mathbf{x}_i$ and the positions of the microphones. Also included in this model are the effects of the directivity characteristics of both the source and the microphones.
- The analysis of chapter 4 is employed to determine the covariance $\Sigma_{\mathbf{x}_i}$ of a least squares

estimate $\hat{\mathbf{x}}_i$ of the speaker positioned at location $\mathbf{x}_i$ using the vector of TDEs $\boldsymbol{\tau}_i$.

- An objective function is defined based on $\Sigma_{\mathbf{x}_i}$ over all speaker positions $i = 1, .., N_s$.
- Finally, a Simulated Annealing (SA) algorithm is proposed to determine the microphone array configuration $\mathcal{M}$ which minimises the objective function.

5.1 Estimating the Localisation Performance of a Microphone Array Configuration

In section 4.2 it was seen that an estimate of TDE-based localisation uncertainty can be made by mapping the covariance of the TDEs into positional space. This technique can be utilised again in this analysis to determine the localisation uncertainty for the set of speaker positions $\mathbf{x}$.

Using the audio-based time-delay measurement function $g(\cdot)$ of equation 4.8 and covariance mapping theory presented in section 4.2, the localisation uncertainty of a least squares estimate $\hat{\mathbf{x}}_i$ of the speaker position $\mathbf{x}_i$ can be determined. The least squares estimate is that which satisfies the criterion function of equation 4.10. Using the microphone positions defined for an array configuration $\mathcal{M}$, the covariance $\Sigma_{\mathbf{x}_i}$ of the location estimate is obtained as

$$\Sigma_{\mathbf{x}_i} = \frac{\partial g^{-1}(E[\boldsymbol{\tau}_i])}{\partial \boldsymbol{\tau}_i} \Sigma_{\boldsymbol{\tau}_i} \frac{\partial g(E[\boldsymbol{\tau}_i])}{\partial \boldsymbol{\tau}_i}^T . \tag{5.11}$$

In this way, a set of localisation covariance matrices $\mathcal{L} = \{\Sigma_{\mathbf{x}_1}, .., \Sigma_{\mathbf{x}_i} .., \Sigma_{\mathbf{x}_{N_s}}, \}$ corresponding to the array configuration $\mathcal{M}$ can be associated with the set of speaker positions $\{\mathbf{x}_1, .., \mathbf{x}_i, .., \mathbf{x}_{N_s}\}$.

5.1.1 Univariate Measures of Localisation Uncertainty

In essence, the set of covariance matrices $\mathcal{L}$ define the localisation performance of the array configuration $\mathcal{M}$. It is necessary however to transform $\mathcal{L}$ into a single measure of performance which are to be optimised. This is a multivariate data reduction problem. The first task is to reduce the covariances $\Sigma_{\mathbf{x}_i}$ to univariate measures of uncertainty. Secondly, these measures of localisation uncertainty must be combined in an appropriate manner to define an objective function over all speaker positions $\mathbf{x}$ which is to be optimised.

In regard to $\Sigma_{\mathbf{x}_i}$, one can consider two possible univariate measures of localisation uncertainty. These are, the trace of $\Sigma_{\mathbf{x}_i}$

$$tr(\Sigma_{\mathbf{x}_i}) = \sigma_x^2 + \sigma_y^2 + \sigma_z^2, \tag{5.12}$$

representing the *total variance* and the determinant of $\Sigma_{\mathbf{x}_i}$

$$|\Sigma_{\mathbf{x}_i}| = \sigma_x^2 \sigma_y^2 \sigma_z^2, \tag{5.13}$$

defining a measure of *generalized variance*, where σ_x^2, σ_y^2 and σ_z^2 are the diagonal elements of $\Sigma_{\mathbf{x}_i}$ defining the variance along the x, y and z axes respectively [155, Chapter 1]. In relation to the error ellipsoid defined by $\Sigma_{\mathbf{x}}$, the $tr(\Sigma_{\mathbf{x}_i})$ can be thought of as the sum of the length of the ellipsoid's principal axes. Therefore, it is directly related to the total size of the error region associated with the location estimate $\mathbf{x}_i$. Likewise, the $|\Sigma_{\mathbf{x}_i}|$ is related to the size of the error region since it is directly proportional to the error ellipsoid's volume [198].

Some remarks in relation to the usefulness of the $tr(\Sigma_{\mathbf{x}_i})$ and the $|\Sigma_{\mathbf{x}_i}|$ as univariate measures of uncertainty are appropriate. It is clear from equation 5.12 and equation 5.13 that both the $tr(\Sigma_{\mathbf{x}_i})$ and the $|\Sigma_{\mathbf{x}_i}|$ neglect the covariance terms. As a consequence, both are *principal invariants* of the covariance matrix $\Sigma_{\mathbf{x}_i}$ [107]. This means that, under similarity transformations of $\Sigma_{\mathbf{x}_i}$ (eg. rotation), $tr(\Sigma_{\mathbf{x}_i})$ and $|\Sigma_{\mathbf{x}_i}|$ are invariant. A geometrical interpretation of this is that both $tr(\Sigma_{\mathbf{x}_i})$ and $|\Sigma_{\mathbf{x}_i}|$ are invariant to the orientation of the error ellipsoid defined by $\Sigma_{\mathbf{x}_i}$. As a result, using equation 5.12 and equation 5.13 as overall measures of uncertainty can fail to distinguish between different covariance measures. Since the aim in optimising microphone array positions in this analysis is to reduce the overall uncertainty irrespective of the orientation of the error ellipsoids, this is not of major concern. It must be considered however if alternatively, the task is that of minimising localisation uncertainty along particular axes. Such a scenario may arise for example in applications where high localisation accuracy is required in the $x-y$ plane but less critical along the z axis.

An additional issue with the use of the $|\Sigma_{\mathbf{x}_i}|$ as a measure of localisation uncertainty is that it can be particularly misleading. Consider an error ellipsoid where σ_x^2 is close to zero but $\sigma_y^2 >> 0$ and $\sigma_z^2 >> 0$. This corresponds to the case where localisation uncertainty is close to zero along the x axis but large in both the y and z axes. In this case due to σ_x^2 being close to zero, $|\Sigma_{\mathbf{x}_i}|$ is small which does not reflect the true uncertainty evidenced by the large values of σ_y^2 and σ_z^2 [131]. In this way minimising $|\Sigma_{\mathbf{x}_i}|$ can result in only minimising uncertainty along one axis.

5.1.2 The Objective Function

It was seen in the previous section that both equation 5.12 and equation 5.13 can be used as a quantitative measure of the size of the error region associated with a location estimate. This section examines the problem of combining these measures of uncertainty over all speaker positions $\mathbf{x}_i$, $i = 1, .., N_s$ in establishing the objective function to be minimised.

A key aim of the proposed optimisation approach is not only to determine the microphone array configuration which minimises localisation uncertainty, but also that which results in the even distribution of uncertainty over all speaker positions. This is so as to avoid determining an optimal microphone array configuration which yields high localisation accuracy over only a subset of the total speaker locations. There are certain circumstances where such a solution may be desirable such as in the case where the minimisation of localisation uncertainty is to be prioritised over a particular subset of speaker positions. In the proposed minimisation approach

however it is desired to incorporate this facility through a prior weighting of speaker positions rather than as an artefact of the minimisation process.

In regard to minimising the overall uncertainty while ensuring that it is evenly distributed over all speaker positions, four objective functions are considered. These are,

$$\mathcal{E}_1(\mathcal{L}) = \sum_{i=1}^{N_s} w_i tr(\Sigma_{\mathbf{x}_i}), \tag{5.14a}$$

$$\mathcal{E}_2(\mathcal{L}) = \sum_{i=1}^{N_s} w_i \sqrt{tr(\Sigma_{\mathbf{x}_i})}, \tag{5.14b}$$

$$\mathcal{E}_3(\mathcal{L}) = \sum_{i=1}^{N_s} w_i |\Sigma_{\mathbf{x}_i}|, \tag{5.14c}$$

$$\mathcal{E}_4(\mathcal{L}) = \sum_{i=1}^{N_s} w_i tr(\Sigma_{\mathbf{x}_i})^2 \tag{5.14d}$$

where w_i is a priority weighting on the ith speaker position, with $0 \leq w_i \leq 1$ and $\sum_i w_i = 1$. Equation 5.14a, equation 5.14b and equation 5.14c have previsouly been used as cost functions in the sensor placement literature by, Neering et al. [95], Brandstein et al. [17] and Erdinc et al. [137] respectively. The cost function of equation 5.14d is proposed in this work to address the issue of ensuring the even distribution of the the total uncertainty over all speaker positions. It is shown in the following, by means of an example, that this is not achieved through the use of either equation 5.14a, equation 5.14b or equation 5.14c. Although the $3D$ positioning of microphone arrays is of concern in this work, the example presented considers only the $2D$ case for illustrative purposes.

Consider the two-dimensional case of a microphone array configuration yielding a solution consisting of a set of localisation covariance matrices $\mathcal{L}_1 = \{\Sigma_{\mathbf{x}_1}, \Sigma_{\mathbf{x}_2}\}$ with $\Sigma_{\mathbf{x}_1} = diag(a, b)$ and $\Sigma_{\mathbf{x}_2} = diag(a, c)$ where $diag(a, b)$ is diagonal matrix formed by elements a and b. For this case, the cost functions of equation 5.14 are determined as,

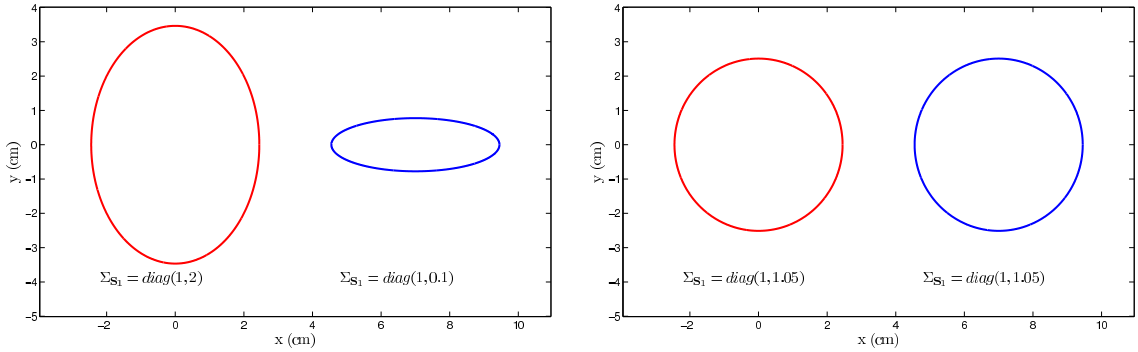
$$\mathcal{E}_1(\mathcal{L}_1) = 2a + b + c \tag{5.15a}$$

$$\mathcal{E}_2(\mathcal{L}_1) = (2a + b + c + 2(a^2 + ab + ac + bc)^{\frac{1}{2}})^{\frac{1}{2}} \tag{5.15b}$$

$$\mathcal{E}_3(\mathcal{L}_1) = ab + ac \tag{5.15c}$$

$$\mathcal{E}_4(\mathcal{L}_1) = (a + b)^2 + (a + c)^2 \tag{5.15d}$$

Consider now a second microphone array configuration yielding the solution resulting in the set of localisation covariance matrices $\mathcal{L}_2 = \{\Sigma_{\mathbf{x}_1}, \Sigma_{\mathbf{x}_2}\}$ where $\Sigma_{\mathbf{x}_1} = \Sigma_{\mathbf{x}_2} = diag(\frac{a+a}{2}, \frac{c+d}{2})$. Essentially, the solution $\mathcal{L}_2$ is formed from $\mathcal{L}_1$ by averaging the localisation uncertainty along both the x and y axes. The error ellipses for an example case of $\mathcal{L}_1$ and $\mathcal{L}_2$ are illustrated in figure 5.2a and figure 5.2b respectively. From this it can be seen that both $\mathcal{L}_1$ and $\mathcal{L}_2$ correspond to



(a) 95 percentile error ellipses of $\mathcal{L}_1$

(b) 95 percentile error ellipses of $\mathcal{L}_2$

Figure 5.2: *In this example the localisation covariance matrices are* $\mathcal{L}_1 = \{diag(1,2), diag(1,0.1)\}$ *and* $\mathcal{L}_2 = \{diag(1,1.05), diag(1,1.05)\}$ *in units of* cm^2. *The resulting cost functions from equation 5.14 for the solution corresponding to* $\mathcal{L}_1$ *as illustrated in (a) are* $\mathcal{E}_1(\mathcal{L}_1) = 4.1cm^2$, $\mathcal{E}_2(\mathcal{L}_1) = 2.78cm^2$, $\mathcal{E}_3(\mathcal{L}_1) = 2.1cm^4$ *and* $\mathcal{E}_4(\mathcal{L}_1) = 10.21cm^4$. *The evaluated cost functions for the solution of* $\mathcal{L}_2$ *as shown in (b) are* $\mathcal{E}_1(\mathcal{L}_2) = 4.1cm^2$, $\mathcal{E}_2(\mathcal{L}_2) = 2.86cm^2$, $\mathcal{E}_3(\mathcal{L}_2) = 2.1cm^4$ *and* $\mathcal{E}_4(\mathcal{L}_2) = 8.40cm^4$. *From this it can be seen that* $\mathcal{E}_1(\mathcal{L}_1) = \mathcal{E}_1(\mathcal{L}_2)$, $\mathcal{E}_2(\mathcal{L}_1) < \mathcal{E}_2(\mathcal{L}_2)$, $\mathcal{E}_3(\mathcal{L}_1) = \mathcal{E}_3(\mathcal{L}_2)$ *and* $\mathcal{E}_4(\mathcal{L}_1) > \mathcal{E}_4(\mathcal{L}_2)$ *therefore only the cost functions of* $\mathcal{E}_2$ *of equation 5.14b and* $\mathcal{E}_4$ *of equation 5.14d assign different costs to each solution. The use of the cost function* $\mathcal{E}_4$, *corresponding to the sum of the squared trace of the covariance matrices, assigns the lower cost to solution* $\mathcal{L}_2$ *where the uncertainty is more evenly distributed over both positions.*

very different localisation error distributions. Despite this however, the resulting cost functions of equation 5.14 for $\mathcal{L}_2$ are found to be,

$$\mathcal{E}_1(\mathcal{L}_2) = 2a + b + c \tag{5.16a}$$

$$\mathcal{E}_2(\mathcal{L}_2) = (2a + b + c + 2(a^2 + ab + ac + bc + \frac{1}{4}(b-c)^2)^{\frac{1}{2}})^{\frac{1}{2}} \tag{5.16b}$$

$$\mathcal{E}_3(\mathcal{L}_2) = ab + ac \tag{5.16c}$$

$$\mathcal{E}_4(\mathcal{L}_2) = (a+b)^2 + (a+c)^2 - \frac{1}{2}(b-c)^2 \tag{5.16d}$$

From this it can be seen that $\mathcal{E}_1(\mathcal{L}_1) = \mathcal{E}_1(\mathcal{L}_2)$ and $\mathcal{E}_3(\mathcal{L}_1) = \mathcal{E}_3(\mathcal{L}_2)$. Therefore the objective functions of $\mathcal{E}_1$ in equation 5.14a and $\mathcal{E}_3$ in equation 5.14c do not distinguish between the two solutions. Careful observation of equation 5.15 and equation 5.16 reveals that since a,b,$c \geq 0$ then $\mathcal{E}_2(\mathcal{L}_1) < \mathcal{E}_2(\mathcal{L}_2)$ and $\mathcal{E}_4(\mathcal{L}_1) > \mathcal{E}_4(\mathcal{L}_2)$. From this, the cost function $\mathcal{E}_2$ in equation 5.14b is seen to assign a lower cost to the solution corresponding to $\mathcal{L}_1$. Only the cost function $\mathcal{E}_2$ in equation 5.14d assigns a lower cost to the solution $\mathcal{L}_2$ where the total localisation uncertainty is evenly distributed over the speaker positions.

5.2 Modelling Uncertainty on the Time-delay Estimates

Determining the localisation covariance at each speaker position through equation 5.11 requires defining the uncertainty on the TDEs $\boldsymbol{\tau}_i$. The CRLB as a bound on the minimum achievable variance of an unbiased TDE was introduced in section 2.1.3.2. This bound was seen to be inversely dependent on the SNR. The accuracy of the TDEs is therefore directly determined by issues affecting the SNR such as the microphone-to-speaker distances and also by the directivity characteristics of both the speakers and the microphones. By adequately modelling for the effects of such on the SNR, the uncertainty of TDEs can be estimated using the CRLB. As was previously state in section 2.1.3.2 the CRLB only models "small" errors in the time-delay estimation problem and does not describe "large" errors due to anomalous TDEs.

This work employs the model of TDE uncertainty as proposed by Gustafsson et al. which more accurately models uncertainty in the time-delay estimation problem than that of the CRLB [175–177]. The model of Gustafsson et al. specifically examines the problem of time-delay estimation in reverberant environments and is based on the Correlator Performance Estimate (CPE) as proposed by Ianniello [97]. In the following, the CPE of Ianneillo is introduced and the use of its extended form as proposed by Gustafsson et al. for reverberant environments is motivated.

5.2.1 The Correlator Performance Estimate (CPE)

Ianniello proposed the CPE to address the shortcomings of the CRLB in describing the *large* error effects on the time-delay estimation problem. Essentially, the CPE is derived by considering two time-delay estimators. The first is affected by *small* errors and provides an unbiased TDE which achieves the CRLB with variance σ^2_{CRLB}. The second is affected by *large* errors and determines the TDE as a random sample from a continuous uniform distribution in the range $[-\tau_{max}, \tau_{max}]$. This corresponds to the case where the received signals are corrupted by noise to the extent that they contain no useful information as to the true time-delay. In the lack of any *a priori* knowledge of the true time-delay, all TDEs in the range $[-\tau_{max}, \tau_{max}]$ are assumed equally likely. The variance of such a TDE is simply equal to the variance of a continuous random variable in the range $[-\tau_{max}, \tau_{max}]$ given by, $\frac{[\tau_{max}-(-\tau_{max})]^2}{12} = \frac{\tau^2_{max}}{3}$ [63, Chapter 2]. Using this, the CPE defines the variance of a TDE made in the presence of both *small* and *large* errors as

$$\sigma^2_{ij} = (1 - Pr[\epsilon])\sigma^2_{CRLB} + Pr[\epsilon]\frac{\tau^2_{max}}{3}, \tag{5.17}$$

where $Pr[\epsilon]$ is the probability of an anomalous estimate.

The CPE can be thought of as describing three regions of operation for an unbiased correlation-based time-delay estimator. The first region corresponds to that where $Pr[\epsilon] = 0$. This is known as the asymptotic region (*small* error region) where the estimator achieves the CRLB asymptotically with increasing SNR. In cases where $Pr[\epsilon] = 1$ corresponding to scenarios of low SNR,

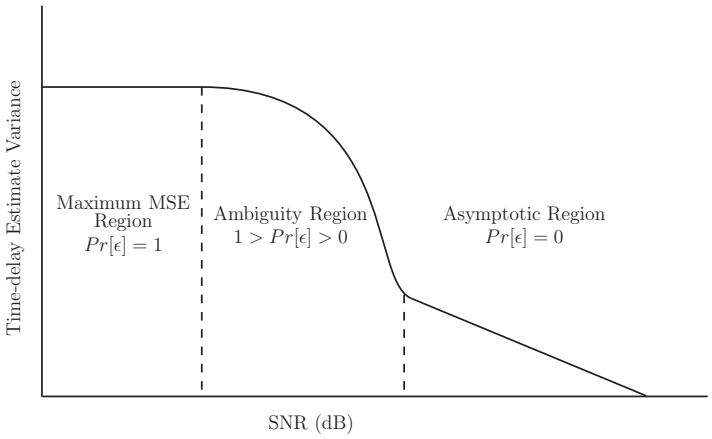


Figure 5.3: *Three regions of operation for an unbiased TDE estimator as defined by the CPE.*

the received signal contains no useful information about the true time-delay and its estimate is restricted to *a priori* knowledge. In this region the variance of the unbiased estimate (Mean Square Error (MSE)) is maximum. This is known as the maximum MSE region (*large* error region). Between these two regions of operation for the case $1 > Pr[\epsilon] > 0$ the estimator is said to operate in the ambiguity region. Within this region the performance of the time-delay estimator is seen to diverge from that predicted by the CRLB and is affected by large errors due to anomalies. The three regions of operation for an unbiased time delay estimator in relation to the SNR are illustrated in figure 5.3.

In proposing the CPE, Ianniello also gives a theoretical definition of the $Pr[\epsilon]$. This definition considers the cross-correlation function $R_{z_m z_n}(\tau)$ of equation 2.28 as consisting of M independent values v_m, $m = 0, .., M-1$ corresponding to the time delays τ_m. Only one of the time-delays τ_m is assumed to correspond to the true time-delay. In the case of a band-limited source signal $s(t)$ with cross power spectrum $G_{ss}(\omega) = 1$ for $-2\pi B \leq \omega \leq 2\pi B$, points $\frac{1}{2B}$ apart in $R_{z_m z_n}(\tau)$ are uncorrelated [99, Chapter 8]. Therefore, the cross-correlation function $R_{x_1 x_2}(\tau)$ defined in the range $[-\tau_{max}, \tau_{max}]$ contains $M = 2\tau_{max}(2B) = 4B\tau_{max}$ independent values.

If the true time-delay corresponds to the value v_0 then the event of an anomalous TDE is defined as [97],

$$\epsilon = v_m > v_0 \text{ for at least one } v_m \quad m \neq 0. \tag{5.18}$$

More simply stated, this event can be thought of as the case where any other peak in the cross correlation function is greater than that of the peak at the true time-delay. By this definition of the event ϵ, the probability of an anomaly is determined as,

$$Pr[\epsilon] = 1 - \int_{-\infty}^{+\infty} p(v_0) \left[\int_{-\infty}^{v_0} p(v_m) dv_m \right]^{M-1} dv_0 \tag{5.19}$$

where $p(v_0)$ and $p(v_m)$ are the probability density functions of the peaks v_0 and v_m of the cross-

correlation function respectively. In Ianniello's formulation, both $p(v_0)$ and $p(v_m)$ are defined as Gaussian with,

$$p(\alpha_0) \sim \mathcal{N}\left(1, \frac{1}{2BT}\left[2 + \frac{1}{SNR} + \frac{1}{SNR^2}\right]\right) \tag{5.20a}$$

$$p(\alpha_m) \sim \mathcal{N}\left(0, \frac{1}{2BT}\left[1 + \frac{1}{SNR} + \frac{1}{SNR^2}\right]\right) \tag{5.20b}$$

where SNR is the signal-to-noise ratio and T is the observation interval over which the cross-correlation function $R_{z_m z_n}(\tau)$ is determined. In Equation 5.20 both $p(v_0)$ and $p(v_m)$ are defined for the case where the cross-correlation function is normalised so that the value of the maximum peak is equal to one. Due to the form of equation 5.19, a value for $Pr[\epsilon]$ can only be determined numerically [2].

Gustafsson et al. extend the CPE to model the performance of time-delay estimation in the presence of room reverberation for the discrete time case [175–177]. They examine a noiseless scenario and assume the only distortion of the source signal $s(t)$ is due to the effects of reverberation. In essence, reverberation is considered as a noise source and in their analysis the SNR is replaced by the Signal-to-Reverberant Ratio (SRR) as defined in equation 2.8.

The CRLB in the case of room reverberation $\sigma^2_{CRLB,rev}$ is obtained by substituting the SNR in equation 2.37 with the SRR and replacing the integral with a summation for the discrete time case. The SRR assumes that the reverberant acoustic energy within the room is constant and not dependent on the relative source-to-microphone distance. Although a simplistic model of reverberant conditions, experimental evidence would suggest that this is a reasonable assumption [42, Chapter 2]. The implications of this assumption in the definition of $Pr[\epsilon]$ is to neglect the effect of early reflections and assume that the probability of an anomaly is constant irrespective of the location of the source and microphones within the room. Experimental evidence suggests that this assumption is likely to be invalid at positions close to walls within the room due to strong earlier reflections [121].

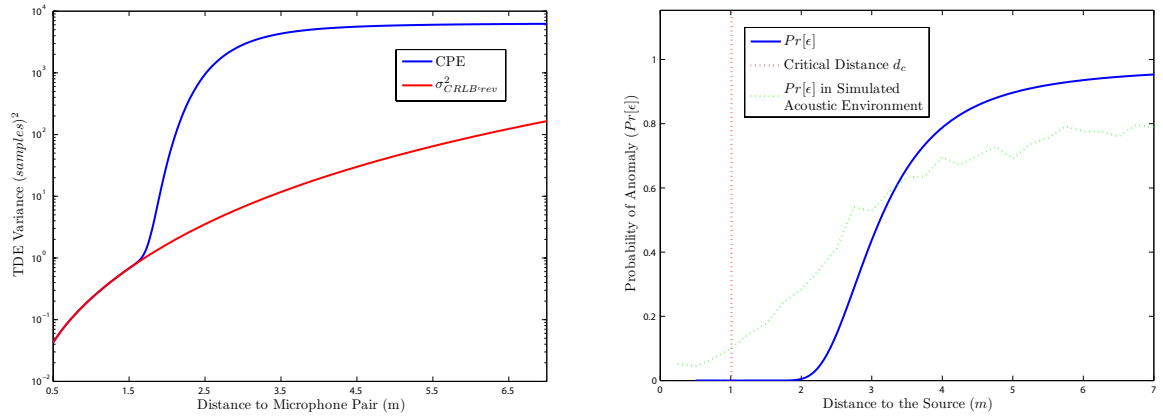
Gustafsson et al. in their work, also redefine the probability densities $p(v_0)$ and $p(v_m)$ as,

$$p(v_0) \sim \mathcal{N}\left(1, \frac{\rho(RT_{60})}{2B}\left[\frac{2}{SRR} + \frac{1}{SRR^2}\right]\right) \tag{5.21a}$$

$$p(v_m) \sim \mathcal{N}\left(0, \frac{\rho(RT_{60})}{2B}\left[\frac{2}{SRR} + \frac{1}{SRR^2}\right]\right) \tag{5.21b}$$

where $\rho(RT_{60})$ is the coherence bandwidth. The coherence bandwidth corresponds to the range of frequencies over which the reverberant components of the signals received at a pair of microphones are correlated. Effectively, reverberation is seen to reduce the number of observed uncorrelated points in the cross-correlation function by $\frac{1}{\rho(RT_{60})}$. In Gustafsson's proposal, the co-

[2] The MATLAB function quadgk [182] which implements the Gauss-Konrod quadrature formula for numerical integration is used to evaluate $Pr[\epsilon]$ of equation 5.19 in the results of this chapter.



(a) (b)

Figure 5.4: *In (a) TDE variance as predicted by the $\sigma^2_{CRLB,rev}$ (red) and CPE (blue) of Gustafsson et al. [177] is shown. This plot presents the variance of determining a cross-correlation based TDE $\hat{\tau}$ resulting from a $300Hz - 5kHz$ band-passed white noise source placed directly in front (i.e. $\hat{\tau} = 0$) of a pair of microphones with spacing $0.4m$. The variance is plotted for varying distance from the source to the microphone pair. The environment considered is a simulated room of the same size as the CHIL room with reverberation time $RT_{60} = 0.4s$. Figure (b) shows the probability of anomaly $Pr[\epsilon]$ (blue) in the estimate $\hat{\tau}$ determined using equation 5.19 and the cross-correlation peak model of equation 5.21. The $Pr[\epsilon]$ obtained through simulating the conditions of the CHIL room using the image method [82] is shown in (green). The $Pr[\epsilon]$ in the simulated environment was determined by 250 Monte Carlo simulations where the microphone pair and source configuration was randomly placed within the simulated room. The critical distance for the room model from equation 2.9 is also shown in red.*

herence bandwidth is defined in an ad-hoc manner as $\rho(RT_{60}) = \frac{10}{RT_{60}}$, but is seen to correspond closely to statistical based analysis of reverberant rooms [124].

In modelling the uncertainty of the TDEs in this optimisation problem, the CPE is used as a measure of TDE uncertainty. Using this model of uncertainty, the TDE covariance is defined as $\Sigma_{\boldsymbol{\tau}_i} = diag(\sigma^2_{i1}, .., \sigma^2_{ij}, ..\sigma^2_{iNp})$ where the variances σ^2_{ij} are determined using equation 5.17 where both σ^2_{CRLB} and $Pr[\epsilon]$ are that proposed by Gustafsson et al. [175–177]. An evaluation of this CPE for a simulated CHIL room with a reverberation time $RT_{60} = 0.4s$ is presented in figure 5.4. The simulated CHIL room is also considered in the latter analysis of this chapter examining the results of the proposed optimisation algorithm.

5.2.2 Speaker and Microphone Directivity Characteristics

As well as employing a detailed model of TDE uncertainty, this analysis aims to also incorporate the effects of the speaker and microphone directivity characteristics into the uncertainty model. To achieve this it is proposed to redefine the SRR in equation 2.8 as dependent on the speaker-to-microphone and microphone-to-speaker angles. This is achieved by scaling the SRR such that

equation 2.8 becomes,

$$SRR = D_j^2(\theta_{(j,\mathbf{x}_i)}, \phi_{(j,\mathbf{x}_i)}) D_{\mathbf{x}_i}^2(\theta_{(\mathbf{x}_i,j)}, \phi_{(\mathbf{x}_i,j)}) \frac{\mathcal{A}\alpha}{16\pi r^2(1-\alpha)}, \tag{5.22}$$

where $D_j(\theta_{(j,\mathbf{x}_i)}, \phi_{(j,\mathbf{x}_i)})$ is the reception amplitude due to the angle of azimuth $\theta_{(j,\mathbf{x}_i)}$ and angle of elevation $\phi_{(j,\mathbf{x}_i)}$ of the speaker relative to the jth microphone pair. Similarly, $D_{\mathbf{x}_i}(\theta_{(\mathbf{x}_i,j)}, \phi_{(\mathbf{x}_i,j)})$ is the speaker source amplitude due to the angles of azimuth $\theta_{(\mathbf{x}_i,j)}$ and elevation $\phi_{(\mathbf{x}_i,j)}$ of the jth microphone pair relative to the speaker position $\mathbf{x}_i$. It should be noted that it is assumed in equation 5.22 that the source and microphone directivity characteristics have no effect on the reverberant acoustic energy present in the room. Modelling the effects of such is beyond the scope of this analysis. It is assumed therefore that the effect of the speaker and microphone directivity characteristics is only observed in the direct path component of the microphone outputs.

The described framework is flexible to the incorporation of any speaker and microphone directivity characteristics. In the following analysis however the microphones are modelled as cardioid receivers where the received signal amplitude is described by,

$$D_m(\theta_{(\mathbf{m}_{kj},\mathbf{x}_i)}, \phi_{(\mathbf{m}_{kj},\mathbf{x}_i)}) = \frac{1}{2}(1 + \cos\theta_{(\mathbf{m}_{kj},\mathbf{x}_i)} \cos\phi_{(\mathbf{m}_{kj},\mathbf{x}_i)}). \tag{5.23}$$

Modelling the directivity characteristics of a speaker is significantly more challenging since the directivity characteristics of the mouth is known to be frequency dependent and also dependent on speech dynamics [89, 93, 189]. There is little evidence in existing literature to suggest the most suitable model for the directivity characteristics of a speaker. Previous research [3] has employed the cardioid pattern of equation 5.23 to model the directivity characteristics of speakers and in the absence of a more appropriate model the cardioid pattern is also employed in this work. The speaker source signal amplitude is therefore defined as,

$$D_S(\theta_{(\mathbf{x}_i,\mathbf{m}_{kj})}, \phi_{(\mathbf{x}_i,\mathbf{m}_{kj})}) = \frac{1}{2}(1 + \cos\theta_{(\mathbf{x}_i,\mathbf{m}_{kj})} \cos\phi_{(\mathbf{x}_i,\mathbf{m}_{kj})}). \tag{5.24}$$

It should be noted that experimental evidence would suggest that modelling speakers as cardioid emitters is only reasonable within the $\pm 30^\circ$ range but outside of this it is overly conservative [89, 93, 189]. The imposed model will therefore be expected to result in microphone array positions directly in front of speakers to be favoured more so than other locations about the speaker. Although in general this is desirable, it may be more restrictive than is necessary.

5.3 A Simulated Annealing based Approach

Optimising a configuration of microphone arrays $\mathcal{M}$ through minimising the objective function of equation 5.14d represents a complex optimisation problem. Typically, the objective function will contain many local minima as well as multiple global minima. Multiple global minima

can be seen to exist by considering the interchange of microphone arrays of a given optimal configuration. The presence of multiple global minima however does not introduce difficulties into the optimisation problem since they correspond to equivalent optimal microphone array configurations.

The existence of local minima however, is detrimental to standard gradient based optimisation approaches. As a result for complex objective functions such as in this problem, one is generally confined to stochastic optimisation techniques such as, Simulated Annealing (SA), simultaneous perturbation stochastic approximation, random search or genetic algorithms [84].

In this work, a simulated annealing algorithm based on the classic SA technique of Kirkpatrick et al. [168] is described for optimising a microphone array configuration. Simulated annealing as an optimisation technique is inspired by the process of annealing in metallurgy. Annealing is a thermal process whereby a material is heated to a high temperature ("melted") and slowly cooled. Atoms of a heated material on cooling will naturally tend towards lower energy states. Therefore, after a process of slow cooling the material attains a lower energy atomic structure. Simulated annealing aims to emulate this process to optimise a set of parameters in minimising some defined objective function. In relation to the annealing problem in metallurgy the value of the objective function is analogous to the energy of the material and the value of the parameters to its atomic structure.

In the implementation of the simulated annealing algorithm, the array parameters $\mathbf{C}_k$ and $\mathbf{O}_k$ are assumed to be discrete. Specifically, the array centre $\mathbf{C}_k$ is assumed to be constrained to a grid $\mathbb{Z}_c$ of feasible user defined positions within the room. Similarly, the orientation angle is assumed to be constrained to a discrete set $\mathbb{Z}_o$ of orientation angles. In the proceeding analysis the $\mathbb{Z}_c$ is assumed to have a resolution of $0.1m$ in each dimension and the $\mathbb{Z}_o$ is defined for orientation angles of pan, tilt and roll in the range $[0°, 360°]$ each with a resolution of $2°$.

