ON THE MODELLING OF MULTI-MODAL DATA USING REDUNDANT DICTIONARIES

THÈSE N^{O} 3741 (2007)

PRÉSENTÉE LE 23 MARS 2007

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Laboratoire de traitement des signaux 2

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Gianluca MONACI

Laurea in ingegneria delle telecomunicazioni, Università degli studi di Siena, Italie
et de nationalité italienne

acceptée sur proposition du jury:

Prof. M. Kunt, président du jury
Prof. P. Vandergheynst, directeur de thèse
Dr S. Bengio, rapporteur
Dr M. Elad, rapporteur
Dr M. Plumbley, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Lausanne, EPFL
2007

Acknowledgments

The first person I want to thank is my advisor, Pierre Vandergheynst. He gave me the opportunity to work on a very interesting subject and he guided and supported me during the four years of thesis. Thank you Pierre for all that and for being the great person you are.

I thank all the members of the jury who carefully read this manuscript and contributed to its improvement with helpful comments and suggestions.

A special thank goes to all the members of the Signal Processing Institute where I have had the opportunity to work during the past four years in a wonderful, stimulating environment. In particular I want to thank the director, Prof. Murat Kunt.

Part of the work present in this dissertation has been done in collaboration with colleagues and students working under my supervision. Working with them has been very important for improving my research work and most of all for my personal development. I would like to express my gratitude to Oscar Divorra Escoda, Philippe Jost, Boris Mailhe, Sylvan Lesage, Anna Llagostera Casanovas, Emilio Maggio and Patricia Besson. I would like to thank in particular Rémi Gribonval for the work we have done together and the interesting and helpful discussions we had. A special thank goes to Andrea Cavallaro for the great time I had in London and for having revealed me the secrets of particle filtering.

I feel very lucky of having found a lot of wonderful friends in Lausanne and all around the world. And I feel even more lucky of having kept so many good friends in Siena and all over the world. I want to thank all these people for being just as they are and for having been with me during all this time.

Infine vorrei ringraziare la mia famiglia per avermi sempre sostenuto e per essere la magnifica famiglia che sono! In particolare, voglio dedicare questo lavoro ai miei genitori, mamma Doretta e babbo Adriano.

Table of contents

Table of contents iv

Abstract ix

Version abrégée xi

List of figures xiv

List of tables xv

Notations and Symbols xvii

1 Introduction 1

1.1 Motivation . . . 1

1.2 Organization of the Thesis and Main Contributions . . . 2

2 Multi-Modal Signal Analysis 5

2.1 What are Multi-Modal Signals . . . 5

2.2 Existing Audiovisual Fusion Methods . . . 8

2.2.1 Seminal Works on Multi-Modal Fusion . . . 8

2.2.2 Information Theoretic Approaches to Multi-Modal Fusion . . . 9

2.2.3 Other Research Directions . . . 10

2.3 Where Are We Now? . . . 11

2.4 Discussion . . . 12

3 Audiovisual Gestalts 13

3.1 Gestalt Theory and Audiovisual Processing . . . 13

3.2 Gestalts in Computer Vision and *Helmholtz Principle* . . . 15

3.3 Audiovisual Gestalts . . . 16

3.4 Sparse Representations on Redundant Dictionaries . . . 17

3.5 Audio-Video Representation and Fusion . . . 18

3.5.1 Representation of the Video Signal . . . 18

3.5.2 Representation of the Audio Signal . . . 21
3.5.3 Audiovisual Fusion . . . 25
3.6 Detection of Audiovisual Meaningful Events . . . 25
3.6.1 An Audiovisual Event Detector Based on the Helmholtz Principle . . . 26
3.7 Experiments . . . 27
3.8 Discussion . . . 34

4 Tracking Atoms with Particles 35
4.1 Tracking Visual Features . . . 35
4.2 Tracking Geometric Video Structures . . . 36
4.2.1 Video Representation . . . 36
4.2.2 Tracking Video Atoms with Particle Filter . . . 37
4.3 Experiments . . . 40
4.3.1 Video Atoms Tracking . . . 42
4.3.2 Audiovisual Source Localization . . . 43
4.4 Discussion . . . 44

5 Blind Audiovisual Source Separation 47
5.1 From Audio to Audiovisual Source Separation . . . 47
5.1.1 Blind Audio Source Separation . . . 48
5.1.2 Audiovisual Source Separation . . . 49
5.1.3 Single-Channel Blind Source Separation : A Difficult Problem . . . 50
5.2 Blind Audiovisual Source Separation (BAVSS) . . . 51
5.2.1 Phase 1 : Spatial Localization of Video Sources . . . 54
5.2.2 Phase 2 : Separation and Reconstruction of Video Sources . . . 59
5.2.3 Phase 3 : Temporal Localization of Audio Sources . . . 60
5.2.4 Phase 4 : Blind Audio Source Separation Aided by Video . . . 62
5.3 Experiments . . . 66
5.4 Discussion . . . 70

6 Learning Multi-Modal Dictionaries 71
6.1 Modelling and Understanding . . . 71
6.1.1 Sparse Approximations of Multi-Modal Signals . . . 72
6.1.2 Synchrony and Shift Invariance in Multi-Modal Signals . . . 73
6.2 Learning Multi-Modal Dictionaries . . . 74
6.3 Experiments . . . 79
6.3.1 Audiovisual Dictionaries . . . 79
6.3.2 Audiovisual Speaker Localization . . . 81
6.4 Discussion . . . 85

7 Conclusion 87
7.1 Discussed Topics and Achievements . . . 87

7.2 Future Research Directions 88

Bibliography **98**

Abstract

Real-world phenomena involve complex interactions between multiple signal modalities. As a consequence, humans are used to integrate at each instant perceptions from all their senses in order to enrich their understanding of the surrounding world. This paradigm can be also extremely useful in many signal processing and computer vision problems involving sets of mutually related signals, called multi-modal signals. The simultaneous processing of multi-modal data can in fact reveal information that is otherwise hidden when considering the different modalities independently.

This dissertation deals with the modelling and the analysis of natural multi-modal signals. The challenge consists in representing sets of data streams of different nature, like audio-video sequences, that are interrelated in some complex and unknown manner, in such a way that useful information shared by the different data modalities can be extracted and intuitively used. In this sense signal representation have to make an effort to model the structural properties of the observed phenomenon, so that data are expressed in terms of few, meaningful elements. In fact, if information can be represented using only few components, this means that such components capture its salient characteristics. In order to efficiently represent multi-modal data, we advocate the use of sparse signal decompositions over redundant sets of functions (called dictionaries).

In this thesis we consider both application-related and theoretical aspects of multi-modal signal processing. We propose two models for multi-modal signals that explain multi-modal phenomena in terms of temporally-proximal events present in the different modalities. A first simple model is inspired by human perception of multi-modal stimuli and it relies on the representation of the different data streams as sparse sums of dictionary elements. This type of representation allows to intuitively define meaningful events present in the different modalities and to discover correlated multi-modal patterns. Taking inspiration by this first model, we introduce a representational framework for multi-modal data based on their sparse decomposition over dictionaries of multi-modal functions. Instead of separately decompose each modality over a dictionary and seek for correlations between the extracted patterns, we impose some correlation between modalities at the model level. Since such correlations are difficult to formalize, we propose as well a method to learn dictionaries of synchronous multi-modal basis elements.

Concerning the applications presented in this dissertation, we tackle two major audiovisual fusion problems, that are audiovisual source localization and separation. Although many of the ideas developed in this work are completely general, we consider this field since it is the one that presents the vastest possibilities of application for this research. The theoretical frameworks developed throughout the thesis are used to localize, separate and extract audio-video sources in audiovisual sequences. Algorithms for cross-modal source localization and blind audiovisual source separation are tested on challenging real-world multimedia sequences. Experiments show that the proposed approach leads to promising results for several newly designed multi-modal signal processing algorithms and

that a careful modelling of data structural properties can convey interesting, useful information to understand complex multi-modal phenomena.

