HAL open science

Conjugate Mixture Models for the Modeling of Visual and Auditory Perception

Vasil Khalidov

To cite this version:

Vasil Khalidov. Conjugate Mixture Models for the Modeling of Visual and Auditory Perception. Human-Computer Interaction [cs.HC]. Université Joseph-Fourier - Grenoble I, 2010. English. NNT : . tel-00584080v2

HAL Id: tel-00584080

https://theses.hal.science/tel-00584080v2

Submitted on 12 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 7 août 2006

Présentée par

Vasil Khalidov

Thèse dirigée par **Stéphane Girard**
et codirigée par **Florence Forbes**

préparée au sein de l' **INRIA Grenoble, Laboratoire Jean Kuntzmann** et de l'école doctorale **Mathematiques, Sciences et Technologies de l'Information (Informatique)**

Modèles de Mélanges Conjugués pour la Modélisation de la Perception Visuelle et Auditive

Thèse soutenue publiquement le **18 octobre 2010**,
devant le jury composé de :

M. Anatoli Juditsky
Professeur, LJK, Université de Grenoble, France, Président
M. Sethu Vijayakumar
Professeur, University of Edinburgh, UK, Rapporteur
M. Jean-Marc Odobez
Senior researcher, Idiap Research Institute, Suisse, Rapporteur
M. Radu Horaud
Directeur de recherche, INRIA, LJK, France, Examinateur
M. Christophe Collet
Professeur, Université de Strasbourg, France, Examinateur
M. Jon Barker
Senior Researcher, University of Sheffield, UK, Examinateur
M. Stéphane Girard
Chargé de recherche, INRIA, LJK, France, Directeur de thèse
Mme. Florence Forbes
Directeur de recherche, INRIA, LJK, France, Co-Directeur de thèse

UNIVERSITY OF GRENOBLE - JOSEPH FOURIER

DOCTORAL SCHOOL MSTII

MATHEMATIQUES, SCIENCES ET TECHNOLOGIES DE L'INFORMATION, INFORMATIQUE

P H D T H E S I S

to obtain the title of

PhD of Science

of the University of Grenoble - Joseph Fourier

Specialty : APPLIED MATHEMATICS

Defended by

Vasil KHALIDOV

Conjugate Mixture Models for the Modelling of Visual and Auditory Perception

Thesis Advisors: Stéphane GIRARD and Florence FORBES

prepared at INRIA Grenoble Rhône-Alpes, MISTIS Team

defended on October 18, 2010

Jury :

Reviewers :	M. Sethu VIJAYAKUMAR	-	Professor, University of Edinburgh (UK)
	M. Jean-Marc ODOBEZ	-	Senior researcher, Idiap Research Institute (Switzerland)
Advisor :	M. Stéphane GIRARD	-	Chargé de recherche, INRIA, LJK (France)
Co-advisor :	Mme. Florence FORBES	-	Directeur de recherche, INRIA, LJK (France)
Examinators :	M. Anatoli JUDITSKY	-	Professor, LJK, University of Grenoble (France)
	M. Christophe COLLET	-	Professor, University of Strasbourg (France)
	M. Radu HORAUD	-	Directeur de recherche, INRIA, LJK (France)
	M. Jon BARKER	-	Senior researcher, University of Sheffield (UK)

This thesis is dedicated
to my dear wife Olga

Acknowledgments

First and foremost I owe my deepest gratitude to my scientific advisors – Florence Forbes, Radu Horaud and Stéphane Girard, who have inspired me and shared their profound experience during these four years. Not only their valuable advices and comments but also their personal qualities made me appreciate my work on the thesis.

I would like to express my gratitude to Anatoli Juditsky to have gently agreed to be the president of my thesis jury. I would like to thank Sethu Vijayakumar and Jean-Marc Odobez for having accepted to be my thesis reviewers, attentively read the manuscript and for all the useful comments and suggestions they provided. I'm very grateful to Christophe Collet and Jon Barker to have agreed to participate in the jury and an interest they took in my thesis.

It is a pleasure to thank my dear friends Caroline and Pierre Mahé who were always near to help, always eager to share, give a useful advice and whose creative projects left so many pleasant memories.

I'm grateful to Juliette Blanchet and Cyril Perot for their help with my first steps in France, support in discovering the country, for all the enjoyable moments that we spent together and for being kind and cordial hosts.

My warm thanks to Lamiae Azizi for sharing her charming smile and cheerful spirit, for those encouraging coffee breaks that would give a charge of energy for the whole day and all kinds of activities that made my stay in France so colourful.

Many thanks to Svetlana Artemova, Sergei Grudinin and Tanya Gordeliy for inspiring me, being caring and joyous friends that always supported me on thesis trail. I thank Karine Kuyumzhiyan for her interesting and original ideas, for being obliging and good-humoured. And, of course, dear colleagues - Laurent, Senan, Darren, Jean-Baptiste, Alexandre, Mathieu, Sophie, Elise, Anne-Françoise, Marie-Jo, Julie, Eugen, – it was a pleasure to work with you, to take lunch, discuss and laugh together.

I thank Inga Paukner-Stojkov, Inna Tsirlin, Andrei Zaharescu, Olga Nagornaya, Ramya Narasimha, Kiran Varanasi, Miles Hansard, Fabio Cuzzolin and Simone Gasparini for sharing their spirit and passion for discoveries and exploration.

Thanks to Amaël Delanoy, Antoine Letouzey, Regis Perrier, Benjamin Petit, Gaëtan, Michael Sapienza, Antoine Delaforge, Xavi Alameda-Pineda, Jordi Sanchez-Riera, Jan Cech for helping me in finding a solution to a cocktail party problem empirically. I'm grateful to Maël Bosson, Noëlle Le Delliou, Thiery Stein and Evelyne Altariba for nice coffee breaks in a cozy and carefree environment.

Thanks to all POP project participants and especially to the members of Speech and Hearing research group at the University of Sheffield. It was a pleasure to collaborate and to go out together. I kept so many pleasant memories of my stay in Sheffield.

Many thanks to Pedro Monteiro, Lionel Atty, Pascale and Julien Diener, Lionel Baboud, Thiago Bellardi, Stéphane Redon, Amaury et Stéphanie Negre, Franck Rochet, Yves Gufflet, Clara Ferley, Xiangwei Zhang, Sophie Jacquard, members of INRIA choir and members of the society “Musique pour tous” for the moments of delight that we shared playing and enjoying the music together.

I would like to warmly thank Bartha Beneddine, Myriam Etienne, Sabrina, Marie-Anne Dauphin, Patricia Oddos, Imma Presseguer, Anne Pasteur and Marie-Eve Morency for having helped me to break through all the paperwork and always keep good mood. Thanks to Aneta, Chika and Marie for their delicious coffee and enchanting smiles.

Finally, my deepest and warmest gratitude to my wife Olga to have brought beauty and inspiration, a vortex of impressions and joy, making my life colourful and full of happiness; to my brother Ildar for being eager to help me in all kinds of situations throughout these years, shared his experience, gave invaluable advices; to my parents and grand parents who always supported, wished all the best and warmed me with their love.

Contents

1 Introduction **1**

1.1 Biological View on Audio-Visual Perception 2

1.2 Overview of Computational Models for Audio-Visual Perception 4

1.3 Modelling Audio-Visual Perception: Ideas and Goals 7

1.4 Outline of the Thesis 8

2 Audio-Visual Scene Analysis Using a Head-like Device **9**

2.1 Audio-Visual Acquisition Devices 9

2.2 Binocular Visual Features 11

2.3 Binaural Hearing 14

2.4 CAVA Database 16

2.5 Discussion 21

3 Spatio-temporal Approach to Audio-Visual Calibration **23**

3.1 Multisensor Calibration Task 23

3.2 Calibration Through Multimodal Trajectory Matching 25

3.3 Trajectory Reconstruction and Parameter Estimation. 26

3.3.1 Problem Discretization and Relaxation 26

3.3.2 Hidden Trajectory Inference Using the EM Algorithm 30

3.3.3 Microphone Locations Inference Using the EM Algorithm. 34

3.3.4 Calibration Algorithm. 34

3.4 Experimental Validation 35

3.4.1 Experiments with Simulated Data 35

3.4.2 Experiments with Real Data 40

3.5 Discussion 41

4 Spatial Multimodal Clustering **43**
4.1 Unsupervised Clustering of Multimodal Data 43
4.2 Conjugate Mixture Models for Multimodal Data 46
4.3 Conjugate KP Algorithm for Clustering Multimodal Data 49
4.3.1 The Penalization Step 50
4.3.2 The Maximization Step 51
4.3.3 Generalized KP for Conjugate Mixture Models 52
4.3.4 Identifiability and Algorithm Convergence 53
4.4 Experimental Evaluation 54
4.5 Discussion 61

5 Conjugate EM Algorithm for Clustering Multimodal Data **65**
5.1 Conjugate EM Algorithm for Clustering Multimodal Data 65
5.1.1 The Expectation Step 66
5.1.2 The Maximization Step 67
5.1.3 Generalized EM for Conjugate Mixture Models 68
5.1.4 Analysis of *Local Search* Procedure 69
5.1.5 Global Search and the *Choose* Procedure 73
5.2 Clustering Using Auditory and Visual Data 74
5.3 Experimental Validation 76
5.3.1 Experiments with Simulated Data 76
5.3.2 Experiments with Real Data 81
5.4 Discussion 89

6 Algorithm Initialization and Model Selection **93**
6.1 Multimodal Cluster Initialization and Model Selection 93
6.2 Initialization 95
6.2.1 EM Initialization for Conjugate Gaussian Mixture Models 96
6.2.2 The *Initialize* Procedure 97
6.2.3 Experimental Validation 98
6.3 Model Selection 103
6.3.1 The *Select* Procedure 107
6.3.2 Weak Consistency of Multimodal Information Criteria 108
6.3.3 Experimental Validation 112
6.4 Discussion 112

7 Spatio-temporal Multimodal Clustering **115**

7.1 Multimodal Multiobject Tracking 115

7.2 Conjugate Filtering for Multimodal Multiobject Tracking 116

7.3 Experimental Results 119

7.4 Discussion 120

8 Conclusion **123**

8.1 Main Contributions 123

8.2 Future Work 125

A Appendix **127**

A.1 Manifold Sampling for the ITD function Pre-image. 127

A.2 Parameter Inference for Student-t Mixtures. 129

Bibliography **133**

CHAPTER 1

Introduction

Sommaire

1.1 Biological View on Audio-Visual Perception 2
1.2 Overview of Computational Models for Audio-Visual Perception 4
1.3 Modelling Audio-Visual Perception: Ideas and Goals 7
1.4 Outline of the Thesis 8

The ambient world is percieved in a unified manner independently of the environment – whether it is in the office, at home, in the street, in the parc, etc. We are not affected by the strongly varying conditions of lighting, acoustics and noise, nor do we pay much attention to the head position and motion at every time instant. We deal with multimodal percepts of the events – an image of a ball jumping off the ground is associated with the hitting sound, distant barking is immediately connected with the image of a running dog. This interpretation may be considered as an audio-visual analysis of the scene. The goal of my research is to develop a computational statistical framework to perform low-level binding of descriptions from different modalities that correspond to the same objects.

The human ability to efficiently extract biologically meaningful events based on independent information from different senses is impressive. Evolution accounted for the development of sophisticated sensory organs linked to specialized brain regions that allow to detect and identify various events or objects of interest. Each of them gives optimal performance under different conditions. Let us consider two examples of a scene shown on Figure 1.1. The first image shows a typical indoor environment with several persons talking in the room. In total five persons are present, four of them are visible and three talk, one of the talking persons is not visible. The auditory and visual signals are significantly corrupted. Numerous occlusions, ambiguous colour information and the fact that the objects are distant makes the extraction of meaningful events difficult. At the same time, such an environment is highly reverberant and allows for shadowing effects, the auditory activity contains interfering sounds – footsteps and motion noise, as well as simultaneous speech of several persons. The more so, auditory and visual scene interpretations are *contradictory* – visual scene contains four persons, while auditory scene contains three and none of the modality scenes is a subset of the other. However, the human brain succeeds in *integrating* information from the two senses and forms correct multimodal percepts of the scene.

The second image shows a dog relay team somewhere in the northern snowy plains. This time the events received from the sensory systems are different, speech and gesture

(a) indoors environment (b) outdoors environment

Figure 1.1: Examples of audio-visual scenes: indoors environment with five persons, four of them are visible and three are speaking, one is to the left, outside the field of view; outdoors environment with a dog relay team, five dogs are present, two of them whine.

cues are absent. Both auditory and visual systems on their own provide weak cues that are not as informative, as in the previous example. But again, the human brain constructs multimodal percepts that allow us to have a stable scene representation.

Therefore, the major role in human audio-visual perception is played by an integrative process that combines information from different senses. This human brain capability is of primary importance for forming unified multimodal representations. Findings in research on neurobiology confirm these ideas and provide more and more evidence that the integrated product reveals more about the nature of the external event leading to faster and better perception.

1.1 Biological View on Audio-Visual Perception

The way the human brain performs audio-visual (AV) integration is amazingly efficient. However, no matter how natural it seems to be in everyday life, the mechanisms leading to such a performance are still a subject of intensive research [Kadunce 2001, Meyer 2001, Spence 2004, Stanford 2007, Stein 2008]. The brain faces a complex task of integrating information that possesses different physical properties. Moreover, the sound and light emitted from a sensory source travel at different speeds and therefore arrive at different times at the sensory organs. The neural processing delay between the auditory and visual systems should also be accounted for. This makes the AV integration problem challenging.

The integration of auditory and visual signals is most commonly assessed by comparing responses to a cross-modal stimulus with those to visual and auditory stimuli alone. The measurements can be performed in various experiments using response speed, such as saccadic reaction times [Colonius 2001], performance improvement, for example in motion prediction task [Hofbauer 2004], or directly on individual multisensory neurons [Stein 2008]. An important consequence of AV integration is *multisensory enhance-*

ment which refers to the phenomenon when the neural response to a multimodal event occurs to be more vigorous than to any of its inputs [Stein 1993, Anastasio 2000]. In certain cases the augmentation can even be superadditive [Stanford 2007], meaning that the response exceeds the sum of its inputs.

Bayesian models of sensory cue integration have been proposed recently in order to account for multisensory enhancement [Anastasio 2000, Knill 2007]. These approaches allow to model the characteristic property of the phenomenon, *inverse effectiveness*, which states that combinations of weak unimodal responses can produce large amount of enhancement.

Certain conditions were found to improve AV integration. Co-localized and co-incident auditory and visual stimuli lead to more effective integration, as shown by single-unit studies [Stein 1988] and detection-based experiments [Meyer 2005]. Sometimes even weaker conditions with co-incident stimuli originate from different points in space are sufficient to ensure integration [Kadunce 2001]. In this case only the overlapping of receptive fields in the superior colliculus is required. Even more complex integration strategies based on stimulus congruence were discovered in cortical multisensory representations [Stein 2008]. The capability of a human brain to perform audio-visual integration under relatively weak conditions gives rise to cross-modal illusions, such as McGurk Effect [McGurk 1976] and Ventriloquism Effect [Howard 1966]; their dependency on spatial, temporal and cognitive factors has been investigated [Lewald 2003].

AV integration is responsible for creating unified percepts, which raises some non-trivial issues and requirements:

- **Processing complexity**: a percept inherits all the complexities related to neural processing in each individual modality;
- **Percept richness**: simultaneous inference of assignment labels and object parameters allows to avoid exponentially hard binding problems;
- **Binding**: appropriate input data should be chosen for binding – these selection processes that are not yet well understood [Stein 2008];
- **Weighting**: binding should be performed based on some strategy, which is often accomplished by weighting the various cues based on the amount of information they are likely to provide [Burr 2006, Knill 2007];
- **Invariance**: a common percept should not be dependent on the current state of the sensory systems. There is a need for multisensory spatial representations and for the means to align receptive fields in case of state changes [Pouget 2002b];

This brief outline shows that there has been extensive research effort aiming at understanding the mechanisms of multisensory integration. The field grows rapidly in both the number and the variety of investigations on multisensory phenomena.

1.2 Overview of Computational Models for Audio-Visual Perception

Advances in research on biological principles of audio-visual (AV) integration influenced the development of computational models. Originally, in most systems that handled multimodal data, audio and visual inputs were first processed by modality-specific subsystems, whose outputs were subsequently combined [Heckmann 2002, Garg 2003]. The performance of such procedures in realistic situations is limited in the following ways. Confusion may arise from factors such as background auditory noise, presence of both speech and non-speech multiple audio sources, acoustic reverberations, rapid changes in the visual appearance of an object, varying illumination conditions, visual occlusions, and so forth. The different attempts that have been made to increase robustness are based on the observation that improved object detection and localization can be achieved by integrating auditory and visual information. The major reason for this was to benefit from multisensory enhancement: weak stimuli from one modality can be potentially reinforced by the other modality. Simultaneous AV processing is particularly critical in complex situations such as the ones encountered when distant sensors (microphones and cameras) are used within realistic AV scenarios. This means that the problems that arise when trying to understand AV integration strategies in human brain should be resolved in a computational model.

The major questions that need to be answered in order to develop a computational audio-visual integration model are:

- Which A and V features to select in order to account for an optimal compromise between single- and cross-modality?
- In which mathematical space the AV data fusion should be performed?
- Once A and V features are detected, which of them should be bound together to form an analogue of AV percept?
- Which strategy could be used to perform the binding?
- How to ensure consistency between modalities?

Different solutions for computational models can be found in the literature. Below we provide an overview of the existing approaches to audio-visual integration.

Features to be selected. The dilemma is the following. On the one hand one wants to extract rich and expressive features which would provide informative event descriptions. This can lead to high-level event detection which is hard to perform, which would be rarely available in noisy conditions, and hard to integrate with other cues due to the event specificity. At the same time, for robust and continuous perception one would like to constantly receive a flow of low-level cues that can be extracted even in noisy conditions. But then these features may occur to be too elementary to provide any significant information. As

usual, the best option is somewhere in the middle: low-level descriptions are processed by *bottom-up* feature detectors to extract meaningful representations which are afterwards combined into high-level patterns by *top-down* processes.

The features that are available through bottom-up detectors depend mostly on the hardware setup. Some methods rely on complex audio hardware such as microphone arrays that are mutually calibrated [Checka 2004, Chen 2004, Perez 2004, Nickel 2005, Gatica-Perez 2007]. This yields an approximate unimodal spatial localization of each audio source. Reducing the number of microphones leads to decrease in localization precision. Two microphones setup [Beal 2003, Kushal 2006, Hospedales 2007, Hospedales 2008] resembles the most the real head, but can only provide approximate localization using binaural localization cues [Wang 2006], such as interaural time difference (ITD), interaural level difference (ILD), interaural phase difference (IPD). A single microphone is simpler to set up, but it cannot, on its own, provide spatial localization.

Several calibrated cameras were used in [Checka 2004, Nickel 2005, Gatica-Perez 2007] that can provide 3D object location estimates. Though in most computational models the 3D scene is further projected onto camera planes to work with a 2D representation. As in the case of a single camera, this can only provide approximate localization. Note that two distinct AV objects may project to nearby locations in an image. The more distant object will be partially or totally occluded in this case, and so purely 2D visual information is not sufficient to solve the localization problem. In this respect it is advantageous to use a pair of stereoscopic cameras. It allows to increase the field of view and at the same time to extract *depth* information through the computation of binocular disparities.

Various modality-specific features can be extracted like spectral auditory features [Wang 2006], photometric visual features such as colour models [Perez 2004], structural templates [Gatica-Perez 2007], etc. These cues are typically used as descriptors for data clusters.

Choosing a fusion space. There are several possibilities. In contrast to the fusion of previous independent processing of each modality [Heckmann 2002], the integration could occur at the feature level. In this case audio and video features are concatenated into larger feature-vectors, which are then processed by a single algorithm. However, owing to the very different physical natures of audio and visual stimuli, direct integration is not straightforward. For example, there is no obvious way to associate dense visual maps with sparse sound sources.

The fusion space should be defined so as to contain common information from auditory and visual features. The most popular choice is the image space [Beal 2003, Kushal 2006, Hospedales 2007, Hospedales 2008, Gatica-Perez 2007]. Though this is usually done under the assumption that there are no occlusions or by considering them as a special case [Gatica-Perez 2007].

We argue here that the fusion space plays an important role in the integration process. The real-world AV data tends to be influenced by the structure of the 3D environment in

which it was generated. Thus we the best choice would be to perform fusion in the physical 3D space.

Feature association. The general association problem that finds the optimal matching within the two sets of data is NP-hard and cannot be easily solved. Certain applications admit simple association strategies based on co-incidence of the cues [Hazen 2004]. However, the conditions under which such a binding can be performed are not common.

Another opportunity is to impose multimodal patterns and thus force association of certain features through a supervised learning strategy [Zeng 2007]. This method is suitable for recognition tasks, but cannot be applied in general tracking scenarious.

Most of the computational models use object-related association models. The essential role here is played by the chosen fusion space. To gain more spatial resolution and increase separation between clusters it is important to keep the dimensionality of the fusion space without projecting the data.

Binding strategies. We identify two major directions depending on the type of *synchrony* being used for binding. The first one focuses on *spatial synchrony* and implies combining those signals that were observed at a given time, or through a short period of time, and correspond to the same location. Generative probabilistic models in [Beal 2003] and [Kushal 2006] for the problem of single speaker tracking achieve this by introducing dependencies of both auditory and visual observations on 2D locations, i.e., in the image plane. The same idea is used in [Hospedales 2007, Hospedales 2008] for the multi-speaker case. The explicit dependency on the source location in these models can be generalized by the use of particle filters. Such approaches have been used for the task of single speaker tracking [Zotkin 2002, Vermaak 2001, Perez 2004, Chen 2004, Nickel 2005] and multiple speaker tracking [Checka 2004, Gatica-Perez 2007, Chen 2004, Bernardin 2007, Brunelli 2007]. In the latter case the parameter space grows exponentially as the number of speakers increases, so efficient sampling procedures may be needed, to keep the problem tractable [Gatica-Perez 2007, Chen 2004].

The second direction focuses on *temporal synchrony*. It efficiently generalizes the previous approach by making no a priori assumption on AV object location. Signals from different modalities are grouped if their evolution is correlated through time. The work in [Fisher III 2004] shows how the principles of information theory can be used to select those features from different modalities that correspond to the same object. Although the setup consists of a single camera and a single microphone and no special signal processing is used, the model is capable of selecting the speaker among several persons that were visible. Another example of this strategy is described in [Barzelay 2007], where matching is performed on the basis of audio and video onsets (times at which sound/motion begins). This model has been shown to work with multiple, as well as with individual, AV objects. Most of these approaches are, however, non-parametric and highly dependent on the choice of appropriate features. Moreover they usually require either learning or ad-hoc tuning of

quantities such as window sizes and temporal resolution. They tend to be quite sensitive to artifacts, and may require careful implementation.

Consistency between modalities. The binding strategy is the core principle of the multimodal integration. However, to show meaningful behaviour it should comply the consistency principle. For spatial synchrony, for instance, the locations to which both modalities are bound should be the same. Thus one should verify that auditory and visual devices used in the setup are calibrated with respect to each other.

Smart room environments [Wilson 2001, Checka 2004, Gatica-Perez 2007, Nickel 2005] require elaborate and complex calibration techniques to align the devices. Displacing one of them would require recalibration of the whole setup. This was the reason for the development of fast and approximate calibration in [Gatica-Perez 2007]. At the same time, a head-like device, while being able to perform binding in the 3D space, offers facilities for fast and exact calibration, and is potentially capable of performing self-calibration.

1.3 Modelling Audio-Visual Perception: Ideas and Goals

Our device, described later in Chapter 2 comprises a pair of stereoscopic cameras and a pair of microphones. Having analyzed major advantages and drawbacks of the existing approaches, we set a number of requirements for the multimodal framework desirable for multiple object tracking and define the following goals:

- **Fusion in the 3D space**: our device allows for 3D scene reconstruction, it is important to reinforce the binding strategy and consider the multimodal integration task in the 3D space;
- **Features extensibility**: the multimodal integration framework should allow to use modality-specific high-level features even if the integration is performed on low-level cues;
- **Modality weighting**: weights for observations should be adjusted automatically based on the amount of information provided by each modality;
- **Multimodal enhancement**: the multimodal framework should enable multimodal enhancement to reinforce weak stimuli from one modality with the stimuli from the other modality;
- **Robust multimodal tracking**: the multimodal framework should be able to perform robust multimodal tracking even when the objects become invisible for a short period of time;
- **Calibration**: the hardware device should allow fast, efficient and precise calibration;

- **Evaluation**: a set of audio-visual scenarios should be developed to mimic natural environments and conditions to evaluate the multimodal multiobject tracking framework;
- **Theoretical integrity**: the proposed multimodal framework should be well-founded, convergence properties and consistency should be verified.

1.4 Outline of the Thesis

The thesis comprises 8 chapters. After this introduction, Chapter 2 presents the hardware devices used in the experiments as well as the feature extraction algorithms used to obtain the data. It introduces some functional models used throughout all the thesis. The database of realistic audio-visual scenarios is described (CAVA database). It was designed and acquired as a part of this work. It is used to validate the results of Chapters 4–7. This database part of Chapter 2 is based on my publication [Arnaud 2008].

Chapter 3 describes the first original contribution of this thesis – it is devoted to the audio-visual head-like device calibration method. It presents the theoretical framework as well as a simulated and real data experimental validation. This chapter is self-consistent and can be read separately.

Chapter 4 contains the second original contribution of this thesis – the conjugate clustering framework and the family of associated optimization algorithms that I developed to perform audio-visual integration. The theoretical framework is introduced, the properties of the algorithms are discussed and verified on simulated data. The chapter is based on my publications [Khalidov 2008b, Khalidov 2010].

Chapter 5 presents the third original contribution of this thesis – it considers one instance of the family of conjugate clustering algorithms, and shows that it can be significantly accelerated and gain attractive theoretical properties. The theoretical results are verified on simulated data and on the CAVA database. The chapter follows my publications [Khalidov 2008a, Khalidov 2010].

Chapter 6 describes the fourth original contribution of this thesis – it introduces the multimodal initialization and model selection procedures that improve the performance of the optimization algorithms considered in previous chapters and are shown to possess the same theoretical properties as their single modality counterparts. Again, the results are verified on the simulated data and CAVA database.

The last original contribution of this thesis is given in Chapter 7 – it combines the developed multimodal clustering framework with some known tracking techniques to perform multimodal multiobject tracking. It shows that our framework can be naturally extended with an object dynamics model. The performance is demonstrated on the CAVA database scenarios.

Finally, Chapter 8 concludes the thesis and discusses future perspectives.

CHAPTER 2

Audio-Visual Scene Analysis Using a Head-like Device

Sommaire

2.1 Audio-Visual Acquisition Devices 9
2.2 Binocular Visual Features 11
2.3 Binaural Hearing 14
2.4 CAVA Database 16
2.5 Discussion 21

In this Chapter we discuss the task of human-centered computational audio-visual (AV) scene analysis. Different hardware configurations that aim at modelling a human perceiver are presented, all of them were used in the experiments. Features that are general enough to be applied to any AV object or scene and informative enough to better suit for the task of AV integration are proposed. The novel database, designed to investigate binaural/binocular fusion strategies of a human and to validate and compare the models of an AV perceiver, is presented.

2.1 Audio-Visual Acquisition Devices

The idea behind the robot head hardware configurations was to create a device that would record data from the perspective of a person, i.e. would try to capture what a person would see and hear while being in natural audio-visual (AV) environment. The three configurations that were used in the experiments are depicted in Figure 2.1, below we give their detailed descriptions.

Figure 2.1(a) shows the device employed for CAVA database acquisition (see Section 2.4). The Brüel & Kjær (B & K) Head and Torso Simulator type 4128C was used to provide a realistic reproduction of the acoustical properties of an average adult human. Two B & K microphones type 4190 (1/2-inch, free-field) are fitted into its ears to record binaural data. The audio signals are then treated by B & K type 2669 1/2-inch preamplifiers and then by B & K type 2690-OS2 Nexus conditional amplifier. Finally, the analog-to-digital (A/D) conversion is performed by Behringer Ultragain Pro-8 Digital ADA8000 A/D and D/A converter. A pair of Point Grey Flea cameras with 6mm Fujinon lenses were fixed to

(a) (b) (c)

Figure 2.1: Audio-visual acquisition devices used in experiments. (a) Mannequin configuration; (b) POPEye Robot Head, high-specification audio configuration; (c) POPEye Robot Head, low-specification audio configuration.

the front of a helmet placed on the mannequin's head to record binocular data. This is a high specification device that was intended for acquiring AV streams that would resemble those obtained by human eyes and ears.

Another type of device shown in Figure 2.1(b) was developed within the European project POP (Perception on Purpose, FP6-IST-027268) [1] by Computer and Robot Vision Laboratory members[2], University of Coimbra, Portugal. The *POPEye Robot Head* uses the same cameras and B & K Head Simulator together with B & K microphones, preamplifiers and the conditional amplifiers, and A/D converter as described in the previous case. They are mounted onto a robot platform with four rotational degrees of freedom, namely neck pan, neck tilt and eyes vergence. The control is performed through four brushed DC motors from Harmonic Drive: one motor PMA-11A-100-01-E500ML for neck pan, one motor PMA-8A-100-01-E500ML for neck tilt and two motors PMA-5A-80-01-E512ML for eyes vergence. They produce much less noise than brushless AC motors, which is essential for experiments involving auditory analysis. The platform allows for adjustment of baseline (distance between the cameras) and camera positions along their optical axes, so that the properties of the configuration can be changed to approach those of human visual system. This device can be controlled in real time and is capable of modelling active perception.

The third device shown in Figure 2.1(c) is a version of the POPEye robot that has a simple polystyrene head and Soundman OKM binaural microphones connected to a Soundman amplifier instead of the B & K head simulator and the B & K audio acquisition system. The summary on the three configurations is given in Table 2.1.

To improve correspondance between the left and right images acquired by the two cameras, the video streams are synchronized by means of an external trigger. Also different calibrations are required to use the data obtained from an AV device. Firstly, the intrinsic and extrinsic camera parameters (see Section 2.2) are estimated through visual

[1] http://perception.inrialpes.fr/POP/
[2] http://labvis.isr.uc.pt/

<table>
<tr><th></th><th>Mannequin</th><th>Robot Head, HSA</th><th>Robot Head, LSA</th></tr>
<tr><td>Audio system</td><td colspan="2">B & K 4128C Head Simulator,
B & K 4190 microphones,
B & K 2669 preamplifiers,
B & K 2690-OS2 conditional amplifier</td><td>Polystyrene head,
Soundman OKM
binaural microphones,
Soundman amplifier</td></tr>
<tr><td>Video System</td><td colspan="3">A pair of Point Grey Flea cameras, external trigger</td></tr>
<tr><td>Platform</td><td>B & K 4128C Torso Simulator</td><td colspan="2">POPEye robot platform</td></tr>
</table>

Table 2.1: Robot configurations. Three columns correspond to three versions of the experimental setup, namely a mannequin, a robot head with high-specification auditory system (HSA) and a robot head with low-specification auditory system (LSA). Each line shows different options for a particular system.

calibration procedure. In our experiments we used the one provided by the image processing library OpenCV[3] with chessboard as a calibration rig. Secondly, the audio calibration is needed to ascertain the exact amplification in the left and right channels. This was done through attaching B & K pure tone generator to each of the microphones and calculating the corresponding normalization factor. Finally, to perform AV integration the AV calibration is required. It consists in determining the microphone coordinates in camera frame. AV calibration method was developed as a part of the current Thesis and is presented in Chapter 3.

2.2 Binocular Visual Features

We would like to extract visual features that would be general enough (not specific to particular object types) and at the same time sufficiently informative to perform AV integration. In this Section we present the technique used to extract and reconstruct in the scene such features called “interest points”.

[3] http://www.intel.com/technology/computing/opencv

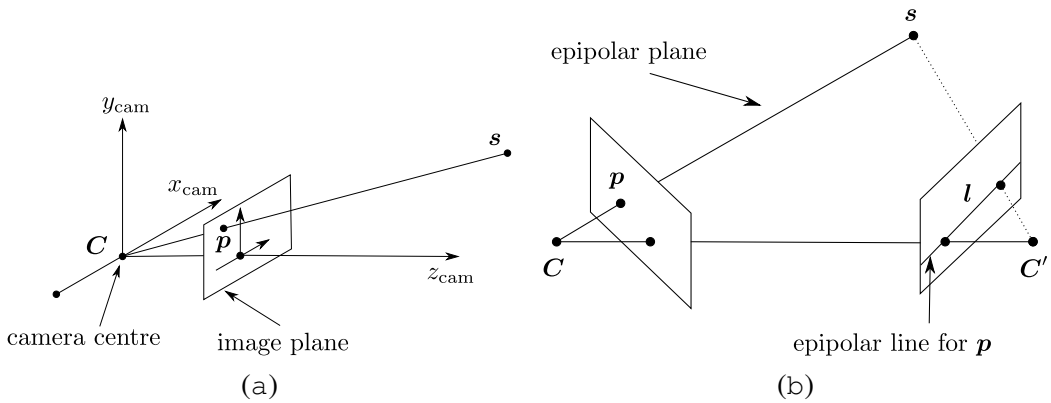


Figure 2.2: Binocular geometry. (a) Basic pinhole camera model. $\boldsymbol{C}$ is the camera centre, $(x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}})$ is the camera frame, $\boldsymbol{s}$ is a point in 3D and $\boldsymbol{p}$ is its projection on the image plane. (b) Point correspondence. The two cameras are indicated by their centers $\boldsymbol{C}$ and $\boldsymbol{C}'$ and image planes. An image point $\boldsymbol{p}$ back-projects to a ray in 3D space defined by $\boldsymbol{C}$ and $\boldsymbol{p}$. This ray is imaged as a line $\boldsymbol{l}$ in the second view.

The visual data is gathered using a pair of stereoscopic cameras, i.e. binocular vision. We assume the *basic pinhole* camera model [Hartley 2003] that establishes a projective mapping

$$\boldsymbol{s} = (x, y, z)^\top \mapsto \boldsymbol{p} = (\boldsymbol{p}_1\boldsymbol{s}/\boldsymbol{p}_3\boldsymbol{s},\ \boldsymbol{p}_2\boldsymbol{s}/\boldsymbol{p}_3\boldsymbol{s}) \tag{2.1}$$

of a point $\boldsymbol{s}$ in 3D onto the image plane. We denoted $\boldsymbol{p}_i$ the ith line of the camera matrix $\mathbf{P} = \mathbf{AR}(\ \mathbb{I}\ |\ -\mathbf{C})$, where $\mathbf{A} = \begin{pmatrix} \alpha_u & \gamma & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix}$ is the matrix of camera intrinsic parameters and $\mathbf{R}$ and $\mathbf{C}$ are the rotation and translation of camera frame respectively with respect to some reference frame (extrinsic parameters); $\mathbb{I}$ is the 3x3 identity matrix. For exact meaning of values in $\mathbf{A}$ matrix we refer to [Hartley 2003]. The extrinsic and intrinsic parameters of a camera are obtained through camera calibration, as mentioned before in Section 2.1. Schematic representation of the basic pinhole camera model is given in Figure 2.2a.

Under the pinhole camera model, image points are represented as rays of light intersecting the image plane on a line running through the camera center. Given a pair of cameras, $\boldsymbol{C}$ and $\boldsymbol{C}'$, and a point $\boldsymbol{p}$ in camera $\boldsymbol{C}$, the location $\boldsymbol{p}'$ of the same point in the other camera can be constrained to an *epipolar line* $\boldsymbol{l}$, as shown in Figure 2.2b. Thus for every scene point $\boldsymbol{s}$ one can introduce the notion of *epipolar disparity* d as a displacement of an image point along the corresponding epipolar line [Hansard 2008]. For a rectified camera pair [Hartley 2003] an invertible function $\mathcal{F} : \mathbb{R}^3 \to \mathbb{R}^3$ can be defined, that maps a scene point $\boldsymbol{s} = (x, y, z)^\top$ onto a cyclopean image point $\boldsymbol{f} = (u, v, d)^\top$ corresponding to a 2D image location (u, v) and to an associated binocular disparity d:

$$\mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{B}{z}\right)^\top \quad \text{and} \quad \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{Bu}{d}, \frac{Bv}{d}, \frac{B}{d}\right)^\top, \tag{2.2}$$

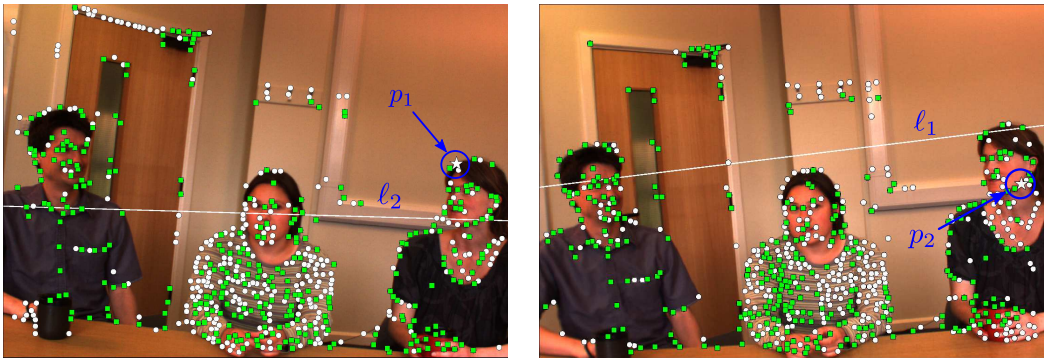


Figure 2.3: Visual observations on the left and right camera images. White circles depict the "interest points", coloured squares show those of them that are matched to some point from the other image. The epipolar lines correspond to a point marked by a star in the opposite image.

where B is the baseline length (distance between camera centres $\boldsymbol{C}$ and $\boldsymbol{C}'$) measured in focal distances of a camera. Without loss of generality we further scale the disparity component and let $B = 1$ to use the following feature space mapping

$$\mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{1}{z}\right)^\top \quad \text{and} \quad \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{u}{d}, \frac{v}{d}, \frac{1}{d}\right)^\top . \tag{2.3}$$

This model can be easily generalized from a rectified camera pair configuration to more complex binocular geometries [Hansard 2007, Hansard 2008]. We use a sensor-centered coordinate system to represent the object locations.

Visual observations $\mathbf{f} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M\}$ in our experiments are obtained as follows. First we detect points of interest (POI) in both the left and right images. Second we perform stereo matching such that a disparity value is associated with each matched point.

In practice we used the POI detector described in [Harris 1988]. This detector is known to have high repeatability in the presence of texture and to be photometric invariant. We analyse each image point detected this way and we select those points associated with a significant motion pattern. Motion patterns are obtained in a straightforward manner. A temporal intensity variance σ_t is estimated at each POI. Assuming stable lighting conditions, the POI belongs to a static scene object if its temporal intensity variance is low and non-zero due to a camera noise only. For image points belonging to a dynamic scene object, the local variance is higher and depends on the texture of the moving object and on the motion speed. In our experiments, we estimated the local temporal intensity variance σ_t at each POI, from a collection of 5 consecutive frames. The point is labelled "motion" if $\sigma_t > 5$ (for 8-bit gray-scale images), otherwise it is labelled as "static". The motion-labelled points are then matched and the associated disparities are estimated using standard stereo methods. The features we use are obtained with the method described in [Hansard 2007]. Examples are shown on Figure 2.3. Alternatively, we could have used the spatiotemporal point detector described in [Laptev 2005]. This method is designed to

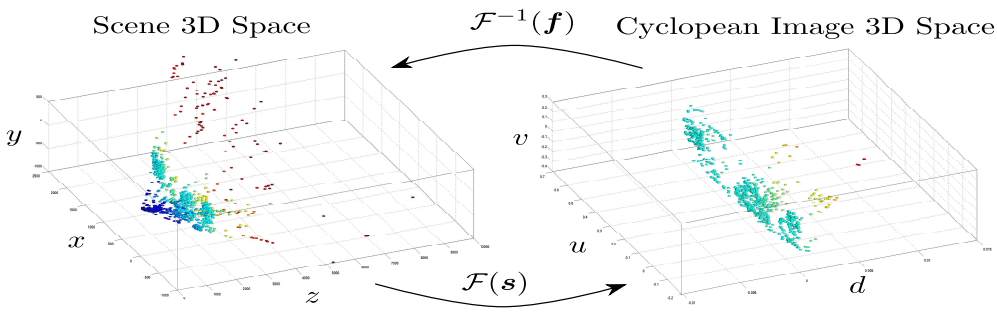


Figure 2.4: Visual observations **f** in the Cyclopean image space (on the right) and their reconstructed correspondances in the scene space (on the left), obtained through applying $\mathcal{F}^{-1}$. Point colour represents the d or z coordinate in Cyclopean image space or scene space respectively.

detect points in a video stream having large local variance in both the spatial and temporal domains, thus representing abrupt events in the stream. However, such points are quite rare in data flows we work with.

An example of visual observation set for a visual scene containing three persons is given in Figure 2.4. The points **f** in the Cyclopean image space (on the right) are obtained through stereo matching of POI in the left and right images. Their reconstruction **s** in the scene space (on the left) can be found through applying the inverse mapping $\mathcal{F}^{-1}$. The point colours are computed from the d or z coordinates in Cyclopean image space or scene space respectively.

The implementation of the visual feature detection algorithm was kindly provided by Miles Hansard, a member of PERCEPTION team[4] at INRIA research institute, France.

2.3 Binaural Hearing

As in the case with binocular vision, we would like the auditory features to be informative and at the same time general enough. This Section is devoted to techniques used to extract the ITD features that fulfil mentioned requirements.

The auditory data is gathered using a pair of microphones, i.e. binaural hearing. A sound emitted at time instant t from a source located at a scene point $\boldsymbol{s} = (x, y, z)^\top$ would be acquired by the left and right microphones located at M_ℓ and M_r at time $t_\ell = t + \frac{1}{c}\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\|$ and $t_\mathrm{r} = t + \frac{1}{c}\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|$ respectively. As soon as the value of t is not known in advance, a good cue for the sound source location would be the time difference $t_\ell - t_\mathrm{r}$. It is called *interaural time difference* (ITD) and plays the role of disparity for binaural hearing. ITD values are widely used by auditory scene analysis methods [Wang 2006]. We

[4] http://perception.inrialpes.fr/

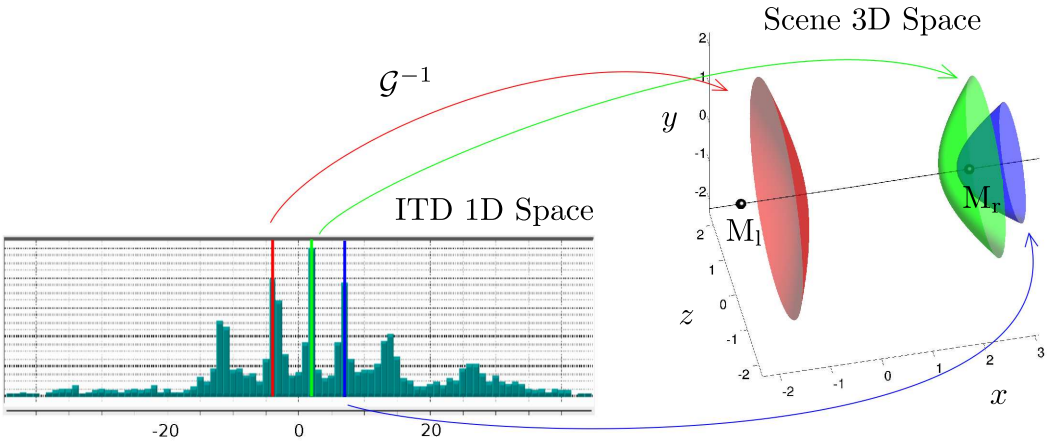


Figure 2.5: Auditory observations $\boldsymbol{g}$ in the ITD space shown as a histogramm (on the left). Three peaks are marked with coloured bars and mapped to the corresponding surfaces in the scene space (on the right), obtained through applying $\mathcal{G}^{-1}$.

introduce the function $\mathcal{G} : \mathbb{R}^3 \to \mathbb{R}$ that maps $\boldsymbol{s} = (x, y, z)^\top$ onto a 1D ITD observation:

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big), \tag{2.4}$$

where $c \approx 343$m/s is the sound speed. Unlike visual observations, an ITD value does not correspond to a unique point in the scene space, but rather to a whole surface of points. In fact, each isosurface defined by (2.4) is represented by one sheet of a two-sheet hyperboloid in 3D, as shown in Figure 2.5. Hence, each audio observation g constrains the location of the auditory source to lie onto a 2D manifold.

Auditory observations $\boldsymbol{g} = \{g_1, \ldots, g_K\}$ in our experiments are obtained using the ITD calculation method described in [Christensen 2007]. First, the left and right microphone signals are processed by a filter bank that separates them into different frequency bands. Second, cross-correlogramm is computed for every frequency band, the results are integrated and analyzed to obtain an ITD value.

In practice we used a bank of biologically inspired *gammatone filters* [Patterson 1992] that model cochlea in the inner ear of a human. The impulse response function of a filter is given by

$$h(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), \tag{2.5}$$

where a is the amplitude, n is the filter order, b is the filter's bandwidth, f_c is the filter centre frequency and ϕ is the phase. It was shown [Patterson 1992] that the choice of $n = 4$ and $b = 1.019 \cdot \mathrm{ERB}$ provides an excellent fit to the human auditory filter shapes, where

$$\mathrm{ERB} = 24.7(4.37 \cdot 10^{-3} f_c + 1) \tag{2.6}$$

is the equivalent rectangular bandwidth (ERB) model proposed by [Glasberg 1990]. Several efficient implementations of the gammatone filterbank are available [Cooke 1993,

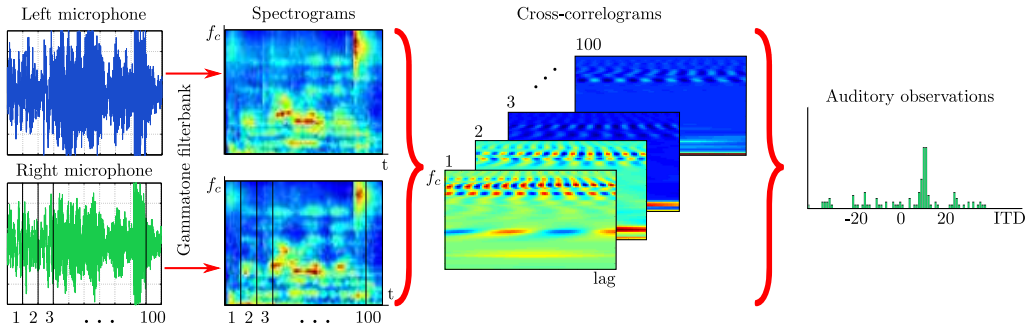


Figure 2.6: Summary of the audio processing system. Signals from the left and right microphones are treated by the gammatone filterbank to obtain time-frequency representations. Spectrograms are then split into 10ms frames (we show 1s fragment of the recording that contains 100 frames) that are cross-correlated to obtain ITD observations, one for each frame. The final observation set $\boldsymbol{g}$ is shown as a histogram of ITD values.

Slaney 1993]. We use the Martin Cooke's digital filters [Cooke 1993] based on impulse invariance transformation[5]. The 64 frequency channels of the bank are uniformly spaced from 50Hz to 8000Hz.

The output of the filterbank for the left and right microphones is split into intervals of 10 ms that are further used to generate cross-correlogramms. The ITD observation is then found as a maximum value of weighted sum of cross-correlogramms for different channels. The processing steps are summarised in Figure 2.6.

Alternative approaches to ITD computation exist, notably [Faller 2004, Mandel 2007]. We've chosen the one proposed by [Christensen 2007] as soon as this method could be extended to the multispeaker case through time-frequency fragment segregation and analysis. The real-time implementation of the algorithm was kindly provided by Heidi Christensen from the Speech and Hearing Laboratory[6] the University of Sheffield, UK.

2.4 CAVA Database

To investigate binaural/binocular fusion strategies of a human and to validate and compare the models of an audio-visual (AV) perceiver, a common data set is required that would satisfy the following conditions:

- data is acquired by a **human head-like device** comprising a pair of calibrated cameras and a pair of calibrated microphones;

[5] http://www.dcs.shef.ac.uk/~ning/resources/gammatone/
[6] http://www.dcs.shef.ac.uk/spandh/

- data is acquired in **natural environment**, so that the recordings contain visual occlusions, lighting changes, auditory reverberations and ambient sounds;
- data contains scenarios of **various complexity**: stationary scenes and simple tracking tasks, as well as complex dynamic scenes.

There already exists a number of databases used by the audio-visual (AV) research community. They can be roughly divided into three groups.

Face/speech-oriented multimodal databases, such as AV-TIMIT [Hazen 2004], GRID [GRI], M2VTS [M2V], XM2VTSDB [Messer 1999], BANCA [Bailly-Bailliére 2003], CUAVE [Patterson 2002], GEMEP [GEM] are typically acquired with one fixed camera and one fixed microphone and include individual speakers or speaker pairs. As actors are recorded in the near field of the sensors, thus these databases are primarily destined for AV verification, AV speech recognition and affect recognition tasks.

Meeting-oriented multimodal databases including AMI [AMI], M4 [McCowan 2003], CHIL [Mostefa 2008], NIST [Michel 2007], VACE [Chen 2005] meeting corpora are acquired using smart room environments comprising distributed camera systems, microphone arrays, individual lapel microphones. The scenes are predominantly stationary and the main accent in the recordings is put on actor interactions and postures.

Finally, *dynamic scene* multimodal corpora AV16 [Lathoud 2004], CHIL [Mostefa 2008] acquired with smart room environments are destined for single/multiple person tracking.

None of the existing databases concerns the challenging task of human-centered audio-visual (AV) scene analysis and thus they do not satisfy the three formulated conditions. In fact, very few studies limit the sensory input to mimic that of humans both in terms of the number of input channels, and especially in terms of the position and dynamics of the perceiver.

The *CAVA database*[7] [Arnaud 2008] was recorded within the POP project by two partners - the University of Sheffield, UK and INRIA, France. The goal was to provide common base for development, verification and comparison of algorithms destined for computational audio-visual analysis (CAVA) by means of a human head-like device. It comprises about 50 sessions of 20 seconds to 3 minutes duration each with varying degrees of visual and auditory complexity.

The entire CAVA corpus was acquired with the mannequin device (see Figure 2.1(a), Table 2.1) in a 7m×5m office-like room with carpets, painted walls and board ceilings. Figure 2.7 shows four photographs from the room depicting parts of the setup and scenario sessions. In addition to the fluorescent lamps in the room, two 500 watt studio lamps with light reflectors were used. To minimise unwanted acoustic noise, all computers were positioned outside the room, and all wires run under a door, which was closed during the recordings.

[7] http://perception.inrialpes.fr/CAVA_Dataset/

Figure 2.7: CAVA database recording environment and setup.

The two audio streams were sampled at 44.1 kHz, the acquisition was done with in-house software from Sheffield University. The resulting file contained wave data for the two microphones and a synchronization timestamp written to its header. The two video streams contained 1024×768 colour images recorded at 25 frames per second. They were stored in raw format, i.e. 8 bits per pixel with Bayer pattern encoding. The two streams were synchronized through an external trigger and each frame was timestamped. Synchronization of the audio and video streams was twofold. Firstly, timers on the computers, on which audio and video acquisition was performed, were aligned through the NTP protocol. This ensured consistency in timestamping for audio and video. Secondly, a device that resembled a clapper board used in movie production was employed.

Below we give details on some scenarios from the database that were designed as verification base for the methods derived in the current Thesis and constitute one of its contributions. The aim was to enable evaluation of audio, video and AV tracking and clustering in scenes with various challenges, such as actors walking in and out of the field of view, walking behind a screen, occluding each other, changing appearance and speaking in presence of multiple simultaneous sound sources.

The considered scenarios are recorded from the point of view of fixed perceiver (the acquisition device doesn't move). Table 2.2 gives an overview of the recordings and Figure 2.8 shows the accompanying "storyboard schematics". The name of each sequence is

sequence name	duration, min:sec	type of head	number of speakers	speaker behaviour	visual occlusion	auditory overlap
TTOS 1	20.84	dummy	1	moving	yes	no
CT1OS 1	18.51	dummy	1	moving	no	no
CT2OS 3	21.76	dummy	1*	moving	no	no
CT3OS 1	19.48	dummy	2†	moving	no	no
NTOS 2	33.02	dummy	1	moving	yes - L	no - M/N/C
TTMS 3	23.28	dummy	3 to 4	moving	yes	yes
CTMS 3	25.34	dummy	1 to 3	moving	yes	yes
DCMS 3	48.40	dummy	2 to 4	moving	yes	yes
NTMS 2	26.62	dummy	2	moving	yes - L	no - M/N/C
CPP 1	2:40.54	dummy	several	seated	yes	yes
M 1	3:47.80	dummy	5	seated	yes‡	yes

* actor changes appearance; † actors speak one at a time; ‡ two speakers are not visible

Table 2.2: List of recorded sequences - the visual occlusion accounts both for (i) an occlusion of a speaker by another speaker or by a wall, and (ii) a speaker outside of the field of view while speaking. In the column "auditory overlap" and "visual occlusion", the tags mean [M]usic, [C]licks, white [N]oise and [L]ight changes.

unique, and is composed of a scenario name and a number e.g. tracking test one speaker, sequence 1 (TTOS 1). Each scenario has been recorded several times. One representative sequence per scenario is currently available. The names used in the table correspond to the names of the sequence on the web site.

TTOS: tracking test; one speaker - Figure 2.8(a). One speaker, walking while speaking continously though the whole scene. The speaker moves in front of the camera and passes behind. He reappears from the right, and turns to the cameras. The purpose of this sequence is to evaluate audio (A), video (V) and audio-visual (AV) speaker tracking on difficult motion cases, and in situations where the speaker is out of the field of view.

CT1OS: clustering test 1; one speaker - Figure 2.8(b). One speaker, walking. The speaker moves while speaking in front of the camera and passes behind it from the left. As soon as he gets out of the field of view, the actor becomes silent. Only on reappearing from the right, does he start speaking again and turns to the cameras. The purpose of this sequence is to evaluate A, V, and AV speaker tracking on difficult motion cases, as well as A, V, and AV recognition test.

CT2OS: clustering test 2; one speaker. Same scenario as CT1OS again with one walking speaker. The main distinction is that, when reappearing, the actor has changed appearance (taken off jacket, put on glasses). An AV recognition test should be able to

detect that it is the same speaker.

CT3OS: clustering test 3; one speaker - Figure 2.8(c). Two actors, only one seen and heard at a time. The first speaker moves towards the camera then disappears from the field of view and stops talking. The second speaker enters the field of view while speaking and faces the cameras. An AV recognition test should be able to distinguish the two speakers.

NTOS: noise test; one speaker - Figure 2.8(d). One speaker, walking. The actor walks behind a wall and returns to his initial position, always speaking. Various audio noises like clicks and music are regularly present. The lighting condition is intentionally modified. This sequence may be use to verify the performance of tracking / recognition / speech analysis algorithms in AV noisy environment, and with visual occlusions.

DCMS: dynamic changes; multiple speakers - Figure 2.8(e). Five actors in total. Initially there are two speakers, then a third joins, one leaves, and later on a fifth joins. Then another two leaves. All actors speak while in the scene and move around.

TTMS: tracking test; multiple speakers - Figure 2.8(f). A more complex tracking scenario than the single speaker TTOS. Four actors are initially in the scene. As they start speaking (and go on speaking throughout the test), they move around; one person exits the visible scene, walks behind the camera while talking, and reappears. To test tracking abilities on speakers when both in and out of the field of view.