5.3.1 Basic Simulated Annealing Algorithm

A basic application of simulated annealing to the problem of optimising a microphone array configuration $\mathcal{M}^{curr}$ with corresponding localisation covariance matrices $\mathcal{L}^{curr}$ can be described as follows. Firstly a temperature Θ is defined and the parameters of $\mathcal{M}^{curr}$ are then perturbed relative to Θ to determine a new configuration $\mathcal{M}^{new}$ with localisation covariance matrices $\mathcal{L}^{new}$. The values of the objective function for both configurations are then compared. If the value of the objective function corresponding to the new configuration is less than that corresponding to the current configuration (i.e. $\mathcal{E}_4(\mathcal{L}^{new}) < \mathcal{E}_4(\mathcal{L}^{curr})$) then $\mathcal{M}^{new}$ is *accepted* and replaces $\mathcal{M}^{curr}$. Otherwise, *acceptance* of $\mathcal{M}^{new}$ is probabilistic and only accepted with probability equal to,

$$\exp\left(-\frac{\mathcal{E}_4(\mathcal{L}^{new}) - \mathcal{E}_4(\mathcal{L}^{new})}{c_b\Theta}\right), \tag{5.25}$$

known as the *Metropolis* criterion. The value c_b is known as the Boltzmann constant but can be incorporated into the temperature definition by redefining Θ as $\Theta = c_b\Theta$. Accepting new configurations where $\mathcal{E}_4(\mathcal{L}^{new}) > \mathcal{E}_4(\mathcal{L}^{curr})$ in a probabilistic manner, enables the algorithm to "escape" local minima. It is seen from equation 5.25 that this is most likely to occur with high probability where $\mathcal{E}_4(\mathcal{L}^{new}) - \mathcal{E}_4(\mathcal{L}^{curr})$ is small or the temperature Θ is relatively large. The above procedure is repeated until a state of equilibrium is reached where no more new configurations are accepted, or after a defined number of evaluations. The algorithm then proceeds iteratively by reducing the temperature Θ and repeating the above process. The manner in which the temperature is reduced at each iteration is termed the *cooling schedule*. A simple and commonly used cooling schedule is to assign the value $\Theta = 0.95\Theta$ to the temperature at each iteration.

5.3.2 Proposed Simulated Annealing Algorithm

The proposed simulated annealing algorithm introduces two adaptations to the basic algorithm described above. The first adaptation occurs in the generation of new microphone configurations at each temperature. Rather than perturb all the parameters of the configuration $\mathcal{M}^{curr}$ at the same time, it is proposed instead to perturb the parameters of each microphone array $\mathcal{M}_k^{curr}$ one at a time. This is implemented by firstly generating a new centre position $\mathbf{C}_k^{curr}$ for $\mathcal{M}_k^{curr}$ while holding all other arrays in the configuration $\mathcal{M}^{curr}$ constant. With the centre position of the array $\mathcal{M}_k$ fixed the array orientation $\mathbf{O}_k^{curr}$ is perturbed N_{ang} times with each defining a new array configuration. This strategy for determining new array configurations is then repeated in a cyclic manner for each array $k = 1, .., K$ in $\mathcal{M}^{curr}$. A single realisation of this process is referred to as an *array cycle*.

In addition to this, the proposed algorithm also incorporates a means for determining when a state of equilibrium has been entered. This is achieved by introducing a variable N_{acc} which records the number of acceptances occurring over the N_{ang} new configurations evaluated over an array cycle. A variable N_{eq} is also used to record the number of consecutive array cycles not yielding any accepted new configurations. The algorithm defines the optimisation process as being in a state of equilibrium at the current temperature if N_{eq} equals some user defined value N_{stop}. At this point, the algorithm progresses to the next temperature defined by the cooling schedule. By setting N_{stop} therefore, the algorithm can be tuned as to how long the optimisation process remains at a given temperature. This also acts to speed up the optimisation process by exiting the current temperature if it is yielding no accepted new configurations.

An overview of the complete algorithm is presented in Algorithm 1. No claim is made as to the convergence of the proposed algorithm to a global optimum since this would require an unfeasible exhaustive search of all possible combinations. The proposed algorithm is seen to tend towards a solution which is seen to be "more" optimal than the initial array configuration. As is common with stochastic optimisation approaches the level of optimality of a particular solution

is determined on the amount of computation time which can be assigned to the optimisation process.

5.4 Results

In this section the simulated annealing algorithm presented in Algorithm 1 is applied to the problem of optimising a configuration $\mathcal{M}$ of $K = 4$ microphone arrays within a room. The criterion for optimisation is that the localisation error is to be minimised over an audience area and a presenter area defined by the set of speaker positions $\mathbf{x}$. The microphone array geometry considered for each array in $\mathcal{M}$ is that of the inverted T-shape array as illustrated in figure 5.1a. The simulated room model considered is the CHIL room as described in section 1.1 with a reverberation time of $RT_{60} = 0.4s$. This corresponds to the same room model examined in figure 5.4b.

The first example presented, describes the case where both the audience and presenter areas are defined by a set of symmetric points within the room such as illustrated in figure 5.5a. This example is used to represent a typical lecture or presentation scenario where each point in the audience area represents an audience member and each point of the presenter area represents a potential location for the presenter. In addition to this, the problem is defined for the case where the presenter is assumed to be facing the audience. Therefore, both the presenter and audience areas face in opposite directions. This is can be seen in figure 5.5a where the cardioid directivity pattern for each speaker position is illustrated.

Shown in figure 5.5b is the assumed initial configuration for $\mathcal{M}$ with the microphone arrays positioned approximately at the top centre location of each wall in the room as in the CHIL room. The first observation in this figure is that the localisation uncertainty is significantly large resulting in the overlapping of the uncertainty regions. This would suggest that localisation accuracy using such a configuration would be poor. It also suggests that due to the overlapping uncertainty regions, associating any location estimate with a particular speaker position would be difficult.

In figure 5.5c the result of the proposed algorithm for optimising the microphone array configuration over the audience area only is shown. This corresponds to the case where the priority weights w_i in equation 5.14d are set to zero over the presenter area. In optimising the array configuration in the described example, the arrays are constrained to the walls and ceiling. From figure 5.5c it is seen that the algorithm determines a configuration which improves significantly upon the initial configuration. The optimised configuration reduces the overall localisation uncertainty such that there is no overlap between uncertainty regions. Once again it is stated that, beyond an unfeasible exhaustive search of all possible configurations, it is difficult to confirm the configuration obtained by the proposed algorithm as optimal. There are some aspects of the obtained solution however which are encouraging. Firstly, the symmetric nature of the returned configuration given the symmetric profile of the audience area, suggests the most

SIMULATED ANNEALING ALGORITHM FOR OPTIMISING MICROPHONE ARRAY POSITIONS

INPUT

Microphone array configuration $\mathcal{M}$

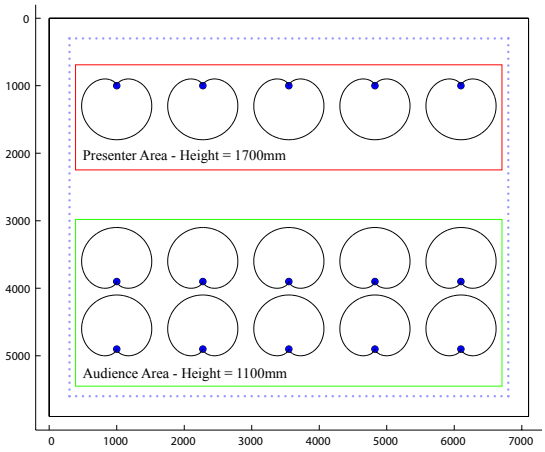
INITIALISATION

Determine $\mathcal{L} = \{\Sigma_{\mathbf{x}_i}, ..\Sigma_{\mathbf{x}_{N_s}}, ..\}$ using equation 5.11
$\Theta = 10^{15}$, $\Theta_0 = 10^{-3}$, $N_{ang} = 10$, $N_{acc} = 0$, $N_{eq} = 0$, $N_{stop} = 3$
$\mathcal{M}^{curr} = \mathcal{M}$, $\mathcal{L}^{curr} = \mathcal{L}$, $\mathcal{M}^{new} = \mathcal{M}$, $\mathcal{L}^{new} = \mathcal{L}$, $\mathcal{M}^{opt} = \mathcal{M}$, $\mathcal{L}^{opt} = \mathcal{L}$

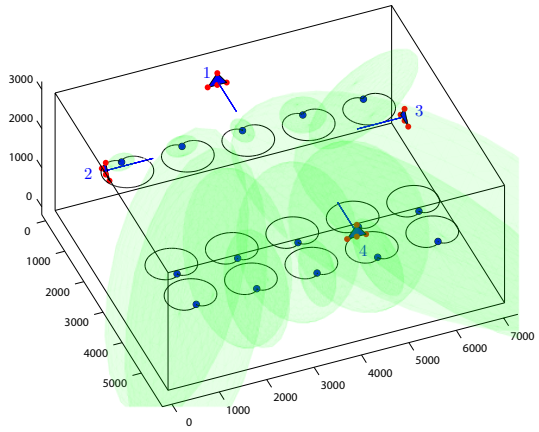
RECURSION

while $\Theta > \Theta_0$ **do**
  while $N_{eq} \neq N_{stop}$ **do**
    for $k = 1 : K$ **do**
      Generate $\mathbf{C}_k$ by perturbing the centre of $\mathcal{M}_k^{curr}$ proportional to Θ
      for $i = 1 : N_{ang}$ **do**
        Generate $\mathbf{O}_k$ by perturbing orientation of $\mathcal{M}_k^{curr}$ proportional to Θ
        Update $\mathcal{M}_k^{new}$ with $\mathcal{M}_k^{new} = \{\mathbf{C}_k, \mathbf{O}_k\}$
        Determine $\mathcal{L}^{new}$ using $\mathcal{M}^{new}$
        if $\mathcal{E}_4(\mathcal{L}^{new}) < \mathcal{E}_4(\mathcal{L}^{curr})$ **then**
          $\mathcal{M}^{curr} = \mathcal{M}^{new}$, $\mathcal{L}^{curr} = \mathcal{L}^{new}$
          $N_{acc} = N_{acc} + 1$
        else
          $P = \exp\left(-\frac{\mathcal{E}_4(\mathcal{L}^{new}) - \mathcal{E}_4(\mathcal{L}^{curr})}{\Theta}\right)$
          Draw random variable $q \sim \mathcal{U}(0, 1)$
          if $P > q$ **then**
            $\mathcal{M}^{curr} = \mathcal{M}^{new}$, $\mathcal{L}^{curr} = \mathcal{L}^{new}$
            $N_{acc} = N_{acc} + 1$
          end if
        end if
        if $\mathcal{E}_4(\mathcal{L}^{curr}) < \mathcal{E}_4(\mathcal{L}^{opt})$ **then**
          $\mathcal{M}^{opt} = \mathcal{M}^{curr}$, $\mathcal{L}^{opt} = \mathcal{L}^{curr}$
        end if
      end for
      if $N_{acc} = 0$ **then**
        $N_{eq} = N_{eq} + 1$
      else
        $N_{eq} = 0$
      end if
    end for
  end while
  $\Theta = 0.95\Theta$
end while
return $\mathcal{M}^{opt}$

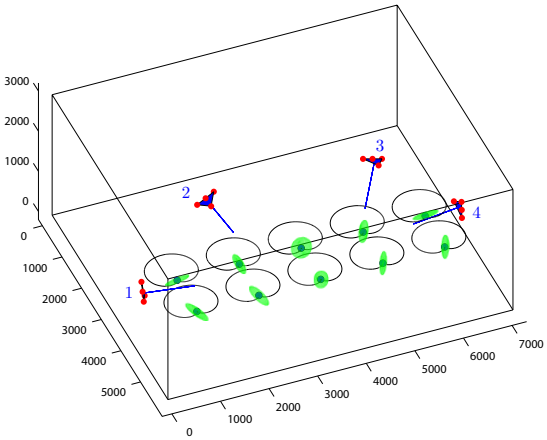
Algorithm 1: *Simulated annealing algorithm for optimising a microphone array configuration* $\mathcal{M}$.



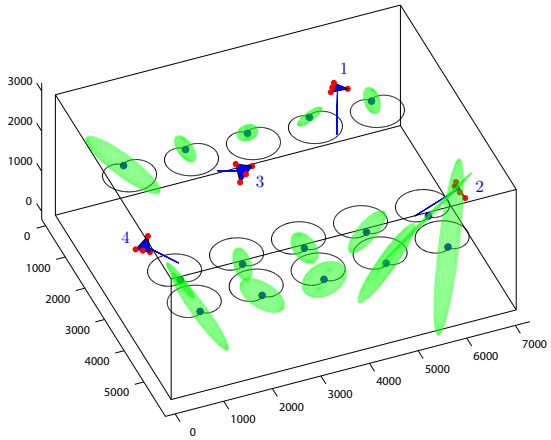
(a) *Symmetric audience (green) and presenter (red) areas facing in opposite directions. The audience is assumed seated at a height of* $1.1m$ *and the presenter is assumed standing at a height of* $1.7m$*. The cardioid directivity patterns are shown in black.*



(b) *Initial microphone array configuration with localisation error ellipsoids shown in green over the audience and presenter areas.*



(c) *Microphone array configuration optimised over the audience area only with localisation error ellipsoids shown in green.*



(d) *Microphone array configuration optimised over both the audience and presenter areas with localisation error ellipsoids shown in green.*

Figure 5.5: *Optimising the configuration of* 4 *inverted T-shaped microphone arrays for symmetric audience and presenter areas facing in opposite directions. In this example the array centres are constrained to the boundaries of the room (i.e. walls, floor and ceiling). The localisation error ellipsoids illustrated, correspond to the* 95 *percentile error regions determined using the CPE uncertainty model of equation 5.17 with* $Pr[\epsilon] = 0$.

intuitive optimal configuration. In addition to this the microphones are all positioned in front of the audience area indicating the effects of including the directionality characteristics into the optimisation problem.

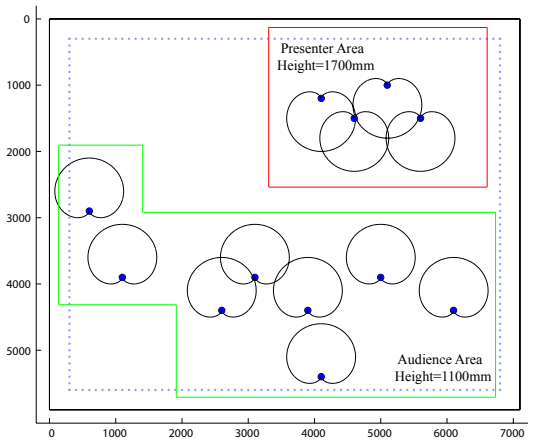
For the case of optimising over both the audience and presenter areas, the optimal configuration obtained, is presented in figure 5.5d. It is seen from this that although the obtained optimal configuration reduces the localisation uncertainty compared to the initial configuration, some of the uncertainty regions overlap. If the configuration determined by the proposed algorithm is assumed to be optimal, then this would suggest that the use of $K = 4$ microphone arrays is insufficient for the described localisation problem. Also, an interesting observation in the returned optimal configuration is that all microphone arrays are positioned on the ceiling and not on the walls. This can be attributed to the dependence of the employed TDE uncertainty model on the microphone-to-speaker distances. Positioning the arrays on the ceiling can be seen to reduce the average microphone-to-speaker distance and therefore reduces the overall TDE uncertainty.

The second scenario examined, is identical to the previous example but considers a non-symmetric audience area. This corresponds to a more realistic situation where the audience members are less constrained to a formal audience layout. The presenter area considered is also non-symmetric but the region within which the presenter is expected to move is smaller than that considered previously. This also corresponds to a more common scenario where the presenter remains within a relatively small area situated close to a laptop or podium for instance. Both non-symmetric audience and presenter areas are illustrated in figure 5.6a.

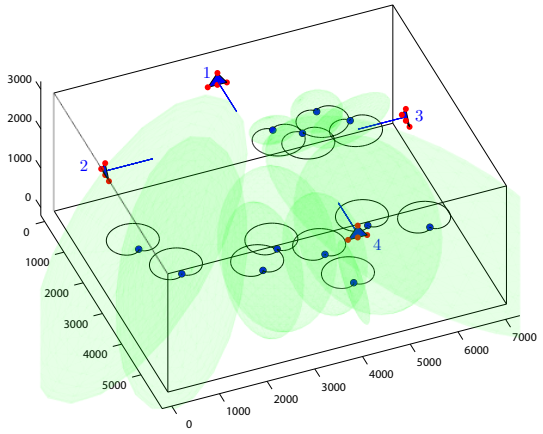
Again as expected, the initial microphone array configuration shown in figure 5.6b is seen to result in poor localisation accuracy. The optimal configuration obtained for localising the audience only for this problem is presented in figure 5.6c. This result is in close correspondence with that of figure 5.5c and a loose symmetry in the configuration of microphone arrays is observed. Shown in figure 5.6d is the obtained configuration by optimising over both the audience and lecturer areas. Once again considerable improvement upon the initial configuration of figure 5.6b is observed. In addition to this it is interesting that the returned optimal configuration is that which distributes microphone array over both the audience and presenter areas in proportions representing their relative sizes (i.e. three arrays over the audience area and a single array over the presenter area).

The final example considers the case of non-symmetric audience and presenter areas as in the previous example. In this case however, the arrays are constrained to the walls and also, additional user defined regions within the room. This example is used to show the practical use of the algorithm to optimise an array configuration for user defined constraints. This has practical uses since it enables the user to exclude regions where microphone arrays are not to be placed, such as in places which would obstruct the presenter or obscure the audience's view of the presenter.

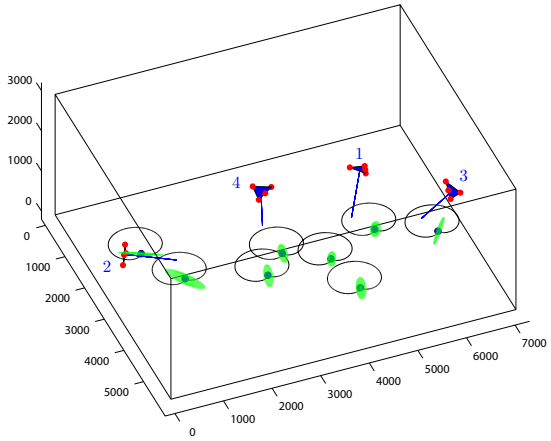
A typical scenario is described and illustrated in figure 5.7a where the grid of feasible locations



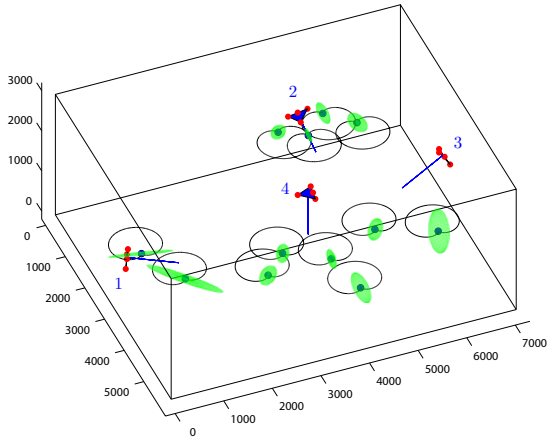
(a) *Non-symmetric audience (green) and presenter (red) areas facing in opposite directions. The audience is assumed seated at a height of* $1.1m$ *and the presenter is assumed standing at a height of* $1.7m$*. The cardioid directivity patterns are shown in black.*



(b) *Initial microphone array configuration with* 95 *percentile localisation error ellipsoids shown in green over the audience and presenter areas.*

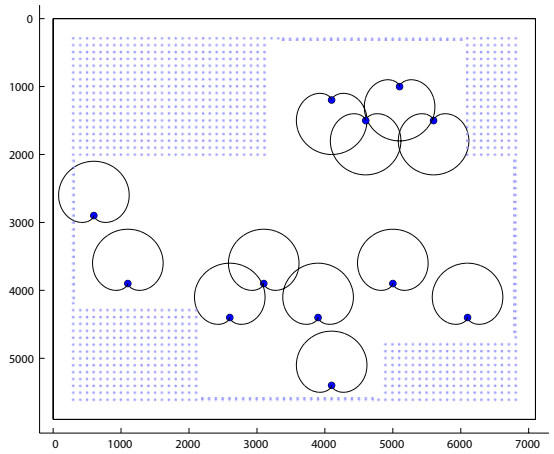


(c) Microphone array configuration optimised over the audience area only with localisation error ellipsoids shown in green.

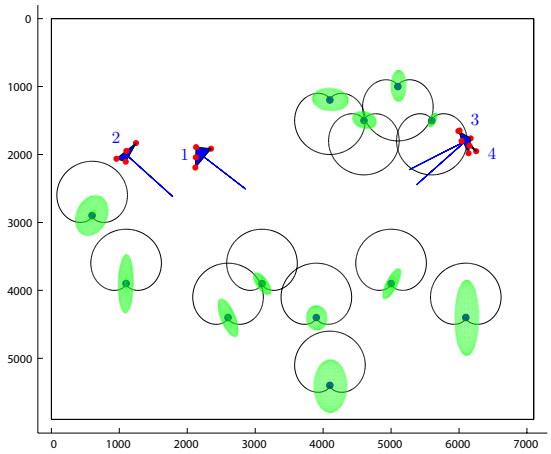


(d) *Microphone array configuration optimised over both the audience and presenter areas with localisation error ellipsoids shown in green.*

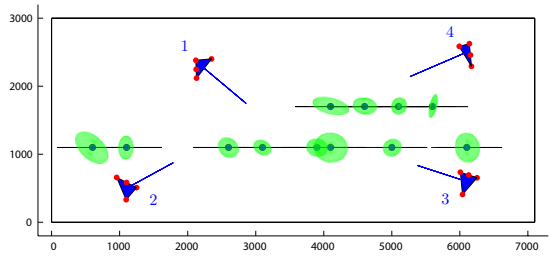
Figure 5.6: *Optimising the configuration of* 4 *inverted T-shaped microphone arrays for the case of non-symmetric audience and presenter regions facing in opposite direction. In this example the centre of the arrays are constrained to the room boundaries (i.e. walls, floor and ceiling). The localisation error ellipsoids illustrated correspond to the* 95 *percentile error region determined using the CPE uncertainty model of equation 5.17 with* $Pr[\epsilon] = 0$.



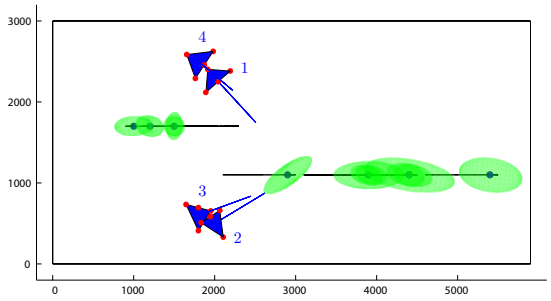
(a) *The optimisation problem as described in figure 5.6a with the microphone array centres constrained to a larger user defined region in the room shown by the light blue dots.*



(b) *View of the xy plane showing the localisation error ellipsoids (green) resulting from the optimised array configuration.*



(c) *View of the xz plane showing the localisation error ellipsoids (green) resulting from the optimised array configuration.*



(d) *View of the yz plane showing the localisation error ellipsoids (green) resulting from the optimised array configuration.*

Figure 5.7: *Optimising the configuration of* 4 *inverted T-shaped microphone arrays within user specified regions for localisation over non-symmetric audience and presenter regions. The localisation error ellipsoids illustrated correspond to the* 95 *percentile error region determined using the CPE uncertainty model of equation 5.17 with* $Pr[\epsilon] = 0$.

$\mathbb{Z}_c$ is defined for locations on the walls and for regions within the room about the audience and presenter areas. In this example the locations on the ceiling are not contained in $\mathbb{Z}_c$. The resulting optimal array configuration showing uncertainty regions in the xy plane, xz plane and yz plane is shown in figure 5.7b, figure 5.7c, figure 5.7d respectively. Once again the algorithm performs well under the more complex constraints. Furthermore, it is interesting in this example with ceiling locations in $\mathbb{Z}_c$ excluded, that the microphone arrays tend towards an even distribution along each axis. This can be clearly seen in figure 5.7c, figure 5.7d where two microphone arrays are positioned above the speaker positions and two below.

5.5 Final Comments

In this chapter the general problem of optimising the placement of microphone arrays was examined. This analysis was based on a theoretical model of TDE uncertainty which accounted for the effects of reverberation. The directivity characteristics of both the speaker and microphones were also incorporated into this model of uncertainty. Using this complete model of TDE uncertainty, a simulated annealing algorithm was proposed which determines an optimal microphone array configuration. This algorithm enables microphone array positions to be optimised automatically within a set of feasible user defined regions.

The results of this chapter also suggest that for each set of possible audience and presenter locations in a lecture room, there is a unique optimal microphone array configuration. Therefore, if the positions of the presenter and audience members are not constrained, a fixed microphone configuration in a lecture room is likely to be sub-optimal. This is one possible explanation for the poor performance of joint audio-video based tracking algorithms which have been reported using data recorded in the CHIL room [133]. This chapter highlights that if a microphone array configuration is not optimal, then poor audio-based localisation is likely. The previous chapter highlighted that when poor audio-based localisation exists it can greatly limit the effectiveness of joint audio-video fusion based tracking. This suggests that any successfully application of joint audio-video based fusion in the CHIL room would be difficult and severely restricted by the quality of the audio data. This supports the use of some other method of combining audio and video for tracking and is used to motivate the manner in which audio and video are combined in the speaker tracking algorithm described in chapter 6.

6

Voxel-based Viterbi Active Speaker Tracking V-VAST

In this chapter, a new algorithm entitled Voxel-based Viterbi Active Speaker Tracking (V-VAST) is introduced for tracking the current active speaker in a multi-microphone and multi-camera recording of a lecture. The particular lecture room environment considered is that of the CHIL lecture room described in section 1.1. The tracking algorithm relies on both video data from multiple camera views and audio data from multiple microphone arrays to infer the $3D$ position of the active speaker over the duration of the captured presentation. The tracking of active speakers is performed off-line and is proposed as a post-production step to visually segment the head region of the active speaker within the scene. The algorithm composes a composite view video sequence of the lecture, consisting of a user defined main view and an inserted view of the active speaker. The segmented view of the active speaker is automatically determined by V-VAST. The focus of this work is towards the automated editing of a multi-view lecture recording into a single view video sequence for presentation over the Internet as in asynchronous eLearning applications or for the purpose of archiving.

V-VAST differs from the traditional joint audio-video based active speaker tracking systems as reviewed in chapter 2. Unlike these, V-VAST does not fuse audio and video location estimates in a statistical sense. Drawing on the evaluation of localisation accuracy in chapter 4, it is argued that the statistical fusion of audio and video based location estimates does not greatly improve localisation accuracy beyond that of video-based localisation. This is due to the poor reliability and accuracy of audio-based localisation. Furthermore, it is argued that audio-based localisation in the CHIL lecture room is likely to be poor since the analysis outlined in chapter 5 revealed that the positions of the microphone arrays within the lecture room are sub-optimal. Taking

this into consideration, V-VAST proposes to use the video-data to detect likely head positions of speakers and to use this to guide an audio-based system for tracking the active speaker.

The complete V-VAST system is outlined in the block diagram of figure 6.1a. At each time instance over the duration of the lecture, a set of candidate speaker positions is obtained using the multi-camera video data. In extracting candidate speaker positions, V-VAST applies a *voxel-based scene analysis* to the $3D$ space of the lecture room. In a top-down manner, each voxel represents a hypothesised speaker position which is confirmed or rejected based on skin colour masks obtained in each camera view. The skin colour masks are obtained using the new skin colour model proposed in section 2.2.3. The result of the voxel-based analysis is a $3D$ foreground denoting possible speaker occupancy in the lecture room. From this $3D$ foreground, individual regions are determined through a $3D$ connected component analysis. A $2D$ connected component analysis on each skin mask enables individual connected $3D$ foreground regions to be associated with connected skin colour regions in each view. The approach taken in detecting head locations is to define an ellipsoidal head model and to fit the ellipsoid not only to the $3D$ foreground but also to its corresponding skin region in each view. Placing such strong constraints on the fitting process enables V-VAST to determine head positions to a high degree of accuracy.

Given the detected head positions, a measure of speaker activity is then determined at each candidate position using TDEs from multiple microphones. This corresponds to the *3D active speaker localisation* block within the system. Exploiting the offline nature of the tracking task, V-VAST uses both past and future audio measurements to improve tracking reliability. This is achieved by examining speaker activity over a window of three time steps centred at the current time instance. By modelling for speaker activity, this allows a prior to be introduced which assigns a high probability to speaker positions that correspond to significant speaker activity over the analysis window. This formulation enforces smoothness on the estimated speaker activity path and penalises transitions to positions where speaker activity is temporally insignificant.

Using the detected head positions and associated measure of speech activity, *speaker activity path tracking* is performed using the Viterbi algorithm [57]. This determines a MAP estimate of the current speaker through the set of candidate speaker head positions over the duration of the recorded lecture. The use of the Viterbi algorithm is inspired by its recent application to MAP estimation problems using Particle Filters [163] and multi-object tracking in video sequences [53].

In applying the Viterbi algorithm for tracking, two different motion priors are examined in this chapter for use with V-VAST. The first focuses on the single speaker tracking task and is modelled simply as a Gaussian white noise process. In order to account for both active speaker motion and active speaker switches however, a second motion prior is examined. The second motion prior only enforces a Gaussian prior on the motion locally about the current active speaker position and a uniform prior density at all other positions.

The final stages implemented by V-VAST relate to the visual segmentation of the active speaker from the available camera views. Using both the detected head regions and estimated

speaker activity path, a *visual segmentation* process extracts the active speaker from each view in which they are visible. V-VAST addresses the *best view selection* problem and automatically chooses the *best* view of the current active speaker by determining in which view the speaker's face is most visible. Finally, the segmented view of the active speaker is inserted into a user defined main view to create a single *composite view* video sequence. Some sample composite view video frames for a CHIL lecture recording created by V-VAST are shown in figure 6.1b. The CHIL lecture recordings are also used in evaluating the performance of the algorithm later in the chapter.

Problem Statement

The task is to determine the position of the current active speaker within the $3D$ space of the room. The multi-speaker case is not considered. Given that the nature of lecture presentations is to follow a single track of communication, few occurrences of simultaneous multi-party speech are observed. Furthermore, when multi-party speech is observed, it is only at brief overlaps occurring at the transitions in conversational exchanges. The following outlines the probabilistic framework for determining a MAP estimate of the path of speaker activity.

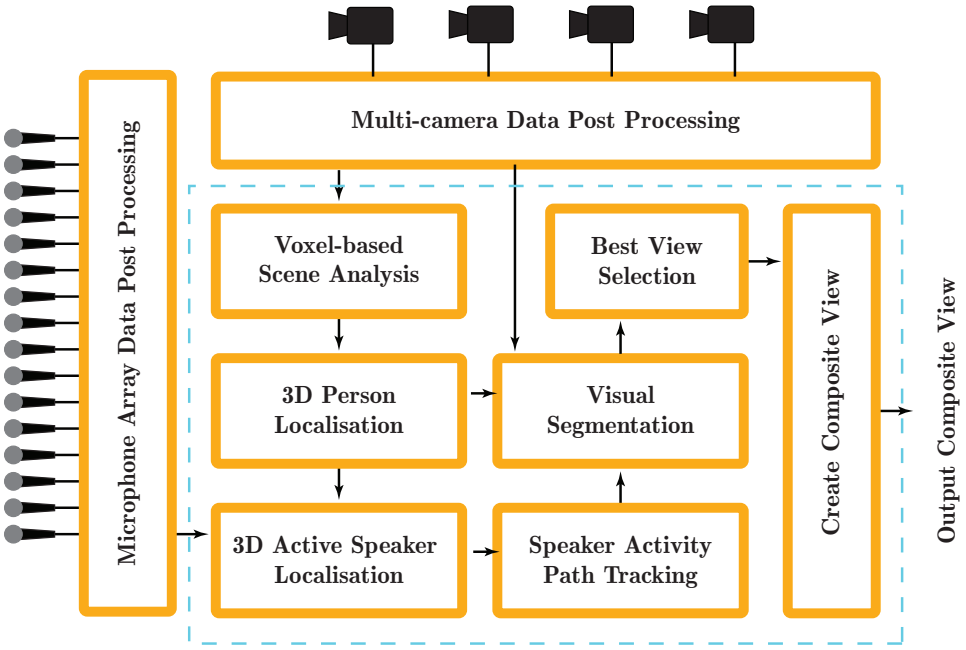
6.1 Probabilistic Framework

Let $\mathbf{x}_k \in \mathbb{R}^3$ denote a speaker position at some time k. Associated with the speaker position $\mathbf{x}_k$ is a binary label $s_k(\mathbf{x}_k) \in [0, 1]$ indicating speaker activity, where $s_k(\mathbf{x}_k) = 1$ labels $\mathbf{x}_k$ as active and $s_k(\mathbf{x}_k) = 0$ labels $\mathbf{x}_k$ as inactive. Speaker activity for $\mathbf{x}_k$ is examined using this label over a window of three time steps centred at time k and combined into a vector,

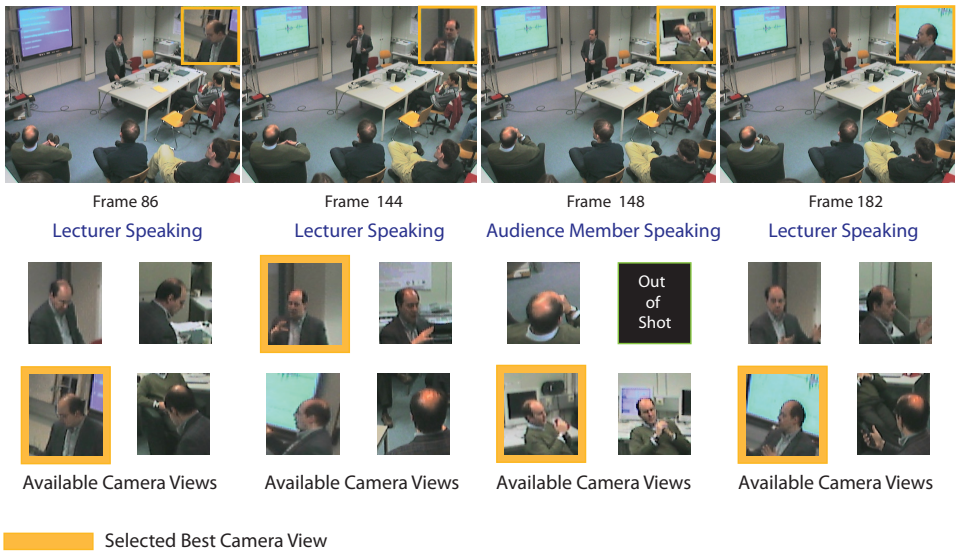
$$\mathbf{s}_k = [s_{k-1}(\mathbf{x}_k), s_k(\mathbf{x}_k), s_{k+1}(\mathbf{x}_k)]. \tag{6.1}$$

Any speaker position $\mathbf{x}_k$ therefore, can be in one of eight possible states. For instance, if a speaker at position $\mathbf{x}_k$ is currently speaking; was speaking at the previous time step $k-1$; and will be speaking at time $k+1$, then for this position $\mathbf{s}_k = [1, 1, 1]$. For clarity, the eight possible states for speaker positions are denoted by $[s_0, s_1, ..., s_7]$ as is shown in table 6.1.