Keywords

Multi-modal signal processing, sparse representation, redundant dictionary, audiovisual blind source separation, cross-modal localization, dictionary learning.

Version abrégée

Les phénomènes réels impliquent des interactions complexes entre plusieurs modalités de signal. Les humains sont habitués à intégrer à chaque instant les perceptions issues de tous leurs sens afin d'enrichir leur compréhension du monde environnant. Ce paradigme peut être extrêmement utile pour beaucoup de problèmes en traitement des signaux impliquant des ensembles de données conjointement corrélés appelés signaux multi-modaux. Le traitement simultané des données multi-modales peut, en fait, révéler de l'information qui est cachée lorsque on considère les signaux indépendamment.

Cette thèse traite de la modélisation et de l'analyse des signaux multi-modaux naturels. Le défi consiste à représenter des ensembles de flux de données de nature différente, tel que les séquences audiovisuelles, qui sont liés de façon complexe et inconnue, de telle manière que l'information utile partagée par différentes modalités puisse être extraite et utilisée. Il faut faire un effort au moment de la représentation des signaux afin de modéliser les propriétés structurales des phénomènes observés, de sorte que les données soient exprimées avec un petit nombre d'éléments significatifs. Si l'information est représentée par un petit nombre de composants, cela signifie que ces composants capturent ses caractéristiques les plus importantes. Afin de représenter efficacement des données multi-modales, nous préconisons l'utilisation des décompositions parcimonieuses dans des ensembles redondants de fonctions (appelés dictionnaires).

Dans cette thèse nous considérons des aspects théoriques et applicatifs du traitement des signaux multi-modaux. Nous proposons deux modèles qui expliquent des phénomènes multi-modaux en termes d'événements temporellement proches dans les différentes modalités. Un premier modèle simple est inspiré par la perception humaine des stimuli multi-modaux et se fonde sur la représentation des différents flux de données par des sommes parcimonieuses d'éléments de dictionnaire. Ce type de représentation permet de définir intuitivement des événements signicatifs dans les différentes modalités et de découvrir des motifs multi-modaux corrélés. Nous présentons aussi un cadre représentatif pour des données multi-modales basé sur leur décomposition parcimonieuse dans un dictionnaire de fonctions multi-modales. Au lieu de décomposer séparément chaque modalité dans un dictionnaire et de rechercher par la suite des corrélations entre les motifs extraits, nous imposons une certaine corrélation entre les modalités au niveau du modèle. Etant donné qu'il est difficile de formaliser de telles corrélations, nous proposons aussi une méthode pour apprendre des dictionnaires d'éléments de base multi-modaux synchrones.

Les applications présentées dans cette thèse traitent deux problèmes majeurs dans le domaine de la fusion des signaux audiovisuels : la localisation et la séparation de sources audiovisuelles. Bien que plusieurs idées développées dans ce travail soient complètement générales, nous considérons ce champ puisqu'il présente le plus grand nombre d'applications. Les cadres théoriques qui ont été développés dans la thèse sont employés pour localiser, séparer et extraire des sources dans des séquences audio-

vidéo. Des algorithmes pour la localisation et la séparation des sources audiovisuelles sont testées sur des séquences naturelles complexes. Les expériences prouvent que l'approche proposée mène à des résultats très prometteurs pour plusieurs nouveaux algorithmes pour le traitement des signaux multi-modaux et que la modélisation des propriétés structurales de données peut fournir de l'information utile et intéressante pour la compréhension des phénomènes multi-modaux complexes.

Liste des mots-clefs

Traitement des signaux multi-modaux, décomposition parcimonieuse,
dictionnaire redondant, séparation de sources audiovisuelles,
localisation de sources, apprentissage de dictionnaires.

List of figures

1.1 Who is giving you her/his telephone number? 1

2.1 Econometric time series : an example of multi-channel signal. 6
2.2 A pair of MR and CT scans. 7
2.3 Audiovisual sequences are multi-modal signals that are heterogeneous both in dimensionality and resolution. 8

3.1 Examples of Gestalt laws of visual perception. 14
3.2 Video component of an audiovisual sequence including a hand playing the piano and a moving toy car. 15
3.3 Generating function $g^{(i)}(x_1, x_2)$. 20
3.4 Approximation of a synthetic scene by means of a 2D time-evolving atom. 22
3.5 Audio signal of a subject uttering digits in English, with its time-frequency energy distribution and the estimated audio feature. 23
3.6 Original audio signal and four corresponding audio features. 24
3.7 Scheme of the proposed audiovisual fusion criterion. 26
3.8 Test sequences Piano 1 and Piano 2. 28
3.9 Results of the proposed algorithm run on clip Piano 1. The most correlated atoms, in white, represent the player's fingers. The moving toy car is not detected. 29
3.10 Results for the sequence Piano 2. The correlated atoms, highlighted in white, are on the player's fingers and the piano keys. The oscillating ventilator is not detected. . . 29
3.11 Results for *Experiment 1* : the first row shows the original video frames, the second row shows the video atoms correlated with the corresponding audio signal, the third row shows the video atoms correlated with an incongruous audio source. 30
3.12 Results for *Experiment 2* : the most correlated 3D atoms are highlighted in white. The mouth and the chin of the correct speaker are detected. 31
3.13 Regions of correct mouth detection. 32
3.14 Sample raw frames of clip g20 and their reconstruction using only video atoms close to the estimated sound source. 33

4.1 Sum of scalar products between the atoms representing the first frame of a sequence, and average scalar product, plotted as a function of the number of functions. 38

4.2 Likelihood function of a candidate atom computed on a region extracted from one of the analyzed clips. . . . 40
4.3 Schematic representation of the Particle Filter algorithm. . . . 41
4.4 Video atoms tracking. Results for the 3D-MP approach and for the MP-PF method are shown. . . . 42
4.5 Frames from clips g19 and g21. The footprints of the most correlated atoms are highlighted. . . . 43

5.1 Example of a sequence analyzed with the BAVSS algorithm. . . . 52
5.2 Schematic representation of the audiovisual source separation algorithm. . . . 53
5.3 Sketches of the audio and video features. . . . 56
5.4 Video atoms location over the image. . . . 57
5.5 Clusters created using different cluster sizes in step 4 of the algorithm. . . . 59
5.6 Example of the video sources reconstruction. . . . 60
5.7 Example of the classification of the audio atoms into the corresponding source. . . . 62
5.8 Estimated frequency probabilities for the two speakers of the considered test sequence. 63
5.9 Estimated temporal probabilities for the two speakers of the test sequence. . . . 64
5.10 Estimated time-frequency probabilities for the two speakers of the considered test sequence. . . . 65
5.11 Source Separation of a real-world mixture. . . . 66
5.12 Comparison between audio atoms resulting of time-frequency analysis in a synthetic mixture with the original ones. . . . 68
5.13 Comparison between estimated and real soundtracks in a synthetic sequence. . . . 69

6.1 Schematic representation of the multi-modal learning algorithm. . . . 77
6.2 Audio-video generating functions of *Dictionary 2*. Twenty learned functions are shown, each consisting on an audio and a video component. . . . 80
6.3 Test sequences Movie 1, Movie 2 and Movie 3. . . . 82
6.4 Sample frames of Movie 1, Movie 2 and Movie 3. The white cross highlights the estimated position of the sound source, which is correctly localized. . . . 83
6.5 Sample frames of Movie 3. The positions of maximal projection between video functions and test sequence are plotted on the image plane. . . . 84