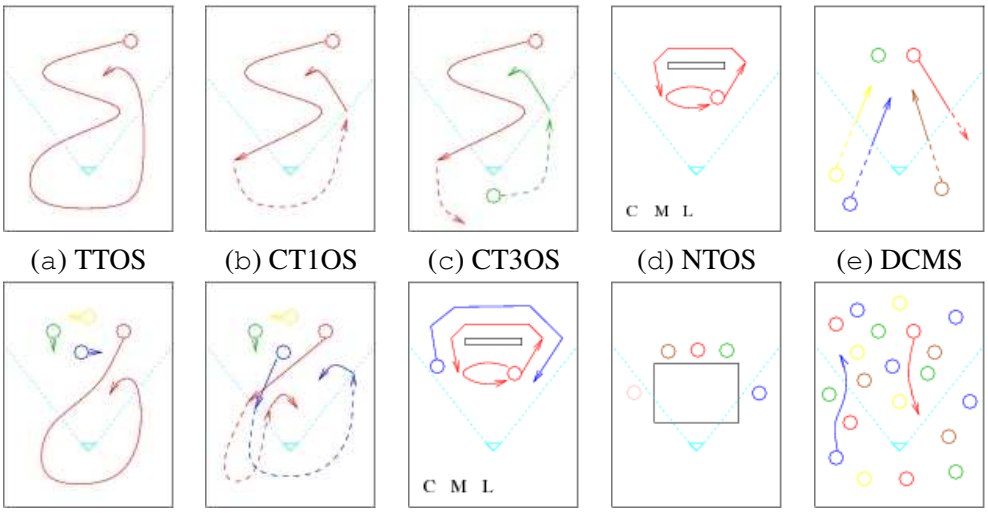
CTMS: clustering test; multiple speaker - Figure 2.8(g). A more complex clustering test scenario than the single speaker CTOS. Here four actors are initially in the scene. As they start speaking and moving around, two people exit the visible scene, stop talking, reappear and start talking again.

NTMS: noise test; multiple speakers - Figure 2.8(h). Similar to the one speaker noise test, NTOS. Two speakers are talking, occasionally walking behind a screen.Meanwhile music and clicks are heard in the background.

M1: meeting - Figure 2.8(i). Five actors are seated around a table, three are visible to the fixed perceiver (dummy head); one is to the left and one is to the right of the dummy. Initially all join into the same conversation and later on two sub-groups of conversations are formed.

CPP: cocktail party problem - Figure 2.8(j). 7 actors in total, 6 in scene and one to the left of the fixed perceiver. Two groups of conversation (one immediately in front of and one further away from the dummy head) are formed. People are seated and generally not moving a lot. At some point one speaker from the furthest away group gets up and joins the conversation of the front group. This setup makes for a very challenging auditory and visual scene.

The TTOS1, CTMS3 and M1 scenarios were annotated: actor 3D positions in camera frame and actor speaking activity were provided for the sequences.



(a) TTOS (b) CT1OS (c) CT3OS (d) NTOS (e) DCMS

(f) TTMS (g) CTMS (h) NTMS (i) M1 (j) CPP

Figure 2.8: Scenario schematics. Actors are depicted with circles, lines indicate 2D actor trajectories in the room. A solid line indicates "speaks while walking", and a dashed line means "quiet while walking". The field of view is drawn in blue. The rectangle accounts for an occluding wall. The tags mean [M]usic, [C]licks, [L]ight changes.

2.5 Discussion

We presented an approach to computational audio-visual (AV) scene analysis using a head-like device. Several hardware configurations are described that possess different properties in terms of recorded signals quality and capabilities of active behaviour. The following advantages of head-like devices with respect to other configurations can be pointed out:

- **Self-sufficiency**: the device, once calibrated, doesn't require any knowledge about the environment it's put to - scene reconstruction and adaptation can be done automatically;
- **Easy calibration**: precise calibration can be performed in short time with well-established techniques;
- **Persistent calibration**: the device can use motors to perform pan, tilt and eye vergence motions, while keeping the calibration valid;
- **Autocalibration**: there is a possibility to make calibration of a head-like device fully-automatic through the use of motor controls;
- **3D reconstruction**: the device is capable of reconstructing the observed scene in the 3D ambient space;

We showed the examples of features that could be extracted from both modalities: "interest points" for binocular vision and interaural time difference (ITD) values for binaural

hearing. Both occur to be general enough to be extracted for almost any kind of AV object and informative enough to perform AV integration. Of course, one can consider other object characteristics, such as local 3D motion, local interest point descriptors, interaural level differences (ILD) [Wang 2006], spectral features of auditory observations etc. In the current Thesis we focus on arbitrary AV object detection, localization and tracking and concentrate on the major principle of AV integration, considering the two described auditory and visual features. The others are left for possible extensions (see Chapter 8).

Finally, we presented a novel CAVA database aimed to investigate binaural/binocular fusion strategies of a human and to validate and compare the models of an audio-visual (AV) perceiver. It was acquired with a head-like device comprising a pair of calibrated cameras and a pair of calibrated microphones. The environment was kept natural, so that auditory reverberations, ambient sounds, lighting changes were not artificially removed. The recorded scenarios vary from almost stationary scenes and single-target tracking tasks to complex dynamnic scenes.

Two contributions of the current Thesis are related to the CAVA database:

- development and implementation of the fixed perceiver part of the scenarios;
- annotation of TTOS1, CTMS3 and M1 scenarios;

CHAPTER 3

Spatio-temporal Approach to Audio-Visual Calibration

Sommaire

3.1 Multisensor Calibration Task **23**
3.2 Calibration Through Multimodal Trajectory Matching **25**
3.3 Trajectory Reconstruction and Parameter Estimation. **26**
3.3.1 Problem Discretization and Relaxation 26
3.3.2 Hidden Trajectory Inference Using the EM Algorithm 30
3.3.3 Microphone Locations Inference Using the EM Algorithm. 34
3.3.4 Calibration Algorithm. 34
3.4 Experimental Validation **35**
3.4.1 Experiments with Simulated Data 35
3.4.2 Experiments with Real Data 40
3.5 Discussion **41**

This Chapter is devoted to a technical, but very important issue of audio-visual (AV) calibration. Indeed, data arriving from different sensors is meaningless, unless there is a common parametrization that ties the observations together. We refer to the task of finding the optimal configuration parameter values as the *calibration task*. Our calibration is based on matching unaligned AV data. The particularity of our approach is that we analyze correspondences between *trajectories* in modality spaces, rather than between separate single points. The approaches based on L_p optimization ($1 \leq p \leq 2$) are compared. We demonstrate the algorithm performance on both, simulated and real data, analyzing accuracy in the estimated values.

3.1 Multisensor Calibration Task

Tendency to use configurations containing multiple sensors is backed by numerous benefits such as robustness to observation noise, increased stability with respect to dynamic changes in the observed scene and better accuracy in estimations derived from the observations. This is achieved through integration of data coming from different sensors. Applications can be found in different domains: speech processing and acoustics [Raykar 2004, McCowan 2008], computer vision [Svoboda 2005, Courchay 2010]

robotics [Yguel 2008, Gould 2008], tracking systems [Cui 2008, Spinello 2008] etc. However, knowledge of inter-sensor parameters of the configuration is required, to benefit from the data integration. Finding the optimal parameters constitutes the *calibration task*.

The major goal in the calibration task is to find a trade-off between the number of parameters and the complexity of the optimization algorithm. On the one hand, the more parameters are included in to the calibration procedure, the better would be the correspondence of optimal parameters to the observations. The extremum of the target function becomes "sharper" but harder to find due to dimensionality increase. On the other hand, reducing the dimensionality of the parameter space leads to more efficient optimization procedures, but less prominent or even ambiguous extremal points. A good solution to this duality problem is to consider rich parameter space and impose various constraints on the observation spaces.

We work with the "robot head" configuration that comprises a pair of stereoscopic cameras and a pair of microphones (see Section 2.1). Thus integration of audio-visual (AV) data to improve AV object detection, localization and tracking is the primary concern. So far there has been no attempt to use a head-like device for this kind of task. Various other AV configurations perform approximate AV calibration by making restrictive assumptions on the observed objects [Beal 2003, Vermaak 2001], or by aligning projected data [Gatica-Perez 2007], or they perform precise AV calibration adding assumptions on the observed environment and using microphone arrays for better auditory localization [Zotkin 2002, Checka 2004, Nickel 2005]. For example, the AV integration models proposed in [Beal 2003] and [Hospedales 2008] perform AV calibration at the same time as AV integration by assuming affine dependency between person's ITD and his location in an image. This approach has an advantage of permanent online correction of the calibration, but at the same time it implicitly assumes that persons are located at a certain distance from the sensors. Moreover, this approximation is not valid for AV objects outside of the field of view and has no direct relation to the geometry of the ambient 3D space. This means that in the case of a mobile robot head with pan, tilt and vergence controls one cannot easily update the calibration using the motor data or determine the angle to turn the head towards a sound source, so that it becomes visible.

The real-world AV data tend to be influenced by the structure of the 3D environment in which they were generated. Thus we would like to use geometric properties of auditory and visual observations and consider the integration task in the 3D ambient space. Approximate projection-based calibration is not sufficient in this case and exact AV calibration is required. At the same time we would like to preserve the original head-like configuration without using additional microphones.

A typical approach to multimodal calibration consists in acquiring observations of the same object (calibration rig) simultaneously by all the sensors for further use in the optimization procedure to find optimal inter-sensor parameters. The optimization task is usually formulated as a least squares problem [Raykar 2004, McCowan 2008].

In our case forcing synchronization of auditory and visual streams would significantly increase the duration of calibration procedure without any improvement in data set. Thus

the goal is to develop a method that performs AV calibration on the two streams without their explicit alignment. This becomes possible if one considers *rig trajectories* in each modality, instead of single observations. Another benefit is that the trajectory-based calibration method is able to take advantage from temporal information, thus augmenting the data set.

In Section 3.2 we formalize the trajectory-based calibration task using the continuous-time notation. The approach is based on geometrical observation models that relate the two modalities. In Section 3.3 we show how to discretize the model and propose the relaxed version that is more robust to various noise types. A general optimization algorithm is formally derived based on the alternating EM procedure, several techniques to accelarate the algorithm are proposed. The experiments presented in Section 3.4 show the algorithm performance for various parameter values and outline the most important optimization steps. The method is demonstrated on both, simulated and real data, acquired with a specially designed device 'Altair'. Discussion of the results and directions for future work in Section 3.5 conclude the Chapter.

3.2 Calibration Through Multimodal Trajectory Matching

Given a head-like device equipped with a calibrated stereo camera pair and a pair of microphones, we would like to relate the auditory and visual frames. The geometry of visual observation model is defined through the visual space mapping $\mathcal{F}$ and given by (2.3):

$$\mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{1}{z}\right)^\top \quad \text{and} \quad \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{u}{d}, \frac{v}{d}, \frac{1}{d}\right)^\top,$$

where $\boldsymbol{s} = (x, y, z)^\top$ is the ambient space 3D position. Similarly, the geometry of auditory observation model is defined through the auditory space mapping $\mathcal{G}$ given by (2.4):

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big),$$

where the speed of sound c should be given in the same units as $\boldsymbol{s}$. Hence to relate the two observation spaces, one needs to determine microphone locations M_ℓ and M_r in visual frame.

Assume an object that is both seen and heard, moves along the trajectory

$$\boldsymbol{s}(t) = \big(x(t), y(t), z(t)\big)^\top, \quad t \in [t_{\min},\ t_{\max}] \tag{3.1}$$

in the 3D space. The object's size is supposed to be negligibly small, so that it can be roughtly considered to be a point. On the one hand, the trajectory maps to visual space into

$$\boldsymbol{f}(t) = \mathcal{F}(\boldsymbol{s}(t)) = \big(u(t), v(t), d(t)\big)^\top, \quad t \in [t_{\min},\ t_{\max}]. \tag{3.2}$$

On the other hand, the image of the trajectory with the auditory space mapping $\mathcal{G}$ gives

$$g(t) = \mathcal{G}(\boldsymbol{s}(t);\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}), \quad t \in [t_{\min},\ t_{\max}]. \tag{3.3}$$

In what follows we denote $\boldsymbol{\theta} = \{\boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\}$ and write $\mathcal{G}(\boldsymbol{s}(t);\ \boldsymbol{\theta})$ instead of $\mathcal{G}(\boldsymbol{s}(t);\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})$ to keep the notation concise.

The task of audio-visual calibration can be stated as follows: given the two observed trajectories $\boldsymbol{f}(t)$ and $g(t)$ that correspond to the same (unobserved) object trajectory $\boldsymbol{s}(t)$ in the ambient space, find the microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ and the 3D object trajectory $\boldsymbol{s}(t)$ such that they minimize the discrepancy simultaneously between $\boldsymbol{f}(t)$ and $\mathcal{F}(\boldsymbol{s}(t))$ and between $g(t)$ and $\mathcal{G}(\boldsymbol{s}(t);\ \boldsymbol{\theta})$. This problem is formalized as

$$\{\boldsymbol{\theta}^*, \boldsymbol{s}^*\} = \underset{\boldsymbol{\theta}\in\Theta,\ \boldsymbol{s}\in\mathscr{S}}{\operatorname{arginf}} \left\{\|\boldsymbol{f} - \mathcal{F}\circ\boldsymbol{s}\|^p_{\mathscr{F}} + \|g - \mathcal{G}(\boldsymbol{\theta})\circ\boldsymbol{s}\|^q_{\mathscr{G}} + \gamma\mathcal{R}(\boldsymbol{s})\right\}, \tag{3.4}$$

for a compact set Θ, positive constants p and q, some functional spaces $\mathscr{S}([t_{\min},\ t_{\max}] \to \mathbb{S})$, $\mathscr{F}([t_{\min},\ t_{\max}] \to \mathbb{F})$ and $\mathscr{G}([t_{\min},\ t_{\max}] \to \mathbb{G})$ with the associated norms $\|\cdot\|_{\mathscr{F}}$, $\|\cdot\|_{\mathscr{G}}$ and $\|\cdot\|_{\mathscr{S}}$, regularization functional $\mathcal{R}$ and regularization parameter γ. The sign '$\circ$' denotes the function composition operation.

For example, one could take quadratic penalties for observed functions with regularization $\mathcal{R}(\boldsymbol{s})$ given by $\int_{t_{\min}}^{t_{\max}} \left\|\frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}t}\right\|^2 \mathrm{d}t$. This would imply the functional spaces $\mathscr{F}$ and $\mathscr{G}$ to be $\mathcal{L}^2([t_{\min},\ t_{\max}], \mathbb{F})$ and $\mathcal{L}^2([t_{\min},\ t_{\max}], \mathbb{G})$ respectively. The trajectory $\boldsymbol{s}(t)$ then belongs to Sobolev space $\mathcal{W}^{1,2}([t_{\min}, t_{\max}])$. In what follows we shall concentrate on the latter class of $\mathcal{W}^{1,2}([t_{\min}, t_{\max}])$ trajectories with the $\mathcal{L}^p([t_{\min},\ t_{\max}], \cdot)$ norm ($1 \le p \le 2$) used for penalty terms.

3.3 Trajectory Reconstruction and Parameter Estimation.

The problem (3.4) includes two optimization tasks to be solved simultaneously - the target function should be minimized with respect to a hidden trajectory $\boldsymbol{s}(t)$ and with respect to the parameters $\boldsymbol{\theta}$. Efficient solutions can be proposed for certain choices of the penalty and regularization terms. In this Section we restrain the general calibration problem (3.4) and adapt it to the particular task of audio-visual (AV) calibration. The variational approach being less suitable in the case of AV data and less evident to derive because of the non-linear mappings $\mathcal{F}$ and $\mathcal{G}$, we develop the discretized analogue of (3.4) and give it the Bayesian interpretation.

3.3.1 Problem Discretization and Relaxation

To narrow down the class of optimization tasks we take norms from $\mathcal{L}^p([t_{\min},\ t_{\max}], \mathbb{F})$ and $\mathcal{L}^p([t_{\min},\ t_{\max}], \mathbb{G})$ with $1 \le p \le 2$ for the penalty terms and the first order regularization term for the trajectory $\boldsymbol{s}(t)$:

$$\{\boldsymbol{\theta}^*, \boldsymbol{s}^*\} = \underset{\boldsymbol{\theta}\in\Theta,\ \boldsymbol{s}\in\mathscr{S}}{\operatorname{arginf}} \left\{\|\boldsymbol{f} - \mathcal{F}(\boldsymbol{s})\|^p_{\mathbb{F},p} + \|g - \mathcal{G}(\boldsymbol{s};\ \boldsymbol{\theta})\|^p_{\mathbb{G},p} + \gamma\mathcal{R}(\boldsymbol{s})\right\}, \tag{3.5}$$

where $\mathscr{S} = \mathcal{W}^{1,2}([t_{\min}, t_{\max}])$ is Sobolev space and

$$\|\boldsymbol{f} - \mathcal{F}(\boldsymbol{s})\|^p_{\mathbb{F},p} = \int_{t_{\min}}^{t_{\max}} \|\boldsymbol{f}(t) - \mathcal{F}(\boldsymbol{s}(t))\|^p_p \mathrm{d}t, \tag{3.6}$$

$$\|g - \mathcal{G}(\boldsymbol{s};\ \boldsymbol{\theta})\|^p_{\mathbb{G},p} = \int_{t_{\min}}^{t_{\max}} |g(t) - \mathcal{G}(\boldsymbol{s}(t);\ \boldsymbol{\theta})|^p \mathrm{d}t, \tag{3.7}$$

$$\text{and} \quad \mathcal{R}(\boldsymbol{s}) = \int_{t_{\min}}^{t_{\max}} \left\| \frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}t} \right\|^2 \mathrm{d}t. \tag{3.8}$$

The minimization problem with respect to $\boldsymbol{s}$ can be solved using the variational approach. The particular case of linear mappings $\mathcal{F}$ and $\mathcal{G}$ with $p = 2$ admits an efficient optimization scheme. Taking the variational derivative leads to a screened Poisson equation that can be solved in Fourier domain, as shown for the 2D case in [Bhat 2008]. However, we do not consider this approach here for several reasons. Firstly, in our case $\mathcal{F}$ and $\mathcal{G}$ are essentially non-linear, approximations are required to reduce the problem to the screened Poisson equation. Secondly, practice shows that the penalty terms given in (3.5) do not fully account for all kinds of noise in the AV data. Finally, the observed data for each modality forms a stream of values arriving at discrete time instants. Thus it would be more natural to discretize the problem, improve penalty terms and develop the optimization method for the general case of non-linear mappings $\mathcal{F}$ and $\mathcal{G}$.

Assume, observations are detected in the two modalities at time instants $t_{\min} \le t_1^{(\mathrm{f})} < \ldots < t_m^{(\mathrm{f})} < \ldots < t_M^{(\mathrm{f})} \le t_{\max}$ and $t_{\min} \le t_1^{(\mathrm{g})} < \ldots < t_k^{(\mathrm{g})} < \ldots < t_K^{(\mathrm{g})} \le t_{\max}$ respectively. We denote the resulting sets

$$\mathbf{f} = \{\boldsymbol{f}_m\}_{m=1}^M, \quad \boldsymbol{f}_m = \boldsymbol{f}(t_m) \in \mathbb{R}^3, \quad \text{and} \quad \boldsymbol{g} = \{g_k\}_{k=1}^K, \quad g_k = g(t_k) \in \mathbb{R}, \tag{3.9}$$

They are not necessarily aligned in time, i.e. M and K can be different and time instants t_m and t_k are not expected to coincide for any m and k. To account for the fact that these observations were generated from the same trajectory $\boldsymbol{s}(t)$ that is discretized

$$\mathbf{s} = \{\boldsymbol{s}_n\}_{n=1}^N, \quad \boldsymbol{s}_n = \boldsymbol{s}(t_n) \in \mathbb{S} \subset \mathbb{R}^3, \quad t_{\min} \le t_1^{(\mathrm{s})} < \ldots < t_n^{(\mathrm{s})} < \ldots < t_N^{(\mathrm{s})} \le t_{\max}, \tag{3.10}$$

we introduce subsequences $n_m^{(f)}$ and $n_k^{(g)}$ that verify $t_m^{(\mathrm{f})} = t_{n_m^{(f)}}^{(\mathrm{s})}$ and $t_k^{(\mathrm{g})} = t_{n_k^{(g)}}^{(\mathrm{s})}$ respectively. Timestamp set $\{t_n^{(\mathrm{s})}\}_{n=1}^N$ can be taken as an ordered union $\{t_m^{(\mathrm{f})}\}_{m=1}^M \cup \{t_k^{(\mathrm{g})}\}_{k=1}^K$. Further we shall omit the upper indicators of the timestamp sets, using t_n, t_m and t_k instead of $t_n^{(\mathrm{s})}$, $t_m^{(\mathrm{f})}$ and $t_k^{(\mathrm{g})}$ respectively. We illustrate how discrete observations sets $\mathbf{f}$ and $\boldsymbol{g}$ are related to the hidden continuous 3D space trajectory $\boldsymbol{s}(t)$ in Figure 3.1.

The discrete analogue of the regularization term $\mathcal{R}(\boldsymbol{s})$ is given by

$$\mathcal{H}_{\mathrm{s}}(\mathbf{s}) = \sum_{n=1}^{N-1} \frac{\|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2}{t_{n+1} - t_n}, \tag{3.11}$$

which engenders a Gaussian process on trajectories $\mathbf{s}$ space:

$$P(\mathbf{s}) \propto \exp\left\{-\gamma \mathcal{H}_{\mathrm{s}}(\mathbf{s})\right\}. \tag{3.12}$$

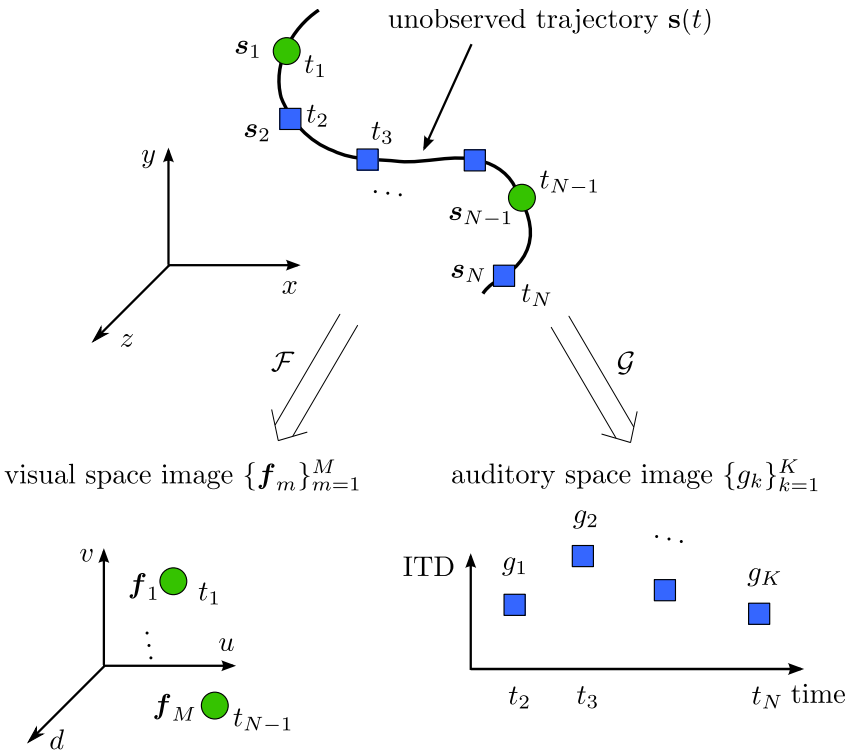


Figure 3.1: Discretization in auditory and visual spaces. The observations $\boldsymbol{f}_m$ and g_k are detected in the two modalities at different time instants, but correspond to the same unobserved 3D space trajectory $\boldsymbol{s}(t)$. The mappings $\mathcal{F}$ and $\mathcal{G}$ that relate 3D space positions to the visual and auditory observations depend on visual and auditory system calibration parameters.

To write the discrete versions of (3.6) and (3.7) we first consider in more detail the way data is generated. The observations in both modalities can be corrupted by two different types of noise. Firstly, the detection process is based on matching two monocular/monaural features into a binocular/binaural one. Matching errors (e.g. in the presence of another visual/auditory observations source) lead to significant deviations of an observation $\boldsymbol{f}_m$ or g_k from the real object's position $\mathcal{F}(\boldsymbol{s}_{n_m})$ or $\mathcal{G}(\boldsymbol{s}_{n_k};\ \boldsymbol{\theta})$ in the corresponding modality. We suppose $\boldsymbol{f}_m$ and g_k to be uniformly distributed on $\mathbb{F}$ or $\mathbb{G}$ respectively in this case. If the pair of features was chosen correctly, there can still exist small deviations from the real object's position and we assume generalized Gaussian noise distributions. To distinguish between the two cases we introduce sets of random variables

$$\boldsymbol{A} = \{A_m\}_{m=1}^M, \quad A_m = \begin{cases} 0, & \text{if visual matching error,} \\ 1, & \text{otherwise,} \end{cases} \tag{3.13}$$

$$\text{and} \quad \boldsymbol{B} = \{B_k\}_{k=1}^K, \quad B_k = \begin{cases} 0, & \text{if auditory matching error,} \\ 1, & \text{otherwise,} \end{cases} \tag{3.14}$$

i.e. each observation $\boldsymbol{f}_m$ and g_k is associated with a matching error flag A_m and B_k.

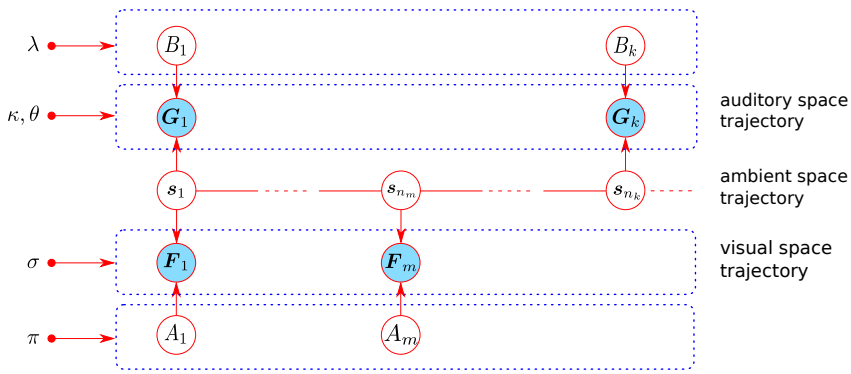


Figure 3.2: Graphical representation of the audio-visual calibration model. Unobserved ambient space trajectory $\boldsymbol{s}(t)$ can have associated visual observations, auditory observations or both at different times. Only the auditory mapping $\mathcal{G}$ depends on calibration parameters $\boldsymbol{\theta}$.

The values of the random variables $\boldsymbol{A}$ and $\boldsymbol{B}$ for a particular realisation are unknown and should be estimated. Then the conditional likelihood of $\boldsymbol{f}_m$ and g_k given the unobserved object position $\boldsymbol{s}_{n_m}$ or $\boldsymbol{s}_{n_k}$ can be written as

$$P(\boldsymbol{f}_m \mid A_m, \boldsymbol{s}_{n_m}) = \mathcal{N}_p(\boldsymbol{f}_m; \mathcal{F}(\boldsymbol{s}_{n_m}), \sigma)\delta_{A_m} + \mathcal{U}(\boldsymbol{f}_m\,; V)(1-\delta_{A_m}), \quad (3.15)$$

$$\text{and} \quad P(g_k \mid B_k, \boldsymbol{s}_{n_k}) = \mathcal{N}_p(g_k; \mathcal{G}(\boldsymbol{s}_{n_k}), \kappa)\delta_{B_k} + \mathcal{U}(g_k\,; U)(1-\delta_{B_k}), \quad (3.16)$$

where for $\boldsymbol{x} \in \mathbb{R}^d$ we let

$$\mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \sigma) = \left(\frac{p}{2\sigma\Gamma(1/p)}\right)^d \exp(-\|\boldsymbol{x}-\boldsymbol{\mu}\|_p^p/\sigma^p) \quad (3.17)$$

and δ is the Kronecker delta:

$$\delta_i = \begin{cases} 1, & \text{if } i = 0, \\ 0, & \text{if } i \neq 0, \end{cases} \quad (3.18)$$

the scale parameters σ and κ do not depend on n and V and U are Lebesgue measures of the support sets in $\mathbb{F}$ and $\mathbb{G}$ respectively. We note that if one takes $p = 2$ and $\boldsymbol{\Sigma} = 0.5\sigma^2\mathbb{I}$, the distribution (3.17) would become the usual multivariate Gaussian distribution. A more general model can be considered with $\sigma = \sigma(n)$ or $\kappa = \kappa(n)$ being parametrically dependent on the position $\boldsymbol{s}_n$, but we do not discuss this case. We suppose the observations $\boldsymbol{f}_m$ and g_k to be conditionally independent, so that

$$P(\mathbf{f} \mid \boldsymbol{A}, \mathbf{s}) = \prod_{m=1}^{M} P(\boldsymbol{f}_m \mid A_m, \boldsymbol{s}_{n_m}), \quad (3.19)$$

$$\text{and} \quad P(\boldsymbol{g} \mid \boldsymbol{B}, \mathbf{s}) = \prod_{k=1}^{K} P(g_k \mid B_k, \boldsymbol{s}_{n_k}). \quad (3.20)$$

The matching error flags are assumed to be independent and identically distributed:

$$P(\boldsymbol{A}) = \prod_{m=1}^{M} P(A_m), \quad P(A_m) = \pi\delta_{A_m} + (1-\pi)(1-\delta_{A_m}), \quad 0 \le \pi \le 1. \tag{3.21}$$

$$P(\boldsymbol{B}) = \prod_{k=1}^{K} P(B_k), \quad P(B_k) = \lambda\delta_{B_k} + (1-\lambda)(1-\delta_{B_k}), \quad 0 \le \lambda \le 1. \tag{3.22}$$

Thus full probabilities for the observation sets $\mathbf{f}$ and $\boldsymbol{g}$ given $\mathbf{s}$ can be written as

$$P(\mathbf{f} \mid \mathbf{s}) = \prod_{m=1}^{M} \left(\pi\mathcal{N}_p(\boldsymbol{f}_m;\, \mathcal{F}(\boldsymbol{s}_{n_m}), \sigma) + (1-\pi)\mathcal{U}(\boldsymbol{f}_m\,;\, V)\right), \tag{3.23}$$

$$P(\boldsymbol{g} \mid \mathbf{s}) = \prod_{k=1}^{K} \left(\lambda\mathcal{N}_p(g_k;\, \mathcal{G}(\boldsymbol{s}_{n_k}), \kappa) + (1-\lambda)\mathcal{U}(g_k\,;\, U)\right). \tag{3.24}$$

We note that (3.23) and (3.24) are discrete analogues of (3.6) and (3.7), where strict observation proximity condition, expressed by a generalized Gaussian distribution, is relaxed by the uniform component.

The calibration problem is formulated as

$$\{\mathbf{s}^*, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*\} = \underset{\mathbf{s}\in\mathbb{S}^N,\, \boldsymbol{\theta}\in\Theta,\, \boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmax}} \log P(\mathbf{f}, \boldsymbol{g}, \mathbf{s}, \boldsymbol{\theta}\,;\, \boldsymbol{\psi}), \tag{3.25}$$

where $\boldsymbol{\psi} = \{\pi, \lambda, \sigma, \kappa\}$ are the model parameters. We use the alternating optimization approach to solve (3.25). The target function is optimized by turns with respect to the ambient space trajectory $\mathbf{s}$ and with respect to microphone locations $\boldsymbol{\theta}$. These two steps are described in detail in Sections 3.3.2 and 3.3.3 respectively.

3.3.2 Hidden Trajectory Inference Using the EM Algorithm

Let's suppose that the parameters $\boldsymbol{\theta}$ are fixed and the task is to carry out optimization with respect to the trajectory $\mathbf{s}$. We formulate the problem of trajectory estimation in Bayesian framework, looking for the optimal $\mathbf{s}^* \in \mathbb{S}^N$ and $\boldsymbol{\psi}^* = \{\pi^*, \lambda^*, \sigma^*, \kappa^*\}$ such that

$$\{\mathbf{s}^*, \boldsymbol{\psi}^*\} = \underset{\mathbf{s}\in\mathbb{S}^N,\, \boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmax}} \log P(\mathbf{f}, \boldsymbol{g}, \mathbf{s}\,;\, \boldsymbol{\psi}). \tag{3.26}$$

The expectation-maximization (EM) algorithm [Dempster 1977, McLachlan 2007] is a standard approach to carry out such a maximization. It is given by an iteration

$$\left\{\mathbf{s}^{(q+1)}, \boldsymbol{\psi}^{(q+1)}\right\} = \underset{\mathbf{s}\in\mathbb{S}^N,\, \boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmax}} Q(\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\psi}^{(q)}, \mathbf{s}^{(q)}), \tag{3.27}$$

$$\text{with} \quad Q(\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\psi}^{(q)}, \mathbf{s}^{(q)}) = \mathbb{E}_{\boldsymbol{A},\boldsymbol{B}}[\log P(\mathbf{f}, \boldsymbol{g}, \mathbf{s}, \boldsymbol{A}, \boldsymbol{B};\, \boldsymbol{\psi}) \,|\, \mathbf{f}, \boldsymbol{g}, \mathbf{s}^{(q)};\, \boldsymbol{\psi}^{(q)}], \tag{3.28}$$

where the expectation is taken over the hidden variables $\boldsymbol{A}$ and $\boldsymbol{B}$. Each iteration q of EM proceeds in two steps.

Expectation. For the current values $\boldsymbol{\psi}^{(q)}$ and $\mathbf{s}^{(q)}$ of the parameters and trajectory, compute the conditional expectation with respect to variables $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\begin{aligned} Q(\boldsymbol{\psi}, \mathbf{s}, \boldsymbol{\psi}^{(q)}, \mathbf{s}^{(q)}) = &- \sum_{m=1}^{M} \alpha_m^{(q)} \left(\sigma^{-p} \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_{n_m})\|_p^p + 3 \log \sigma - \log \frac{\pi}{1-\pi} \right) - \\ &- \sum_{k=1}^{K} \beta_k^{(q)} \left(\kappa^{-p} |g_k - \mathcal{G}(\boldsymbol{s}_{n_k};\ \boldsymbol{\theta})|^p + \log \kappa - \log \frac{\lambda}{1-\lambda} \right) + \\ &+ M \log(1-\pi) + K \log(1-\lambda) - \gamma \sum_{n=1}^{N-1} \frac{\|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2}{t_{n+1} - t_n} + C^{(q)}, \end{aligned} \tag{3.29}$$

where $C^{(q)}$ is a term that does not depend on $\boldsymbol{\psi}$ and $\mathbf{s}$, $\alpha_m^{(q)} = P(A_k = 0 \mid \boldsymbol{f}_m, \mathbf{s}^{(q)};\ \boldsymbol{\psi}^{(q)})$ and $\beta_k^{(q)} = P(B_k = 0 \mid g_k, \mathbf{s}^{(q)};\ \boldsymbol{\psi}^{(q)})$ are the posterior probabilities. Their expressions can be derived straightforwardly from Bayes' theorem:

$$\alpha_m^{(q)} = \frac{\pi^{(q)} \mathcal{N}_p(\boldsymbol{f}_m\,;\, \mathcal{F}(\boldsymbol{s}_n^{(q)}), \sigma^{(q)})}{\pi^{(q)} \mathcal{N}_p(\boldsymbol{f}_m\,;\, \mathcal{F}(\boldsymbol{s}_n^{(q)}), \sigma^{(q)}) + (1-\pi^{(q)}) \mathcal{U}(\boldsymbol{f}_m\,;\, V)}, \tag{3.30}$$

$$\text{and} \quad \beta_k^{(q)} = \frac{\lambda^{(q)} \mathcal{N}_p(g_k\,;\, \mathcal{G}(\boldsymbol{s}_n^{(q)};\ \boldsymbol{\theta}), \kappa^{(q)})}{\lambda^{(q)} \mathcal{N}_p(g_k\,;\, \mathcal{G}(\boldsymbol{s}_n^{(q)};\ \boldsymbol{\theta}), \kappa^{(q)}) + (1-\lambda^{(q)}) \mathcal{U}(g_k\,;\, U)}. \tag{3.31}$$

Maximization. Update the parameter set $\boldsymbol{\psi}^{(q)}$ and the trajectory $\mathbf{s}^{(q)}$ by performing maximization (3.27). We set the derivatives of the conditional expectation (3.29) with respect to model parameters to zero to obtain the update expressions. For priors one gets the usual empirical formulas:

$$\pi^{(q+1)} = \frac{1}{M} \sum_{m=1}^{M} \alpha_m^{(q)}, \tag{3.32}$$

$$\text{and} \quad \lambda^{(q+1)} = \frac{1}{K} \sum_{k=1}^{K} \beta_k^{(q)}. \tag{3.33}$$

Scale parameters are expressed as functions of the hidden trajectory

$$\sigma^{(q+1)} = \left(\frac{p}{3} \sum_{m=1}^{M} \alpha_m^{(q)} \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_{n_m}^{(q+1)})\|_p^p \right)^{1/p}, \tag{3.34}$$

$$\text{and} \quad \kappa^{(q+1)} = \left(p \sum_{k=1}^{K} \beta_k^{(q)} |g_k - \mathcal{G}(\boldsymbol{s}_{n_k}^{(q+1)};\ \boldsymbol{\theta})|^p \right)^{1/p}. \tag{3.35}$$

The trajectory $\mathbf{s}^{(q+1)}$ is found as a solution for a system of optimization problems

$$\begin{aligned}\boldsymbol{s}_n^{(q+1)} = \operatorname*{argmin}_{\boldsymbol{s}\in\mathbb{S}} \gamma \left(\frac{\|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2}{t_{n+1} - t_n} + \frac{\|\boldsymbol{s}_n - \boldsymbol{s}_{n-1}\|^2}{t_n - t_{n-1}} \right) + \\ + \delta_{m_n} \alpha_{m_n}^{(q)} \left(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|_p / \sigma^{(q)} \right)^p + \delta_{k_n} \beta_{k_n}^{(q)} \left(|g_k - \mathcal{G}(\boldsymbol{s}_n;\ \boldsymbol{\theta})| / \kappa^{(q)} \right)^p ,\end{aligned} \tag{3.36}$$

where δ_{m_n} and δ_{k_n} are defined as

$$\delta_{m_n} = \begin{cases} 1, & \text{if} \quad \exists m = m_n : n_m = n, \\ 0, & \text{otherwise,} \end{cases} \tag{3.37}$$

$$\text{and} \quad \delta_{k_n} = \begin{cases} 1, & \text{if} \quad \exists k = k_n : n_k = n, \\ 0, & \text{otherwise.} \end{cases} \tag{3.38}$$

Thus the trajectory is optimized taking into account the regularization term in (3.36) for every $n = 1 \dots N$ and observation discrepancy terms for those n that are observed in at least one modality. Taking the timestamp sequence $\{t_n^{(\mathrm{s})}\}_{n=1}^N$ as an ordered union $\{t_m^{(\mathrm{f})}\}_{m=1}^M \cup \{t_k^{(\mathrm{g})}\}_{k=1}^K$ ensures that there is always one observation corresponding to a hidden variable $\boldsymbol{s}_n$.

The optimization task (3.36) is performed using the *method of generations* that efficiently combines local and global optimization methods [Zhigljavsky 2008]. We make use of the fact that the mapping $\mathcal{F}$ is injective and sample the trajectory space using the preimage of the regularized visual space trajectory $\mathcal{F}^{-1}(\tilde{\mathbf{f}})$. Afterwards, we perform local coordinate-wise optimization of the trajectory $\mathbf{s}$.

Trajectory sampling. The sampling method we consider here is based on visual data. To draw a trajectory in the ambient space, we take the visual observation sequence $\mathbf{f}$, regularize it into $\tilde{\mathbf{f}}$ using a random $\tilde{\gamma}$ parameter value and map to $\mathbb{S}^N$ to get $\mathbf{s}^{(q+1,0)}$. The visual trajectory regularization method we consider resembles the one, given by (3.11) and (3.15) for $p = 2$. Though now the observation space and hidden trajectory space coincide, so the non-linear mapping is no longer present in the formulas:

$$P(\tilde{\mathbf{f}}) \propto \exp\left(-\tilde{\gamma} \sum_{m=1}^{M-1} \frac{\|\tilde{\boldsymbol{f}}_{m+1} - \tilde{\boldsymbol{f}}_m\|^2}{t_{m+1} - t_n} \right), \tag{3.39}$$

$$P(\mathbf{f} \mid \tilde{\mathbf{f}}) = \prod_{m=1}^{M} \left(\pi \mathcal{N}(\boldsymbol{f}_m;\ \tilde{\boldsymbol{f}}_m, \boldsymbol{\Sigma}) + (1-\pi)\mathcal{U}(\boldsymbol{f}_m\,;\, V) \right). \tag{3.40}$$

The solution to the problem is again acquired using the EM algorithm, but this time both steps admit closed form expressions. The E-step is given by

$$\tilde{\alpha}_m^{(q)} = \frac{\tilde{\pi}^{(q)} \mathcal{N}(\boldsymbol{f}_m\,;\, \tilde{\boldsymbol{f}}_m, \boldsymbol{\Sigma}^{(q)})}{\tilde{\pi}^{(q)} \mathcal{N}(\boldsymbol{f}_m\,;\, \tilde{\boldsymbol{f}}_m, \boldsymbol{\Sigma}^{(q)}) + (1 - \tilde{\pi}^{(q)}) \mathcal{U}(\boldsymbol{f}_m\,;\, V)} \tag{3.41}$$

M-step update expressions are

$$\tilde{\boldsymbol{f}}_m^{(q+1)} = \boldsymbol{\Xi}^{(q)} \boldsymbol{v}^{(q)}, \tag{3.42}$$

$$\tilde{\pi}^{(q+1)} = \frac{1}{M} \sum_{m=1}^{M} \tilde{\alpha}_m^{(q)}, \tag{3.43}$$

$$\boldsymbol{\Sigma}^{(q+1)} = \frac{1}{\sum\limits_{m=1}^{M} \tilde{\alpha}_m^{(q)}} \sum_{m=1}^{M} \tilde{\alpha}_m^{(q)} (\boldsymbol{f}_m - \tilde{\boldsymbol{f}}_m^{(q+1)})(\boldsymbol{f}_m - \tilde{\boldsymbol{f}}_m^{(q+1)})^\top. \tag{3.44}$$

Here $\boldsymbol{\Xi}^{(q)}$ is the inverse of a sparse $3M \times 3M$ matrix made of M blocks of size 3×3:

$$\boldsymbol{\Xi}^{(q)} = \begin{pmatrix} -\mathbb{I} & \mathbf{R}_1^{(q)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{L}_2^{(q)} & -\mathbb{I} & \mathbf{R}_2^{(q)} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_3^{(q)} & -\mathbb{I} & \dots & \mathbf{0} \\ \vdots & & & \dots & \mathbf{R}_{M-1}^{(q)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & -\mathbb{I} \end{pmatrix}^{-1}. \tag{3.45}$$

We denoted the 3×3 identity and zero matrices by $\mathbb{I}$ and $\mathbf{0}$ respectively and $\mathbf{L}_m^{(q)}$ and $\mathbf{R}_m^{(q)}$ are 3×3 matrices defined by

$$\mathbf{L}_m^{(q)} = \left[\frac{t_m - t_{m-1}}{\tilde{\gamma}} \tilde{\alpha}_m^{(q)} \boldsymbol{\Sigma}^{(q)^{-1}} + \frac{t_{m+1} - t_{m-1}}{t_{m+1} - t_m} \mathbb{I} \right]^{-1}, \quad m = 2, \dots, M, \tag{3.46}$$

$$\text{and} \quad \mathbf{R}_m^{(q)} = \left[\frac{t_{m+1} - t_m}{\tilde{\gamma}} \tilde{\alpha}_m^{(q)} \boldsymbol{\Sigma}^{(q)^{-1}} + \frac{t_{m+1} - t_{m-1}}{t_m - t_{m-1}} \mathbb{I} \right]^{-1}, \quad m = 1, \dots, M-1. \tag{3.47}$$

The vector $\boldsymbol{v}^{(q)} \in \mathbb{R}^{3M}$ in (3.42) is given by

$$\boldsymbol{v}^{(q)} = \left(\boldsymbol{v}_1^{(q)^\top}, \dots, \boldsymbol{v}_M^{(q)^\top} \right)^\top, \tag{3.48}$$

$$\text{with} \quad \boldsymbol{v}_m^{(q)} = \left[\frac{(t_{m+1} - t_m)(t_m - t_{m-1})}{\tilde{\gamma}(t_{m+1} - t_{m-1})} \tilde{\alpha}_m^{(q)} \boldsymbol{\Sigma}^{(q)^{-1}} + \mathbb{I} \right]^{-1} \boldsymbol{f}_m - \boldsymbol{f}_m. \tag{3.49}$$

We note that equations (3.45)-(3.49) define a variant of EM that uses the covariance matrix $\boldsymbol{\Sigma}^{(q)}$ from the previous step instead of $\boldsymbol{\Sigma}^{(q+1)}$. Using the arguments similar to those presented in [Xu 1997], we can argue that the resulting algorithm has the same convergence properties as its basic version.

Coordinate-wise trajectory optimization. Once the initial sampled solution $\mathbf{s}^{((q+1),0)}$ is obtained, we apply the local optimization procedure to (3.36). This procedure involves iterative updates of the target function with respect to $\boldsymbol{s}_n$, $n = 1, \dots, N$. Given the current trajectory $\mathbf{s}^{(q+1,i)}$, its update $\mathbf{s}^{(q+1,i+1)}$ is computed as follows: for $n = n(i)$ chosen

according to some scan strategy, the initial guess is computed as the one minimizing the regularization component:

$$\boldsymbol{s}_n^{(q+1,i+1)} = \frac{1}{t_{n+1} - t_{n-1}} \big((t_n - t_{n-1}) \boldsymbol{s}_{n+1} + (t_{n+1} - t_n) \boldsymbol{s}_{n-1} \big). \tag{3.50}$$

This position $\boldsymbol{s}_{n(i)}^{(q+1,i+1)}$ is then improved using the simultaneous perturbation stochastic approximation (SPSA) optimization algorithm [Spall 2003], other nodes $\{\boldsymbol{s}_j\}_{j \neq n(i)}$ on iteration i remain unchanged.

3.3.3 Microphone Locations Inference Using the EM Algorithm.

We assume now that the ambient space trajectory $\mathbf{s}$ is fixed and consider the optimization task

$$\{\boldsymbol{\theta}^*, \boldsymbol{\psi}^*\} = \underset{\boldsymbol{\theta} \in \Theta,\, \boldsymbol{\psi} \in \boldsymbol{\Psi}}{\operatorname{argmax}} \log P(\mathbf{f}, \boldsymbol{g}, \boldsymbol{\theta}\ ;\ \boldsymbol{\psi}) \tag{3.51}$$

to find microphone locations $\boldsymbol{\theta}^* = \{\boldsymbol{s}^*_{\mathrm{M}_\ell}, \boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}}\}$. As in the case of the ambient trajectory, the inference is performed with the EM algorithm that proceeds in two steps.

Expectation. For the current values $\boldsymbol{\psi}^{(q)}$ and $\boldsymbol{\theta}^{(q)}$ of the parameters and microphone locations, compute the conditional expectation with respect to variables $\boldsymbol{A}$ and $\boldsymbol{B}$. It is given by (3.29), but this time it is considered as a function $Q(\boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{\psi}^{(q)}, \boldsymbol{\theta}^{(q)})$ of microphone locations.

Maximization. Update the parameter set $\boldsymbol{\psi}^{(q)}$ and microphone locations $\boldsymbol{\theta}^{(q)}$ by performing maximization (3.51). As previously, we get the formulas (3.32)- (3.35) for the optimal parameters $\boldsymbol{\psi}^{(q+1)}$. The microphone locations $\boldsymbol{\theta}^{(q+1)}$ can be found as a solution to

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \sum_{k=1}^{K} \beta_k^{(q)} \left(|g_k - \mathcal{G}(\boldsymbol{s}_{n_k};\ \boldsymbol{\theta})| / \kappa \right)^p - \log P(\boldsymbol{\theta}), \tag{3.52}$$

where $P(\boldsymbol{\theta})$ is some prior distribution on the parameter values. Depending on the information we possess on the configuration, we take either uniform prior on some known domain Θ or a Gaussian prior centered at some supposed parameter value $\hat{\boldsymbol{\theta}}$. Performance of models, based on different priors is compared in Section 3.4. The minimization (3.52) is performed using the method of generations, based on sampling from the prior distribution $P(\boldsymbol{\theta})$ and local optimization through the SPSA algorithm.

3.3.4 Calibration Algorithm.

We provide the summary of a head-like device calibration algorithm. Given observation sequences $\mathbf{f} = \{\boldsymbol{f}_m\}_{m=1}^M$ and $\boldsymbol{g} = \{g_k\}_{k=1}^K$ from a calibrated camera pair and a microphone pair respectively, and the associated timestamp sequences $\{t_m\}_{m=1}^M$ and $\{t_k\}_{k=1}^K$

1. Calculate $\{t_n\}_{n=1}^N$ as an ordered union of $\{t_m\}_{m=1}^M$ and $\{t_k\}_{k=1}^K$;
2. Initialize the ambient space trajectory $\mathbf{s}^{(0)}$ from the visual space trajectory $\mathbf{f}$ using the regularization procedure (3.39)- (3.49) and interpolate it to the timestamps $\{t_n\}_{n=1}^N$;
3. Initialize the microphone locations $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}\}$ and parameters $\boldsymbol{\psi}^{(0)}$ using the EM algorithm, as described in Section 3.3.3;
4. $q \leftarrow 1$
5. Compute the ambient space trajectory $\mathbf{s}^{(q)}$ and parameters $\boldsymbol{\psi}^{(q-1/2)}$ from $\{\mathbf{s}^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\psi}^{(q-1)}\}$ using the EM algorithm, as described in Section 3.3.2;
6. Compute the microphone locations $\boldsymbol{\theta}^{(q)}$ and parameters $\boldsymbol{\psi}^{(q)}$ from $\{\mathbf{s}^{(q)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\psi}^{(q-1/2)}\}$ using the EM algorithm, as described in Section 3.3.3;
7. Terminate on convergence, otherwise $q \leftarrow q + 1$ and go to Step 5;

To improve calibration quality one can perform trajectory sampling on Step 2 instead of considering only one trajectory $\mathbf{s}^{(0)}$. Procedure proposed in Section 3.3.2 could be used to initialize multiple trajectories. The overall complexity of the proposed algorithm is $\mathcal{O}(N^2)$.

3.4 Experimental Validation

To verify performance of our model, we tested the algorithm on simulated and real datasets. Parameter values close to the ones observed for real configurations were used in simulated experiments. Different versions of the calibration algorithm for various penalty terms with $p \in [1; 2]$ are compared. Real-life experiment part contains calibration rig description, shows data obtained for both modalities and calibration results.

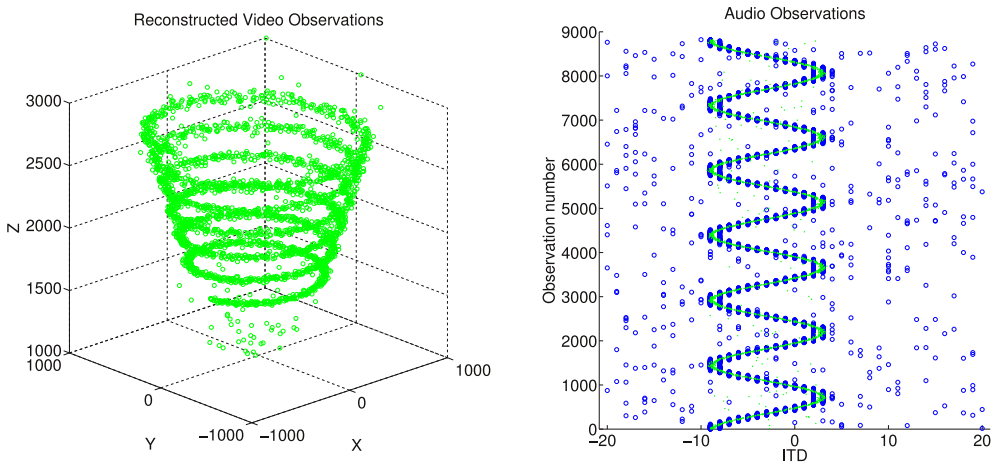
3.4.1 Experiments with Simulated Data

We aim at modelling the multimodal data as close as possible to the real data. Assume the calibration rig follows a spiral trajectory, given by

$$\boldsymbol{s}(t) = (30t\cos(3t), 30t\sin(3t), 100t)^\top, \quad t \in [5\pi,\, 9\pi]. \tag{3.53}$$

This trajectory was chosen to get the ITD values and associated visual disparities at various depths and angles. We imitated the natural limits to the visual field of view that restricts visual observations to lie within a fixed conic volume. The observations in visual and auditory spaces were produced according to models (3.23) and (3.24). Detector failure levels $1 - \pi_*$ and $1 - \lambda_*$ are taken to be equal to 0.05 for both modalities. Detector noise is taken normally distributed with (co)variances

$$\boldsymbol{\Sigma} = \begin{pmatrix} 10^{-4} & 0 & 0 \\ 0 & 10^{-4} & 0 \\ 0 & 0 & 10^{-11} \end{pmatrix} \quad \text{and} \quad \kappa = 10^{-1/2}, \tag{3.54}$$



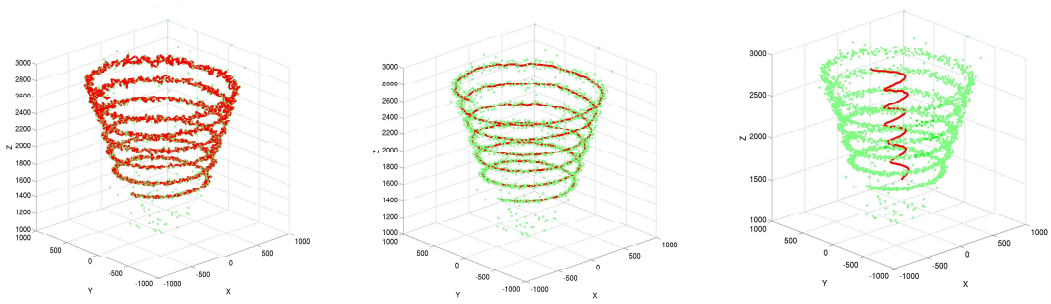
(a) (b)

Figure 3.3: Simulated experiments data. (a) Reconstructed video observations $\mathbf{f} = \{\boldsymbol{f}_m\}_{m=1}^{M}$ in the ambient space, and (b) audio observations $\boldsymbol{g} = \{g_k\}_{k=1}^{K}$ in the ITD space for a spiral trajectory $\boldsymbol{s}(t) = (30t\cos(3t), 30t\sin(3t), 100t)^\top$. Data is simulated using generative observation models, visual data is mapped to auditory domain using ground truth microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$.

for visual and auditory data respectively. Microphones are located at $\boldsymbol{s}^*_{\mathrm{M}_\ell} = (-85, 120, 10)^\top$ and $\boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}} = (75, 110, -15)^\top$. These coordinates are given in millimeters, so the inter-microphone distance is about 16cm. The generated data in auditory and visual domains is shown in Figure 3.3. Visual observations are mapped into the ambient 3D space and into the ITD domain using ground truth microphone locations $\boldsymbol{s}^*_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^*_{\mathrm{M}_\mathrm{r}}$. Auditory data is taken rounded to imitate the discretization effect.