From N_{cam} camera views, a set of video frame data $\mathcal{I}_k = \{\mathbf{I}_1, .., \mathbf{I}_{N_{cam}}\}$ is obtained at each time instance k. Also at time k from M_k microphone pairs, a set of TDEs $\mathbf{y}_k = \{y_k^1, ..y_k^m .., y_k^{M_k}\}$ is obtained where $y_k^m \in \mathbb{R}$ for $m = 1, .., M_k$. Given the complete set of video-based observations $\mathcal{I}_{0:K}$, where K is the duration of the sequence and audio-based observations $\mathbf{y}_{0:K}$, we wish to estimate the posterior distribution $p(\mathbf{x}_{0:K}, \mathbf{s}_{0:K} | \mathcal{I}_{0:K}, \mathbf{y}_{0:K})$. This represents our joint belief in the speaker position $\mathbf{x}_k$ and speaker activity state $\mathbf{s}_k$ over the duration of the sequence based on the complete set of audio and video based observations. In this derivation the path of speaker activity is defined as the joint MAP estimate of both $\mathbf{x}_k$ and $\mathbf{s}_k$ which is determined from the



(a) *Block diagram of V-VAST, an algorithm for tracking the active speaker in a multi-camera, multi-microphone recording of a lecture. The system outputs a composite view video sequence consisting of a user defined main view and an automatically segmented view of the current active speaker. The blocks encased in the dotted blue line are described in this chapter.*



(b) *Sample composite view video sequence output by V-VAST. The top row shows sample composite view video frames of a CHIL lecture recording created by V-VAST. The second row shows the current active speaker segmented in each available camera view. The best view selected by V-VAST is highlighted with an orange border. In this particular example, in frame 86, 144 and 182 the presenter is the active speaker however, in frame 148 an audience member asks a question and becomes the active speaker.*

Figure 6.1: *Block diagram of the Voxel-based Viterbi Active Speaker Tracking (V-VAST) algorithm and sample output composite view video sequence.*

Speaker Activity State Labels	
$\mathbf{s}_k = \mathbf{s}_0 = [0,0,0]$	$\mathbf{s}_k = \mathbf{s}_4 = [1,0,0]$
$\mathbf{s}_k = \mathbf{s}_1 = [0,0,1]$	$\mathbf{s}_k = \mathbf{s}_5 = [1,0,1]$
$\mathbf{s}_k = \mathbf{s}_2 = [0,1,0]$	$\mathbf{s}_k = \mathbf{s}_6 = [1,1,0]$
$\mathbf{s}_k = \mathbf{s}_3 = [0,1,1]$	$\mathbf{s}_k = \mathbf{s}_7 = [1,1,1]$

Table 6.1: *The 8 possible speaker activity states* $\mathbf{s}_k(\mathbf{x}_k) = [s_{k-1}, s_k, s_{k+1}]$ *for speaker position* $\mathbf{x}_k$.

posterior distribution as,

$$[\mathbf{x}_{0:K}^{MAP}, \mathbf{s}_{0:K}^{MAP}] = \underset{\mathbf{x}_{0:K}, \mathbf{s}_{0:K}}{\arg\max}\, p(\mathbf{x}_{0:K}, \mathbf{s}_{0:K} | \mathcal{I}_{0:K}, \mathbf{y}_{0:K}) \tag{6.2}$$

If the position of the active speaker $\mathbf{x}_k$ is assumed to be a Markov process of order 1, i.e.

$$p(\mathbf{x}_k | \mathbf{x}_{0:k-1}, \mathbf{s}_k) = p(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{s}_k), \tag{6.3}$$

then the posterior distribution at a single time instance k can be obtained in using Bayes' law. The full Bayesian expression for this estimation problem is

$$p(\mathbf{x}_k, \mathbf{s}_k | \mathcal{I}_{0:K}, \mathbf{y}_{0:K}, \mathbf{x}_{k-1}) = \frac{p(\mathcal{I}_{0:K}, \mathbf{y}_{0:K} | \mathbf{x}_k, \mathbf{s}_k) p(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{s}_k)}{p(\mathcal{I}_{0:K}, \mathbf{y}_{0:k})}. \tag{6.4}$$

Since the denominator of equation 6.4 does not depend explicitly on the parameters of interest $\mathbf{x}_k$ and $\mathbf{s}_k$ it is sufficient to express this relation as a proportionality by,

$$p(\mathbf{x}_k, \mathbf{s}_k | \mathcal{I}_{0:K}, \mathbf{y}_{0:K}, \mathbf{x}_{k-1}) \propto p(\mathcal{I}_{0:K}, \mathbf{y}_{0:K} | \mathbf{x}_k, \mathbf{s}_k) p(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{s}_k). \tag{6.5}$$

V-VAST only considers the current video measurements $\mathcal{I}_k$ and only audio-based measurements $\mathbf{y}_{k-1:k+1}$ at any time step k. The *posterior* is therefore simplified to,

$$p(\mathbf{x}_k, \mathbf{s}_k | \mathcal{I}_k, \mathbf{y}_{k-1:k+1}, \mathbf{x}_{k-1}) \propto p(\mathcal{I}_k, \mathbf{y}_{k-1:k+1} | \mathbf{x}_k, \mathbf{s}_k) p(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{s}_k). \tag{6.6}$$

6.1.1 Audio-based and Video-based Likelihood Functions

In this derivation it is assumed that given $\mathbf{x}_k$, the audio-based observations $\mathbf{y}_{k-1:k+1}$ and video observations $\mathcal{I}_k$ are independent. The importance of this assumption among existing joint audio-video trackers was examined in section 3.2. It is sensible to assume independent measurements since the only common dependence between the audio and video measurements is the speaker position $\mathbf{x}_k$. Through this, the likelihood function of equation 6.6 then becomes,

$$p(\mathcal{I}_k, \mathbf{y}_{k-1:k+1} | \mathbf{x}_k, \mathbf{s}_k) = p(\mathcal{I}_k | \mathbf{x}_k) p(\mathbf{y}_{k-1:k+1} | \mathbf{x}_k, \mathbf{s}_k) \tag{6.7}$$

which expresses the likelihood as a product of an audio-based likelihood $p(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k)$ and a video-based likelihood $p(\mathcal{I}_k|\mathbf{x}_k)$. These are designed next.

6.1.1.1 Audio-based Likelihood

The audio-based likelihood relates the speaker position $\mathbf{x}_k$ and speech activity state $\mathbf{s}_k$ to the TDEs measured at each of the microphone pairs. In this section the likelihood model is first defined for a single microphone pair yielding a single TDE and later for the complete likelihood over all microphone pairs.

For the mth microphone pair the audio-based likelihood function is defined as follows,

$$p(y^m_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) \propto \prod_{j=0}^{2} \exp\left(\frac{(g_m(\mathbf{x}_k) - y^m_{k-1+j})^2}{2\sigma^2_{g_m(\mathbf{x}_k)}} \mathbf{s}_k(j) + \frac{\alpha^2}{2\sigma^2_{g_m(\mathbf{x}_k)}} (1 - \mathbf{s}_k(j)) \right) \tag{6.8}$$

where $g_m(\cdot)$ is the time delay measurement function for the mth microphone pair and $g_m(\mathbf{x}_k)$ is the expected time delay associated with the position $\mathbf{x}_k$. The variance $\sigma^2_{g_m(\mathbf{x}_k)}$ is the estimated variance of a time delay estimate associated with the position $\mathbf{x}_k$ obtained through the covariance mapping technique described in section 4.1.1. The parameter α in this likelihood is $\alpha = 2.76\sigma_{g_m(\mathbf{x}_k)}$ which relates to the 99 percentile region for a Gaussian distribution. The notation $\mathbf{s}_k(j)$ is introduced to indicate the jth element of the speaker activity state vector as defined in equation 6.1. In essence, the likelihood of equation 6.8 can be thought of as the likelihood of $\mathbf{x}_k$ being the current active speaker position evaluated for each of the eight possible speaker activity states of $\mathbf{s}_k$. An example illustrating the form which the likelihood takes is illustrated in figure 6.2.

The problem of combining multiple measurement data in estimation problems encompasses many complex tasks from defining a meaningful manner in which the information can be fused, to determining useful observations from possible anomalous [1] estimates. The common data fusion strategy in problems relating to tracking, is to make the assumption that the likelihoods are independent. Hence the likelihood is a product of individual likelihood densities. To apply such an approach to the problem presented here, the TDEs observed at each pair of microphones are assumed to be independent such that,

$$p(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) = \prod_{m}^{M_k} p(y^m_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k). \tag{6.9}$$

One problem which can arise through this approach is that, due to the product across all observations, a single anomalous TDE can result in an overall likelihood expression which does not reflect the majority of the observations. This can be seen by a simple example.

Consider the hypothetical example of three TDEs, $y^1_{k-1:k+1}$, $y^2_{k-1:k+1}$, $y^3_{k-1:k+1}$ where due to

[1] The term "anomolous TDE" is more commonly used than "TDE outlier" in the time-delay estimation literature however they have equal meaning.

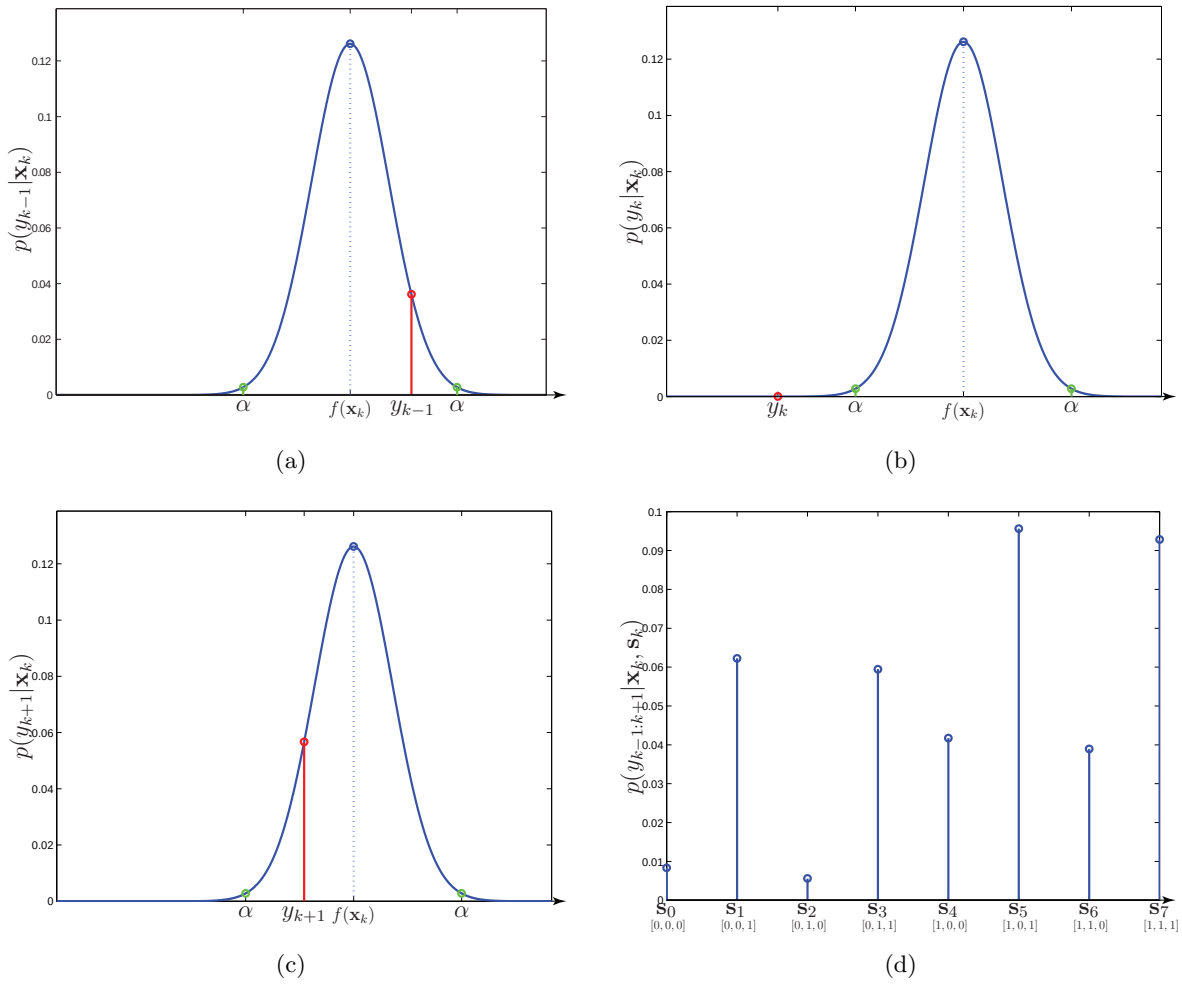


Figure 6.2: *Example of the audio-based likelihood for a single microphone pair determined over a window of three time-steps. The likelihood of the TDE measurement given the speaker position* $\mathbf{x}_k$ *at time* $k-1$ *is shown in* (a)*.* (b) *shows the TDE measurement likelihood at time* k *and* (c) *is the TDE measurement likelihood at time* $k+1$*. The overall likelihood evaluated at each speaker activity state* $\mathbf{s}_k$ *is presented in* (d)*. The given case illustrates the scenario of a speaker position with speaker activity state* $\mathbf{s}_k = s_5$ *from table 6.1 which can be seen from* (d) *to yield a maximum in the likelihood* $p(y_{k-1:k+1}|\mathbf{x}_k, \mathbf{s})$*.*

some anomaly the TDE $y^3_{k-1:k+1}$ is erroneous. For a single speaker position $\mathbf{x}_k = \mathbf{a}$ where the speaker activity state is known to be $\mathbf{s}_k = s_7$, the three likelihood functions are found to be,

$$p(y^1_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k) = [0.0668, 0.0662, 0.0668, 0.0669, 0.0667, 0.0665, 0.0668, 0.5333] \quad (6.10a)$$

$$p(y^2_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k) = [0.0780, 0.0775, 0.0759, 0.0765, 0.0780, 0.0750, 0.0770, 0.4621] \quad (6.10b)$$

$$p(y^3_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k) = [0.1200, 0.1223, 0.2568, 0.1255, 0.1245, 0.1266, 0.1241, 0.0002]\,. \quad (6.10c)$$

An examination of the MAP estimate of the speaker activity state at microphone pairs 1, 2 and

3 yields $\mathbf{s}_k = \mathbf{s}_7$, $\mathbf{s}_k = \mathbf{s}_7$ and $\mathbf{s}_k = \mathbf{s}_2$ respectively. If the independent likelihood model is applied in the described case, the overall likelihood function is determined as,

$$p(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k) = p(y^1_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k)p(y^2_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k)p(y^3_{k-1:k+1}|\mathbf{x}_k = \mathbf{a}, \mathbf{s}_k) \tag{6.11}$$

$$= [0.0006, 0.0006, 0.0013, 0.0006, 0.0006, 0.0006, 0.0006, 0.0001]. \tag{6.12}$$

The independent likelihood therefore indicates that the speaker activity state is $\mathbf{s}_k = \mathbf{s}_2$ which contradicts the majority of the sensor observations. This occurs since measurements in the independent likelihood model are assumed to be equally reliable and a single erroneous measurement can result in a likelihood function which does not reflect that of the majority of microphone pairs.

In order to account for anomalous TDEs, it is necessary to only select a subset $\mathbf{M}_k \subseteq \{1, ..., M_k\}$ of the observations which meet some measure of reliability. In this new definition the likelihood over all microphone pairs becomes,

$$p(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) = \prod_{m \in \mathbf{M}_k} p(y^m_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) \tag{6.13}$$

The problem of determining the reliability of TDEs however is not straightforward.

6.1.1.2 Approximate Robust Likelihood

Strategies for defining the reliability of TDEs were reviewed in section 2.1.3.2. Simple measures of the strength of the correlation peak [33] can be effective in identifying reliable TDEs but there is no definitive reliability measure at present. In an attempt to address the task of combining the individual likelihood functions of equation 6.8, an approximate likelihood is proposed which is robust in the presence of anomalous TDEs and reflects the number of microphone pair observations which are in agreement. This is achieved by determining a MAP estimate of the speaker activity state $\mathbf{s}_k$ for $\mathbf{x}_k$ at each microphone pair and then defining a histogram over all microphones pairs to represent the overall likelihood. The histogram defines a function $h_k(\mathbf{s}_k, \mathbf{x}_k) = c_M$ where c_M is the number of microphone pairs whose MAP estimate of speaker activity is $\mathbf{s_k}$. In order for the histogram to be a valid approximation of the overall likelihood function it must be normalised. For the example of equation 6.10, using this method to approximate the likelihood results in,

$$p_h(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) = [0, 0, 0.333, 0, 0, 0, 0, 0.667]. \tag{6.14}$$

where the notation p_h is used to denote this new form of the likelihood probability density. From this it can be seen that the approximate likelihood reflects the observations of the three TDEs more accurately than the independent likelihood model of equation 6.12 showing robustness to erroneous TDEs.

6.1.1.3 Video-based Likelihood

A set of head positions $\mathbf{z}_k = \mathbf{z}_k^1, .., \mathbf{z}_k^i, ..\mathbf{z}_k^{N_k}$, $\mathbf{z}_k^i \in \mathbb{R}^3$ are determined in the space of the room by pre-processing the multi-view video data $\mathcal{I}$. The details of this pre-process are discussed later in section 6.2. The video likelihood given $\mathbf{z}_k$ can be considered as a set of delta functions of equal height, one at each estimated head position in $\mathbf{z}_k$. This creates a grid based estimate of the likelihood distribution of 6.7 as

$$p(\mathcal{I}_k, \mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) \propto \sum_{i=1}^{N_k} \delta(z_k^i - \mathbf{x}_k) p_h(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) \tag{6.15}$$

Estimated active speaker positions are therefore constrained to that of video-based estimates of head positions. Hence, the posterior distribution of equation 6.6 becomes,

$$p(\mathbf{x}_k, \mathbf{s}_k|\mathbf{y}_{k-1:k+1}) \propto \sum_{i=1}^{N_k} \delta(z_k^i - \mathbf{x}_k) p_h(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k, \mathbf{s}_k) p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{s}_k). \tag{6.16}$$

Equivalently, this formulation reduces the continuous state space of speaker positions $\mathbf{x}_k$ to a set of discrete positions $\mathbf{x}_k = \{\mathbf{x}_k^1, .., \mathbf{x}_k^i .., \mathbf{x}_k^{N_k}\}$. In this case, the viterbi algorithm can be used to determine the MAP sequence of joint estimates $[\mathbf{x}_{1:k}^{MAP}, \mathbf{s}_{1:k}^{MAP}]$ as defined in equation 6.2. This is discussed later in section 6.3.

6.1.2 Priors

In order to account for both the movement of speakers and also to enable the changing of speaker activity between speaker positions, both a motion prior and speaker activity prior is introduced. This follows from the reasonable assumption that $\mathbf{x}_k$ and $\mathbf{s}_k$ are independent so that the prior density $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{s}_k)$ in equation 6.6 becomes,

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{s}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_k|\mathbf{s}_k). \tag{6.17}$$

6.1.2.1 Motion Model

Modelling states of motion such as constant velocity, constant acceleration and constant position is straightforward. People however, are not constrained to any particular motion and therefore appropriately modelling their motion in tracking applications is challenging. Over a tracking sequence, it is typical to observe a person progress through multiple different states of motion. As described in section 3.1.2, inaccurate motion modelling can cause tracking divergence resulting in poor performance. Usually, in people tracking applications, motion is typically modelled as a Gaussian-Markov random process and the motion prior is defined as,

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) \propto \exp(-\frac{1}{2}(\mathbf{x}_k - \mathbf{x}_{k-1})' \Sigma^{-1} (\mathbf{x}_k - \mathbf{x}_{k-1})). \tag{6.18}$$

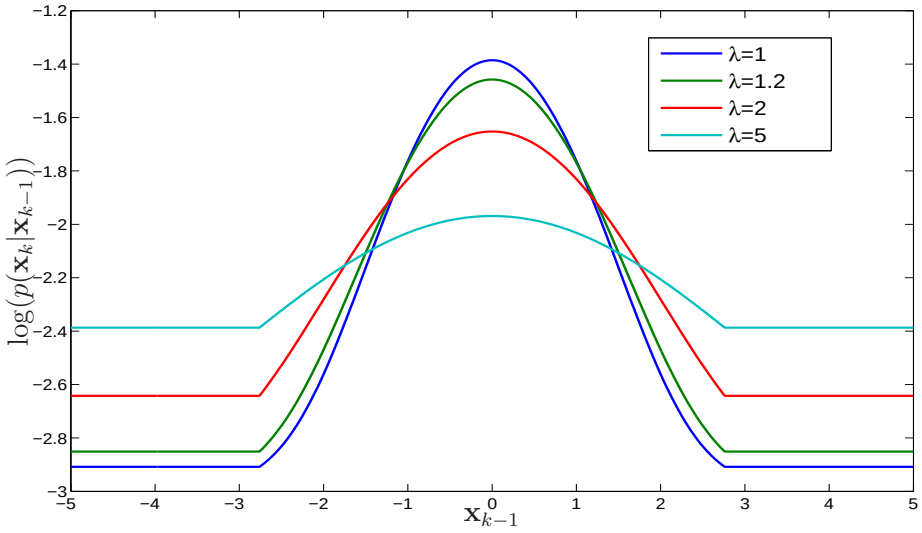


Figure 6.3: *Example of the motion prior of equation 6.19 for the* $1D$ *case with* $\mathbf{x}_{k-1} = 0$ *and* $\Sigma = 1$ *for* $\lambda = 1$, $\lambda = 1.2$, $\lambda = 2$ *and* $\lambda = 5$.

Since it is desired to track switches in speaker activity, a Gaussian prior is inappropriate since the assumption that the next active speaker location is close to the current active speaker position does not apply. This is because switches in speaker activity are not dictated by proximity. A different speaker position at any point in the room may become the current active speaker at the next time step.

In order to address the speaker switching scenario, another prior motion model is considered. This prior can be considered as a piecewise probability density consisting of Gaussian and uniform density components. The piecewise density is constructed such that within a defined region close to the previous speaker location $\mathbf{x}_{k-1}$ the current speaker location $\mathbf{x}_k$ has a Gaussian distribution whereas outside of this region $\mathbf{x}_k$ is uniformly distributed within the space of the room. More formally this can be defined as,

$$p(\mathbf{x}_k, |\mathbf{x}_{k-1}) \propto \begin{cases} \exp(-\frac{d}{2}) & d < \frac{\beta^2}{\lambda} \\ \exp(-\frac{\beta^2}{2\lambda}) & d \geq \frac{\beta^2}{\lambda} \end{cases} \tag{6.19a}$$

$$\text{where} \quad d = (\mathbf{x}_k - \mathbf{x}_{k-1})'(\lambda\Sigma)^{-1}(\mathbf{x}_k - \mathbf{x}_{k-1}) \tag{6.19b}$$

and $\beta = 3.37$ relates to the 99 percentile region to the $3D$ covariance matrix Σ. In this definition the weight λ is introduced which is used to vary the weighting between the uniform and Gaussian density components. A typical value for λ is 1. An example of the prior of 6.19 in the $1D$ case and the effects of the weight λ is illustrated in figure 6.3.

6.1.2.2 Speaker Activity Prior

A prior on the speaker position given the speaker activity state $\mathbf{s}_k$ is also incorporated into the estimation problem. The purpose of this prior is to assign a high probability to speaker positions where speaker activity is observed to be temporally significant. This prior reflects the intuition that speaker positions with speaker activity states $\mathbf{s}_k = \{\mathbf{s}_3, \mathbf{s}_5, \mathbf{s}_6, \mathbf{s}_7\}$ should be assigned a higher probability of being the current active speaker than positions with speaker activity states $\mathbf{s}_k = \{\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_4\}$. Therefore, the speaker activity prior causes V-VAST to favour speaker positions which have been actively speaking for greater than or equal to two time-instances over a three time-step window.

The prior is configured as follows,

$$p(\mathbf{x}_k, |\mathbf{s}_k) \propto \exp(-\gamma(1 - \mathbf{s}_k(0)) - \gamma(1 - \mathbf{s}_k(1)) - \gamma(1 - \mathbf{s}_k(2))) \tag{6.20}$$

where $\gamma = 1$, although it can be varied to increase the weighting of the prior. The aim of this prior is to make the algorithm insensitive to speaker positions or possible noise sources which are not observed as significantly active.

6.1.2.3 Choosing an Appropriate Time-step Duration

The choice of the time step duration over which the V-VAST is applied is critical since it determines the temporal length of the window used to analyse speaker activity. Too small a time step would result in the algorithm being sensitive to brief speech utterances and all benefits in using the window to enforce smoothness on the estimated speaker activity path would be lost. Ideally, the time step should be chosen to reflect some measure of temporal significance in relation to speech activity. Previous research has reported the minimal vocalisation in a multiple person conversation as ≈ 1.64 seconds [91]. In order to reflect this a time step of duration of 1 second is chosen.

Although this time step may be suitable for the tracking of speaker activity, such a large time step is often not suitable in motion tracking applications. This is so, because a $1sec$ update rate could be too slow to track a person if their movements are fast and complex. Given the offline nature of tracking in this application, it is assumed that estimating speaker positions every $1sec$ gives a sufficient representation of the motion trajectory to enable the speaker's position to be determined at any point in time by spline interpolation. In order to account for discontinuities in the active speaker path due to speaker switches a shape preserving method of spline interpolation is employed [52].

6.2 Determining Candidate Speaker Positions

Possible speaker positions are estimated by head detection through a voxel based analysis in $3D$ space and skin colour detection in multiple camera views. This detection process is described in the following sections.

6.2.1 Extracting Connected Skin Regions

Determining candidate speaker positions relies heavily on accurately detecting skin regions within each of the camera views. To detect head positions in $3D$ space, a skin colour mask needs to be determined for each camera view. V-VAST employs the deterministic technique for skin colour detection which models for the non-linear dependence of skin-tone on luminance, as introduced in section 2.2.3.

Using this method of skin detection, a skin mask $\mathbf{S}_i$ for the ith camera view can be obtained using equation 2.57. The obtained skin mask $\mathbf{S}_i$ can be equivalently defined as the set of pixels $\mathbf{p}$ which correspond to skin regions as

$$\mathbb{S}_i = \{\mathbf{p}\ ;\ C_1(\mathbf{p}) = true, C_2(\mathbf{p}) = true\} \tag{6.21}$$

where C_1 and C_2 are as defined in equation 2.57. This definition is useful in analysing the problem of extracting connected skin regions. Connected skin regions are extracted by a $2D$ connected component analysis on the skin mask of each view. This in effect partitions the set of skin colour pixels $\mathbb{S}_i$ into,

$$\mathbb{S}_i = \{\mathbb{S}_{i1}, .., \mathbb{S}_{iN_i}\} \tag{6.22}$$

where $\mathbb{S}_{ij}$, $j = 1, .., N_i$ are disjoint sets of connected skin pixel regions and N_i is the total number of connected pixel regions. Knowing the relation of pixel locations $\mathbf{p}$ to that of sets of connected skin colour regions enables an indexing function $M_i(\mathbf{p})$ to be defined as,

$$M_i(\mathbf{p}) = \{\mathbb{S}_{ij}; \mathbf{p} \in \mathbb{S}_{ij}, j = 1, .., N_i\} \tag{6.23}$$

which returns the set of skin colour pixels connected to pixel site $\mathbf{p}$.

6.2.2 $3D$ Voxel-based Head Detection

The proposed approach for head detection determines likely head positions through a voxel-based analysis of the lecture room. The voxels represent hypotheses for head positions which are confirmed as occupied or unoccupied based on their relation to the skin colour masks obtained from the multiple camera views. The result of this analysis is a $3D$ foreground indicating volumes of skin regions in $3D$ space.

The voxelization of the room is performed by defining a uniform grid in $3D$ space. For the results presented in this chapter V-VAST employed a grid with a fixed resolution of $0.05m$ in

each dimension. The $3D$ grid is defined to fully occupy the width and depth of the lecture room but only the height region between $0.8m$ and $2.0m$. This grid represents the likely head positions of both seated and standing people in a lecture scenario.

A voxel is represented by its centre position $\mathbf{x}^{'}$ in the room. The $2D$ pixel location $\mathbf{p}_i^{'}$ corresponding to the voxel's centre in the ith view can be determined using the camera projection matrix $\mathbf{P}_i$ by,

$$\tilde{\mathbf{p}}_i^{'} = \mathbf{P}_i \tilde{\mathbf{x}}^{'}. \tag{6.24}$$

where $\tilde{\mathbf{p}}_i^{'}$ and $\tilde{\mathbf{x}}^{'}$ are the equivalent representations of $\mathbf{p}_i^{'}$ and $\mathbf{x}^{'}$ in homogeneous space (See section 2.2.4). In the actual implementation of V-VAST, the images of the pixel centres as obtained from equation 6.24 are rounded to integer pixel locations within the frame. Also, in an effort to reduce the computational complexity, pixel-to-voxel centre relations are maintained in a *look-up table* and calculated as a pre-process. This is possible because the cameras are fixed. The number of views indicating that a voxel is occupied by a skin region is defined as,

$$C_3(\mathbf{x}^{'}) = \sum_{k=1}^{N_{cam}} \mathbf{S}_i(\mathbf{p}_i^{'}). \tag{6.25}$$

Given this measure of voxel occupancy, a set of voxels $\mathbb{V}$ corresponding to the $3D$ foreground can be defined as,

$$\mathbb{V} = \{\mathbf{x}^{'}; C_3 = n_{cam}\}. \tag{6.26}$$

The minimum requirement for extracting a $3D$ foreground uses two views (i.e. $n_{cam} = 2$), however the more views which are used the more accurately the estimated $3D$ foreground represents the object's true form in $3D$ space. This is assuming that the camera views are positioned at relatively different angles about the object from each other. Since skin regions are typically not visible from every viewed angle of a head, using all cameras (i.e. $n_{cam} = N_{cam}$) is likely to result in a poor $3D$ foreground estimate. To address this issue the algorithm determines multiple $3D$ foreground estimates from all combinations of two or more views (i.e. $n_{cam} = [2, .., N_{cam}]$).

6.2.3 Extracting Connected Voxel Regions

Similar to the manner in which connected skin pixel regions are extracted within the skin mask of each view, occupied connected voxel regions in $3D$ space are also determined. This is achieved using a connected component analysis in $3D$ space. This enables the partitioning of the $3D$ foreground $\mathbb{V}$ into

$$\mathbb{V} = \{\mathbb{V}_1, .., \mathbb{V}_N\} \tag{6.27}$$

consisting of disjoint subsets $\mathbb{V}_i$, where $i = 1, .., N$ and N is the number of detected connected regions in the $3D$ foreground.

6.2.4 Ellipsoid Fitting

In the detection of head regions, the head is assumed to be an ellipsoid with four degrees of freedom. The four degrees of freedom correspond to a $3D$ translation and a rotation in the xy plane. This model is then fitted to each $3D$ foreground region and also to its corresponding $2D$ region in the skin mask of each camera view.

The parameters of the head model to be fitted to each of the $j = 1, .., N$ voxels region consist of the ellipsoid centre $\mathbf{C}_j$ and a rotation matrix

$$\mathbf{R}_j = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6.28}$$

where θ denotes the angle of head rotation. It is not the aim of the head fitting process to inherently estimate head pose. The incorporation of a head rotation is modelled so as not to overly constrain the fitting process. The size of the ellipsoid and hence the assumed size of the head is defined by a covariance matrix Σ_0^H. This covariance matrix is chosen such that its 95 percentile ellisoid has x, y and z axis dimensions of $194mm$, $145mm$ and $241mm$ respectively. These dimensions are in-line with reported head size statistics [77]. The covariance of the head associated with a connected $3D$ foreground region $\mathbb{V}_j$ therefore is $\Sigma_j^H = \mathbf{R}_j \Sigma_0^H$.

The fitting of the head model is driven by two error functions. The first of these enforces a weak constraint on the head fitting in that the head is loosely constrained to within the $3D$ foreground region $\mathbb{V}_j$. This error function is defined as,

$$E_{0j} = \min_{\mathbf{x}' \in \mathbb{V}_j} ((\mathbf{x}' - \mathbf{C}_j)^T \Sigma_j^H (\mathbf{X}' - \mathbf{C}_j)) \tag{6.29}$$

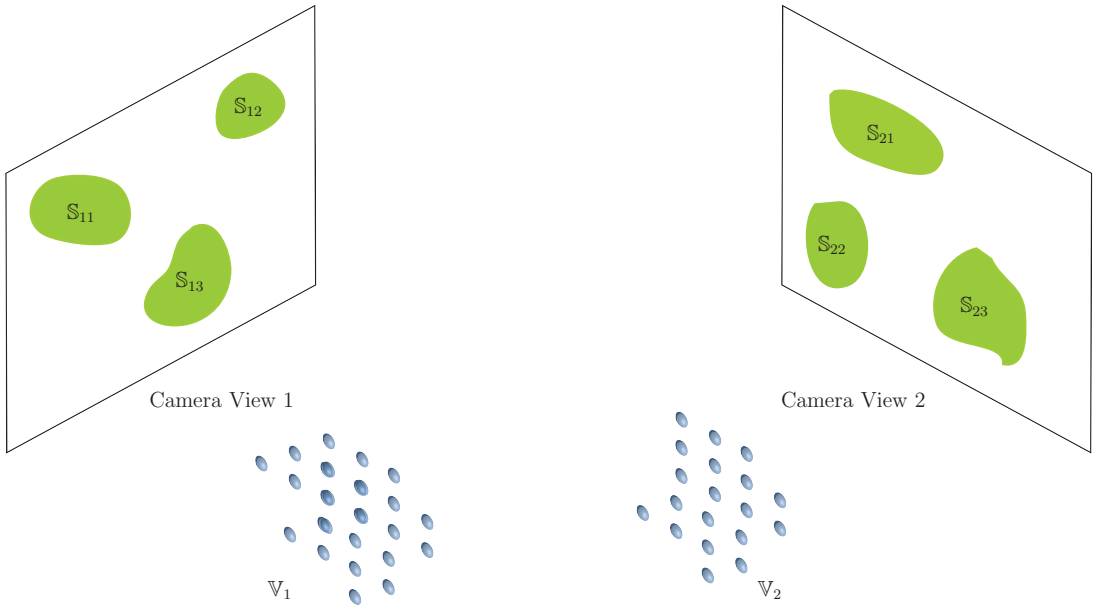
where $\mathbf{x}'_{jk}$ is the kth voxel of the jth $3D$ foreground region. This energy function is minimum if the centre of the ellipsoid is equal to that of a voxel location.