List of tables

3.1 Audiovisual source localization results expressed in percentage of correct detections. 33

4.1 Audiovisual localization results expressed in percentage of correct detections. 44

5.1 Example of the list of correlation values between one audio atom and the correlated video atoms 61

5.2 Results obtained with synthetic sequences generated for different clips of CUAVE database. 68

6.1 Summary of the source localization results for all the tested sequences. 85

Notations and Symbols

$\mathbb{R}$	The set of real numbers
$\mathbb{Z}$	The set of integer numbers
$a(t)$	Digital audio signal
$I(x_1, x_2)$	Digital image
$\mathbf{V}(x_1, x_2, t)$	Digital video signal
$(s^{(1)}, \ldots, s^{(M)})$	Multi-modal signal (M modalities)
$\mathcal{D}$	General (redundant) dictionary of unit norm atoms
$\mathcal{D}^{(m)}$	Dictionary of unit norm atoms for modality m
ϕ	General normalized atom
$\phi^{(m)}$	Normalized atom for modality m
$(\phi^{(1)}, \ldots, \phi^{(M)})$	Multi-modal atom (M modalities)
g	generating function
$g^{(m)}$	generating function for modality m
$(g^{(1)}, \ldots, g^{(M)})$	Multi-modal generating function (M modalities)
$W\phi(t, \omega)$	Wigner-Ville distribution of atom ϕ
T_p	Multi-modal translation operator (continuous)
$\mathcal{T}_p$	Multi-modal translation operator (discrete)
$\mathsf{y}(t)$	Activation vector
$\mathbf{s}(t)$	Synchronization vector
$\mathbf{x}[n]$	State vector of the n-th target
$\mathbf{z}[n]$	Measurement vector associated to the n-th target
$\mathcal{L}(\cdot)$	Likelihood function
χ	Correlation score between audio and video atoms
κ	Confidence value of a video atom
K_C	Confidence value of a cluster C

1 Introduction

1.1 Motivation

Figure 1.1 shows four sample frames taken from an audiovisual sequence. The movie involves a boy and a girl speaking in front of a camera. Actually they are saying series of digits in English. It should be clear, looking at the pictures, that both of them are moving their lips as if they were uttering some words. However, if one could listen to the movie soundtrack, it would be immediately clear that only one person is speaking. But how can one know which of the two is the speaker? This could be quite an interesting information, for example if the speaker is telling us her/his telephone number.

Figure 1.1 – *Who is giving you her/his telephone number? Sequence taken from the CUAVE database [88].*

In this particular case a human listener could recognize, in absence of significant noise, whether the speaker is a boy or a girl and associate the speech to the correct person. However, not only this reasoning would fail if both people are of the same gender, but it also involves a complex, high-level process of gender recognition both in the audio and video domain. It is interesting to notice that this type of approach requires an independent analysis of the audio and video signals. The information deduced (the gender of the persons on the video and the gender of the speaker) is then trivially combined to associate speech and speaker.

On the other hand, it has been shown that there is a more simple and basic mechanism that strongly contributes to the integration of acoustic and visual stimuli, the synchrony between the presence of a sound and a visible movement [11, 39, 58, 75, 113]. Such process is not cognitive but

it exploits the physical nature of the observed phenomenon : we are hearing a sound and thus it is likely that some mechanical, visible action has produced it. Interestingly, this type of mechanism acts at the stimuli level and no complex inference has to be done. In contrast, the different signals have to be analyzed together.

This thesis deals with a family of signals, called *multi-modal*, that like audiovisual sequences are constituted of different data streams (or *modalities*) that have a certain degree of correlation since they describe the same physical phenomenon. The interest of studying these type of signals resides in the fact that many useful information can be extracted from the joint analysis of the different modalities that is otherwise unavailable if the signal modalities are considered independently. In the example above, it is difficult to say if it is the girl or the boy who is speaking, if one looks at the audio and video signals separately. However, if one observes the two modalities together and seeks for synchronicity between sound and lips movements, it becomes possible to understand that the speaker is in fact the boy.

In this manuscript the attention will be focused on a broad class of multi-modal data that exhibit correlations along time. In fact, throughout this dissertation we will consider the case study of audiovisual sequences. The are several reasons to do that. The first and more prosaic one is that the analysis of audiovisual data has been the starting point of the research underlying this work, providing tools and ideas for a study that later has led to the definition of more general techniques. Secondly, audiovisual sequences represent well most of the challenges involved in the analysis of multi-modal signals. Finally, audiovisual data processing is the most important field of application for this research and insights in this field would help facing many multimedia signal processing problems.

The main objective of this work is to *understand and model* correlated multi-modal data arrays in order to develop effective and intuitive techniques to jointly analyze this type of signals and thus to extract the useful information "hidden" in the data. As we were underlying before, multi-modal signals describe different aspects of a same physical phenomenon. In our understanding, if we want to retrieve correlations between different signal modalities, it is of paramount importance to capture the structure of such phenomenon. In this sense effective data modelling should be able to represent signals in terms of few, important data structures, in such a way that dimensionality gets reduced and only relevant signal information is used. In fact it seems that this is what the human brain does when it localizes the source of a sound in the space by associating salient features like visual motion and presence of a sound. In addition, advances in the understanding and modelling of this type of data can be extremely valuable in a relatively young and barely explored research field like that of multi-modal signal processing. In order to effectively model multi-modal data we want to exploit the structural properties of the considered signals, and in this thesis we will show how this can be done using *sparse* signal representations over *redundant dictionaries of functions*.

To summarize, in this research work we consider and develop these three main issues :

- Why it is useful to jointly analyze correlated multi-modal data;
- Why it is important to carefully model the structural properties of such data;
- How redundant dictionaries can be used to effectively model multi-modal signals.

1.2 Organization of the Thesis and Main Contributions

At this point let us introduce the outline of the thesis. This dissertation pivots around the central idea that in order to catch the correlations between complex signal modalities we need to model the

observed phenomenon in such a way that few significant data structures are highlighted. This idea is developed through four main parts presented in Chapters 3, 4, 5 and 6, that are preceded by an introductory chapter.

More precisely, **Chapter 2** analyzes the first two points stated at the end of previous section. We start by defining what are in our understanding multi-modal signals and we present examples of multi-modal data analysis borrowed from different disciplines, from economics to medical imaging and audiovisual data processing. This highlights the importance that has been given in the last years to the study of this type of signals. We focus then our attention on recent advances brought up in the field of audiovisual signal fusion, with particular emphasis on the audiovisual source localization and separation problems, that are the principal applications targeted in this thesis. A detailed literature survey is carried out and advantages and limitations of existing audiovisual fusion methods are discussed, motivating the choice done in this work to adopt an approach that aims at modelling audiovisual signals as synchronous salient audio-video structures.

Chapter 3 introduces an audiovisual localization framework based on the detection of correlated audio-video events in multimedia sequences. The problem that is faced in this chapter is the one proposed at the beginning of this introduction : if one has a video sequence showing several possible video sources and an audio signal associated to one of these sources, how can we link the acoustic stimulus to the correct visual structure and thus localize on the video the sound source? As discussed above, this task is trivial for humans, but it is a real challenge for automatic systems. In this chapter we propose a source localization model inspired by human perception and that thus exploits the synchrony between audio and video *events*. Audio and video signals are represented as *sparse* sums of few representative functions taken from a large set of candidate basis waveforms (called *redundant dictionaries*). In this way salient signal features are extracted and perceptually meaningful audiovisual events are defined. We will show how this principle can be used to detect existing cross-modal correlations between audio-video signals even in presence of distracting motion and acoustic noise. Results show that temporal proximity between audiovisual events is a key ingredient for the integration of information across modalities and that it can be effectively exploited for the design of multi-modal analysis algorithms.

The proposed approach is based on signal representation methods that decompose multi-modal signals over redundant dictionaries of functions, obtaining concise descriptions of the structural properties of the data. Audio and video representation techniques are analyzed more in details in Chapters 4 and 5, where their characteristics, flaws and strengths are studied.

In **Chapter 4** the video representation algorithm is considered. This chapter presents a framework and an algorithm for tracking relevant visual structures. Important image contours to be tracked are picked up from a redundant dictionary and ranked. Based on the ranking, the contours are automatically selected to initialize a *Particle Filtering* tracker. The proposed algorithm deals with salient video entities whose behavior has an intuitive meaning, related to the physics of the signal. Moreover, as the interactions between such structures can be easily defined, the inference of higher level signal configurations can be made intuitive. We will see how the proposed method improves the performance of existing video structures trackers, while reducing the computational complexity.