We assume the auditory and visual data to be acquired at different frequencies: 25Hz for video and 75Hz for audio. This results in total of about $M = 3000$ video and $K = 9000$ audio observations that are not aligned. Below we provide details on the stages of the optimization algorithm.

The initial sampling of the trajectory $\mathbf{s}^{(0)}$ follows the procedure described in Section 3.3.2. The regularization parameter $\tilde{\gamma}$ is taken uniformly distributed on $[10^{-7}, 10^{-3}]$. Sometimes when the regularization term is overweighted, the model tends to infer trajectories $\tilde{\mathbf{f}}$ that are too smooth. As a side effect, the algorithm assigns all the observations to the uniform component, considering them as erroneous and converges to very small values of the prior $\tilde{\pi}$, as can be seen from (3.43). Two solutions can be proposed in this case. Firstly, one can reduce the support of the distribution for $\tilde{\gamma}$. Secondly, if one has some a priori knowledge on the amount of detector failures, it is possible to include prior distribution on the values of $\tilde{\pi}$ into (3.39) and (3.40). Then the E-step of the EM algorithm is still given by (3.41). The M-step expression (3.43) for the optimal value of $\tilde{\pi}$ would change. Assuming



$\tilde{\gamma} = 10^{-7}$ $\tilde{\gamma} = 10^{-3}$ $\tilde{\gamma} = 6 \cdot 10^{-3}$

Figure 3.4: Regularized sampled trajectories for various parameter $\tilde{\gamma}$ values, resulting in underweighted (left), normal (centre) and overweighted (right) regularization term cases, are shown in the ambient 3D space $\mathbb{S}$. Green dots correspond to visual observations mapped into $\mathbb{S}$. Trajectories are depicted with red lines.

Gaussian prior $\mathcal{N}(\tilde{\pi};\ \pi_0, \varrho)$ on $\tilde{\pi}$ leads to a cubic equation

$$\tilde{\pi}\left(M + 2\varrho^{-2}(\tilde{\pi} - \tilde{\pi}_0)(1 - \tilde{\pi})\right) = \sum_{m=1}^{M} \tilde{\alpha}_m^{(q)}, \tag{3.55}$$

for $0 < \tilde{\pi} < 1$. In our experiments we used the second approach and took $\pi_0 = 0.9$ and $\varrho = 0.01$ for the prior distribution. Examples of regularized trajectories for $\tilde{\gamma} = 10^{-7}$, $\tilde{\gamma} = 10^{-3}$ and $\tilde{\gamma} = 6 \cdot 10^{-3}$ are given in Figure 3.4.

The left and right microphone locations were initialized to $\boldsymbol{s}_{\mathrm{M}_\ell} = (-120, 0, 0)$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}} = (120, 0, 0)$ respectively. To set up the initial algorithm parameters $\boldsymbol{\psi}^{(0)}$ and find $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}\}$ we ran 10 iterations of the EM algorithm (3.51) with 100 optimization iterations of the M-step (3.52). Two possibilities were considered for the prior distribution $P(\boldsymbol{\theta})$ in (3.52).

Uniform microphone locations prior $\mathcal{U}(\boldsymbol{\theta}, \Theta)$ was based on the assumption that the microphone pair center $\frac{1}{2}(\boldsymbol{s}_{\mathrm{M}_\ell} + \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})$ is known up to 10cm in each coordinate, and microphone rotations with respect to the center lie within a range of $\pi/4$. These conditions are naturally verified, as soon as a real-life head-like device has microphones that are physically located close to the cameras and the mentioned precision in support domain estimation can be easily achieved.

Normal prior $\mathcal{N}(\boldsymbol{\theta};\ \hat{\boldsymbol{\theta}}, \boldsymbol{\Gamma})$ was taken centered at the initial microphone locations guesses $\boldsymbol{s}_{\mathrm{M}_\ell} = (-120, 0, 0)$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}} = (120, 0, 0)$ with the covariance matrix $\boldsymbol{\Gamma} = 50\mathbb{I}$, where $\mathbb{I}$ is the identity matrix.

The results on the initialized microphone positions $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}\}$ are presented in Table 3.1. The estimated initial left and right microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}^{(0)}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{(0)}$ are compared to the corresponding ground truth values $\boldsymbol{s}_{\mathrm{M}_\ell}^{*}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}^{*}$. The absolute errors e_ℓ

		$p=1$			$p=1.5$			$p=2$		
		x	y	z	x	y	z	x	y	z
$\mathcal{N}$	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$	-87.69	70.22	9.91	-92.78	45.34	7.56	-75.3	67.62	11.56
	$\boldsymbol{s}^{*}_{\mathrm{M}_\ell}$	-85	120	10	-85	120	10	-85	120	10
	e_ℓ	2.69	49.78	0.09	7.78	74.66	2.44	9.7	52.38	1.56
	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$	75.48	60.02	-15.08	69.37	34.24	-17.64	85.76	57.24	-12.5
	$\boldsymbol{s}^{*}_{\mathrm{M}_\mathrm{r}}$	75	110	-15	75	110	-15	75	110	-15
	e_r	0.48	49.98	0.08	5.63	75.86	2.64	10.76	62.76	2.5
	$\mathcal{L}^{(0)}$	1400.18			2931.79			3523.09		
	$\mathcal{L}^{*}$	1237.75			2828.76			3442.6		
$\mathcal{U}$	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$	-81.39	139.89	5.03	-89.82	120.72	-7.07	-86.36	146.88	-9.47
	$\boldsymbol{s}^{*}_{\mathrm{M}_\ell}$	-85	120	10	-85	120	10	-85	120	10
	e_ℓ	3.61	19.89	4.97	4.82	0.72	17.07	1.36	26.88	19.47
	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$	82.56	129.73	-19.81	73.54	110.54	-32.35	76.17	137.05	-34.63
	$\boldsymbol{s}^{*}_{\mathrm{M}_\mathrm{r}}$	75	110	-15	75	110	-15	75	110	-15
	e_r	7.56	19.73	4.81	1.46	0.54	17.35	1.17	27.05	19.63
	$\mathcal{L}^{(0)}$	1437.45			2970.74			3563.53		
	$\mathcal{L}^{*}$	1351.75			2942.76			3556.6		

Table 3.1: Microphone locations initialization results for simulated data. Tables show estimated initial left and right microphone locations $\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$, their ground truth values $\boldsymbol{s}^{*}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^{*}_{\mathrm{M}_\mathrm{r}}$ and the absolute errors e_ℓ and e_r (in mm), evaluated for each coordinate between ground truth and estimated microphone positions. Dependency on the generalized Gaussian distribution parameter p and the type of prior $P(\boldsymbol{\theta})$ - uniform ($\mathcal{U}$) or Gaussian ($\mathcal{N}$), is shown. For every p we also give the log-likelihoods $\mathcal{L}^{(0)}$ and $\mathcal{L}^{*}$ of the observed data based on estimated microphone locations and the ground truth ones respectively.

and e_r are evaluated for each coordinate. The dependency on the generalized Gaussian distribution parameter p and the prior distribution on the misrophone locations $P(\boldsymbol{\theta})$ - uniform ($\mathcal{U}$) or Gaussian ($\mathcal{N}$), is shown. For every p we provide the log-likelihoods $\mathcal{L}^{(0)}$ and $\mathcal{L}^{*}$ of the observed data based on estimated microphone locations and the ground truth ones respectively.

Further optimization was performed by 10 iterations of the alternating EM algorithm with 100 optimization iterations on each M-step. The final results are given in Table 3.2. Initial left and right microphone locations $\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$ are compared to the corresponding final estimated locations $\hat{\boldsymbol{s}}_{\mathrm{M}_\ell}$ and $\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}}$. Changes in absolute errors with respect to initial estimates are colour-coded: improvement is shown in green, deterioration - in red.

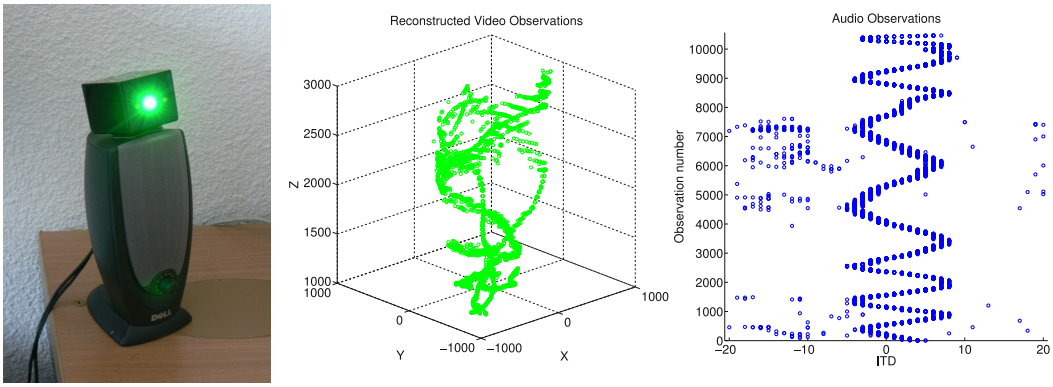
These results show that in general a uniform prior with reasonable bounds gives much better results than a Gaussian prior, unless the mean of the latter lies in proximity of the ground truth microphone locations, which one cannot presume. There is no clear dependency on the generalized Gaussian distribution parameter p, so preference should be given to the standard Gaussian distribution ($p = 2$) to gain in computation speed.

		$p=1$			$p=1.5$			$p=2$		
		x	y	z	x	y	z	x	y	z
$\mathcal{N}$	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$	-87.69	70.22	9.91	-92.78	45.34	7.56	-75.3	67.62	11.56
	$\hat{\boldsymbol{s}}_{\mathrm{M}_\ell}$	-88.8	55.15	6.61	-84.92	59.62	5.46	-75.3	67.62	11.56
	$\boldsymbol{s}^{*}_{\mathrm{M}_\ell}$	-85	120	10	-85	120	10	-85	120	10
	e_ℓ	3.8	64.85	3.39	0.08	60.38	7.56	9.7	52.38	1.56
	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$	75.48	60.02	-15.08	69.37	34.24	-17.64	85.76	57.24	-12.5
	$\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}}$	75.04	43.65	-18.37	77.32	49.17	-19.22	85.76	57.24	-12.5
	$\boldsymbol{s}^{*}_{\mathrm{M}_\mathrm{r}}$	75	110	-15	75	110	-15	75	110	-15
	e_r	0.04	66.35	3.37	2.32	60.83	4.22	10.76	62.76	2.5
	$\mathcal{L}^{(0)}$	1400.18			2931.79			3523.09		
	$\hat{\mathcal{L}}$	1402.02			2935.57			3523.09		
$\mathcal{U}$	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$	-81.39	139.89	5.03	-89.82	120.72	-7.07	-86.36	146.88	-9.47
	$\hat{\boldsymbol{s}}_{\mathrm{M}_\ell}$	-86.7	132.42	-1.82	-89.82	120.72	-7.07	-85.06	140.9	-7.74
	$\boldsymbol{s}^{*}_{\mathrm{M}_\ell}$	-85	120	10	-85	120	10	-85	120	10
	e_ℓ	1.7	12.42	11.82	4.82	0.72	17.07	0.06	20.9	17.74
	$\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$	82.56	129.73	-19.81	73.54	110.54	-32.35	76.17	137.05	-34.63
	$\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}}$	77.61	122.23	-26.99	73.54	110.54	-32.35	77.33	131	-32.78
	$\boldsymbol{s}^{*}_{\mathrm{M}_\mathrm{r}}$	75	110	-15	75	110	-15	75	110	-15
	e_r	2.61	12.23	11.99	1.46	0.54	17.35	2.33	21	17.78
	$\mathcal{L}^{(0)}$	1437.45			2970.74			3563.53		
	$\hat{\mathcal{L}}$	1437.59			2970.74			3563.62		

Table 3.2: Estimated microphone locations for simulated data. Tables show initial left and right microphone locations $\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$, the corresponding final estimated locations $\hat{\boldsymbol{s}}_{\mathrm{M}_\ell}$ and $\hat{\boldsymbol{s}}_{\mathrm{M}_\mathrm{r}}$ and ground truth values $\boldsymbol{s}^{*}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}^{*}_{\mathrm{M}_\mathrm{r}}$ and the absolute errors e_ℓ and e_r (in mm), evaluated for each coordinate between ground truth and estimated microphone positions. Colour designates whether the initial result was improved (green) or not (red). For every p we also give the log-likelihoods $\mathcal{L}^{(0)}$ and $\hat{\mathcal{L}}$ of the observed data based on initial microphone locations and the estimated ones respectively.

The ground truth parameters do not represent a stationary point of the log-likelihood due to ITD observations discretization effect that is not included explicitly into the model. So the optimal estimates are likely to lie in some neighbourhood of the ground truth parameters, but not coincide with them. The size of this neighbourhood tends to be smaller for the uniform distribution, as soon as it influences less the target function (3.52) and the log-likelihood. The resulting precision is about 1-2cm for each microphone, which is comparable to the sensor size. This gives perfect observation alignment in the auditory domain.

Likelihood values comparison shows that the principal role in the optimization procedure is played by the microphone locations inference (3.51) and the initial trajectory sampling (see Table 3.1). Further point-by-point trajectory optimization alternated with



(a) (b) (c)

Figure 3.5: Calibration rig *Altair* and the acquired data used for audio-visual head-like calibration. (a) The rig consists of a speaker with an LED light bulb mounted on its top. Small source of bright light emitting white noise provides high quality of auditory and visual observations in most of the environments. (b) Reconstructed visual observations $\mathbf{f} = \{\boldsymbol{f}_m\}_{m=1}^{M}$ in the ambient space, and (c) auditory observations $\boldsymbol{g} = \{g_k\}_{k=1}^{K}$ in the ITD space. Trajectory is chosen so as to produce as much correspondences between auditory and visual domains, as possible.

microphone locations optimization does not improve the results much, (see Table 3.2).

3.4.2 Experiments with Real Data

When performing real-data experiments, it is essential to assure the best possible precision of observations in both modalities. Therefore in our case the calibration rig should fulfil two requirements. Firstly, it should be clearly detected in both camera images in regular lighting conditions and should be small enough to be considered as a point in 3D. Secondly, it should emit sound such that the ITD calculation method is robust to natural reverberations and acoustic noise.

The calibration rig *Altair* used in our experiments is presented in Figure 3.5(a). It consists of a speaker with an LED light bulb mounted on its top. While Altair was moving inside the room white noise played through its speaker. Together with a bright light source this ensures high quality of auditory and visual observations for most of the environments. The spiral trajectory was chosen, as in the simulated data case, to better cover the hidden space locations.

We used feature detection algorithms described in detail in Sections 2.2 and 2.3. The extracted data is shown in Figure 3.5(b) and (c). The auditory and visual observation sets resemble the ones that were generated for simulated data experiments, Figure 3.3. To synchronize auditory and visual streams we use a clapper device, as when recording the CAVA database. Images from the stereoscopic camera pair are acquired at 25fps, each one of them is timestamped. The two audio streams are sampled at 44.1kHz.

(a) (b)

Figure 3.6: Calibration results for the audio-visual data acquired in real environment. (a) Reconstructed visual observations $\mathbf{f} = \{\boldsymbol{f}_m\}_{m=1}^{M}$ (green) and the estimated trajectory $\mathbf{s} = \{\boldsymbol{s}_n\}_{n=1}^{N}$ (red) are shown in the ambient space, (b) Auditory observations $\boldsymbol{g} = \{g_k\}_{k=1}^{K}$ (blue) with visual observations and the estimated trajectory mapped through $\mathcal{F}$ and $\mathcal{G}$ into the ITD space. Estimated microphone locations $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ correspond well to the configuration used in the experiments, perfect alignment of the trajectories in the ITD space is achieved.

The estimated trajectory and microphone locations are shown on Figure 3.6. Left microphone position $\boldsymbol{s}_{\mathrm{M}_\ell} = (-101.866021, 27.905023, 48.918872)^\top$ and right microphone position $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}} = (51.322268, 75.526212, 56.759726)^\top$ that were found using the calibration procedure correspond well to the configuration we used to gather the data. Figure 3.6 shows that auditory and visual domains are well aligned for all the positions that lie on the spiral trajectory.

3.5 Discussion

Multimodal multisensor calibration is a challenging task that is characterized by high dimensionality of the parameter space, diversity of modality spaces and considerable observation noise levels. In such conditions both, the calibration rig and the calibration method should be developed so as to reduce noise effects. We presented a general approach to multimodal multisensor calibration based on calibration rig simultaneous tracking in multiple modalities.

The problem formulation and the algorithm that we proposed in this Chapter possess several benefits that we outline below

- **Geometry Consistency**: as opposed to a number of methods built over approximating assumptions on affine dependency between auditory and visual observations that

hold only for the visible region and only at a certain distance from the sensors, we base on physical models of the sensors, so that our assumptions are consistent with the 3D geometry of the ambient space;

- **Robustness**: various kinds of noise are taken into account by the model, such as detector failures and observation noise; the outliers are automatically detected and not included into the calibration;
- **Persistent Target Function Increase**: the alternating optimization approach uses the EM algorithm for the two maximization steps that possesses the non-decrease property;
- **Acceleration**: various techniques to speed-up the calibration can be employed, in case of AV head-like device we propose visual space trajectory sampling as to increase the convergence speed;
- **Prior Information**: there is a possibility to include prior information on calibration parameters that does not affect much the optimization procedure, but can significantly improve the results;

The comparison of performance on simulated data showed that 1-2cm precision on microphone locations is achieved. These results are confirmed by the real-data experiments that produce very good alignment of the auditory and visual spaces. Considering the calibration task in the ambient 3D space, we found a good trade-off between the expressiveness of the calibration process model and efficiency of the calibration procedure. Multiple constraints allow us to improve the convergence speed, while keeping the results precise.

Future developments for AV head-like device calibration can address the issue of visual field restriction. Indeed, the more different audio-visual correspondances one gets, the more precise are the obtained results. In case of a motor-controlled robot head one could verge the cameras while keeping the microphones still. This would increase the set of possible trajectories $\mathbf{s}$ allowing for better calibration results.

CHAPTER 4

Spatial Multimodal Clustering

Sommaire

4.1 Unsupervised Clustering of Multimodal Data 43
4.2 Conjugate Mixture Models for Multimodal Data 46
4.3 Conjugate KP Algorithm for Clustering Multimodal Data 49
4.3.1 The Penalization Step 50
4.3.2 The Maximization Step 51
4.3.3 Generalized KP for Conjugate Mixture Models 52
4.3.4 Identifiability and Algorithm Convergence 53
4.4 Experimental Evaluation 54
4.5 Discussion 61

The general problem of how to determine properties of several objects that are observed simultaneously by different sensors is considered in this Chapter. The goal is to split each sensor's data set into groups that correspond to the objects, so that these groups stay coherent across data sets from different sensors. In what follows we refer to this task as 'multimodal clustering'. We formalize the problem considering it in the framework of *conjugate mixture models*. We discuss convergence properties of the proposed algorithm and consider different strategies to infer the object properties. The algorithms are verified on simulated data, their performance in various cases depending on object and detector properties is compared.

4.1 Unsupervised Clustering of Multimodal Data

The unsupervised clustering of multimodal data is a key capability whenever the goal is to group observations that are gathered using several physically different sensors. A typical example is the computational modelling of biological *multisensory perception*. This includes the issues of how a human detects objects that are both seen and touched [Pouget 2002a, Ernst 2002], seen and heard [Anastasio 2000, King 2004, King 2005] or how a human localizes one source of sensory input in a natural environment in the presence of competing stimuli and of a variety of noise sources [Haykin 2005]. More generally, *multisensory fusion* [Hall 2004, Mitchell 2007] is highly relevant in various other research domains, such as target tracking [Smith 2006] based on radar and

sonar data [Naus 2004, Coiras 2007], mobile robot localization with laser rangefinders and cameras [Castellanos 1999], robot manipulation and object recognition using both tactile and visual data [Allen 1995, Joshi 1999], underwater navigation based on active sonar and underwater cameras [Majumder 2001], audio-visual speaker detection [Beal 2003, Perez 2004, Fisher III 2004], speech recognition [Heckmann 2002, Nefian 2002, Shao 2008], and so forth.

When the data originates from a single object, finding the best estimates for the object's characteristics is usually referred to as a *pure fusion* task and it reduces to combining multisensor observations in some optimal way [Beal 2003, Kushal 2006, Smith 2006]. For example, land and underwater robots fuse data from several sensors to build a 3D map of the ambient space not considering individual objects present in the environment [Castellanos 1999, Majumder 2001]. The problem is much more complex when several objects are present and when the task implies their detection, identification, and localization. In this case one has to consider two processes simultaneously: *(i) segregation* [Fisher III 2001] which assigns each observation either to an object or to an *outlier* category and *(ii) estimation* which computes the parameters of each object based on the group of observations that were assigned to that object. In other words, in addition to fusing observations from different sensors, multimodal analysis requires the assignment of each observation to one of the objects.

This observation-to-object association problem can be cast into a probabilistic framework. Recent multisensor data fusion methods able to handle several objects are based on particle filters [Checka 2004, Chen 2004, Gatica-Perez 2007]. Notice, however, that the dimensionality of the parameter space grows exponentially with the number of objects, causing the number of required particles to increase dramatically and augmenting computational costs. A number of efficient sampling procedures were suggested [Chen 2004, Gatica-Perez 2007] to keep the problem tractable. Of course this is done at the cost of loss in model generality, and hence these attempts are strongly application-dependent. Another drawback of such models is that they cannot provide estimates of accuracy and importance of each modality with respect to each object. The sampling and distribution estimation are performed in the parameter space, but no statistics are gathered for the observation spaces. Recently [Hospedales 2008] extended the single-object model of [Beal 2003] to multiple objects: several trained single-object models are incorporated into the multiple-object model that uses an additional type of audio association to detect situations where audio signal is speech, but does not correspond to person's location in an image. This method's complexity is linear in the number of objects. However, we would like to address the problem of clustering of AV data in a completely unsupervised context and rely only on spatial and temporal coherence of the observations, but not on any trained parameters.

In the case of unimodal data, the problems of grouping observations and of associating groups with objects can be cast into the framework of standard data clustering which can be solved using a variety of parametric or non-parametric techniques. The problem of *clustering multimodal data* raises the difficult question of how to group together observations that belong to different physical spaces with different dimensionalities, e.g., how to group visual data with auditory data? When the observations from two different modalities can

be *aligned* pairwise, a natural solution is to consider the Cartesian product of two unimodal spaces. Unfortunately, such an alignment is not possible in most practical cases. Different sensors operate at different frequency rates and hence the number of observations gathered with one sensor can be quite different from the number of observations gathered with another sensor. Consequently, there is no obvious way to align the observations pairwise. Considering all possible pairs would result in a combinatorial blow-up and typically create abundance of erroneous observations corresponding to inconsistent solutions.

Alternatively, one may consider several unimodal clusterings, provided that the relationships between a common object space and several observation spaces can be explicitly specified. *Multimodal clustering* then results in a number of unimodal clusterings that are jointly governed by the same unknown parameters characterizing the object space. We note that binding unimodal clusters through common parameters allows to correctly model situations where object is not observed in one of the observation spaces (e.g. a person that is visible and silent).

The original contribution of this Chapter is to show how the problem of *clustering multimodal data* can be addressed within the framework of mixture models [McLachlan 2000]. We consider the Kullback-Proximal (KP) algorithm [Chrétien 2000, Chrétien 2008] specifically designed to estimate object-space parameters that are indirectly observed in several sensor spaces. As a special case it includes the expectation-maximization (EM) algorithm [Dempster 1977] for a certain choice of the gain sequence. The convergence properties of the proposed KP algorithm are thoroughly investigated and several efficient implementations are described in detail. The proposed model is composed of a number of modality-specific mixtures. These mixtures are jointly governed by a set of common *object-space parameters* (which will be referred to as the *tying parameters*), thus ensuring consistency between the sensory data and the object space being sensed. This is done using explicit transformations from the unobserved parameter space (object space) to each of the observed spaces (sensor spaces). Hence, the proposed model is able to deal with observations that live in spaces with different physical properties such as dimensionality, space metric, sensor sampling rate, etc. We believe that linking the object space with the sensor spaces based on object-space-to-sensor-space transformations has more discriminative power than existing multisensor fusion techniques and hence performs better in terms of multiple object identification and localization. To the best of our knowledge, there has been no attempt to use a generative model, such as ours, for the task of multimodal data interpretation.

In Section 4.2 we formally introduce the concept of *conjugate mixture models*. Standard Gaussian mixture models (GMM) are used to model the data in each modality. The parameters of these Gaussian mixtures are governed by the object parameters through a number of object-space-to-sensor-space transformations (one transformation for each sensing modality). In Section 4.3 we cast the multimodal data clustering problem in the framework of maximum likelihood estimation and we explicitly derive the penalization and maximization steps of the associated KP algorithm that lead to a set of fixed point equations. The convergence properties are discussed and various existing optimization methods [Polyak 1987, Zhigljavsky 1991, Zhigljavsky 2008] are considered to solve the

equations. Section 4.4 illustrates the proposed method with the task of audio-visual object detection and localization using binocular vision and binaural hearing and analyses in detail the performances of the proposed model under various practical conditions on simulated data. Finally, Section 4.5 concludes the Chapter and provides directions for further improvements.

4.2 Conjugate Mixture Models for Multimodal Data

We consider N objects $n = 1 \dots N$. Each object n is characterized by a parameter vector of dimension d, denoted by $\boldsymbol{s}_n \in \mathbb{S} \subseteq \mathbb{R}^d$. The set $\mathbf{s} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n, \dots, \boldsymbol{s}_N\}$ corresponds to the unknown *tying parameters*. The objects are observed with a number of physically different sensors. Although, for the sake of clarity, we will consider two modalities, generalization is straightforward. Therefore, the observed data consists of two sets of observations denoted respectively by $\mathbf{f} = \{\boldsymbol{f}_1, \dots, \boldsymbol{f}_m, \dots, \boldsymbol{f}_M\}$ and $\mathbf{g} = \{\boldsymbol{g}_1, \dots, \boldsymbol{g}_k, \dots, \boldsymbol{g}_K\}$ lying in two different observation spaces of dimensions r and p, $\boldsymbol{f}_m \in \mathbb{F} \subseteq \mathbb{R}^r$ and $\boldsymbol{g}_k \in \mathbb{G} \subseteq \mathbb{R}^p$.

We introduce the *conjugate mixture models* framework that explicitly takes into account dependencies between the observation spaces. One key ingredient of our approach is that we consider the transformations:

$$\begin{cases} \mathcal{F} : \mathbb{S} \to \mathbb{F} \\ \mathcal{G} : \mathbb{S} \to \mathbb{G} \end{cases} \tag{4.1}$$

that map $\mathbb{S}$ into the observation spaces $\mathbb{F}$ and $\mathbb{G}$ respectively. These transformations are defined by the physical and geometric properties of the sensors and they are supposed to be known. We treat the general case when both $\mathcal{F}$ and $\mathcal{G}$ are non-linear.

An assignment variable is associated with each observation, thus indicating the object that generated the observation: $\boldsymbol{A} = \{A_1, \dots, A_m, \dots, A_M\}$ and $\boldsymbol{B} = \{B_1, \dots, B_k, \dots, B_K\}$. Hence, the segregation process is cast into a hidden variable problem. The notation $A_m = n$ (resp. $B_k = n$) means that the observation $\boldsymbol{f}_m$ (resp. $\boldsymbol{g}_k$) was generated by object n. In order to account for erroneous observations, an additional $N+1$-th fictitious object is introduced to represent an outlier category. The notation $A_m = N+1$ (resp. $B_k = N+1$) means that $\boldsymbol{f}_m$ (resp. $\boldsymbol{g}_k$) is an outlier. Note that we will also use the following standard convention: upper case letters for random variables ($\boldsymbol{A}$ and $\boldsymbol{B}$) and lower case letters for their realizations ($\boldsymbol{a}$ and $\boldsymbol{b}$). The usual conditional independence assumption leads to:

$$P(\mathbf{f}, \mathbf{g} | \boldsymbol{a}, \boldsymbol{b}) = \prod_{m=1}^{M} P(\boldsymbol{f}_m | a_m) \prod_{k=1}^{K} P(\boldsymbol{g}_k | b_k). \tag{4.2}$$

In addition, all assignment variables are assumed to be independent, i.e.:

$$P(\boldsymbol{a}, \boldsymbol{b}) = \prod_{m=1}^{M} P(a_m) \prod_{k=1}^{K} P(b_k). \tag{4.3}$$

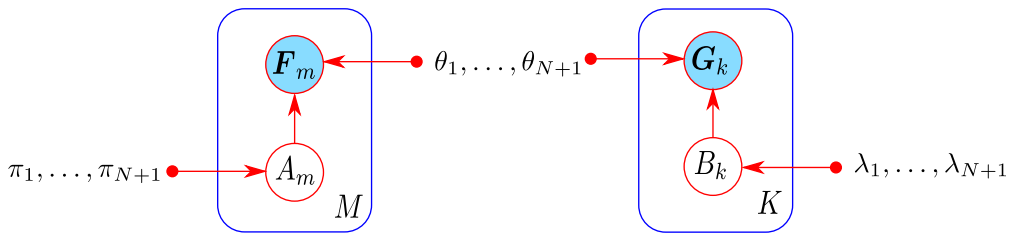


Figure 4.1: Graphical representation of a general conjugate mixture model. Circles denote random variables, plates (rectangles) around them represent multiple similar nodes, their number being given in the plates.

As discussed in Section 4.5, more general cases could be considered. However, we focus on the independent case for it captures most of the features relevant to the conjugate clustering task and because more general dependence structures could be reduced to the independent case via the use of appropriate variational approximation techniques [Jordan 1998, Celeux 2003].

Next we define the following probability density functions, for all $n = 1 \ldots N, N+1$, for all $\boldsymbol{f}_m \in \mathbb{F}$ and for all $\boldsymbol{g}_k \in \mathbb{G}$:

$$P_n^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}_n) = P(\boldsymbol{f}_m | A_m = n;\ \boldsymbol{\theta}_n), \tag{4.4}$$

$$\text{and} \quad P_n^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}_n) = P(\boldsymbol{g}_k | B_k = n;\ \boldsymbol{\theta}_n), \tag{4.5}$$

with parameters $\boldsymbol{\theta}_n$ that describe cluster properties. We introduce the prior probabilities $\pi = (\pi_1, \ldots, \pi_n, \ldots, \pi_{N+1})$ and $\lambda = (\lambda_1, \ldots, \lambda_n, \ldots, \lambda_{N+1})$:

$$\pi_n = P(A_m = n), \quad \forall m = 1 \ldots M, \tag{4.6}$$

$$\lambda_n = P(B_k = n), \quad \forall k = 1 \ldots K. \tag{4.7}$$

Therefore, $\boldsymbol{f}_m$ and $\boldsymbol{g}_k$ are distributed according to two $(N+1)$-component mixture models:

$$P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}) = \sum_{n=1}^{N+1} \pi_n P_n^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}_n), \tag{4.8}$$

$$\text{and} \quad P^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}) = \sum_{n=1}^{N+1} \lambda_n P_n^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}_n), \tag{4.9}$$

where $\boldsymbol{\theta} = \{\pi_n, \ldots, \pi_{N+1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N+1}\}$. The log-likelihood of the observed data is:

$$\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}) = \sum_{m=1}^{M} \log P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}) + \sum_{k=1}^{K} \log P^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}) =$$

$$= \sum_{m=1}^{M} \log \left(\sum_{n=1}^{N+1} \pi_n P_n^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}_n) \right) + \sum_{k=1}^{K} \log \left(\sum_{n=1}^{N+1} \lambda_n P_n^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}_n) \right). \tag{4.10}$$

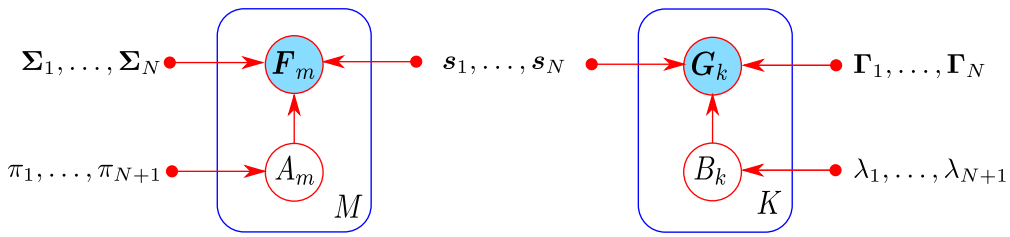


Figure 4.2: Graphical representation of Gaussian conjugate mixture models. Circles denote random variables, plates (rectangles) around them represent multiple similar nodes, their number being given in the plates.

The graphical representation of our conjugate mixture model is shown in Figure 4.1. We adopted the graphical notation introduced in [Bishop 2006] to represent similar nodes in a more compact way: the M (resp. K) similar nodes are indicated with a *plate*. The two sensorial modalities are linked by the *tying parameters* $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_{N+1}$ shown in between the two plates.

Various choices can be made for parameters $\boldsymbol{\theta}_n$ and distributions $P_n^{\mathbb{F}}$ and $P_n^{\mathbb{G}}$. In this Chapter we consider Gaussian distribution family for both modalities and $n = 1, \ldots, N$

$$P_n^{\mathbb{F}}(\boldsymbol{f}_m, \boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{f}_m;\ \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n), \tag{4.11}$$

$$\text{and} \quad P_n^{\mathbb{G}}(\boldsymbol{g}_k, \boldsymbol{\theta}_n) = \mathcal{N}(\boldsymbol{g}_k;\ \mathcal{G}(\boldsymbol{s}_n), \boldsymbol{\Gamma}_n), \tag{4.12}$$

where $\boldsymbol{\theta}_n = \{\boldsymbol{s}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\Gamma}_n\}$. We denoted

$$\mathcal{N}(\boldsymbol{f}_m;\ \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n) = \frac{1}{(2\pi)^{r/2}|\boldsymbol{\Sigma}_n|^{1/2}} \exp\left(-\frac{1}{2}\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n}\right). \tag{4.13}$$

The notation $\|\boldsymbol{v} - \boldsymbol{w}\|^2_{\boldsymbol{\Sigma}}$ stands for the Mahalanobis distance $(\boldsymbol{v} - \boldsymbol{w})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{v} - \boldsymbol{w})$ and $^\top$ stands for the transpose of a matrix. Formula analogous to (4.13) is taken for $P_n^{\mathbb{G}}$. Of course, other distribution families could also have been employed. In fact, the model permits to parametrize each cluster in any observation space in its own manner. Though without any prior knowledge on the objects we chose distributions from the same family for the same observation space.

The outlier class is taken to be uniform

$$P_{N+1}^{\mathbb{F}}(\boldsymbol{f}_m) = \mathcal{U}(\boldsymbol{f}_m; V), \tag{4.14}$$

$$P_{N+1}^{\mathbb{G}}(\boldsymbol{g}_k) = \mathcal{U}(\boldsymbol{g}_k; U), \tag{4.15}$$

where V and U denote the respective support volumes. In what follows we consider Gaussian mixture models with uniform outliers. However, an example of how to employ Student t-distribution mixtures for the same task is presented in Appendix A.2. The graphical model for the conjugate Gaussian mixtures is given on Figure 4.2.

We rewrite the mixtures (4.8) and (4.9) for the case of Gaussian distribution

$$P^{\mathbb{F}}(\boldsymbol{f}_m, \boldsymbol{\theta}_n) = \sum_{n=1}^{N} \pi_n \mathcal{N}(\boldsymbol{f}_m;\ \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n) + \pi_{N+1}\mathcal{U}(\boldsymbol{f}_m; V), \tag{4.16}$$

$$P^{\mathbb{G}}(\boldsymbol{g}_k, \boldsymbol{\theta}_n) = \sum_{n=1}^{N} \lambda_n \mathcal{N}(\boldsymbol{g}_k;\ \mathcal{G}(\boldsymbol{s}_n), \boldsymbol{\Gamma}_n) + \lambda_{N+1}\mathcal{U}(\boldsymbol{g}_k; U), \tag{4.17}$$

and the log-likelihood function

$$\begin{aligned}\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}) &= \sum_{m=1}^{M} \log\left(\sum_{n=1}^{N} \pi_n \mathcal{N}(\boldsymbol{f}_m;\ \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n) + \pi_{N+1}\mathcal{U}(\boldsymbol{f}_m; V)\right) + \\ &+ \sum_{k=1}^{K} \log\left(\sum_{n=1}^{N} \lambda_n \mathcal{N}(\boldsymbol{g}_k;\ \mathcal{G}(\boldsymbol{s}_n), \boldsymbol{\Gamma}_n) + \lambda_{N+1}\mathcal{U}(\boldsymbol{g}_k; U)\right)\end{aligned} \tag{4.18}$$

where:

$$\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_N, \pi_{N+1}, \lambda_1, \ldots, \lambda_N, \lambda_{N+1}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_N\} \tag{4.19}$$

denotes the set of all unknown parameters to be estimated.

4.3 Conjugate KP Algorithm for Clustering Multimodal Data

Given the probabilistic model just described, we wish to determine the parameter vectors $\boldsymbol{\theta}_n$ associated with the objects that generated observations in two different sensory spaces. The problem is stated as maximum likelihood (ML) estimation:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}), \tag{4.20}$$

and considered in the Kullback proximal framework.The basic idea of the KP algorithm resembles that of the Levenberg-Marquardt method [Polyak 1987]: it is an iterative optimization technique where the target function is penalized by an additional distance term at every iteration.

Definition 1 *Let $(h_q)_{q\in\mathbb{N}}$ be a sequence of positive real numbers. Then, the Kullback-proximal algorithm is defined by*

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, \mathcal{L}_{\mathrm{pen}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}), \tag{4.21}$$

with $$\mathcal{L}_{\mathrm{pen}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}) - h_q H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}), \tag{4.22}$$

and $$H(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = -\mathbb{E}[\log P(\boldsymbol{A}, \boldsymbol{B} \mid \mathbf{f}, \mathbf{g}; \boldsymbol{\theta}) | \mathbf{f}, \mathbf{g}; \tilde{\boldsymbol{\theta}}]. \tag{4.23}$$

The expectation in (4.23) is taken over the hidden variables $\boldsymbol{A}$ and $\boldsymbol{B}$. The following two results can be easily shown [Chrétien 2008]:

Proposition 1 *For any iteration $q \in \mathbb{N}$, the sequence $\boldsymbol{\theta}^{(q)}$ satisfies*

$$\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}^{(q+1)}) - \mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}^{(q)}) \geq h_q(H(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) - H(\boldsymbol{\theta}^{(q)}, \boldsymbol{\theta}^{(q)})) \geq 0. \tag{4.24}$$

Proposition 2 *The EM algorithm is a special instance of the Kullback-proximal algorithm with $h_q \equiv 1$.*

Remark 1 *In case $h_q \equiv 0$ the Kullback-proximal algorithm reduces to direct optimization of the log-likelihood function $\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta})$.*

Each iteration q of KP proceeds in two steps:

- *Penalization*. For the current values $\boldsymbol{\theta}^{(q)}$ of the parameters, compute the penalization term as the conditional expectation with respect to variables $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = - \sum_{\boldsymbol{a} \in \{1 \ldots N+1\}^M} \sum_{\boldsymbol{b} \in \{1 \ldots N+1\}^K} P(\boldsymbol{a}, \boldsymbol{b} | \mathbf{f}, \mathbf{g};\ \boldsymbol{\theta}^{(q)})\ \log P(\boldsymbol{a}, \boldsymbol{b} | \mathbf{f}, \mathbf{g};\ \boldsymbol{\theta}) \tag{4.25}$$

- *Maximization*. Update the parameter set $\boldsymbol{\theta}^{(q)}$ by performing maximization (4.21).

Proposition 1 shows that KP algorithm always increases the target function $\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta})$ in (4.18). Though the closed-form solution for (4.21) exists only in special cases. When the maximization (4.21) is difficult to achieve, various generalizations of KP are proposed. The maximization step can be relaxed by requiring just an increase rather than an optimum. This yieds Generalized KP (GKP) procedures that search for some $\boldsymbol{\theta}^{(q+1)}$ such that $\mathcal{L}_{\text{pen}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \geq \mathcal{L}_{\text{pen}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}^{(q)}, \boldsymbol{\theta}^{(q)})$. Therefore it provides a sequence of estimates that still verifies the non-decreasing likelihood property (4.24) although the convergence speed is likely to decrease and global optimality is not guaranteed. In the case of conjugate mixture models, we describe in more detail the specific forms of the two steps of the algorithm in the following Sections.

4.3.1 The Penalization Step

Using independency assumptions (4.2)-(4.3) the penalization term (4.25) can be decomposed as:

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = H_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) + H_{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}), \tag{4.26}$$

with

$$H_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = - \sum_{m=1}^{M} \sum_{n=1}^{N+1} \alpha_{mn}(\boldsymbol{\theta}^{(q)}) \log \alpha_{mn}(\boldsymbol{\theta}), \tag{4.27}$$

$$H_{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = - \sum_{k=1}^{K} \sum_{n=1}^{N+1} \beta_{kn}(\boldsymbol{\theta}^{(q)}) \log \beta_{kn}(\boldsymbol{\theta}), \tag{4.28}$$

where α_{mn} and β_{kn} denote the posterior probabilities as functions of parameters $\alpha_{mn}(\boldsymbol{\theta}) = P(A_m = n|\boldsymbol{f}_m;\boldsymbol{\theta})$ and $\beta_{kn}(\boldsymbol{\theta}) = P(B_k = n|\boldsymbol{g}_k;\boldsymbol{\theta})$. Their expressions can be derived straightforwardly from Bayes' theorem, $\forall n = 1\dots N$:

$$\alpha_{mn}(\boldsymbol{\theta}) = \frac{\pi_n P_n^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}_n)}{P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta})} = \frac{\pi_n \mathcal{N}(\boldsymbol{f}_m;\mathcal{F}(\boldsymbol{s}_n),\boldsymbol{\Sigma}_n)}{\sum\limits_{i=1}^{N}\pi_i\mathcal{N}(\boldsymbol{f}_m;\mathcal{F}(\boldsymbol{s}_i),\boldsymbol{\Sigma}_i) + V^{-1}\pi_{N+1}}, \tag{4.29}$$

$$\beta_{kn}(\boldsymbol{\theta}) = \frac{\lambda_n P_n^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}_n)}{P^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta})} = \frac{\lambda_n \mathcal{N}(\boldsymbol{g}_k;\ \mathcal{G}(\boldsymbol{s}_n),\boldsymbol{\Gamma}_n)}{\sum\limits_{i=1}^{N}\lambda_i\mathcal{N}(\boldsymbol{g}_k;\mathcal{G}(\boldsymbol{s}_i),\boldsymbol{\Gamma}_i) + U^{-1}\lambda_{N+1}}, \tag{4.30}$$

and $\alpha_{m,N+1}(\boldsymbol{\theta}) = 1 - \sum\limits_{n=1}^{N}\alpha_{mn}(\boldsymbol{\theta})$ and $\beta_{k,N+1}(\boldsymbol{\theta}) = 1 - \sum\limits_{n=1}^{N}\beta_{kn}(\boldsymbol{\theta})$.

4.3.2 The Maximization Step

We start with rewriting the expression for the penalized likelihood (4.21) using the expressions (4.18) and (4.26):

$$\begin{aligned}
\mathcal{L}_{\text{pen}}(\mathbf{f},\mathbf{g},\boldsymbol{\theta},\boldsymbol{\theta}^{(q)}) &= \sum_{m=1}^{M}\left(\log P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}) + h_q\sum_{n=1}^{N+1}\alpha_{mn}(\boldsymbol{\theta}^{(q)})\log\alpha_{mn}(\boldsymbol{\theta})\right) + \\
&+ \sum_{k=1}^{K}\left(\log P^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}) + h_q\sum_{n=1}^{N+1}\beta_{kn}(\boldsymbol{\theta}^{(q)})\log\beta_{kn}(\boldsymbol{\theta})\right) = \\
&= \sum_{m=1}^{M}\left((1-h_q)\log P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}) + h_q\sum_{n=1}^{N+1}\alpha_{mn}(\boldsymbol{\theta}^{(q)})\log(\pi_n P_n^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}_n))\right) + \\
&+ \sum_{k=1}^{K}\left((1-h_q)\log P^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}) + h_q\sum_{n=1}^{N+1}\beta_{kn}(\boldsymbol{\theta}^{(q)})\log(\lambda_n P_n^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}_n))\right),
\end{aligned} \tag{4.31}$$

In order to carry out the maximization of (4.31), its derivatives with respect to the model parameters are set to zero. In case of priors one obtains

$$\frac{\partial\mathcal{L}_{\text{pen}}(\mathbf{f},\mathbf{g},\boldsymbol{\theta},\boldsymbol{\theta}^{(q)})}{\partial\pi_n} = \pi_n^{-1}\sum_{m=1}^{M}\left((1-h_q)\alpha_{mn}(\boldsymbol{\theta}) + h_q\alpha_{mn}(\boldsymbol{\theta}^{(q)})\right). \tag{4.32}$$

We denote

$$\alpha_{mn}(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) = (1-h_q)\alpha_{mn}(\boldsymbol{\theta}) + h_q\alpha_{mn}(\tilde{\boldsymbol{\theta}}), \tag{4.33}$$

and repeat the same steps for λ_n to obtain the usual update expressions, i.e. $\forall n = 1,\dots,N+1$:

$$\pi_n^{(q+1)} = \frac{1}{M}\sum_{m=1}^{M}\alpha_{mn}(\boldsymbol{\theta}^{(q+1)},\boldsymbol{\theta}^{(q)}), \tag{4.34}$$

$$\lambda_n^{(q+1)} = \frac{1}{K}\sum_{k=1}^{K}\beta_{kn}(\boldsymbol{\theta}^{(q+1)},\boldsymbol{\theta}^{(q)}). \tag{4.35}$$

We note that these are not the closed form solutions, but rather fixed point equations that the solution should verify. However, for the EM algorithm $h_q \equiv 1$ the expressions (4.34) and (4.35) are closed form solutions, as soon as the first term in (4.33) disappears. Similar equations are derived for the optimal covariance matrices, $\forall n = 1, \ldots, N+1$:

$$\boldsymbol{\Sigma}_n^{(q+1)} = \frac{\sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \left(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q+1)})\right)\left(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q+1)})\right)^\top}{\sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})}, \quad (4.36)$$

$$\text{and} \quad \boldsymbol{\Gamma}_n^{(q+1)} = \frac{\sum_{k=1}^{K} \beta_{kn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \left(\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n^{(q+1)})\right)\left(\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n^{(q+1)})\right)^\top}{\sum_{k=1}^{K} \beta_{kn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})}. \quad (4.37)$$

For every $n = 1, \ldots, N$, the optimal tying parameter vector $\boldsymbol{s}_n^{(q+1)}$ is such that:

$$\begin{aligned} &\bar{\alpha}_n(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})(\bar{\boldsymbol{f}}_n - \mathcal{F}(\boldsymbol{s}_n^{(q+1)}))^\top \left(\boldsymbol{\Sigma}_n^{(q+1)}\right)^{-1} \mathcal{F}'(\boldsymbol{s}_n^{(q+1)}) \quad + \\ &+\bar{\beta}_n(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})(\bar{\boldsymbol{g}}_n - \mathcal{G}(\boldsymbol{s}_n^{(q+1)}))^\top \left(\boldsymbol{\Gamma}_n^{(q+1)}\right)^{-1} \mathcal{G}'(\boldsymbol{s}_n^{(q+1)}) \quad = \quad \mathbf{0}, \end{aligned} \quad (4.38)$$

where we denoted $\mathcal{F}'$ and $\mathcal{G}'$ the Jacobian matrices of $\mathcal{F}$ and $\mathcal{G}$ respectively and

$$\bar{\alpha}_n(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) = \sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}), \quad (4.39)$$

$$\text{and} \quad \bar{\beta}_n(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) = \sum_{k=1}^{K} \beta_{kn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}), \quad (4.40)$$

$$\text{and} \quad \bar{\boldsymbol{f}}_n = \bar{\alpha}_n(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})^{-1} \sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \boldsymbol{f}_m, \quad (4.41)$$

$$\text{and} \quad \bar{\boldsymbol{g}}_n = \bar{\beta}_n(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})^{-1} \sum_{k=1}^{K} \beta_{kn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \boldsymbol{g}_k. \quad (4.42)$$

The optimal solution $\boldsymbol{\theta}^{(q+1)}$ for the penalized likelihood maximization problem (4.21) is found as a fixed point of the system of equations (4.34)-(4.37) and a solution to the implicit function equation (4.38). If one of the mappings, $\mathcal{F}(\boldsymbol{s})$ or $\mathcal{G}(\boldsymbol{s})$ is injective such that its differential $\mathrm{d}\mathcal{F}(\boldsymbol{s})$ or $\mathrm{d}\mathcal{G}(\boldsymbol{s})$ is isomorphic, it is possible to transform (4.38) into the fixed point equation (FPE) as well and apply existing methods to solve the FPE problem [Polyak 1987]. Another possibility is to apply general iterative techniques to (4.34)-(4.38).

4.3.3 Generalized KP for Conjugate Mixture Models

Assume the total number of clusters N is known beforehands and the initial values for parameters

$$\boldsymbol{\theta}^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_{N+1}^{(0)}, \lambda_1^{(0)}, \ldots, \lambda_{N+1}^{(0)}, \boldsymbol{s}_1^{(0)}, \ldots, \boldsymbol{s}_N^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_N^{(0)}, \boldsymbol{\Gamma}_1^{(0)}, \ldots, \boldsymbol{\Gamma}_N^{(0)}\} \quad (4.43)$$

are chosen. The procedures *Select* to estimate N and *Initialize* to initialize parameter values $\boldsymbol{\theta}^{(0)}$ are described in detail in Chapter 6. Then the overall KP algorithm is outlined below:

1. *P step*: compute $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ using equations (4.26) to (4.30);
2. *M step*: find $\boldsymbol{\theta}^{(q+1)}$ as a fixed point of (4.34)-(4.38);
3. *Check for convergence*: go to Step 1, if convergence not achieved;

If the number of optimization iterations is taken constant, the overall complexity of the Generalized KP algorithm is $\mathcal{O}(N(M+K))$.

4.3.4 Identifiability and Algorithm Convergence

Before actually solving the optimization problem (4.21) we would like to verify that the method is capable of finding the unbiased parameter estimates, provided the number of observations is sufficiently large. As soon as the original problem is symmetric with respect to cluster permutations, we would like to split possible parameter values into classes of equivalence and introduce the following definition.

Definition 2 *The number of components N in a model $\boldsymbol{\theta}_N$ (called also the model order) is the smallest integer such that the triplets $\{(\boldsymbol{s}_n, \boldsymbol{\Sigma}_n, \sigma_k)\}_{n=1}^N$ are all different and the associated priors satisfy $\pi_n + \lambda_n > 0, \quad n = 1, \ldots, N$.*

Remark 2 *When deriving asymptotical properties, we consider the behaviour of fraction $\frac{M}{M+K}$ as $M, K \to \infty$. We'll use the word 'sequence' for $\frac{M}{M+K}$, meaning that there exists an enumeration scheme $\{M(i), K(i)\}_{i=1}^{\infty}$, such that $M(i) \to \infty$ and $K(i) \to \infty$ as $i \to \infty$. When considering asymptotical properties as $M, K \to \infty$ we would implicitly assume asymptotical properties as $i \to \infty$.*

Theorem 1 (Asymptotical Identifiability) *Assume the true number of objects N_* is known and denote $\boldsymbol{\theta}_{N_*}$ the true model. Then $\boldsymbol{\theta}_{N_*}$ belongs to the set of fixed points of the algorithm (4.21) a.s. for any choice of h_q, as $M, K \to \infty$.*

Proof: By the strong law of large numbers, the normalized log-likelihood $\frac{M}{M+K}\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta})$ has a.s. finite accumulation points of the form

$$\gamma \mathbb{E}_{\boldsymbol{\theta}_{N_*}} \log P^{\mathbb{F}}(\boldsymbol{F}, \boldsymbol{\theta}) + (1-\gamma)\mathbb{E}_{\boldsymbol{\theta}_{N_*}} \log P^{\mathbb{G}}(\boldsymbol{G}, \boldsymbol{\theta}), \tag{4.44}$$

where γ is an accumulation point of the sequence $\frac{M}{M+K}$ as $M, K \to \infty$. At the same time

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}_{N_*}) \geq H(\boldsymbol{\theta}_{N_*}, \boldsymbol{\theta}_{N_*}), \tag{4.45}$$

and the equality holds if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_{N_*}$ by Definition 2. The inequality would still hold for any accumulation point of sequence $\frac{M}{M+K}H(\boldsymbol{\theta}, \boldsymbol{\theta}_{N_*})$, as $M, K \to \infty$. This can be

seen from (4.22) taking into account that both, the log-likelihood (4.18) and the penalized likelihood (4.31) converge a.s. to finite accumulation points as $M, K \to \infty$ by the strong law of large numbers.

From here we can conclude that $\boldsymbol{\theta} = \boldsymbol{\theta}_{N_*}$ maximizes (4.44) and thus belongs to the set of stationary points of (4.21) a.s. for any choice of h_q, as $M, K \to \infty$. ■

This result shows that the generalized KP algorithm is capable of finding the true parameter values asymptotically for any choice of the sequence h_q. However, other stationary points may exist and one should avoid convergence to local maxima of the penalized log-likelihood $\mathcal{L}_{\text{pen}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$. This could be achieved through the proper choice of h_q or by standard methods involving stochastic perturbations [Spall 2003].

Another important factor is the convergence speed, which again can be improved by adjusting the sequence h_q. We've seen that in the case $h_q \equiv 0$ the algorithm would perform direct constrainted optimization of the log-likelihood (4.20) by iteratively solving pure fixed point problem for various $\boldsymbol{\theta}^{(q)}$. At the same time, Proposition 2 shows that in case of the EM algorithm $h_q \equiv 1$ the solutions of equations (4.34)-(4.37) are available in closed form, so it is only (4.38) that should be solved iteratively for each $\boldsymbol{\theta}^{(q)}$. It was pointed out by [Chrétien 2000] that in some particular cases (e.g. strictly concave log-likelihood), choosing $h_q \to 0$ may increase convergence speed from linear to quadratic. In our case, however, the results mentioned above are not applicable because of the likelihood function that does not satisfy the necessary conditions (for finite M and K there can exist several maximizers of the log-likelihood). We compare the performance of different versions of the GKP algorithm on the task of audio-visual integration in the next Section.

4.4 Experimental Evaluation

We illustrate the method in the case of audio-visual (AV) objects. Objects could be characterized both by their locations in space and by their auditory status, i.e., whether they are emitting sounds or not. These object characteristics are not directly observable and hence they need to be inferred from sensor data, e.g., cameras and microphones. These sensors are based on different physical principles, they operate with different bandwidths and sampling rates, and they provide different types of information. On one side, light waves convey useful visual information only indirectly, on the premise that they reflect onto the objects' surfaces. A natural scene is composed of many objects/surfaces and hence the task of associating visual data with objects is a difficult one. On the other side, acoustic waves convey auditory information directly from the emitter to the receiver but the observed data is perturbed by the presence of reverberations, of other sound sources, and of background noise. Moreover, very different methods are used to extract information from these two sensor types. A wide variety of computer vision principles exist for extracting 3D points from a single image or from a pair of stereoscopic cameras [Forsyth 2003] but practical methods are strongly dependent on the lighting conditions and on the properties of the objects' surfaces (presence or absence of texture, color, shape, reflectance,

etc.). Similarly, various algorithms were developed to locate sound sources using a microphone pair based on interaural time differences (ITD) and on interaural level differences (ILD) [Wang 2006, Christensen 2007], but these cues are difficult to interpret in natural settings due to the presence of background noise and of other reverberant objects. A notable improvement consists in the use a larger number of microphones [Dibiase 2001]. Nevertheless, the extraction of 3D sound source positions from several microphone observations results in inaccurate estimates. We show below that our method can be used to combine visual and auditory observations to detect and localize objects. A typical example where the conjugate mixture models framework may help is the task of locating several speaking persons.