The second error function aims at enforcing the requirement that the position of the head within the $3D$ foreground region should also satisfy the projected view of the head in all camera views. This energy function is defined as follows,

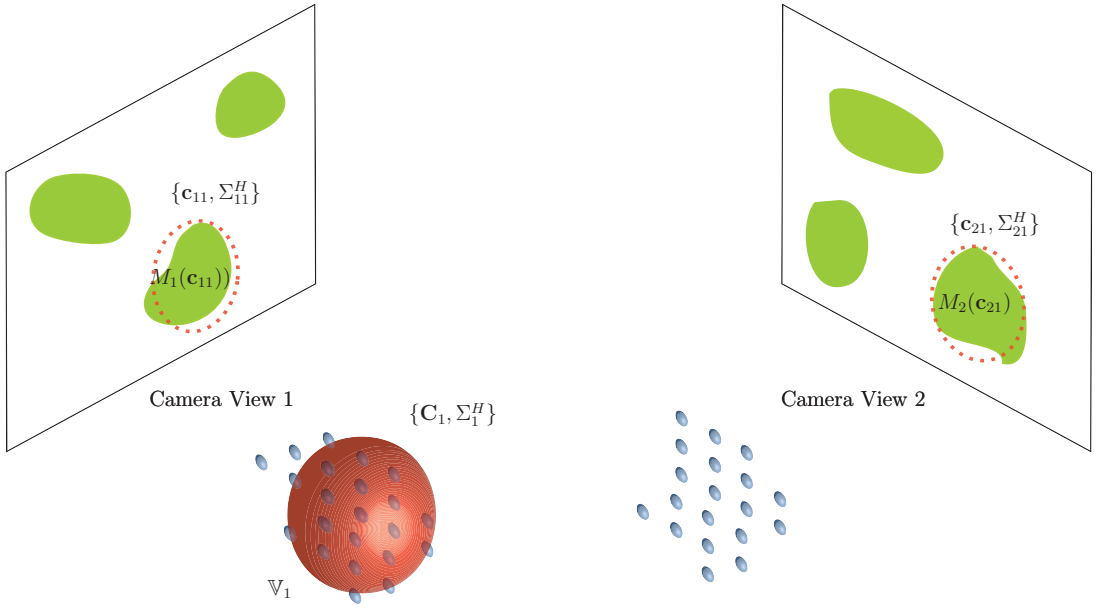
$$E_{ij} = \sum_{\mathbf{p}_i \in M_i(\mathbf{c}_{ij})} (\mathbf{p}_i - \mathbf{c}_{ij})^T \Sigma_{ij}^H (\mathbf{p}_i - \mathbf{c}_{ij}) \tag{6.30}$$

where Σ_{ij}^H is the projection of Σ_j^H into the ith view, $\mathbf{c}_{ij}$ is the pixel location of the ellipsoid centre $\mathbf{C}_j$ projected into the ith view and $\mathbf{p}_i$ is a skin pixel location in the ith view. Given these two error functions, the overall error function for the fitting is,

$$E_j = \sum_{i=0}^{N_{cam}} E_{ji}. \tag{6.31}$$

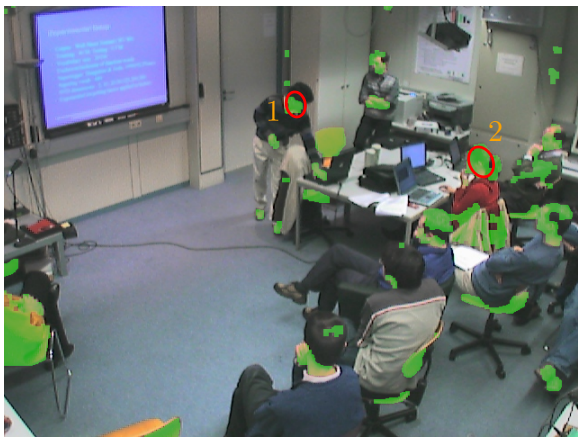


(a) *An example of head detection for a two view scenario with connected skin regions* $\mathbb{S}_1 = \{\mathbb{S}_{11}, \mathbb{S}_{12}, \mathbb{S}_{13}\}$ *in camera view* 1 *and* $\mathbb{S}_2 = \{\mathbb{S}_{21}, \mathbb{S}_{22}, \mathbb{S}_{23}\}$ *in camera view* 2. *A* 3*D foreground* $\mathbb{V} = \{\mathbb{V}_1, \mathbb{V}_2\}$ *consisting of two connected voxel regions* $\mathbb{V}_1$ *and* $\mathbb{V}_2$ *is also shown.*

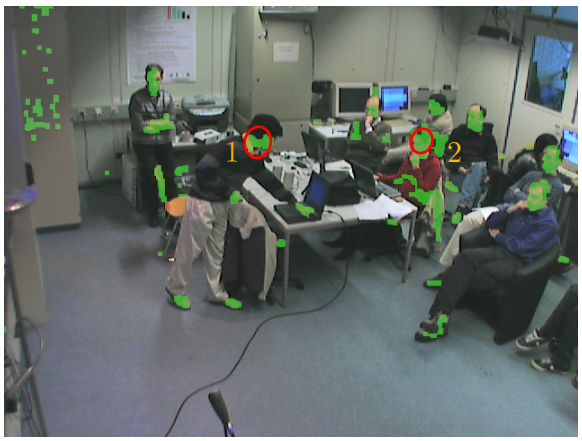


(b) *Fitting the ellipsoidal head model* $\{\mathbf{C}_1, \Sigma_1^H\}$ *to connected voxel region* $\mathbb{V}_1$. *The projections of the head model into camera views* 1 *and* 2 *are denoted* $\{\mathbf{c}_{11}, \Sigma_{11}^H\}$ *and* $\{\mathbf{c}_{21}, \Sigma_{21}^H\}$ *respectively.* $M_1(\mathbf{c}_{11}) = \mathbb{S}_{13}$ *identifies the connected skin region in camera view* 1 *which pixel location* $\mathbf{c}_{11}$ *occupies and* $M_2(\mathbf{c}_{21}) = \mathbb{S}_{23}$ *identifies the connected skin region in camera view* 2 *which* $\mathbf{c}_{21}$ *occupies.*

Figure 6.4: *Illustrative example of fitting an ellipsoial head model to the detected* 3*D foreground.*

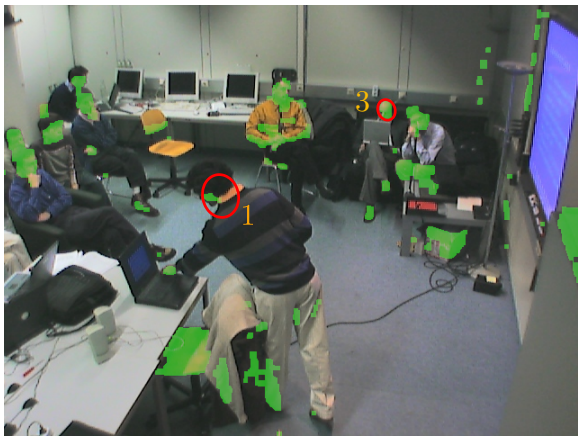


(a) *Camera 1: Skin colour mask with example detected heads.*

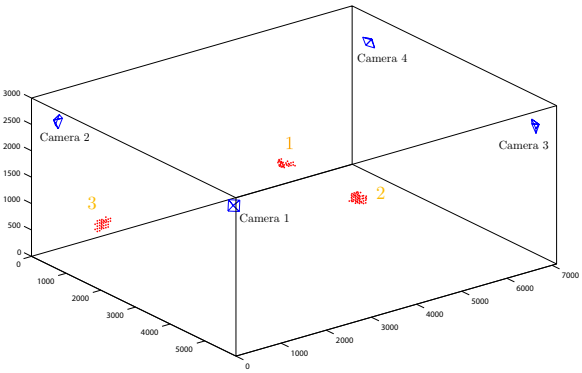


(b) *Camera 2: Skin colour mask with example detected heads.*

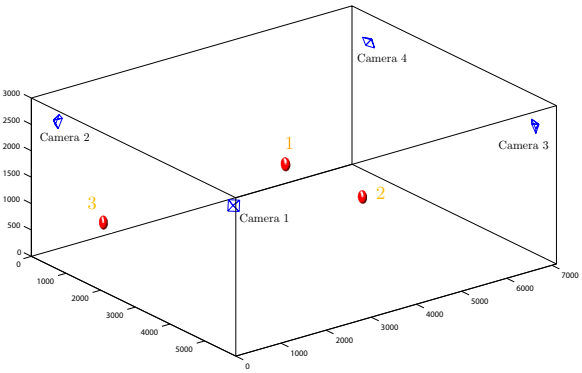
(c) *Camera 3: Skin colour mask with example detected heads.*



(d) *Camera 4: Skin colour mask with example detected heads.*



(e) *3D foreground region corresponding to the example detected heads.*



(f) *Ellpsoid head model fitted to each 3D foreground region.*

Figure 6.5: *Head detection example using detected skin colour regions in four views and an ellipsoidal head model. The heads labelled* 1, 2 *and* 3 *are visible in four views, three views and two views respectively. Note that for clarity in this figure, only three of the detected head are shown.*

The process of fitting the ellipsoidal head model to voxel regions is illustrated in figure 6.4. In fitting the head model, V-VAST employs the simplex algorithm.

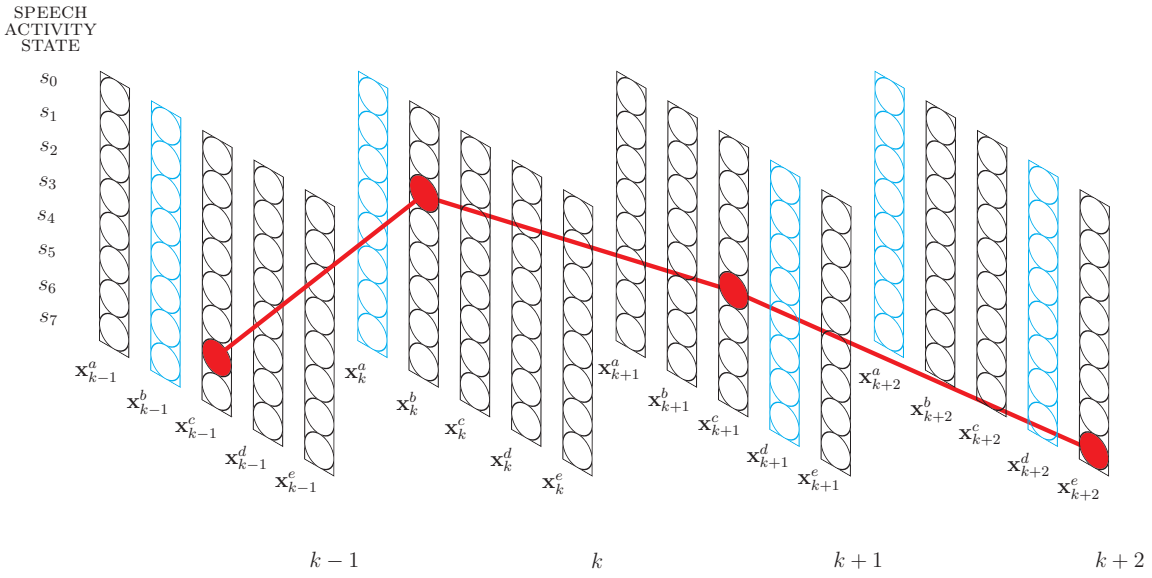
Some example results of the head detection process are shown in figure 6.5. In this example camera view one, two, three and four are shown in figure 6.5a, 6.5b, 6.5c and 6.5d respectively. In these views, three examples of detected heads within the CHIL lecture room are shown. The three detected heads labelled 1, 2 and 3 represent three different detection scenarios. Head 1 is visible in four views but heads 2 and 3 are only visible in three and two views respectively. The resulting $3D$ foreground corresponding to the detected heads is shown in figure 6.5e. The result of the ellipsoidal head model fitting to the detected $3D$ foreground is shown in figure 6.5f. The fitted ellipsoids to each detected head is projected back into the four views of figure 6.5a, 6.5b, 6.5c and 6.5d.

6.3 Joint MAP Estimation using the Viterbi Algorithm

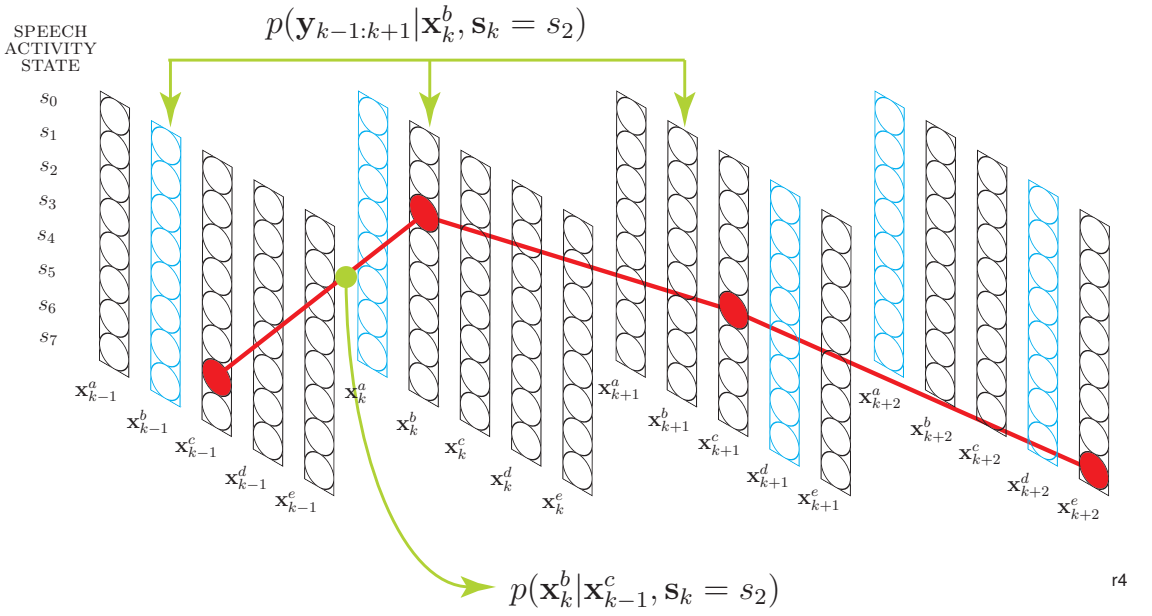
As described in section 6.1, determining the speaker activity path requires estimating the joint state sequence $[\mathbf{x}_{0:K}, \mathbf{s}_{0:K}]$. V-VAST applies the Viterbi algorithm to this estimation problem. This is possible since the position of an active speaker is constrained to a discrete set of possible speaker positions $\mathbf{x}_k^i$ through head detection using the video data. In addition to this, by its definition in this work, the speaker activity state $\mathbf{s}_k$ is also discrete.

If only the positional state sequence was to be estimated, the function of the Viterbi algorithm would be to navigate the optimal path through the trellis structure defined by the discrete positional states $\mathbf{x}_k^i$. In the joint estimation problem, where the speaker activity state $\mathbf{s}_k$ is also to be estimated, the trellis structure is three-dimensional. An illustration of the $3D$ trellis diagram for the joint tracking problem and an example path through the trellis is presented in figure 6.6a.

In this diagram, some states in the Viterbi trellis are illustrated in a light blue colour. These correspond to estimated head positions which have disappeared due to failed detection or occlusion. Since the likelihood function is defined over three time steps, if a detected head position disappears at time k, an estimate of the head position at k must still be made in order to evaluate the likelihood. This is achieved by propagating the estimated head position at the previous time $k-1$ forward to time k and evaluating the likelihood function based on this approximate head position. Therefore, if a head is occluded or head detection fails, the current head position estimate is taken as the closest head position estimated at the previous time step. An example of the likelihood function and prior in relation to the $3D$ Viterbi trellis is illustrated in figure 6.6b. The complete Viterbi algorithm for determining the optimal path through the $3D$ trellis in the joint tracking problem is presented in Algorithm 2.



(a) *Illustration of the 3D trellis structure navigated by the Viterbi algorithm. Also shown is an example path through the trellis corresponding to the state sequence,* $[\mathbf{x}^c_{k-1}, \mathbf{s}_{k-1} = s_6]$,$[\mathbf{x}^b_k, \mathbf{s}_k = s_2]$,$[\mathbf{x}^c_{k+1}, \mathbf{s}_k = s_4]$,$[\mathbf{x}^e_{k+2}, \mathbf{s}_{k+2} = s_7]$



(b) *Example of the prior probability density and likelihood function in relation to the 3D trellis. The light blue states correspond to cases where estimate head positions appear and disappear over time due to detection failure or occlusion. For example, the position* $\mathbf{x}^b$ *exists at time step* k *and* $k+1$ *but is not present at time step* $k-1$.

Figure 6.6: *Illustration of the joint trellis structure of the Viterbi tracking problem in V-VAST.*

The Viterbi Algorithm

Initialization: For $i = 1, .., N_0$

$$\delta_0(i) = \log(p(\mathbf{y}_0|\mathbf{x}_0^i)) + \log(p(\mathbf{x}_0^i)) \tag{6.32}$$

Recursion: For $2 \leq k \leq t$ and $1 \leq j \leq N_k$

$$\Phi_k(j,n) = \max_{i,m}[\delta_{k-1}(i,n) + \log(p(\mathbf{x}_k^j|\mathbf{x}_{k-1}^i, \mathbf{s}_{k-1} = s_m, \mathbf{s}_k = s_n)] \tag{6.33}$$

$$\Psi_k(j,n) = \operatorname*{argmax}_{i,m}[\delta_{k-1}(i,n) + \log(p(\mathbf{x}_k^j|\mathbf{x}_{k-1}^i, \mathbf{s}_{k-1} = s_m, \mathbf{s}_k = s_n)] \tag{6.34}$$

$$\delta_k(j,n) = \log(p(\mathbf{y}_{k-1:k+1}|\mathbf{x}_k^j, \mathbf{s}_k = s_n)) + \Phi_k(j,n) \tag{6.35}$$

$$\tag{6.36}$$

Termnation:

$$[j_k, n_k] = \operatorname*{argmax}_{j,n}[\delta_k(j,n)] \tag{6.37}$$

$$\mathbf{x}_k^{MAP} = \mathbf{x}_k^{j_k} \tag{6.38}$$

$$\mathbf{s}_k^{MAP} = s_{n_k} \tag{6.39}$$

Back-Tracking: For $k = t-1, .., 1$

$$[j_k, n_k] = \Psi_{k+1}(j_{k+1}, n_{k+1}) \tag{6.40}$$

$$\mathbf{x}_k^{MAP} = \mathbf{x}_k^{j_k} \tag{6.41}$$

$$\mathbf{s}_k^{MAP} = s_{n_k} \tag{6.42}$$

Algorithm 2: *Joint Viterbi Algorithm which returns the MAP estimate of both the speaker position* $\mathbf{x}_k$ *and speaker activity state* $\mathbf{s}_k$.

6.4 Visually Segmenting the Active Speaker

The second task which is examined in the evaluation of V-VAST is that of visually segmenting the *best* view of the active speaker from the available camera views. The aim in doing so, is to compose a composite video sequence consisting of a user defined main view of the lecture and an automatically inserted view of the active speaker.

6.4.1 Determining the *Best* Camera View

Determining the *best* camera view of the active speaker is based on a simple but effective measure of the amount of visible skin in the head view. The estimation of the active speaker's head position by the technique described in section 6.2.4 results in an estimate of ellipses defined

by $\{\mathbf{c}_{ij}, \Sigma_{ij}^{H}\}$ in each view where j denotes the index of the active speaker's head position and $i = 1, .., N_{cam}$. The ellipses define the region in each view where the active speaker's head is contained. An example of the estimated $2D$ ellipses for an active speaker head position in four views in shown in figure 6.7a to figure 6.7d.

Given the $2D$ ellipses in each view corresponding to the active speaker, the criteria for determining the *best* view examines the total area of detected skin within the area of the ellipse in relation to the total area of the ellipse. More formally this can be defined as follows. Let the set of pixels within the ellipse defined in the ith view be denoted by $\mathbf{A}_i$ and let the total number of pixels in this set be n_i. The measure of the visibility of the head for the ith view is then defined as,

$$B_i = \frac{1}{n_i} \sum_{\mathbf{p} \in \mathbf{A}_i} \mathbf{S}_i(\mathbf{p}) \tag{6.43}$$

where $\mathbf{S}_i(\mathbf{p})$ is the skin mask at pixel location $\mathbf{p}$ as defined in equation 2.57. In essence, what B_i measures is the ratio of the area of skin within the $2D$ ellipse of the head in the ith view to the total area of the ellipse. Given this measure of visibility the *best* view of the active speaker is defined as the view in which the speaker is most visible. This corresponds to the view,

$$i = \arg\max_{i} [B_i]. \tag{6.44}$$

An example of the effectiveness of this measure for determining the most suitable camera view is shown in figure 6.7e to figure 6.7l.

6.5 Evaluation of Tracking Accuracy

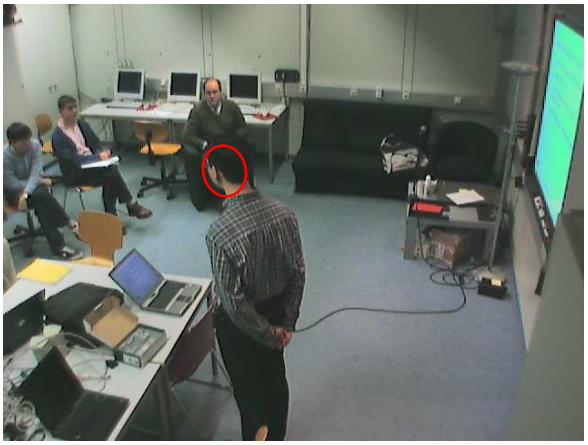
This section presents an evaluation of the accuracy of V-VAST for tracking the presenter in a lecture presentation. For this, the database of seminar recordings from the 2005 CHIL evaluation package as described in section 1.1 is used [171]. Even though the 2005 evaluation package contains both multi-channel audio and multi-channel video recordings it was not specifically designed for the evaluation of joint audio-visual based tracking algorithms [2]. Instead it was designed for the separate evaluation of tracking tasks such as the visual tracking of the presenter and also acoustic based person tracking. For the evaluation of acoustic based person tracking, the package contains the complete set of audio recordings for all seminars. For the evaluation of the visual tracking of the presenter however, only segments of the seminar recordings are provided. Therefore the package contains multi-channel audio recordings for which there is no

[2] The evaluation of joint-audio video based tracking algorithms was the focus of later CHIL evaluations in both 2006 and 2007. Evaluation workshops were held in conjunction with these evaluations complete with published proceedings. In addition to the seminar recordings contained in the 2005 evaluation package, additional recordings were included in the 2006 and 2007 packages of interactive meetings and presentations. At the time of this work however only the 2005 evaluation package was available through European Language Resource Association (ELRA).

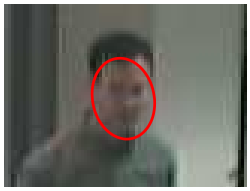
(a) *Active speaker located in camera 1*

(b) *Active speaker located in camera 2*

(c) *Active speaker located in camera 3*



(d) *Active speaker located in camera 4*



(e) *Camera view 1 zoom*

(f) *Camera view 2 zoom*

(g) *Camera view 3 zoom*

(h) *Camera view 4 zoom*

(i) *Camera view 1 with skin mask*

(j) *Camera view 2 with skin mask*

(k) *Camera view 3 with skin mask*

(l) *Camera view 4 with skin mask*

Figure 6.7: *Example of the criteria for determining the best view of the active speaker from the available camera views. In this example camera view 1 is determined as the* best *camera view using equation 6.44*

corresponding video data. As a result, since V-VAST requires both audio and video data, it can only be evaluated on the visual tracking task. The algorithm is firstly evaluated on the visual tracking task outlined in the CHIL 2005 evaluation and secondly in relation to general active speaker tracking and segmenting the *best* view of the speaker.

The visual tracking task of the CHIL 2005 evaluation is that of tracking the position of a presenter giving a presentation in front of an audience in the CHIL room. The ground truth of the speaker's position which is defined as the centroid location of the head in $3D$ is provided for evaluation purposes. This ground truth data was extracted by human annotators by locating the head centroid in each view in which the face was visible in every 10th frame of each sequence. A face is defined as visible only if the nose of the face can be seen. A script file is provided with the CHIL 2005 Evaluation package which evaluates tracking results against the ground truth for different metrics. The proposed metrics for consideration in the evaluation are,

- **The $2D$ global mean error:** Mean of the Euclidean distance in millimetres between the estimated position of the head centre and the ground truth. The mean is only determined over frames which are positively labelled with a ground truth position. In the evaluation, only $2D$ distances are examined and the height of the head from the ground is not included in the metric.

- **Percentage of Misses:** Percentage of frames where no estimate of the head position is given by the algorithm even though a positively labelled frame with ground truth exists.

- **Percentage of False Positives:** Percentage of frames where a position estimate is given by the algorithm even though the frame is not positively labelled and has no ground truth.

In the test sequences provided for the presenter tracking task, it is predominantly the case that the presenter is the only active speaker. Tracking the speaker activity path using V-VAST therefore is equivalent to that of tracking the presenter. Some of the test sequences however contain some periods where the presenter is not the active speaker such as during periods where an audience member asks a question. Despite this V-VAST, will be examined on the task of tracking the presenter only. Later analysis will examine the problem of tracking both the presenter and questions from audience members.

Shown in table 6.2 is an overview of the performance of V-VAST evaluated on the visual tracking task of the 2005 CHIL evaluation. In the implementation of the tracking algorithm for the single speaker case, the Gaussian motion prior of equation 6.18 is used with $\Sigma = \mathbf{I}[\sigma_x^2, \sigma_y^2, \sigma_z^2]^T$ where $\sigma_x = 0.5m$, $\sigma_y = 0.5m$, $\sigma_z = 0.08m$ and $\mathbf{I}$ is a 3×3 identity matrix. Also, in all of the presented results, TDEs were obtained using the GCC-PHAT algorithm. Only the 4 inverted T-shaped microphone arrays of the CHIL room were used. The time-delays were obtained from all inner-array microphone pair combinations. Pair combinations between different arrays were not used. Therefore in total the 4 inverted T-shape arrays yielded 24 TDEs.

Since no proceedings were published or are available for the CHIL 2005 evaluation[3], direct comparison of the proposed algorithm against different tracking approaches on the same evaluation task can not be made. The results however are presented so as to be consistent with the guidelines of the evaluation package and so as to be comparable with other tracking algorithms evaluated on the same dataset.

Table 6.2 provides a good insight into the performance of the proposed tracking algorithm. From this table, it can be seen that estimated head positions can be highly accurate with a minimum error no greater than $0.02m$ in any seminar sequence and in the majority of sequences it is less than $0.01m$. This level of performance is seen to be consistent over the total duration of the track sequences with the results showing an overall mean error of $0.20m$ on approximately $53mins$ of recorded seminar footage. To put the overall mean error into context it can be compared to the head model of size $0.194m \times 0.145m \times 0.241m$ used in detecting head locations within the room. The tracking error therefore corresponds on average to slightly over one head width against the true position of the head.

The metric of the percentage of false positives as quoted in the evaluation can be misleading. The ground truth data only positively labels a frame if the nose position of the head is visible in two or more camera views. Therefore, if an estimate of a head position is given at negatively labelled frames it is regarded as a false positive. This would be reasonable if V-VAST required the nose to be visible in two or more views to provide a head position estimate, but this is not the case. The quoted false positive rate in the evaluation therefore has little meaning and can only be interpreted as the percentage of frames where V-VAST provided an estimate of the head position even though a frontal view of the presenter's face was not visible in two or more views.

The $2D$ global mean cannot give any indication as to the accuracy of the tracking algorithm in estimating the $3D$ position of the presenter within the room. To address this, an additional metric of the $3D$ global mean is presented in table 6.3. This incorporates the estimated height of the head from the ground in the mean error analysis. From this table it can be seen that including the estimated height of the head from the ground into the error analysis increases the overall mean error from $0.20m$ to $0.23m$. What this indicates is that the largest component of the error occurs in the $2D$ $x-y$ plane. This seems reasonable since it is in this plane where the most significant motion is observed. There is little variation in the height of the presenter's head from the ground over the duration of the seminar. This is because the presenter remains standing throughout the seminar and the only variation in head height from the ground occurs at a few points in the sequence when the presenter is stooping to operate a laptop or placing/removing objects from a table.

The worst performance of the tracking algorithm occurs for sequence **seminar_2004-11-12_segment2**. This is due to the occurrence of a number of significant active speaker switches where the presenter is not talking but some other person in the room is. As previously stated V-VAST is designed to track the speaker activity path and use of the motion prior of equation

[3]Personal communication with Keni Bernardin, Universität Karlsruhe, Germany.

Presenter Tracking Results using the 2D Global Mean Metric						
Seminar_2004-11-11_A	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	5m 07s	0.02	0.51	0.11	0.00	0.00
Segment2	5m 04s	0.00	0.53	0.13	0.00	0.00
Seminar_2004-11-11_B	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	6m 49s	0.01	1.70	0.29	0.00	0.00
Segment2	5m 58s	0.01	2.39	0.35	0.00	1.30
Seminar_2004-11-11_C	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	5m 00s	0.01	1.02	0.17	0.00	23.06
Segment2	5m 03s	0.00	1.97	0.15	0.00	6.97
Seminar_2004-11-12_A	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	3m 49s	0.00	0.60	0.06	0.00	1.16
Segment2	5m 40s	0.00	0.14	0.04	0.00	0.00
Seminar_2004-11-12_B	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	4m 44s	0.00	0.17	0.04	0.00	0.70
Segment2	5m 46s	0.00	3.55	0.51	0.00	4.43
Overall	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
	53m 00s	0.00	3.55	0.20	0.00	3.63

Table 6.2: *Results corresponding to the visual person tracking task of the CHIL 2005 evaluation campaign using the* 2D *global mean metric as proposed in the evaluation package.*

6.18 therefore does not suppress active speaker switches. This sequence is examined later in tracking speaker activity in the presence of speaker switches.

A more detailed presentation on the accuracy of the proposed tracking algorithm is given in appendix C where tracking results in terms of the x, y and z axes are plotted for each of the 10 seminar segments used in the evaluation. In analysing this data it is clear that a consistent negative standard error is present in the estimated z position in relation to the true speaker position. This is attributed to a slight bias in the fitting of the head model when skin is detected at exposed neck regions as well as face regions. Detected skin at the neck position below the head results in the head model being fitted to this region as well as the head.

6.6 Visual Segmentation Results

In this section the complete V-VAST algorithm with its *best* view selection criteria as described in section 6.4.1 is applied to visually segment the current active speaker over the duration of the lecture presentation. The presented results examine two different classes of presentations. The first examines the single speaker tracking case such as the task of tracking the presenter as described in section 6.5. The second class of presentation examined is the more interactive scenario with switches in speaker activity between lecture participants. The video sequences created by V-VAST which are described in the following sections can be found on the accompanying *DVD*.

Presenter Tracking Results using the $3D$ Global Mean Metric						
Seminar_2004-11-11_A	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	5m 07s	0.06	0.53	0.13	0.00	0.00
Segment2	5m 04s	0.06	0.54	0.17	0.00	0.00
Seminar_2004-11-11_B	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	6m 49s	0.02	1.71	0.36	0.00	0.00
Segment2	5m 58s	0.04	2.40	0.42	0.00	1.30
Seminar_2004-11-11_C	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	5m 00s	0.01	1.02	0.20	0.00	23.06
Segment2	5m 03s	0.00	1.98	0.18	0.00	6.97
Seminar_2004-11-12_A	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	3m 49s	0.00	0.61	0.07	0.00	1.16
Segment2	5m 40s	0.01	0.14	0.05	0.00	0.00
Seminar_2004-11-12_B	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
Segment1	4m 44s	0.01	0.17	0.05	0.00	0.70
Segment2	5m 46s	0.00	3.57	0.54	0.00	4.43
Overall	duration	min (m)	max (m)	avg (m)	% misses %	% false pos. %
	53m 00s	0.00	3.57	0.23	0.00	3.63

Table 6.3: *Results corresponding to the visual person tracking task of the CHIL 2005 evaluation campaign where the average error examined is the $3D$ global mean.*

6.6.1 Single Speaker Case

In the single active speaker case, the visual segmentation results examined correspond to the sequences ***seminar_2004-11-12_A_segment2***, ***seminar_2004-11-12_B_segment1*** and ***seminar_2004-11-11_C_segment1***. The extracted *best* view of the active speaker over the duration of each of these sequences is shown in figure 6.8, figure 6.9 and figure 6.10 respectively. The results presented in this section relate directly to the tracking accuracy evaluation as presented in table 6.2 and table 6.3. In the figures both the tracking results in x, y and z coordinates together with the ground truth and the extracted *best* view of the active speaker are shown. The regions highlighted in green in these figures correspond to points in the sequence where the presenter's head is not visible in two or more views as is given in the ground truth of the CHIL database described in section 6.5.

Figure 6.8 shows the results for sequence ***seminar_2004-11-12_A_segment2***. In this particular sequence over the duration of the seminar, the presenter undergoes little motion and remains within an area of approximately $1m^2$. This does not represent a challenging tracking scenario and as expected the tracking is highly accurate with table 6.3 showing a mean $3D$ positional error of $0.05m$. Although this sequence does not represent a difficult tracking problem it does enable the performance of the *best* view selection to be examined. The criteria for *best* view selection as defined in 6.4.1 is reliant on an accurate estimate of the head position. Therefore, under known accurate tracking, the performance of the *best* view selection criteria

can be evaluated in isolation to that of the tracking performance.