In **Chapter 5** instead our attention is turned to the audio signal. In this chapter we introduce a new concept, Audiovisual Source Separation, that lies on the edge of two very different research areas : audiovisual fusion and one-microphone blind audio source separation. These two fields are typically considered to be separated, but we will see in this chapter how the ideas developed in these two areas can be helpful to extract correlated audio-video sources exploiting the information

contained in the *mono* soundtrack and in the associated video sequence. The method builds correlations between acoustic and visual structures that are represented using functions retrieved from redundant dictionaries. Video structures that exhibit strong correlations with the audio track and that are spatially close are grouped together using a robust clustering algorithm that can confidently count and localize audiovisual sources on the image plane. Then, using such information and exploiting the coherence between audio and video signals, audio sources are localized as well and separated.

In Chapters 3, 4 and 5 audio and video modalities are represented with basic forms taken from redundant dictionaries. Audio-video structures are extracted separately using general codebooks of functions, and then correlations between them are searched. We argue that a more efficient strategy would be to jointly extract meaningful multi-modal structures, introducing cross-modal correlations *at the model level*. **Chapter 6** explores this paradigm introducing a completely new model for multi-modal signals. The model considers multi-modal data to be composed of a sum of recurrent synchronous multi-modal structures retrieved from a dictionary of functions. Since it is not trivial to design a dictionary of meaningful multi-modal basis functions, we propose as well an algorithm to learn a collection of such basis waveforms from training data, enforcing synchrony between the different modalities and de-correlation between the dictionary elements. The model and the learning method are completely general, but we have employed them to represent audiovisual sequences. The learned audio-video dictionaries seem to effectively capture underlying structures present in the data. The dictionary functions are used to analyze complex multimedia clips, showing the ability to detect meaningful correlated audio-video structures and to localize the sound source in the video sequence.

Finally, in **Chapter 7** the entire thesis is discussed and conclusions are drawn. We propose as well possible developments and future research directions for the presented work.

To summarize, the main contributions presented in this dissertation are the following :

- We propose the use of redundant dictionaries of functions to represent, in terms of salient signal structures, audio and video data. This representation allows the intuitive definition of multi-modal correlated structures that can be effectively detected and extracted;
- We introduce a novel framework for the tracking of visual structures. The tracker follows relevant image contours defined with functions retrieved from a redundant dictionary using the Particle Filtering method;
- We propose a new approach to audiovisual source separation that exploits audiovisual coherence between a single-microphone audio signal and the associated video sequence to separate correlated sources;
- We define a model of multi-modal signals that are represented as sparse sums of recurrent multi-modal functions taken from codebooks of functions. A learning algorithm to build such multi-modal dictionaries from training data is proposed as well.

Multi-Modal Signal Analysis

2

2.1 What are Multi-Modal Signals

We continuously combine stimuli from our senses to enhance our perception of the world. In fact several research works have investigated this issue, demonstrating that humans are used to integrate acoustic and visual signals [39, 75, 113, 116] or tactile and visual inputs [15, 112]. Several signal processing algorithms used to analyze sets of interrelated data successfully exploit this same principle.

Multi-modal signals are sets of *heterogeneous* data arrays that exhibit some mutual dependency, since they represent the same physical phenomenon. Different modalities in fact are often captured by different sensors, and thus they can have different dimensionality and resolution, which often makes the definition of cross-modal correlations difficult and the joint analysis of this type of signals challenging. However, the simultaneous processing of different signal modalities allows to discover structures in the data revealing information that is unavailable when considering the modalities independently. Several researchers in various fields have tackled the problem, suggesting different definitions of multi-modal signals and different techniques to represent and jointly analyze signal modalities.

We would like to start this dissertation by clearly defining what multi-modal signals are in our understanding. We term multi-modal signals sets of ***correlated multi-channel heterogeneous signals***. Each channel is considered to represent a modality. Signal modalities can be heterogeneous both in resolution (e.g. if modalities are captured with different devices) and in dimensionality (they can be data arrays in 1D, 2D, 3D...), but they are supposed to describe the same phenomenon and thus to be somehow correlated. Following this definition, we can classify multi-component signals according to their *degree of multi-modality* :

Signals homogeneous in dimensionality and resolution (*multi-channel signals*) - These signals are often termed as multi-modal, even if we prefer to call them *multi-channel signals*. This type of signals are typically analyzed in economics, where the joint processing of different economic time series (1D) is of paramount importance to build effective macroeconomics or

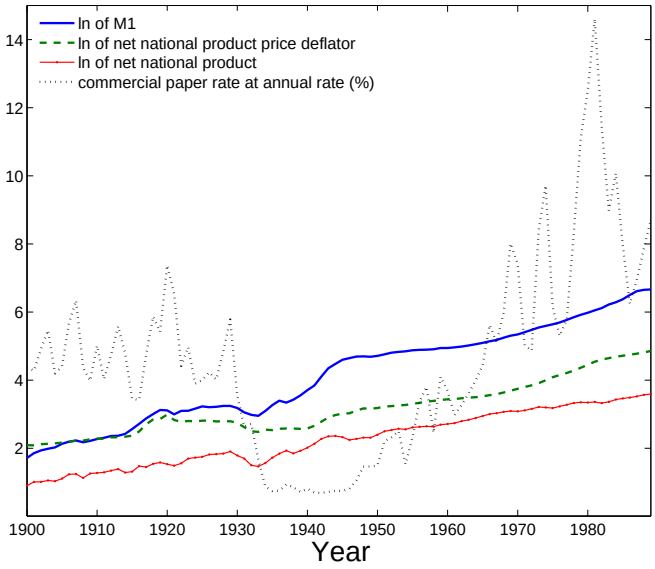


Figure 2.1 – *Econometric time series are multi-channel signals. These are 1D signals typically sampled at an homogeneous frequency (one sample per year in this case). Source : http://www.e.u-tokyo.ac.jp/~hayashi/hayashi_econometrics.htm.*

financial economics models [1, 53, 71]. Other examples come from remote sensing, where satellite images are segmented and classified using image versions at different wavelengths [44, 100]. In Fig. 2.1 an example of multi-channel signal taken from [53] is shown. Here the values of four econometric indexes are plotted as a function of time : the natural logarithm of US M1 money stock, of the US net national product price deflator and of the US net national product, together with the commercial paper rate in percent at an annual rate (see [53] or `http://www.federalreserve.gov` for further details on the meaning of these values). These four indexes are jointly studied to analyze the US money market and build models of "money supply". All signals are 1D and they are sampled at an homogeneous frequency of one sample per year.

Signals homogeneous in dimensionality and heterogeneous in resolution - These signals are typically captured using different sensors and they can be considered a simple case of multi-modal signals. This type of data are extensively analyzed in medical imaging, where the spatial correlation between different modalities is exploited for the segmentation and registration of magnetic resonance (MR) and computed tomography (CT) scans [18, 72]. In remote sensing as well, multi-spectral satellite images are jointly segmented using measurements from visible, infra-red and radar sensors [41] or ice charts are built combining information from satellite images captured with very high resolution radiometer, synthetic aperture radar, operational line scanner and sensor microwave/imager [86]. Figure 2.2 shows a pair of corresponding sections of MR [Left] and CT [Right] scans. The two modalities in this case are both 3D volumes that are however acquired using different devices and at different resolutions. In this example the MR scan is composed of 150 slices of 256×256 pixels images like the one shown in Figure 2.2, while the CT scan has a higher resolution being made up of 150 slices of 512×512 pixels. In addition, the patients usually lie in different positions when the scans are acquired, making the correspondence between the two signals even more complex to be established. MR

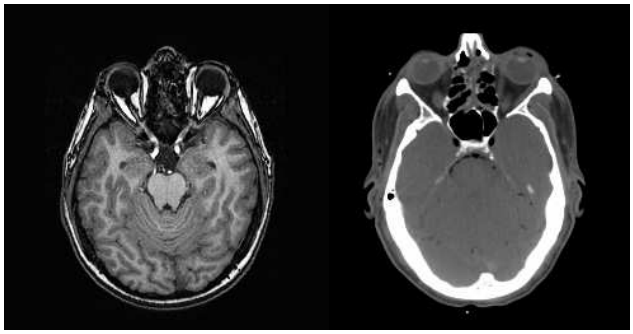


Figure 2.2 – *A pair of MR [Left] and CT [Right] scans is a multi-modal signal whose modalities have the same dimensionality (3D) but different characteristics and resolutions. Two corresponding sections of the MR and CT volumes are shown.*

and CT images are jointly used for registration-segmentation purposes since MR images are best suited for soft, non-calcified tissues, while CT scanners represent well dense tissues (e.g. bones), offering thus complementary information.