Using the same notations as above, we consider two sensor spaces. The multimodal data consists of M visual observations $\mathbf{f}$ and of K auditory observations $\boldsymbol{g}$. We consider data that are recorded over a short time interval $[t_1, t_2]$, such that one can reasonably assume that the AV objects have a stationary spatial location. Nevertheless, it is not assumed here that the AV objects, e.g., speakers, are static: lip movements, head and hand gestures are tolerated. Generalization of multimodal clustering to multimodal tracking for dynamic scenes will be considered further.

We address the problem of estimating the spatial locations of all the objects that are both seen and heard. Let N be the number of objects and in this case each object is described by a three dimensional parameter vector $\boldsymbol{s}_n = (x_n, y_n, z_n)^\top$.

The AV data are gathered using a pair of stereoscopic cameras and a pair of omnidirectional microphones, i.e., binocular vision and binaural hearing. A visual observation vector $\boldsymbol{f}_m = (u_m, v_m, d_m)^\top$ corresponds to a 2D image location (u_m, v_m) and to an associated binocular disparity d_m. Considering a projective camera model [Faugeras 1993] it is straightforward to define an invertible function $\mathcal{F} : \mathbb{R}^3 \to \mathbb{R}^3$ that maps $\boldsymbol{s} = (x, y, z)^\top$ onto $\boldsymbol{f} = (u, v, d)^\top$:

$$\mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{1}{z}\right)^\top \quad \text{and} \quad \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{u}{d}, \frac{v}{d}, \frac{1}{d}\right)^\top. \tag{4.46}$$

This model, introduced in Chapter 2, corresponds to a rectified camera pair [Hartley 2003] and it can be easily generalized to more complex binocular geometries [Hansard 2007, Hansard 2008]. Without loss of generality one can use a sensor-centered coordinate system to represent the object locations.

Similarly one can use the auditory equivalent of disparity, namely the *interaural time difference* (ITD) widely used by auditory scene analysis methods [Wang 2006]. The function $\mathcal{G} : \mathbb{R}^3 \to \mathbb{R}$, introduced in Chapter 2, maps $\boldsymbol{s} = (x, y, z)^\top$ onto a 1D audio observation:

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big). \tag{4.47}$$

Here c is the sound speed and $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ are the 3D locations of the two microphones in the sensor-centered coordinate system. The setup is supposed to be calibrated, so that the left and right microphone positions $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ are known. To simplify the notation we would further write $\mathcal{G}(\boldsymbol{s})$ instead of $\mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})$.

	Object 1	Object 2	Object 3	Outliers
$\boldsymbol{s}_{\mathbf{WS}}$	$(300, -400, 1800)$	$(100, -300, 2050)$	$(-150, -181, 1950)$	–
$\boldsymbol{s}_{\mathbf{MS}}$	$(200, -400, 1800)$	$(100, -300, 2050)$	$(-50, -181, 1950)$	–
$\boldsymbol{s}_{\mathbf{PS}}$	$(150, -400, 1800)$	$(100, -300, 2050)$	$(0, -181, 1950)$	–
$\boldsymbol{\Sigma}$	$\begin{pmatrix} 2.7\cdot10^{-4} & 0 & 0 \\ 0 & 9\cdot10^{-4} & 0 \\ 0 & 0 & 7.5\cdot10^{-10} \end{pmatrix}$	$\begin{pmatrix} 3.3\cdot10^{-4} & 0 & 0 \\ 0 & 9.4\cdot10^{-4} & 0 \\ 0 & 0 & 8.6\cdot10^{-11} \end{pmatrix}$	$\begin{pmatrix} 1.4\cdot10^{-4} & 0 & 0 \\ 0 & 7.9\cdot10^{-4} & 0 \\ 0 & 0 & 1.5\cdot10^{-10} \end{pmatrix}$	$V = 10^{-4}$
π	0.32	0.28	0.35	0.05
σ	0.1	0.2	0.1	$U = 90$
λ	0.3	0.2	0.3	0.2

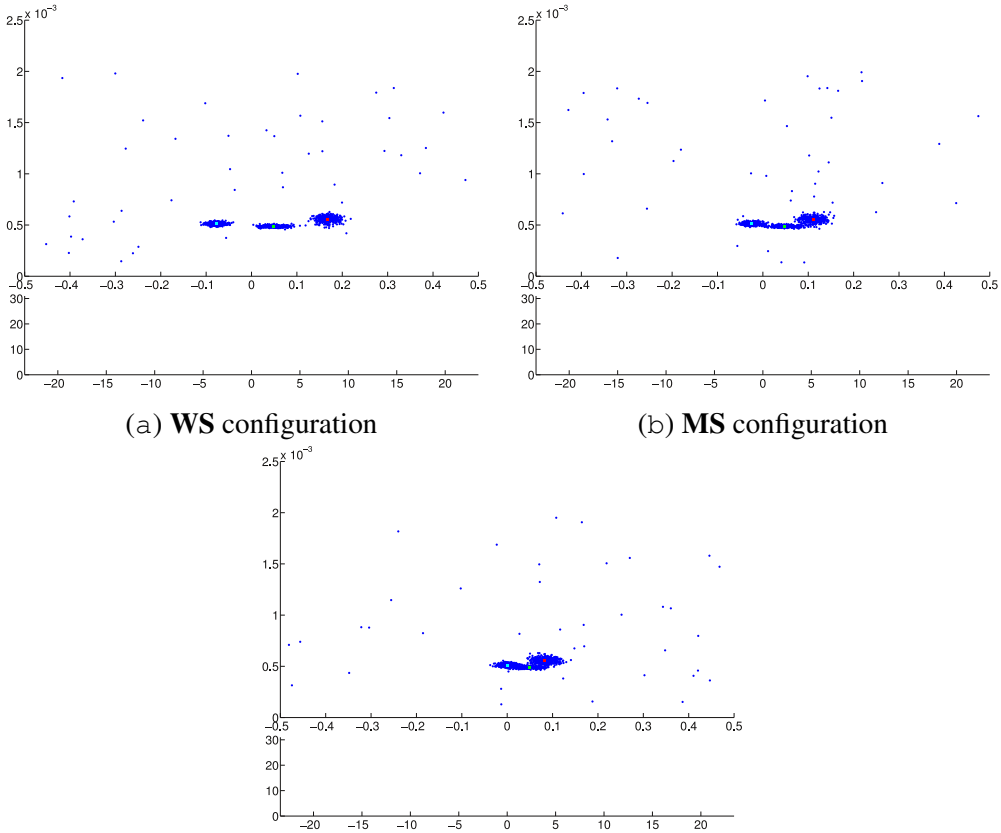
Table 4.1: Ground truth parameter values $\boldsymbol{\theta}$ for the well separated (**WS**), moderately separated (**MS**) and poorly separated (**PS**) object configurations. Cluster means in the observation spaces are calculated through applying mappings $\mathcal{F}$ and $\mathcal{G}$ to locations $\boldsymbol{s}_n$, $n = 1, 2, 3$. Observations are then sampled from mixture models of Gaussians with an outlier (uniform) component in auditory and visual spaces.

The performance of the conjugate KP algorithm is verified on the simulated audio-visual localization task. We generate the data using the conjugate mixture models that were introduced in Section 4.2. Three objects that are both seen and heard are supposed to be present in the scene. The tying parameters in this case are object locations in the 3D ambient space $\mathbf{s}_n \in \mathbb{S}$, $n = 1, 2, 3$. The rest of parameters characterize object images in the observation spaces. Their values that were used to generate the auditory and visual data are summarized in Table 4.1. The parameters were taken so as to imitate observations obtained in real-world scenarios.

We suppose three objects that are present in the scene are defined in $\mathbb{S}$ by $\mathbf{s}_n$, $n = 1, 2, 3$. Different configurations were considered: well separated (**WS**), moderately separated (**MS**) and poorly separated (**PS**) cases. The modality-associated parameters $\boldsymbol{\Sigma}_n$, σ_k, π_n and λ_n for $n = 1, 2, 3$ that account for object and detector properties are kept the same across the configurations, whereas the object ambient space positions $\mathbf{s}_n$ vary.

The data was sampled in the visual and auditory observation spaces using the mapped mean values $\mathcal{F}(\boldsymbol{s}_n)$ and $\mathcal{G}(\boldsymbol{s}_n)$, covariance matrices $\boldsymbol{\Sigma}_n$ and variances σ_n and priors π_n and λ_n respectively. Microphone locations were taken to be $\boldsymbol{s}_{\mathrm{M}_\ell} = (-85.9, -80.3, 20.4)^\top$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}} = (85.8, -80, -15)^\top$.

In total $M = 1000$ and $K = 100$ samples were drawn from the corresponding mixture models. The simulated data for the three configurations is shown in Figure 4.3. In each case a scatter plot of visual observation projections on the (u, d) coordinates is shown in the upper part of a plot. The corresponding means and covariance matrices are depicted with points and ellipsoids. Auditory observations are shown as an ITD domain histogram in the lower part of each plot. Auditory means and variances are depicted with coloured bars, their height designates prior probabilities. The same colour is used for each object in both domains. Dashed black lines show the boundaries of the field of view mapped to the visual and auditory spaces.



(a) **WS** configuration

(b) **MS** configuration

(c) **PS** configuration

Figure 4.3: Simulated data for the (a) well separated, **WS**, (b) moderately separated, **MS** and (c) poorly separated, **PS** object configurations. Scatter plot of visual observation projections on the (u, d) coordinates is shown in the upper part of each plot. The corresponding means and covariance matrices are depicted with points and ellipsoids. Auditory observations are shown as an ITD domain histogram in the lower part of each plot. Auditory means and variances are depicted with coloured bars, their height designates prior probabilities. The same colour is used for each object in both domains. Each one of the two mixtures models (associated with each sensorial modality) contains four components: three objects and one outlier class. Dashed black lines show the boundaries of the field of view in visual and auditory spaces.

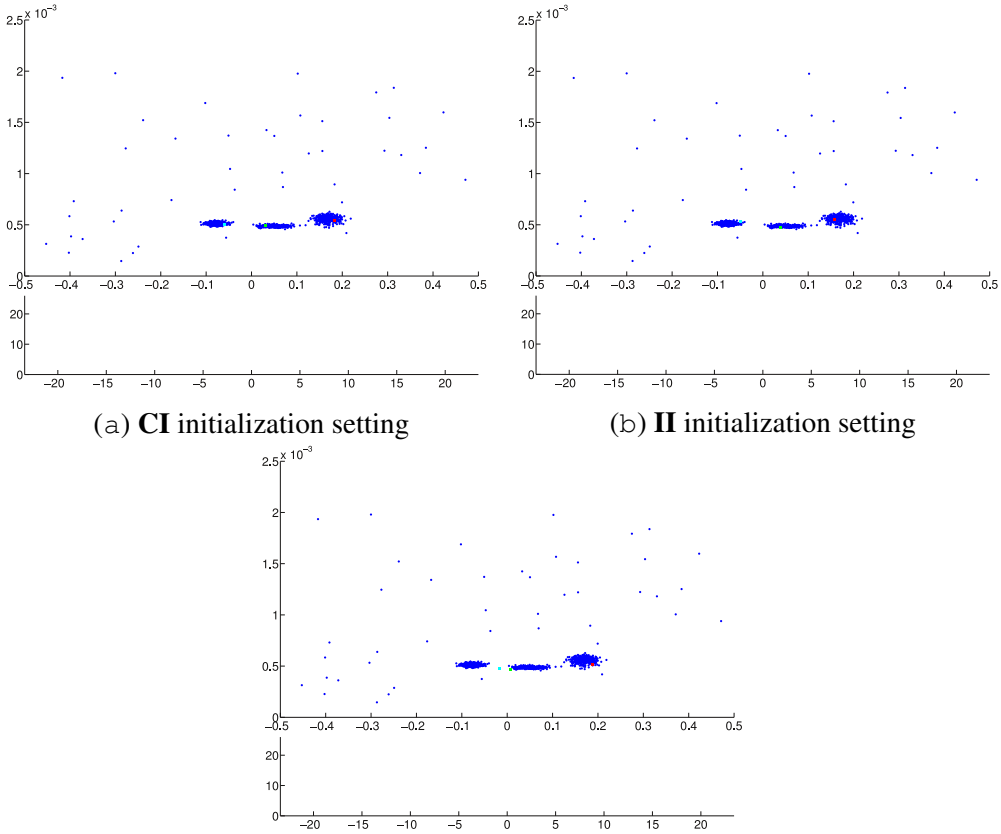
We compare the performance of different versions of the KP algorithm with gain sequence $h_q \equiv 1$ (EM algorithm, **EM**), $h_q = 1/q$ (relaxed EM algorithm, **REM**) and $h_q \equiv 0$ (direct log-likelihood optimization, **DO**). Every run of the KP algorithm comprised 30 iterations. To carry out the maximization step, we performed 5 fixed point iterations, recalculating $\alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})$ every time through (4.33) and updating the parameters by (4.34)-(4.38). The optimization (4.38) with respect to tying parameters $\boldsymbol{s}_n$ was performed using 1000 iterations of the simultaneous perturbation stochastic approximation (SPSA) algorithm [Spall 2003]. The performance of the fixed point iteration is defined majorly by theoretical properties of the system itself (contracting property, conditions on the Jacobian matrix, we refer to [Polyak 1987] for more details). But it is also largely dependent on the quality of its optimization subtask solutions. Thus we needed to perform more optimization iterations for (4.38) to ensure the fixed point iteration uses an appropriate value.

To investigate the dependency of the results on initialization, the parameters values $\boldsymbol{\theta}^{(0)}$ were sampled from the ground truth values using close (**CI**), intermediate (**II**) and far (**FI**) initialization settings. The x, y and z coordinates of object locations $\boldsymbol{s}_n^{(0)}$ were drawn from Gaussian distributions centered in the corresponding ground truth values with variances $\varrho_{\mathrm{x}}^2 = 10^3, \varrho_{\mathrm{y}}^2 = 10^3, \varrho_{\mathrm{z}}^2 = 10^3$ (**CI**), $\varrho_{\mathrm{x}}^2 = 10^3, \varrho_{\mathrm{y}}^2 = 10^3, \varrho_{\mathrm{z}}^2 = 10^4$ (**II**) and $\varrho_{\mathrm{x}}^2 = 10^4, \varrho_{\mathrm{y}}^2 = 10^4, \varrho_{\mathrm{z}}^2 = 10^4$ (**FI**) respectively. Eigenvalues of visual space covariance matrices $\boldsymbol{\Sigma}_n^{(0)}$ and auditory space variances $\sigma_n^{(0)}$ were sampled using Rice distribution $\mathrm{Rice}(\nu, \kappa)$ to ensure their positiveness. This distribution always has a mode (which is not always the case for the Gamma distribution) and is fairly easy to simulate. Its density function is given by

$$p(x \mid \nu, \kappa) = \frac{x}{\kappa^2} \exp\left(\frac{-(x^2+\nu^2)}{2\kappa^2}\right) I_0\left(\frac{x\nu}{\kappa^2}\right), \tag{4.48}$$

where $I_0(x)$ is the modified Bessel function of the first kind with order zero. The ν parameter in each case was taken to be the corresponding ground truth value, the κ parameter was chosen for different settings to be $\kappa_u = 10^{-3}, \kappa_v = 10^{-3}, \kappa_d = 10^{-8}, \kappa_g = 1$ (**CI**), $\kappa_u = 5 \cdot 10^{-3}, \kappa_v = 5 \cdot 10^{-3}, \kappa_u = 5 \cdot 10^{-8}, \kappa_g = 2.5$ (**II**) and $\kappa_u = 10^{-2}, \kappa_v = 10^{-2}, \kappa_d = 10^{-7}, \kappa_g = 5$ (**FI**). The initial priors in both modalities are always taken to be equal. Examples of the obtained parameter values $\boldsymbol{\theta}^{(0)}$ for the **WS** configuration are shown in Figure 4.4.

The detailed results of the **REM** version of the KP algorithm with $h_q = 1/q$ for the **WS**, **MS** and **PS** configurations in the **CI** initialization setting are presented in Table 4.3. Estimated locations $\hat{\boldsymbol{s}}_{\mathrm{WS}}$ and their images in visual $\hat{\boldsymbol{f}}_{\mathrm{WS}}$ and auditory $\hat{g}_{\mathrm{WS}}$ spaces are compared to the ground truth values $\boldsymbol{s}_{\mathrm{WS}}$, $\boldsymbol{f}_{\mathrm{WS}}$ and g_{WS} respectively. Absolute and relative errors $\varepsilon_{\mathrm{abs}} = \|\hat{\boldsymbol{s}}_{\mathrm{WS}} - \boldsymbol{s}_{\mathrm{WS}}\|$ and $\varepsilon_{\mathrm{rel}} = \|\hat{\boldsymbol{s}}_{\mathrm{WS}} - \boldsymbol{s}_{\mathrm{WS}}\| / \|\boldsymbol{s}_{\mathrm{WS}}\|$ are calculated for object locations $\hat{\boldsymbol{s}}_{\mathrm{WS}}$ and their observation space images $\hat{\boldsymbol{f}}_{\mathrm{WS}}$ and $\hat{g}_{\mathrm{WS}}$ using similar formulas. The results are averaged over 10 runs of the algorithm with random initializations. The estimated errors show that when given good initial values, all versions of the KP algorithm converge to a solution that is very close to the ground truth. Matching in the observations spaces is perfect. In our simulations the distances in ambient space are measured in millimetres, so 3-4mm precision is typically achieved.



(a) **CI** initialization setting

(b) **II** initialization setting

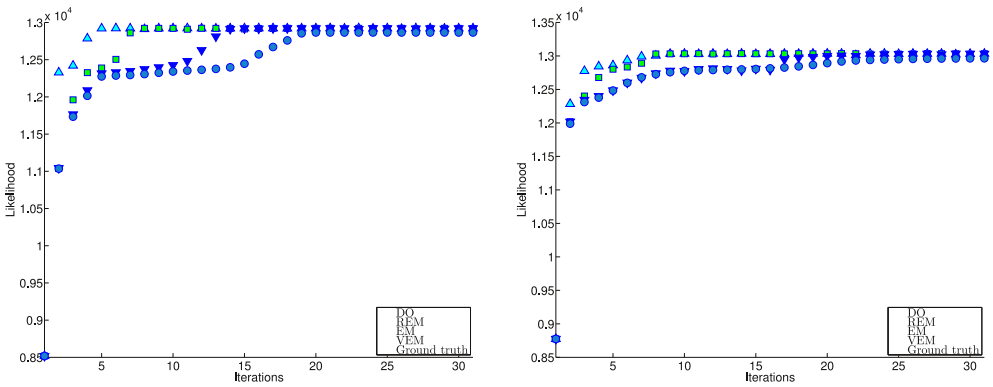
(c) **FI** initialization setting

Figure 4.4: Sampled initializations $\boldsymbol{\theta}^{(0)}$ for the well-separated **WS** configuration: (a) close initialization, **CI**, (b) intermediate initialization, **II** and (c) far initialization, **FI**. The visual means and covariance matrices are depicted with points and ellipsoids in (u, d) coordinates in the upper part of each plot. The corresponding auditory means and variances are shown with the same colour in the lower part of each plot. Dashed black lines show the boundaries of the field of view in visual and auditory spaces.

We compared different versions of the multimodal KP algorithm (**EM**, **REM** and **DO**) to a unimodal EM algorithm based on visual observations only (**VEM**). Summary of average absolute error values ε_{abs} for object location estimates $\hat{\boldsymbol{s}}$ for **WS**, **MS** and **PS** object configurations and **IC**, **II** and **IF** initial values settings is given in Table 4.2. All the algorithms show acceptable performance (3-5mm precision) on various configurations. Thus the primary criteria on the algorithm choice would be (i) convergence speed; (ii) computation speed; (iii) algorithm stability. To compare the convergence speeds we plotted likelihood evolution graphs shown in Figure 4.5. **DO** and **REM** are the most efficient on the initial stage of the optimization (bootstrap period). They admit certain fluctuations in the likelihood values on the final stage (see the close-up in Figure 4.6), which is the consequence of complexity of the optimization task (4.21). These fluctuations are acceptable

	IC				II				IF			
	EM	REM	DO	VEM	EM	REM	DO	VEM	EM	REM	DO	VEM
WS	4.62	3.92	4.48	3.92	4.33	4.16	4.05	3.92	4.5	3.96	4.38	3.92
MS	3.22	3.63	3.3	4.66	2.81	3.56	3.33	4.66	3.58	3.08	3.57	4.66
PS	2.93	2.94	3.08	3.72	3.13	3.38	3.44	3.73	3.1	3.4	3.26	3.74

Table 4.2: Summary of average absolute error ε_{abs} values for object location estimates $\hat{s}$ for **WS**, **MS** and **PS** object configurations. Dependency on initial values setting (**IC**, **II** or **IF**) and the optimization algorithm version (**EM**, **REM**, **DO**, **VEM**) is shown.



(a) **WS** configuration (b) **PS** configuration

Figure 4.5: Likelihood evolution graphs for (a) **WS** object configuration and (b) **PS** object configuration for **FI** initialization setting. The faster the sequence h_q decreases, the more efficient the bootstrap period of the KP algorithm is and the less stable the behaviour of the estimate becomes afterwards. Dashed black lines show the likelihood of the ground truth parameters.

in the case of well separated objects, though in the case of more complex object configurations they can become crucial for performance. At the same time the **VEM** strategy does well in the case when visual information is rich and not too corrupted, but heavily relies on the data quality. Computation speed depends directly on the optimization method complexity, which favours the **VEM** and **EM** algorithms.

The conjugate **EM** algorithm has an advantage over the EM algorithms operating on single modalities (like **VEM** algorithm) as soon as it can perform optimization in cases when data from an object in one modality is almost absent (see Figure 4.6). At the same time, single modality EM algorithms are a particular cases of **ConjEM** with zero auditory weights α_{mn} or β_{kn}, so we would further concentrate on **ConjEM** and try to improve its convergence speed to that of **DO** and **REM** and reduce the complexity.

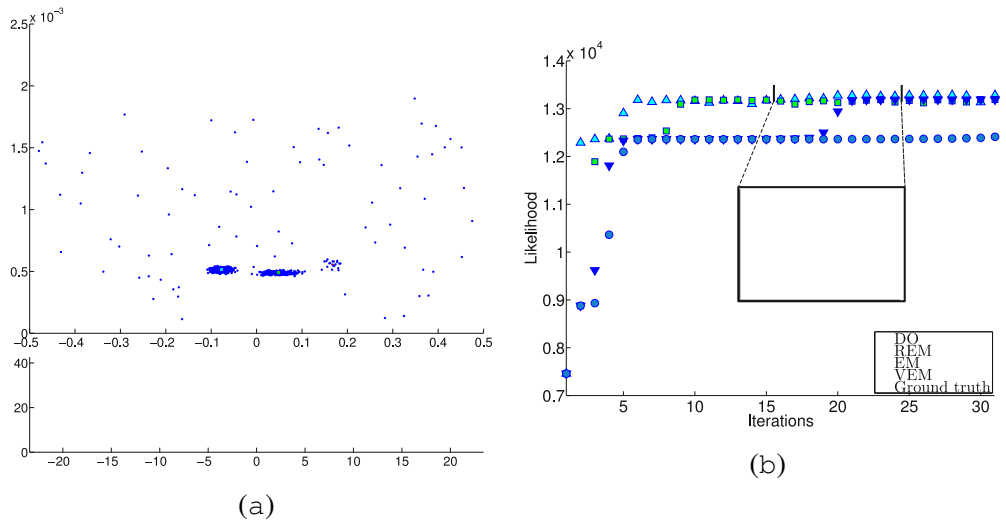


Figure 4.6: Missing data case showing the best the advantages of the ConjEM algorithm over single modality EM algorithms and other algorithms from the KP family. (a) ConjEM converges properly for configurations where certain objects are absent in one modality, whereas single modality EM algorithms fail in this case. Data in auditory and visual domains contains two strong and one weak cluster. Ground truth is depicted with red, green and cyan colours in both domains, the corresponding estimated clusters have darker colours (they are quite accurate and overlap with the ground truth in the image). ConjEM extends the EM algorithms that operate on a single modality, and can be reduced to them in special cases. (b) ConjEM shows stable behaviour always increasing the likelihood function, which is not always the case for other algorithms from the KP family.

4.5 Discussion

We proposed a novel framework to cluster heterogeneous data gathered with physically different sensors. Our approach differs from other existing approaches in that it combines in a single statistical model a number of clustering tasks while ensuring the consistency of their results. In addition, the fact that the clustering is performed in observation spaces allows one to get useful statistics on the data (i.e. variances and covariance matrices, priors), which is an advantage of our approach over particle filtering models. The task of simultaneous clustering in spaces of different nature, related through known functional dependencies to a common parameter space, was formulated as a likelihood maximization problem. We built the conjugate KP algorithm to perform the multimodal clustering task using the standard KP theory.

One of the strong points of the formulated model is that it is open to different useful extensions. It can be easily extended to an arbitrary number J of observation spaces $\mathbb{F}_1, \ldots, \mathbb{F}_J$. The sum of two terms, related to spaces $\mathbb{F}$ and $\mathbb{G}$, would have to be replaced by a sum of J terms corresponding to $\mathbb{F}_1, \ldots, \mathbb{F}_J$ in the formulas of Sections 4.2 and 4.3. Additional features can be added to the unimodal mixture models. This would increase

the dimensionality of $\mathbb{F}$ and $\mathbb{G}$ spaces, but the KP algorithm would stay unchanged. Also, the assumption that assignment variables $\boldsymbol{a}$ and $\boldsymbol{b}$ are independent could be relaxed. An appropriate approach to perform inference in a non independent case would be to consider variational approximations [Jordan 1998] and in particular a variational EM framework. The general idea would be to approximate the joint distribution $P(\boldsymbol{a})$ by a distribution from a restricted class of probability distributions that factorize as $\tilde{P}(\boldsymbol{a}) = \prod_{m=1}^{M} \tilde{P}(a_m)$. For any such distribution, our model would be applicable without any changes, so that for a variational version of the conjugate EM algorithm, all the results from this Chapter would hold. Thus one can consider generalizations, such as conjugate random fields and conjugate point processes. Finally, each of the Gaussian mixtures (4.16) and (4.17) can be replaced by any other mixture. Some distribution choices would not require any significant changes. We consider Student t-distribution mixtures as an example in Appendix A.2.

A non trivial audio-visual localization task was considered to illustrate the conjugate KP performance on simulated data. These experiments allowed us to assess the average method behaviour for different algorithm options, various object configurations and initialization properties. They showed that the obtained clustering results were precise in determining object locations in the hidden ambient space and in the observation spaces. Certain peculiarities regarding bootstrap time and solution stability were revealed. It occured that though KP algorithms with fastly decreasing gain sequence show better convergence on the initial stage, they demonstrate worse stability on the final stage, which leads to precision loss. The conjugate EM (ConjEM) version of the algorithm takes more time to converge, but is more stable. At the same time, the simple EM algorithm based on visual data only (VEM), which is the particular case of ConjEM, showed good results on simple object configurations. Thus the best would be to find the way to accelerate ConjEM to benefit from both, speed and accuracy.

To summarize we outline the major advantages of the proposed framework:

- **Information integration**: the use of tying parameters guarantees coherent results in both modality spaces;
- **Efficient optimization**: simultaneous inference of assignment labels and object parameters allows to avoid exponentially hard binding problems;
- **Identifiability**: the method is proved to be capable of finding the optimal parameters asymptotically;
- **Automatic modality weighting**: data statistics are estimated, so that more precise and richer data is automatically assigned greater weight;
- **Integration reinforcement**: the algorithm does not rely on the quality of a particular modality and efficiently combines the data, showing good performance on various object configurations;

- **Extensibility**: the model can be extended onto any number of modalities, various modality-specific properties can be included, different statistical models can be incorporated;

It appears that as a generalization of Gaussian mixture models, our approach has larger modelling capabilities. It is entirely based on a mathematical framework in which each step is theoretically well-founded. Its ability to provide good results in a non trivial multimodal clustering task is particularly promising for applications requiring the integration of several heterogenious information sources. Therefore, it has advantages over other methods that include ad-hoc processing while being open to incorporation of more task dependent information.

The proposed framework aligns well with the findings from neurobiology in what concerns multisensory integration. Adopting the assumption that the objects are co-localized and co-incident the model reinforces the integration process the way the multisensory enhancement phenomenon does. The case of weaker modality signals (less observations or more complex object configurations) shows greater improvement of multimodal algorithm with respect to the unimodal one. At the same time the model is quite flexible and can automatically weight the modalities, which is an important feature for multimodal applications. The single-modality algorithms are the particular cases of our approach, so strong unimodal signals are treated in the unimodal way and are almost not enhanced. The multimodal binding happens naturally in our model and the consistency is verified across the modalities.

			Object 1	Object 2	Object 3
Well Separated, **WS**	Location	$\boldsymbol{s}_{\mathbf{WS}}$	$(300, -400, 1800)$	$(100, -300, 2050)$	$(-150, -181, 1950)$
		$\hat{\boldsymbol{s}}_{\mathbf{WS}}$	$(300.8, -397.25, 1794.46)$	$(99.25, -297.47, 2053.58)$	$(-149.66, -182.03, 1950.07)$
		ε_{abs}	6.23	4.45	1.09
		ε_{rel}	0.0033	0.0021	0.00056
	Visual	$\boldsymbol{f}_{\mathbf{WS}}$	$(0.16, -0.22, 0.00056)$	$(0.49, -0.15, 0.00049)$	$(-0.077, -0.09, 0.00051)$
		$\hat{\boldsymbol{f}}_{\mathbf{WS}}$	$(0.17, -0.22, 0.00056)$	$(0.048, -0.14, 0.00049)$	$(-0.077, -0.09, 0.00051)$
		ε_{abs}	0.0013	0.0016	0.0006
		ε_{rel}	0.0046	0.01	0.0046
	Auditory	$g_{\mathbf{WS}}$	-0.88	-3.59	-6.47
		$\hat{g}_{\mathbf{WS}}$	-0.86	-3.6	-6.46
		ε_{abs}	0.021	0.01	0.004
		ε_{rel}	0.024	0.003	0.0006
Moderately Separated, **MS**	Location	$\boldsymbol{s}_{\mathbf{MS}}$	$(200, -400, 1800)$	$(100, -300, 2050)$	$(-50, -181, 1950)$
		$\hat{\boldsymbol{s}}_{\mathbf{MS}}$	$(201.05, -402.79, 1802.13)$	$(96.73, -299.34, 2049.69)$	$(-47.11, -181.83, 1947.55)$
		ε_{abs}	3.66	3.35	3.88
		ε_{rel}	0.002	0.0016	0.002
	Visual	$\boldsymbol{f}_{\mathbf{MS}}$	$(0.11, -0.22, 0.00056)$	$(0.049, -0.15, 0.00049)$	$(-0.026, -0.09, 0.00051)$
		$\hat{\boldsymbol{f}}_{\mathbf{MS}}$	$(0.11, -0.22, 0.00056)$	$(0.047, -0.15, 0.00049)$	$(-0.024, -0.09, 0.00051)$
		ε_{abs}	0.0014	0.0016	0.0016
		ε_{rel}	0.0055	0.0105	0.0161
	Auditory	$g_{\mathbf{MS}}$	-2.13	-3.59	-5.31
		$\hat{g}_{\mathbf{MS}}$	-2.12	-3.62	-5.27
		ε_{abs}	0.011	0.037	0.033
		ε_{rel}	0.005	0.01	0.006
Poorly Separated, **PS**	Location	$\boldsymbol{s}_{\mathbf{MS}}$	$(150, -400, 1800)$	$(100, -300, 2050)$	$(0, -181, 1950)$
		$\hat{\boldsymbol{s}}_{\mathbf{MS}}$	$(150.57, -398.07, 1799.01)$	$(100.93, -301.48, 2047.57)$	$(1.65, -182.76, 1952.68)$
		ε_{abs}	2.24	2.99	3.6
		ε_{rel}	0.0012	0.0014	0.0018
	Visual	$\boldsymbol{f}_{\mathbf{MS}}$	$(0.08, -0.22, 0.00056)$	$(0.05, -0.15, 0.00049)$	$(0, -0.09, 0.00051)$
		$\hat{\boldsymbol{f}}_{\mathbf{MS}}$	$(0.08, -0.22, 0.00056)$	$(0.05, -0.15, 0.00049)$	$(0.0008, -0.09, 0.00051)$
		ε_{abs}	0.001	0.001	0.0011
		ε_{rel}	0.0043	0.0067	0.0123
	Auditory	$g_{\mathbf{MS}}$	-2.77	-3.59	-4.72
		$\hat{g}_{\mathbf{MS}}$	-2.76	-3.6	-4.7
		ε_{abs}	0.008	0.012	0.019
		ε_{rel}	0.003	0.003	0.004

Table 4.3: Results of the **REM** version ($h_q = 1/q$) of the KP algorithm for the well separated (**WS**), moderately separated (**MS**) and poorly separated (**PS**) object configurations in the close initialization (**CI**) setting. They resemble the results of other versions of the KP algorithm, REM was chosen as an 'average' representative. Estimated locations $\hat{\boldsymbol{s}}_{\mathbf{WS}}$ and their images in visual $\hat{\boldsymbol{f}}_{\mathbf{WS}}$ and auditory $\hat{g}_{\mathbf{WS}}$ spaces are compared to the ground truth values $\boldsymbol{s}_{\mathbf{WS}}$, $\boldsymbol{f}_{\mathbf{WS}}$ and $g_{\mathbf{WS}}$ respectively. Absolute ε_{abs} and relative ε_{rel} errors are calculated for object locations and their observation space images.

CHAPTER 5

Conjugate EM Algorithm for Clustering Multimodal Data

Sommaire

5.1 Conjugate EM Algorithm for Clustering Multimodal Data 65
5.1.1 The Expectation Step 66
5.1.2 The Maximization Step 67
5.1.3 Generalized EM for Conjugate Mixture Models 68
5.1.4 Analysis of *Local Search* Procedure 69
5.1.5 Global Search and the *Choose* Procedure 73
5.2 Clustering Using Auditory and Visual Data 74
5.3 Experimental Validation 76
5.3.1 Experiments with Simulated Data 76
5.3.2 Experiments with Real Data 81
5.4 Discussion 89

The general approach to multimodal clustering based on Kullback-Proximal (KP) framework considered in the previous Chapter cannot be significantly improved. As we've seen, in practice it is quite difficult to ensure both, the stability and the algorithm efficiency. In this Chapter we concentrate on one particular instance from the KP family, namely the conjugate Expectation-Maximization (ConjEM) algorithm. It is shown to guarantee the increase of target function for a large class of observation space mappings and a number of possibilities are proposed to accelerate the convergence. We demonstrate the performance of the ConjEM algorithm on the task of audio-visual localization considering both, simulated and real data.

5.1 Conjugate EM Algorithm for Clustering Multimodal Data

Performing direct optimization of the observed data log-likelihood function (4.18) and applying general penalized optimization techniques presents certain difficulties. The optimization methods do not guarantee permanent increase of the target function which leads to undesirable fluctuations. At the same time, one instance of the KP family showed more regular behaviour, though slower convergence speed.

The expectation-maximization (EM) algorithm [Dempster 1977, McLachlan 2007] is a standard approach to maximize likelihood functions of type (4.18). It is a particular case of the considered previously KP algorithm with $h_q \equiv 1$, as stated in Proposition 2. Each iteration q of EM proceeds in two steps:

- *Expectation*. For the current values $\boldsymbol{\theta}^{(q)}$ of the parameters, compute the conditional expectation with respect to variables $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \sum_{\boldsymbol{a} \in \{1 \dots N+1\}^M} \sum_{\boldsymbol{b} \in \{1 \dots N+1\}^K} P(\boldsymbol{a}, \boldsymbol{b} | \mathbf{f}, \mathbf{g};\ \boldsymbol{\theta}^{(q)})\ \log P(\mathbf{f}, \mathbf{g}, \boldsymbol{a}, \boldsymbol{b};\ \boldsymbol{\theta}) \tag{5.1}$$

- *Maximization*. Update the parameter set $\boldsymbol{\theta}^{(q)}$ by maximizing (5.1) with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(q+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) \tag{5.2}$$

As soon as the EM algorithm is an instance of the KP algorithm family, the increasing property (4.24) stated in Proposition 1 remains valid for the cases when the M-step 5.2 has a closed form solution. This is the case for standard EM that deals with the parameter estimation of a single mixture model. For the case when the maximization (5.2) is difficult to achieve, various generalizations of EM are proposed. As in the case of KP algorithm, the M step can be relaxed by requiring just an increase rather than an optimum. This yields Generalized EM (GEM) procedures [McLachlan 2007] (see [Boyles 1983] for a result on the convergence of this class of algorithms). GEM occurs to be more stable than the GKP algorithm in the sense that increases in the target function are easier to achieve. Below we describe in more detail the specific forms of the E and M steps for the case of conjugate mixture models.

5.1.1 The Expectation Step

Using the independency assumptions (4.2)-(4.3), the conditional expectation (5.1) can be decomposed as:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = Q_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) + Q_{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}), \tag{5.3}$$

with

$$Q_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \sum_{m=1}^{M} \sum_{n=1}^{N+1} \alpha_{mn}^{(q)} \log \left(\pi_n P(\boldsymbol{f}_m | A_m = n;\ \boldsymbol{\theta})\right), \tag{5.4}$$

$$Q_{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \sum_{k=1}^{K} \sum_{n=1}^{N+1} \beta_{kn}^{(q)} \log \left(\lambda_n P(\boldsymbol{g}_k | B_k = n;\ \boldsymbol{\theta})\right), \tag{5.5}$$

where $\alpha_{mn}^{(q)}$ and $\beta_{kn}^{(q)}$ denote the posterior probabilities $\alpha_{mn}^{(q)} = P(A_m = n | \boldsymbol{f}_m; \boldsymbol{\theta}^{(q)})$ and $\beta_{kn}^{(q)} = P(B_k = n | \boldsymbol{g}_k; \boldsymbol{\theta}^{(q)})$ that can be easily computed by equations (4.29) and (4.30).

Using likelihood expressions (4.11)-(4.15) further leads to:

$$Q_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{mn}^{(q)}\left(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n} + \log((2\pi)^r|\boldsymbol{\Sigma}_n|\pi_n^{-2})\right) - \\ -\frac{1}{2}\sum_{m=1}^{M}\alpha_{m,N+1}^{(q)}\log(V^2\pi_{N+1}^{-2}), \tag{5.6}$$

$$Q_{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\beta_{kn}^{(q)}\left(\|\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Gamma}_n} + \log((2\pi)^p|\boldsymbol{\Gamma}_n|\lambda_n^{-2})\right) - \\ -\frac{1}{2}\sum_{k=1}^{K}\beta_{k,N+1}^{(q)}\log(U^2\lambda_{N+1}^{-2}). \tag{5.7}$$

5.1.2 The Maximization Step

In order to carry out the maximization (5.2) of the conditional expectation (5.1), its derivatives with respect to the model parameters are set to zero. This leads to the standard update expressions for priors, more specifically $\forall n = 1, \ldots, N+1$:

$$\pi_n^{(q+1)} = \frac{1}{M}\sum_{m=1}^{M}\alpha_{mn}^{(q)}, \tag{5.8}$$

$$\lambda_n^{(q+1)} = \frac{1}{K}\sum_{k=1}^{K}\beta_{kn}^{(q)}. \tag{5.9}$$

The covariance matrices are governed by the tying parameters $\boldsymbol{s}_n^{(q+1)} \in \mathbb{S}$ through the functions $\mathcal{F}$ and $\mathcal{G}$, $\forall n = 1, \ldots, N$:

$$\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s}_n^{(q+1)}) = \frac{1}{\sum_{m=1}^{M}\alpha_{mn}^{(q)}}\sum_{m=1}^{M}\alpha_{mn}^{(q)}(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q+1)}))(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q+1)}))^\top, \tag{5.10}$$

$$\boldsymbol{\Gamma}_n^{(q+1)}(\boldsymbol{s}_n^{(q+1)}) = \frac{1}{\sum_{k=1}^{K}\beta_{kn}^{(q)}}\sum_{k=1}^{K}\beta_{kn}^{(q)}(\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n^{(q+1)}))(\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n^{(q+1)}))^\top. \tag{5.11}$$

For every $n = 1, \ldots, N$, the parameter vector $\boldsymbol{s}_n^{(q+1)}$ is computed such that:

$$\boldsymbol{s}_n^{(q+1)} = \underset{\boldsymbol{s}}{\operatorname{argmax}}\, Q_n^{(q)}(\boldsymbol{s}), \tag{5.12}$$

where

$$Q_n^{(q)}(\boldsymbol{s}) = -\sum_{m=1}^{M}\alpha_{mn}^{(q)}(\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s})\|^2_{\boldsymbol{\Sigma}_n(\boldsymbol{s})} + \log|\boldsymbol{\Sigma}_n(\boldsymbol{s})|) - \\ -\sum_{k=1}^{K}\beta_{kn}^{(q)}(\|\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s})\|^2_{\boldsymbol{\Gamma}_n(\boldsymbol{s})} + \log|\boldsymbol{\Gamma}_n(\boldsymbol{s})|). \tag{5.13}$$

We stress that the covariances $\boldsymbol{\Sigma}_n(\boldsymbol{s})$ and $\boldsymbol{\Gamma}_n(\boldsymbol{s})$ in (5.10) and (5.11) are considered as functions of $\boldsymbol{s} \in \mathbb{S}$. Hence, at each iteration of the algorithm, the overall update of the tying parameters can be split into N identical optimization tasks of the form (5.13). These tasks can be solved in parallel. In general, $\mathcal{F}$ and $\mathcal{G}$ are non-linear transformations and hence there is no simple closed-form expression for the estimation of the tying parameters.

5.1.3 Generalized EM for Conjugate Mixture Models

We assume the initial parameters $\boldsymbol{\theta}^{(0)}$ to be selected for the conjugate EM (ConjEM) algorithm. An efficient procedure *Initialize* would be proposed in Chapter 6 to choose $\boldsymbol{\theta}^{(0)}$. The maximization step uses two procedures, referred to as *Choose* and *Local Search* which are explained in detail in Sections 5.1.4 and 5.1.5 respectively. To determine the number of objects N we define the procedure *Select* that is derived in Chapter 6. The overall EM procedure is outlined below:

1. Apply procedure *Initialize* to initialize the parameter vector:
   $\boldsymbol{\theta}^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_{N+1}^{(0)}, \lambda_1^{(0)}, \ldots, \lambda_{N+1}^{(0)}, \boldsymbol{s}_1^{(0)}, \ldots, \boldsymbol{s}_N^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_N^{(0)}, \boldsymbol{\Gamma}_1^{(0)}, \ldots, \boldsymbol{\Gamma}_N^{(0)}\}$;
2. *E step*: compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ using equations (4.29), (4.30), (5.6) and (5.7);
3. *M step*: estimate $\boldsymbol{\theta}^{(q+1)}$ using the following sub-steps:
   (a) *The priors.* Compute $\pi_1^{(q+1)}, \ldots, \pi_{N+1}^{(q+1)}$ and $\lambda_1^{(q+1)}, \ldots, \lambda_{N+1}^{(q+1)}$ using (5.8) and (5.9);
   (b) *The tying parameters.* For each $n = 1 \ldots N$:
   - Apply procedure *Choose* to determine an initial value, denoted by $\tilde{\boldsymbol{s}}_n^{(0)}$, as proposed in Section 5.1.5;
   - Apply procedure *Local Search* to each $Q_n^{(q)}(\boldsymbol{s})$ as defined in (5.13) starting from $\tilde{\boldsymbol{s}}_n^{(0)}$ and set the result to $\boldsymbol{s}_n^{(q+1)}$ using the equation (5.14) specified below;

   (c) *The covariance matrices.* For every $n = 1 \ldots N$, use (5.10) and (5.11) to compute $\boldsymbol{\Sigma}_n^{(q+1)}$ and $\boldsymbol{\Gamma}_n^{(q+1)}$;
4. *Check for convergence*: Terminate, otherwise go to Step 2;
5. Apply procedure *Select*, use the criterion from Chapter 6 specified below to determine the best N;

If the number of optimization iterations on Step 3b is taken constant, the overall complexity of the Generalized EM algorithm is $\mathcal{O}(N(M + K))$. This algorithm uses the following procedures:

- *Initialize*: this procedure aims at providing the initial parameter values $\boldsymbol{\theta}^{(0)}$. Its performance has a strong impact on the time required for the algorithm to converge. In Chapter 6 we propose an efficient initialization strategy based on single-space probability density estimation and cluster detection.

- *Select*: this procedure applies the information criterion for conjugate mixture models to determine the number of objects N. In Chapter 6 we show that the proposed criterion is consistent in the case of conjugate mixture models.

- *Choose*: the goal of this procedure is to provide at each M step initial values $\tilde{\boldsymbol{s}}_1^{(0)}, \ldots, \tilde{\boldsymbol{s}}_N^{(0)}$ which are likely to be close to the global maxima of the functions $Q_n^{(q)}(\boldsymbol{s})$ in (5.13). The exact form of this procedure is important to ensure the ability of the subsequent *Local Search* procedure to find these global maxima. We will use results on global search algorithms [Zhigljavsky 2008] and propose different variants in Section 5.1.5.

- *Local Search*: an important requirement of this procedure is that it finds a local maximum of the $Q_n^{(q)}(\boldsymbol{s})$'s starting from any arbitrary point in $\mathbb{S}$. We will consider procedures that consist in iterating a local update of the form (ν is the iteration index):

$$\tilde{\boldsymbol{s}_n}^{(\nu+1)} = \tilde{\boldsymbol{s}_n}^{(\nu)} + \mathbf{H}_n^{(q,\nu)} \boldsymbol{\nabla} Q_n^{(q)}(\tilde{\boldsymbol{s}_n}^{(\nu)}), \tag{5.14}$$

  with $\mathbf{H}_n^{(q,\nu)}$ being a positive definite matrix that may vary with ν. When the gradient $\boldsymbol{\nabla} Q_n^{(q)}(\boldsymbol{s})$ is Lipschitz continuous with some constant $L_n^{(q)}$, an appropriate choice that guarantees the increase of $Q_n^{(q)}(\tilde{\boldsymbol{s}}^{(\nu)})$ at each iteration ν, is to choose $\mathbf{H}_n^{(q,\nu)}$ such that it verifies $\|\mathbf{H}_n^{(q,\nu)}\| \leq 2/L_n^{(q)}$.

  Different choices for $\mathbf{H}_n^{(q,\nu)}$ are possible and they correspond to different optimization methods that belong, in general, to the variable metric class. For example $\mathbf{H}_n^{(q,\nu)} = \frac{2}{L_n^{(q)}}\mathbf{I}$ leads to gradient ascent, while taking $\mathbf{H}_n^{(q,\nu)}$ as a scaled inverse of the Hessian matrix would lead to a Newton-Raphson optimization step. Other possibilities include Levenberg-Marquardt and quasi-Newton methods [Polyak 1987].

5.1.4 Analysis of *Local Search* Procedure

Each instance of (5.13) for $n = 1, \ldots, N$ can be solved independently. In this Section we focus on providing a set of conditions under which each iteration of our algorithm guarantees that the objective function $Q_n^{(q)}(\boldsymbol{s})$ in (5.13) is increased. We start by rewriting (5.13) more conveniently in order to perform the optimization with respect to $\boldsymbol{s} \in \mathbb{S}$. To simplify the notation, the iteration index q is sometimes omitted. We simply write $Q_n(\boldsymbol{s})$ for $Q_n^{(q)}(\boldsymbol{s})$.

Let $\bar{\alpha}_n = \sum_{m=1}^{M} \alpha_{mn}^{(q)}$ and $\bar{\beta}_n = \sum_{k=1}^{K} \beta_{kn}^{(q)}$ denote the average object weights in each one of the two modalities. We introduce $\alpha_n = \bar{\alpha}_n^{-1}(\alpha_{1n}^{(q)}, \ldots, \alpha_{Mn}^{(q)})$ and $\beta_n = \bar{\beta}_n^{-1}(\beta_{1n}^{(q)}, \ldots, \beta_{Kn}^{(q)})$ the discrete probability distributions obtained by normalizing the object weights. We denote by $\boldsymbol{F}$ and $\boldsymbol{G}$ the random variables that take their values in the discrete sets $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\}$ and $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k, \ldots, \boldsymbol{g}_K\}$. It follows that the expres-

sions for the optimal variances (5.10) and (5.11) as functions of $\boldsymbol{s}$, can be rewritten as:

$$\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s}) = \mathbb{E}_{\alpha_n}[\big(\boldsymbol{F} - \mathcal{F}(\boldsymbol{s})\big)\big(\boldsymbol{F} - \mathcal{F}(\boldsymbol{s})\big)^\top], \tag{5.15}$$

$$\boldsymbol{\Gamma}_n^{(q+1)}(\boldsymbol{s}) = \mathbb{E}_{\beta_n}[\big(\boldsymbol{G} - \mathcal{G}(\boldsymbol{s})\big)\big(\boldsymbol{G} - \mathcal{G}(\boldsymbol{s})\big)^\top], \tag{5.16}$$

where $\mathbb{E}_{\alpha_n}$ and $\mathbb{E}_{\beta_n}$ denote the expectations with respect to the distributions α_n and β_n. Using some standard projection formula, it follows that the covariances are:

$$\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s}) = \mathbf{V}_f + \boldsymbol{v}_f \boldsymbol{v}_f^\top, \tag{5.17}$$

$$\boldsymbol{\Gamma}^{(q+1)}(\boldsymbol{s}) = \mathbf{V}_g + \boldsymbol{v}_g \boldsymbol{v}_g^\top, \tag{5.18}$$

where $\mathbf{V}_f$ and $\mathbf{V}_g$ are the covariance matrices of $\boldsymbol{F}$ and $\boldsymbol{G}$ respectively under distributions α_n and β_n, and $\boldsymbol{v}_f$ and $\boldsymbol{v}_g$ are vectors defined by:

$$\boldsymbol{v}_f = \mathbb{E}_{\alpha_n}[\boldsymbol{F}] - \mathcal{F}(\boldsymbol{s}), \tag{5.19}$$

$$\boldsymbol{v}_g = \mathbb{E}_{\beta_n}[\boldsymbol{G}] - \mathcal{G}(\boldsymbol{s}). \tag{5.20}$$

For convenience we omit the index n for $\mathbf{V}_f$, $\mathbf{V}_g$, $\boldsymbol{v}_f$ and $\boldsymbol{v}_g$. Let $\bar{\boldsymbol{f}}_n = \mathbb{E}_{\alpha_n}[\boldsymbol{F}]$ and $\bar{\boldsymbol{g}}_n = \mathbb{E}_{\beta_n}[\boldsymbol{G}]$. This yields:

$$\bar{\boldsymbol{f}}_n = \bar{\alpha}_n^{-1} \sum_{m=1}^{M} \alpha_{mn}^{(q)} \boldsymbol{f}_m, \tag{5.21}$$

$$\bar{\boldsymbol{g}}_n = \bar{\beta}_n^{-1} \sum_{k=1}^{K} \beta_{kn}^{(q)} \boldsymbol{g}_k, \tag{5.22}$$

$$\mathbf{V}_f = \bar{\alpha}_n^{-1} \sum_{m=1}^{M} \alpha_{mn}^{(q)} \boldsymbol{f}_m \boldsymbol{f}_m^\top - \bar{\boldsymbol{f}}_n \bar{\boldsymbol{f}}_n^\top, \tag{5.23}$$

$$\mathbf{V}_g = \bar{\beta}_n^{-1} \sum_{k=1}^{K} \beta_{kn}^{(q)} \boldsymbol{g}_k \boldsymbol{g}_k^\top - \bar{\boldsymbol{g}}_n \bar{\boldsymbol{g}}_n^\top. \tag{5.24}$$

Next we derive a simplified expression for $Q_n(\boldsymbol{s})$ in (5.13) in order to investigate its properties. Notice that one can write (5.13) as the sum $Q_n(\boldsymbol{s}) = Q_{n,\mathcal{F}}(\boldsymbol{s}) + Q_{n,\mathcal{G}}(\boldsymbol{s})$, with:

$$Q_{n,\mathcal{F}}(\boldsymbol{s}) = -\sum_{m=1}^{M} \alpha_{mn}^{(q)} (\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s})\|^2_{\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})} + \log|\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})|), \tag{5.25}$$

and a similar expression for $Q_{n,\mathcal{G}}(\boldsymbol{s})$. Equation (5.25) can be written:

$$Q_{n,\mathcal{F}}(\boldsymbol{s}) = -\bar{\alpha}_n (\mathbb{E}_{\alpha_n}[(\boldsymbol{F} - \mathcal{F}(\boldsymbol{s}))^\top \boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1} (\boldsymbol{F} - \mathcal{F}(\boldsymbol{s}))] + \log|\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})|). \tag{5.26}$$

The first term of (5.26) can be further divided into two terms:

$$\begin{aligned} &\mathbb{E}_{\alpha_n}[(\boldsymbol{F} - \mathcal{F}(\boldsymbol{s}))^\top \boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1} (\boldsymbol{F} - \mathcal{F}(\boldsymbol{s}))] = \\ =&\mathbb{E}_{\alpha_n}[(\boldsymbol{F} - \bar{\boldsymbol{f}}_n)^\top \boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1} (\boldsymbol{F} - \bar{\boldsymbol{f}}_n)] + \boldsymbol{v}_f^\top \boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1} \boldsymbol{v}_f. \end{aligned} \tag{5.27}$$

The Sherman-Morrison formula applied to (5.17) leads to

$$\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1} = \mathbf{V}_f^{-1} - \mathbf{V}_f^{-1}\boldsymbol{v}_f\boldsymbol{v}_f^\top\mathbf{V}_f^{-1}/(1 + D_{n,\mathcal{F}}(\boldsymbol{s})), \tag{5.28}$$

with:

$$D_{n,\mathcal{F}}(\boldsymbol{s}) = \|\mathcal{F}(\boldsymbol{s}) - \bar{\boldsymbol{f}}_n\|^2_{\mathbf{V}_f}. \tag{5.29}$$

It follows that (5.27) can be written as the sum of:

$$\mathbb{E}_{\alpha_n}[(\boldsymbol{F} - \bar{\boldsymbol{f}}_n)^\top\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1}(\boldsymbol{F} - \bar{\boldsymbol{f}}_n)] = C_f - \frac{D_{n,\mathcal{F}}(\boldsymbol{s})}{1 + D_{n,\mathcal{F}}(\boldsymbol{s})}, \tag{5.30}$$

and of

$$\boldsymbol{v}_f^\top\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})^{-1}\boldsymbol{v}_f = \frac{D_{n,\mathcal{F}}(\boldsymbol{s})}{1 + D_{n,\mathcal{F}}(\boldsymbol{s})}. \tag{5.31}$$

Hence the first term of (5.26), namely (5.27) is equal to C_f which is constant with respect to $\boldsymbol{s}$. Moreover, applying the matrix determinant lemma to the second term of (5.26) we successively obtain:

$$\begin{aligned}\log|\boldsymbol{\Sigma}_n^{(q+1)}(\boldsymbol{s})| &= \log|\mathbf{V}_f + \boldsymbol{v}_f\boldsymbol{v}_f^\top| = \log|\mathbf{V}_f| + \log(1 + \boldsymbol{v}_f^\top\mathbf{V}_f^{-1}\boldsymbol{v}_f) = \\ &= \log|\mathbf{V}_f| + \log(1 + D_{n,\mathcal{F}}(\boldsymbol{s})).\end{aligned} \tag{5.32}$$

It follows that there is only one term depending on $\boldsymbol{s}$ in (5.26):

$$Q_{n,\mathcal{F}}(\boldsymbol{s}) = -\bar{\alpha}_n\left(C_f + \log|\mathbf{V}_f| + \log(1 + D_{n,\mathcal{F}}(\boldsymbol{s}))\right). \tag{5.33}$$

Repeating the same derivation for the second sensorial modality we obtain the following equivalent form of (5.13):

$$Q_n(\boldsymbol{s}) = -\bar{\alpha}_n\log(1 + D_{n,\mathcal{F}}(\boldsymbol{s})) - \bar{\beta}_n\log(1 + D_{n,\mathcal{G}}(\boldsymbol{s})) + C, \tag{5.34}$$

where C is some constant not depending on $\boldsymbol{s}$.