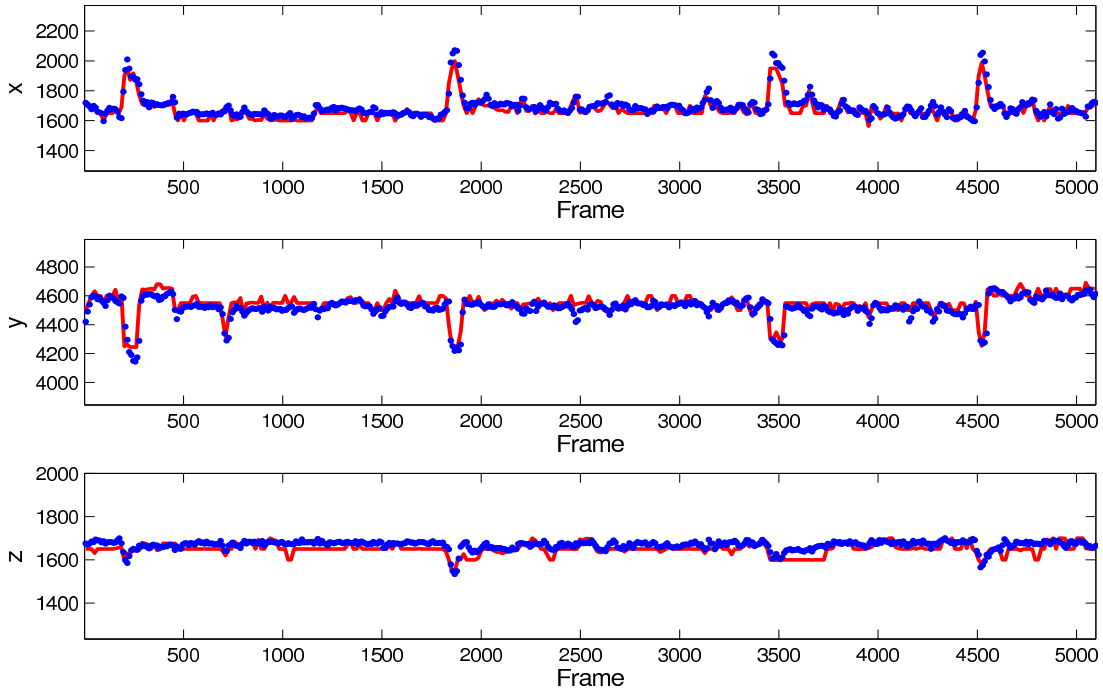
The segmented *best* view of the speaker over the duration of the seminar is shown in figure 6.8b. The choice of the *best* camera view for viewing the active speaker is largely subjective. There are few objective measures on which to determine the performance of the proposed view selection criteria. Two simple conditions which can be referred to in relation to the *best* view of a speaker are firstly, whether the face of the speaker is visible and secondly, whether the head is framed in the centre of the segmented view. In this regard, the results of figure 6.8b show that the proposed *best* view selection criteria is effective in choosing a suitable view of the active speaker. Of the presented sample frames, due to the high tracking accuracy, there are no poorly framed head views. In most cases, the view of the head is a frontal face view although some occurrences of profile face views can be seen in frame 1540, 1625 and 3410. Although these particular frames are not visually poor as a head view, they do show the limitations of a purely skin based *best* view selection criteria. Using skin only cannot guarantee a frontal face view.

The results for the sequence ***seminar_2004-11-12_B_segment1*** shown in figure 6.9 represent a more difficult tracking scenario. In this sequence the observed motion is relatively more complex and the speaker moves within an area of approximately $9m^2$. Even with the increase in the range of the observed motion, the algorithm maintains a tracking accuracy of $0.05m$ as seen in table 6.3.

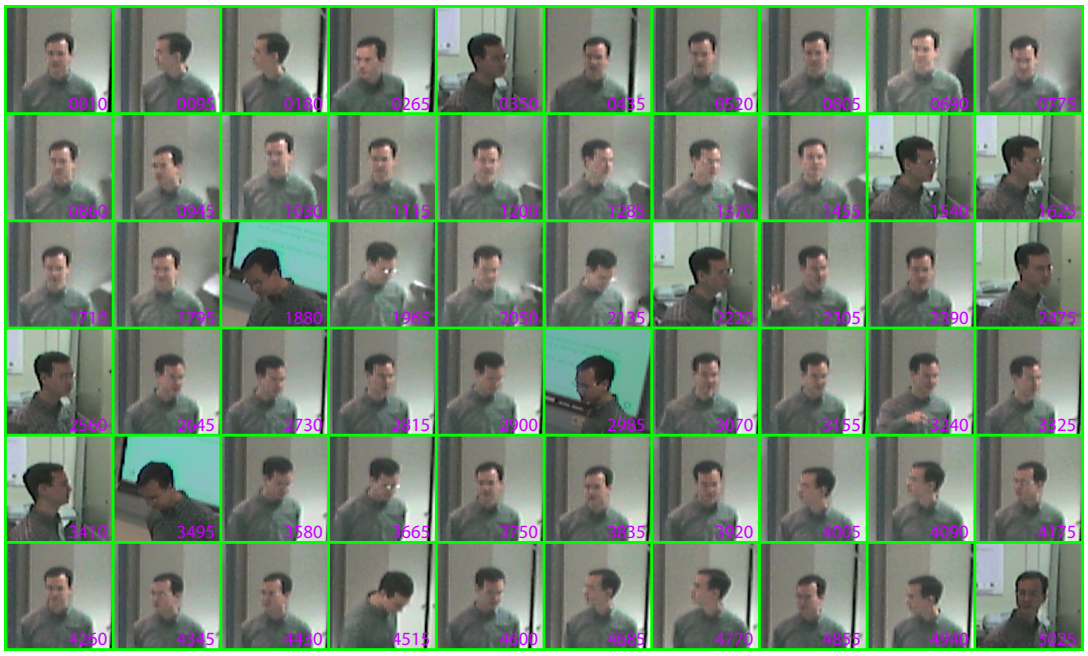
In this example the benefits of camera view switching can be seen clearly in frames 0791; 0933; 1643; 2353; 2566; 3276; 3489; 3560 and 3844. At these particular frames the presenter turns from the audience towards the projector screen. The *best* view determined before these points is from camera 1 showing a frontal face view. When the presenter turns towards the projector screen he is facing camera 2. In this case, the *best* view selection criteria appropriately determines the frontal face view now in camera 2 as the being *best* view.

In the majority of the presented frames, the *best* view selection criteria produces favourable results in ensuring a framed and frontal view of the speaker's face. Again however, as in the previous example there is some evidence as to the limitations of using the visibility of skin in the head region as a *best* view selection criteria. This is evident in frames 0578, 2069 and 3347.

The sequence examined in figure 6.10 presents a complex tracking scenario where the presenter moves over an area of $12m^2$. In this particular sequence, there is also an interesting occurrence of head occlusion where the presenter's head is only visible in one view. This occurs at the point in the frame sequence illustrated by the light blue coloured bar at the top of the plot in figure 6.10a. Even though the head over this period can not be located due to the occlusion, V-VAST is still able to provide a reasonably accurate estimate of the head position and maintain track. This is achieved through the occlusion handling process employed in the Viterbi algorithm as described in section 6.3. In essence, the handling of occlusions is managed by maintaining the last visible head location in the set of head candidates at the point at which the occlusion occurs. For instance, if a candidate head is not located close to a previous head location at the current time instance, then the previous location is set to be current head posi-

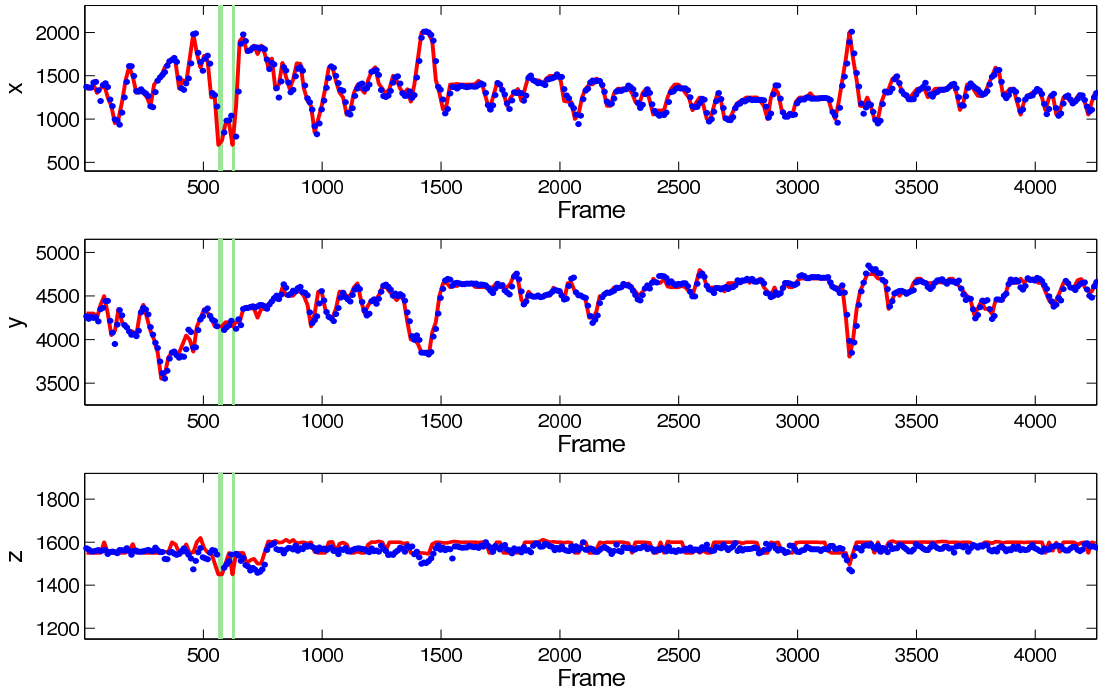


(a) *Tracking results for* ***seminar_2004-11-12_A_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*



(b) *Segmented active speaker for uniform sample of* 60 *frames over the sequence. The progression of the frames is shown from top to bottom and from left to right.*

Figure 6.8: *Visually segmenting the active speaker in the* ***seminar_2004-11-12_A_segment2*** *sequence of the CHIL database.*



(a) *Tracking results for* ***seminar_2004-11-12_B_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*

(b) *Segmented active speaker for uniform sample of* 60 *frames over the sequence. The progression of the frames is shown from top to bottom and from left to right.*

Figure 6.9: *Visually segmenting the active speaker in the* ***seminar_2004-11-12_B_segment1*** *sequence.*

tion. An example frame where the presenter is occluded is shown in figure 6.11. In this figure the presenter is only visible in camera views 3 and 4, but the head is only visible in camera view 4. The ellipses in both camera views in this figure correspond to the previously detected head location as introduced by the occlusion handling process.

Although the handling of the occlusion results in an estimate of the head position which is close to the true head position, it is not accurate enough to determine a *best* view of the presenter's head. Clearly, from figure 6.11 the *best* view is that of camera 4 since the head is not visible in camera 3. In this case however, the measure of the visibility of the head as defined by equation 6.44 for each view is, $\mathbf{B}_3 = \mathbf{B}_4 = 0$. Therefore, both views are considered equal. By default, the camera view with the lowest camera index (i.e. camera 3), is chosen by V-VAST as the *best* view. During such periods of occlusion, V-VAST fails to return a *best* view of the presenter. This can be seen in figure 6.10b where frames over periods of occlusion are highlighted by a light blue border.

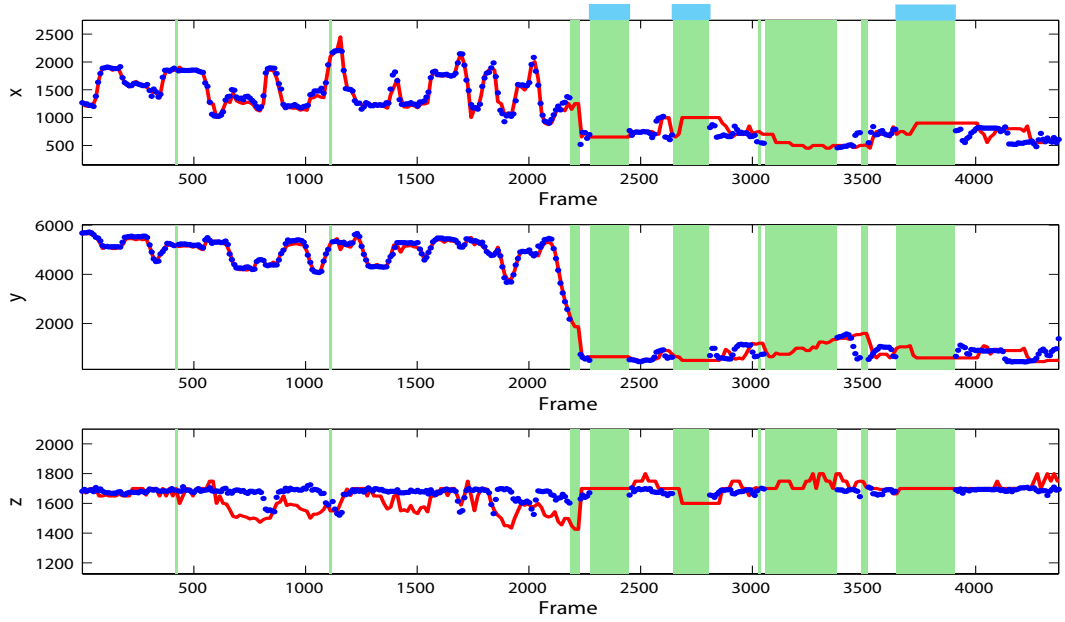
6.6.2 Speaker Switching Case

In this section tracking the path of speaker activity where speaker switches occur is examined. The presented results relate to seminar sequences **seminar_2004-11-11_A_segment4**, **seminar_2004-11-12_B_segment3** and **seminar_2004-11-12_B_segment2**. The seminar sequence **seminar_2004-11-12_B_segment2** was previously examined in the presenter tracking evaluation of section 6.5 and resulted in the worst performance. This was due to the presence of numerous speaker switches where the presenter was not speaking. The sequence is therefore re-examined in this section to track the speaker activity path.

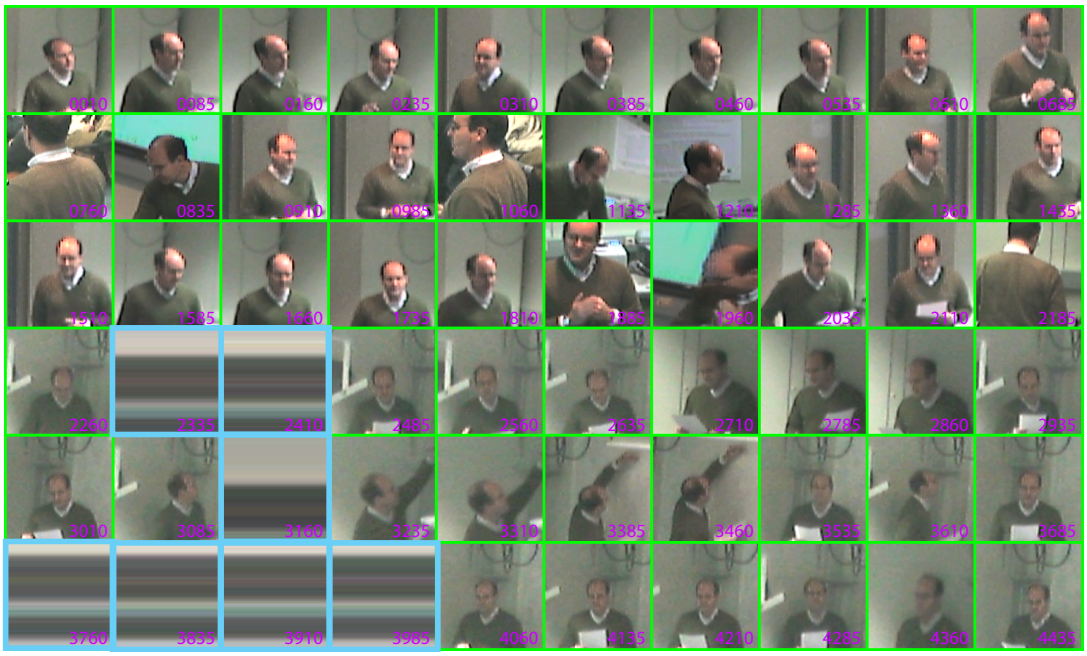
The results of tracking the speaker activity path and extracting the *best* view of the speaker for each of the seminars, are presented in figure 6.12, figure 6.13 and figure 6.14. Shown in each of the figures are the tracking results in x, y and z coordinates together with the ground truth and the extracted *best* view of the active speaker. In the cases presented in this section, the ground truth has been re-evaluated to correspond to the position of the active speaker and not only the presenter. In the implementation of V-VAST in this evaluation, the motion prior of equation 6.19 is used with $\lambda = 1.2$ and $\Sigma = \mathbf{I}[\sigma_x^2, \sigma_y^2, \sigma_z^2]^T$ where $\sigma_x = 0.5m$,$\sigma_y = 0.5m$,$\sigma_z = 0.08m$ and $\mathbf{I}$ is a 3×3 identity matrix.

Figure 6.12 and figure 6.13 show the results for V-VAST in tracking the speaker activity path for the sequences **seminar_2004-11-11_A_segment4** and **seminar_2004-11-12_B_segment3**. Both of these sequences have periods where the presenter takes a question from an audience member. These periods are illustrated by orange coloured bars at the top of the plots in figure 6.12a and 6.13a. Over these periods in the sequences, there are many quick conversational exchanges between the presenter and the audience member.

V-VAST however, is insensitive to temporally brief speech utterances and tends to follow a path of temporally smooth dominant speaker activity. This can be seen to be the case in the



(a) *Tracking results for* ***seminar_2004-11-11_C_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*



(b) *Segmented active speaker for uniform sample of* 60 *frames over the sequence. The progression of the frames is shown from top to bottom and from left to right.*

Figure 6.10: *Visually segmenting the active speaker in the* ***seminar_2004-11-11_C_segment1*** *sequence of the CHIL database.*

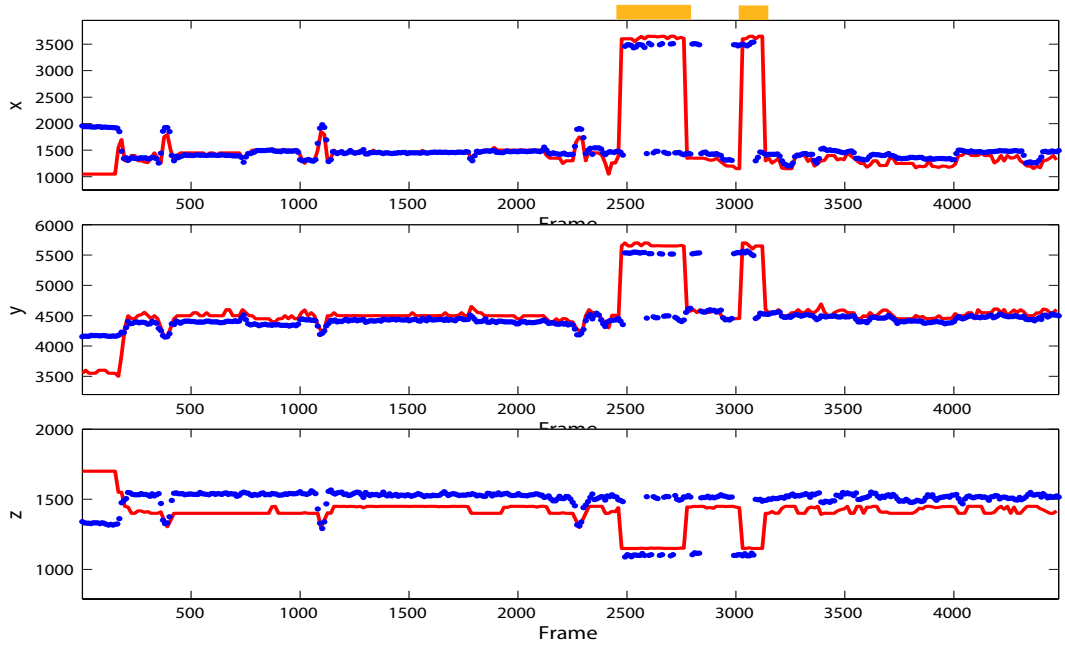
(a) *Camera view* 3 *with skin mask.* (b) *Camera view* 4 *with skin mask.*

Figure 6.11: *Frame* 3825 *from the sequence* ***seminar_2004-11-11_C_segment1****. The head of the active speaker is only visible in one view and therefore no head position is detected. The estimated head position is the last visible location of the head corresponding to the candidate head position introduced by occlusion handling in the Viterbi algorithm. In this case, the measure of visibility of the head for each view as defined in equation 6.43 is* $B_3 = B_4 = 0$*. With the visibility of the head in each view equal, the* best *view returned by default is that corresponding to the camera view with the lowest camera index which is camera view* 3*. Over this period in the sequence therefore, V-VAST can not determine the best view of the presenter as can be seen in figure 6.10b at the frames highlighted with a light blue border.*

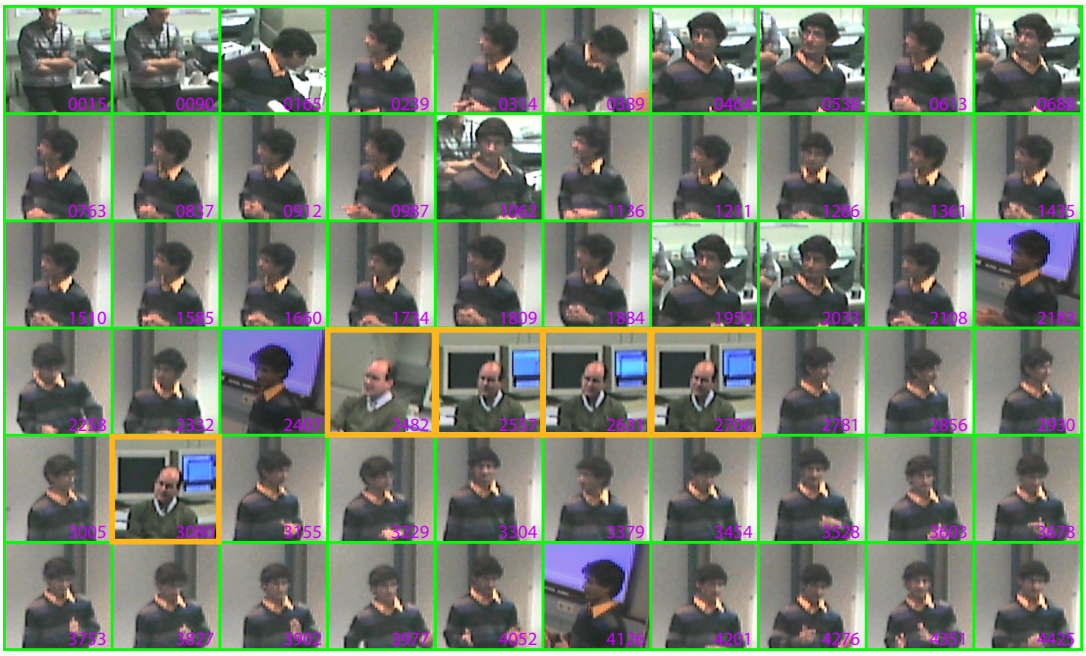
estimated speaker activity path in both of the sequences. It is particularly evident at frame 4100 in figure 6.13a where the audience member is actively speaking for less than three time instances. V-VAST in this case smoothes over this brief period of speaker activity. This reflects the effect of the prior as defined in equation 6.20 which enforces temporal smoothness on the speaker activity path.

The consequence is that the number of switches between the audience member and presenter are fewer which results in a more visually pleasing video sequence of the extracted active speaker. The extracted sequence of frames is more visually pleasing in the sense that visual switching does not occur abruptly at each speech utterance or for very short time instances. This can be seen in the extracted view of the active speaker shown in figure 6.12b and figure 6.13b where the active speaker switches are indicated for each frame by an orange coloured border. The accuracy of the head position estimate can also be seen in these examples where the head in each view is accurately framed in the centre.

The sequence of ***seminar_2004-11-12_B_segment2*** represents a more complex active speaker tracking problem than the previous examples. In this sequence there are conversational exchanges of short and long duration between the presenter and an audience member including at one point, the motion of two speakers. The periods over which the algorithm detected the audience member as the active speaker are shown by an orange coloured bar over the plot in figure 6.14a. Shown in figure 6.14b is the extracted view of the active speaker where the

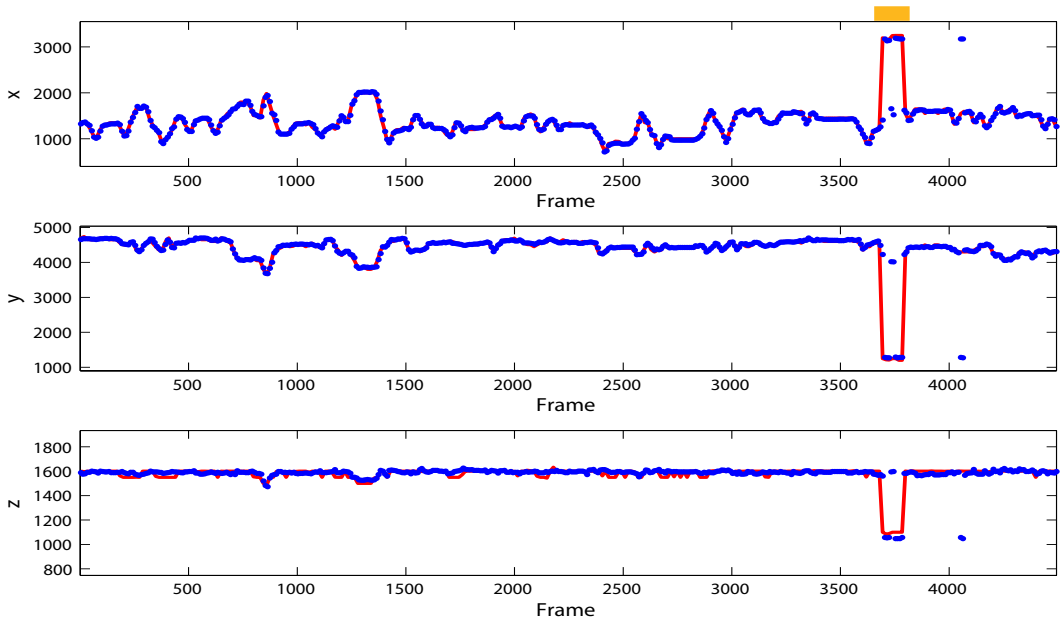


(a) *Tracking results with switches for* ***seminar_2004-11-11_A_segment4*** *in x, y and z coordinates (red) against ground truth (blue).*

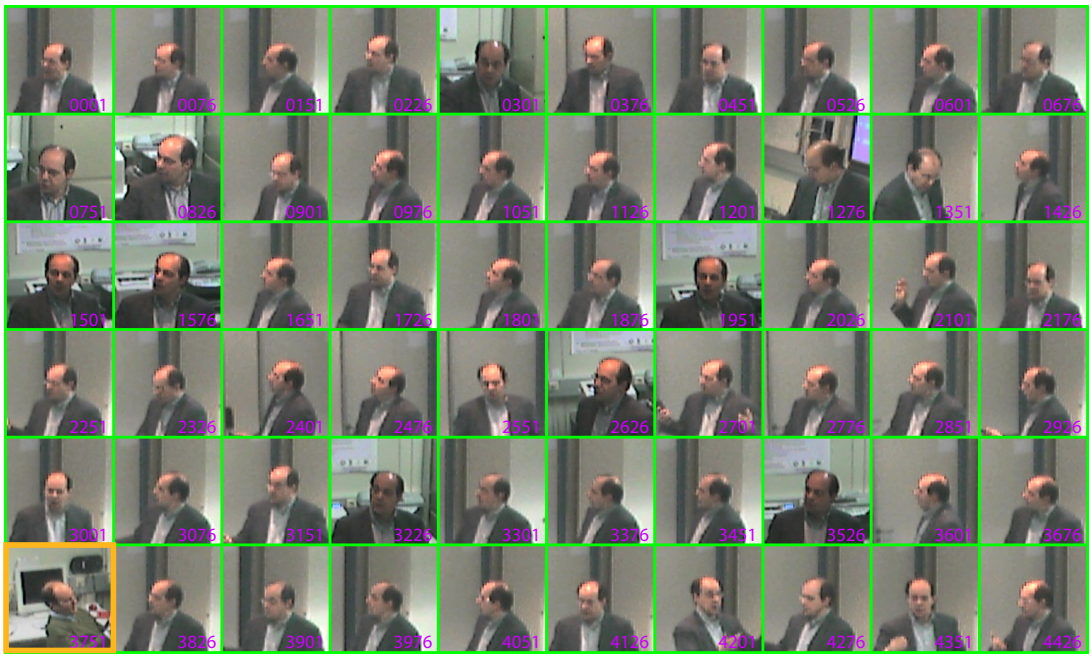


(b) *Segmented active speaker for uniform sample of* 60 *frames over the sequence. The progression of the frames is shown from top to bottom and from left to right. Speaker switches are shown in yellow.*

Figure 6.12: *Visually segmenting the active speaker in the* ***seminar_2004-11-11_A_segment4*** *sequence of the CHIL database.*



(a) *Tracking results with switches for* ***seminar_2004-11-12_B_segment3*** *in x, y and z coordinates (red) against ground truth (blue).*



(b) *Segmented active speaker for uniform sample of* 60 *frames over the sequence. The progression of the frames is shown from top to bottom and from left to right. Speaker switches are shown in yellow.*

Figure 6.13: *Visually segmenting the active speaker in the* ***seminar_2004-11-12_B_segment3*** *sequence of the CHIL database.*

audience member, when active, is indicated in the frames by an orange coloured border. Again in this example the smoothness on the estimated speaker activity path is evident at frame 3500 where a brief period of speech activity is not estimated in the speaker activity path.

The majority of the extracted views of the active speaker's head are seen to be correctly framed in figure 6.14b. This indicates the accuracy of the head detection process which is consistent throughout the sequence. This is notably the case in frames 3844 to 3930 where the audience member stands up and moves towards the projector screen while speaking.

There is a period from frame 4017 to frame 4534 where head detection fails and the algorithm returns a poor estimate of the *best* head view. This period of the sequence is illustrated by the light blue coloured bar at the top of the plot in figure 6.14a. During this time the presenter is actively speaking however the head detection process fails due to a failure to detect skin in the head region. This occurs as a result of the presenter moving into the path of the projector. Due to the colour of the slide, the skin colour is distorted. A frame from this period in the sequence in shown in figure 6.15.

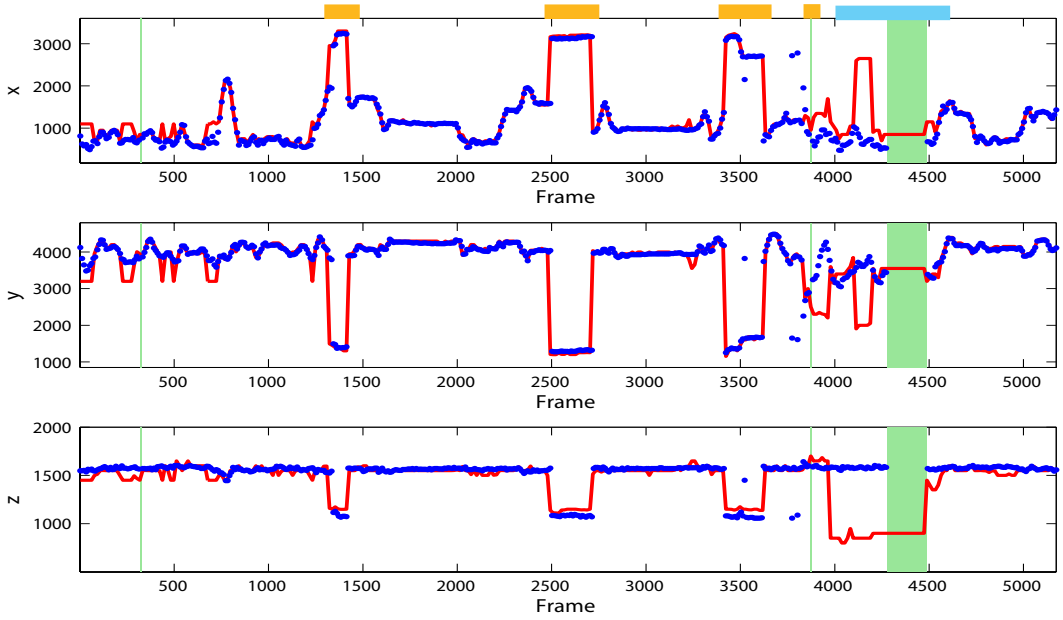
In normal circumstances the algorithm would attempt to correct this occlusion using the occlusion handling scheme as described in figure 6.15. In this instance however the presenter moves into a large region of the room where skin detection fails due to colour distortion from the projector. The last visible head location therefore is at the boundary of this region and far away from the presenter's current position. In this case, an erroneous estimate of the head is made which corresponds to the presenter's hand position. The extracted *best* view over this period of anomalous head estimates in the sequence is shown in figure 6.8b with a light blue border.

6.7 Final Comments

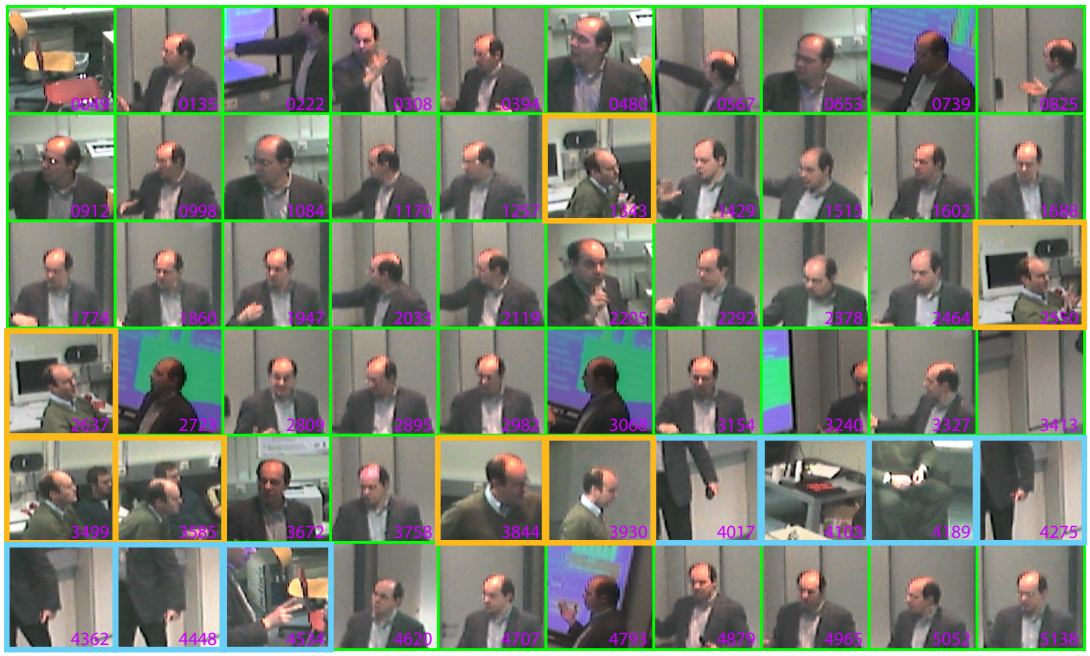
This chapter presented V-VAST, an algorithm for composing a composite video sequence of the current active speaker from multiple camera recordings of a lecture. The algorithm relies on both multi-camera video data to determine likely speaker positions and multi-channel audio to monitor speaker activity. Using this information V-VAST estimates the speaker activity path between speakers and segments the *best* view of the speaker from the available cameras.