Signals heterogeneous in dimensionality and resolution - This is the most general type of multi-modal signals. This type of data can be very difficult to study, due to the differences in resolution and dimensionality. However the joint analysis of the different channels can be extremely fruitful, since each modality provides information about the observed phenomenon that is typically very different and often complementary with respect to the other modalities. Examples of these techniques come again from medical imaging, where 3D magnetic resonance angiography is aligned with 2D X-ray angiographic images to obtain a richer visualization [55]. In neuroscience, 1D electroencephalogram (EEG) and 4D functional magnetic resonance imaging (fMRI) data are jointly analyzed to study brain activation patterns [74]. In environmental science, connections between local and global climatic phenomena are discovered by correlating different spatio-temporal measurements [21]. Finally, the class of multi-modal signals that has been investigated more in detail in the last years is surely that of audiovisual sequences. Many multimedia signal processing problems involve the simultaneous analysis of audio and video data, e.g. talking heads creation and animation [24], speech-speaker recognition [70, 92] and detection [13, 120], audio filtering and enhancement based on video [31, 45, 46], or sound source localization [18, 27, 42, 52, 54, 62, 82, 102, 103]. Figure 2.3 shows an example of audiovisual signal. The audio component [Left] is a 1D signal that is typically sampled at $\mathcal{O}(10^4)$ samples/sec, while the video component [Right] is a 3D signal sampled with considerably lower temporal resolution ($\mathcal{O}(10^1)$ frames/sec) and with a spatial resolution typically of $\mathcal{O}(10^2) \times \mathcal{O}(10^2)$ pixels.

In this thesis we study this last type of multi-modal signals, i.e. multi-channel heterogeneous signals. In particular we will develop algorithms to analyze a broad class of signals exhibiting correlations *along time*, like EEG-fMRI data or audiovisual sequences. The case studies that will be considered throughout all the manuscript are the cross-modal audiovisual source localization and source separation problems, since they represent well the challenges involved in the analysis of multi-modal data and because they constitute some of the major fields of application for this research.

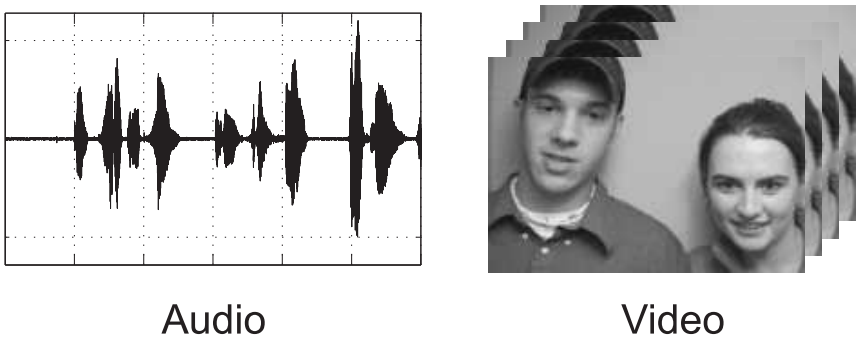


Figure 2.3 – *Audiovisual sequences are multi-modal signals that are heterogeneous both in dimensionality and resolution. The audio component [Left] is a 1D signal typically sampled at $\mathcal{O}(10^4)$ samples/sec, while the video component [Right] is a 3D signal sampled with considerably lower temporal resolution ($\mathcal{O}(10^1)$ frames/sec) and with a spatial resolution typically of $\mathcal{O}(10^2) \times \mathcal{O}(10^2)$ pixels.*

2.2 Existing Audiovisual Fusion Methods

Even though some of the techniques that we will present are completely general, we will target applications in the audiovisual signal processing field. There exist several methods that face the audiovisual source localization and separation problem using multi-microphone systems : stereo triangulation is used to estimate the spatial location of sounds [8, 91] while in [97, 104] video information is integrated in Blind Audio Source Separation (BASS) algorithms to perform speech separation. Instead, here we want to achieve cross-modal source localization and separation using only an image sequence and one microphone, exploiting thus the correlation between audio and video at the *signal level*. We believe in fact that when modality-fusion takes place at the *decision level*, many of the joint (and useful) signal characteristics get lost. Although decision level fusion schemes may be based on simple statistical measures, such simplification typically results in a reduced capability of modelling the observed phenomena.

In the next sections we will review the main contributions in the field of audiovisual fusion and localization. In particular, the attention will be focused on the features used to represent audio-video data and on the techniques that are adopted to estimate correlations among them.

2.2.1 Seminal Works on Multi-Modal Fusion

The problem we are challenging is that of correlating audio and video signals in multimedia sequences to detect consistent audiovisual pairs that could originate from the same physical phenomenon. The topic was first faced by Hershey and Movellan [54], that proposed to measure the correlation between audio and video using an estimate of the Mutual Information (MI) [26] between the energy of an audio track and the values of single pixels. Since a per-pixel measure is used, the hypothesis that pixels are independent of each other conditioned on the speech signal is introduced. In [54], Mutual Information is derived from the Pearson's correlation coefficient [4], assuming thus that the joint statistics are Gaussian and that audio-video representations are linearly related.

Slaney and Covell [102] generalized this approach looking for a method able to measure the synchrony between audio signals and video facial images. In order to deduce a relationship between the cepstral representation of the audio and the video pixels, the authors use Canonical Correlation Analysis (CCA), which is equivalent to maximum Mutual Information projection in the jointly Gaussian case [42]. CCA allows to compare sequences of different dimensions, allowing thus to estimate correlations between an audio feature and the whole video frames (and not just the single

pixels). Several audio descriptors are investigated in this work, such as Mel-frequency Cepstral Coefficients (MFCC) [93], linear-predictive coding [93], line spectral frequencies [105], spectrograms, and raw signal energy. The authors report similar results for the first three methods, while raw energy and spectrograms result less effective since more noisy. They end up using MFCC analysis, like many other researchers, because it is a favorite front-end for speech-recognition systems.

Recently a CCA-based approach was suggested in [62]. The authors perform a rigorous analysis of the multi-modal localization problem and propose an algorithm that provides a unique solution to the problem imposing sparsity of the result. The video signal is represented using the wavelet coefficients of difference images while the audio feature is the average acoustic energy per frame. The algorithm is tested on two sequences with substantial audio-video distractors and it exhibits promising localization results.

Nock and co-workers [82] carried out a first extensive study evaluating three different audiovisual synchrony measures and several video representations in a speaker localization context. Two of the considered methods are based on MI : one assumes discrete distributions [81] and the other one considers multivariate Gaussian distributions as in [54]. These two measures, like those proposed in [54, 102] are general statistical measures of correlation between random variables and make no assumption about the structure of the analyzed signals. The third synchrony measure instead is a face-speech specific measure and makes use of Hidden Markov Models (HMMs) trained on audiovisual data to define the likelihood of a video configuration given a certain sound. Audio features are extracted by MFCC analysis, while different video features are tested : coefficients of the Discrete Cosine Transform (DCT), pixel intensities and pixel intensity changes [18]. Tests are performed on a large database of audiovisual sequences, the CUAVE dataset [88], and the Gaussian MI method achieved superior results when using the DCT-based representation of the video.