Using this form of $Q_n(\boldsymbol{s})$, we can now investigate the properties of its gradient $\boldsymbol{\nabla}Q_n(\boldsymbol{s})$. It appears that under some regularity assumptions on $\mathcal{F}$ and $\mathcal{G}$, the gradient $\boldsymbol{\nabla}Q_n(\boldsymbol{s})$ is bounded and Lipschitz continuous. The corresponding theorem is formulated and proved. First we establish as a lemma some technical results, required to prove the theorem. In what follows, for any matrix $\mathbf{V}$, the matrix norm used is the operator norm $\|\mathbf{V}\| = \sup_{\|\boldsymbol{v}\|=1}\|\mathbf{V}\boldsymbol{v}\|$. For simplicity, we further omit the index n.

Lemma 1 *Let* $\mathbf{V}$ *be a symmetric positive definite matrix. Then the function*

$$\varphi(\boldsymbol{v}) = \|\mathbf{V}\boldsymbol{v}\|/(1 + \boldsymbol{v}^\top\mathbf{V}\boldsymbol{v})$$

is bounded by $\varphi(\boldsymbol{v}) \leq C_\varphi(\mathbf{V})$ *with* $C_\varphi(\mathbf{V}) = \sqrt{\|\mathbf{V}\|}/2$ *and is Lipschitz continuous:*

$$\forall\boldsymbol{v}, \tilde{\boldsymbol{v}} \quad \|\varphi(\boldsymbol{v}) - \varphi(\tilde{\boldsymbol{v}})\| \leq L_\varphi(\mathbf{V})\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|,$$

where $L_\varphi(\mathbf{V}) = \|\mathbf{V}\|(1 + \mu(\mathbf{V})/2)$ *is the Lipschitz constant and* $\mu(\mathbf{V}) = \|\mathbf{V}\|\|\mathbf{V}^{-1}\|$ *is the condition number of* $\mathbf{V}$.

Proof: We start by introducing $\boldsymbol{w} = \mathbf{V}\boldsymbol{v}$ so that $\varphi(\boldsymbol{v}) = \tilde{\varphi}(\boldsymbol{w}) = \|\boldsymbol{w}\|/(1 + \boldsymbol{w}^\top\mathbf{V}^{-1}\boldsymbol{w})$. As soon as $\boldsymbol{w}^\top\mathbf{V}^{-1}\boldsymbol{w} \geq \lambda_{\min}\|\boldsymbol{w}\|^2$ (where we denoted by $\lambda_{\min}$ the smallest eigenvalue of $\mathbf{V}^{-1}$, so that in fact $\lambda_{\min} = \|\mathbf{V}\|^{-1}$), to find the maximum of $\tilde{\varphi}(\boldsymbol{w})$ we should maximize the expression $t/(1+\lambda_{\min}t^2)$ for $t = \|\boldsymbol{w}\| \geq 0$. It is reached at the point $t^* = \lambda_{\min}^{-1/2}$. Substituting this value into the original expressions gives $\varphi(\boldsymbol{v}) \leq \sqrt{\|\mathbf{V}\|}/2$.

To compute the Lipschitz constant L_φ we consider the derivative:

$$\|\boldsymbol{\nabla}\tilde{\varphi}'(\boldsymbol{w})\| = \frac{\left\|(1+\boldsymbol{w}^\top\mathbf{V}^{-1}\boldsymbol{w})\boldsymbol{w} - 2\|\boldsymbol{w}\|^2\mathbf{V}^{-1}\boldsymbol{w}\right\|}{\|\boldsymbol{w}\|(1+\boldsymbol{w}^\top\mathbf{V}^{-1}\boldsymbol{w})^2} \leq 1 + \frac{2\|\mathbf{V}^{-1}\|\|\boldsymbol{w}\|^2}{(1+\boldsymbol{w}^\top\mathbf{V}^{-1}\boldsymbol{w})^2},$$

from where we find that $\|\boldsymbol{\nabla}\tilde{\varphi}'(\boldsymbol{w})\| \leq 1 + \mu(\mathbf{V})/2$, and so $L_\varphi = \|\mathbf{V}\|(1+\mu(\mathbf{V})/2)$. ■

This lemma yields the following main result for the gradient $\boldsymbol{\nabla}Q$:

Theorem 2 *Assume functions $\mathcal{F}$ and $\mathcal{G}$ and their derivatives $\mathcal{F}'$ and $\mathcal{G}'$ are Lipschitz continuous with constants $L_\mathcal{F}$, $L_\mathcal{G}$, $L'_\mathcal{F}$ and $L'_\mathcal{G}$ respectively. Then the gradient $\boldsymbol{\nabla}Q$ is bounded and Lipschitz continuous with some constant L.*

Proof: From (5.34) the gradient $\boldsymbol{\nabla}Q$ can be written as:

$$\begin{aligned}\boldsymbol{\nabla}Q(\boldsymbol{s}) &= \boldsymbol{\nabla}Q_\mathcal{F}(\boldsymbol{s}) + \boldsymbol{\nabla}Q_\mathcal{G}(\boldsymbol{s}) = \\ &= \frac{2\bar{\alpha}\mathcal{F}'^\top(\boldsymbol{s})\mathbf{V}_f^{-1}(\bar{\boldsymbol{f}} - \mathcal{F}(\boldsymbol{s}))}{1+D_\mathcal{F}(\boldsymbol{s})} + \frac{2\bar{\beta}\mathcal{G}'^\top(\boldsymbol{s})\mathbf{V}_g^{-1}(\bar{\boldsymbol{g}} - \mathcal{G}(\boldsymbol{s}))}{1+D_\mathcal{G}(\boldsymbol{s})}. \end{aligned} \tag{5.35}$$

It follows from Lemma 1 that $\|\boldsymbol{\nabla}Q_\mathcal{F}(\boldsymbol{s})\| \leq 2L_\mathcal{F}\bar{\alpha}C_\varphi(\mathbf{V}_f^{-1})$ and $\|\boldsymbol{\nabla}Q_\mathcal{G}(\boldsymbol{s})\| \leq 2L_\mathcal{G}\bar{\beta}C_\varphi(\mathbf{V}_g^{-1})$. The norm of the gradient is then bounded by:

$$\|\boldsymbol{\nabla}Q(\boldsymbol{s})\| \leq 2L_\mathcal{F}\bar{\alpha}C_\varphi(\mathbf{V}_f^{-1}) + 2L_\mathcal{G}\bar{\beta}C_\varphi(\mathbf{V}_g^{-1}). \tag{5.36}$$

Considering the norm $\|\boldsymbol{\nabla}Q_\mathcal{F}(\boldsymbol{s}) - \boldsymbol{\nabla}Q_\mathcal{F}(\tilde{\boldsymbol{s}})\|$, we introduce $\boldsymbol{v}_1 = \bar{\boldsymbol{f}} - \mathcal{F}(\boldsymbol{s})$ and $\boldsymbol{v}_2 = \bar{\boldsymbol{f}} - \mathcal{F}(\tilde{\boldsymbol{s}})$. Then we have:

$$\begin{aligned}\|\boldsymbol{\nabla}Q_\mathcal{F}(\boldsymbol{s}) - \boldsymbol{\nabla}Q_\mathcal{F}(\tilde{\boldsymbol{s}})\| \leq 2\bar{\alpha}\Bigg(&\left\|\frac{(\mathcal{F}'(\boldsymbol{s}) - \mathcal{F}'(\tilde{\boldsymbol{s}}))^\top\mathbf{V}_f^{-1}\boldsymbol{v}_1}{1+\|\boldsymbol{v}_1\|^2_{\mathbf{V}_f}}\right\| + \\ &+ \left\|\frac{\mathcal{F}'^\top(\tilde{\boldsymbol{s}})\mathbf{V}_f^{-1}\boldsymbol{v}_2}{1+\|\boldsymbol{v}_2\|^2_{\mathbf{V}_f}} - \frac{\mathcal{F}'^\top(\tilde{\boldsymbol{s}})\mathbf{V}_f^{-1}\boldsymbol{v}_1}{1+\|\boldsymbol{v}_1\|^2_{\mathbf{V}_f}}\right\|\Bigg).\end{aligned} \tag{5.37}$$

Using Lemma 1 with $\mathbf{V}_f^{-1}$ we have:

$$\|\boldsymbol{\nabla}Q_\mathcal{F}(\boldsymbol{s}) - \boldsymbol{\nabla}Q_\mathcal{F}(\tilde{\boldsymbol{s}})\| \leq 2\bar{\alpha}\big(L'_\mathcal{F}C_\varphi(\mathbf{V}_f^{-1}) + L^2_\mathcal{F}L_\varphi(\mathbf{V}_f^{-1})\big)\|\boldsymbol{s} - \tilde{\boldsymbol{s}}\|.$$

The same derivations can be performed for $\boldsymbol{\nabla}Q_\mathcal{G}(\boldsymbol{s})$, so that finally we get:

$$\|\boldsymbol{\nabla}Q_\mathcal{G}(\boldsymbol{s}) - \boldsymbol{\nabla}Q_\mathcal{G}(\tilde{\boldsymbol{s}})\| \leq L\|\boldsymbol{s} - \tilde{\boldsymbol{s}}\|, \tag{5.38}$$

where the Lipschitz constant is given by:

$$L = 2\bar{\alpha}\left(L'_\mathcal{F}C_\varphi(\mathbf{V}_f^{-1}) + L^2_\mathcal{F}L_\varphi(\mathbf{V}_f^{-1})\right) + 2\bar{\beta}\left(L'_\mathcal{G}C_\varphi(\mathbf{V}_g^{-1}) + L^2_\mathcal{G}L_\varphi(\mathbf{V}_g^{-1})\right). \tag{5.39}$$

■

To actually construct the non-decreasing sequence in (5.14), we make use of the following fundamental result on variable metric gradient ascent algorithms.

Theorem 3 ([Polyak 1987]) *Let the function $Q : \mathbb{R}^d \to \mathbb{R}$ be differentiable on $\mathbb{R}^d$ and its gradient $\boldsymbol{\nabla} Q$ be Lipschitz continuous with constant L. Let the matrix $\mathbf{H}$ be positive definite, such that $\|\mathbf{H}\| \leq \frac{2}{L}$. Then the sequence $Q(\tilde{\boldsymbol{s}}^{(\nu)})$, defined by $\tilde{\boldsymbol{s}}^{(\nu+1)} = \tilde{\boldsymbol{s}}^{(\nu)} + \mathbf{H}\boldsymbol{\nabla} Q(\tilde{\boldsymbol{s}}^{(\nu)})$ is non-decreasing.*

This result shows that for any functions $\mathcal{F}$ and $\mathcal{G}$ that verify the conditions of Theorem 2, using (5.14) with $\mathbf{H} = \frac{2}{L}\mathbf{I}$, we are able to construct a non-decreasing sequence and an appropriate *Local Search* procedure. Notice however, that its guaranteed theoretical convergence speed is linear. It can be improved in several ways.

First, the optimization *direction* can be adjusted. For certain problems, the matrix $\mathbf{H}$ can be chosen as in variable metric algorithms, such as Newton-Raphson method, quasi-Newton methods or Levenberg-Marquardt method, provided that it satisfies the conditions of Theorem 3. Second, the optimization *step size* can be increased based on local properties of the target function. For example, at iteration ν, if when considering the functions $\mathcal{F}$ and $\mathcal{G}$ on some restricted domain $\mathbb{S}^{(\nu)}$ there exist smaller local Lipschitz constants $L_{\mathcal{F}}^{(\nu)}$, $L_{\mathcal{G}}^{(\nu)}$, $L_{\mathcal{F}}^{\prime(\nu)}$ and $L_{\mathcal{G}}^{\prime(\nu)}$, $\mathbf{H}$ can be set to $\mathbf{H} = \frac{2}{L^{(\nu)}}\mathbf{I}$ with $L^{(\nu)}$ smaller than L. It follows that $\|\tilde{\boldsymbol{s}}^{(\nu+1)} - \tilde{\boldsymbol{s}}^{(\nu)}\| \leq \frac{2}{L^{(\nu)}}\|\boldsymbol{\nabla} Q(\tilde{\boldsymbol{s}}^{(\nu)})\|$, which means that one can take the local constants, $L_{\mathcal{F}}^{(\nu)}$, $L_{\mathcal{G}}^{(\nu)}$, $L_{\mathcal{F}}^{\prime(\nu)}$ and $L_{\mathcal{G}}^{\prime(\nu)}$ if they are valid in the ball $\mathbb{B}_{\rho^{(\nu)}}(\tilde{\boldsymbol{s}}^{(\nu)})$ with

$$\rho^{(\nu)} = \frac{2}{L^{(\nu)}}\left(2L_{\mathcal{F}}^{(\nu)}\bar{\alpha}C_{\varphi}(\mathbf{V}_f^{-1}) + 2L_{\mathcal{G}}^{(\nu)}\bar{\beta}C_{\varphi}(\mathbf{V}_g^{-1})\right). \tag{5.40}$$

5.1.5 Global Search and the *Choose* Procedure

Theorem 2 allows us to use the improved global random search techniques for Lipschitz continuous functions [Zhigljavsky 1991]. These algorithms are known to converge, in the sense that generated point sequences fall infinitely often into an arbitrarily small neighbourhood of the optimal points set. For more details and convergence conditions see Theorem 3.2.1 and the discussion that follows in [Zhigljavsky 1991]. A proper choice of the initial value $\tilde{\boldsymbol{s}}^{(0)}$ not only guarantees to find the global maximum, but can also be used to increase the convergence speed. A basic strategy is to draw samples in $\mathbb{S}$, according to some sequence of distributions over $\mathbb{S}$, that verifies the convergence conditions of global random search methods. However, the speed of convergence of such an algorithm is quite low.

Global random search methods can also be significantly improved by taking into account some specificities of the target function. Indeed, in our case, function (5.34) is made of two parts for which the optimal points are known and are respectively $\bar{\boldsymbol{f}}$ and $\bar{\boldsymbol{g}}$. If there exists $\tilde{\boldsymbol{s}}^{(0)}$ such that $\tilde{\boldsymbol{s}}^{(0)} \in \mathcal{F}^{-1}(\bar{\boldsymbol{f}}) \cap \mathcal{G}^{-1}(\bar{\boldsymbol{g}})$, then it is the global maximum and the M step solution is found. Otherwise, one can sample $\mathbb{S}$ in the vicinity of the set $\mathcal{F}^{-1}(\bar{\boldsymbol{f}}) \cup \mathcal{G}^{-1}(\bar{\boldsymbol{g}})$

to focus on a subspace that is likely to contain the global maximum. This set is, generally speaking, a union of two manifolds. For sampling methods on manifolds we refer to [Zhigljavsky 1991]. An illustration of this technique is given in the Appendix A.1.

Another possibility is to use a heuristic that function (5.34) does not change much after one iteration of the EM algorithm. Then, the initial point $\tilde{\boldsymbol{s}}^{(0)}$ for the current iteration can be set to the optimal value computed at the previous iteration. However, in general, this simple strategy does not yield the global maximum, as can be seen from the results in Section 5.3.

5.2 Clustering Using Auditory and Visual Data

As in the previous Chapter, we illustrate the method in the case of audio-visual (AV) objects. The objects are characterized both by their locations in space $\boldsymbol{s}_n$ and by their auditory status, i.e., whether they are emitting sounds or not. These object characteristics are not directly observable and hence they need to be inferred from sensor data. A typical example where the conjugate mixture models framework may help is the task of locating several speaking persons.

Using the same notations as above, we consider two sensor spaces. The multimodal data consists of M visual observations $\mathbf{f}$ and of K auditory observations $\boldsymbol{g}$. We consider data that are recorded over a short time interval $[t_1, t_2]$, such that one can reasonably assume that the AV objects have a stationary spatial location. Nevertheless, it is not assumed here that the AV objects, e.g., speakers, are static: lip movements, head and hand gestures are tolerated. We address the problem of estimating the spatial locations of all the objects that are both seen and heard. Let N be the number of objects and in this case each object is described by a three dimensional parameter vector $\boldsymbol{s}_n = (x_n, y_n, z_n)^\top$.

As in previous Chapters we define an invertible function $\mathcal{F} : \mathbb{R}^3 \to \mathbb{R}^3$ that maps a 3D location $\boldsymbol{s} = (x, y, z)^\top$ onto a visual space 3D point $\boldsymbol{f} = (u, v, d)^\top$:

$$\mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{1}{z}\right)^\top \quad \text{and} \quad \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{u}{d}, \frac{v}{d}, \frac{1}{d}\right)^\top. \tag{5.41}$$

We remind that this model, introduced in Chapter 2, corresponds to a rectified camera pair [Hartley 2003] and can be easily generalized to more complex binocular geometries [Hansard 2007, Hansard 2008]. Without loss of generality one can use a sensor-centered coordinate system to represent the object locations.

Similarly we introduce the ITD function $\mathcal{G} : \mathbb{R}^3 \to \mathbb{R}$ defined in Chapter 2 that maps a 3D location $\boldsymbol{s} = (x, y, z)^\top$ onto a 1D auditory observation:

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big). \tag{5.42}$$

Here c is the sound speed and $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ are the 3D locations of the two microphones in the sensor-centered coordinate system. Again, the setup is supposed to be calibrated,

so that the left and right microphone positions $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ are known. To simplify the notation we would further write $\mathcal{G}(\boldsymbol{s})$ instead of $\mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})$.

In order to perform audio-visual clustering based on the conjugate EM algorithm, Theorem 2 (Section 5.1.4) must hold for both (5.41) and (5.42), namely the functions $\mathcal{F}$ and $\mathcal{G}$ and their derivatives are Lipschitz continuous. We prove the following theorem:

Theorem 4 *The functions $\mathcal{F}$, $\mathcal{F}'$, $\mathcal{G}$ and $\mathcal{G}'$ are Lipschitz continuous with constants $L_{\mathcal{F}} = z_{\min}^{-1}\sqrt{3}$, $L'_{\mathcal{F}} = z_{\min}^{-2}$, $L_{\mathcal{G}} = \|\boldsymbol{s}_{\mathrm{M}_1} - \boldsymbol{s}_{\mathrm{M}_2}\|(cR)^{-1}$ and $L'_{\mathcal{G}} = 3(cR)^{-1}$ in the domain $\mathbb{S} = \{|z| > z_{\min} > 1\} \cap \Big\{ \min\{\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}\|, \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}\|\} > R > 1 \Big\}$.*

Proof: The derivatives of $\mathcal{F}$ and $\mathcal{G}$ are given by:

$$\mathcal{F}'(\boldsymbol{s}) = \frac{1}{z}\begin{bmatrix} 1 & 0 & -x/z \\ 0 & 1 & -y/z \\ 0 & 0 & -1/z \end{bmatrix} \tag{5.43}$$

$$\mathcal{G}'(\boldsymbol{s}) = \frac{1}{c}\left(\frac{\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}}{\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}\|} - \frac{\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}}{\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}\|}\right). \tag{5.44}$$

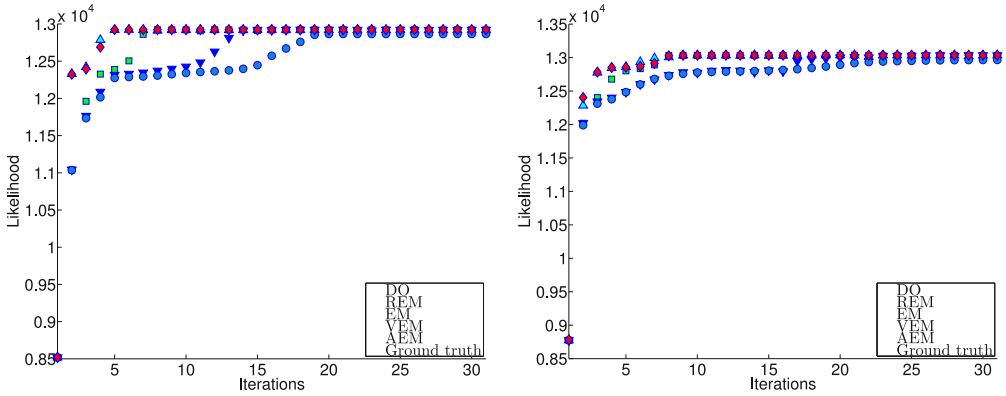
The eigenvalues of $\mathcal{F}'(\boldsymbol{s})$ are $1/z$ and $-1/z^2$, so $\|\mathcal{F}'(\boldsymbol{s})\| \leq \max\{z^{-1}, z^{-2}\} \leq z_{\min}^{-1}$, from which it follows that $L_{\mathcal{F}}$ can be taken as $L_{\mathcal{F}} = z_{\min}^{-1}\sqrt{3}$. Also $\|\mathcal{F}'(\boldsymbol{s}) - \mathcal{F}'(\tilde{\boldsymbol{s}})\| \leq \max\{|z^{-1} - \tilde{z}^{-1}|, |z^{-2} - \tilde{z}^{-2}|\} \leq z_{\min}^{-2}\|\boldsymbol{s} - \tilde{\boldsymbol{s}}\|$, so that $L'_{\mathcal{F}}$ can be set to $L'_{\mathcal{F}} = z_{\min}^{-2}$.

Introducing $\boldsymbol{e}_1 = \frac{\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M1}}}{\|\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M}_1}\|}$ and $\boldsymbol{e}_2 = \frac{\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M2}}}{\|\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M}_2}\|}$, it comes $\|\boldsymbol{e}_1\| = \|\boldsymbol{e}_2\| = 1$ and $\mathcal{G}'(\boldsymbol{s}) = \frac{1}{c}(\boldsymbol{e}_1 - \boldsymbol{e}_2)$. Provided that $\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}\|$ and $\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}\|$ are both greater than R, it follows $\|\mathcal{G}'(\boldsymbol{s})\| = \frac{1}{c}\|\boldsymbol{e}_1 - \boldsymbol{e}_2\| \leq \|\boldsymbol{s}_{\mathrm{M}_1} - \boldsymbol{s}_{\mathrm{M}_2}\|(cR)^{-1}$ and so $L_{\mathcal{G}} = \|\boldsymbol{s}_{\mathrm{M}_1} - \boldsymbol{s}_{\mathrm{M}_2}\|(cR)^{-1}$. Then, the second derivative of $\mathcal{G}$ is given by

$$\mathcal{G}''(\boldsymbol{s}) = \frac{1}{c\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}\|}(\mathbf{I} - \boldsymbol{e}_1\boldsymbol{e}_1^\top) - \frac{1}{c\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}\|}(\mathbf{I} - \boldsymbol{e}_2\boldsymbol{e}_2^\top).$$

so that $\|\mathcal{G}''(\boldsymbol{s})\| \leq \left|\frac{1}{c\|\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M1}}\|} - \frac{1}{c\|\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M2}}\|}\right| + \sup_{\|\boldsymbol{v}\|=1} \frac{2\boldsymbol{e}_1\boldsymbol{e}_1^\top\boldsymbol{v}}{c\min\{\|\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M1}}\|,\|\boldsymbol{s}-\boldsymbol{s}_{\mathrm{M2}}\|\}} \leq 3(cR)^{-1}$, and $L'_{\mathcal{G}}$ can be set to $L'_{\mathcal{G}} = 3(cR)^{-1}$. ■

This result shows that under some natural conditions (The AV objects should not be too close to the sensors) the conjugate EM algorithm described in Section 5.1.3 can be applied. The constant L given by Lemma 2 guarantees a certain (worst-case) convergence speed. In practice, we can use the techniques mentioned in Sections 5.1.4 and 5.1.5 to accelerate the algorithm. First, to speed up the local optimization step, local Lipschitz constants can be computed based on the current value of parameter $\tilde{\boldsymbol{s}}^{(\nu)}$. Equation (5.40) gives the largest possible step size $\rho^{(\nu)}$, so setting $z_{\min}^{(\nu)} = z^{(\nu)} - \rho^{(\nu)}$ and $R^{(\nu)} = \min\{\|\tilde{\boldsymbol{s}}^{(\nu)} - \boldsymbol{s}_{\mathrm{M}_2}\|, \|\tilde{\boldsymbol{s}}^{(\nu)} - \boldsymbol{s}_{\mathrm{M}_1}\|\} - \rho^{(\nu)}$, provides local Lipschitz constants that ensure the update not to quit $\mathbb{S}^{(\nu)} = \{|z| > z_{\min}^{(\nu)}\} \cap \Big\{ \min\{\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}\|, \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}\|\} > R^{(\nu)} \Big\}$. Second, we propose four possibilities to set the initial object parameter values $\tilde{\boldsymbol{s}}_n^{(0)}$: (i) it can be taken



(a) **WS** configuration (b) **PS** configuration

Figure 5.1: Likelihood function evolution for the ConjKP and ConjEM algorithms for the cases of (a) well-separated objects, and (b) poorly separated objects.

to be the previously estimated object position $\boldsymbol{s}_n^{(q-1)}$, (ii) it can be set to $\mathcal{F}^{-1}(\bar{\boldsymbol{f}})$ (as soon as $\mathcal{F}$ is injective in $\mathbb{S}$), (iii) it can be found through sampling of the manifold $\mathcal{G}^{-1}(\bar{\boldsymbol{g}})$ by selecting the sampled value which gives the largest Q value, or (iv) similarly through sampling directly in $\mathbb{S}$. Comparisons are reported in the following Sections.

5.3 Experimental Validation

5.3.1 Experiments with Simulated Data

Our algorithm is first illustrated on the simulated data described in Section 4.4. The goal is to compare the performance to that of the algorithms from the Chapter 4. As previously, three cases are considered: well separated (**WS**), moderately separated (**MS**) and poorly separated (**PS**) object configurations, the observations are shown in Figure 4.3. The initialization settings are the same as in the experiments with the KP algorithm: close (**CI**), intermediate (**II**) and far (**FI**).

The convergence speed of the accelerated ConjEM algorithm was verified on the **WS** and **PS** object configurations. The likelihood evolution graphs are presented in Figure 5.1 (cf. Figure 4.5). The three versions of the ConjKP algorithm from the previous Chapter – direct optimization (**DO**), relaxed ConjEM (**REM**), ConjEM (**EM**), – along with simple visual data-based EM (**VEM**) are compared The convergence speed of the accelerated version of ConjEM algorihtm is the same as the convergence speed of direct optimization, whereas the complexity of the algorithm was significantly reduced. Also, the newly designed algorithm is well established from the point of view of optimization theory. It is guaranteed to improve the log-likelihood function value, even in the case of the generalized ConjEM algorithm.

Average absolute error values ε_{abs} for object location estimates $\hat{\boldsymbol{s}}$ are summarized in Table 5.2 (cf. Table 4.2). Three versions are compared: normal conjugate EM algorithm (**EM**), accelerated conjugate EM algorithm (**AEM**) and a simple EM based on visual data only (**VEM**).

To determine, which acceleration strategy is the best, we compare the performance of several versions of the ConjEM algorithm based on various *Choose* and *Local Search* strategies. For the initial values $\tilde{\boldsymbol{s}}_n^{(0)}$ in (5.14), we considered the following possibilities: the optimal value computed at a previous run of the algorithm (IP), the value predicted from visual data (IV), the value predicted from audio data (IA) and the value obtained by global random search (IG). More specifically:x

- When initializing from visual data (IV), the average value $\bar{\boldsymbol{f}}_n$, calculated in the current E-step of the algorithm for every n, was mapped to the parameter space and $\tilde{\boldsymbol{s}_n}^{(0)}$ set to $\tilde{\boldsymbol{s}_n}^{(0)} = \mathcal{F}^{-1}(\bar{\boldsymbol{f}}_n)$ using the injectivity of $\mathcal{F}$.

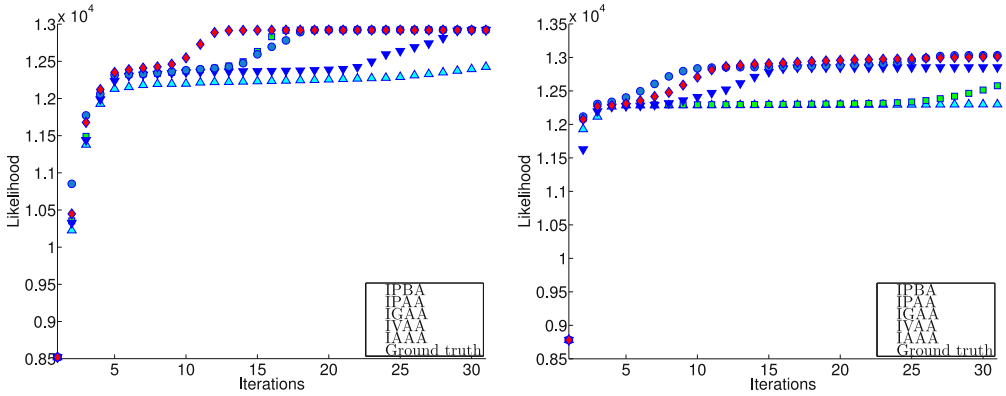
- When initializing from audio data (IA), $\mathcal{G}^{-1}(\bar{\boldsymbol{g}}_n)$ defines a manifold. The general strategy here would be to find the optimal point that lies on this surface. We achieved this through random search based on a uniform sampling on the corresponding part of the hyperboloid (see [Zhigljavsky 1991] for details on sampling from an arbitrary distribution on a manifold and Appendix A.1 for details on sampling the surface defined by $\mathcal{G}^{-1}(\bar{\boldsymbol{g}}_n)$); in our experiments we used 50 samples to select the one providing the largest Q (likelihood) value.

- The most general initialization scheme (IG) was implemented using global random search in the whole parameter space $\mathbb{S}$; 200 samples were used in this case.

Local optimization was performed either using basic gradient ascent (BA) or the locally accelerated gradient ascent (AA). The latter used the local Lipschitz constants to augment the step size, as described in Section 5.1.4. Each algorithm run consisted of 30 iterations of the EM algorithm with 10 non-decreasing iterations during the M step.

To check the convergence speed of different versions of the algorithm for the **WS** and **PS** object configurations we compared the likelihood evolution graphs that are presented in Figure 5.2. Each graph contains several curves that correspond to five different versions of the algorithm. The acronyms we use to refer to the different versions (for example, IPAA) consist of two parts encoding the initialization (IP) and the local optimization (AA) types. The black dashed line on each graph shows the 'ground truth' likelihood level, that is the likelihood value for the parameters used to generate the data. The meaning of the acronyms is recalled in Table 5.1.

As expected, the simplest version IPBA that uses none of the proposed acceleration techniques appears to be the slowest. The other variants using basic gradient ascent are then not reported. Predicting a single object parameter value from visual observations (IVAA) gives certain improvement over IPAA, where $\tilde{\boldsymbol{s}}^{(0)}$ is taken from the previous EM iteration. When $\tilde{\boldsymbol{s}}^{(0)}$ is obtained by sampling the hyperboloid predicted from audio observations



(a) **WS** configuration (b) **PS** configuration

Figure 5.2: Likelihood function evolution for five variants of the algorithm for the cases of (a) well-separated objects, and (b) poorly separated objects.

Acronym	$\tilde{\boldsymbol{s}}^{(0)}$ initialization (*Choose*)	Local optimization (*Search*)
IPBA	previous iteration value	basic gradient ascent
IGAA	global random search	accelerated gradient ascent
IVAA	predicted value from visual data	accelerated gradient ascent
IPAA	previous iteration value	accelerated gradient ascent
IAAA	audio predicted manifold sampling	accelerated gradient ascent

Table 5.1: Acronyms used for five variants of the conjugate EM algorithm. Variants correspond to different choices for the *Choose* and *Local search* procedures.

(IAAA), a significant impact on the convergence speed is observed, especially on early stages of the algorithm, where the predicted value can be quite far from the optimal one. However, 'blind' sampling of the whole parameter space does not bring any advantage: it is much less efficient regarding the number of samples required for the same precision. This suggests that in the general case, the best strategy would be to sample the manifolds $\mathcal{F}^{-1}(\bar{\boldsymbol{f}}_n)$ and $\mathcal{G}^{-1}(\bar{\boldsymbol{g}}_n)$ with possible small perturbations to find the best $\tilde{\boldsymbol{s}}^{(0)}$ estimate and to perform an accelerated gradient ascent afterwards (IAAA). We note that IAAA succeeds in all the cases to find parameter values that are well-fitted to the model in terms of likelihood function (likelihood is greater or equal than that of real parameter values).

Parameter evolution trajectories for the IAAA version of the algorithm in the **WS** case are shown in Figures 5.3-5.4. The estimate changes are reflected by the node sizes (from smaller to bigger) and colours (from darker to lighter). The final values are very close to the real cluster centers in all three audio, visual and object spaces. The convergence speed is quite dependent on the initialization. In the provided example the algorithm spent certain number of iterations to disentangle the estimates trying to decide which one corresponds to which class. Another possibility here would be to predict the initial values through sampling in the audio domain. We demonstrate this strategy further when working with

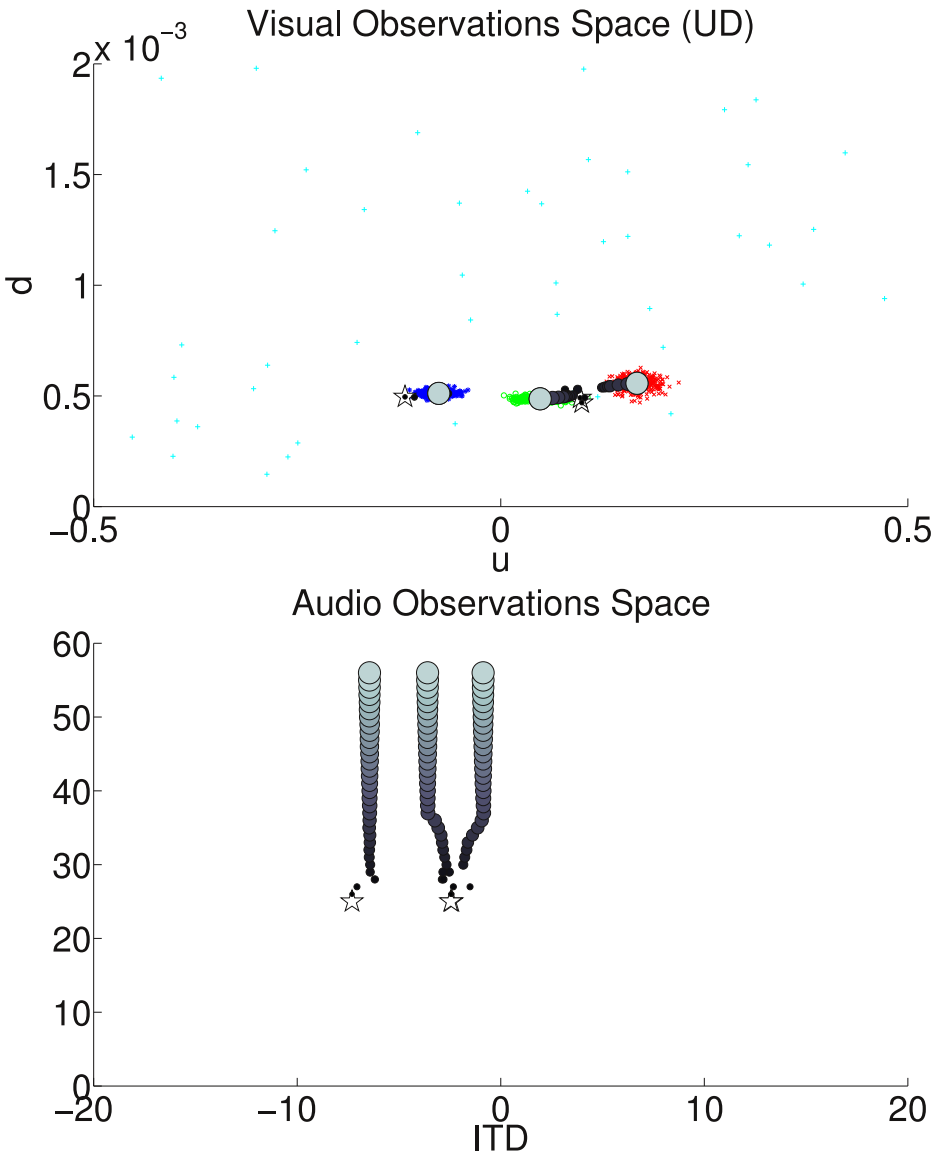


Figure 5.3: IAAA algorithm: parameter evolution and assignment results for the **WS** case in audio and visual spaces (note the scale change which corresponds to a zoom on the cluster centers). Ground truth means are marked with squares. The evolution is shown by circles from smaller to bigger, from darker to brighter. Observations assignments are depicted by different markers ($\circ$, $*$ and $\times$ for the three object classes) in visual space and are colour-coded in audio space. Due to the zoom, outliers are not visible on these figures.

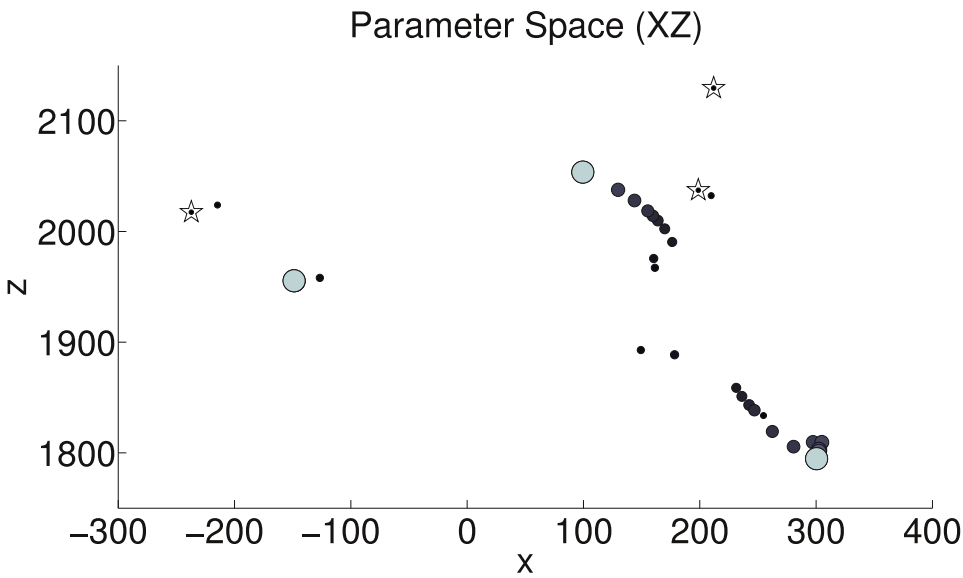


Figure 5.4: IAAA algorithm: parameter evolution for the **WS** case in object space. Ground truth means are marked with squares. The evolution is shown by circles from smaller to bigger, from darker to brighter.

	IC			**II**			**IF**		
	EM	AEM	VEM	EM	AEM	VEM	EM	AEM	VEM
WS	4.62	3.9	3.92	4.33	3.9	3.92	4.5	3.9	3.92
MS	3.22	4.48	4.66	2.81	4.48	4.66	3.58	4.48	4.66
PS	2.93	3.71	3.72	3.13	3.71	3.73	3.1	3.71	3.74

Table 5.2: Summary of average absolute error ε_{abs} values for object location estimates $\hat{s}$ for **WS**, **MS** and **PS** object configurations. The conjugate EM (**EM**), accelerated ConjEM (**AEM**) and the EM based on visual data only (**VEM**) are compared, dependency on initial values setting (**IC**, **II** or **IF**) is shown.

real data.

It appears that the localization precision is quite high. In a realistic setting such as that of Section 5.3.2, the measurement unit can be set to a millimeter. In that case, the observed precision, in a well-separated objects configuration, it is at worse about 5mm. However, precision in the z coordinate is quite sensitive to the variance of the visual data and the object configuration. To get a better idea of the relationship between the variance in object space and the variance in visual space, $\mathcal{F}^{-1}$ can be replaced by its linear approximation given by a first order Taylor expansion. Assuming then that visual data are distributed according to some probability distribution with mean $\mu_{\mathcal{F}}$ and variance $\Sigma_{\mathcal{F}}$, it follows that through the linear approximation of $\mathcal{F}^{-1}$, the variance in object space is

$\frac{\partial \mathcal{F}^{-1}(\mu_{\mathcal{F}})}{\partial \boldsymbol{f}} \boldsymbol{\Sigma}_{\mathcal{F}} \frac{\partial \mathcal{F}^{-1}(\mu_{\mathcal{F}})}{\partial \boldsymbol{f}}^{\top}$. Then, the z coordinate covariance for an object n is approximately proportional to the d covariance for the object multiplied by z_n^4. For distant objects, a very high precision in d is needed to get a satisfactory precision in z. At the same time we observe that the likelihood of the estimate configuration often exceeds the likelihood for real parameter values. This suggests that the model performs well for the given data, but cannot get better precision than that imposed by the data. The same reasoning, however, can be applied to the EM algorithms that work on single modalties.

These results on simulated data show that the ConjEM algorithm allows an efficient implementation and can be significantly accelerated with respect to the basic version presented in Chapter 4. At the same time it keeps all the advantages outlined previously for the KP algorithm family.

5.3.2 Experiments with Real Data

In this section we evaluate the effectiveness of our algorithms in estimating the 3D locations of AV objects, i.e., a person localization task. The examples used below are from the CAVA database described in detail in Chapter 2.

The experimental setup consists of a *mannequin* equipped with a pair of microphones fixed into its ears and a pair of stereoscopic cameras mounted onto its forehead (this device was developed within the POP[1] project). Each data set comprises two audio tracks, two image sequences, as well as the calibration information. All the recordings were performed in an ordinary room with no special adjustments to its acoustics or appearance. Thus the data contain both visual background information, and auditory noise, reverberations in particular. This configuration best mimics what a person would hear and see in a standard indoor environment.

We tested our multimodal clustering method with three scenarios: a *meeting*, a *moving target* and a *cocktail party*, Table 5.3:

- The meeting scenario[2] is a recording of a discussion held by five persons sitting around a table, only three of them being visible. It lasts 25 seconds and contains a total of about 8000 visual and 600 audio observations. The three visible persons perform head and body movements while taking speech turns. Sometimes two persons (visible or not) speak simultaneously.

- The moving target scenario[3] involves a person walking along a zig-zag trajectory towards the camera while speaking. It also contains various ambient sounds such as footsteps. The scenario lasts 9 seconds and contains a total of about 3500 visual and 260 audio observations.

[1] http://perception.inrialpes.fr/POP/
[2] http://perception.inrialpes.fr/CAVA_Dataset/Site/data.html#M1
[3] http://perception.inrialpes.fr/CAVA_Dataset/Site/data.html#TTOS1

scenario	visible persons	speaking persons	visual background	audio noise	occluded speakers	audio overlap
meeting	3	5	yes	yes	no	yes
tracking	1	1	yes	yes	no	no
cocktail party	3	3	yes	yes	yes	yes

Table 5.3: Summary of the main characteristics of the three scenarios used to evaluate the multimodal clustering algorithm.

- The cocktail party scenario[4] shows a dynamic scene with three persons walking in a room and taking speech turns. Occasionally, one speaker is hidden by another person and two persons may speak simultaneously. Speakers may go in and out of the two cameras field of view. Moreover, there are sounds emitted by the persons' steps. The recording lasts 30 seconds and contains a total of about 12500 visual and 3400 audio observations.

Visual and auditory observations $\mathbf{f}$ and $\mathbf{g}$ were obtained using the methods described in detail in Chapter 2. In order to initialize the algorithm's parameter values we used the *Initialize* procedure based on random data-driven sampling and bootstrap techniques. Further details on this procedure are given in Chapter 6. Although real-data distributions do not strictly correspond to the case of Gaussian mixtures, the initialization strategy that we have adopted remains relevant. This originates from the fact that parameter space sampling with configuration restrictions plays the role of a global optimization method similar to Monte-Carlo sampling in the method of generations [Zhigljavsky 2008]. It helps to avoid local maxima and allows to quickly find a set of appropriate initial parameters. Local distribution density modes occur to be good candidates to initialize cluster centers. As in the case of simulated data, we used a BIC-like information criterion to select the optimal number of audio-visual clusters. Details on the selection procedure would be given further in Chapter 6.

The experimental validation described below was performed with two goals in mind. Firstly, we wanted to check that our method was stable and robust with real data gathered in complex situations, that it correctly finds the number of clusters and that it efficiently determines the model's parameters, i.e., the 3D positions of the audio-visual objects composing a scene. Secondly, we wanted to test the model's capability to deal with dynamic changes in the scene, yet in the presence of acoustic noise/reverberations and visually occluded persons, etc. Below we provide a detailed account of the results obtained with the meeting and cocktail-party audio-visual sequences.

The audio-visual recordings are split into "segments", each segment lasts 0.3 seconds. At 25 frames/second this corresponds to approximately eight video frames. The initializa-

[4] http://perception.inrialpes.fr/CAVA_Dataset/Site/data.html#CTMS3

tion method described in Section 6.2 and the model selection method described in Section 6.3 are combined and applied to the first segment in order to find initial parameter values and to estimate the number of components (the number of audio-visual objects) to be used by the conjugate EM algorithm. Consequently, the parameters estimated for one segment are used to initialize the parameters for the next segment, while the number of components remains constant.

- Quasi-static scene. The meeting situation corresponds to the well-separated case which is referred to as **WS** in the previous Section. The initialization strategy performs well and the candidate configuration obtained by the initialization step is relatively close to the optimal one found by the EM algorithm described in detail in Section 5.1.3. In fact, the likelihood evolution reported in Figure 5.5 shows that convergence is reached in about 20 iterations of EM, which is comparable to the simulated **WS** case reported in Figure 5.2. The 3D position estimates are quite accurate, in particular the natural alignment of the speakers along the table is clearly seen in the XZ plane. Even though in practice, the data are not piecewise Gaussian and the outliers are not uniformly distributed, our method performs quite well, which illustrates its robustness when dealing with real-data distributions. Figure 5.6 shows sequential results obtained in this case. The speech sources are correctly detected even in the case when two persons are simultaneously active, the overall statistics on auditory activity are presented in Table 5.4.

- Simple dynamic scene. The tracking scenario was included to check whether the algorithm can cope with tracking an audio-visual object that moves in the scene without any special tuning. Figure 5.7 shows sequential results obtained in this case, Table 5.4 contains auditory activity detection statistics and Figure 5.9 shows the estimated trajectory in ambient space.

- Dynamic scene. The cocktail party situation corresponds to the partially occluded case which is referred to as **PS** in the previous Chapter. In this case, the audio-visual object locations vary over time, as well as their number. Nevertheless, we assume that these changes are rather slow. We did not attempt to tune our algorithm to the dynamic case. Hence, we use the same initialization strategy as in the quasi-static case which is briefly summarized above. Figure 5.8 shows the results obtained in this case, Table 5.4 summarizes auditory activity detection statistics and Figure 5.9 shows the estimated audio-visual object trajectories in ambient space.

Overall, the proposed method performs well on data collected in a natural environment. The initialization strategy and the model selection criterion proved to be robust to noise and to minor deviations from the Gaussian distribution assumption. It possesses the features of a global optimization method which enables to find initial parameter values that are close to the optimal ones. In all the examples, the parameter initialization and model selection were performed on the first audio-visual data segment. This certainly biases the overall results. Indeed, in all the cases, the initialization and model selection algorithms dealt with a case

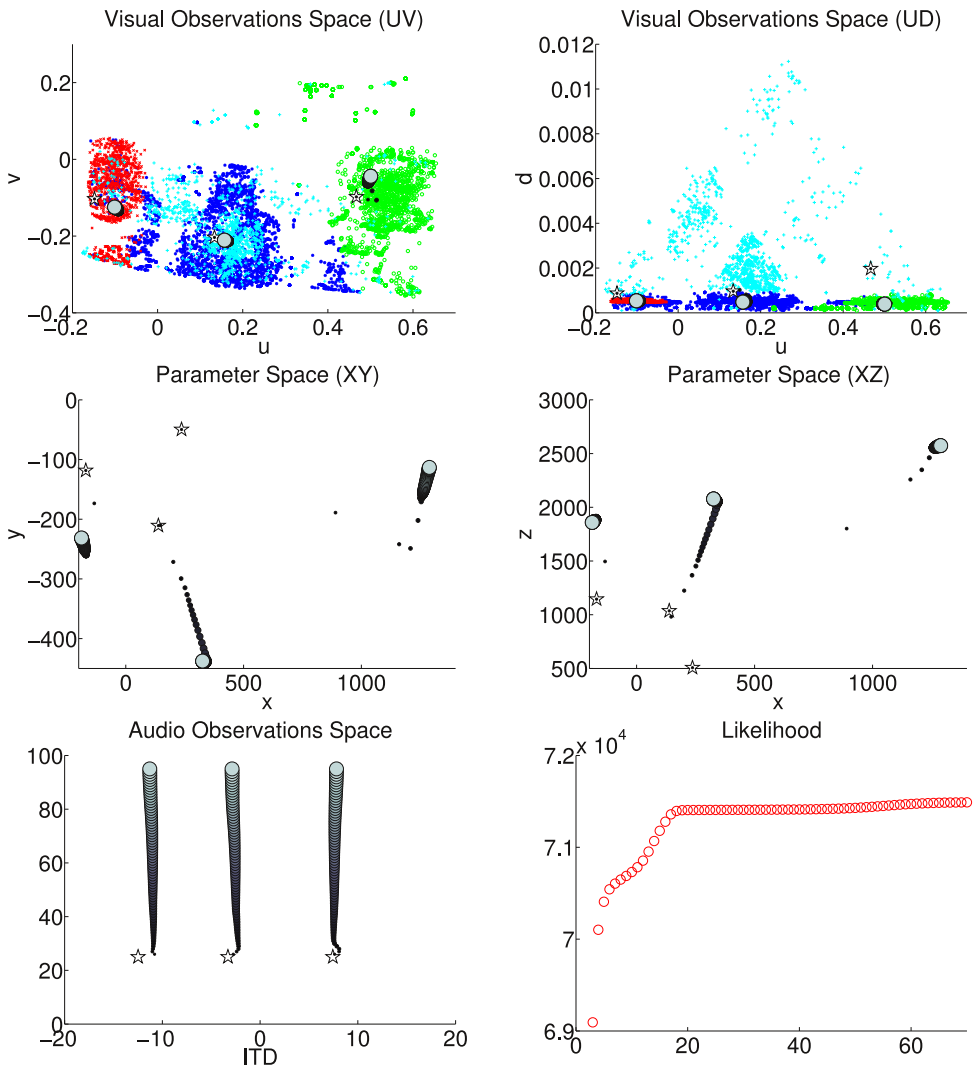
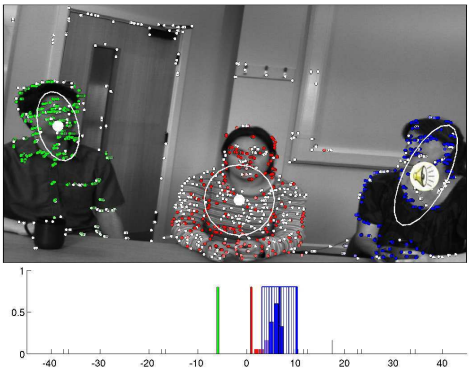
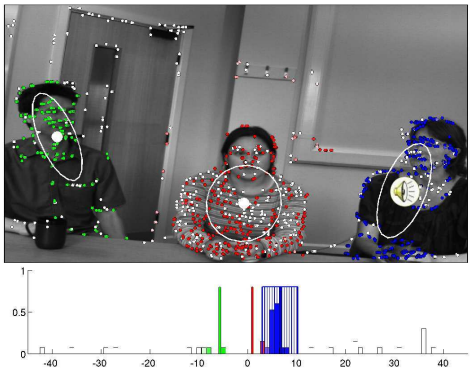


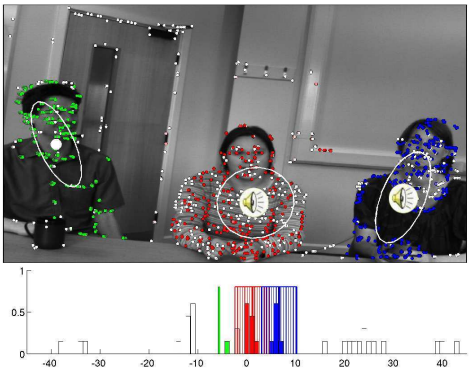
Figure 5.5: An example of applying the proposed EM algorithm to a time interval of 20 seconds of the meeting scenario. The results are shown in the visual and auditory observation spaces as well as in the parameter space. The initial parameter values are shown with three stars while the parameter evolution trajectories are shown with circles of increasing size. The final observation-to-cluster assignments are shown in colour: red, blue, and green for the three Gaussian components and light-blue for the outlier component. The log-likelihood curve (bottom-right) shows that the algorithm converged after 20 iterations.