V-VAST is focused towards extracting a temporally smooth speaker activity path over the duration of the lecture recording. Therefore, the estimated path does not strictly adhere to speaker activity but rather to a smoother path which is more suited to its visual presentation. By smoothing the speaker activity path, the algorithm is also robust to the occurrence of noise. It is not only this aspect which contributes to the robustness of the algorithm, but also the nature in which active speaker positions are constrained to detected head locations. Reliable head detection therefore is an important aspect of the algorithm.

The process of head detection employs a voxelization of the space of the lecture room from which a $3D$ foreground is extracted. This $3D$ foreground is extracted using skin colour masks obtained in each available camera view. In detecting head locations, an ellipsoidal model of the



(a) *Tracking results with switches for* ***seminar_2004-11-12_B_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*



(b) *Segmented active speaker for uniform sample of* 60 *frames over the sequence. The progression of the frames is shown from top to bottom and from left to right. Speaker switches are shown in yellow.*

Figure 6.14: *Visually segmenting the active speaker in the* ***seminar_2004-11-12_B_segment2*** *sequence of the CHIL database.*

(a) *Camera view* 1 *with skin mask.* (b) *Camera view* 3 *with skin mask.*

Figure 6.15: *Frame* 4425 *from the sequence* ***seminar_2004-11-12_B_segment2****. The presenter's head is in the path of the projector. The colour of the projected slide distorts the skin colour and no skin is detected in the head region. Therefore no head is detected. An erroneous estimate for the head is given at the position of the presenter's hand.*

head is assumed which is fitted to connected voxel regions and to their corresponding connected skin regions in each view. In order to improve the reliability of the head detection process, the algorithm also examines multiple $3D$ foregrounds from every combination of two or more camera views. In this way, the detection of head positions does not require that skin regions of the head are visible in every view. Although this greatly increases the computational demands of the algorithm it was found to increase the robustness of head detection to anomalies and inaccuracies in the skin colour masks. This process of head detection was found to locate all heads visible in at least two views in the majority of the examined cases. However, only using the visual cue of skin colour in locating heads introduced a lot of false positives in the detection process. Typically these occurred at other visible skin regions such as hand and arm locations. These falsely detected head locations were found to only introduce anomalies in the speaker activity path when they corresponded to noise source locations. Eradicating falsely detected head locations using additional cues such as face detection could possibly improve the reliability of the algorithm. This would also reduce the number of candidate speaker locations and therefore reduce the algorithm's computational complexity.

The accuracy by which heads can be estimated is also an important aspect of the algorithm since it determines how well the active speaker is framed in the segmented view. The use of a strong ellipsoidal model for the head and also the fitting of the model in both the $3D$ space of the room and $2D$ space of the camera views was shown to resolve $3D$ head locations to a high degree of accuracy. In the evaluation of the algorithm positively detected head locations were found to be accurate to less than $0.06m$ at best. This level of accuracy was found to be sufficient to accurately segment a well framed view of the speaker.

The process of detecting head locations relies heavily on skin colour detection to obtain a

skin mask from each view. This is achieved through a deterministic approach to skin colour modelling. The determination of the *best* view of the active speaker also uses the extracted skin colour masks in each view. The criteria for selecting the *best* view aims to determine the view of the active speaker in which the face is most visible. The underlying assumption in the approach is that the face is the region of head with the most visible skin. Of course, in some cases this assumption is not true and in these situations the proposed *best* view selection criteria was seen to fail. It was also seen however, that this simple criteria is effective in producing the view in which the face is most visible. The incorporation of face detection in addition to the proposed *best* view selection criteria would clearly improve the algorithm's performance.

7

Discussion & Conclusion

This thesis examined the problem of applying joint audio and video based techniques for localising and tracking an active speaker in a lecture room environment. The main focus in the work was towards the evaluation of the accuracy of both audio and video techniques for localisation, the factors that limit their performance and how they can be best combined in their current state of development to the problem of tracking an active speaker.

Chapter 2 focused on the localisation problem through the extraction of audio and video based features. In the audio domain, both TDE and DOA based localisation techniques were introduced. These were examined in relation to their suitability in the lecture room environment. The nature of lecture room acoustics was discussed and how it affects localisation was quantified. In particular, measures for characterising the level of reverberation in a room were established and it was described how reverberation affects signal coherence enforcing a lower limit on the achievable localisation accuracy.

In the video domain features for localisation such as faces and foreground detection through background modelling were introduced and the challenges facing their successful application in a multicamera lecture room environment were analysed. Spatially varying illumination was identified as a significant problem and a new model for skin colour was presented which models for the nonlinear dependence of skin colour on luminance. The suitability of the new skin colour model in detecting skin regions under low illumination was demonstrated.

Within a Bayesian tracking framework, audio-based active speaker tracking techniques were explored in chapter 3 and it was shown how these techniques can be extended to include video based observations. The assumption of independent audio and video measurements was iden-

tified as the key component facilitating the combination of both modalities in existing joint audio-visual tracking systems. A review of existing systems based on KFs and PFs was presented. The occurrence of filter divergence due to poor motion modelling was highlighted as a persistent problem which can significantly affect tracking performance not helping in the task of fusing audio and video measurements.

Chapter 4 was dedicated to the evaluation of a multi-camera and multi-microphone joint audio-video based source localisation system in a typical lecture room environment. Motivated by the lack of clarity in existing literature as to the improvement in accuracy through joint audio-video based techniques, this chapter examined localising a moving audio-visual source in 3D using audio only, video only and different audio-video fusion strategies. Fusion was examined as applied in the audio domain, the video domain and the positional domain with the latter found to be most accurate for the evaluated system. Within this analysis, the ML localisation problem was evaluated to remove any motion modelling errors from the analysis. It was found that little accuracy beyond the accuracy obtained through video alone could be achieved through the fusion of both modalities. It was concluded therefore that audio contributed little to the accuracy of a video based location estimate. Existing techniques for the joint calibration of both audio and video tracking spaces, together with the lack of appropriate techniques for measuring the uncertainty associated with TDEs, contributed significantly to this observation. The accuracy of audio localisation in comparison to that of video was found to be the limiting factor in achieving improved accuracy through fusion.

The problem of optimising the positions of microphone arrays within a room so as to minimise localisation uncertainty was analysed in chapter 5. This followed the direction of chapter 4 that audio-based localisation accuracy must be improved in order to justify its fusion with video to improve localisation accuracy. This analysis brought together existing bounds theory on time-delay estimation performance in a reverberant environment with that of the framework for uncertainty estimation developed in chapter 4. The employed bounds better reflected TDE-based localisation in a realistic reverberant room in the optimal microphone placement problem in comparison to existing approaches in the literature. The directionality characteristics of both the speaker and microphones were also accounted for in optimising the microphone arrays. A theoretical evaluation of localisation accuracy was performed on the CHIL lecture room under this analysis. It was found that for the given microphone array positions in the lecture room, the expected performance of audio-based localisation would be poor. A simulated annealing algorithm was proposed for automatically optimising the microphone array positions within the room to minimise localisation uncertainty. Under the employed bounds theory, the algorithm was seen to automatically determine positions significantly improving the overall localisation accuracy. This highlighted the sub-optimal microphone array positions in the CHIL lecture room.

Building on the analysis of previous chapters, an algorithm for tracking the current active speaker in the multi-camera and multi-microphone recording of a lecture in the CHIL room was

proposed. Given the analysis of joint audio-visual localisation accuracy in chapter 4 and the sub-optimal placement of microphone arrays in the CHIL room as established in chapter 5, it was concluded that audio-based localisation would contribute little to the expected accuracy of a video-based location estimate. Therefore, the statistical fusion of both audio and video location estimates was rejected in favour of a different strategy.

Voxel-based head detection using the new skin detection algorithm as proposed in chapter 2 was used to determine likely speaker positions. TDEs from the microphone arrays were then used to assign a state of speaker activity to detected speaker locations. The Viterbi algorithm was employed to determine a joint MAP estimate of the active speaker position and speaker activity state over the duration of the recorded lecture. A prior on the speaker activity states enabled positions of temporally significant speaker activity to be weighted highly in the returned estimate. This tracking algorithm was proposed within a system called V-VAST which used the estimated active speaker position over the duration of the recorded lecture to extract the best view of the speaker from the available camera views. This was then used by the system to create a composite view video sequence output consisting of a user defined view and an automatically inserted view of the active speaker. The system was evaluated on 10 segments from different lecture recordings totalling over $53min$ and was found to reliably estimate the active speaker position with an average 3D Euclidean error of $0.2m$.

7.1 Future Work

The future success of joint audio video based tracking systems relies on the continuing efforts of researchers in each individual domain to strive towards improvements in tracking reliability and accuracy.

In the audio domain, significant challenges prevail. The most dominant of these is the detrimental effect of reverberation on TDE-based localisation. Continuous efforts in developing methods to evaluate the reliability of TDEs are essential since this is what currently limits the contribution of audio in joint audio-video tracking systems. The importance of optimising the positions of microphone arrays for improving audio-based localisation accuracy was highlighted in this thesis. It is believed that this will become an important requirement for joint audio-video based systems. The technique proposed in chapter 5 for optimising microphone array positions has the limitation that it does not incorporate any measure of early reflections in the employed model of reverberation. Future work on optimising microphone positions which does consider this problem is necessary.

In the video-tracking domain, conditions of temporally and spatially varying illumination together with the problem of detecting faces at various poses still represent significant difficulties. These issues are strongly voiced in the video tracking literature and continuous efforts in regard to these will contribute to improvement in the multi-modal tracking problem.

Although there is much work to be done in both the individual research domains of audio

and video based tracking, it is important that there is a communication of ideas and technologies between each of the domains. At present, there is an obvious divide between the audio-based tracking and video-based tracking research communities. As a result, significant barriers face researchers wishing to develop joint audio-video based systems. Experts in one domain may not have the required expertise in the other domain to achieve their goal. There is a clear trend in much of the current literature which reflects this. Research originating from experts in video-based tracking tend to use sophisticated state-of-the-art video-based tracking systems but relatively primitive audio-based tracking techniques. Contrary trends are observed with research originating for audio-based experts.

As was highlighted in chapter 4, it is common for proposed joint audio-video tracking systems to be shown to improve upon single modality trackers. If primitive single modality trackers are being used, the significance of this observation is lessened. The evaluation of joint audio-visual tracking using state-of-the-art techniques from each domain would give a greater indication as to whether joint audio video fusion has an application beyond that of just improving the reliability of primitive trackers. Therefore, there is a need for the transfer of state-of-the-art technologies between the audio-based tracking and video-based tracking research communities. It is felt that without this, progress towards improving both accuracy and not only reliability through joint audio-video based tracking techniques will be hindered significantly.

There is much potential to extend the capabilities of the V-VAST system for the post-production of recorded lectures. A natural extension of the system would be to evaluate the *best* audio stream from the available microphones within the lecture room. Since the system provides a *3D* estimate of the active speaker's head position, the relative distance of the speaker to microphones within the lecture room can be determined. A simple criterion for choosing the best audio stream could be to select the output of the microphone which is closest to the speaker. This would improve the overall quality of the audio in the created lecture presentation and increase the intelligibility of the recorded speech.

A completely understated problem in relation to the implementation of audio and video based tracking systems, is the relative calibration between the audio and video tracking spaces. There is no adequate treatment of this topic in the existing literature. Currently, sophisticated semi-automated to fully automated techniques exist for the calibration of multiple cameras in 3D space. There are no existing equivalent techniques for calibrating multiple microphones. At present in audio-video based systems, highly sophisticated calibration techniques are employed for cameras however the calibration of microphones is generally achieved by manual measurements of the microphone positions. The author's experience in relation to this through the experimental analysis in chapter 4 is that minute errors in manually measuring microphone positions can introduce significant resultant localisation errors. This arises since small measurement errors when propagated over multiple microphones will translate into significant localisation inaccuracies. The strategy employed in chapter 4 to address this was to localise the microphones within the tracking space of the cameras, effectively using the cameras to calibrate the micro-

phone positions. This was implemented by minimising the relative error between audio-based localisation over a trajectory in *3D* to that of the estimated trajectory obtained through video-based localisation. Ideally, the development of more appropriate, more accurate and automated techniques is desired.

An interesting recent development which could facilitate this is through the use of array geometries which enforce the same epipolar constraints on measurements as observed by cameras [8]. This offers the potential to view microphone arrays as generalised cameras meaning that existing camera calibration techniques can be used for calibrating microphone arrays. Further studies into the feasibility of this could see the seamless integration of both microphones and cameras sensors for audio-visual tracking.

A

Approximation of the First Order Derivative of the Inverse Time-Delay Measurement Function

In this appendix, it is described how the implicit functions theorem can be used to determine the first order derivative of the inverse time-delay measurement function. Given $G(\tau, \mathbf{x})$ and C_g as defined in Sec. 4.2 in equation 4.10 the implicit functions theorem implies [138, chapter 5],

$$\frac{\partial g^{-1}(\tau)}{\partial \tau} = -\mathbf{H}^{-1}\frac{\partial \Phi}{\partial \tau} \tag{A.1}$$

where Φ is defined as

$$\Phi = \left(\frac{\partial C_g}{\partial \mathbf{x}}\right)^T = 2\sum_i^N G_i(\mathbf{x}, \tau)\left(\frac{\partial G_i}{\partial \mathbf{x}}\right)^T. \tag{A.2}$$

and $\mathbf{H}$ is derived as,

$$\mathbf{H} = \frac{\partial \Phi}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial C_g}{\partial \mathbf{x}} \right)^T \tag{A.3}$$

$$= 2 \sum_i^N G_i(\mathbf{x}, \tau) \frac{\partial}{\partial \mathbf{x}} \left(\frac{\partial G_i}{\partial \mathbf{x}} \right)^T + 2 \sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{x}} \right)^T \frac{\partial G_i}{\partial \mathbf{x}} \tag{A.4}$$

$$\approx 2 \sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{x}} \right)^T \frac{\partial G_i}{\partial \mathbf{x}} \tag{A.5}$$

$$= 2 \sum_i^N \begin{bmatrix} \frac{\partial G_i}{\partial X}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial X}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial X}\frac{\partial G_i}{\partial Z} \\ \frac{\partial G_i}{\partial Y}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial Y}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial Y}\frac{\partial G_i}{\partial Z} \\ \frac{\partial G_i}{\partial Z}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial Z}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial Z}\frac{\partial G_i}{\partial Z} \end{bmatrix} \tag{A.6}$$

$$= 2 \left(\frac{\partial G}{\partial \mathbf{x}} \right)^T \frac{\partial G}{\partial \mathbf{x}}, \tag{A.7}$$

Using equation A.2 $\frac{\partial \Phi}{\partial \tau}$ is obtained by,

$$\frac{\partial \Phi}{\partial \tau} = \frac{\partial}{\partial \tau} \left(\frac{\partial C_g}{\partial \mathbf{x}} \right)^T \tag{A.8}$$

$$= 2 \sum_i^N G_i(\mathbf{x}, \tau) \frac{\partial}{\partial \tau} \left(\frac{\partial G_i}{\partial \mathbf{x}} \right)^T + 2 \sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{x}} \right)^T \frac{\partial G_i}{\partial \tau} \tag{A.9}$$

$$\approx 2 \sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{x}} \right)^T \frac{\partial G_i}{\partial \tau} \tag{A.10}$$

$$= 2 \sum_i^N \begin{bmatrix} \frac{\partial G_i}{\partial \tau_1}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial \tau_1}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial \tau_1}\frac{\partial G_i}{\partial Z} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \frac{\partial G_i}{\partial \tau_N}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial \tau_N}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial \tau_N}\frac{\partial G_i}{\partial Z}] \end{bmatrix} \tag{A.11}$$

$$= 2 \left(\frac{\partial G}{\partial \mathbf{x}} \right)^T \frac{\partial G}{\partial \tau}. \tag{A.12}$$

Substituting (A.7) and (A.12) into (A.1), $\frac{\partial g^{-1}(\tau)}{\partial \tau}$ is determined as,

$$\frac{\partial g^{-1}(\tau)}{\partial \tau} \approx - \left(\frac{\partial G}{\partial \mathbf{x}} \right)^\dagger \frac{\partial G}{\partial \tau} \tag{A.13}$$

where † is used to denote the pseudo inverse.

B

Audio-Video Localisation: Experimental Setup

This appendix presents some additional details in relation to the experimental setup used in chapter 5 for analysing the accuracy of audio-visual source localisation. Section B.1 documents the model numbers of both the video cameras and the microphones used in the experimental analysis together with their positions within the room. Also included in section B.1 are the calibration points used to calibrate the video cameras in the experiment. Presented in section B.2 is a brief note which considers the optimality of the microphone positions which were used in the analysis.

B.1 Video Cameras and Microphones

Three video cameras and six microphones were used in the experimental setup. The positioning of the microphones and the video cameras is given below in table B.1. In relation to the microphones, the quoted positions refer to the position of the centre of the microphone capsule within the room. The quoted video camera positions refer to the camera centres as obtained through the calibration procedures. Also documented in table B.1 are the model details of the equipment used. A list of both the calibration training points and calibration test points are presented in table B.2.

B.2 Note on the Optimality of the Experimental Setup

This section presents a retrospective note which considers the optimality of the microphone array placement used in the experiment. The main point of the experiment was to examine what the

Equipment	**Position**			**Model**
	$x\ (mm)$	$y\ (mm)$	$z\ (mm)$	
Camera 1	1416.3	38.0	2340.8	Canon DM-XM2
Camera 2	2741.6	157.8	2312.7	Canon DM-XM2
Camera 3	3976.1	101.1	2242.1	Panasonic NV GS250EB
Microphone 1	1037.6	5550.7	1304.7	Rode $NT5$
Microphone 2	1019.6	5204.9	1310.1	Rode $NT5$
Microphone 3	1016.8	5381.5	1614.7	Rode $NT5$
Microphone 4	3997.4	5557.4	1301.7	Rode $NT5$
Microphone 5	4000.0	5210.8	1315.0	Rode $NT5$
Microphone 6	4001.3	5394.3	1603.4	Rode $NT5$

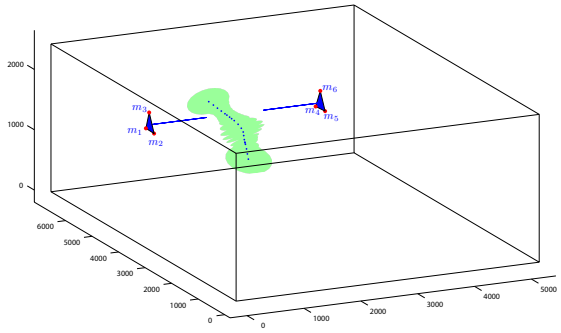
Table B.1: *Description and positions of the six microphones and three video cameras used in the analysis of audio-visual source localisation.*

expected accuracy of audio-visual source localisation is in the absence of any consideration for optimal sensor placement. Using the algorithm for optimising the placement of microphone arrays as presented in algorithm 1 however, it is possible to determine the optimal placement of the microphones for the given experimental scenario.

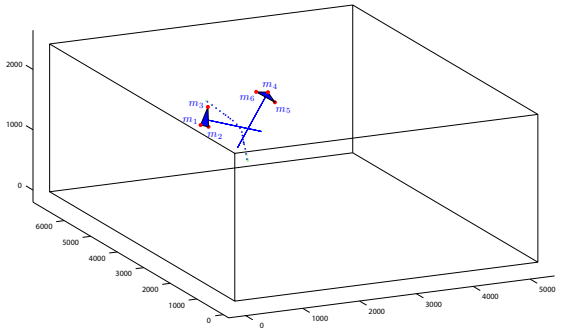
Shown in figure B.1a is the original configuration of the six microphones used in experiment together with a visual illustration of the localisation error associated with this configuration. The optimised microphone positions are illustrated in figure B.1b and listed in figure B.1c. Comparing both figure B.1a and figure B.1b it is clear that there is a significant improvement in the expected localisation accuracy where the microphone positions are optimised. This further motivates the need for optimising the positions of microphones for localisation when considering joint audio-visual fusion for tracking.

Calibration Points					
Training Points			Test Points		
$x\ (mm)$	$y\ (mm)$	$z\ (mm)$	$x\ (mm)$	$y\ (mm)$	$z\ (mm)$
1220	6980	2369	1517	6980	2159
1620	6980	2369	1917	6980	2159
2020	6980	2369	2317	6980	2159
2420	6980	2369	2717	6980	2159
2820	6980	2369	3117	6980	2159
3220	6980	2369	3517	6980	2159
3620	6980	2369	3917	6980	2159
1220	6980	1969	1517	6980	1759
1620	6980	1969	1917	6980	1759
2020	6980	1969	2317	6980	1759
2420	6980	1969	2717	6980	1759
2820	6980	1969	3117	6980	1759
3220	6980	1969	3517	6980	1759
3620	6980	1969	3917	6980	1759
1220	6980	1569	1517	6980	1359
1620	6980	1569	1917	6980	1359
2020	6980	1569	2317	6980	1359
2420	6980	1569	2717	6980	1359
2820	6980	1569	3117	6980	1359
3220	6980	1569	3517	6980	1359
3620	6980	1569	3917	6980	1359
1220	6980	1169	1517	6980	959
1620	6980	1169	1917	6980	959
2020	6980	1169	2317	6980	959
2420	6980	1169	2717	6980	959
2820	6980	1169	3117	6980	959
3220	6980	1169	3517	6980	959
3620	6980	1169	3917	6980	959
1822	5795	9	2119	5585	9
2172	5795	9	2469	5585	9
2522	5795	9	2819	5585	9
1822	5545	9	2119	5335	9
2172	5545	9	2469	5335	9
2522	5545	9	2819	5335	9
1822	5295	9	2119	5085	9
2172	5295	9	2469	5085	9
2522	5295	9	2819	5085	9

Table B.2: *Positions of both the calibration training points and test points used in calibrating the video cameras*



(a) Original microphone array positions of the experimental setup. The 95 percentile error ellipsoids representing localisation error are shown in green.



(b) Optimised microphone array positions using the optimisation algorithm described in chapter 5. The 95 percentile error ellipsoids representing localisation error are shown in green.

Microphone	**Position**		
	x (mm)	y (mm)	z (mm)
Microphone 1	2082.6	5755.0	1172.1
Microphone 2	2063.0	5414.6	1232.6
Microphone 3	2154.3	5630.5	1495.3
Microphone 4	3156.1	5519.2	1648.0
Microphone 5	3123.3	5182.6	1571.0
Microphone 6	2870.6	5348.2	1731.0

(c) Optimised x, y and z positions of the microphones for the experimental setup.

Figure B.1: *Illustration of the unoptimised (a) and optimised microphone array positions (b) for the experimental setup used in chapter 5. Shown in (c) are the optimised x, y and z locations of the microphones in the room for the given experimental setup. The optimisation algorithm used in this analysis corresponds to that of algorithm 1.*

C

Complete Set of Seminar Tracking Results

This appendix presents the complete set of results obtained for the visual tracking task of the CHIL 2005 evaluation. These results give a more detailed view of the tracking results which yielded tables 6.2 and 6.3 in chapter 6. As was outlined in chapter 6, the green in the following figures corresponds to points in the video sequence where the presenter's face was not visible in at least two views. The plotted results correspond to the ground truth shown in blue and the V-VAST tracking results shown in red. In these results the V-VAST algorithm is applied to the single speaker tracking scenario (see section 6.6.1 for more information on the particular configuration of the V-VAST algorithm used to generate these results).

The results of table 6.3 show the overall performance of the V-VAST algorithm on the visual tracking task of the CHIL 2005 evaluation package. From this table it can be seen that on average the algorithm accurately tracks the speaker with a mean *3D* global error of $0.24m$. From the complete set of results presented in figures C.1 to C.10 however it can be seen that some of the results have a consistent bias along the z axis. This bias is particularly noticeable in figures C.1, C.2, C.3 and C.4 where although the x and y tracking results are accurate, there is a clear offset in the tracking results along the z axis in relation to the ground truth. This offset reveals that at times, heads are tracked by the V-VAST algorithm at positions slightly lower than their true positions. It can also be seen in figures C.1, C.2, C.3 and C.4 that this offset is most often small and not greater than $0.25m$. Given that V-VAST models heads as having a height of $0.241m$, the apparent offset in z corresponds to just over 1 head height. This offset effects the quality of framing the speaker in the *best view* output of the V-VAST algorithm. Effectively, when the offset occurs the view of the speaker is framed about a point which is 1

head height below the true head position. This was not deemed to significantly affect the visual quality of the results.

The observed tracking bias in the results is attributed to the primitive head model which the V-VAST algorithm employs. Tracking using such a primitive model can be affected by the presence of exposed neck regions. V-VAST fits a head model to detected skin regions. If skin is detected at exposed neck regions below a head position, then the head model fitting process will be distorted. Under such circumstances V-VAST will fit the head model to a region below the true head position. The offset in the z axis is attributed mainly to this tracking distortion.

The head model which V-VAST employs has further limitations. By only considering skin colour in detecting head regions, V-VAST is unable to discern skin regions corresponding to a face from that of skin colour regions relating to hands and arms. This can lead to inaccurate tracking results where hand or arm positions occur at locations close to head positions.

The results shown in this appendix evaluate the ability of V-VAST to track the presenter in various seminar recordings. Figures C.4 and C.10 reveal some large discrepancies between the V-VAST tracking results and the ground truth. The ground truth corresponds to the presenter's position only, irrespective of whether they are speaking or not. There are some instances in these particular recordings however where the active speaker is not the presenter but that of an audience member. The large divergences in figures C.4 and C.10 between the V-VAST tracking results and the ground truth arise at points in the recording where an audience member begins talking. These large divergences show V-VAST tracking the conversational switches between active speakers. The seminar recording corresponding to C.10 is examined in section 6.6.2 where the results are examined against ground truth which does incorporate conversational switches between different speakers. Direct comparison between figures C.10 and 6.14 respectively show the tracking results of V-VAST on the same seminar recording configured for single speaker tracking and the tracking of conversational switches.

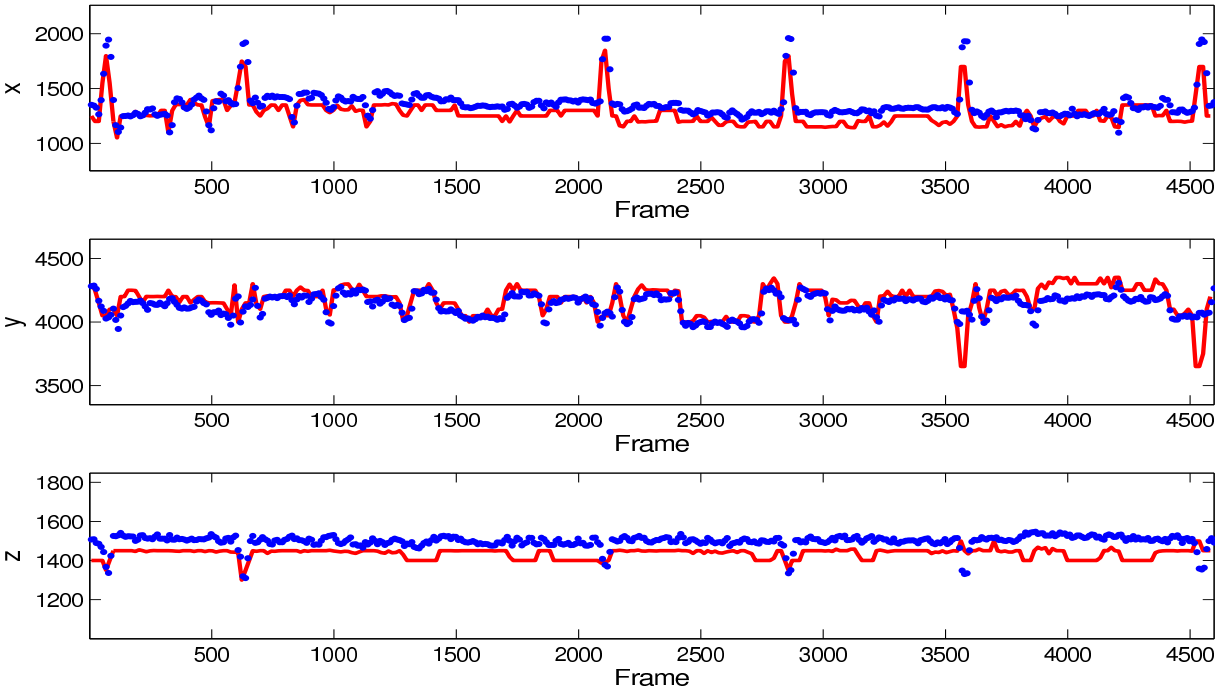


Figure C.1: *Tracking results for* ***seminar_2004-11-11_A_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*

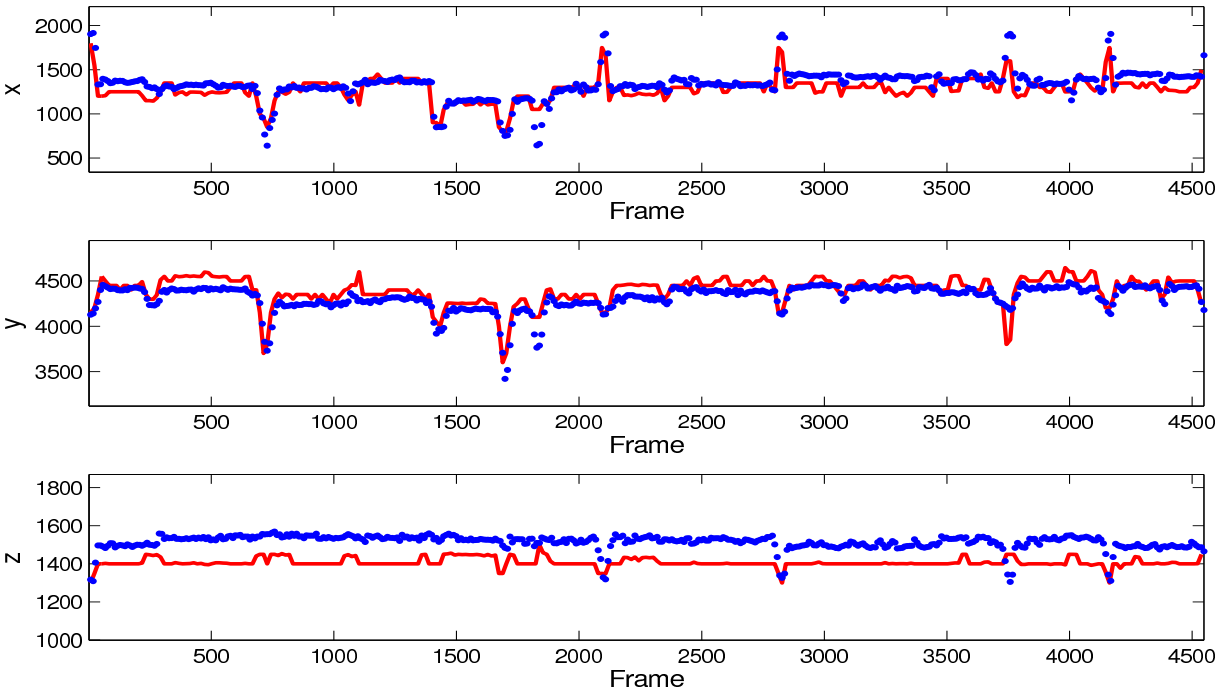


Figure C.2: *Tracking results for* ***seminar_2004-11-11_A_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*

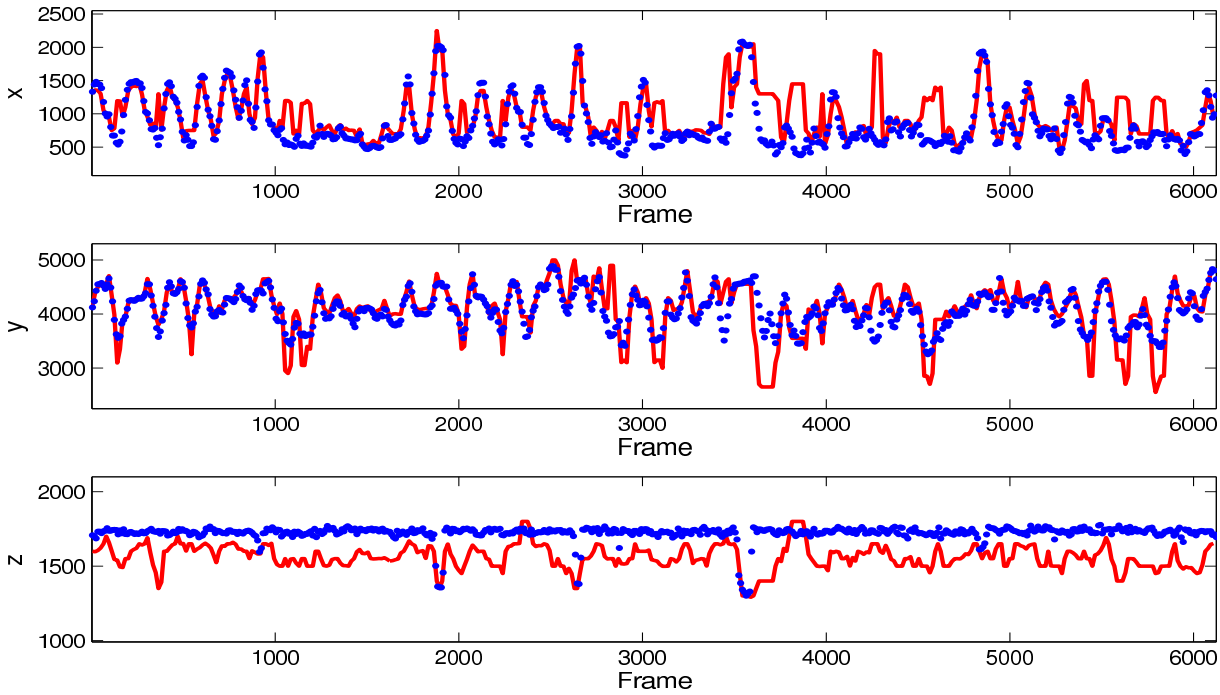


Figure C.3: *Tracking results for* ***seminar_2004-11-11_B_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*