The methods described until here share several characteristics and limitations :

- All algorithms except [62] require training in order to build *a priori* models;
- Correlation measures are computed between data array that are considered to be random variables (except for the HMMs-based method in [82]) under more or less restrictive assumptions (linearity, independence, mutual Gaussianity);
- Visual information is represented using basic features that are barely meaningful from the point of view of the signal structures, like pixel-based features (intensities or intensity changes) or DCT coefficients;
- Audio representations are typically based on MFCC analysis inherited from the speech processing field. Mono-dimensional audio features are built by concatenating or combining cepstral coefficient, making somehow difficult the interpretation of the representation.

In the next sections we review audiovisual localization algorithms that address some of the matters brought up by the above-described studies.

2.2.2 Information Theoretic Approaches to Multi-Modal Fusion

In the last years several algorithms based on information theoretic features optimization have been introduced. Fisher and colleagues have developed in [43] a multi-modal fusion framework that has been extended in their latest work [42]. The algorithm is based on a probabilistic generation model that is used to define projection rules on *maximally informative subspaces*. Such subspaces are defined as linear combinations of input signals that maximize MI between different modalities. Here no hypothesis is made on the distributions of the random variables representing audio-video

signals and MI is calculated using a Parzen estimator [87]. The audiovisual features used are still simple, pixel intensities for the video and audio periodograms. This approach is used to solve a conversational audiovisual correspondence problem, obtaining interesting results. However, the use of Parzen windows to estimate Mutual Information requires multiple tune-up parameters and a considerable amount of data to make the estimation reliable.

A similar approach based on Markov chains modelling of audio and video signals is proposed by Butz and Thiran [18]. The audiovisual consistency is assessed by maximizing the *efficiency coefficient*, i.e. the ratio between audiovisual MI and the audio-video joint entropy. The distributions of audio-video features are again estimated using Parzen density estimators. The video is represented by pixel intensity change and the audio feature is the linear combination of the power spectrum coefficients that brings biggest entropy. The framework developed in [18] is used in [13] to extract optimal audio features with respect to video, that is represented using the optical flow extracted from target regions identified using a face tracker. These audiovisual features are then correlated by maximizing the efficiency coefficient in order to locate the active speaker among several candidates.

Gurban and Thiran [52] have recently proposed a slightly different approach. While the above presented works do not use any learning procedure, in [52] a training step is required to learn the parameters of the Gaussian Mixture Model (GMM) that is used to estimate the joint *pdf* of audio-video features. This should allow to speed up the correlation computation *at test time* and to foresee a possible real-time implementation. The video feature that is employed is specific to speech : it consists in the difference between the average optical flow on the top and bottom halves of the central part of the mouth region. Audio sources are localized on the video by finding the image regions where samples have highest likelihood to have been generated by the learned joint *pdf*. Results on the CUAVE database demonstrate the effectiveness of the approach, even though the training step needs to be tuned to the type of analyzed sequences, since the learned *pdf* depends on the geometry of the scene (mouth size and orientation, audio energy). Moreover, at training stage mouth regions have to be cropped to compute the video feature and a sufficient amount of data needs to be considered in order to accurately estimate the parameters of the mixture model.

Some considerations can be done at this point concerning information theoretic frameworks and the evolution of audiovisual fusion-localization methods :

- The methods reviewed in this section remove the assumptions on the joint distribution of audiovisual features estimating the *pdf* using Parzen windows or GMM;
- Audio-video representations are still simple, even though video features are becoming more accurate, considering optical flow estimations [13] and even speech specific representations [52];
- With the exception of [52], the training stage is no more considered while the modelling of signal cross-correlations becomes more and more important;
- Information theoretic frameworks suffer from the major problem of the estimation of joint *pdf*s. While on-line estimation using Parzen kernels [13, 18, 42] is parameter-sensitive and requires an amount of data often unavailable in real conversational systems, the learning of GMM [52] seems to depend strongly on the analyzed sequences, resulting somehow little robust and flexible.

2.2.3 Other Research Directions

One of the first work in the field was by Cutler and Davis [27] who conceived a time-delay neural network approach. Audio-video correlations are learned on positive and negative examples using the

neural network, which is then used to find in time and space a speaking person on the input data. Normalized cross correlation between consecutive images is employed as video feature, while cepstral representation is used for the audio signal. Localization capabilities of the system are demonstrated on a small test set including a single person speaking in front of the camera.

An interesting approach was proposed by Smaragdis and Casey [103] who suggested a method to localize and extract audio-video sources applying methodologies borrowed from the BASS field. An optimal modelling and fusion criteria of data are found in a joint manner. Principal Components Analysis (PCA) and Independent Component Analysis (ICA) are performed on audio and video features at the same time, in order to find the maximally independent audio-video subspaces, and thus extract audiovisual independent components. The video is represented with pixel intensities and the audio signal using the amplitude spectrum. This work is interesting since the proposed technique allows not only to localize sources on the video, but to explicitly link and extract meaningful audiovisual structures. However, this method is not able to deal with dynamic scenes.

2.3 Where Are We Now?

The retrieval of correlation between audio and video signals is a non-trivial challenge, since complex relationships between complex signals of different nature have to be modelled. The first works in the field faced the problem using very simple signal representations and correlating them imposing simplifying assumptions on the relationships between audio-video modalities. Since then, researchers have investigated several directions, adopting more accurate signal representations than pixel intensities or audio energy and developing increasingly complex and effective measures to describe audiovisual interdependencies. However, in our opinion there are at least two major shortcomings in the above mentioned fusion schemes that in this thesis we want to analyze and possibly alleviate :

Audiovisual features - Reviewed methods dealing with audiovisual fusion problems basically attempt to build statistical models to capture the relationships between audio and video features. Surprisingly enough however, the features employed are quite simple and barely connected with the physics of the problem : we refer in particular to pixel-related features typically used for video representations. Efficient signal modelling and representation require the use of methods able to capture particular characteristics of each signal kind. A question that arises at this point is : why should we use a representation of video based on a basis of *deltas* (i.e. pixel wise features), if video is made of moving regions surrounded by contours with high geometrical content? Pixel-related quantities seem to us a relatively poor sources of information that have huge dimensionality, that are quite sensitive to noise and that have low semantic content, which makes it impractical to extract and manipulate correlated audiovisual structures. Moreover it is difficult to deal with dynamic scenes, since the variables that are observed (pixel values or related quantities) are static. A very simple example can clarify this concept. If a person is moving back and forth while speaking in front of a camera, pixel intensities on the mouth region change depending on the lips movements *and* on the speaker movement. In this case the evolution of pixel values conveys few or even misleading visual information about the sound source.

Audiovisual fusion criteria - Audio and video features are considered as *random variables* whose degree of correlation is estimated using statistical measures under more or less restrictive assumptions. The estimation of statistical cross-modal correlations forces one to consider an uncomfortable trade-off. Either the statistical relationships between different modalities are supposed to be very simple, assuming for example linearity [54, 82], independence [103] or

mutual Gaussianity [62, 102]. Either complex models involving the estimation of MI [42], HMMs parameters [82] or GMM features [52] have to be conceived if no strong assumption is made, incurring in problems of parameter sensitivity and lack of data. We argue again that in order to better understand audiovisual mechanisms and to improve existing fusion frameworks, an effort should be done to model the structure of the phenomenon looking more in depth into the physics of the problem.

In contrast to previous research works, in this thesis the attention is focused on modelling the observed phenomena, i.e. multi-modal stimuli and their correlations. In particular, we propose models of multi-modal sequences that concisely describe the structure of the considered signals. Such representations allow as well the design of intuitive and precise fusion criteria that do not require the formulation of any complex statistical model describing the relationships between the different modalities. The idea is basically that of defining *meaningful* representations for signals, instead of defining a complex statistical fusion model that has, however, to find correspondences between barely meaningful features. In the case of multimedia sequences for example, if accurate descriptions of the scene are available we can actually think of detecting consistent audiovisual pairs generated by the same phenomenon by simply observing the co-occurrence of interesting audio and video "events". In the next chapter we will develop this topic and we will propose a framework that allows to build relationships between correlated audio and visual data.