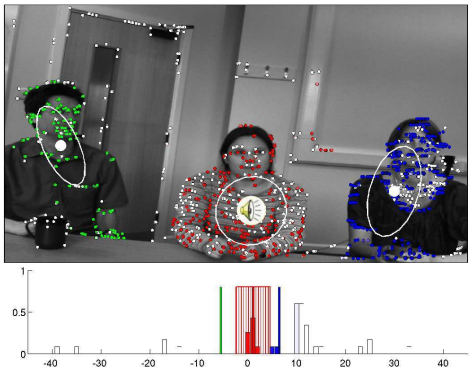
(a) frames 1001-1010



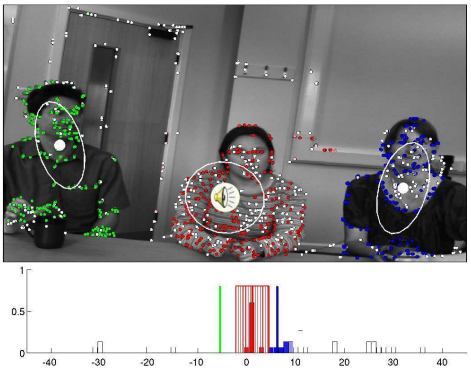
(b) frames 1011-1020



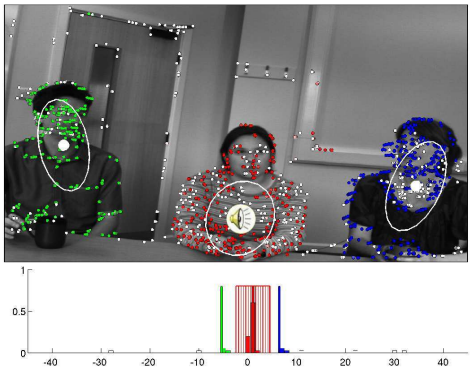
(c) frames 1021-1030



(d) frames 1031-1040

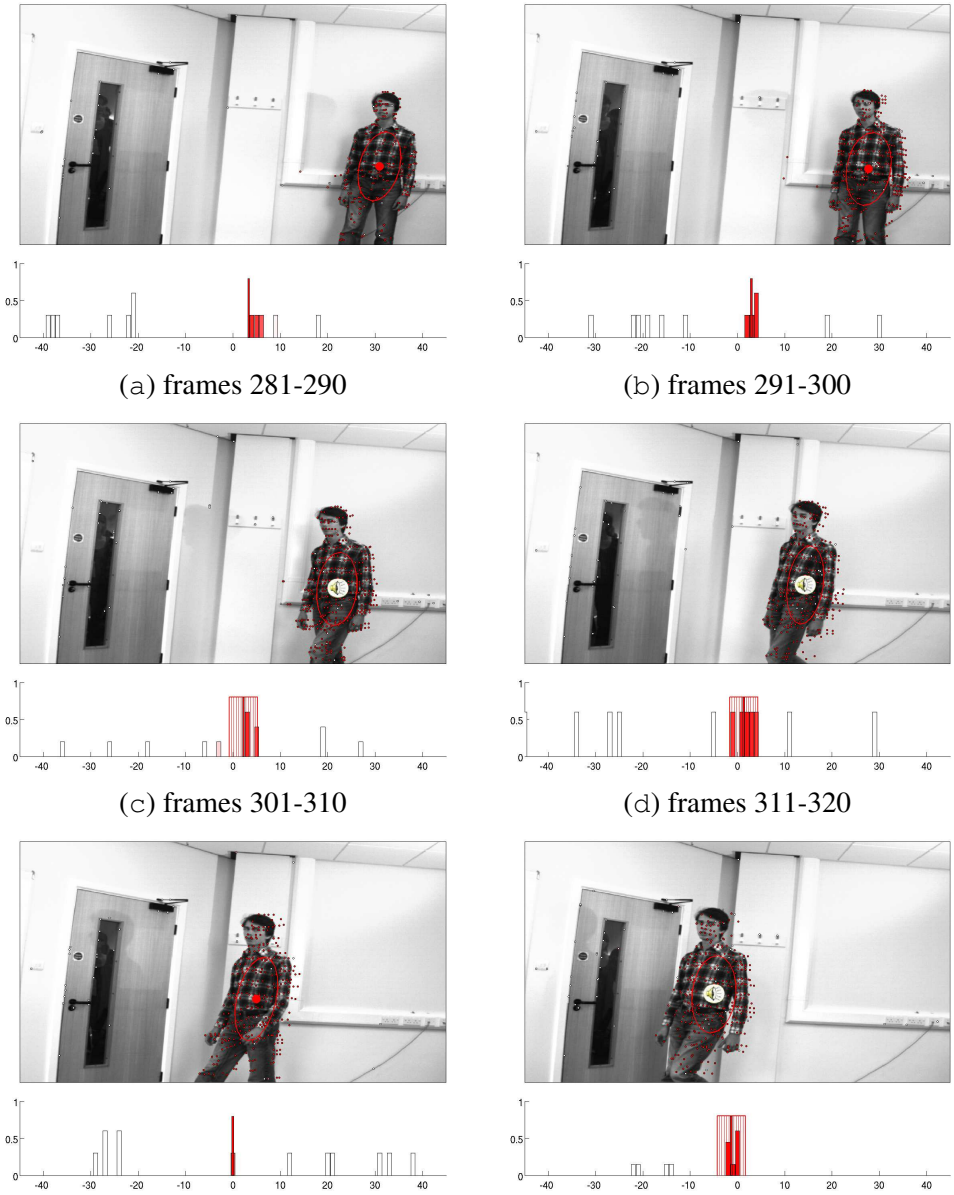


(e) frames 1041-1050



(f) frames 1051-1060

Figure 5.6: Results obtained in the case of the meeting scenario shown overlapped onto the left image. Sixty frames (1001 to 1060) were split into six segments. Parameter initialization and model selection were performed on the first segment (frames 1-10) and are not shown. The "visual" covariance matrices associated with the 3 Gaussian components are projected onto the image plane. The white dots correspond to the projected 3D locations estimated by the algorithm. The blue, green, and red colours encode the observation-to-cluster assignments and the active speaker is marked with a corresponding symbol. The algorithm correctly estimates speech sources, even in the case when two speakers are active.

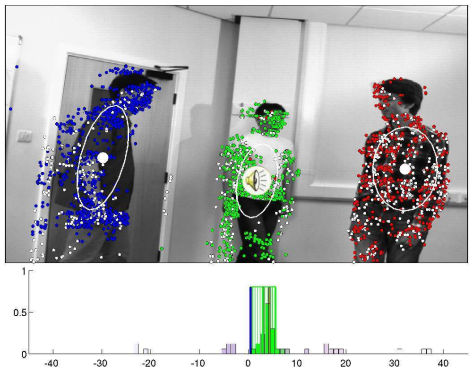


(a) frames 281-290 (b) frames 291-300

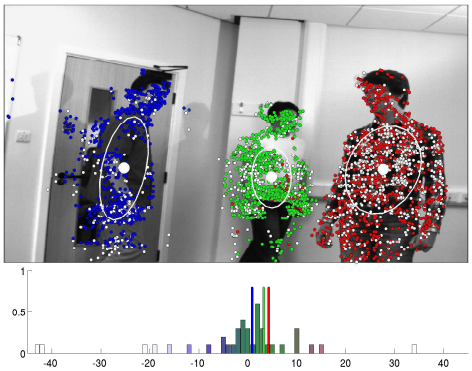
(c) frames 301-310 (d) frames 311-320

(e) frames 321-330 (f) frames 331-340

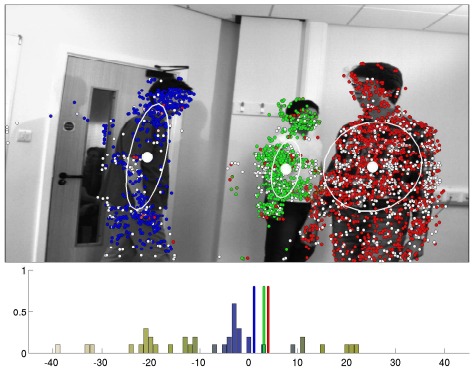
Figure 5.7: Results obtained in the case of the tracking scenario shown overlapped onto the left image. Sixty frames (281 to 340) were split into six segments. Parameter initialization and model selection were performed on the first segment (frames 1-10) and are not shown. The "visual" covariance matrix associated with the Gaussian component is projected onto the image plane. The red color encodes the observation-to-cluster assignments and the auditory activity is shown with the speaker symbol. The algorithm correctly tracks a moving target without any special tuning.



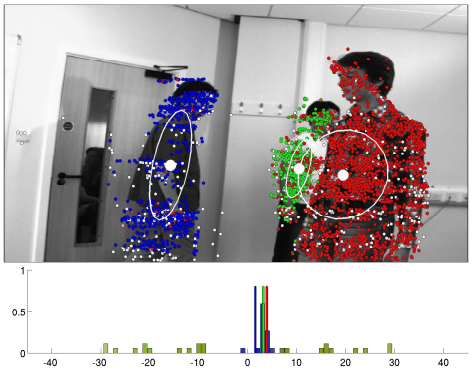
(a) frames 181-190



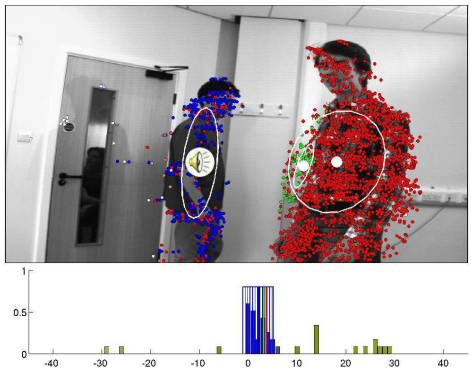
(b) frames 191-200



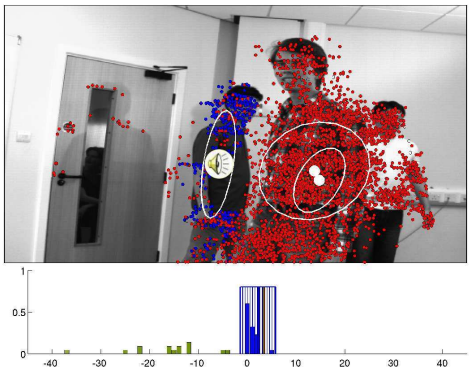
(c) frames 201-210



(d) frames 211-220



(e) frames 221-230

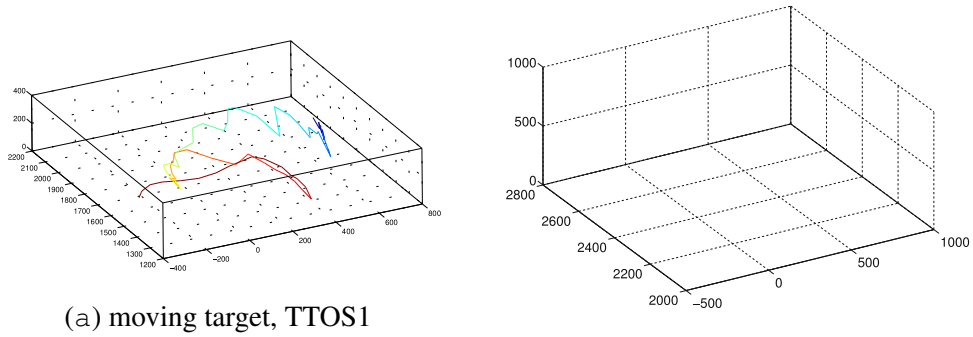


(f) frames 231-240

Figure 5.8: Results obtained in the case of the cocktail party scenario shown overlapped onto the left image. As in the previous case, sixty frames (181 to 240) were split into six segments. Parameter initialization and model selection were performed on the first segment (frames 1-10) and are not shown. As expected, well separated objects, (a)-(c), are correctly handled. While partial occlusion, (d)-(e) is also handled correctly, the algorithm fails to deal with a complete occlusion, (f).

	intervals	speaking	detected	E1	E2
M1	166	89	75	0.16	0.14
TTOS1	76	69	60	0.13	0.43
CTMS3	219	97	62	0.36	0.52

Table 5.4: Comparative results of the algorithm for the three scenarios: meeting (M1), moving target (TTOS1) and cocktail party (CTMS3). In each case the total number of frames, ground truth on the total number of auditory activity events to be detected, the total number of actually detected auditory activity and the probabilities of 'missed target' and 'false alarm' errors are given.



(a) moving target, TTOS1

(b) cocktail party, CTMS3

Figure 5.9: Estimated ambient space trajectories for the (a) moving target (TTOS1), and (b) cocktail party (CTMS3) scenarios. Motion is shown with colour gradient: from blue to red for a single target in TTOS1 and from darker to lighter colours in CTMS3. Dashed lines in the right image show the estimated trajectories after complete occlusion.

were the objects were well separated. One could rerun initialization and model selection on every data segment, at the cost of a less efficient procedure.

The conjugate clustering method automatically weights the auditory and visual modalities, in terms of precision and amount of observations, to infer the parameter values. We noticed that, in general, the visual data are considered by the algorithm as more reliable. This can be explained by the fact that, in practice, the auditory signals are contaminated with noise and reverberations. This typically smooths the histogram peaks in the ITD domain and adds false peaks, as can be seen in Figures 5.6-5.8. As reverberations are natural for most of the environments and sound sources, we added auditory cluster variances to model the local smoothing effect, as well as an outlier category to treat false peaks. In general, if the data is gathered using a small time interval, reverberations and noise have higher effect, the observations are scattered and auditory spatial localization is poor. At the same time, widening the time interval would result in sharper peaks for sound sources that are smoothed due to reverberations and dynamics of the scene, and hence the auditory temporal localization will be less accurate. Thus the auditory data are typically sparse both

in time and space. The temporal discontinuity of the auditory data together with the lack of resolution makes it less reliable than the visual data.

One advantage of the proposed conjugate clustering method is that while it performs audio-visual object ambient space position estimation, its auditory activity is detected simultaneously based on the number of auditory observations associated to the object at each time interval. We use maximum a posteriori (MAP) association principle based on calculated posterior probabilities α_{mn} and β_{kn} to assign each observation to an object or to an outlier class. The results of auditory activity detection are summarized in Table 5.4. For each scenario the first column contains the total number of time intervals being considered. The second one gives the total number of persons involved in auditory activity (it was counted for each time interval separately and then summed). The third column contains the total number of correctly detected speakers (which should ideally be equal to the previous value). And finally, the two last numbers are the probability of 'missed target' (i.e. the probability of a speaking person being marked as non-speaking) and the probability of 'false alarm' (i.e. the probability of a non-speaking person being marked as speaking).

As expected, the estimates contain less errors in case of well-separated objects. The 'false alarm'-type errors are typically generated by reverberations that tend to smooth out the histogram peaks in the auditory domain and generate false peaks, as mentioned earlier. Another reason is ambient sounds that originate from the same direction as the considered audio-visual object. The 'missed target' errors are most of the time due to the discretization effect (artificial splitting of the time line into intervals) and reverberations that sometimes produce stronger localization cues than the real signals. One way to eliminate these errors is to adapt the proposed ConjEM algorithm to the dynamic case, so that the instantaneous noise (such as reverberations) is smoothed out through considering larger time scales and the discretization effect is no longer present. Another possibility is to make more assumptions on auditory sources and include high-level detectors that consider only sounds of certain type, such as speech.

Although our multimodal clustering model has no built-in dynamic capability, as is the case with target-tracking methods based on the Kalman filter, the implemented algorithm performs quite well in the case of simple tracking tasks as well as more complex dynamic scenes, as shown in Figure 5.9. In particular, it is capable to deal with partial visual occlusions, as illustrated in the cocktail party scenario, Figure 5.8. In general the object position estimates are precise (within 10cm from the ground truth object position). However, Figure 5.9 shows that the algorithm admits high estimate fluctuations and can fail on complete visual occlusions, Figure 5.8(`f`).

5.4 Discussion

We proposed an efficient acceleration technique for the conjugate EM (ConjEM) algorithm from the KP family considered in the previous Chapter. Using the ideas underlying the classical EM algorithm we built the ConjEM algorithm to perform the multimodal clustering task, while keeping attractive convergence properties. The analysis of the conjugate

EM algorithm and, more specifically, of the optimization task arising in the M-step, revealed several possibilities to increase the convergence speed. We proposed to decompose the M-step into two procedures, namely the *Local Search* and *Choose* procedures, which allowed us to derive a number of acceleration strategies. We exhibited appealing properties of the target function which induced several implementations of these procedures resulting in a significantly improved convergence speed.

We used both simulated and real data to illustrate the performance of ConjEM on a non trivial audio-visual localization task. Simulated data experiments allowed us to assess the average method behaviour in various configurations and initialization settings compared to the other KP algorithms and single-modality EM algorithms. These experiments showed that while keeping the important stability and precision qualities found in the previous Chapter, we were able to significantly improve the convergence speed to match the other KP algorithms. They also illustrated the theoretical dependency between the precisions in observation and parameter spaces. Real data experiments then showed that the observed data precision was high enough to guarantee high precision in the parameter space.

One strong point of our approach is that it allows to detect object activity in each modality along with its parameters estimation. This feature was used when detecting auditory activity of several persons in the audio-visual localization task. The results obtained on real data from various scenarios show that the proposed model is able to perform robust detection in the case of well-separated objects without any special assumptions on sound sources. However, in the case of dynamic scene the error rates are increased. It is argued here that they are likely to be improved by incorporating the scene dynamics into model.

The strong points of the KP framework, such as extensibility to an arbitrary number of feature spaces and various clustering models (likelihood distributions) are inherited by the ConjEM algorithm. The main results, including *Local Search* and *Choose* acceleration strategies stay valid with minor changes. At the same time, one important feature of the ConjEM algorithm is that it can be easily extended to the dynamic case. In particular, adding Gaussian priors on parameters (i.e., priors, covariance matrices and objet locations) would not essentially change the formulae. For a large class of dynamics equations, the update expressions (5.8)-(5.11) for priors and variances will remain in closed form, whereas the function $Q_n^{(q)}(\boldsymbol{s})$ in (5.13) will receive an additional term $\log P(\boldsymbol{s})$. For instance, multimodal dynamic inference of parameter values for Brownian dynamics [van Kampen 2007] can be performed by means of the formulated model. Gaussian priors would add a quadratic term similar to the others in (5.13), that can be viewed as an 'observation' from the ambient space modality. Thus the optimization algorithm would not require any changes and would give an unbiased estimate. Various possibilities to adapt the ConjEM algorithm to the dynamic case are discussed in Chapter 7.

The major advantages of the proposed algorithm are summarized below:

- **Acceleration possibilities**: several efficient and theoretically well-founded acceleration strategies were proposed to improve the convergence speed of ConjEM;

- **Activity detection**: object activity in every modality is estimated along with its parameters, the detection was shown to be robust in various real data scenarios;
- **Inherent to the KP family**: all the strong points formulated for the KP family in general remain valid for the accelerated version of the ConjEM algorithm;
- **Extensibility to the dynamic case**: system dynamics can be included into the model so that the fast convergence property is kept;

The conjugate clustering framework together with the ConjEM algorithm offers a powerful tool to perform multimodal clustering that possesses most of the features that characterize the integration processes in mammals responsible for the creation of unified percepts that were outlined in the introduction in Chapter 1. The complex modality processing algorithms can be used to extract high-level features, the proposed framework keeps all the information without stripping parts that are not required for the integration. The principle of co-localization and co-incidence is used to bind the high-level features based on low-level localization cues. The modalities are weighted automatically based on the amount of information provided by each modality. The ConjEM algorithm can be reduced to single modality EM algorithms in the limiting cases. The state of sensory systems is encoded into the mappings $\mathcal{F}$ and $\mathcal{G}$, so that the algorithm occurs to be invariant to changes in sensory systems states (or, using the terminology of Chapter 3, inter-system calibration data).

CHAPTER 6

Algorithm Initialization and Model Selection

Sommaire

6.1 Multimodal Cluster Initialization and Model Selection 93
6.2 Initialization 95
6.2.1 EM Initialization for Conjugate Gaussian Mixture Models 96
6.2.2 The *Initialize* Procedure 97
6.2.3 Experimental Validation 98
6.3 Model Selection 103
6.3.1 The *Select* Procedure 107
6.3.2 Weak Consistency of Multimodal Information Criteria 108
6.3.3 Experimental Validation 112
6.4 Discussion 112

So far we have considered the task of multimodal clustering under the assumption that the number of objects as well as their initial parameter values were known. The algorithms proposed previously are then viewed as local optimization methods that, given a current point, converge to some stationary solution. In practice, however, the efficiency of such algorithms is highly dependent on the choice of that point.

In this Chapter we address the problem of how to choose initial parameter values for the multimodal clustering algorithms. The procedure *Initialize* based on predictive probability density function in the object space is proposed. We also develop a framework to compare conjugate mixture models using Bayesian information criterion (BIC). The initialization and model selection methods are verified on simulated and real data.

6.1 Multimodal Cluster Initialization and Model Selection

Multimodal approaches that generate observations from different sensors in different spaces are more and more common in real-world applications. The need for efficient algorithms that are capable of consistent treatment of several modalities increases. Multimodal clustering algorithms are proposed in Chapters 4 and 5 within the framework of Gaussian

mixture models. They are detailed for two modalities in the context of audio-visual object detection but their extension to more than one modality is straightforward. The associated conjugate Expectation-Maximization (ConjEM) performs several unimodal clustering tasks that are coupled using explicit relationships between a common object space and each one of the observation spaces. Objects are shared by all modalities. Each object is assumed to be responsible for a possibly different number of observations in each modality. The clustering task it therefore recast as that of recovering the observations assignments to the different objects. The parameters of the modality-specific Gaussian mixtures are conditioned by a common set of object-space parameters through explicit object-space-to-observation-space mappings, one mapping for each modality. It follows that each modality-specific mixture shares the same number of components corresponding to the number of objects.

While the E-step of ConjEM is rather standard, the M-step implies non-linear optimization with respect to the model parameters. In Chapter 5 we proved that if the object-to-sensor mappings and their first derivatives are Lipschitz continous functions (which they are in the audio-visual example), the gradient of the expected complete-data log-likelihood function is Lipschitz continous as well. Consequently, the recently proposed optimization algorithm specifically designed to solve Lipschitzian global optimization problems can be used within the M-step of ConjEM [Zhigljavsky 2008]. This implies that the ConjEM algorithm has guaranteed convergence properties.

However, like any other EM procedure, ConjEM suffers from two limitations: (i) the number of components must be determined in advance, and (ii) the parameters must be properly initialized. The first one of these problems, referred to as model selection, is critical in our case because it means that one needs to know in advance the number of multimodal objects under consideration (e.g. the number of audio-visual objects composing a scene, like the number of speaking persons in a complex meeting scenario). The second problem is also very important because without a proper initialization, ConjEM is likely to be trapped in a local maximum.

This Chapter contains two original contributions. First, we propose to extend information criteria for model selection for multimodal data and show that such criteria provide consistent estimators of the number of objects. To our knowledge, there has been no procedure so far that properly selects the model dimensionality for multimodal case in a consistent manner. Standard results on information criteria are shown for identically distributed data, which is typically not the case in the multimodal setting. Second, we introduce the initialization algorithm based on predictive probability density function that has two benefits: (i) it provides parameters that are close to local maxima, so that the ConjEM algorithm converges faster; (ii) the algorithm samples the object space in a way global optimization methods do, which increases the chance to find a global optimal solution with the ConjEM algorithm. With these two contributions we are able to derive an appropriate information criterion with a BIC-like penalty and illustrate the performance of the conjugate EM algorithm on the task of detecting and localizing audio-visual objects.

In Section 6.2 we present the method to initialize multimodal clusters. The model se-

lection criteria are introduced and investigated further in Section 6.3, a consistency result is given. We demonstrate the performance of model selection and initialization on simulated data and realistic scenarios. A discussion of the results and future work concludes the Chapter.

6.2 Initialization

We consider the maximum likelihood (ML) estimation problem formulated in Chapter 4. The aim is to perform optimization (4.20) of the log-likelihood function $\mathcal{L}$ given by (4.18). The Expectation-Maximization (EM) algorithm is a popular technique to compute ML estimates for such problems with incomplete data. Though the solution depends a lot on its starting position. The EM algorithm for Gaussian mixtures can be viewed as a local optimization method similar to a variable metric method [Ma 2000], so it can get stuck in a local maximum point.

At the same time, finding the parameters that maximize the likelihod is important for two reasons. Firstly, one would like to obtain sensible values for the model. Secondly, the ML estimates are often used in model selection procedures.

The problem of choosing initial values for the EM algorithm for multivariate Gaussian mixtures received considerable attention during the past years due to its importance. Several methods were proposed.

Random initialization is one of the most popular techniques employed for the EM initialization [Meila 2001, Biernacki 2003]. It consists in initializing the parameters or some part of them at random. Sampling can be performed based on some prior distribution or it can also be data-dependent. This way random initialization resembles search algorithms in global optimization [Zhigljavsky 2008].

Bootstrap initialization is a technique that includes either several iterations of other algorithms (K-means, K-medoids, CEM, SEM) [Biernacki 2003], or a sequence of solutions to relaxed problems [Ueda 1998] as the initial solution.

Our problem is different from the ones considered in the above cited papers. Firstly, we work with conjugate mixture models that perform simultaneous clustering in several physically different observation spaces. This implies that the initialization algorithm should be multimodal as well. Secondly, certain restrictions apply to the object configurations and we would like to come up with a more efficient technique than the general ones.

The method we propose combines the features of both, random data-driven and bootstrap initializations. We use data points to compute the *predictive density* that we sample subsequently to select the initial parameters values and perform a short run of the ConjEM algorithm.

6.2.1 EM Initialization for Conjugate Gaussian Mixture Models

Consider conjugate Gaussian mixture models with outliers introduced in Section 4.2. They are governed by the parameters $\boldsymbol{\theta}$ given by

$$\boldsymbol{\theta} = \{\pi_1, \dots, \pi_{N+1}, \lambda_1, \dots, \lambda_{N+1}, \boldsymbol{s}_1, \dots, \boldsymbol{s}_N, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_N\}, \tag{6.1}$$

of which the tying parameters $\boldsymbol{s}_n$ are inferred using both modalities and the rest of $\boldsymbol{\theta}$ is modality-specific parameters that govern cluster shape in auditory and visual spaces. We would like sometimes to consider these groups of parameters separately, so we denote

$$\boldsymbol{\theta}_{\mathrm{S}} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n, \dots, \boldsymbol{s}_N\}, \tag{6.2}$$

$$\boldsymbol{\theta}_{\mathrm{F}} = \{\pi_1, \dots, \pi_N, \pi_{N+1}, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N\}, \tag{6.3}$$

$$\boldsymbol{\theta}_{\mathrm{G}} = \{\lambda_1, \dots, \lambda_N, \lambda_{N+1}, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_N\}. \tag{6.4}$$

Then the initialization task can be formulated as follows: given the observation sets $\mathbf{f} = \{\boldsymbol{f}_1, \dots, \boldsymbol{f}_m, \dots, \boldsymbol{f}_M\}$ and $\mathbf{g} = \{\boldsymbol{g}_1, \dots, \boldsymbol{g}_k, \dots, \boldsymbol{g}_K\}$ initialize $\boldsymbol{\theta}$, so that for every object n located at $\boldsymbol{s}_n$ its cluster shapes $\{\pi_n, \boldsymbol{\Sigma}_n\}$ and $\{\lambda_n, \boldsymbol{\Gamma}_n\}$ align well with the observed visual and auditory data respectively.

This task can be viewed as simultaneous probability density estimation through parametrized density families. That is, given $\rho_{\mathrm{F}}(\boldsymbol{f})$ and $\rho_{\mathrm{G}}(\boldsymbol{g})$ one has to find $\hat{\rho}_{\mathrm{F}}(\boldsymbol{f};\ \boldsymbol{\theta}_{\mathrm{S}}, \boldsymbol{\theta}_{\mathrm{F}})$ and $\hat{\rho}_{\mathrm{G}}(\boldsymbol{g};\ \boldsymbol{\theta}_{\mathrm{S}}, \boldsymbol{\theta}_{\mathrm{G}})$ that correspond to the observed densities in certain sense.

Our approach is based on iterative local approximations of $\rho_{\mathrm{F}}(\boldsymbol{f})$ and $\rho_{\mathrm{G}}(\boldsymbol{g})$. At iteration n classes $1, \dots, n-1$ are supposed to be initialized, therefore we sample object location $\boldsymbol{s}_n$ from the *predictive distribution* $\rho_{\mathrm{S}}^{(n)}(\boldsymbol{s})$ and choose optimal $\{\pi_n, \boldsymbol{\Sigma}_n, \lambda_n, \boldsymbol{\Gamma}_n\}$ to construct $\hat{\rho}_{\mathrm{F}}^{(n)}(\boldsymbol{f};\ \boldsymbol{\theta}_{\mathrm{S}}, \boldsymbol{\theta}_{\mathrm{F}})$ and $\hat{\rho}_{\mathrm{G}}^{(n)}(\boldsymbol{g};\ \boldsymbol{\theta}_{\mathrm{S}}, \boldsymbol{\theta}_{\mathrm{G}})$. The predictive distribution is calculated through kernel estimators of the observation space distributions $\rho_{\mathrm{F}}(\boldsymbol{f})$ and $\rho_{\mathrm{G}}(\boldsymbol{g})$. Every observation $\boldsymbol{f}_m$ and $\boldsymbol{g}_k$ is assigned a weight $\alpha_m^{(n)}$ and $\beta_k^{(n)}$ respectively, that corresponds to the posterior probability of an observation to belong to the outlier class:

$$\alpha_m^{(n)} = \frac{\pi_{n+1}\mathcal{U}(\boldsymbol{f}_m;\ V)}{\sum\limits_{i=1}^{n-1} \pi_i \mathcal{N}(\boldsymbol{f}_m; \mathcal{F}(\boldsymbol{s}_i), \boldsymbol{\Sigma}_i) + \pi_{n+1}\mathcal{U}(\boldsymbol{f}_m;\ V)}, \tag{6.5}$$

$$\text{and} \quad \beta_k^{(n)} = \frac{\lambda_{n+1}\mathcal{U}(\boldsymbol{g}_k;\ U)}{\sum\limits_{i=1}^{n-1} \lambda_i \mathcal{N}(\boldsymbol{g}_k; \mathcal{G}(\boldsymbol{s}_i), \boldsymbol{\Gamma}_i) + \lambda_{n+1}\mathcal{U}(\boldsymbol{f}_m;\ V)}. \tag{6.6}$$

The kernel estimators are then computed by

$$\tilde{\rho}_{\mathrm{F}}(\boldsymbol{f}) = \frac{1}{\sum\limits_{m=1}^{M} \alpha_m^{(n)}} \sum_{m=1}^{M} \alpha_m^{(n)} \mathcal{N}(\boldsymbol{f};\ \boldsymbol{f}_m, \boldsymbol{\Lambda}), \tag{6.7}$$

$$\text{and} \quad \tilde{\rho}_{\mathrm{G}}(\boldsymbol{g}) = \frac{1}{\sum\limits_{k=1}^{K} \beta_k^{(n)}} \sum_{k=1}^{K} \beta_k^{(n)} \mathcal{N}(\boldsymbol{g};\ \boldsymbol{g}_k, \boldsymbol{\Upsilon}), \tag{6.8}$$

where the choice of the bandwidths $\mathbf{\Lambda}$ and $\mathbf{\Upsilon}$ is based on the properties of the feature detector algorithms.

We therefore sample L particles $\{\boldsymbol{s}_l^{(n)}\}_{l=1}^L$ in the parameter space $\mathbb{S}$. They are obtained through drawing $\tilde{\boldsymbol{f}}_l \sim \tilde{\rho}_{\mathrm{F}}(\boldsymbol{f})$ or $\tilde{\boldsymbol{g}}_l \sim \tilde{\rho}_{\mathrm{G}}(\boldsymbol{g})$, and subsequently applying the corresponding inverse mapping, $\mathcal{F}^{-1}$ or $\mathcal{G}^{-1}$. Each particle $\boldsymbol{s}_l^{(n)}$ is then assigned a weight

$$\gamma_l^{(n)} = \tilde{\rho}_{\mathrm{F}}\left(\mathcal{F}(\boldsymbol{s}_l^{(n)})\right)\tilde{\rho}_{\mathrm{G}}\left(\mathcal{G}(\boldsymbol{s}_l^{(n)})\right), \quad l = 1, \ldots, L. \tag{6.9}$$

Without loss of generality, we suppose the weights to be normalized, so that $\sum_{l=1}^{L} \gamma_l^{(n)} = 1$. The object location $\boldsymbol{s}_n$ is sampled from a discrete probability distribution defined by $\left\{(\boldsymbol{s}_l^{(n)}, \gamma_l^{(n)})\right\}_{l=1}^L$.

After having obtained the new estimate $\boldsymbol{s}_n$, we need to initialize the associated modality-specific parameters π_n, $\mathbf{\Sigma}_n$, λ_n and $\mathbf{\Gamma}_n$. We use the standard empirical covariance matrix formulas for $\mathbf{\Sigma}_n$ and $\mathbf{\Gamma}_n$:

$$\mathbf{\Sigma}_n = \frac{1}{\sum_{m=1}^{M} \alpha_m^{(n)}} \sum_{m=1}^{M} \alpha_m^{(n)} (\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n))(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n))^\top, \tag{6.10}$$

$$\text{and} \quad \mathbf{\Gamma}_n = \frac{1}{\sum_{k=1}^{K} \beta_k^{(n)}} \sum_{k=1}^{K} \beta_k^{(n)} (\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n))(\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s}_n))^\top, \tag{6.11}$$

and the priors are set to be equal

$$\pi_1 = \ldots = \pi_n = \pi_{n+1} = 1/(n+1), \tag{6.12}$$

$$\text{and} \quad \lambda_1 = \ldots = \lambda_n = \lambda_{n+1} = 1/(n+1). \tag{6.13}$$

Finally, we run several iterations of the ConjEM algorithm with $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ being fixed to bootstrap and improve the computed parameter values $\boldsymbol{\theta}$.

It is possible to choose covariance matrices and priors that provide a better local fit to the data than the simple empirical formulas (6.10)-(6.13). For example, one can use the fitted local likelihood (FLL) technique, presented in [Katkovnik 2008]. It proposes a method based on hypothesis testing to choose the appropriate scale for parameters estimation. This approach is likely to give even better initial values, but we choose here to use simpler formulas for a fixed scale that provide satisfactory results.

6.2.2 The *Initialize* Procedure

The overall procedure for parameters $\boldsymbol{\theta}$ initialization is outlined below:

1. Set $\alpha_m^{(1)} = 1, \ \ \forall m = 1, \ldots, M$ and $\beta_k^{(1)} = 1, \ \ \forall k = 1, \ldots, K$;

2. For $n = 1, \ldots, N$ do
   (a) Sample particles $\boldsymbol{s}_l^{(n)}$ from $\tilde{\rho}_{\mathrm{F}}(\boldsymbol{f})$ and $\tilde{\rho}_{\mathrm{G}}(\boldsymbol{g})$ given by (6.7) and (6.8);
   (b) Compute particle weights $\gamma_l^{(n)}$ through (6.9);
   (c) Sample $\boldsymbol{s}_n$ from the discrete distribution $\left\{(\boldsymbol{s}_l^{(n)}, \gamma_l^{(n)})\right\}_{l=1}^{L}$;
   (d) Compute covariance matrices $\boldsymbol{\Sigma}_n$ and $\boldsymbol{\Gamma}_n$ using (6.10) and (6.11);
   (e) Reset the priors $\pi_1, \ldots, \pi_{n+1}$ and $\lambda_1, \ldots, \lambda_{n+1}$ using (6.12) and (6.13);
   (f) Bootstrap current parameter values $\boldsymbol{\theta}$ running several iterations of the ConjEM algorithm with fixed or slowly varying $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$;
   (g) Compute weights $\alpha_m^{(n+1)}$ and $\beta_k^{(n+1)}$ using (6.5) and (6.6);

6.2.3 Experimental Validation

The proposed initialization method in verified on the audio-visual (AV) data, letting, as before, $\mathbb{F}$, $\mathbb{G}$ and $\mathbb{S}$ denote the visual, auditory and ambient spaces respectively. The multimodal data consists of M visual observations $\mathbf{f}$ and of K auditory observations $\boldsymbol{g}$. Each object is described by a 3D parameter vector $\boldsymbol{s}_n = (x_n, y_n, z_n)^\top$. As previously, we suppose the AV device to be calibrated and use the projective visual feature space mapping $\mathcal{F}$ defined by

$$\mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{1}{z}\right)^\top \quad \text{and} \quad \mathcal{F}^{-1}(\boldsymbol{f}) = \left(\frac{u}{d}, \frac{v}{d}, \frac{1}{d}\right)^\top, \tag{6.14}$$

and the auditory feature space mapping $\mathcal{G}$ defined by

$$g = \mathcal{G}(\boldsymbol{s}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big). \tag{6.15}$$

The choice of the kernel estimator matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Upsilon}$ is based entirely on detector properties and in our case they are taken to be

$$\boldsymbol{\Lambda} = \begin{pmatrix} 10^{-4} & 0 & 0 \\ 0 & 10^{-4} & 0 \\ 0 & 0 & 10^{-10} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Upsilon} = 0.1. \tag{6.16}$$

6.2.3.1 Experiments with Simulated Data

We consider two simulated configurations described in detail in Chapter 4: well separated (**WS**) and poorly separated (**PS**). In these experiments we used the *Initialize* procedure with the inhibited update of position parameters $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. The step by step initialization results for **WS** and **PS** configurations are shown in Figures 6.1 and 6.2 respectively. Images in the left column show the probability densities $\tilde{\rho}_{\mathrm{F}}(\boldsymbol{f})$ (upper part) and $\tilde{\rho}_{\mathrm{G}}(\boldsymbol{g})$ (lower part) in the corresponding feature spaces. The visual space density is colour-coded, blue colour

corresponds to lower values, red colour – to higher values. Below the distribution density $\tilde{\rho}_{\mathrm{G}}(\boldsymbol{g})$ is plotted in the ITD space.

The obtained initialization results are depicted in the right column. Clusters are shown with coloured ellipsoids (visual covariance matrix projections onto the (u, d) coordinates) centered in the visual space mean values (upper part) and with rectangles of the corresponding colour (auditory variance) centered in the auditory space mean values (lower part).

We note that the results obtained for the WS and PS configurations correspond to the intermediate initialization (II) sampling setting considered in Chapters 4 and 5. Thus we conclude that the proposed strategy is relevant to the task of multimodal initialization, even in the case of poorly separated (PS) objects. Of course, we relied heavily on the assumptions concerning object configurations and detector properties. Firstly, the objects are supposed to be sufficiently separated (at least as in the PS case). Secondly, the detector properties are supposed to be known, so that the kernel estimator matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Upsilon}$ can be chosen respectively. In return, the formulated method occurred to be more efficient than simple random algorithms. The 'no free lunch' principle can be stated: the quality of the initialization results depends on the validity of the assumptions.

One advantage of the proposed method is that it is able to treat correctly situations with partially observed objects. The initialization process, for the example with missing data given in Figure 4.6, is depicted in Figure 6.3. A cluster that is almost invisible in one modality is compensated with high particle weights from another modality and the overall predictive density for the cluster occurs to be strong. Thus the initialization strategy complies with the multisensory enhancement principle discussed in the introduction in Chapter 1.

Another advantage is that the *Initialize* procedure is naturally integrated into the ConjEM framework as an extension that gives the EM – typically local optimization algorithm, – the features of a global optimization procedure. By making restrictive assumptions mentioned before we narrow down the parameter search domain increasing the algorithm efficiency.

The initialization strategy that we propose has an iterative nature, so provided the object tying parameters are forced to have small deviation from their current locations, one can track dynamic changes in scene formation. We use this observation to construct the multimodal multiobject tracking method in Chapter 7. The assumption of slow system evolution there plays the role of the inhibiting force for the object tying parameters.

In general, this kind of parameter space sampling can be considered as a marked point process with leading measure given by the predictive density. Such a point process can be used in jump-diffusion [Grenander 1994, Jacobsen 2006] optimization schemes, which proves to be useful when considering dynamic tracking tasks. We further discuss this in Chapter 7.

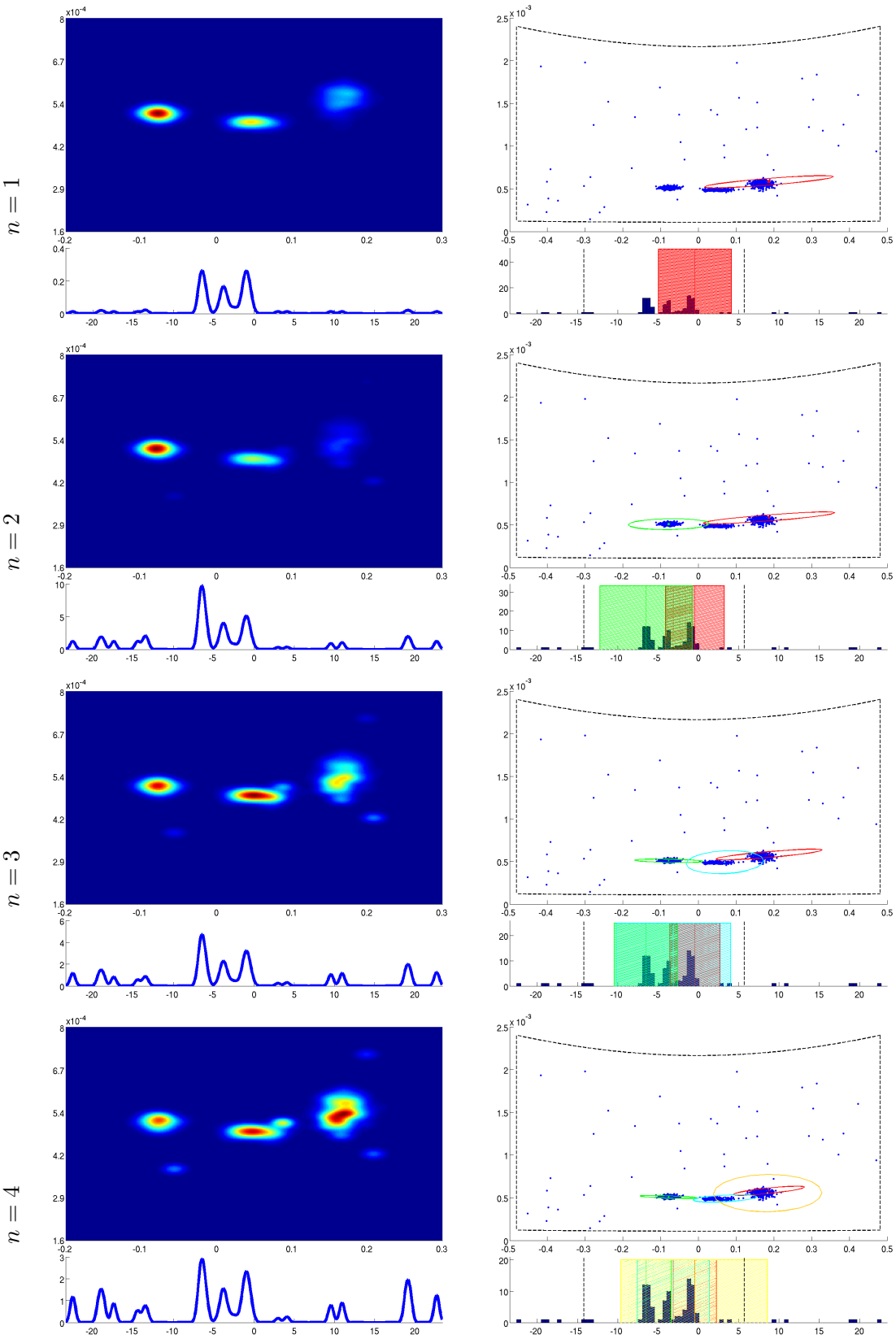


Figure 6.1: Iterations $n = 1, \ldots, 4$ of the *Initialize* procedure for the well separated (**WS**) object configuration. The left column shows the predictive distributions $\tilde{\rho}_{\mathrm{F}}(\boldsymbol{f})$ and $\tilde{\rho}_{\mathrm{G}}(\boldsymbol{f})$ in the corresponding feature spaces. The initialization result obtained using the densities on the left are shown on the right.

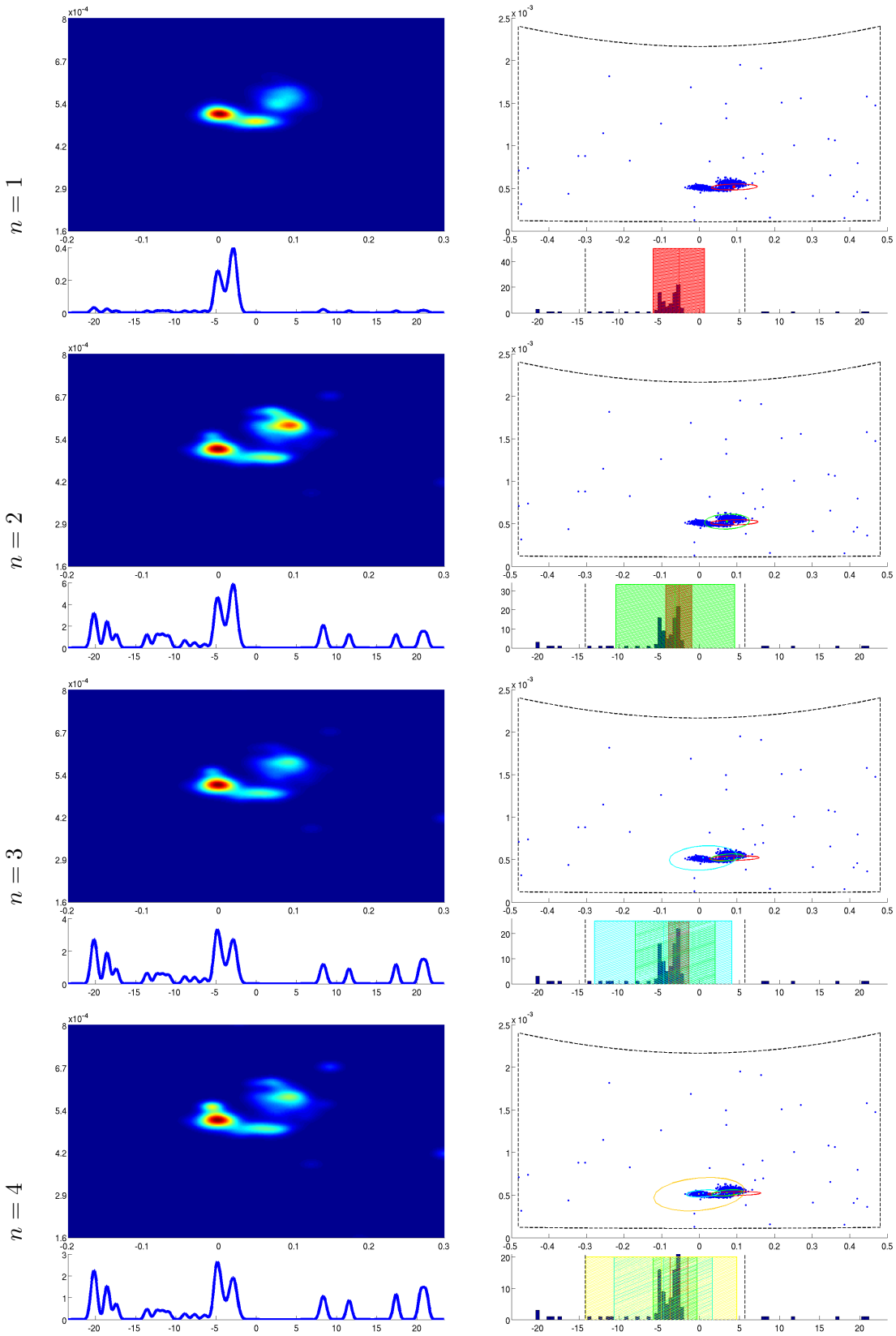


Figure 6.2: Iterations $n = 1, \ldots, 4$ of the *Initialize* procedure for the poorly separated (**PS**) object configuration. The left column contains the predictive distributions $\tilde{\rho}_{\mathrm{F}}(\boldsymbol{f})$ and $\tilde{\rho}_{\mathrm{G}}(\boldsymbol{f})$ in the corresponding feature spaces. The initialization result obtained using the densities on the left are shown on the right.

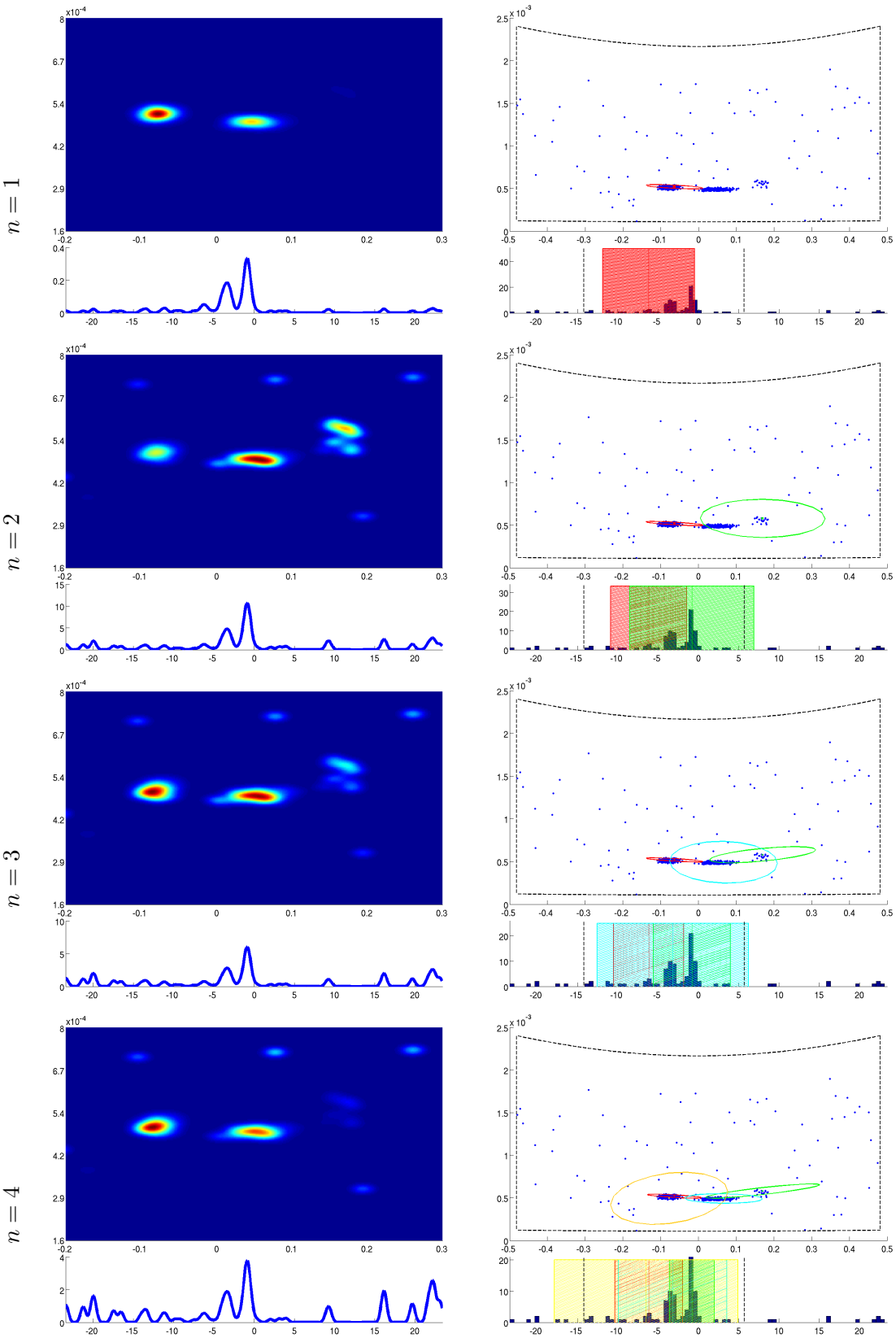


Figure 6.3: Iterations $n = 1, \ldots, 4$ of the *Initialize* procedure for partially observed objects show the multisensory enhancement behaviour. Left column contains predictive distributions $\tilde{\rho}_{\mathrm{F}}(\boldsymbol{f})$ and $\tilde{\rho}_{\mathrm{G}}(\boldsymbol{f})$ in the corresponding feature spaces. The initialization result obtained using the densities on the left are shown on the right.

6.2.3.2 Experiments with Real Data

We demonstrate the performance of the *Initialize* procedure on different time intervals of the cocktail party (CTMS3) scenario from the CAVA database described in detail in Chapter 2. This scenario is the most interesting benchmark as it contains various objects configurations including well separated, poorly separated (partially occluded) and completely occluded objects.

The initialization procedure for the well separated AV objects configuration (frames 181–190 of the scenario) is illustrated in Figure 6.4. The results for iterations $n = 1, \ldots, 6$ are shown projected onto the left image plane. Visual covariance matrices are represented by ellipses drawn around projected cluster mean values. The colours encode the observation-to-cluster assignments.

The performance on a more complicated configuration which corresponds to the simulated poorly separated case is demonstrated in Figure 6.5. The same objects as in the previous case are partially occluded (frames 211–220 of the scenario). One by one the initialization procedure adds 6 clusters using the predictive densities.

These results show that the *Initialize* procedure is quite robust to changes in object configurations and is capable of providing meaningful starting values for the ConjEM algorithm that are close to the optimal ones. At the same time we note that in real scenarios the performance is dominated by visual data as soon as it is richer and more precise.

6.3 Model Selection

We address the problem of consistent model selection in the conjugate mixture model framework introduced in Section 4.2. Conjugate Gaussian mixture models with outliers are governed by N groups of parameters $\{\boldsymbol{\theta}_n\}_{n=1}^N$ that define conjugate clusters in the two modality spaces $\mathbb{F}$ and $\mathbb{G}$. The problem of model selection in this case consists in estimating the number of clusters N.

The model selection is a well known but difficult problem in statistics. Numerous approaches aiming to solve this task are based on penalized likelihood maximization: Akaike Information Criterion (AIC) [Akaike 1973], Bayesian Information Criterion (BIC) [Schwarz 1978], Minimum Description Length (MDL) [Rissanen 1978], Normalized Entropy Criterion (NEC) [Celeux 1996], Integrated Complete Likelihood (ICL) [Biernacki 2000] etc. Though they are usually developed for the case of independent and identically distributed (i.i.d) observations. Thus we need to adopt the developed theory to our case of multiple modalities and observations sets lying in different spaces $\mathbb{F}$ and $\mathbb{G}$ and being bound through common parameters.

The aim is to generalize the existing approaches to prove consistency in the case of conjugate mixture models. We start with introducing some basic notations. Let $\mathfrak{F}$ and $\mathfrak{G}$ denote respectively the sets of probability densities on $\mathbb{F} \subseteq \mathbb{R}^r$ and $\mathbb{G} \subseteq \mathbb{R}^p$ with respect to

$n = 1$ $n = 2$

$n = 3$ $n = 4$

$n = 5$ $n = 6$

Figure 6.4: Iterations $n = 1, \ldots, 6$ of the *Initialize* procedure for frames 181–190 of the cocktail party scenario (well separated objects). The 'visual' covariance matrices associated with the Gaussian components are projected onto the image plane. They are shown with ellipses around the projected cluster mean values. The colours encode the observation-to-cluster assignments.

$n = 1$ $n = 2$

$n = 3$ $n = 4$

$n = 5$ $n = 6$

Figure 6.5: Iterations $n = 1, \ldots, 6$ of the *Initialize* procedure for frames 211–220 of the cocktail party scenario (partially occluded objects). The 'visual' covariance matrices associated with the Gaussian components are projected onto the image plane. They are shown with ellipses around the projected cluster mean values. The colours encode the observation-to-cluster assignments.

some positive measures ν_{F} and ν_{G} and some positive integers r and p. Let $\mathbf{F}$ and $\mathbf{G}$ be two sets of i.i.d. random variables with respective densities $P_0^{\mathbb{F}} \in \mathfrak{F}$ and $P_0^{\mathbb{G}} \in \mathfrak{G}$:

$$\mathbf{F} = \{\boldsymbol{F}_1, \ldots, \boldsymbol{F}_m, \ldots, \boldsymbol{F}_M\}, \quad \boldsymbol{F}_m \sim P_0^{\mathbb{F}} \mathrm{d}\nu_{\mathrm{F}} \quad \forall m = 1, \ldots, M \tag{6.17}$$

$$\text{and} \quad \mathbf{G} = \{\boldsymbol{G}_1, \ldots, \boldsymbol{G}_k, \ldots, \boldsymbol{G}_K\}, \quad \boldsymbol{G}_k \sim P_0^{\mathbb{G}} \mathrm{d}\nu_{\mathrm{G}} \quad \forall k = 1, \ldots, K \tag{6.18}$$

The densities $P_0^{\mathbb{F}}$ and $P_0^{\mathbb{G}}$ correspond to the *true* models for the $\boldsymbol{F}_m$'s and $\boldsymbol{G}_k$'s. They are not generally mixture densities. For any $P^{\mathbb{F}} \in \mathfrak{F}$ and $P^{\mathbb{G}} \in \mathfrak{G}$, let $\mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}})$ be the log-likelihood of all the observed data:

$$\mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) = \sum_{m=1}^{M} \log P^{\mathbb{F}}(\boldsymbol{f}_m) + \sum_{k=1}^{K} \log P^{\mathbb{G}}(\boldsymbol{g}_k) = \mathcal{L}_M(P^{\mathbb{F}}) + \mathcal{L}_K(P^{\mathbb{G}}), \tag{6.19}$$

where $\mathcal{L}_M(P^{\mathbb{F}})$ and $\mathcal{L}_K(P^{\mathbb{G}})$ denote the log-likelihoods for observations $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\}$ and $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k, \ldots, \boldsymbol{g}_K\}$ respectively. Note that it is the sum of two terms, that correspond to data from different spaces in (6.19), that prevents us to use standard results on information criteria that are derived for log-likelihoods of i.i.d data.