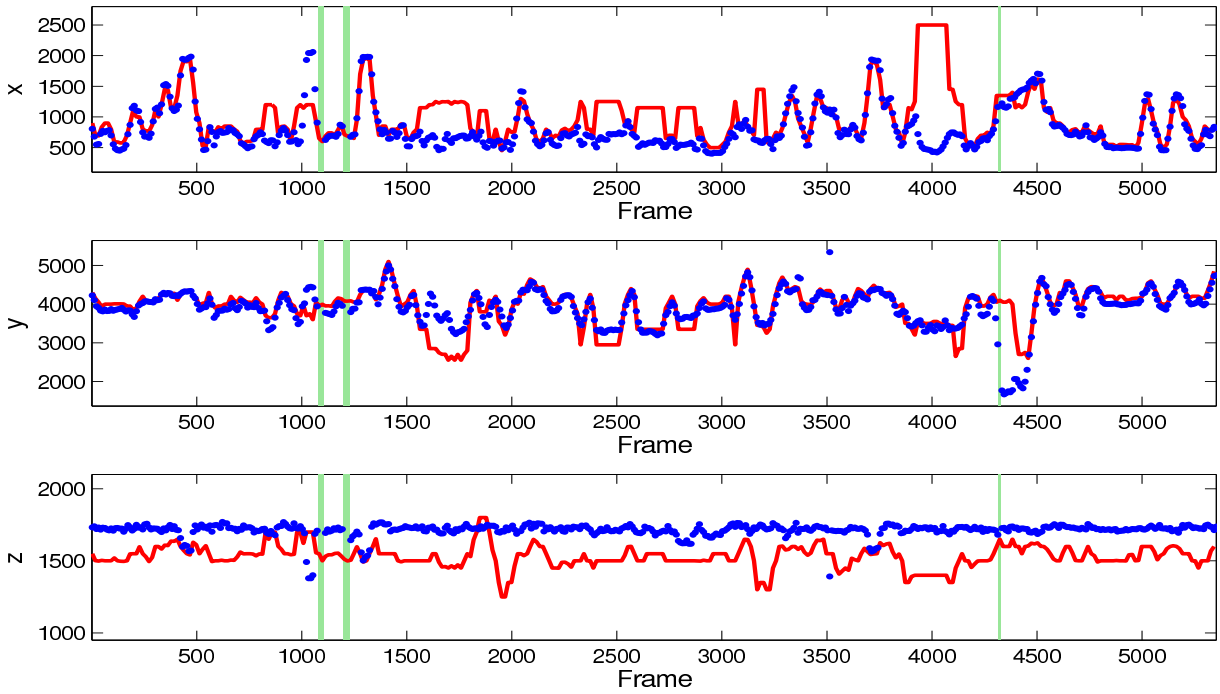


Figure C.4: *Tracking results for* ***seminar_2004-11-11_B_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*

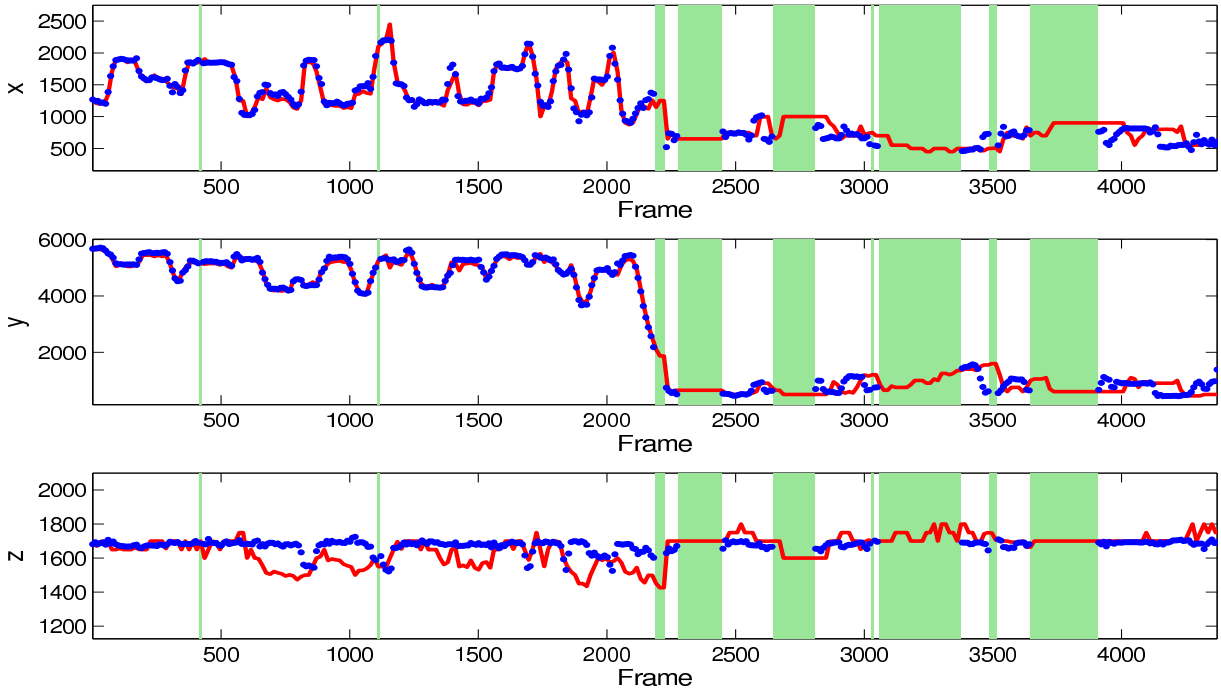


Figure C.5: *Tracking results for* ***seminar_2004-11-11_C_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*

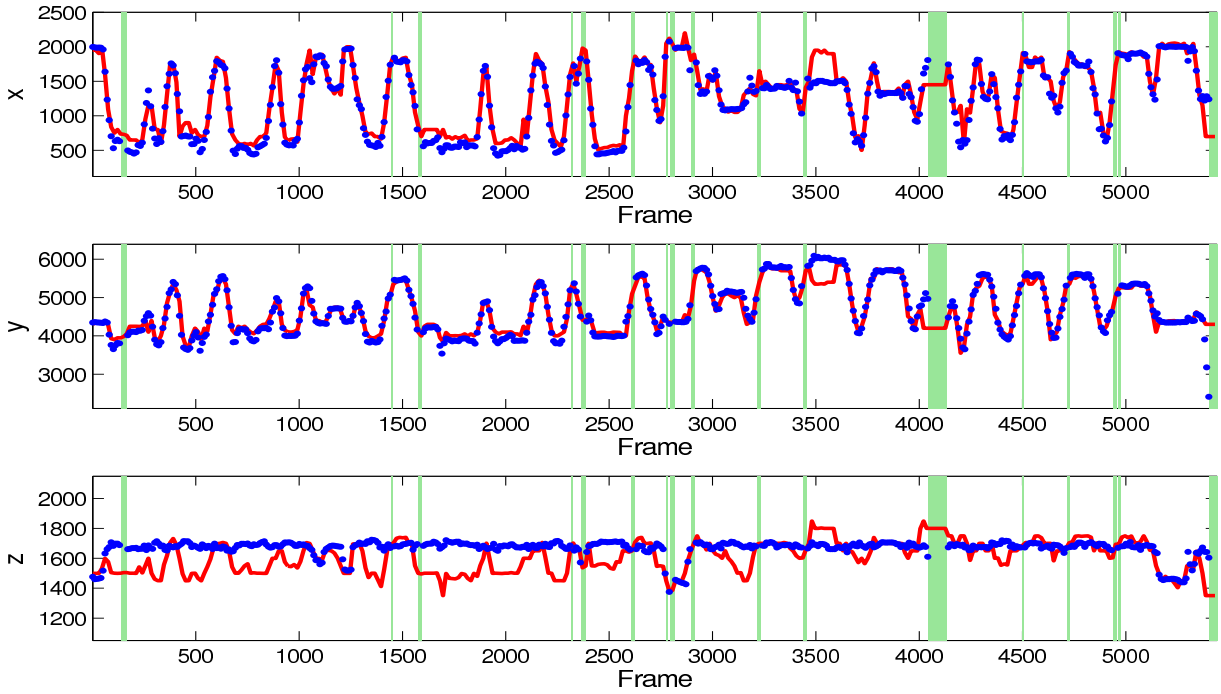


Figure C.6: *Tracking results for* ***seminar_2004-11-11_C_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*

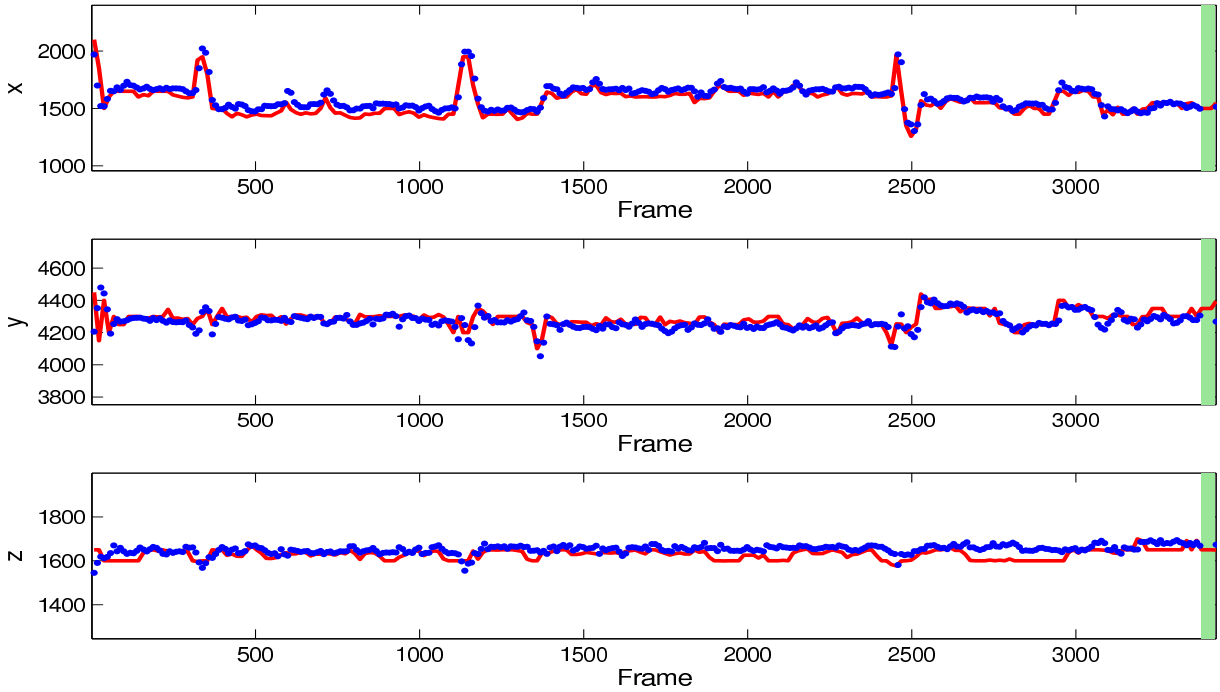


Figure C.7: *Tracking results for* ***seminar_2004-11-12_A_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*

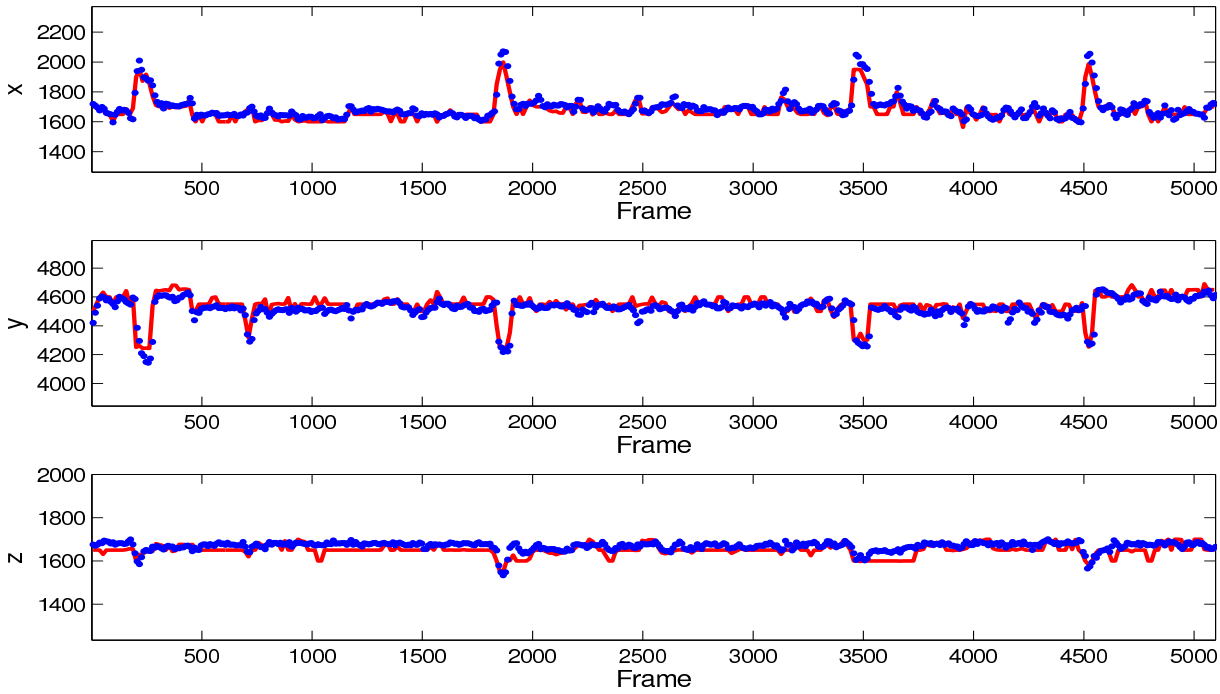


Figure C.8: *Tracking results for* ***seminar_2004-11-12_A_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*

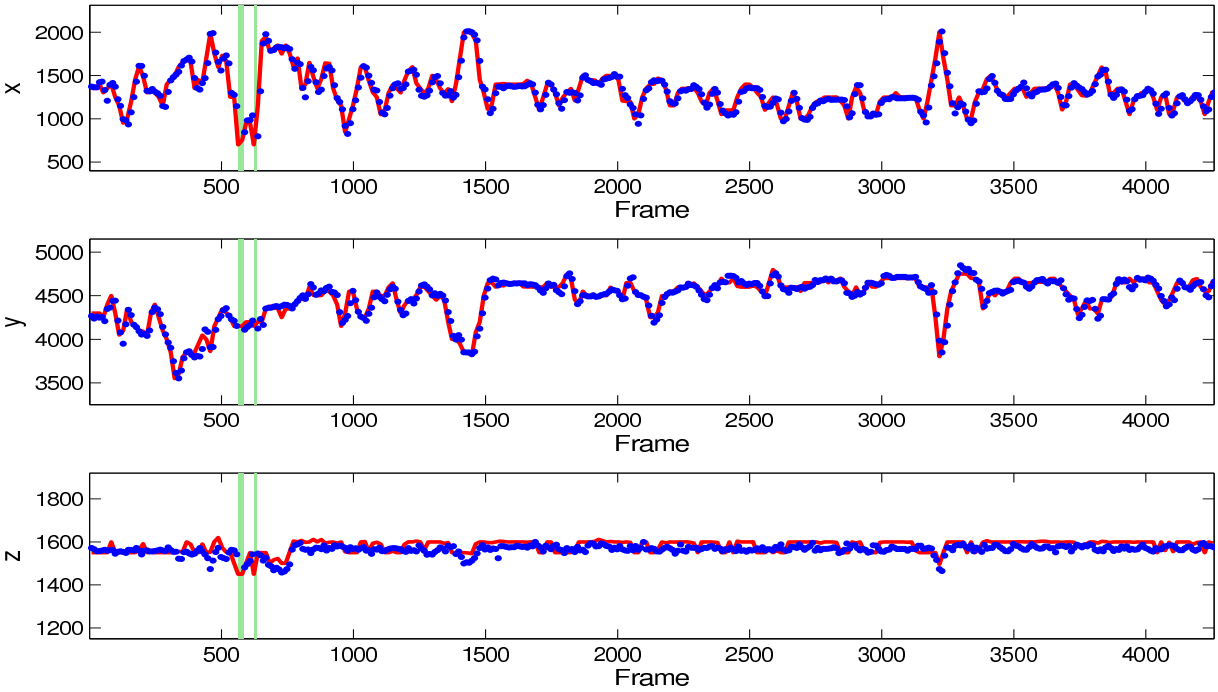


Figure C.9: *Tracking results for* ***seminar_2004-11-12_B_segment1*** *in x, y and z coordinates (red) against ground truth (blue).*

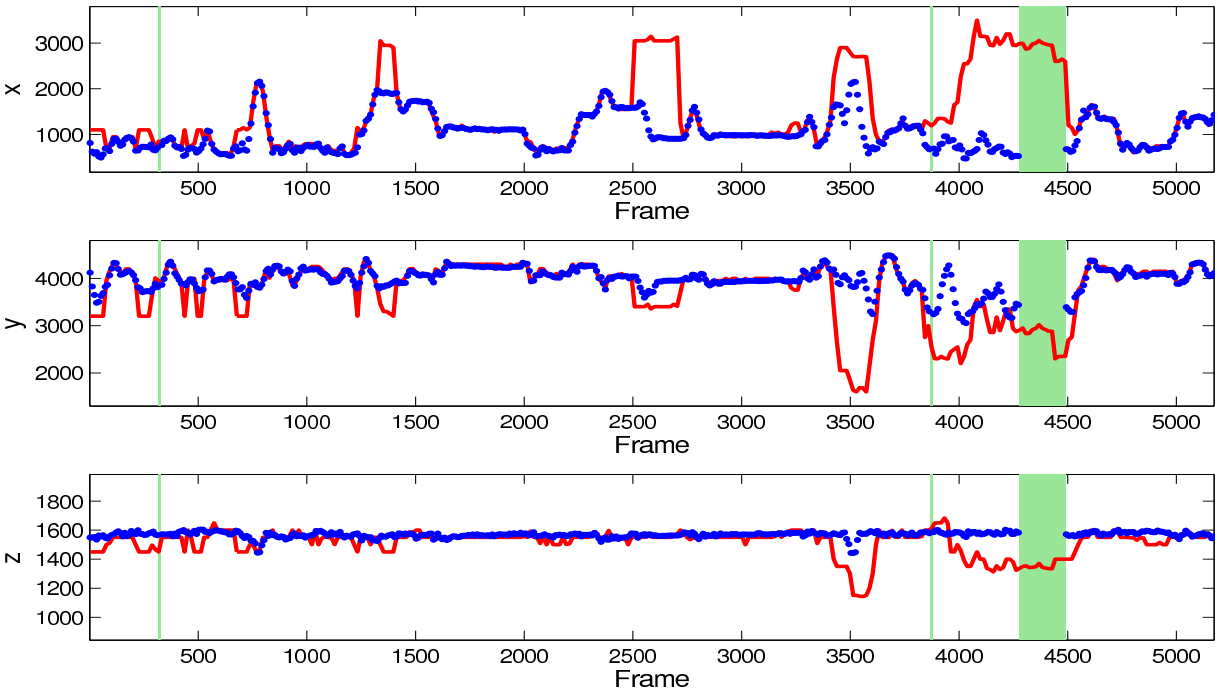


Figure C.10: *Tracking results for* ***seminar_2004-11-12_B_segment2*** *in x, y and z coordinates (red) against ground truth (blue).*

D

Multi-camera Calibration Procedure

This appendix is provided to present the reader with more details relating to the multi-camera calibration technique employed in the analysis of chapter 4. It is assumed in this presentation that the reader has some familiarity with multi-view geometry and camera calibration methods. Background to the various multi-view signal processing techniques described in this appendix can be found in the textbooks of O. Faugeras [138] and Hartley et al. [154].

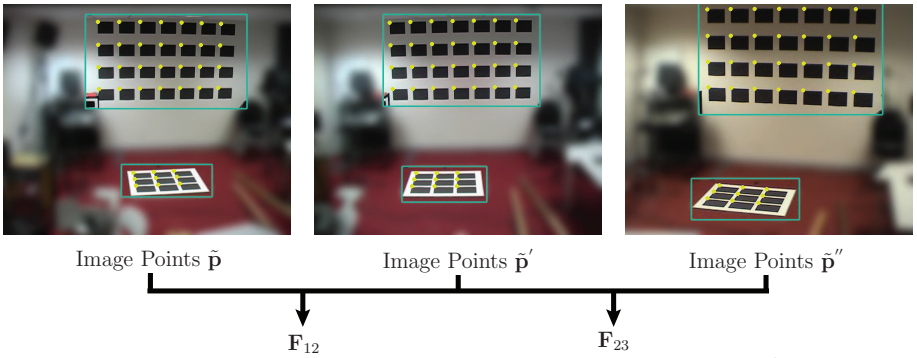
Chapter 4 required the calibration of three video cameras for the purpose of localising points in *3D* space. This appendix briefly outlines the calibration procedure which was implemented in achieving this. A *stratified* approach [154, pg. 267] to reconstructing a set of *3D* points $\tilde{\mathbf{x}}$ from a set of image point correspondences was used to fully calibrate the three cameras. Both the intrinsic and extrinsic camera parameters were obtained using manual measurements of the *3D* points $\tilde{\mathbf{x}}$ and their extracted images $\tilde{\mathbf{p}}$, $\tilde{\mathbf{p}}^{'}$ and $\tilde{\mathbf{p}}^{''}$ in cameras 1, 2 and 3 respectively. Recall that in this thesis, $\tilde{\mathbf{p}}$ refers to the homogeneous representation of an image point. In the following, it is described how this information can be used to obtain the set of camera matrices $\{\mathbf{P}_E, \mathbf{P}^{'}_E, \mathbf{P}^{''}_E\}$, corresponding to the three cameras which is necessary for localising imaged points in *3D* space.

The first procedure in the calibration process required obtaining two fundamental matrices $\mathbf{F}_{12}$ and $\mathbf{F}_{23}$, from the point correspondences $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}^{'} \leftrightarrow \tilde{\mathbf{p}}^{''}\}$ across the three camera views [154, Chapter 11]. The fundamental matrix $\mathbf{F}_{12}$ was obtained using the point correspondence $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}^{'}\}$ and $\mathbf{F}_{23}$ was obtained using point correspondences $\{\tilde{\mathbf{p}}^{'} \leftrightarrow \tilde{\mathbf{p}}^{''}\}$. Two sets of camera matrices $\{\mathbf{P}, \mathbf{P}^{'}\}$ and $\{\mathbf{P}^{'}, \mathbf{P}^{''}\}$ were obtained from $\mathbf{F}_{12}$ and $\mathbf{F}_{23}$ respectively [154, pg. 253]. Using these camera matrices, two projective reconstructions $\tilde{\mathbf{x}}_{12}$ and $\tilde{\mathbf{x}}_{23}$ were respectively obtained from the point correspondences $\{\tilde{\mathbf{p}}^{'} \leftrightarrow \tilde{\mathbf{p}}^{''}\}$ and $\{\tilde{\mathbf{p}}^{'} \leftrightarrow \tilde{\mathbf{p}}^{''}\}$. The reconstructions $\tilde{\mathbf{x}}_{12}$ and $\tilde{\mathbf{x}}_{23}$

as well as the camera matrices $\{\mathbf{P}, \mathbf{P}^{'}, , \mathbf{P}^{''}\}$ are obtained through this approach with projective ambiguity. This projective ambiguity was corrected using manual measurements of the *3D* points $\tilde{\mathbf{x}}$.

Given $\tilde{\mathbf{x}}$, two *3D* homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ were obtained using the Direct Linear Transformation (DLT) algorithm [154, pg. 88] such that $\tilde{\mathbf{x}} = \mathbf{H}_1\tilde{\mathbf{x}}_{12}$ and $\tilde{\mathbf{x}} = \mathbf{H}_2\tilde{\mathbf{x}}_{23}$. These homographies were then applied to correct the projective ambiguity in the camera matrices $\{\mathbf{P}, \mathbf{P}^{'}, \mathbf{P}^{''}\}$ to determine the set of camera matrices $\{\mathbf{P}_E, \mathbf{P}^{'}_E, \mathbf{P}^{''}_E\}$ where $\mathbf{P}_E = \mathbf{P}\mathbf{H}_1^{-1}$, $\mathbf{P}^{'}_E = \mathbf{P}^{'}\mathbf{H}_1^{-1}$ and $\mathbf{P}^{''}_E = \mathbf{P}^{''}\mathbf{H}_2^{-1}$ [154, pg. 266]. From the new set of camera matrices $\{\mathbf{P}_E, \mathbf{P}^{'}_E, \mathbf{P}^{''}_E\}$ and point correspondences $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}^{'} \leftrightarrow \tilde{\mathbf{p}}^{''}\}$, linear triangulation [154, pg. 312] was then employed to estimate a Euclidean reconstruction $\hat{\tilde{\mathbf{x}}}$ of the points $\tilde{\mathbf{x}}$.

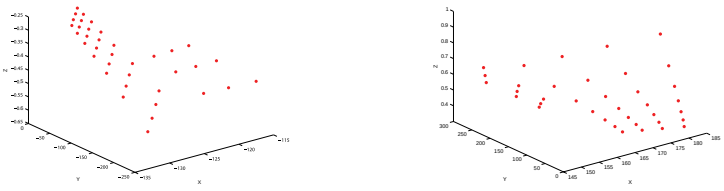
In order to refine the estimates of the camera matrices, bundle adjustment was applied over the camera matrices $\{\mathbf{P}_E, \mathbf{P}^{'}_E, \mathbf{P}^{''}_E\}$ and estimated reconstruction $\hat{\tilde{\mathbf{x}}}$ [154, pg. 434]. This required projecting $\hat{\tilde{\mathbf{x}}}$ back into each camera view to determine the set of backprojected point correspondences $\{\hat{\tilde{\mathbf{p}}} \leftrightarrow \hat{\tilde{\mathbf{p}}}^{'} \leftrightarrow \hat{\tilde{\mathbf{p}}}^{''}\}$. Bundle adjustment was then performed by minimising the geometric error between the point correspondences $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}^{'} \leftrightarrow \tilde{\mathbf{p}}^{''}\}$ and the backprojected point correspondences $\{\hat{\tilde{\mathbf{p}}} \leftrightarrow \hat{\tilde{\mathbf{p}}}^{'} \leftrightarrow \hat{\tilde{\mathbf{p}}}^{''}\}$. An overview of the described calibration procedure for obtaining the set of camera matrices $\{\mathbf{P}_E, \mathbf{P}^{'}_E, \mathbf{P}^{''}_E\}$ is illustrated in figure D.1.



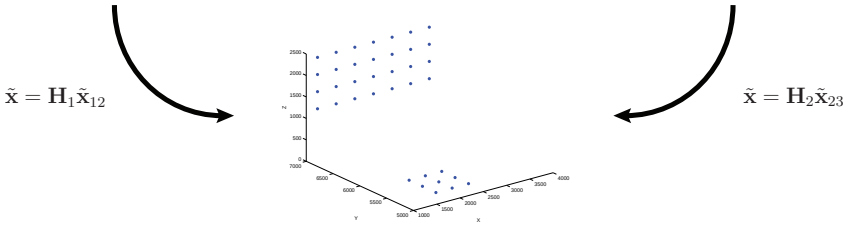
Step 1: Obtain fundamental matrix $\mathbf{F}_{12}$ from image point correspondences $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}'\}$ and fundamental matrix $\mathbf{F}_{23}$ from image point correspondences $\{\tilde{\mathbf{p}}' \leftrightarrow \tilde{\mathbf{p}}''\}$.

$\{\mathbf{P}, \mathbf{P}'\}$ $\{\mathbf{P}', \mathbf{P}''\}$

Step 2: Determine camera matrices $\{\mathbf{P}, \mathbf{P}'\}$ from $\mathbf{F}_{12}$ and $\{\mathbf{P}', \mathbf{P}''\}$ from $\mathbf{F}_{23}$ with $\mathbf{P}' = [I|0]$.



Step 3. Determine the projective reconstruction $\tilde{\mathbf{x}_{12}}$ from $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}'\}$ and $\{\mathbf{P}, \mathbf{P}'\}$. Also determine a projective reconstruction $\tilde{\mathbf{x}}_{23}$ using $\{\tilde{\mathbf{p}}' \leftrightarrow \tilde{\mathbf{p}}''\}$ and $\{\mathbf{P}', \mathbf{P}''\}$.



Known 3D points $\tilde{\mathbf{x}}$.

Step 4: Calculate the $3D$ homographies $\mathbf{H}_1$ and $\mathbf{H}_2$ where $\tilde{\mathbf{x}} = \mathbf{H}_1\tilde{\mathbf{x}}_{12}$ and $\tilde{\mathbf{x}} = \mathbf{H}_2\tilde{\mathbf{x}}_{23}$.

Step 5: Correct projective ambiguity in camera matrices through $\mathbf{P}_E = \mathbf{P}\mathbf{H}_1^{-1}$, $\mathbf{P}'_E = \mathbf{P}'\mathbf{H}_1^{-1}$ and $\mathbf{P}''_E = \mathbf{P}''\mathbf{H}_2^{-1}$.

Step 6: Using point correspondences $\{\tilde{\mathbf{p}} \leftrightarrow \tilde{\mathbf{p}}' \leftrightarrow \tilde{\mathbf{p}}''\}$ and camera matrices $\{\mathbf{P}_E, \mathbf{P}'_E, \mathbf{P}''_E\}$ determine $\hat{\tilde{\mathbf{x}}}$ an estimate of $\tilde{\mathbf{x}}$ by triangulation.

Step 7: Reproject $\hat{\tilde{\mathbf{x}}}$ back into each view to define $\hat{\tilde{\mathbf{p}}} = \mathbf{P}_E\hat{\tilde{\mathbf{x}}}$, $\hat{\tilde{\mathbf{p}}}' = \mathbf{P}'_E\hat{\tilde{\mathbf{x}}}$ and $\hat{\tilde{\mathbf{p}}}'' = \mathbf{P}''_E\hat{\tilde{\mathbf{x}}}$.

Step 8: *Bundle Adjustment:* Minimize $\sum_i d(\tilde{\mathbf{p}}_i, \hat{\tilde{\mathbf{p}}}_i)^2 + d(\tilde{\mathbf{p}}_i, \hat{\tilde{\mathbf{p}}}'_i)^2 + d(\tilde{\mathbf{p}}_i, \hat{\tilde{\mathbf{p}}}''_i)^2$ over $\{\mathbf{P}_E, \mathbf{P}'_E, \mathbf{P}''_E\}$ and $\hat{\tilde{\mathbf{x}}}$ where $d(\mathbf{a}, \mathbf{b})$ denotes the Euclidean distance between points $\mathbf{a}$ and $\mathbf{b}$.