2.4 Discussion

In this chapter we have defined the class of signals that we will study in this thesis, i.e. multi-modal (or *heterogeneous multi-channel*) signals. The target applications of this research work will be in the field of audiovisual signal processing, and in particular audiovisual single-microphone source localization and separation. An exhaustive survey of the literature in these fields have been carried out and it has highlighted that, despite an increasing interest in these topics, major issues still have to be addressed. Existing audiovisual fusion techniques typically consider the different signal modality representations as random variables and propose complex statistical models describing relationships between these features, renouncing somehow to model the structural properties of the considered phenomenon.

The next chapter will develop this issue and it will introduce a multi-modal signal model grounded on signal representation techniques that provide an interpretation of the data in terms of salient signal structures. Such multi-modal structures will constitute as well the core of the analysis tools that will be developed throughout all the thesis.

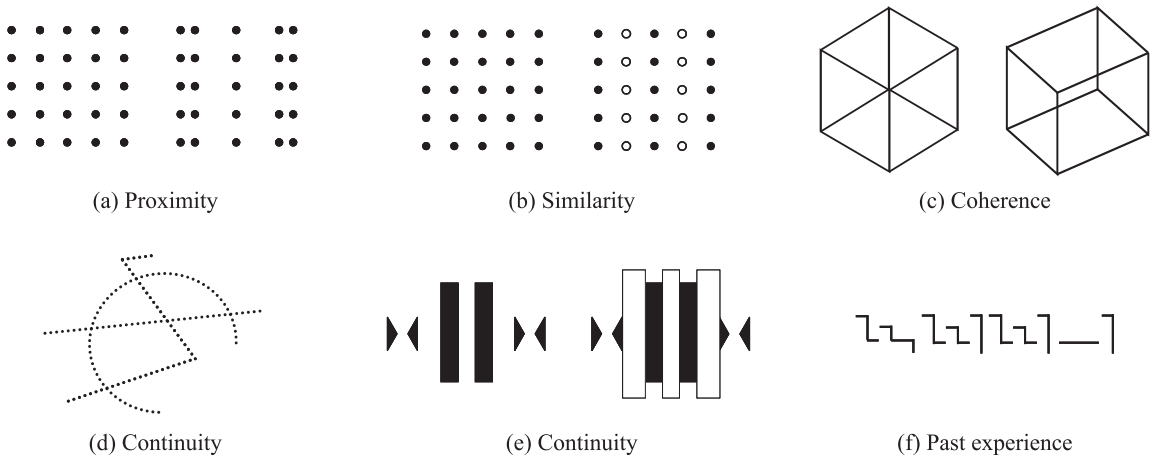
3 Audiovisual Gestalts

In this chapter we introduce and discuss a new framework for detecting correlated audiovisual events in multi-modal signals. In particular, we want to localize the source of a sound in a video sequence. Such task is quite trivial for humans, while it is particularly challenging for automatic systems. It is for this reason that we have decided to study a perceptually-driven approach to audiovisual fusion inspired by the research of Desolneux, Moisan and Morel on *Gestalt theory* and Computer Vision [32–34].

3.1 Gestalt Theory and Audiovisual Processing

First of all, let us briefly introduce what Gestalt theory is. Starting from the first decades of past century, Gestaltists have tried to express all the basic laws that rule human visual perception [61, 117]. The basic set of such laws consists of *grouping laws* : starting from local data, objects are formed by recursively building larger visual objects, i.e. *gestalts*, that share one or more common properties. Such properties represent specific, simple qualities of visual objects. The list of qualities according to which gestalts are built includes proximity, similarity, continuity of direction, amodal completion, closure, constant width, tendency to convexity, symmetry, common motion, past experience [61]. Examples of some of these Gestalt laws "in action" taken from Kanizsa's book [61] are shown in Fig. 3.1. Clearly, such simple rules are not able, alone, to explain the human perception of the world. Thus, more complex principles governing the collaboration and the contrast between gestalt laws and that are active at cognitive level have also been introduced. Here, we will focus our attention on the basic set of simple grouping laws, called by Desolneux and coworkers *partial gestalts* [34]. The interested reader is referred to [61] for an exhaustive presentation of the Gestalt theory of perception.

Interestingly, Gestalt laws have been demonstrated to hold not only for visual perception, but also for other type of sensorial experiences, like acoustic and tactile perception [61]. Moreover, several works in psychophysics and neuroscience have also shown that Gestalt-like rules, and notably temporal proximity, contribute to integrate cross-modal information in humans, both in the case of audiovisual [11, 39, 58, 75, 101, 113, 116] and tactile-visual stimuli [15, 112]. In one of



(a) Proximity (b) Similarity (c) Coherence

(d) Continuity (e) Continuity (f) Past experience

Figure 3.1 – *Examples of Gestalt laws of visual perception (after [61]). In (a), the 25 points on the left could be grouped into many different configurations. However, if two of the columns of points are slightly shifted, the configuration of points acquires a unique, well defined structure. In (b) the same type of effect is obtained by modifying the aspect of two of the columns. In (c), the hexagon on the left becomes a cube by slightly changing the point of view. Interestingly, the figure on the left can be a cube as well, but its configuration as an hexagon is already well defined, while the figure on the right becomes more coherent if it is perceives as a cube. In (d) the continuity of direction contributes to perceive the set of points as three lines. Figure (e) shows an example of the closure law. In the left picture objects are grouped according to their similarity and proximity (pairs of triangles and rectangles). However, if the three white rectangles are superimposed to the drawing as on the right picture, objects are forced to be grouped in such a way to form a partially covered hexagon. This effect is due to the closure law. Finally, in (e) one typically sees a group of independent segments. However, if the page is rotated clockwise by 90°, the word FEEL appears. Since it seems unlikely that the same effect could arise if one does not know the alphabet's letters, in this case the past experience seems to play a determinant role.*

the pioneering studies in the field, Jack and Thurlow [58] discovered that synchronization of visible movements with peaks of speech intensity is the main condition for considering that audiovisual stimuli are originated by the same generating event. In a 1976 paper, McGurk and MacDonald [75] reported an amazing demonstration of the multi-modal nature of speech perception which represents a milestone in the field of sensory integration. They discovered that if the audio syllable "ba" is dubbed onto a visual "ga", one perceives a "da". The effect is induced by the human brain, that integrates audiovisual stimuli that are incongruous but simultaneous. Examples of the McGurk effect can be downloaded through `http://www.media.uio.no/personer/arntm/McGurk_english.html` or `http://www.faculty.ucr.edu/~rosenblu/VSMcGurk.html`. The effect is strong, and it works even if audio and video sequences are from speakers of different genders [48], if the speaking lips are shown upside-down [12] or if the face representation is extremely schematic and if subjects are unaware that they are looking at a face [98]. Demonstrations of these two effects can be linked respectively through `http://www.faculty.ucr.edu/~rosenblu/VSinvertedspeech.html` and `http://www.faculty.ucr.edu/~rosenblu/VSlipreadingdots.html`. Another surprising example of audio-video information integration has been presented in [101] and it represents the first observed illusion induced by a non-visual stimulus, namely sound : when a single flash of light is accompanied by multiple auditory beeps, the single flash is perceived as multiple flashes. A demonstration of the effect can be seen at `http://shamslab.psych.ucla.edu/demos/`.