In practice, the general unknown $P_0^{\mathbb{F}}$ and $P_0^{\mathbb{G}}$ are often approximated by parametric densities. In our case these parametric densities are families of mixtures (4.8) and (4.9):

$$\mathfrak{M}^{\mathbb{F}} = \bigcup_{N=1}^{\infty} \mathfrak{M}_N^{\mathbb{F}}, \tag{6.20}$$

$$\text{and} \quad \mathfrak{M}^{\mathbb{G}} = \bigcup_{N=1}^{\infty} \mathfrak{M}_N^{\mathbb{G}}. \tag{6.21}$$

We denoted $\mathfrak{M}_N^{\mathbb{F}}$ and $\mathfrak{M}_N^{\mathbb{G}}$ the sets of all N-mixtures with outliers

$$\mathfrak{M}_N^{\mathbb{F}} = \left\{ P^{\mathbb{F}} = \sum_{n=1}^{N} \pi_n \rho^{\mathbb{F}}(\boldsymbol{\theta}_n) + \pi_{N+1}\mathcal{U},\ \boldsymbol{\theta}_n \in \Theta,\ 0 \le \pi_n \le 1,\ \sum_{n=1}^{N+1} \pi_n = 1 \right\}, \tag{6.22}$$

$$\text{and} \quad \mathfrak{M}_N^{\mathbb{G}} = \left\{ P^{\mathbb{G}} = \sum_{n=1}^{N} \lambda_n \rho^{\mathbb{G}}(\boldsymbol{\theta}_n) + \lambda_{N+1}\mathcal{U},\ \boldsymbol{\theta}_n \in \Theta,\ 0 \le \lambda_n \le 1,\ \sum_{n=1}^{N+1} \lambda_n = 1 \right\}, \tag{6.23}$$

where $\boldsymbol{\theta}_n = \{\boldsymbol{s}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\Gamma}_n\}$. These definitions imply $\mathfrak{M}_1^{\mathbb{F}} \subset \ldots \subset \mathfrak{M}_{N-1}^{\mathbb{F}} \subset \mathfrak{M}_N^{\mathbb{F}}$ and $\mathfrak{M}_1^{\mathbb{G}} \subset \ldots \subset \mathfrak{M}_{N-1}^{\mathbb{G}} \subset \mathfrak{M}_N^{\mathbb{G}}$. We introduce

$$\mathfrak{M}_N = \left\{ (P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_N^{\mathbb{F}} \times \mathfrak{M}_N^{\mathbb{F}} \mid \boldsymbol{\theta}_n \text{ is common for } P^{\mathbb{F}} \text{ and } P^{\mathbb{G}}, n = 1, \ldots, N \right\}, \tag{6.24}$$

so that again $\mathfrak{M}_1 \subset \ldots \subset \mathfrak{M}_{N-1} \subset \mathfrak{M}_N$.

Suppose the upper bound for N is fixed, we denote it $N_{\max}$. Since in the general case the true densities $P_0^{\mathbb{F}}$ and $P_0^{\mathbb{G}}$ are not necessarily mixtures, we define the so-called *quasi-true* densities $P_*^{\mathbb{F}}$ and $P_*^{\mathbb{G}}$ as follows. Let

$$\mathrm{KL}^* = \inf_{(P^{\mathbb{F}},P^{\mathbb{G}})\in\mathfrak{M}_{N_{\max}}} \left\{\mathrm{KL}(P_0^{\mathbb{F}}||P^{\mathbb{F}}) + \mathrm{KL}(P_0^{\mathbb{G}}||P^{\mathbb{G}})\right\}, \tag{6.25}$$

where KL denotes the Kullback-Leibler divergence. Then we define

$$N_* = \min\left\{N \mid \exists(P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_N \text{ s.t. } \mathrm{KL}(P_0^{\mathbb{F}}||P^{\mathbb{F}}) + \mathrm{KL}(P_0^{\mathbb{G}}||P^{\mathbb{G}}) = \mathrm{KL}^*\right\}, \tag{6.26}$$

and

$$(P_*^{\mathbb{F}}, P_*^{\mathbb{G}}) = \operatorname*{argmin}_{(P^{\mathbb{F}},P^{\mathbb{G}})\in\mathfrak{M}_{N_*}} \left\{\mathrm{KL}(P_0^{\mathbb{F}}||P^{\mathbb{F}}) + \mathrm{KL}(P_0^{\mathbb{G}}||P^{\mathbb{G}})\right\}. \tag{6.27}$$

We note that additional assumptions are required for the set in (6.26) to be non-empty, they would be considered further.

Definition 3 *The maximum penalized likelihood estimator of N_* is a maximizer $\hat{N}$ over $\{1,\dots,N_{\max}\}$ of*

$$T_{M,K}(N) = \sup_{(P^{\mathbb{F}},P^{\mathbb{G}})\in\mathfrak{M}_N} \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) - a_{M,K}(N), \tag{6.28}$$

where $a_{M,K}(N)$ is some penalty term.

The task is to determine the estimator $\hat{N}$ consistency, that is whether $\hat{N}$ converges to N_* in some sense as $M, K \to \infty$ (we refer to Remark 2 in Chapter 4 for the exact meaning of convergence of "sequences" under $M, K \to \infty$). The existing approaches are not directly applicable to the case of conjugate mixture models $(P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_N$ and associated log-likelihoods $\mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}})$, so the existing consistency proofs need to be generalized. We provide the proof of consistency of $\hat{N}$ for multimodal information criteria in Section 6.3.2. The multimodal information criterion is tested on the simulated and real data in Section 6.3.3.

6.3.1 The *Select* Procedure

The overall procedure to determine the number of multimodal clusters N is outlined below:

1. For $n = 1,\dots,N_{\max}$ do
   (a) Initialize n clusters using the *Initialize* procedure;
   (b) Apply the ConjEM algorithm to converge to maximum likelihood (ML) parameter estimates $\{\hat{\boldsymbol{\theta}}_1,\dots,\hat{\boldsymbol{\theta}}_n,\hat{\boldsymbol{\theta}}_{n+1}\}$;
2. Apply the criterion (6.28) to determine the number of clusters N_*;

6.3.2 Weak Consistency of Multimodal Information Criteria

We start with providing a general inequality on likelihood ratios proved in [Gassiat 2002]. For any $P^{\mathbb{F}}, P^{\mathbb{F}}_* \in \mathfrak{F}$ consider the subset of the unit sphere in $L_2(P^{\mathbb{F}}_0 \mathrm{d}\nu_{\mathrm{F}})$ defined as

$$\left\{ S^{\mathbb{F}}_P = \frac{P^{\mathbb{F}}/P^{\mathbb{F}}_* - 1}{\|P^{\mathbb{F}}/P^{\mathbb{F}}_* - 1\|_{\mathfrak{F},2}},\ P^{\mathbb{F}} \in \mathfrak{F} \setminus P^{\mathbb{F}}_* \right\}, \tag{6.29}$$

where $\|.\|_{\mathfrak{F},2}$ denotes the norm in $L_2(P^{\mathbb{F}}_0 \mathrm{d}\nu_{\mathrm{F}})$ and we let $S^{\mathbb{F}}_{P_*} \equiv 1$. Such functions $S^{\mathbb{F}}_P$ satisfy

$$\|S^{\mathbb{F}}_P\|^2_{\mathfrak{F},2} = \int (S^{\mathbb{F}}_P)^2 P^{\mathbb{F}}_0 \mathrm{d}\nu_{\mathrm{F}} = 1, \tag{6.30}$$

$$\text{and} \quad \int S^{\mathbb{F}}_P P^{\mathbb{F}}_* \mathrm{d}\nu_{\mathrm{F}} = 0. \tag{6.31}$$

The inequality 1.2 of [Gassiat 2002] indicates that

$$\sup_{P^{\mathbb{F}} \in \mathfrak{F}} \mathcal{L}_M(P^{\mathbb{F}}) - \mathcal{L}_M(P^{\mathbb{F}}_*) \leq \frac{1}{2} \sup_{P^{\mathbb{F}} \in \mathfrak{F}} \frac{\left(\sum_{m=1}^{M} S^{\mathbb{F}}_P(\boldsymbol{f}_m) \right)^2}{\sum_{m=1}^{M} (S^{\mathbb{F}}_P)^2_-(\boldsymbol{f}_m)}, \tag{6.32}$$

where $(S^{\mathbb{F}}_P)_-(\boldsymbol{f}) = \min\{0, S^{\mathbb{F}}_P(\boldsymbol{f})\}$ is the negative part of $S^{\mathbb{F}}_P$. A similar set can be constructed for densities $P^{\mathbb{G}}, P^{\mathbb{G}}_* \in \mathfrak{G}$. We denote $\mathfrak{S}_{\mathbb{F}}$ and $\mathfrak{S}_{\mathbb{G}}$ the set of functions $S^{\mathbb{F}}_P$ for $P^{\mathbb{F}} \in \mathfrak{M}^{\mathbb{F}}_{N_{\max}}$ and $S^{\mathbb{G}}_P$ for $P^{\mathbb{G}} \in \mathfrak{M}^{\mathbb{G}}_{N_{\max}}$ respectively.

The consistency derivation would require the following assumptions.

(A1) *The mixture components $\rho^{\mathbb{F}}(\boldsymbol{f}, \boldsymbol{\theta})$ and $\rho^{\mathbb{G}}(\boldsymbol{g}, \boldsymbol{\theta})$ are continuous functions of their second argument $\boldsymbol{\theta}$ for all values of $\boldsymbol{f}$ and $\boldsymbol{g}$. Moreover, there exist functions $\phi \in L_1(P^{\mathbb{F}}_0 \mathrm{d}\nu_F)$ and $\psi \in L_1(P^{\mathbb{G}}_0 \mathrm{d}\nu_G)$ such that for any $(P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_{N_{\max}}$ one has $|\log P^{\mathbb{F}}| \leq \phi$ and $|\log P^{\mathbb{G}}| \leq \psi$.*

This assumption restricts possible types of mixtures. The likelihood of a Gaussian mixture, does not verify (A1) unless the parameter space is bounded. In practice one has to lower bound the variances.

(A2) *The parameter support Θ is compact.*

The assumptions (A1), (A2) and the fact that the KL function in (6.25) is continuous imply the existence of the optimal N_* in (6.26). This assumption is not satisfied for general mixtures of Gaussians without any constraint on the parameter space. However, as mentioned in [Ciuperca 2003] (see also references therein), it is common to consider restrained parameter spaces.

(A3) *The penalty term* $a_{M,K}(N)$ *is increasing and* $a_{M,K}(N) = o(M+K)$. *Also,* $a_{M,K}(N_1) - a_{M,K}(N_2) \to \infty$ *as* $M, K \to \infty$ *for all* $N_1 > N_2$.

This assumption on the penalty term is rather standard and is verified by all the information criteria mentioned in the beginning, except for AIC, for which the last condition does not hold.

(A4) $\mathfrak{S}^2_{\mathbb{F}}$ *and* $\mathfrak{S}^2_{\mathbb{G}}$ *are Glivenko-Cantelli classes.*

Definition 4 *A class* $\mathfrak{S}$ *of measurable functions* S *is called Glivenko-Cantelli with respect to a probability measure* μ*, if*

$$\|\mathbb{P}_m - \mathbb{E}_\mu\|_{\mathfrak{S}} = \sup_{S \in \mathfrak{S}} |\mathbb{P}_m(S) - \mathbb{E}_\mu S| \to 0 \quad a.s. \tag{6.33}$$

(A5) *The true densities* $(P_0^{\mathbb{F}}, P_0^{\mathbb{G}}) \in \mathfrak{M}_{N_{\max}}$.

(A6) *The "sequence"* $\frac{M}{M+K}$ *converges to the limit* $\kappa = 1/2$.

The last two assumptions are rather strong, but we shall use one of them to prove the consistency of the penalized maximum likelihood criterion family. However, as can be seen from the proof, one can think of some less restrictive assumptions. We discuss this later in the Section.

Now everything is ready for the proof of the main result.

Theorem 5 *Let assumptions (A1), (A2), (A3), (A4) hold and one of the conditions (A5) or (A6) be satisfied. Then the estimator* $\hat{N}$ *converges in probability to the true number of components* N_* *when* $M, K \to \infty$ *with* $\frac{M}{M+K}$ *converging to some finite limit* $0 \le \kappa \le 1$.

Proof: 1. *Underestimation.* We start from the traditionally easier part and show that N_* cannot be underestimated. Suppose the opposite, $\hat{N} < N_*$ and consider $T_{M,K}(\hat{N}) - T_{M,K}(N_*)$. By definition of $\hat{N}$ and $T_{M,K}(N)$ one has

$$\sup_{(P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_{\hat{N}}} \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) - a_{M,K}(\hat{N}) \ge \sup_{(P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_{N_*}} \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) - a_{M,K}(N_*). \tag{6.34}$$

Under (A1) and (A2) (see example 19.8, p.272 of [van der Vaart 2004]), for all N the set $\left\{(\log \frac{P^{\mathbb{F}}}{P_0^{\mathbb{F}}}, \log \frac{P^{\mathbb{G}}}{P_0^{\mathbb{G}}}), (P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_N\right\}$ is Glivenko-Cantelli. This means that if we divide (6.34) by $M+K$ and consider the limit under $M, K \to \infty$, we obtain

$$\sup_{(P^{\mathbb{F}}, P^{\mathbb{G}}) \in \mathfrak{M}_{\hat{N}}} \Big(\kappa(\mathrm{KL}(P_0^{\mathbb{F}} || P_*^{\mathbb{F}}) - \mathrm{KL}(P_0^{\mathbb{F}} || P^{\mathbb{F}})) + \\ + (1-\kappa)(\mathrm{KL}(P_0^{\mathbb{G}} || P_*^{\mathbb{G}}) - \mathrm{KL}(P_0^{\mathbb{G}} || P^{\mathbb{G}})) \Big) \ge 0, \tag{6.35}$$

where we used (A3), the definition of the model $(P_*^{\mathbb{F}}, P_*^{\mathbb{G}})$ given by (6.27) and the definition of κ. Suppose now that (A5) is fulfiled. Then $P_0^{\mathbb{F}} = P_*^{\mathbb{F}}$ and (6.35) must be strictly negative, which leads to a contradiction. If (A6) is verified, from the definition (6.27) it follows again that (6.35) is strictly negative. Thus the estimator $\hat{N}$ is a.s. greater than N_*.

2. *Overestimation.* We now prove that N_* cannot be overestimated by showing that $P(\hat{N} > N_*)$ tends to zero as $M, K \to \infty$. Indeed,

$$P(\hat{N} > N_*) = P(\exists N_{\max} \geq N > N_* \text{ s.t. } T_{M,K}(N) \geq T_{M,K}(N_*)) \leq$$
$$\leq \sum_{N=N_*+1}^{N_{\max}} P(T_{M,K}(N) \geq T_{M,K}(N_*)). \tag{6.36}$$

By definition of $T_{M,K}(N)$ each event in the sum (6.36) can be written as

$$\sup_{(P^{\mathbb{F}},P^{\mathbb{G}})\in\mathfrak{M}_N} \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) - a_{M,K}(N) \geq \sup_{(P^{\mathbb{F}},P^{\mathbb{G}})\in\mathfrak{M}_{N_*}} \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) - a_{M,K}(N_*). \tag{6.37}$$

Using the definition (6.27) of the optimal densities, (6.37) is equivalent to

$$\sup_{(P^{\mathbb{F}},P^{\mathbb{G}})\in\mathfrak{M}_N} \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) - \mathcal{L}_{M,K}(P^{\mathbb{F}}, P^{\mathbb{G}}) \geq a_{M,K}(N) - a_{M,K}(N_*), \tag{6.38}$$

since $(P_*^{\mathbb{F}}, P_*^{\mathbb{G}})$ belongs to $\mathfrak{M}_{N_*}$. We can further bound (6.36) from above by

$$\sum_{N=N_*+1}^{N_{\max}} P\Bigg(\sup_{P^{\mathbb{F}}\in\mathfrak{M}_N^{\mathbb{F}}} \mathcal{L}_M(P^{\mathbb{F}}) - \mathcal{L}_M(P_*^{\mathbb{F}}) + \sup_{P^{\mathbb{G}}\in\mathfrak{M}_N^{\mathbb{G}}} \mathcal{L}_K(P^{\mathbb{G}}) - \mathcal{L}_K(P_*^{\mathbb{G}}) \geq$$
$$\geq a_{M,K}(N) - a_{M,K}(N_*)\Bigg). \tag{6.39}$$

Then we use inequality (6.32) and the definitions of $\mathfrak{S}_{\mathbb{F}}$ and $\mathfrak{S}_{\mathbb{G}}$ to write an upper bound for (6.39)

$$\sum_{N=N_*+1}^{N_{\max}} P\Bigg(\frac{1}{2} \sup_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{\left(\sum_{m=1}^{M} S(\boldsymbol{f}_m)\right)^2}{\sum_{m=1}^{M} S_-^2(\boldsymbol{f}_m)} + \sup_{S\in\mathfrak{S}_{\mathbb{G}}} \frac{\left(\sum_{k=1}^{K} S(\boldsymbol{g}_k)\right)^2}{\sum_{k=1}^{K} S_-^2(\boldsymbol{g}_k)} \geq$$
$$\geq a_{M,K}(N) - a_{M,K}(N_*)\Bigg). \tag{6.40}$$

We would like to show that $\sup_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{\left(\sum_{m=1}^{M} S(\boldsymbol{f}_m)\right)^2}{\sum_{m=1}^{M} S_-^2(\boldsymbol{f}_m)}$ and $\sup_{S\in\mathfrak{S}_{\mathbb{G}}} \frac{\left(\sum_{k=1}^{K} S(\boldsymbol{g}_k)\right)^2}{\sum_{k=1}^{K} S_-^2(\boldsymbol{g}_k)}$ are bounded. Indeed,

$$\sup_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{\left(\sum_{m=1}^{M} S(\boldsymbol{f}_m)\right)^2}{\sum_{m=1}^{M} S_-^2(\boldsymbol{f}_m)} \leq \frac{\sup_{S\in\mathfrak{S}_{\mathbb{F}}} \left(\sum_{m=1}^{M} S(\boldsymbol{f}_m)\right)^2}{\inf_{S\in\mathfrak{S}_{\mathbb{F}}} \sum_{m=1}^{M} S_-^2(\boldsymbol{f}_m)} \leq \frac{\sup_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{1}{M} \sum_{m=1}^{M} S^2(\boldsymbol{f}_m)}{\inf_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{1}{M} \sum_{m=1}^{M} S_-^2(\boldsymbol{f}_m)}. \tag{6.41}$$

Using the assumption (A4) that $\mathfrak{S}_{\mathbb{F}}^2$ is Glivenko-Cantelli we get

$$\lim_{M\to\infty} \sup_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{1}{M}\sum_{m=1}^{M} S^2(\boldsymbol{f}_m) = \sup_{S\in\mathfrak{S}_{\mathbb{F}}} \int S^2 P_0^{\mathbb{F}} \mathrm{d}\nu_{\mathbb{F}} = 1,$$

$$\text{and} \quad \lim_{M\to\infty} \inf_{S\in\mathfrak{S}_{\mathbb{F}}} \frac{1}{M}\sum_{m=1}^{M} S_-^2(\boldsymbol{f}_m) = \inf_{S\in\mathfrak{S}_{\mathbb{F}}} \|S_-\|_{\mathfrak{F},2}^2. \tag{6.42}$$

By construction we have $\|S_-\|_{\mathfrak{F},2}^2 > 0$. Indeed, if it were not the case, there would exist a function $S \in \mathfrak{S}_{\mathbb{F}}$ such that $\|S_-\|_{\mathfrak{F},2}^2 = 0$, as soon as $\mathfrak{S}_{\mathbb{F}}$ is compact. Assuming $P_0^{\mathbb{F}} > 0$ would imply $S \equiv 0$. But this contradicts the definition of $S \in \mathfrak{S}_{\mathbb{F}}$ that must satisfy $\|S\|_{\mathfrak{F},2} = 1$.

Thus we proved that $\sup\limits_{S\in\mathfrak{S}_{\mathbb{F}}} \dfrac{\left(\sum\limits_{m=1}^{M} S(\boldsymbol{f}_m)\right)^2}{\sum\limits_{m=1}^{M} S_-^2(\boldsymbol{f}_m)}$ was bounded. Similar developments show that $\sup\limits_{S\in\mathfrak{S}_{\mathbb{G}}} \dfrac{\left(\sum\limits_{k=1}^{K} S(\boldsymbol{g}_k)\right)^2}{\sum\limits_{k=1}^{K} S_-^2(\boldsymbol{g}_k)}$ is bounded as well. But $a_{M,K}(N) - a_{M,K}(N_*) \to \infty$ by (A3), from where using (6.40) we conclude that $P(\hat{N} > N_*)$ converges to zero as M and K tend to infinity. ■

This result shows that criteria that belong to the penalized maximum likelihood class provide asymptotically consistent estimates. The proof required one of the assumptions, (A5) or ((A6)) to be made. The first of them restrains the unknown distributions $P_0^{\mathbb{F}}$ and $P_0^{\mathbb{G}}$ to belong to the considered parametrized class. The second one assumes certain asymptotic relation between the number of observations in the modalities. These assumptions are quite restrictive. However, the issue that required such a coarse decision is worth to pay attention to. We are convinced that these assumptions, in fact, can be significantly weakened.

Indeed, the only place where we had to use them is the equation (6.35), to show that its left-hand side was negative. The latter consists of two weighted differences $\mathrm{KL}(P_0^{\mathbb{F}}||P_*^{\mathbb{F}}) - \mathrm{KL}(P_0^{\mathbb{F}}||P^{\mathbb{F}})$ and $\mathrm{KL}(P_0^{\mathbb{G}}||P_*^{\mathbb{G}}) - \mathrm{KL}(P_0^{\mathbb{G}}||P^{\mathbb{G}})$. It can be easily proved that at most one of them is positive. If this was not true, we could have estimated the sum from below by a similar expression, where instead of κ and $1-\kappa$, the minimum of the two is used. On the one hand, that expression is supposed to be positive. From the other hand, the definition (6.27) implies that it is strictly negative.

Thus we deal with the case where a pair of conjugate distributions $(P_*^{\mathbb{F}}, P_*^{\mathbb{G}})$ is an optimal approximation to $(P_0^{\mathbb{F}}, P_0^{\mathbb{G}})$ when both modalities are equally observable. But in real conditions, when the number of observations in different modalities is proportional to κ and $1-\kappa$, distributions $(P_*^{\mathbb{F}}, P_*^{\mathbb{G}})$ may lose this optimality property. One solution is to make the assumption (A6) and fix the proportions, which we consider to be inappropriate when dealing with real-world scenarios. At the same time, we note that so far no assumption was made on relation between the two distributions $P_0^{\mathbb{F}}$ and $P_0^{\mathbb{G}}$. It would be reasonable to

suppose that as they correspond to the same object configuration in the hidden space, there should be some connection between them. Assumption (A5) is one possibility to impose such a relation. Of course, one can think of milder conditions.

6.3.3 Experimental Validation

The *Select* procedure was tested on the audio-visual (AV) object localization task where the aim is to determine the number of AV objects present in the scene. We first considered the simulated data examples that were used to verify the initialization algorithm, namely the well separated (WS), poorly separated (PS) and partially observed (PO) object configurations.

For each configuration we take the parameter initializations $\boldsymbol{\theta}_n$, $n = 1, \ldots, N_{\max}$ and run the ConjEM algorithm till convergence. The obtained parameter estimates $\hat{\boldsymbol{\theta}}_n$ are used to compute the 'model score' – a penalized likelihood value given by (6.28). The corresponding results are depicted in Figure 6.6. The graphs show model scores for different numbers of objects. The selection criterion chooses the model with the best (maximal) score, which is marked by a white circle.

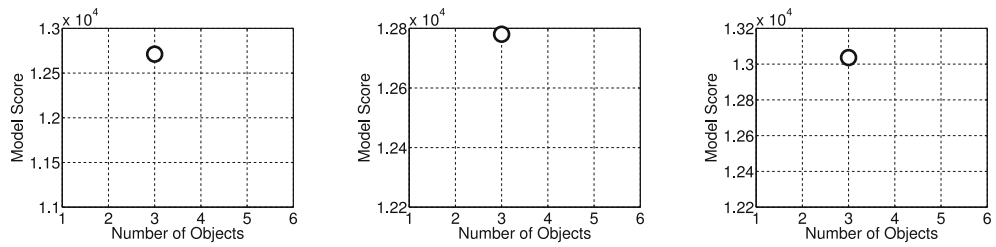
We note that the criterion performed well both, in the case of well separated and poorly separated objects and did not underestimate or overestimate their number. From here we conclude that such a criterion is robust to configuration changes and could be used for real data, where the distributions are not necessarily Gaussian.

At the same time, the *Select* procedure choses correctly the number of objects in the case of partially observed data. This is an important property, since the multimodal criterion is the *only* way to detect all the objects: each modality contains two strong clusters and the third cluster appears only in the case of multimodal integration. This way the selection criterion together with the initialization strategy and the ConjEM algorithm implements the multimodal enhancement principle.

Next we consider scenarios with real data from the CAVA database. Two time intervals from the cocktail party (CTMS3) scenario are taken that were used in the previous Section to check the *Initialize* procedure: the well separated objects case (frames 181–190) and the partially occluded objects case (frames 211–220). The ConjEM algorithm is run on the initialization results to get the maximum likelihood (ML) parameter estimates. The latter are used to compute model scores for various numbers of objects. The score graphs are given in Figure 6.7.

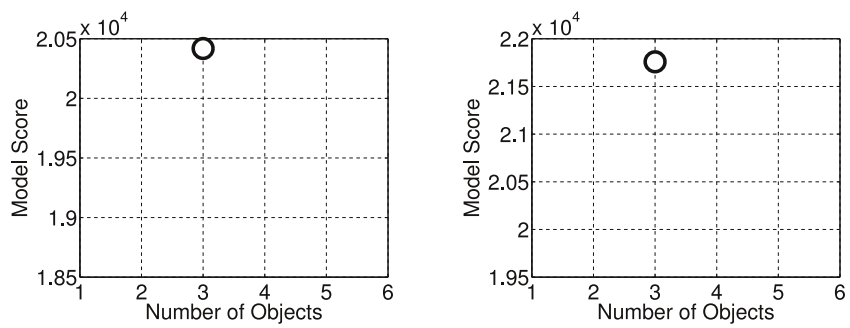
6.4 Discussion

The two procedures proposed in this Chapter - *Initialize* and *Select*, play important role in the multimodal clustering task. Efficiency of the ConjEM algorithm derived in Chapter 5 is highly dependent on the initial parameters and model choice strategies. Moreover, in order to fully benefit from the ConjEM multimodal integration capabilities, they need to



(a) well separated, WS (b) poorly separated, PS (c) partially observed, PO

Figure 6.6: Simulated data experiments. Model scores obtained by the multimodal *Select* procedure for the three configurations of object. Each model is characterized by the number of objects it contains and is assigned a score based on the penalized likelihood value. The selection criterion chooses the model with the best (maximal) score. In all the cases the correct number of objects (three) is chosen. For the partially observed data multimodal criterion is the *only* way to detect all the objects, as soon as each modality contains only two strong clusters.



(a) well separated, WS (b) partially occluded, PO

Figure 6.7: Real data experiments. Model scores obtained by the multimodal *Select* procedure for the two configurations of objects that correspond to frames 181–190 (well separated objects) and 211–220 (partially occluded objects) of the cocktail party scenario. Each model is characterized by the number of objects it contains and is assigned a score based on the penalized likelihood value. The selection criterion chooses the model with the best (maximal) score. Even in the case of partially occluded objects the correct number of objects (three) is chosen.

be fully multimodal and consider observations from different sensors on equal basis. The *Initialize* and *Select* procedures fulfil these requirements, being symmetric with respect to the modalities.

Assuming certain detector properties to be known we developed an original *cluster sampling* technique based on multimodal predictive distributions. It proved to be efficient on both simulated and real data. The initial estimates are close to the optimal parameter values, so only a few iterations of the ConjEM algorithm are required afterwards to converge.

The multimodal selection criterion was developed to choose the best model matching the data. It was inspired by the existing model selection strategies. Though standard consistency results could not be applied directly to the multimodal case. Thus we prove the multimodal consistency of our criterion and show its performance on simulated and real data. One important feature of our criterion is that it allows for multimodal enhancement: weak cluster from one modality can be enhanced by a cluster from another modality leading to multimodal object detection where single modality models fail.

The *Select* and *Initialize* procedures are based on the same framework as the ConjEM algorithm, so that they can be naturally integrated. This way ConjEM obtains the good properties of a global optimization method. The more so, even though in the examples we used the initialization, optimization and model selection procedures in an offline manner, it is easy to make them work online. By analogy with jump-diffusion processes [Grenander 1994, Jacobsen 2006], one can consider the ConjEM algorithm as a diffusion, the *Initialize* procedure as the jump proposal method and the *Select* procedure as the jump acceptance criterion. This leads to efficient optimization schemes and multimodal tracking algorithms. We use such an approach to perform multimodal multiobject audio-visual tracking in Chapter 7.

We outline the strong points of our initialization and model selection approach below:

- **Fully multimodal**: both procedures do not assume any of the modalities as the leading one, they are completely symmetric with respect to observation spaces;
- **Multimodal enhancement**: the *Select* procedure is able to enhance stimuli from one modality with stimuli from other modalities to detect weak clusters;
- **Efficient sampling**: the initialization strategy is based on assumptions on data that are always verified, which leads to a more efficient *Initialize* procedure;
- **Consistent selection**: the model selection criterion is theoretically well-founded and possesses asymptotic consistency property;
- **Online use**: the *Initialize* and *Select* procedures are designed in such a way that they can be incorporated into the dynamic scene model and used to track multiple multimodal objects;

CHAPTER 7

Spatio-temporal Multimodal Clustering

Sommaire

7.1 Multimodal Multiobject Tracking 115
7.2 Conjugate Filtering for Multimodal Multiobject Tracking 116
7.3 Experimental Results 119
7.4 Discussion 120

Multimodal multiobject tracking is a difficult problem that involves various types of noise, complex object configuration dynamics and nontrivial object appearance changes. Some of these events can completely ruin the observed short-term data. This may confuse an algorithm that relies on data obtained for short time intervals. The most efficient way to deal with abrupt changes and short-term data corruption is to incorporate an appropriate dynamics model.

The case of unaligned multimodal observations gives rise to the 'multimodal filtering' task (by analogy with the multimodal clustering task for the stationary case). We use the formalism of conjugate mixture models together with the associated optimization procedures and show how it can be adjusted to the dynamic case. The ways to perform multimodal filtering are discussed. Real-data results are provided that show the dynamic model superiority over the simple stationary model adjustments presented in Chapter 5.

7.1 Multimodal Multiobject Tracking

The task of multimodal multiobject tracking arises naturally whenever several objects need to be tracked based on observations arriving from different sensors. Numerous potential applications can be found in perception modelling ([Pouget 2002a, Ernst 2002, Anastasio 2000, King 2004, King 2005, Haykin 2005]), military target tracking ([Luo 2002, Pannetier 2008]), robotics ([Castellanos 1999, Allen 1995, Joshi 1999]) and various other research domains.

The tracking problem is usually formulated as a *filtering task* that aims at inferring the most recent hidden object state by all the available observations. This treatment has received much attention because of the development of efficient filtering techniques. Kalman

filter [Kalman 1961] and its various extensions [Wan 2000, Lefebvre 2001] are popular and efficient parametric techniques that use a multivariate Gaussian distribution for the posterior over the hidden state (belief). Their generalizations to the multiple objects case exist [Mahler 2005]. Particle filters [Doucet 2001] is a popular nonparametric technique that uses weighted particle sets to approximate the belief. This approach is more general, though it typically requires more computational effort. Particle filters can also be generalized to the multiple objects case [Khan 2005].

There have been several models that aimed at extending the particle filters approach to the multimodal case [Checka 2004, Chen 2004, Gatica-Perez 2007]. However, as already mentioned in Chapter 4, the dimensionality of the parameter space grows exponentially with the number of objects, causing the number of required particles to increase dramatically and augmenting computational costs. A number of efficient sampling procedures have been suggested [Chen 2004, Gatica-Perez 2007] to keep the problem tractable. Of course this is done at the cost of a loss in model generality, and hence these attempts are strongly application-dependent. Another drawback of such models is that they cannot provide estimates of accuracy and importance of each modality with respect to each object. The sampling and distribution estimation are performed in the parameter space, but no statistics are gathered for the observation spaces.

So far there has been no attempt to apply parametric approaches to the task of multimodal multiobject tracking. The single-object localization model of [Beal 2003] was extended for single-object tracking tasks in [Kushal 2006] and for multiple-object localization tasks in [Hospedales 2007, Hospedales 2008]. The latter approach incorporated several single-object models into the multiple-object model and tracking was performed through inference of filtering distributions. However, a learning phase for each of the object was required to perform multiobject tracking.

In this Chapter we show how the conjugate muxture model can be extended to the non-stationary case to perform robust multimodal multiobject tracking. We note that in Chapter 5 we have already tried to apply our conjugate muxture model directly to the multimodal multiobject tracking task relying on the model's attractor stability. It performed well on simple scenarios, though failed in the case of complete occlusion. We formally introduce the non-stationary case extension in Section 7.2, show its performance on real data in Section 7.3 and conclude the Chapter with a discussion on the results.

7.2 Conjugate Filtering for Multimodal Multiobject Tracking

As in Chapter 4, we assume that the system consists of N objects observed in two feature spaces $\mathbb{F} \subseteq \mathbb{R}^r$ and $\mathbb{G} \subseteq \mathbb{R}^p$. The objects are described by *tying parameters* $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N \in \mathbb{S} \subseteq \mathbb{R}^d$. We assume that transformations

$$\begin{cases} \mathcal{F} : \mathbb{S} \to \mathbb{F} \\ \mathcal{G} : \mathbb{S} \to \mathbb{G} \end{cases} \tag{7.1}$$

are known, that map $\mathbb{S}$ into the observation spaces $\mathbb{F}$ and $\mathbb{G}$ respectively. These transformations are defined by the physical and geometric properties of the sensors and they are supposed to be known. We treat the general case when both $\mathcal{F}$ and $\mathcal{G}$ are non-linear.

One way to account for configurations of varying number of objects N_t that exhibit dynamic behaviour in space $\mathbb{S}$ is to introduce a *jump diffusion* process [Grenander 1994, Jacobsen 2006]. This is a marked point process defined on $\left(T, \bigcup_{N=0}^{\infty} \Theta_N\right)$ consisting of the timestamp set T and the set of configurations, we denoted Θ_N the configuration space containing N objects, $N = 0, 1, \ldots$. Jumps between configurations are performed according to a point process that generates their times $\tau \in T$ and destinations $\theta \in \bigcup_{N=0}^{\infty} \Theta_N$. Between the jumps objects are supposed to follow some system dynamics. When dealing with parametric estimation tasks, a popular way to describe dynamics with random effects is to define an associated stochastic differential equation (SDE) [Rozovskii 1990]. The general SDE for the task of multimodal multiobject tracking is given by the following Itō equation

$$\mathrm{d}\boldsymbol{\theta}_{\mathrm{S}}(t) = \boldsymbol{\mu}(t, \boldsymbol{\theta}_{\mathrm{S}}(t))\mathrm{d}t + \boldsymbol{\Lambda}(t, \boldsymbol{\theta}_{\mathrm{S}}(t))\mathrm{d}W_t^{\Theta}, \tag{7.2}$$

where $\boldsymbol{\theta}_{\mathrm{S}} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\}$ is the set of tying parameters, $\boldsymbol{\mu}$ is the drift field, $\boldsymbol{\Lambda}$ is the diffusion field and W^{Θ} is a multidimensional Brownian motion. In case when the dynamic models for different objects can be assumed to be independent, the equation (7.2) can be split into a system of N simpler equations

$$\mathrm{d}\boldsymbol{s}_n(t) = \boldsymbol{\mu}_n(t, \boldsymbol{s}_n(t))\mathrm{d}t + \boldsymbol{\Lambda}_n(t, \boldsymbol{s}_n(t))\mathrm{d}W_t^{(n)}, \tag{7.3}$$

where $W^{(n)}$ are independent d-dimensional Brownian motions. Assuming, as previously, independent observation models for every object n, we can write

$$\mathrm{d}\boldsymbol{Z}_{\mathbb{F}}^{(n)}(t) = \mathcal{F}(\boldsymbol{s}_n(t))\mathrm{d}t + \boldsymbol{\Sigma}_n(t, \boldsymbol{s}_n(t))\mathrm{d}W_t^{\mathbb{F}}, \tag{7.4}$$

$$\text{and} \quad \mathrm{d}\boldsymbol{Z}_{\mathbb{G}}^{(n)}(t) = \mathcal{G}(\boldsymbol{s}_n(t))\mathrm{d}t + \boldsymbol{\Gamma}_n(t, \boldsymbol{s}_n(t))\mathrm{d}W_t^{\mathbb{G}}, \tag{7.5}$$

where $\boldsymbol{Z}_{\mathbb{F}}^{(n)}(t) = \int_0^t \boldsymbol{f}(\tau)\mathrm{d}\tau$ and $\boldsymbol{Z}_{\mathbb{G}}^{(n)}(t) = \int_0^t \boldsymbol{g}(\tau)\mathrm{d}\tau$ are the observed processes in modality spaces $\mathbb{F}$ and $\mathbb{G}$ respectively, and $\boldsymbol{\Sigma}_n$ and $\boldsymbol{\Gamma}_n$ are the corresponding diffusion fields. The task is thus to compute the filtering density $\mathrm{d}P\left(\boldsymbol{\theta}_{\mathrm{S}}(t) \in \mathrm{d}\boldsymbol{\theta} | \mathcal{Z}_t^{\mathbb{F}}, \mathcal{Z}_t^{\mathbb{G}}\right)/\mathrm{d}\boldsymbol{\theta}$, where $\mathcal{Z}_t^{\mathbb{F}}$ and $\mathcal{Z}_t^{\mathbb{G}}$ are the σ-algebras, generated by $\boldsymbol{Z}_{\mathbb{F}}^{(n)}(\tau)$ and $\boldsymbol{Z}_{\mathbb{G}}^{(n)}(\tau)$ respectively for $n = 1, \ldots, N$ and all $\tau \in [0, t]$. The equations for $\mathrm{d}P\left(\boldsymbol{\theta}_{\mathrm{S}}(t) \in \mathrm{d}\boldsymbol{\theta} | \mathcal{Z}_t^{\mathbb{F}}, \mathcal{Z}_t^{\mathbb{G}}\right)/\mathrm{d}\boldsymbol{\theta}$ for a system defined by (7.2), (7.4) and (7.5) do not admit a closed-form solution even in the simpler case of one observation space. The derivation and general solution of the corresponding differential equation can be found, for example, in [Rozovskii 1990]. In fact, the explicit solution is rarely available and some kind of approximation is required.

Approximate filtering. One way to perform inference of the approximate filtering distribution is to derive the multimodal analogue of one of the extensions of the Kalman

filter based on weighted approximations of feature space densities $\mathrm{d}P\left(\boldsymbol{f}(t) \in \mathrm{d}\boldsymbol{f}\right)/\mathrm{d}\boldsymbol{f}$ and $\mathrm{d}P\left(\boldsymbol{g}(t) \in \mathrm{d}\boldsymbol{g}\right)/\mathrm{d}\boldsymbol{g}$ in the parameter space $\mathbb{S}$. This leads, for example, to multimodal extended Kalman filter (parametric approach), multimodal unscented Kalman filter (semi-parametric approach) schemes. We do not consider these algorithms here in detail, since their derivation is similar to the case of a single modality. However, we would like to point out one advantage of our framework for such algorithms.

Suppose that the filtering algorithm at time instant $t_1 < t_2$ approximates the distribution $P(\boldsymbol{s}_n(t_2) \in \mathrm{d}\boldsymbol{s})|\mathcal{Z}^{\mathbb{F}}_{t_1}, \mathcal{Z}^{\mathbb{G}}_{t_1})$ by a Gaussian distribution in the model given by (7.3), (7.4) and (7.5) with $\boldsymbol{\Sigma}_n(t, \boldsymbol{s}_n(t)) = \boldsymbol{\Sigma}_n(t)$ and $\boldsymbol{\Gamma}_n(t, \boldsymbol{s}_n(t)) = \boldsymbol{\Gamma}_n(t)$. Then the maximum a posteriori (MAP) estimate of the position $\hat{\boldsymbol{s}}_n(t_2)$ can be found using the conjugate EM (ConjEM) algorithm, all the acceleration techniques from Chapter 5 apply.

Indeed, instead of the log-likelihood function (4.18) introduced in Chapter 5, we consider the log-posterior function:

$$\mathcal{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\theta}) = \sum_{m=1}^{M} \log P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}) + \sum_{k=1}^{K} \log P^{\mathbb{G}}(\boldsymbol{g}_k;\ \boldsymbol{\theta}) + \sum_{n=1}^{N} \log P^{\mathbb{S}}(\boldsymbol{s}_n;\ \boldsymbol{\theta}), \tag{7.6}$$

where $P^{\mathbb{S}}(\boldsymbol{s}_n;\ \boldsymbol{\theta})$ is a Gaussian. Then the E step of the accelerated ConjEM algorithm given by (5.6) and (5.6) would remain the same. The only change in the M step would concern the function $Q_n^{(q)}(\boldsymbol{s})$ given by (5.13) which would become

$$\begin{aligned} Q_n^{(q)}(\boldsymbol{s}) = \ & - \sum_{m=1}^{M} \alpha_{mn}^{(q)} (\|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s})\|^2_{\boldsymbol{\Sigma}_n(\boldsymbol{s})} + \log|\boldsymbol{\Sigma}_n(\boldsymbol{s})|) - \\ & - \sum_{k=1}^{K} \beta_{kn}^{(q)} (\|\boldsymbol{g}_k - \mathcal{G}(\boldsymbol{s})\|^2_{\boldsymbol{\Gamma}_n(\boldsymbol{s})} + \log|\boldsymbol{\Gamma}_n(\boldsymbol{s})|) - \|\tilde{\boldsymbol{s}} - \boldsymbol{s}\|^2_{\boldsymbol{\Upsilon}}, \end{aligned} \tag{7.7}$$

where $\tilde{\boldsymbol{s}}$ and $\boldsymbol{\Upsilon}$ are the mean and the variance of the Gaussian distribution $P^{\mathbb{S}}(\boldsymbol{s}_n;\ \boldsymbol{\theta})$. We note that the expression (7.7) has the same form as (5.13) from Chapter 5. That is adding a Gaussian prior on the object position would be treated as making the object observed in the parameter space. Thus this change is equivalent to the increase in modality spaces number and all the results concerning acceleration strategies of the ConjEM algorithm apply.

Stochastic approximation. Another possibility for the parameter $\boldsymbol{\theta}(t)$ inference consists in applying *stochastic approximation* [Wasan 1969, Nevelson 1976, Benveniste 1990]. This method, designed to solve statistical estimation problems, updates iteratively the existing estimator $\hat{\boldsymbol{\theta}}(t)$ based on new information $\boldsymbol{f}(t)$ and $\boldsymbol{g}(t)$. The general algorithm takes the form

$$\mathrm{d}\hat{\boldsymbol{\theta}}(t) = \gamma(t)\left(H(\hat{\boldsymbol{\theta}}(t), \boldsymbol{f}(t), \boldsymbol{g}(t))\mathrm{d}t + \boldsymbol{\Upsilon}(t, \hat{\boldsymbol{\theta}}(t))\mathrm{d}\hat{W}_t^{\mathbb{S}}\right), \tag{7.8}$$

where $\gamma(t)$ is the gain function, $H(\hat{\boldsymbol{\theta}}(t), \boldsymbol{f}(t), \boldsymbol{g}(t))$ is the drift field, usually chosen so that $\mathbb{E}_{\boldsymbol{F},\boldsymbol{G}} H(\hat{\boldsymbol{\theta}}(t), \boldsymbol{F}(t), \boldsymbol{G}(t)) = 0$ if and only if the estimate $\hat{\boldsymbol{\theta}}(t)$ is equal to the real parameter values $\boldsymbol{\theta}_*(t)$, and $\boldsymbol{\Upsilon}(t, \hat{\boldsymbol{\theta}}(t))$ is the diffusion field accounting for small algorithm perturbations. We refer to the sources cited above for specific choices of γ, H and $\boldsymbol{\Upsilon}$ and convergence conditions.

	intervals	speaking	detected	E1	E2
M1	166	89	82	0.08	0.06
TTOS1	76	69	65	0.06	0.23
CTMS3	219	97	71	0.27	0.39

Table 7.1: Comparative results of the dynamic algorithm for the three scenarios: meeting (M1), moving target (TTOS1) and cocktail party (CTMS3). In each case the total number of frames, ground truth on the total number of auditory activity events to be detected, the total number of actually detected auditory activity and the probabilities of 'missed target' and 'false alarm' errors are given.

7.3 Experimental Results

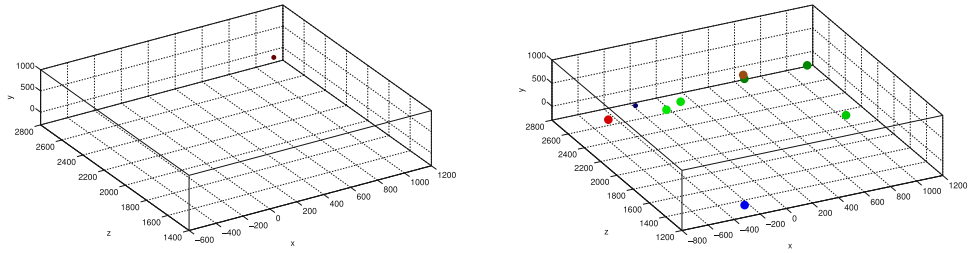
We evaluated the dynamic version of the ConjEM algorithm on the *meeting*, *tracking* and *cocktail party* scenarios (sequences M1, TTOS1 and CTMS3 of the CAVA database presented in Chapter 2). Both auditory activity estimation and tracking accuracy were considered, as previously in Chapter 5.

Since the exact system dynamics in the general audio-visual tracking task are not known and one can only assume the speed of dynamic scene changes, we adopt the stochastic approximation approach to multimodal multiobject tracking for the diffusion part. The gain function was taken to be constant $\gamma(t) \equiv 0.1$. We assumed independent object dynamics, and took the drift term that coincided with the direction of the ConjEM algorithm optimization. Thus the stationarity condition is asymptotically fulfiled. The diffusion part of (7.8) was not included.

To account for scene configuration changes (objects that enter and exit the scene, comlete occlusions), we run the *Initialize* and *Select* procedures to propose new clusters and accept/reject them or to delete existing clusters that no longer receive observations. This strategy resembles a jump-diffusion process [Grenander 1994, Jacobsen 2006], where diffusion is carried out through (7.8) and jumps are generated by the initialization and selection procedures. Similar approaches can be found in video-based tracking [Yao 2008].

One advantage of considering the dynamic model is that different time scales can be used for different modalities to estimate the object activity. Considering longer time intervals for auditory data leads to the auditory activity detection improvement, see Table 7.1 and Table 5.4. Some short-term effects of ambient sounds and reverberations are eliminated which decreases 'false alarm' probabilities.

Spatial localization results are also improved with respect to those from Chapter 5. The dynamic version of the ConjEM algorithm can handle not only partial, but also complete occlusions. Different cases are demonstrated on the cocktail party (CTMS3) sequence in Figure 7.2. After the objects are initialized (a), one of them gets completely occluded (b)-(c) which results in track loss and consequent detection (d). Another occlusion happens to be more rapid (e)-(f), so that the object reappears before the cluster was eliminated.



(a) moving target, TTOS1 (b) cocktail party, CTMS3

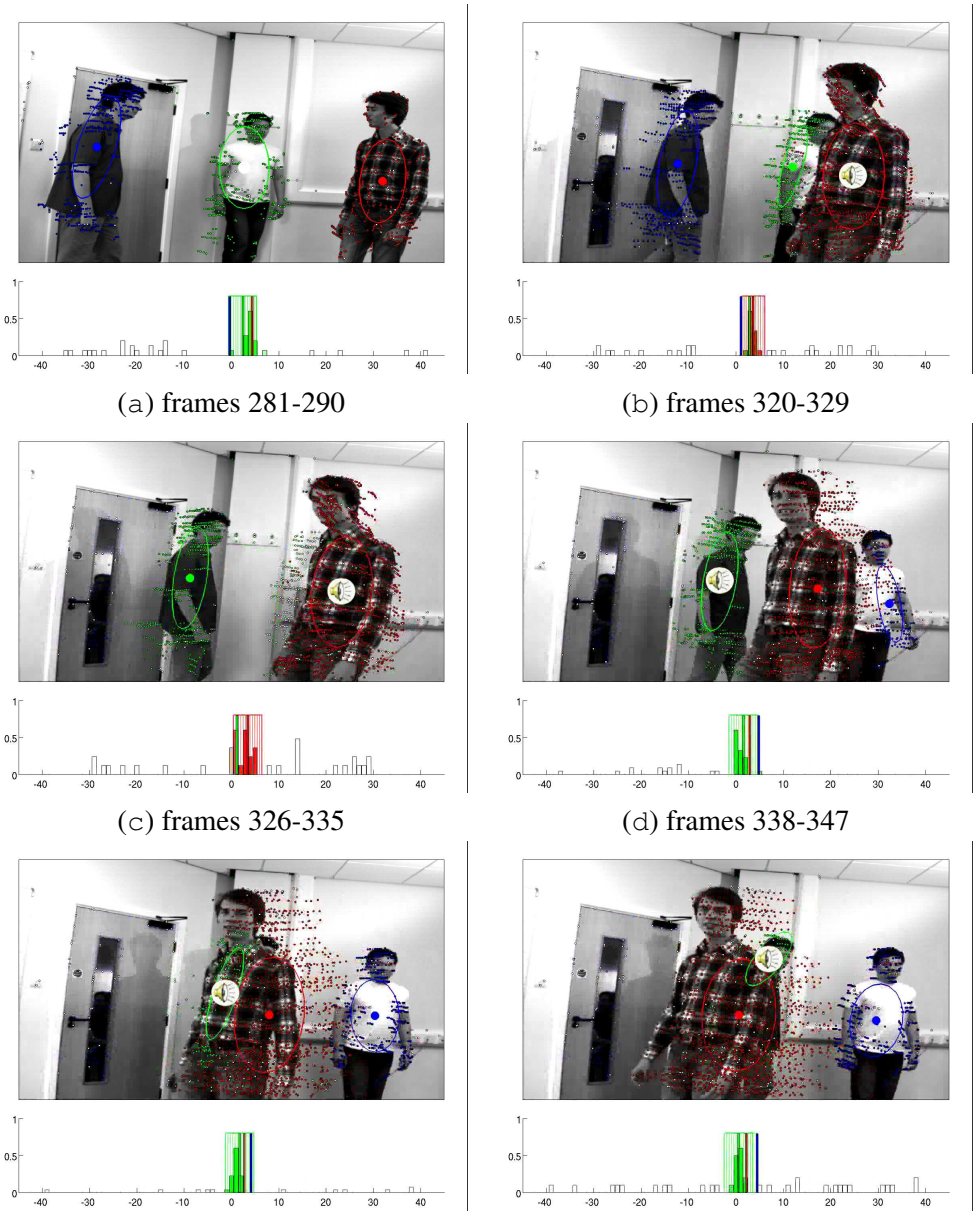
Figure 7.1: Estimated ambient space trajectories for the (a) moving target (TTOS1), and (b) cocktail party (CTMS3) scenarios. Motion is shown with colour gradient from darker to lighter colours. Points where the algorithm lost/regained track of an object are marked with coloured points. Green object is occluded several times. The first time track is lost and as it is regained, the estimate captures part of the red object that goes nearby and follows it (right green segment). But after the second occlusion the green object is redetected and properly tracked (middle green segment). The third occlusion (leftmost green segment) does not spoil the estimate. Blue object was not lost even after one occlusion.

In general the trajectories obtained with the dynamic version of the ConjEM algorithm are smoother and more precise - the position estimates are within 5cm from the object location in the XZ-plane. The precision in the Y coordinate (vertical axis) is typically worse because of the cluster shapes that are typically elongated in the scenarios we consider and admit greater variability in vertical direction. The summary on estimated trajectories for moving target (TTOS1) and cocktail party (CTMS3) scenarios is given in Figure 7.1. See Figure 5.9 for comparison. Object motion is shown with colour gradient from darker to lighter colours. Points where the algorithm lost/regained track of an object in the CTMS3 sequence are marked with coloured points.

7.4 Discussion

The multimodal multiobject tracking task is a hard problem due to various strong noise contaminating the observations and scene dynamics that are usually hard to estimate even without noise. In this Chapter we addressed this problem within the ConjEM framework. We showed how our approach could be efficiently combined with different tracking techniques to benefit from integration of both spatial information coming from multiple modalities and temporal information kept by a system.

On the one hand, the powerful ConjEM framework with efficient *Initialize* and *Select* procedures provides parameter inference from multiple modalities, automatically weighting the data according to the amount of information it contains. It enhances weak multimodal clusters that can be then detected and tracked and hence is responsible for the scene configuration representation.



(a) frames 281-290 (b) frames 320-329

(c) frames 326-335 (d) frames 338-347

(e) frames 346-355 (f) frames 351-360

Figure 7.2: Cocktail party scenario tracking results. After the objects are initialized (a), one of them gets completely occluded (b)-(c) which results in track loss and a detection that follows (d). Another occlusion happens to be more rapid (e)-(f), so that the object reappears before the cluster is eliminated. The results are shown projected onto the left image plane. Colours encode the observation-to-cluster assignments and the auditory activity is shown with the speaker symbol. The "visual" covariance matrix associated with the Gaussian component is projected onto the image plane.

On the other hand, a well-established framework for parameter inference in dynamic systems accounts for proper temporal tracks of multiple multimodal objects. They can become completely invisible for a short period of time and nevertheless still be followed using the estimated trajectory information.

The results show a clear advantage for the joint multimodal tracking over the single ConjEM algorithm, as well as potential benefits over single modality tracking techniques through multimodal enhancement. Both, object auditory activity and ambient space position estimates were improved with respect to the ConjEM results presented in Chapter 5.

We outline the advantages of the dynamic ConjEM framework:

- **Fully multimodal**: the framework benefits from the ConjEM capability of putting all the modalities on equal basis, weighting them based on the amount of information they provide and integrating the multimodal data;
- **Multimodal enhancement**: the ability of the *Initialize* and *Select* procedure to detect and enhance stimuli from one modality with stimuli from other modalities to reinforce weak clusters can also be exploited within the dynamic ConjEM framework;
- **Extensibility**: as with the ConjEM framework, various multimodal features can be added to the dynamic ConjEM to improve tracking;
- **Robust tracking**: ConjEM allows for efficient integration with well-established tracking techniques that can handle temporal invisibility of an object;

CHAPTER 8

Conclusion

Sommaire

8.1 Main Contributions 123
8.2 Future Work 125

The goal of my thesis was to develop a full and efficient framework for audio-visual integration and, in particular, for audio-visual object detection and localization.