Figure D.1: *Flow diagram of the steps towards reconstructing 3D points from image point correspondences. The process describes a stratified approach to 3D reconstruction where a projective reconstruction of points is refined to a Euclidean reconstruction. This technique is uses to obtain the set of cameras matrices* $\{\mathbf{P}_E, \mathbf{P}'_E, \mathbf{P}''_E\}$ *required for the* 3D *reconstruction of imaged points.*

Bibliography

[1] A. Abad, C. Canton-Ferrer, C. Segura, J. L. Landabasso, D. Macho, J. R. Casas. UPC Audio, Video and Multimodal Person Tracking Systems in the CLEAR Evaluation Campaign. In *Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton U.K., April 2006. 3.2.2

[2] A. Albiol, L. Torres, E. J. Delp. Optimum Color Spaces for Skin Detection. In *Int. Conf. on Image Processing (ICIP)*, pages 122–124, 2001. 2.2.3

[3] A. Brutti, M. Omologo, P. Svaizer, C. Zieger. Classification of Acoustic Maps to Determine Speaker Position and Orientation from a Distributed Microphone Network. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, volume 4, pages 493–496, 2007. 5.2.2

[4] A. D. Pierce. *Acoustics: An Introduction to its Physical Principle and Applications*. McGraw-Hill, 1981. 2.1, 2.1.1, 2.1.1, 2.1.1

[5] A. Doucet, N. de Freitas, N. Gordon. An Introduction to Sequential Monte Carlo Methods. In A. Doucet, N. de Freitas, N. Gordon, editor, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001. 3.1, 3.1

[6] A. H. Quazi. An Overview on the Time Delay Estimate in Active and Passive Systems for Target Localization. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 29:527–533, 1981. 2.1.3.2, 5

[7] A. Mahajan, M. Walworth. 3-D Position Sensing using the Differences in the Time-of-Flights from a Wave Source to Various Receivers. *IEEE Transactions on Robotics and Automation*, 17(1), Feb. 2001. 2.1.3, 2.2a

[8] A. O' Donovan, R. Duraiswami, J. Neumann. Microphone Arrays as Generalized Cameras for Automated Audio Visual Processing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 7.1

[9] A. Pnevmatikakis, G. Talantzis, J. Soldatos, L. Polmenakos. Robust Multimodal Audio-Visual Processing for Advanced Context Awareness in Smart Spaces. *Personal and Ubiquitous Computing*, 13(1):3–14, 2009. 3.2.2

[10] A. Yilmaz, O. Javed, M. Shah. Object Tracking: A Survey. *ACM Computing Surveys (CSUR)*, 38(13), 2006. 3

[11] Apple. iTunes EDU. www.itunes.com. Last Accessed, 24 Oct. 2009. 1

[12] B. Champagne, S. Bédard, A. Stéphenne. Performance of Time-Delay Estimation in the Presence of Room Reverberation. *IEEE Transactions on Speech and Audio Processing*, 4(2), Mar. 1996. 2.1.3.1, 2.1.3.1

[13] B. Kapralos, M. Jenkin, E. Milios, J. Tsotsos. Eyes 'n Ears: Face Detection Utilizing Audio And Video Cues. In *IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, pages 106–112, Vancouver, Canada, 2001. 3

[14] B. Kapralos, M. R. M. Jenkin, E. Milios. Audiovisual Localization of Multiple Speakers in a Video Teleconferencing Setting. *International Journal Imaging Systems and Technology*, 13:95–105, 2003. 3

[15] B. Martinkauppi, M. Soriano, M. Pietikainen. Detection of Skin Color under Changing Illumination: A Comparative Study. In *12th International Conference on Image Analysis and Processing (ICIAP)*, pages 652–657, September 2003. 2.2.3

[16] B. Yang, J. Scheuing. Cramer-Rao Bound and Optimum Sensor Array for Source Localization from Time Differences of Arrival. In *IEEE Int. Conference on Acoustics Speech and Signal Processing (ICASSP '05)*, pages 961–964, 2005. 5

[17] M. Brandstein, J. Adcock, and H. Silverman. Microphone Array Localization Error Estimation with Application to Sensor Placement. *Journal of the Acoustical Society of America (JASA)*, 99:3807–3816, 1996. 4, 4.2, 5, 5.1.2

[18] C. Brown, H. Durrant-Whyte, J. Leonard, B. Rao, B. Steer. Distributed data fusion using kalman filtering: A robotics approach. In M. A. Abidi and R. C. Gonzalez, editor, *Data Fusion in Robotics and Machine Intelligence*, pages 267–309. Academic Press, 1992. 3.2.2

[19] C. Choi, D. Kong, S. Lee, K. Park, S. -G. Hong, H. -K. Lee, S. Bang, Y. Lee,. S. Kim. Real-Time Audio-Visual Localization of User Using Microphone Array and Vision Camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems 2005*, pages 1935–1940, 2005. 3

[20] C. Knapp, G. Carter. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans. on Acoustics Speech and Signal Processing*, ASSP-24:320–327, 1976. 2.1.2, 2.1.3.1, 2.1.3.1, 2.1.3.1, 4.4, 5

[21] C. Kotropoulos, I. Pitas. Rule-based Face Detection in Frontal Views. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 2537–2540, 1997. 2.2.2

[22] C. Stauffer, W. E. L. Grimson. Adaptive Background Models for Real-Time Tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, 2:246–252, 1999. 2.2.1

[23] C. Wang, M. S. Brandstein. A Hybrid Real-Time Face Tracking System. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3737–3740, 1997. 3

[24] C. Wang, M. S. Brandstein. Multi-source Face Tracking with Audio and Visual Data. In *IEEE 3rd Workshop on Multimedia Signal Processing*, pages 169–174, 1999. 3

[25] C. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland. Pfinder Real-Time Tracking of the Human Body. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 19:780–785, 1997. 2.2.1

[26] C. Zhang, D. Florêncio, Z. Zhang. Why Does PHAT Work Well in Low Noise, Reverberant Environments? In *International Conference on Acoustics, Speech and Signal Processing*, pages 2565–2568, 2008. 2.1.3.1

[27] C. Zhang, Y. Rui, J. Crawford, L. -W. He. An Automated End-to-end Lecture Capture and Broadcast System. *ACM Transactions on Multimedia Computing, Communications and Applications*, 4(1), Jan. 2008. Article 6. 1

[28] Cambridge University. CamTV. http://mediaplayer.group.cam.ac.uk/. Last Accessed, 24 Oct. 2009. 1

[29] G. C. Carter. Variance Bounds for Passively Locating an Acoustic Source with a Symmetric Line Array. *J. Acoust. Soc. Am.*, 62:922–926, 1977. 5

[30] CHIL. Computer in the Human Interaction Loop. http://chil.server.de/. Last Accessed March 26, 2010. 1.1

[31] D. B. Ward, E. A. Lehmann, R. C. Williamson. Particle Filtering Algorithms for Acoustic Source Localization. *IEEE Transactions on Speech and Audio Processing*, 11(6):826–836, 2003. 3.2.3, 3.2.3

[32] D. B. Ward, R. C. Williamson. Particle Filter Beamforming for Acoustic Localisation in a Reverberant Environment. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1777–1780, 2002. 3.1.4

[33] D. Bechler, K. Kroschel. Three different criteria for time-delay estimates. In *European Conference on Signal Processing (Eusipco)*, pages 1987–1990, 2004. 2.1.3.2, 6.1.1.2

[34] D. Bechler, M. Grimm, K. Kroschel. Speaker Tracking with a Microphone Array using Kalman filtering. *Advances in Radio Science*, 1:113–117, 2003. 3.1.3

[35] D. E. Sturim, M. S. Brandstein, H. F. Silverman. Tracking Multiple Talkers Using Microphone-Array Measurements. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 371–374, 1997. 3.1.3

[36] D. Gatica-Perez, G. Lathoud, I. McCowan, J. -M. Odobez, D. Moore. Audio-Visual Speaker Tracking with Importance Particle Filters. In *International Conference on Image Processing (ICIP)*, Barcelona, 2003. 3.2.3

[37] D. Gatica-Perez, G. Lathoud, J. -M. Odobez, I. McCowan. Multimodal Multispeaker Probabilistic Tracking in Meetings. In *International Conference on Multimodal Interfaces (ICMI)*, Trento, Italy, 2005. 3.2.3

[38] D. Gatica-Perez, G. Lathoud, J.-M Odobez, I. McCowan. Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings. *IEEE Trans. on Audio Speech and Language Processing*, pages 601–616, Feb. 2007. 3.2.3

[39] D. Kelly, F. Boland. Motion Model Selection in Tracking Humans. *Irish Signals and Systems Conf. (ISSC)*, pages 363–368, 2006. 1.3.1, 3.1.2

[40] D. Kelly, F. Boland. Optimal Microphone Placement for Active Speaker Localization. In *8th IMA International Conference on Mathematics in Signal Processing*, Cirencester, England UK, Dec. 16th-18th 2008. 1.3.1, 1

[41] D. Kelly, F. Pitie, A. Kokaram, F. Boland. A Comparative Error Analysis of Audio-Visual Source Localization. In *Workshop on Multi-camera Multi-modal Sensor Fusion Algorithms and Applications* (M^2SFA^2) *in conjunction with 10th European Conference on Computer Vision (ECCV)*, 2008. 1.3.1, 1

[42] D. L. McCarty. *Microphone Array Processing Techniques for Classroom-Based Videoconferencing.* PhD thesis, Department of Electronic and Electrical Engineering, Trinity College Dublin, 2007. 2.1.1, 5.2.1

[43] D. N. Zotkin, R. Duraiswami, L. S. Davis. Joint Audio Visual Tracking Using Particle Filters. *EURASIP Journal on Applied Signal Processing*, 11:1154–1164, 2002. 3.2.3, 3.2.3, 3.2.3, 3.2.3

[44] D. Singh. *Audio Signal Analysis for Classification and Source Localization in e-Learning Applications.* PhD thesis, Department of Electronic and Electrical Engineering, Trinity College Dublin, 2007. 2.1.3.2

[45] E. A. Lehmann, R. C. Williamson. Experimental Comparison of Particle Filtering Algorithms for Acoustic Source Localization in a Reverberant Room. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 147–150, 2003. 3.1.4

[46] E. A. P. Habets, P. C. W. Sommen. Optimal Microphone Placement for Source Localization using Time Delay Estimation. In *Proc. of the 13th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC)*, Nov. 28-29 2002. 5

[47] E. Hjelmås, B. K. Low. Face Detection@ A Survey. *Computer Vision and Image Understanding*, 83(3), Sept. 2001. 2.2.2

[48] E. Osuna, R. Freund, F. Girosi. Training Support Vector Machines: An Application to Face Detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–136, 1997. 2.2.2

[49] Echo360. EchoSystem. http://www.echo360.com/the-echosystem/. Last Accessed, 24 Oct. 2009. 1

[50] ELRA. European Language Resources Association. http://www.elra.info/. Last Accessed March 26, 2010. 1.1

[51] F. M. Boland, H. Nicholson. Control of Divergence in Kalman Filters. *Electronics Letters*, 12:367–369, 1976. 3.1.2

[52] F. N. Fritsch, R. E. Carlson. Monotone Piecewise Cubic Interpolation. *SIAM Journal on Numerical Analysis*, 17:238–246, 1980. 6.1.2.3

[53] F. Pitié, S.-A. Berrani, A. Kokaram, R. Dayhot. Off-line Multiple Object Tracking using Candidate Selection and the Viterbi Algorithm. In *IEEE International Conference on Image Processing (ICIP '05)*, volume 3, pages 109–112, 2005. 6

[54] F. van der Heijden, R. P. W. Duin, D. de Ridder, D. M. J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach using MATLAB*. Wiley, 2004. 3.1.1

[55] R. J. Fitzgerald. Divergence of the Kalman Filter. *IEEE Transactions on Automatic Control*, 16:736–747, 1971. 3.1.2

[56] G. C. Carter. Coherence and Time Delay Estimation. In *Proceedings of the IEEE*, volume 75, February 1987. 2.1.2, 2.1.3.2, 5

[57] G. D. Forney. The Viterbi Algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. 3.1.5, 6

[58] G. Jacovitti, G. Scarano. Discrete Time Techniques for Time Delay Estimation. *IEEE Transactions on Signal Processing*, 41:525–533, 1993. 4.4

[59] G. Lathoud, J. -M. Odobez, D. Gatica-Perez. AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking. In *Workshop on Machine Learning and Multimodal Interaction (MLMI)*, 2004. http://www.idiap.ch/av16_3corpus/. 2.5

[60] G. Olague, R. Mohr. Optimal Camera Placement for Accurate Reconstruction. *Pattern Recognition*, 35:927–944, 2002. 4, 4.3

[61] G. Pingali. Integrated Audio-Visual Processing for Object Localization and Tracking. In *Proceedings of Multimedia Computing and Networking, SPIE*, volume 3310, pages 206–213, 1998. 3

[62] G. Pingali, G. Tunali, I. Carlbom. Audio-Visual Tracking for Natural Interactivity. In *ACM International Conference on Multimedia (MM)*, pages 373–382, 1999. 3

[63] G. R. Cooper, C. D. McGillem. *Probabilistic Methods of Signal and System Analysis*. HRW Series in Electrical Engineering, Electronics and Systems, Holt, Rhinehart and Winston, Inc., NY, 1971. 5.2.1

[64] G. Wang, R. Rabenstein, N. Strobel, S. Spors. Object Localization by Joint Audio-Video Signal Processing. In *Proceedings of Vision, Modeling, and Visualization (VMV)*, pages 97–104, Saarbruecken, Germany, Nov. 2000. 3.2.2

[65] G. Welch, G. Bishop. An Introduction to the Kalman Filter. Technical Report TR 95-041, University of North Carolina at Chapel Hill, Department of Computer Science, 1995. 3.1.1

[66] G. Welch, G. Bishop. SCAAT: Incremental Tracking with Incomplete Information. In *SIGGRAPH*, pages 333–344, 1997. 3.2.2

[67] Google. Youtube EDU. www.youtube.com/edu. Last Accessed, 24 Oct. 2009. 1

[68] H. A. Rowley. *Neural Network-based Face Detection*. PhD thesis, Carnegie Mellon University, 1999. 2.2.2

[69] H. Anton, C. Rorres. *Elementary Linear Algebra: Applications Version*. Wiley, Seventh edition, 1994. 2.2.4.2

[70] H. C. Schau, A. Z. Robinson. Passive Source Localization Employing Intersecting Spherical Surfaces from Time-of-Arrival Difference. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(8), Aug. 1987. 2.1.3

[71] H. Durrant-Whyte. Consistent Integration and Propagation of Disparate Sensor Observations. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 1464–1469, 1986. 4.1

[72] H. Krim, M. Viberg. Two Decades of Array Signal Processing. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996. 2.1.2, 3

[73] H. Kuttruff. *Room Acoustics*. Spon Press, London and NY, Fifth edition, 2009. 2.1, 2.1.1

[74] H. Rowley, S. Baluja, T. Kanade. Neural Network-based Face Detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–208, 1996. 2.2.2

[75] H. Wang, P. Chu. Voice Source Localization for Automatic Camera Pointing System in Videoconferencing. In *Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 187–190, 1997. 2.1.3.1

[76] H. Zhang. Two-Dimensional Optimal Sensor Placement. *IEEE Transactions on Systems, Man and Cybernetics*, 25(5), May 1995. 5

[77] Human Factors Engineering Technical Advisory Group. Human Engineering Design Data Digest. http://www.hfetag.com/docs/pocket_guide.doc. Last Accessed, 24 Oct. 2009. 6.2.4

[78] I. Craw, D. Tock. Finding Face Features. In *European Conference on Computer Vision (ECCV)*, pages 92–96, 1992. 2.2.2

[79] J. -C. Terrillon, H. Fukamachi, S. Akamatsu, M. N. Shirazi. Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images. *Automatic Face and Gesture Recognition, IEEE International Conference on*, pages 54–61, March 2000. 2.2.3

[80] J. A. Beracoechea, S. Torres-Guijarro, L. García, F. J. Casajús-Quirós. On Building Immersive Audio Applications Using Robust Adaptive Beamforming and Joint Audio-Video Source Localization. *EURASIP Journal on Applied Signal Processing*, 2006:1–12, 2006. 3

[81] J. A. Brotherton, G. A. Abowd. Lessons Learned for eClass: Assessing Automated Capture and Access in the Classroom. *ACM Transactions on Computer-Human Interaction*, 11(12):121–155, 2004. 1

[82] J. B. Allen, D. A. Berkley. Image Method for Efficiently Simulating Small-room Acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, 1979. 5.4

[83] J. Brand, J. S. Mason. A Comparative Assessment of Three Approaches to Pixel-Level Human Skin Detection. In *15th Int. Conf. on Pattern Recognition*, volume 1, pages 1056–1059, 2000. 2.2.3

[84] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. Wiley-Interscience, New York, 2003. 5.3

[85] J. Chen, J. Benesty, Y. A. Huang. Time Delay Estimation in Room Acoustic Environments: An Overview. *EURASIP Journal on Applied Signal Processing*, 2006:1–19, 2006. 2.1.3.1

[86] J. Chen, Y. A. Huang, J. Benesty. A Comparative Study on Time Delay Estimation in Reverberant and Noisy Environments. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 21–24, Oct. 16-19 2005. 2.1.3.1

[87] J. DiBiase, H. Silverman, M. Brandstein. Robust Localization in Reverberant Rooms. In M. Brandstein and D. Ward, editor, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 157–180. Springer-Verlag, 2001. 2.1.5

[88] J. H. DiBiase. *A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments using Microphone Arrays*. PhD thesis, Brown University, Providence RI, USA, May 2000. 2.1.3.1, 2.1.5, 4.4

[89] J. Huopaniemi, K. Kettunen, J. Rahkonen. Measurement and Modeling Techniques for Directional Sound Radiation from the Mouth. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 183–186, Oct. 17-20 1999. 2.1.1, 5.2.2, 5.2.2

[90] J. J. Wang, S. Singh. Video Analysis of Human Dynamics - A Survey. *Real-Time Imaging*, 9(5):321–346, 2003. 3

[91] J. Jaffe, S. Feldstein. *Rhythms of Dialogue*, chapter Chapter 2. Academic Press, New York, 1970. 6.1.2.3

[92] J. K. Aggarwal, Q. Cai. Human Motion Analysis: A Review. *Computer Vision and Image Understanding*, 73(3):428–440, Mar. 1999. 3

[93] J. L. Flanagan. Analog Measurements of Sound Radiation from the Mouth. *The Journal of the Acoustical Society of America*, 32(12):1613–1620, Dec. 1960. 2.1.1, 5.2.2, 5.2.2

[94] J. Manyika, H. Durrant-Whyte. *Data Fusion and Sensor Management, A Decentralized Information Theoretic Approach*. Ellis Horwood Series in Electrical and Electronic Engineering, Ellis Horwood, 1994. 3.2.3

[95] J. Neering, M. Bordier, N. Maizi. Optimal Passive Source Localization. In *Int. Conf. on Sensor Technologies and Applications*, pages 295–300, Oct 2007. 5, 5.1.2

[96] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, second edition, 1985. 3.2.3, 3.2.3

[97] J. P. Ianniello. Time Delay Estimation via Cross-Correlation in the Presence of Large Estimation Errors. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-30(6), December 1982. 5.2, 5.2.1

[98] J. S. Abel. Optimal Sensor Placement for Passive Source Localization. *International Coference on Acoustics Speech and Signal Processing (ICASSP)*, 5:2927–2930, 1990. 5

[99] J. S. Bendat, A. G. Piersol. *Random Data: Analysis and Measurement Procedures*. Wiley, Third edition, 2000. 5.2.1

[100] J. Vermaak, A. Blake. Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 3021–3024, 2001. 3.1.4, 3.2.3

[101] J. Vermaak, M. Gagnet, A. Blake, P. Perez. Sequential Monte Carlo Fusion of Sound and Vision for Speaker Tracking. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 741–746, 2001. 3.2.3, 3.2.3, 3.2.3

[102] J. Yang, A. Waibel. A Real-Time Face Tracker. In *Third IEEE Workshop on Applications of Computer Vision*, pages 142–147, 1996. 2.2.3

[103] K. Bernardin, T. Gehrig, R. Stiefelhagen. Multi-level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In *Workshop on Classification of Events, Actions and Relations (CLEAR)*, Baltimore, MD, USA, May 2007. 3.2.1, 3.2.3

[104] K. C. Yow, R. Cipolla. A probabilistic Framework for Perceptual Grouping of Features. In *International Conference on Automatic Face and Gesture Recognition*, pages 16–21, 1996. 2.2.2

[105] K. Nickel, T. Gehrig, H. K. Ekenel, J. McDonough, R. Stiefelhagen. An Audio-Visual Particle Filter for Speaker Tracking on the CLEAR'06 Evaluation Dataset. In *Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, U.K., April 2006. 3.2.3, 3.2.3, 3.2.3

[106] K. Nickel, T. Gehrig, R. Steifelhagen, J. McDonough. A Joint Particle Filter for Audio-visual Speaker Tracking. In *International Conference on Multimodal Interfaces (ICMI)*, pages 61–68, 2005. 3.2.3, 3.2.3, 3.2.3, 3.3c, 3.2.3

[107] K. Rohr. Extraction of 3D Anatomical Point Landmarks based on Invariance Principles. *Pattern Recognition*, 32:3–15, 1999. 5.1.1

[108] K. Wilson, T. Darrell. Improving Audio Source Localization by Learning the Precedence Effect. In *International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 1125–1128, 2005. 2.1.3.2

[109] G. Lathoud. *Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays*. PhD thesis, École Polytechnique Fédérale de Lausanne, December 2006. 4.4

[110] M. -H. Yang, D. J. Kriegman, N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), Jan. 2002. 2.2.2

[111] M. -H. Yang, N. Ahuja. Detecting Human Faces in Color Images. In *International Conference on Image Processing (ICIP)*, volume 1, pages 127–130, 1996. 2.2.3

[112] M. Bianchi. Automatic Video Production of Lectures using an Intelligent and Aware Environment. In *3rd International Conference on Mobile and Ubiquitous Multimedia*, volume Mobile and Ubiquitous Multimedia: Vol.83, pages 117–123, 2004. 1

[113] M. Brandstein, H. Silverman. A Practical Methodology for Speech Source Localization with Microphone Arrays. *Computer, Speech and Language*, 11(2):91–126, 1997. 2.1.3, 2.1.3.1

[114] M. Bregonzio, M. Taj, A. Cavallaro. Multi-modal Particle Filtering Tracking using Appearance, Motion and Audio Likelihoods. In *International Conference on Image Processing (ICIP)*, pages 33–36, 2007. 3.2.3, 3.2.3

[115] M. C. Shin, K. I. Chang, L. V. Tsap. Does Colorspace Transformation make any Difference on Skin Detection. In *IEEE Workshop on Applications of Computer Vision*, pages 275–279, 2002. 2.2.3

[116] M. Collobert, R. Feraud, G. Le Tourneour, O. Bernier. LISTEN: A System for Locating and Tracking Individual Speakers. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG '96)*, pages 283–288, 1996. 3

[117] M. F. Berger, H. F. Silverman. Microphone Array Optimization by Stochastic Region Contraction. *IEEE Trans. on Signal Processing*, 39:2377–2386, Nov. 1991. 5

[118] M. H. Kalos, P. A. Whitlock. *Monte-Carlo Methods*. Wiley-VCH, 2004. 4.4.2

[119] M. Hartle, H. Bär, C. Trompler, G. Rößling. Perspectives for Lecture Videos. *Lecture Notes in Computer Science*, 3648/2005: Euro-Par 2005 Parallel Processing: 11th International Euro-Par Conference:901–908, 2005. 1

[120] M. J. Jones. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*, 46(1):81–96, 2002. (document), 2.2.3, 2.10c, 2.10e, 2.10

[121] M. Jian, A. C. Kot, M. H. Er. Performance Study of Time Delay Estimation in a Room Environment. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 5, pages 554–557, 1998. 3.2.3, 5.2.1

[122] M. Murakami, S. Nishiguchi, Y. Kameda, M. Minoh. Effect on Lecturer and Students by Multimedia Lecture Archive System, 2003. 1

[123] M. Omologo, P. Svaizer. Acoustic Event Localization Using A Crosspower-Spectrum Phase Based Technique. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 273–276, 1994. 2.1.5

[124] M. R. Schroeder. Frequency-Correlation Functions. *Journal of the Acoustical Society of America*, 34(12):1819–1823, Dec. 1962. 5.2.1

[125] M. S. Brandstein, J. E. Adcock, H. F. Silverman. A Closed-Form Location Estimator for Use with Room Environment Microphone Arrays. *IEEE Transactions on Speech and Audio Processing*, 5(1), 1997. 2.2b

[126] M. Turk, A. Pentland. Face Recognition using EigenFaces. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991. 2.2.2

[127] Massachusetts Institute of Technology, USA. OpenCourseWare. http://web.mit.edu/ocw/. Last Accessed, 24 Oct. 2009. 1

[128] N. Checka, K. W. Wilson, M. R. Sicacusa, T. Darrell. Multiple Person and Speaker Activity Tracking with a Particle Filter. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 881–884, 2004. 3.2.3, 3.2.3, 4

[129] N. Checka, K. Wilson, V. Rangarajan, T. Darrell. A Probabilistic Framework for Multi-modal Multi-person Tracking. In *Workshop on Multi-Object Tracking in conjunction with Conference Computer Vision and Pattern Recognition (CVPR)*, 2003. 3.2.3

[130] N. Flores, S. Savage. Student Demand for Streaming Lecture Video: Empirical Evidence from Undergraduate Economics Classes. In *International Review of Economics Education*, volume 6, pages 57–78, 2007. 1

[131] N. H. Timm. *Applied Multivariate Analysis*. Springer-Verlag, 2002. 5.1.1

[132] N. J. Gordon, D. J. Salmond, A. F. M. Smith. Novel Approach to Nonlinear/non-Guassian Bayesian state estimation. In *IEE Proceedings F, Radar and Signal Processing*, volume 140, pages 107–113, 1993. 3.1.4

[133] N. Katsarakis, G. Souretis, F. Talantzis, A. Pnevmatikakis, L. Polymenakos. 3D Audio-visual Person Tracking using Kalman Filtering and Information Theory. In *Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton U.K., April 2006. 3.2.2, 4, 5.5

[134] N. M. Oliver, B. Rosario, A. P. Pentland. A Bayesian Computer Vision System for Modeling Human Interactions. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 22:831–843, 2000. 2.2.1

[135] N. Strobel, S. Spors, R. Rabenstein. Joint Audio-Video Object Localization and Tracking: A Presentation of General Methodology. *IEEE Signal Processing Magazine*, 18(1):22–31, 2001. 3.2.2

[136] N. Strobel, S. Spors, R. Rabenstein. Joint Audio-Video Signal Processing for Object Localization and Tracking. In M. Brandstein and D. Ward, editor, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 203–225. Springer-Verlag, 2001. 3.1.3, 3.2b, 3.2.2, 4

[137] O. Erdinc, P. Willet, S. Coraluppi. Multistatic Sensor Placement: A Tracking Approach. *Journal of Advances in Information Fusion*, 2(1), June 2007. 5, 5.1.2

[138] O. Faugeras. *Three-Dimensional Computer Vision: A Geomteric Viewpoint*. MIT Press, 1993. 2.2.4, 4.1.1, A, D

[139] P. Aarabi. The Fusion of Distributed Microphone Arrays for Sound Localization. *EURASIP Journal on Applied Signal Processing*, 2003:338–347, 2003. 2.1.3.2

[140] P. Aarabi, S. Zaky. Integrated Vision and Sound Localization. In *Third International Conference on Information Fusion, (FUSION 2000)*, volume 2, pages 21–27, July 2000. 3.2.3

[141] P. Aarabi, S. Zaky. Robust Sound Localization using Multi-source Audiovisual Information Fusion. *Information Fusion*, 3(2):209–223, 2001. 3.2.3

[142] P. G. Barker, I. D. Benest. The On-line Lecture Content- A Comparison of Two Approaches. In *IEE Colloquium on Learning at a Distance: Developments in Media Technologies*, volume Digest No. 1996/148, pages 1–7, June 1996. 1

[143] P. Viola, M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. *IEEE Conference on Computer Vision and Pattern Recognition*, 1:511–518, 2001. 2.2.2, 3.2.3

[144] P. Viola, M. Jones. Fast Multi-View Face Detection. Technical Report TR2003-96, Mitsubishi Electric Research Laboratories (MERL), 2003. 2.2.2

[145] Panopto. CourseCast. http://www.panopto.com/products_coursecast.aspx. Last Accested, 24 Oct. 2009. 1

[146] Peter Hancock. Psychological Image Collection at Stirling. http://pics.psych.stir.ac.uk. Last accessed March 26, 2010. 2.2.3, 2.6

[147] M. Piccardi. Background Subtraction Techniques: A Review. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 4, 2004. 2.2.1

[148] C. F. Price. An Analysis of the Divergence Problem in the Kalman Filter. *IEEE Trans. on Automatic Control*, 13:699–702, 1968. 3.1.2

[149] Princeton University, New Jersey, USA . Webmedia. http://www.princeton.edu/WebMedia/lectures/. Last Accessed, 24 Oct. 2009. 1

[150] Q. Liu, Y. Rui, A. Gupta, J. J. Cadiz. Automating Camera Management for Lecture Room Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 442–449, Seattle, Washington, USA, 2001. 1

[151] R. Brunelli, A. Brutti, P. Chipendale, O. Lanz, M. Omologo, P. Svaizer, F. Tobia. A Generative Approach to Audio-Visual Person Tracking. In *Workshop on Classification of Events, Actions and Relations (CLEAR)*, Southampton, U.K., April 2006. 3.3d, 3.2.3, 3.2.3, 3.2.3

[152] R. C. Gonzalez, R. E. Woods. *Digital Image Processing*. Prentice Hall, second edition, 2002. 4.4

[153] R. Cucchiara, C. Grana, M. Piccardi, A. Prati. Detecting Moving Objects, Ghosts and Shadows in Video. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 25:1337–1442, 2003. 2.2.1

[154] R. Hartley, A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Second edition, 2004. 2.2.4, 4.4, D

[155] R. Khattree, D. N. Naik. *Applied Multivariate Statistics with SAS Software*. SAS Institute and Wiley, Second edition, 2002. 5.1.1

[156] R. L. Hsu, M. Abdel-Mottaleb, A. K. Jain. Face Detection in Color Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:696–706, 2002. (document), 2.2.3, 2.2.3, 2.10b, 2.10

[157] R. Martin. Small Microphone Arrays with Postfilters for Noise and Acoustic Reduction. In M. Brandstein and D. Ward, editor, *Microphone Arrays: Signal Processing Techniques and Applications*, pages 256–279. Springer-Verlag, 2001. 5

[158] R. Phillimore. Face to Face Lectures or eContent: Students and Staff Perspective. In *International Conference on Computers in Education (ICCE)*, volume 1, pages 211–212, 2002. 1

[159] R. Stiefelhagen, K. Bernardin, H. K. Ekenel, J. McDonough, K. Nickel, M. Voit, M. Wölfel. Audio-visual Perception of a Lecturer in a Smart Seminar Room. *Signal Processing - Special Issue on Multimodal Interfaces*, 86(12), Dec. 2006. 3.2.3, 3.2.3, 3.2.3

[160] S. Arulampalam, S. Maskell, N. Gordon, T. Clapp. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Trans. on Signal Processing*, 50:174–188, 2002. 3.1.4, 3.1.4

[161] S. Gannot, T. G. Dvorkind. Microphone Array Speaker Localizers Using Spatial-Temporal Information. *EURASIP Journal on Applied Signal Processing*, pages 1–17, 2006. 3.1.3

[162] S. Gazor, Y. Grenier. Criteria for Positioning of Sensors for a Microphone Array. *IEEE Transactions on Speech and Audio Processing*, 3:294–303, 1995. 5

[163] S. Godsill, A. Doucet, M. West. Maximum A Posteriori Sequence Estimation Using Monte Carlo Particle Filters. *Annals of the Institute of Statistical Mathematics*, 53:82–96, 2001. 6

[164] S. J. Julier, J. K. Ulhmann. Unscented Filtering and Nonlinear Estimation. In *Proceedings of the IEEE*, volume 92, pages 401–422, 2004. 3.1.2, 3.1.3, 4.4.2

[165] S. J. McKenna, S. Gong, Y. Raja. Modeling Facial Colour and Identity with Gaussian Mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998. 2.2.3

[166] S. Julier, J. K. Uhlmann. Data Fusion in Nonlinear Systems. In D. L. Hall and J. Llinas, editors, *Handbook of Multisensor Data Fusion*, chapter 13. CRC Press, 2001. 4.1.1, 4.4.2

[167] S. Julier, J. Uhlmann, H. F. Durrant-Whyte. A New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators. *IEEE Transactions on Automatic Control*, 45:477–482, March 2000. 3.1.3, 4.4.2, 4.4.2

[168] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220:671–680, 1983. 5.3

[169] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume 2: Estimation and Theory*. Prentice Hall, 1993. 4.1.1

[170] S. Nishiguchi, K. Higashi, Y. Kameda, M. Minoh. A Sensor-Fusion Method for Detecting a Speaking Student. In *International Conference on Multimedia and Expo (ICME '03)*, volume 1, pages 129–132, 2003. 3

[171] S. Surcin, R. Steifelhagen, J. McDonough. D7.4 Evaluation Packages for the First CHIL Evaluation Campaign. http://chil.server.de/servlet/is/2712/. Last Accessed March 26, 2010. 3.3, 6.5

[172] S. T. Birchfield, G. Gangishetty. Acoustic Localization by Interaural Level Difference. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1109–1112, 2205. 2.1.4

[173] Sonic Foundary. MediaSuite. http://www.sonicfoundry.com/mediasite/. Last Accessed, 24 Oct. 2009. 1

[174] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, J. McDonough. Kalman Filters for Audio-Video Source Localization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2005. 3.2.2

[175] T. Gustafsson, B. D. Rao, M. Trivedi. Source Localization in Reverberant Environments: Part I - Modeling. *Technical Report, University of California, San Diego*, 2000. 5, 5.2, 5.2.1, 5.2.1

[176] T. Gustafsson, B. D. Rao, M. Trivedi. Source Localization in Reverberant Environments: Part II - Statistical Analysis. *Technical Report, University of California, San Diego*, 2000. 5.2, 5.2.1, 5.2.1

[177] T. Gustafsson, R. D. Bhaskar, M. Trivedi. Source Localization in Reverberant Environments: Modeling and Statistical Analysis. *IEEE Transactions on Speech and Signal Processing*, 11, Nov. 2003. 2.1.1, 2.1.3.1, 5.2, 5.2.1, 5.4, 5.2.1

[178] T. L. Tung, K. Yao, C. W. Reed, R. E. Hudson, D. Chen, J. Chen. Source Localization and Time Delay Estimation using Constrained Least Squares and Best Path Smoothing. In *SPIE*, volume 3807, pages 220–233, July 1999. 3.1.5

[179] T. S. Caetano, S. D. Olabarriaga, D. A. C. Barone. Do Mixture Models in Chromaticity Space Improve Skin Detection? *Pattern Recognition*, 12:3019–3021, 2003. 2.2.3

[180] T. Wark, S. Sridharan. A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification. In *Int. Conf. Acoustics Speech and Signal Processing*, volume 6, pages 3693–3696, 1998. 2.2.3

[181] Tegrity. Tegrity Campus. http://www.tegrity.com/tegrity-campus-20.html. Last Accessed, 24 Oct. 2009. 1

[182] The Math Works Inc. Natick, MA. USA. MATLAB R2008a. `http://www.mathworks.com`. Last Accessed, 24 Oct. 2009. 2

[183] U. Bub, M. Hunke, A. Waibel. Knowing Who to Listen to in Speech Recognition: Visually Guided Beamforming. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:848–851, 1995. 3

[184] U. Klee, T. Gehrig, J. McDonough. Kalman Filters for Time Delay of Arrival-based Source Localization. *EURASIP Journal on Applied Signal Processing*, 2006:1–15, 2006. 3.1.3

[185] University of California, Berkley, USA. Webcasts. `http://webcast.berkeley.edu/`. Last Accessed, 24 Oct. 2009. 1

[186] V. Govindaraju, D. B. Sher, R. K. Srihari, S. N. Srihari. Locating Human Faces in Newspaper Photographs. In *Computer Vision and Pattern Recognition (CVPR)*, pages 549–554, 1989. 2.2.2

[187] V. Vezhnevets, V. Sazonov, A. Andreeva. A Survey on Pixel-based Skin Color Detection Techniques. In *Int. Conf. on Computer Graphics and Vision (Graphicon)*, pages 85–92, 2003. 2.2.3

[188] W. D. Blair, Y. Bar-Shalom. Tracking Maneuvering Targets with Multiple Sensors: Does more data always mean better estimates? *IEEE Transactions on Aerospace and Electronic Systems*, 32(1):450–456, 1996. 3

[189] W. T. Chu, A. C. Warnock. Detailed Directivity of Sound Fields Around Human Talkers. Technical Report IRC-RR-104, Institute for Research in Construction, National Research Council Canada, 2002. 2.1.1, 5.2.2, 5.2.2

[190] X. R. Li, V. P. Jilkov. Survey of Maneuvering Target Tracking. Part 1: Dynamic Models. *IEEE Trans. on Aerospace and Electronic Systems*, 39(1):1333–1364, 2003. 3.1.2

[191] X. Rong Li, A. Zhao. Evaluation of Estimation Algorithms: Part 1. *IEEE Transactions on Aerospace and Electronic Systems*, 42(4), Oct. 2006. 2

[192] Y. A. Huang, J. Benesty, G. W. Elko. Source Localization. In Y. A. Huang and J. Benesty, editor, *Audio Signal Processing for Next Generation Multimedia Communication Systems*, pages 229–253. Kluwer Academic, 2004. 2.1.3

[193] Y. Bar-Shalom. *Tracking and Data Association*. Mathematics in Science and Engineering, Volume 179, Academic Press, 1988. 3.1.3

[194] Y. Bar-Shalom, X. Rong Li, T. Kirubarajan. *Estimation with Applications to Tracking and Navigation*. Wiley-Interscience, 2001. 3.1.1, 3.1.2, 3.1.2, 3.1.2, 4.1.1

[195] Y. Chen, Y. Rui. Real-Time Speaker Tracking Using Particle Filter Sensor Fusion. In *Proceedings of IEEE*, volume 92, pages 485–494, March 2004. 3.2.3, 3.2.3, 3.2.3

[196] Y. Kameda, M. Minoh. A Human Motion Estimation Method Using 3-Successive Video Frames. In *Proceedings of the International Conference on Virtual Systems and Multimedia*, pages 135–140, 1996. 2.2.1

[197] Y. Lee, R. Mersereau. A Bayesian 3D People Tracking Using Multiple Cameras and a Microphone Array. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 937–940, 2007. 3.2.3, 3.2.3, 3.2.3, 3.2.3

[198] Y. Nakamura. Geometric fusion: Minimizing uncertainty ellipsoid volumes. In M. A. Abidi and R. C. Gonzalez, editor, *Data Fusion in Robotics and Machine Intelligence*, pages 457–479. Academic Press, 1992. 5.1.1

[199] Y. T. Chan, K. C. Ho. A Simple and Efficient Estimator for Hyperbolic Location. *IEEE Transactions on Signal Processing*, 42(8), Aug. 1994. 2.1.3