I first address the problem making a simplifying assumption of a quasi-stationary or slowly varying object configuration. Under this assumption, I developed a full framework possessing attractive theoretical properties that solves a number of important issues: *i*) hardware calibration (Chapter 3); *ii*) estimation of the number of objects (Chapter 6); *iii*) efficient and accurate initialization (Chapter 6); *iv*) consistent multimodal integration (Chapters 4 and 5) with *v*) guaranteed accuracy and reliability (Chapter 5). The ideas and models that I developed in this framework are general and can be potentially applied to any multimodal clustering task. All the theoretical facts proved about the models are application-independent. However, in the experimental results sections of this thesis I demonstrate how to tune every proposed technique for the particular case of audio-visual integration.

Then I show that this framework could be still used without the assumption on scene dynamics and address the problem of inclusion of object dynamics in the multimodal integration model (Chapter 7). Again, the proposed approach is general and uses a well-established methodology. It can be applied to various multimodal tasks. I believe that this combination of multimodal integration model with system dynamics is very promising in that it further improves the conjugate clustering approach towards the conjugate filtering framework. The latter offers broader range of applications and better performance in terms of robustness to configuration changes (such as visual occlusion) and track losses.

We proceed with the summary of major contributions of the thesis and discussion of perspectives for future research.

8.1 Main Contributions

This thesis contains a number of original contributions that can be split into two groups: *i*) the theoretical models and facts on multimodal integration, and *ii*) their versions tuned for

the task of the audio-visual integration. Below we provide the summary of both groups.

Theoretical models and facts

- **Conjugate mixture models (CMM)**: the formalism of conjugate mixture models was introduced to address the multimodal integration task. It allows to preserve characteristics that are specific to the modalities, while reinforcing integration through the features that are common. Asymptotic identifiability of CMM's is proved and various extensions are proposed concerning different choices of single modality mixtures, conjugate random fields and conjugate point processes;
- **Kullback Proximal optimization algorithm family for CMM**: a class of optimization algorithms for Gaussian CMM was derived within the Kullback Proximal (KP) framework, their convergence properties are discussed;
- **Efficient conjugate EM implementation for CMM**: the multimodal EM algorithm (ConjEM) that belongs to the KP family was improved by transforming the optimization problem to a more convenient form. Several acceleration strategies were proposed. Attractive convergence properties were proved for a large class of CMM models;
- **CMM initialization based on predictive densities**: an efficient method for CMM initialization was proposed based on predictive densities constructed from multimodal data. This method is fully multimodal in the sense that it puts all the modalities on equal footing. It plays role of a sampling technique for an optimization algorithm for CMM, providing the characteristics of a global optimization method and improving the convergence speed and the final estimate.
- **Multimodal criterion for model selection**: a multimodal criterion for CMMs was formulated, its consistency properties were proved. Together with the multimodal initialization strategy and the ConjEM optimization algorithm it provides an efficient multimodal integration strategy that enables multimodal enhancement;
- **Multimodal filtering algorithms**: several possibilities for extending CMMs to the multimodal tracking tasks were offered; their way to efficiently combine filtering algorithms with the CMM initialization and model selection algorithms is described.

Audio-visual (AV) integration contributions

- **CAVA database**: a set of realistic AV scenarios was designed and acquired to provide the evaluation ground for multimodal algorithms that work with head-like devices comprising two microphones and two cameras; annotation was performed for certain scenarios;
- **AV calibration**: the AV calibration algorithm was developed to ensure proper alignment of A and V data; its evaluation on synthetic and real data is provided;

- **AV localization and activity detection**: the theoretical CMM framework was applied to the task of localization of multiple AV objects; we consider the AV integration task in the 3D space which reinforces the integration; the acceleration strategies for the case of AV data are derived and demonstrated, different implementational aspects of the optimization algorithms are discussed; the performance is shown on simulated data and CAVA database scenarios; localization is verified for both, quasi-static and dynamic scenes;

- **AV object detection**: the proposed AV object detection method is based on the multimodal initialization and model selection strategies; it demonstrates AV enhancement, efficiently combining input AV data to detect objects that are poorly represented in one of the modalities; AV object detection is demonstrated on simulated and real data from the CAVA database;

- **AV object tracking**: the AV object tracking task is addressed within the proposed framework of multimodal filtering algorithms; our approach uses all the techniques developed for the case of AV data for multimodal object localization and detection; the verification is performed on CAVA database recordings, among which we included the challenging cocktail party scenario.

8.2 Future Work

The work presented in the current thesis is inspired by biological principles of multimodal integration and contains models that implement low-level multimodal integration bases. There are numerous directions in which these models can be extended for the task of audio-visual (AV) multiobject tracking or adjusted for other types of applications. Below we outline the prospective directions of research.

Motion cues for AV integration. In our multimodal integration approach we used colocalization as the core principle, binding different modalities through the 3D object location. Dynamics information could also be included into the common unobserved parameter space. Motion cues can be extracted from both, auditory [Lu 2010] and visual [Shi 1994] data. On the one hand, this would reinforce multimodal integration by increasing the dimensionality of the parameter space and better separating the objects being observed. On the other hand, these cues could occur to be less reliable in the realistic setting, such as found in CAVA database scenarios. As mentioned in Chapter 5, the increase in observations covariance leads to significant losses in precision.

Modality-specific features. Multimodal tracking can be improved by extending the model with various modality-specific features. Low-level photometric and spectral characteristics and high-level appearance and acoustic models can be added to the audio-visual

integration framework to reinforce clustering and perform more robust tracking. However, the increase of dimensionality of the observation spaces can increase the risk of track losses.

Adding feature spaces. The conjugate clustering model that we developed allows for an arbitrary number of feature spaces. One can include detectors of different nature, such as sonar or infra-red range finders for the localization task, to improve the model performance.

Object statistical models. In this thesis we performed AV tracking under the assumption that objects are represented by feature distribution and features are independently generated. Other statistical models can be used. One possible generalization would be to consider features to be generated by a marked cluster point process, where the child point processes are governed by some potential function. In fact, conjugate mixtures is the particular case of such a model. This allows for more sophisticated object shapes and appearances. The optimization is usually performed using the variational approach, such as mean field (or force field in physics), simulated field, etc. This kind of model is good to account for sophisticated spatial scene structures with known statistical properties.

Another possibility is to consider partially observed particle diffusion models, governed by drift and diffusion fields, as those considered in Chapter 7, but without any independency assumptions. These models are potentially capable of reconstructing dependencies between spatial points and thus restituating object forms. Moreover, clustering can be performed based on regularity assumption for drift and diffusion fields. Though inference in such models is a hard problem, that requires efficient numerical approximations to be developed.

Considering other applications. The multimodal integration can be useful in various other domains, where temporal parameter inference is performed based on unaligned data arriving from physically different sensors. Examples could include tracking of chemical reaction state in biophysics, airplane tracking by sonar and turbulence data from several independent stations, disease state tracking by multiple biological factors etc.

APPENDIX A

Appendix

Sommaire

A.1 Manifold Sampling for the ITD function Pre-image. 127
A.2 Parameter Inference for Student-t Mixtures. 129

A.1 Manifold Sampling for the ITD function Pre-image.

The goal is to develop a method to sample isosurfaces of the auditory observation space (ITD) function $\mathcal{G}$ defined by (2.4):

$$\mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = \frac{1}{c}\Big(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\Big). \tag{A.1}$$

We assume the system to be fully calibrated and microphones $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ to be fixed. Thus to simplify the notation we further write $\mathcal{G}(\boldsymbol{s})$ instead of $\mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})$. The sampling technique proposed below follows the general principle of sampling method construction described in Chapter 6 of [Zhigljavsky 1991].

Let's take the orthonormal coordinate system such that its x axis goes through the two microphones $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$, from the left to the right microphone, and its center is located at $(\boldsymbol{s}_{\mathrm{M}_\ell} + \boldsymbol{s}_{\mathrm{M}_\mathrm{r}})/2$. The orientation of the y and z axes can be arbitrary. Microphone coordinates $\boldsymbol{s}_{\mathrm{M}_\ell}$ and $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$ are then $(-x_\mathrm{F}, 0, 0)$ and $(x_\mathrm{F}, 0, 0)$ respectively for some $x_\mathrm{F} \geq 0$. The locus $\mathcal{G}(\boldsymbol{s}) = g_0$ is defined by equation

$$\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\| = cg_0, \tag{A.2}$$

that can be written in the (xyz) coordinates

$$\sqrt{(x + x_\mathrm{F})^2 + y^2 + z^2} - \sqrt{(x - x_\mathrm{F})^2 + y^2 + z^2} = cg_0, \tag{A.3}$$

which after some basic algebraic transformations leads to the surface equation

$$-\frac{y^2}{x_\mathrm{F}^2 - (cg_0/2)^2} - \frac{z^2}{x_\mathrm{F}^2 - (cg_0/2)^2} + \frac{x^2}{(cg_0/2)^2} = 1. \tag{A.4}$$

A surface $\mathcal{S}$ defined by (A.4) is a hyperboloid of two sheets with microphone locations being its foci. The sign of g_0 defines which part of the hyperboloid to consider, left or

right. From (A.2) we find that $x_{\mathrm{F}}^2 \geq (cg_0/2)^2$, so by letting $a^2 = x_{\mathrm{F}}^2 - (cg_0/2)^2$ and $b^2 = (cg_0/2)^2$, the equation (A.4) becomes

$$-\frac{y^2}{a^2} - \frac{z^2}{a^2} + \frac{x^2}{b^2} = 1, \tag{A.5}$$

which is the canonical representation of a two sheet hyperboloid. Its asymptotic cone, known also in auditory analysis as "the cone of confusion" is given by

$$-\frac{y^2}{a^2} - \frac{z^2}{a^2} + \frac{x^2}{b^2} = 0. \tag{A.6}$$

We parametrize the surface (A.5) by

$$\begin{cases} x &= bt, \\ y &= a\sqrt{t^2-1}\cos\phi, \\ z &= a\sqrt{t^2-1}\sin\phi, \end{cases} \tag{A.7}$$

where $t \geq 1$ and $\phi \in [0;\ 2\pi]$. We denote $\theta = (t, \phi) \in \Theta$ the 2D surface coordinates and $\boldsymbol{s}(\theta) = (x(\theta), y(\theta), z(\theta))^\top$ the associated mapping (A.7).

The goal is to establish a distribution $P(\mathrm{d}\boldsymbol{s}) = p(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$ of some pre-defined density $p(\boldsymbol{s})$ on a hyperboloid (A.5), where $p \geq 0$ is such that $\int_{\mathcal{S}} p(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = 1$ and $\mathrm{d}\boldsymbol{s}$ is the surface measure on $\Omega = \boldsymbol{s}(\Theta)$. We make use of a well-known fact on measure transform (see [Schwarz 1993], §2 of Chapter 6)

$$\int_{\Omega} p(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = \int_{\Theta} p(\boldsymbol{s}(\theta))D(\theta)\mu(\mathrm{d}\theta), \tag{A.8}$$

where

$$D(\theta) = \sqrt{\det(\mathbf{J}\mathbf{J}^\top)}, \tag{A.9}$$

$\mathbf{J}$ is the Jacobian matrix of $\boldsymbol{s}(\theta)$ and μ is the Lebesgue measure on $\mathbb{R}^2$. In particular, for the mapping $\boldsymbol{s}(\theta)$ defined by (A.7) one has

$$D(\theta) = \sqrt{a^2b^2(t^2-1) + a^4t^2}. \tag{A.10}$$

We can define the sampling algorithm for $P(\mathrm{d}\boldsymbol{s})$ on the hyperboloid surface. For that one has to draw realizations of a random vector ζ with distribution

$$P_2(\mathrm{d}\theta) = p(\boldsymbol{s}(\theta))D(\theta)\mu(\mathrm{d}\theta), \tag{A.11}$$

and consider a random vector $\xi = \boldsymbol{s}(\zeta)$ that is distributed according to $P(\mathrm{d}\boldsymbol{s})$.

For the important case of ξ being distributed uniformly on a hyperboloid $\boldsymbol{s}(\theta)$, $\theta \in \Theta$ for parameter domain $\Theta = [1, T] \times [0, 2\pi]$, one should consider

$$\zeta \sim \alpha\sqrt{a^2b^2(t^2-1) + a^4t^2}\mathrm{d}\theta \tag{A.12}$$

with

$$\alpha = \left(2\pi|a| \int_1^T \sqrt{t^2(a^2+b^2)-b^2}\mathrm{d}t\right)^{-1}. \tag{A.13}$$

The latter integral can be readily computed, which gives the following expression

$$\alpha = \left(\pi|a|\left[T\sqrt{(a^2+b^2)T^2-b^2} - |a| - \frac{a^2}{\sqrt{a^2+b^2}}\log\frac{T+\sqrt{T^2-b^2/(a^2+b^2)}}{1+\sqrt{a^2/(a^2+b^2)}}\right]\right)^{-1}. \tag{A.14}$$

The most natural way to sample the random variable ζ by (A.12) is the acceptance-rejection method [Ermakov 1975].

A.2 Parameter Inference for Student-t Mixtures.

In Chapter 4 we mentioned that distributions $P_n^{\mathbb{F}}(\boldsymbol{f};\ \boldsymbol{\theta}_n)$ and $P_n^{\mathbb{G}}(\boldsymbol{g};\ \boldsymbol{\theta}_n)$ in mixtures (4.8) and (4.9) for modalities $\mathbb{F}$ and $\mathbb{G}$ respectively should not be necessarily Gaussian. In the most general case $P_n^{\mathbb{F}}(\boldsymbol{f};\ \boldsymbol{\theta}_n)$ is different for every n and for every modality. Though when no additional information on clusters is available, it is reasonable to consider the same distribution family for all n. But one can choose different families for modalities $\mathbb{F}$ and $\mathbb{G}$ according to statistical properties of feature detectors. In certain cases the algorithm derivation presented in Chapters 4 and 5 would not change significantly.

In this Section we show how the Student t-distribution can be used in the context of inference for conjugate mixture models. The obtained optimization scheme resembles the one developed for Gaussian mixtures.

Without loss of generality we consider the modality $\mathbb{F}$. Let's take $P_n^{\mathbb{F}}(\boldsymbol{f};\ \boldsymbol{\theta}_n)$ to belong to a Student t-distribution family for $n = 1,\dots,N$ and keep the outlier class $P_{N+1}^{\mathbb{F}}(\boldsymbol{f})$ uniform

$$P^{\mathbb{F}}(\boldsymbol{f};\ \boldsymbol{\theta}) = \sum_{n=1}^{N+1} \pi_n P_n^{\mathbb{F}}(\boldsymbol{f};\ \boldsymbol{\theta}_n), \tag{A.15}$$

$$\text{with} \quad P_n^{\mathbb{F}}(\boldsymbol{f},\boldsymbol{\theta}_n) = \mathrm{St}(\boldsymbol{f};\ \mathcal{F}(\boldsymbol{s}_n),\boldsymbol{\Sigma}_n,\vartheta_n), \quad n = 1,\dots,N, \tag{A.16}$$

$$\text{and} \quad P_{N+1}^{\mathbb{F}}(\boldsymbol{f}) = \mathcal{U}(\boldsymbol{f};\ V). \tag{A.17}$$

Here $\mathrm{St}(\boldsymbol{f};\ \mathcal{F}(\boldsymbol{s}_n),\boldsymbol{\Sigma}_n,\vartheta_n)$ is the Student t-distribution density function given by

$$\mathrm{St}(\boldsymbol{f};\ \mathcal{F}(\boldsymbol{s}_n),\boldsymbol{\Sigma}_n,\vartheta_n) = \frac{\Gamma(\frac{\vartheta_n+r}{2})}{\Gamma(\frac{\vartheta_n}{2})}\frac{|\boldsymbol{\Sigma}_n|^{-1/2}}{(\pi\vartheta_n)^{r/2}}\left(1+\frac{1}{\vartheta_n}\|\boldsymbol{f}-\mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n}\right)^{-\frac{\vartheta_n+r}{2}}, \tag{A.18}$$

where $\boldsymbol{s}_n$, $\boldsymbol{\Sigma}_n$, and ϑ_n are included into $\boldsymbol{\theta}_n$.

Following [Peel 2000] we introduce two sets of latent variables. The assignment variables $\boldsymbol{A} = \{A_1,\dots,A_m,\dots,A_M\}$ define the component of origin for each observation, the notation and the meaning are the same as in Chapter 4. The auxiliary variables

$\boldsymbol{U} = \{U_1, \ldots, U_m, \ldots, U_M\}$ are taken such that for $n = 1, \ldots, N$

$$\boldsymbol{F}_m \mid u_m, a_m = n;\ \boldsymbol{\theta}_n \ \sim\ \mathcal{N}(\mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n/u_m), \tag{A.19}$$

$$\text{and} \quad U_m \mid a_m = n;\ \boldsymbol{\theta}_n \ \sim\ \mathrm{Gamma}(\vartheta_n/2, \vartheta_n/2), \tag{A.20}$$

and $U_m, m = 1, \ldots, M$ are conditionally independent given $\boldsymbol{a}$ and the density function of the gamma distribution $\mathrm{Gamma}(\vartheta, \tilde{\vartheta})$ is given by

$$p(u;\ \vartheta, \tilde{\vartheta}) = \{\tilde{\vartheta}^{\vartheta} u^{\vartheta-1}/\Gamma(\vartheta)\} \exp(-\tilde{\vartheta} u) I_{(0,\infty)}(u), \tag{A.21}$$

where $\vartheta, \tilde{\vartheta} > 0$, the indicator function $I_{(0,\infty)}(u) = 1$ for $u > 0$ and is zero elsewhere and $\Gamma(\vartheta)$ is the Gamma function. Then $\boldsymbol{F}_m$ are distributed by the Student law (A.18). We adopt the following standard convention: upper case letters for random variables ($\boldsymbol{A}$ and $\boldsymbol{U}$) and lower case letters for their realizations ($\boldsymbol{a}$ and $\boldsymbol{u}$).

The penalization term $H_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ of the general KP algorithm is given by

$$\begin{aligned} H_{\mathcal{F}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = -\frac{1}{P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta})} \sum_{m=1}^{M} \Bigg[\sum_{n=1}^{N+1} \alpha_{mn}(\boldsymbol{\theta}^{(q)}) \Big(\log \pi_n + \frac{\vartheta_n}{2} \log \frac{\vartheta_n}{2} - \frac{1}{2} \log |\boldsymbol{\Sigma}_n| + \\ + \frac{\vartheta_n + r - 2}{2} \Big(\psi((\vartheta_n^{(q)} + r)/2) - \log((\vartheta_n^{(q)} + \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q)})\|^2_{\boldsymbol{\Sigma}_n^{(q)}})/2) \Big) - \\ - \log \Gamma\left(\frac{\vartheta}{2}\right) - \frac{r}{2} \log(2\pi) - \frac{1}{2} \gamma_{mn}(\boldsymbol{\theta}^{(q)}) (\vartheta + \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n}) \Big) + \\ + \alpha_{m,N+1}(\boldsymbol{\theta}^{(q)}) \big(\log \pi_{N+1} + \log \mathcal{U}(\boldsymbol{f}_m;\ V) \big) \Bigg], \end{aligned} \tag{A.22}$$

where α_{mn} and γ_{mn} denote posterior probabilities $\alpha_{mn}(\boldsymbol{\theta}) = P(A_m = n | \boldsymbol{f}_m; \boldsymbol{\theta})$ and $\gamma_{mn}(\boldsymbol{\theta}) = P(U_m | \boldsymbol{f}_m, A_m = n; \boldsymbol{\theta})$ as functions of parameters, and

$$\psi(t) = \frac{\partial \Gamma(t)}{\partial t} \frac{1}{\Gamma(t)} \tag{A.23}$$

is the Digamma function.

The expression for α_{mn} can be derived straightforwardly from Bayes' theorem, $\forall n = 1 \ldots N$:

$$\alpha_{mn}(\boldsymbol{\theta}) = \frac{\pi_n P_n^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta}_n)}{P^{\mathbb{F}}(\boldsymbol{f}_m;\ \boldsymbol{\theta})} \quad = \frac{\pi_n \mathrm{St}(\boldsymbol{f}_m; \mathcal{F}(\boldsymbol{s}_n), \boldsymbol{\Sigma}_n, \vartheta_n)}{\sum_{i=1}^{N} \pi_i \mathrm{St}(\boldsymbol{f}_m; \mathcal{F}(\boldsymbol{s}_i), \boldsymbol{\Sigma}_i, \vartheta_n) + V^{-1} \pi_{N+1}}, \tag{A.24}$$

and $\alpha_{m,N+1}(\boldsymbol{\theta}) = 1 - \sum_{n=1}^{N} \alpha_{mn}(\boldsymbol{\theta})$. The derivation for γ_{mn} is presented in [Peel 2000]

$$\gamma_{mn}(\boldsymbol{\theta}) = \frac{\vartheta_n + r}{\vartheta_n + \|\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n)\|^2_{\boldsymbol{\Sigma}_n}}. \tag{A.25}$$

The update expressions for $\{\pi_1, \ldots, \pi_{N+1}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N\}$ are similar to those derived in Chapter 4:

$$\pi_n^{(q+1)} = \frac{1}{M} \sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}), \tag{A.26}$$

$$\boldsymbol{\Sigma}_n^{(q+1)} = \frac{\sum_{m=1}^{M} \kappa_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \left(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q+1)})\right)\left(\boldsymbol{f}_m - \mathcal{F}(\boldsymbol{s}_n^{(q+1)})\right)^\top}{\sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})}, \tag{A.27}$$

where we introduced

$$\alpha_{mn}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = (1 - h_q)\alpha_{mn}(\boldsymbol{\theta}) + h_q \alpha_{mn}(\tilde{\boldsymbol{\theta}}), \tag{A.28}$$

$$\text{and} \quad \kappa_{mn}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = (1 - h_q)\alpha_{mn}(\boldsymbol{\theta})\gamma_{mn}(\boldsymbol{\theta}) + h_q \alpha_{mn}(\tilde{\boldsymbol{\theta}})\gamma_{mn}(\tilde{\boldsymbol{\theta}}). \tag{A.29}$$

Moreover, the equation for optimal tying parameters $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ resembles that from Chapter 4, only the weights are adjusted so that the first part of (4.38) becomes

$$\bar{\kappa}_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})(\bar{\boldsymbol{f}}_n - \mathcal{F}(\boldsymbol{s}_n^{(q+1)}))^\top \left(\boldsymbol{\Sigma}_n^{(q+1)}\right)^{-1} \mathcal{F}'(\boldsymbol{s}_n^{(q+1)}), \tag{A.30}$$

where we denoted $\mathcal{F}'$ and $\mathcal{G}'$ the Jacobian matrices of $\mathcal{F}$ and $\mathcal{G}$ respectively and

$$\bar{\kappa}_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) = \sum_{m=1}^{M} \kappa_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}), \tag{A.31}$$

$$\text{and} \quad \bar{\boldsymbol{f}}_n = \bar{\kappa}_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)})^{-1} \sum_{m=1}^{M} \kappa_{mn}(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \boldsymbol{f}_m. \tag{A.32}$$

If the 'degrees of freedom' parameters $\vartheta_1, \ldots, \vartheta_N$ are fixed to some values, the algorithms presented in Chapters 4 and 5 would require only minor changes, so we suppose that their performance would be essentially the same. Otherwise, if one considers $\vartheta_1, \ldots, \vartheta_{N+1}$ as parameters to estimate, their optimal values should be found from an equation that does not admit a closed form solution. In the case of the efficient EM algorithm described in Chapter 5 (with $h_q \equiv 1$), this equation is given below

$$\frac{1}{\sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q)})} \sum_{m=1}^{M} \alpha_{mn}(\boldsymbol{\theta}^{(q)}) \left(\log \gamma_{mn}(\boldsymbol{\theta}^{(q)}) - \gamma_{mn}(\boldsymbol{\theta}^{(q)})\right) + 1 =$$

$$= \log\left((\vartheta_n^{(q)} + r)/2\right) - \psi\left((\vartheta_n^{(q)} + r)/2\right) - (\log(\vartheta_n/2) - \psi(\vartheta_n/2)). \tag{A.33}$$

We note that the left-hand side of (A.33) is non-positive, as soon as $\log t \leq t - 1$. At the same time, the function $\varphi(t) = \log t - \psi(t)$ is strictly decreasing and strictly convex for $t > 1$. Indeed, one can use the expression

$$\psi(t) = \log t - \int_0^1 \int_0^1 \frac{1 - x}{(1 - xy)(-\log(xy))} (xy)^{t-1} \mathrm{d}x\mathrm{d}y, \tag{A.34}$$

to show that $\varphi'(t) < 0$ and $\varphi''(t) > 0$ for $t > 1$. This means that the optimization problem (A.33) is convex. Moreover, the optimization domain can be restrained as soon as $\vartheta_n^{(q+1)} < \vartheta_n^{(q)} + r$. Thus one can consider this restrained domain and apply efficient techniques that solve covex optimization problems [Polyak 1987] to find the optimal 'degrees of freedom' parameters $\vartheta_1, \ldots, \vartheta_N$.

Bibliography

[Akaike 1973] H. Akaike. *Information theory as an extension of the maximum likelihood principle*. In B.N. Petrov and F. Csaki, editeurs, Second International Symposium on Information Theory, pages 267–281. Akademiai Kiado, Budapest, 1973. 103

[Allen 1995] P.K. Allen. *Integrating Vision and Touch for Object Recognition Tasks*. In Luo and Kay, editeurs, Multisensor Integration and Fusion for Intelligent Machines and Systems, pages 407–440. Ablex Publishing Corporation, 1995. 44, 115

[AMI] *AMI meeting corpus*. http://corpus.amiproject.org/. 17

[Anastasio 2000] T. J. Anastasio, P. E. Patton and K. E. Belkacem-Boussaid. *Using Bayes' Rule to Model Multisensory Enhancement in the Superior Colliculus*. Neural Computation, vol. 12, no. 5, pages 1165–1187, 2000. 3, 43, 115

[Arnaud 2008] E. Arnaud, H. Christensen, Y.C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, F. Forbes and R. Horaud. *The CAVA Corpus: Synchronized Stereoscopic and Binaural Datasets with Head Movements*. In Proc. of ACM/IEEE Tenth International Conference on Multimodal Interfaces, October 2008. 8, 17

[Bailly-Bailliére 2003] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz and J.-P. Thiran. *The BANCA Database and Evaluation Protocol*. In AVBPA, http://www.ee.surrey.ac.uk/CVSSP/banca/, 2003. 17

[Barzelay 2007] Z. Barzelay and Y. Schechner. *Harmony in Motion*. In CVPR, 2007. 6

[Beal 2003] M. Beal, N. Jojic and H. Attias. *A graphical model for audiovisual object tracking*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pages 828–836, 2003. 5, 6, 24, 44, 116

[Benveniste 1990] A. Benveniste, M. Métivier and P. Priouret. Adaptive algorithms and stochastic approximations, volume 22 of *Applications of Mathematics*. Springer Verlag, Berlin, Heidelberg, New York, 1990. 118

[Bernardin 2007] K. Bernardin and R. Stiefelhagen. *Audio-Visual Multi-Person Tracking and Identification for Smart Environments*. In ACM Multimedia, Augsburg, Germany, 2007. 6

[Bhat 2008] P. Bhat, B. Curless, M.F. Cohen and C.L. Zitnick. *Fourier Analysis of the 2D Screened Poisson Equation for Gradient Domain Problems*. In Proc. of ECCV (2), volume 5303, pages 114–128. Springer, 2008. 27

[Biernacki 2000] C. Biernacki, G. Celeux and G. Govaert. *Assessing a mixture model for clustering with the integrated completed likelihood*. Trans. on PAMI, vol. 22, pages 719–725, 2000. 103

[Biernacki 2003] C. Biernacki, G. Celeux and G. Govaert. *Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models*. Comp. Statistics and Data Analysis, vol. 41, pages 561–575, 2003. 95

[Bishop 2006] C.M. Bishop. Pattern recognition and machine learning. Springer, 2006. 48

[Boyles 1983] R.A. Boyles. *On the convergence of EM algorithms*. Journal of the Royal Statistical Society: Series B, vol. 45, no. 1, pages 47–50, 1983. 66

[Brunelli 2007] R. Brunelli, B. Alessio, P. Chippendale, O. Lanz, M. Omologo, S. Piergiorgio and F. Tobia. *A Generative Approach to Audio-Visual Person Tracking*. In R. Stiefelhagen and J. Garofolo, editeurs, Multimodal Technologies for Perception of Humans, LNCS. Springer, Berlin / Heidelberg, 2007. 6

[Burr 2006] D. Burr and D. Alais. *Combining visual and auditory information*. Progress in Brain Research, vol. 155, no. 2, pages 243–258, 2006. 3

[Castellanos 1999] J.A. Castellanos and J.D Tardos. Mobile robot localization and map building: A multisensor fusion approach. Kluwer, Boston, MA, 1999. 44, 115

[Celeux 1996] G. Celeux and G. Soromenho. *An entropy criterion for assessing the number of clusters in a mixture model*. Journal of Classification, vol. 13, pages 195–212, 1996. 103

[Celeux 2003] G. Celeux, F. Forbes and N. Peyrard. *EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation*. Pattern Recognition, vol. 36, no. 1, pages 131–144, 2003. 47

[Checka 2004] N. Checka, K. Wilson, M. Siracusa and T. Darrell. *Multiple person and speaker activity tracking with a particle filter*. In Proc. of IEEE Conference on Acoustics, Speech, and Signal Processing, pages 881–884. IEEE, 2004. 5, 6, 7, 24, 44, 116

[Chen 2004] Y. Chen and Y. Rui. *Real-time speaker tracking using particle filter sensor fusion*. Proceedings of IEEE, vol. 92, no. 3, pages 485–494, 2004. 5, 6, 44, 116

[Chen 2005] Lei Chen, R. Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, Yingen Xiong, David Mcneill, Ronald Tuttle and Thomas Huang. *VACE multimodal meeting corpus*. In MLMI, 2005. 17

[Chrétien 2000] S. Chrétien and A. Hero. *Kullback Proximal Algorithms for Maximum Likelihood Estimation*. IEEE Transactions on Information Theory, vol. 46, pages 1800–1810, 2000. 45, 54

[Chrétien 2008] S. Chrétien and Alfred Hero. *On EM Algorithms and Their Proximal Generalizations*. ESAIM. P&S, vol. 12, pages 308–326, 2008. 45, 49

[Christensen 2007] H. Christensen, N. Ma, S.N. Wrigley and J. Barker. *Integrating Pitch and Localisation Cues at a Speech Fragment Level*. In Proc. of Interspeech, pages 2769–2772, 2007. 15, 16, 55

[Ciuperca 2003] G. Ciuperca, A. Ridolfi and J. Idier. *Maximum likelihood estimator for normal mixtures*. Scandinavian Journal of Statistics, vol. 30, no. 1, pages 45–59, 2003. 108

[Coiras 2007] E. Coiras, F. Baralli and B. Evans. *Rigid Data Association for Shallow Water Surveys*. IET Radar, Sonar and Navigation, vol. 1, no. 5, pages 354–361, October 2007. 44

[Colonius 2001] H. Colonius and P. Arndt. *A two-stage model for visual and auditory interaction in saccadic latencies*. Perception & Psychophysics, vol. 63, pages 126–147, 2001. 2

[Cooke 1993] M. Cooke. Modelling auditory processing and organisation. Cambridge University Press, New York, NY, USA, 1993. 16

[Courchay 2010] J. Courchay, A. Dalalyan, R. Keriven and P. Sturm. *A Global Camera Network Calibration Method with Linear Programming*. In Proc. of Int. Symp. on 3D Data Processing, Visualization and Transmission, 2010. 23

[Cui 2008] J. Cui, H. Zha, H. Zhao and R. Shibasaki. *Multi-modal Tracking of People Using Laser Scanners and Video Camera*. Image Vision Comput., vol. 26, no. 2, pages 240–252, 2008. 24

[Dempster 1977] A. P. Dempster, N. M. Laird and D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*. Journal of the Royal Statistical Society: Series B, vol. 39, no. 1, pages 1–38, 1977. 30, 45, 66

[Dibiase 2001] J. Dibiase, H. Silverman and M. Brandstein. *Robust localization in reverberant rooms*. In M. Brandstein and D. Ward, editeurs, Microphone Arrays: Signal Processing Techniques and Applications, chapitre 8. Springer, 2001. 55

[Doucet 2001] A. Doucet, N. de Freitas and N. Gordon, editeurs. Sequential Monte Carlo methods in practice. Statistics for Emgineering and Information Science. Springer, 2001. 116

[Ermakov 1975] S.M. Ermakov. Die Monte-Carlo methode und verwandte fragen (deutsch). VEB Deutscher Verlag der Wissenschaften, Berlin, 1975. 129

[Ernst 2002] M. O. Ernst and M. S. Banks. *Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion.* Nature, vol. 415, pages 429–433, 2002. 43, 115

[Faller 2004] C. Faller and J. Merimaa. *Sound Localization in Complex Listening Situations: Selection of binaural Cues Based on Interaural Coherence*. J. Acoust. Soc. Amer, vol. 116, no. 5, pages 3075–3089, 2004. 16

[Faugeras 1993] O. D. Faugeras. Three dimensional computer vision: A geometric viewpoint. MIT Press, Boston, 1993. 55

[Fisher III 2001] J. W. Fisher III, T. Darrell, W. T. Freeman and P. Viola. *Learning Joint Statistical Models for Audio-Visual Fusion Segregation*. In Proceedings of Annual Conference on Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 2001. 44

[Fisher III 2004] J. W Fisher III and T. Darrell. *Speaker association with signal-level audiovisual fusion*. IEEE Transactions on Multimedia, vol. 6, no. 3, pages 406–413, 2004. 6, 44

[Forsyth 2003] D.A. Forsyth and J. Ponce. Computer vision – a modern approach. Prentice Hall, New Jersey, 2003. 54

[Garg 2003] A. Garg, V. Pavlović and J. Rehg. *Boosted learning in dynamic Bayesian networks for multimodal speaker detection*. Proceedings of IEEE, vol. 91, no. 9, pages 1355–1369, 2003. 4

[Gassiat 2002] E. Gassiat. *Likelihood ratio inequalities with applications to various mixtures*. Ann. I. H. Poincare, vol. 38, no. 6, pages 897–906, 2002. 108

[Gatica-Perez 2007] D. Gatica-Perez, G. Lathoud, J.-M. Odobez and I. McCowan. *Audio-visual probabilistic tracking of multiple speakers in meetings*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 2, pages 601–616, 2007. 5, 6, 7, 24, 44, 116

[GEM] *GEMEP: Geneva multimodal emotion portrayals corpus*. http://www.affective-sciences.org/gemep/. 17

[Glasberg 1990] B.R. Glasberg and B.C.J. Moore. *Derivation of Auditory Filter Shapes From Notched-Noise Data*. Hearing Research, vol. 47, pages 103–138, 1990. 15

[Gould 2008] S. Gould, P. Baumstarck, M. Quigley, A.Y. Ng and D. Koller. *Integrating Visual and Range Data for Robotic Object Detection*. In ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2), 2008. 24

[Grenander 1994] U. Grenander and M. Miller. *Representations of Knowledge in Complex Systems*. J. of Royal Stat. Soc., vol. 56, no. 4, pages 549–603, 1994. 99, 114, 117, 119

[GRI] *GRID Audio-Visual sentence corpus*. http://www.dcs.shef.ac.uk/spandh/gridcorpus/. 17

[Hall 2004] D. L. Hall and S. A. H. McMullen. Mathematical techniques in multisensor data fusion. Artech House, New York, 2004. 43

[Hansard 2007] M. Hansard and R. Horaud. *Patterns of Binocular Disparity for a Fixating Observer*. In Proc. of Second International Symposium of Advances in Brain, Vision, and Artificial Intelligence, pages 308–317. Springer, 2007. 13, 55, 74

[Hansard 2008] M. Hansard and R. Horaud. *Cyclopean Geometry of Binocular Vision*. Journal of the Optical Society of America A, vol. 25, no. 9, page 2357 2369, September 2008. 12, 13, 55, 74

[Harris 1988] C. Harris and M. Stephens. *A Combined Corner and Edge Detector*. In Proc. of Fourth Alvey Vision Conference, pages 147–151, 1988. 13

[Hartley 2003] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, Cambridge, UK, 2003. 12, 55, 74

[Haykin 2005] S. Haykin and Z. Chen. *The Cocktail Party Problem*. Neural Computation, vol. 17, pages 1875–1902, 2005. 43, 115

[Hazen 2004] T. Hazen, E. Saenko, C. La and J. Glass. *A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development and Initial Experiments*. In ICMI, 2004. 6, 17

[Heckmann 2002] M. Heckmann, F. Berthommier and K. Kroschel. *Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition*. EURASIP Journal on Applied Signal Processing, vol. 11, pages 1260–1273, 2002. 4, 5, 44

[Hofbauer 2004] M. Hofbauer, S.M. Wuerger, G.F. Meyer, F. Roehrbein, K. Schill and C. Zetzsche. *Catching audio-visual mice: predicting the arrival time of audio-visual motion signals*. Cognitive, Affective, & Behavioral Neuroscience, vol. 4, pages 241–250, 2004. 2

[Hospedales 2007] T.M. Hospedales, J.J.Cartwright and S. Vijayakumar. *Structure Inference for Bayesian Multisensory Perception and Tracking*. In Proc. of IJCAI, pages 2122–2128, 2007. 5, 6, 116

[Hospedales 2008] T.M. Hospedales and S. Vijayakumar. *Structure Inference for Bayesian Multisensory Scene Understanding*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, pages 2140–2157, 2008. 5, 6, 24, 44, 116

[Howard 1966] I.P. Howard and W.B. Templeton. Human spatial orientation. Wiley, London, 1966. 3

[Jacobsen 2006] M. Jacobsen. Point process theory and applications. marked point and piecewise deterministic processses. Probability and Its Applications. Birkhäuser, 2006. 99, 114, 117, 119

[Jordan 1998] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola and L.K. Saul. *An introduction to variational methods for graphical models*. In M.I. Jordan, editeur, Learning in Graphical Models, pages 105–162. Kluwer Academic, 1998. 47, 62

[Joshi 1999] R. Joshi and A.C. Sanderson. Multisensor fusion: A minimal representation framework. World Scientific, 1999. 44, 115

[Kadunce 2001] D.C. Kadunce, J.W. Vaughan, M.T. Wallace and B.E. Stein. *The influence of visual and auditory receptive field organization on multisensory integration in the superior colliculus*. Experimental Brain Research, vol. 139, pages 303–310, 2001. 2, 3

[Kalman 1961] R.E. Kalman and R.S. Bucy. *New results in linear filtering and prediction theory*. Trans. ASME Ser. D.J. Basic Eng., vol. 83, pages 95–108, 1961. 116

[Katkovnik 2008] V. Katkovnik and V. Spokoiny. *Spatially Adaptive Estimation via Fitted Local Likelihood Techniques*. IEEE Trans. Signal Proc., vol. 56, pages 873–886, 2008. 97

[Khalidov 2008a] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud and R. Horaud. *Detecion and Localization of 3D Audio-Visual Objects Using Unsupervised Clustering*. In Proc. of ICMI, 2008. 8

[Khalidov 2008b] V. Khalidov, F. Forbes, H. Miles, E. Arnaud and R. Horaud. *Audio-Visual Clustering for Multiple Speaker Localization*. In MLMI, pages 86–97, 2008. 8

[Khalidov 2010] V. Khalidov, F. Forbes and R. Horaud. *Conjugate Mixture Models for clustering Multimodal Data*. Neural Computation, vol. 6, no. 1-2, pages 48–83, 2010. 8

[Khan 2005] Z. Khan, T. Balch and F. Dellaert. *MCMC-based particle filtering for tracking a variable number of interacting targets*. IEEE Trans. on PAMI, vol. 27, no. 11, pages 1805–1918, 2005. 116

[King 2004] A. J. King. *The superior colliculus*. Current Biology, vol. 14, pages 335–338, 2004. 43, 115

[King 2005] A. J. King. *Multisensory integration: strategies for synchronization*. Current Biology, vol. 15, pages 339–341, 2005. 43, 115

[Knill 2007] D.C. Knill. *Bayesian Models of Sensory Cue Integration*. In K. Doya, S. Ishii, A. Pouget and R.P.N. Rao, editeurs, Bayesian Brain. Probabilistic approaches to neural coding, pages 189–206. 2007. 3

[Kushal 2006] A. Kushal, M. Rahurkar, L. Fei-Fei, J. Ponce and T. Huang. *Audio-Visual Speaker Localization Using Graphical Models*. In Proc. of the Eighteenth International Conference on Pattern Recognition, pages 291–294, 2006. 5, 6, 44, 116

[Laptev 2005] Ivan Laptev. *On Space-Time Interest Points*. International Journal of Computer Vision, vol. 64, no. 2-3, pages 107–123, 2005. 13

[Lathoud 2004] G. Lathoud, J.-M. Odobez and D. Gatica-Perez. *AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking*. In MLMI, pages 182–195, 2004. 17

[Lefebvre 2001] T. Lefebvre, H. Bruyninckx and J. de Schutter. *Kalman filters for non-linear systems: a comparison of performance*. Int. J. of Control, vol. 77, pages 639–653, 2001. 116

[Lewald 2003] J. Lewald and R. Guski. *Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli*. Cognitive Brain Research, vol. 16, pages 468–478, 2003. 3

[Lu 2010] Y.-C. Lu and M. Cooke. *Motion Strategies for Binaural Localisation of Speech Sources in Azimuth and Distance by Artificial Listeners*. Speech Communication, 2010. in press. 125

[Luo 2002] R.C. Luo, Y. Chih-Chen and L.S. Kuo. *Multisensor fusion and integration: approaches, applications, and future research directions*. Sensors Journal, IEEE, vol. 2, no. 2, pages 107–119, 2002. 115

[M2V] *M2VTS database*. http://www.tele.ucl.ac.be/PROJECTS/M2VTS/. 17

[Ma 2000] J. Ma, L. Xu and M. Jordan. *Asymptotic Convergence Rate of the EM Algorithm for Gaussian Mixtures*. Neural Computation, vol. 12, no. 12, pages 2881–2907, 2000. 95

[Mahler 2005] R. Mahler. *A General Theory of Multitarget Extended Kalman filters*. In Proc. of SPIE, volume 5809, 2005. 116

[Majumder 2001] S. Majumder, S. Scheding and H. Durrant-Whyte. *Multisensor data fusion for underwater navigation*. Robotics and Autonomous Systems, vol. 35, pages 97–108, 2001. 44

[Mandel 2007] M.I. Mandel, D.P.W. Ellis and T. Jebara. *An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments*. In B. Schölkopf, J. Platt and T. Hoffman, editeurs, Advances in Neural Information Processing Systems 19, pages 953–960. MIT Press, Cambridge, MA, 2007. 16

[McCowan 2003] I. McCowan, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner and H. Bourlard. *Modelling Human Interaction in Meetings*. In ICASSP, 2003. 17

[McCowan 2008] I. McCowan, M. Lincoln and I. Himawan. *Microphone Array Shape Calibration in Diffuse Noise Fields*. IEEE Trans. on Audio, Speech and Language Processing, vol. 16, no. 3, pages 666–670, 2008. 23, 24

[McGurk 1976] H. McGurk and J. MacDonald. *Hearing lips and seeing voices*. Nature, vol. 264, pages 746–748, 1976. 3

[McLachlan 2000] G. J. McLachlan and D. Peel. Finite mixture models. Wiley, New-York, 2000. 45

[McLachlan 2007] G.J. McLachlan and T. Krishnan. The EM algorithm and extensions. Wiley, New York, second édition, 2007. 30, 66

[Meila 2001] M. Meila and D. Heckerman. *An Experimental Comparison of Model-based Clustering Methods*. Machine Learning, vol. 42, pages 9–29, 2001. 95

[Messer 1999] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre. *XM2VTSDB: the Extended M2VTS Database*. In AVBPA, 1999. 17

[Meyer 2001] G. Meyer and S. Wuerger. *Cross-modal integration of auditory and visual motion signals*. NeuroReport, vol. 12, no. 11, pages 2557–2600, 2001. 2

[Meyer 2005] G.F. Meyer, S.M. Wuerger, F. Röhrbein and C. Zetzsche. *Low-level integration of auditory and visual motion signals requires spatial co-localisation*. Experimental Brain Research, vol. 166, pages 538–547, 2005. 3

[Michel 2007] M. Michel, J. Ajot and J. Fiscus. *The NIST Meeting Room Corpus 2 Phase 1*. In MLMI, 2007. 17

[Mitchell 2007] H.B. Mitchell. Multi-sensor data fusion. Springer, Berlin Heidelberg, 2007. 43

[Mostefa 2008] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin and C. Rochet. *The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms*, 2008. 17

[Naus 2004] H.W.L. Naus and C.V. van Wijk. *Simultaneous Localization of Multiple Emitters*. IEE Proceedings Radar Sonar and Navigation, vol. 151, pages 65–70, 2004. 44

[Nefian 2002] A. V. Nefian, L. Liang, X. Pi, X. Liu and K. Murphy. *Dynamic Bayesian networks for audio-visual speech recognition*. EURASIP Journal of Applied Signal Processing, vol. 2002, no. 1, pages 1274–1288, 2002. 44

[Nevelson 1976] M.B. Nevelson and R.Z. Khasminskii. Stochastic approximation and recursive estimation. Amer. Math. Soc., 1976. Translated from Russian. 118

[Nickel 2005] K. Nickel, T. Gehrig, R. Stiefelhagen and J. McDonough. *A Joint Particle Filter for Audio-Visual Speaker Tracking*. In Proc. of ICMI, pages 61–68, New York, NY, USA, 2005. ACM. 5, 6, 7, 24

[Pannetier 2008] B. Pannetier, J. Dezert and E. Pollard. *Improvement of multiple ground targets tracking with GMTI sensor and fusion of identification attributes*. In IEEE Aerospace Conf., pages 1–13, Big Sky, MT, 2008. 115

[Patterson 1992] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M. Allerhand. *Complex Sounds and Auditory Images*. In Y. Cazals, L. Demany and K. Horner, editeurs, Auditory Physiology and Perception, pages 429–446. Pergamon, Oxford, 1992. 15

[Patterson 2002] E.K. Patterson, S. Gurbuz, Z. Tufekci and J.N. Gowdy. *CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research*. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), http://www.ece.clemson.edu/speech/cuave.htm, 2002. 17

[Peel 2000] D. Peel and G.J. McLachlan. *Robust mixture modelling using the t distribution*. Statistics and Computing, vol. 10, pages 339–348, 2000. 129, 130

[Perez 2004] P. Perez, J. Vermaak and A. Blake. *Data Fusion for Visual Tracking with Particles*. Proceedings of IEEE, vol. 92, no. 3, pages 495–513, 2004. 5, 6, 44

[Polyak 1987] B.T. Polyak. Introduction to optimization. New York: Optimization Software, Publications Division, 1987. 45, 49, 52, 58, 69, 73, 132

[Pouget 2002a] A. Pouget, S. Deneve and Duhamel J.-R. *A Computational Perspective on the Neural Basis of Multisensory Spatial Representations*. Nature Review Neuroscience, vol. 3, pages 741–747, 2002. 43, 115

[Pouget 2002b] A. Pouget, J.C. Ducom, J. Torri and D. Bavelier. *Multisensory spatial representations in eye-centered coordinates for reaching*. Cognition, vol. 83, pages 1–11, 2002. 3

[Raykar 2004] V.C. Raykar and R. Duraiswami. *Automatic Position Calibration of Multiple Microphones*. In Proc. of ICASSP, pages 69–72, 2004. 23, 24

[Rissanen 1978] J. Rissanen. *Modelling by the shortest data description*. Automatica, vol. 14, pages 465–471, 1978. 103

[Rozovskii 1990] B.L. Rozovskii. Stochastic evolution systems. Mathematics and its Applications. Kluwer Academic Publishers, 1990. 117

[Schwarz 1978] G. Schwarz. *Estimating the dimension of a model*. Annals of Statistics, vol. 6, pages 461–464, 1978. 103

[Schwarz 1993] L. Schwarz. Analyse iv. applications à la théorie de la mesure. Hermann, Paris, 1993. 128

[Shao 2008] X. Shao and J. Barker. *Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment*. Speech Communication, vol. 50, no. 4, pages 337–353, April 2008. 44

[Shi 1994] J. Shi and C. Tomasi. *Good Features to Track*. In CVPR, pages 593–600, 1994. 125

[Slaney 1993] M. Slaney. *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank*. Rapport technique, Apple Computer Inc., 1993. 16

[Smith 2006] D. Smith and S. Singh. *Approaches to Multisensor Data Fusion in Target Tracking: A Survey*. IEEE Transactions on Knowledge and Data Engineering, vol. 18, pages 1041–4347, 2006. 43, 44

[Spall 2003] J.C. Spall. Introduction to stochastic searchand optimization: Estimation, simulation and control. Wiley, 2003. 34, 54, 58

[Spence 2004] C. Spence and J. Driver. Crossmodal space and crossmodal attention. Oxford Univ. Press, Oxford, 2004. 2

[Spinello 2008] L. Spinello, R. Triebel and R. Siegwart. *Multimodal Detection and Tracking of Pedestrians in Urban environments with Explicit Groung Plane Extraction*. In IROS, pages 1823–1829, 2008. 24

[Stanford 2007] T.R. Stanford and B.E. Stein. *Superadditivity in multisensory integration: putting the computation in context*. Neuroreport, vol. 18, no. 8, pages 787–792, 2007. 2, 3

[Stein 1988] B.E. Stein, W.S. Huneycutt and M.A. Meredith. *Neurons and behavior: the same rules of multisensory integration apply*. Brain Research, vol. 448, no. 2, pages 355–358, 1988. 3

[Stein 1993] B.E. Stein and M.A. Meredith. The merging of the senses. MIT Press, Cambridge, MA, 1993. 3

[Stein 2008] B.E. Stein and T.R. Stanford. *Multisensory integration: current issues from the perspective of the single neuron*. Nature Reviews Neuroscience, vol. 9, pages 255–266, 2008. 2, 3

[Svoboda 2005] T. Svoboda, D. Martinec and T. Pajdla. *A Convenient Multi-Camera Self-Calibration for Virtual Environments*. PRESENCE: Teleoperators and Virtual Environments, vol. 14, no. 4, pages 407–422, August 2005. 23

[Ueda 1998] N. Ueda and R. Nakano. *Deterministic Annealing EM Algorithm*. Neural Networks, vol. 11, pages 271–282, 1998. 95

[van der Vaart 2004] A.W. van der Vaart. Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2004. 109

[van Kampen 2007] N.G. van Kampen. Stochastic processes in physics and chemistry. North Holland, 3 édition, 2007. 90

[Vermaak 2001] J. Vermaak, M. Ganget, A. Blake and P. Pérez. *Sequential Monte Carlo fusion of sound and vision for speaker tracking*. In Proceedings of the Eighth International Conference on Computer Vision, pages 741–746. IEEE, 2001. 6, 24

[Wan 2000] E.A. Wan and R. van der Merwe. *The unscented Kalman filter for non-linear estimation*. In Proc. Symp. on Adaptive Syst. for Sign. Proc., Comm. nd Control, Lake Louise, Alberta, Canada, 2000. 116

[Wang 2006] D. Wang and G. J. Brown, editeurs. Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE Press, September 2006. 5, 14, 22, 55

[Wasan 1969] M.T. Wasan. Stochastic approximation. Cambridge Univ. Press, 1969. 118

[Wilson 2001] K. Wilson, N. Checka, D. Demirdjian and T. Darrell. *Audio-video array source separation for perceptual user interfaces*. In Proceedings of the 2001 Workshop on Perceptive User Interfaces, pages 1–7, New York, NY, USA, 2001. ACM Press. 7

[Xu 1997] Lei Xu. *Comparative Analysis on Convergence Rates of The EM Algorithm and Its Two Modifications for Gaussian Mixtures*. Neural Process. Lett., vol. 6, no. 3, pages 69–76, 1997. 33

[Yao 2008] J. Yao and J.-M. Odobez. *Multi-Camera Multi-Person 3D Space Tracking with MCMC in Surveillance Scenarios*. In M2SFA2, Marseille, France, 2008. 119

[Yguel 2008] M. Yguel, O. Aycard and C. Laugier. *Efficient GPU-based Construction of Occupancy Grids Using Several Laser Range-finders*. Int. J. of Autonomous Vehicles, vol. 6, no. 1-2, pages 48–83, 2008. 24

[Zeng 2007] Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth and S. Levinson. *Audio-visual affect recognition*. IEEE Transactions on Multimedia, vol. 9, no. 2, pages 424–428, 2007. 6

[Zhigljavsky 1991] A.A. Zhigljavsky. Theory of global random search. Kluwer Academic Publishers, 1991. 45, 73, 74, 77, 127

[Zhigljavsky 2008] A.A. Zhigljavsky and A. Žilinskas. Stochastic global optimization. Springer, 2008. 32, 45, 69, 82, 94, 95

[Zotkin 2002] D. N. Zotkin, R. Duraiswami and L. S. Davis. *Joint audio-visual tracking using particle filters*. EURASIP Journal on Applied Signal Processing, vol. 11, pages 1154–1164, 2002. 6, 24

Modèles de Mélanges Conjugués pour la Modélisation de la Perception Visuelle et Auditive

Résumé:
Dans cette thèse, nous nous intéressons à la modélisation de la perception audio-visuelle avec une tête robotique. Les problèmes associés, notamment la calibration audio-visuelle, la détection, la localisation et le suivi d'objets audio-visuels sont étudiés. Une approche spatio-temporelle de calibration d'une tête robotique est proposée, basée sur une mise en correspondance probabiliste multimodale des trajectoires. Le formalisme de modèles de mélange conjugué est introduit ainsi qu'une famille d'algorithmes d'optimisation efficaces pour effectuer le regroupement multimodal. Un cas particulier de cette famille d'algorithmes, notamment l'algorithme EM conjugué, est amélioré pour obtenir des propriétés théoriques intéressantes. Des méthodes de détection d'objets multimodaux et d'estimation du nombre d'objets sont développées et leurs propriétés théoriques sont étudiées. Enfin, la méthode de regroupement multimodal proposée est combinée avec des stratégies de détection et d'estimation du nombre d'objets ainsi qu'avec des techniques de suivi pour effectuer le suivi multimodal de plusieurs objets. La performance des méthodes est démontrée sur des données simulées et réelles issues d'une base de données de scénarios audio-visuels réalistes (base de données CAVA).

Mots clés : modèles de mélanges conjugués, analyse audio-visuel de scène, calibration audio-visuelle, détection multimodale d'objets, suivi multimodal d'objets

Conjugate Mixture Models for the Modelling of Visual and Auditory Perception

Abstract:
In this thesis, the modelling of audio-visual perception with a head-like device is considered. The related problems, namely audio-visual calibration, audio-visual object detection, localization and tracking are addressed. A spatio-temporal approach to the head-like device calibration is proposed based on probabilistic multimodal trajectory matching. The formalism of conjugate mixture models is introduced along with a family of efficient optimization algorithms to perform multimodal clustering. One instance of this algorithm family, namely the conjugate expectation maximization (ConjEM) algorithm is further improved to gain attractive theoretical properties. The multimodal object detection and object number estimation methods are developed, their theoretical properties are discussed. Finally, the proposed multimodal clustering method is combined with the object detection and object number estimation strategies and known tracking techniques to perform multimodal multiobject tracking. The performance is demonstrated on simulated data and the database of realistic audio-visual scenarios (CAVA database).

Keywords: conjugate mixture models, audio-visual scene analysis, audio-visual calibration, multimodal object detection, multimodal object tracking