All this evidence suggests that the synchrony between modalities is an important, strong cue that can be valuably exploited when simultaneously processing multi-modal signals. In fact, as discussed in the previous section, several researchers have exploited audio-video coherence in particular to

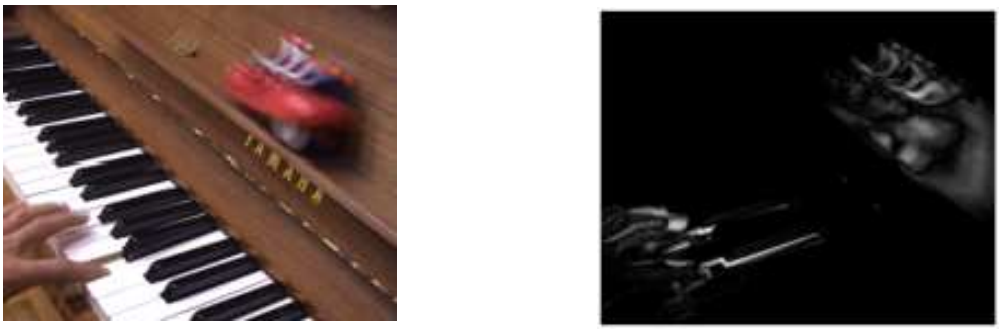


Figure 3.2 – *Video component of an audiovisual sequence including a hand playing the piano and a moving toy car. One sample frame is shown on the left and the corresponding dynamic pixels are on the right : gray-levels represent the absolute value of the difference between the luminance components of two successive frames. Black pixels indicate thus no motion. Observing only the visual motion it is clearly not possible to deduce where the sound source is.*

design audio source localization algorithms in audiovisual sequences [18, 27, 42, 52, 54, 62, 77–79, 82, 102, 103]. It is worth underlying that such task is not trivial and it can be extremely challenging if one decides to consider audio and video modalities separately. It is clear that the audio stream provides no information concerning the spatial location of the source if only a one-microphone signal is available. However, also video information alone can be barely helpful. In Fig. 3.2 it is shown the video component of an audiovisual sequence including a hand playing the piano and a moving toy car. One sample frame is shown on the left and the corresponding dynamic pixels are on the right : gray-levels represent the absolute value of the difference between the luminance components of two successive frames. It is clear that observing only the visual motion it is impossible to decide whether the sound has been produced by the hand playing the piano key (on the bottom left part of the scene) or by the toy car passing by (on the upper right part). Instead, such ambiguity can be solved by using the information present in the audio channel, and in particular by searching for synchronous audiovisual patterns.

Therefore, the idea here is to design an audiovisual source localization algorithm that exploits cross-modal information just like humans do. The localization is accomplished by detecting synchronous audiovisual events, i.e. *audiovisual gestalts*. This detector relies on a simple principle that will be introduced in the next section. Then, in section 3.3 we will discuss more in detail how we can build a model of audiovisual phenomena that will allow the definition of meaningful audiovisual gestalts.

3.2 Gestalts in Computer Vision and *Helmholtz Principle*

A great effort to apply Gestalt theory to Computer Vision has been done in the last years by several researchers [19, 32–34]. Desolneux and colleagues have shown that it exists a very simple and general principle, that they have called *Helmholtz principle*, which allows to decide whether a gestalt is reliable or not. This principle was introduced to try to describe how perception decides to group objects according to a certain quality. It roughly states that an event is perceptually meaningful if it has very low probability to be observed by chance. In [32] this principle is formalized in the following manner. Assume that we are observing r objects $O_1, \ldots, O_r$. Assume that q of them, $O_1, \ldots, O_q$, share a common quality. Is the presence of this common feature a coincidence, or is there a better explanation for it? To answer this question, we do this mental experiment : we assume *a contrario* that the considered quality was uniformly and independently distributed

on all the objects $O_1, \ldots, O_r$. Clearly, the independence assumption is not realistic, but here we are defining an *a contrario* model that grossly represents the absence of relevant events. Then we (mentally) assume that the observed objects are distributed according to this random process. Finally, we ask the question : is the observed set of points probable or not? The Helmholtz principle states that if the expectation of the observed configuration $O_1, \ldots, O_q$ is small, then we are observing a meaningful event, that is, a gestalt.

The Helmholtz principle, conversely to classical statistical methods, does not require a precise model of the observed phenomenon. In fact it coarsely models a statistical background that represents the absence of significant events. An event is considered to be relevant if it has, according to this generic model, a very low probability. In this case, we suppose that such a particular event has a better explanation than chance alone, it is a meaningful gestalt. These events have to be defined so that they correspond qualitatively to some perceptually meaningful structures. We will see in the next section how this can be achieved in the case of audiovisual signals.

3.3 Audiovisual Gestalts

As already stated, the audiovisual gestalt we want to detect is the co-occurrence of an acoustic and a visual event. Such synchronization of events is the main manifestation of a physical phenomenon (utterance of a sound by a speaker for example), whose effects are recorded over different channels (audio and video in this case). As underlined at the end of the previous section, the audiovisual configuration to detect has to be defined in such a manner that it depicts a structure with a certain perceptual meaning. We observe that a visual signal is mainly made of moving regions surrounded by contours with high geometrical content. Measures that consider video pixels independently seem thus a relatively poor source of information that moreover has a huge dimensionality and does not exploit structures in images. In contrast we aim at modelling audiovisual data such that dimensionality gets reduced and salient signal structures are extracted.

Instead of considering raw pixel data, one interesting option is to represent visual information using video approximation techniques that can express an image sequence as a set of video components intended to capture meaningful geometric features (like oriented edges) and their temporal evolution [35, 77–79]. Note that representing the video signal as a set of edge-like patterns that are tracked trough time, we try to define meaningful video structures that obey Gestalt principles. In particular, sets of individual pixels are grouped together and represented with a 3D moving edge according to the rules of proximity, similarity and common motion, which are three of the basic Gestalt grouping laws postulated by Kanizsa [61] (see section 3.1).

Here we employ the video representation algorithm developed by Divorra [35]. The use of geometric video decomposition has at least two main advantages :

- Unlike the case of simple pixel-based representations, when considering image structures that evolve through time we deal with dynamic features that have a true geometrical meaning [35, 77–79];
- Geometric video decompositions provide compact representations of information, allowing a considerable dimensionality reduction of the input signals. This property is particularly appealing in this context, since we have to process signals of very high dimensionality.

In the next section we introduce a signal representation technique based on sparse decompositions over redundant dictionaries of functions. This type of approach will be the basis of the audio and video representation methods that will be used throughout this thesis.

3.4 Sparse Representations on Redundant Dictionaries

Our aim is to represent a digital signal $f \in \mathbb{R}^d$, in terms of its most salient and meaningful structures. In order to capture the large variety of structures present in natural signals, redundant sets of basis functions have to be considered. Let $\mathcal{D} = \{\phi_n\}_{n\in\Omega}$ be a *dictionary* of vectors called *atoms*, with $\phi_n : \mathbb{R}^d \to \mathbb{R}$ and $\|\phi_n\| = 1$ and Ω indexes the (finite) set of all functions composing $\mathcal{D}$. We propose to approximate f by means of a linear expansion into a redundant family of functions :

$$f \approx \sum_{n\in\Gamma} c_n \phi_n \,, \tag{3.1}$$

where Γ is a set of functions s.t. $\Gamma \subset \Omega$ and c_n is a coefficient weighting each component. In this thesis a dictionary is understood as a generic set of atoms containing all available vectors for representing signals based on the model described in (3.1). The dictionary is said to be redundant if the number of functions that compose it is larger than the dimensionality d of the space where f lives. The interest of considering redundant dictionaries resides in their capacity to supply signal representations that are *sparse*, i.e. the cardinality of Γ is much smaller than the dimension of the signal. Given a certain class of signals, one may define a dictionary of functions such that they have a rich collection of shapes in order to adapt better to the characteristics of the signals to represent. Such kind of dictionaries are typically redundant sets of atoms able to provide sparse signal representations of the form of (3.1).

Unfortunately, the decomposition of a signal f on a general redundant dictionary $\mathcal{D}$ is not unique, and several decomposition approaches have been proposed. One popular approach is the method of frames [29], that however does not guarantee the sparsity of the solution. Interesting alternatives are the FOCUSS algorithm [47] or the Basis Pursuit algorithm (BP) [22]. However, these two techniques are basically unusable when dealing with high-dimensional signals like video sequences because of their computational complexity. Here we use the Matching Pursuit algorithm (MP) [73], a simple iterative method to represent a signal according to the model (3.1). MP is a greedy algorithm that decomposes any signal into a linear expansion of waveforms that belong to a dictionary of functions. These waveforms are iteratively chosen to best match signal structures. This characteristic makes the MP approach viable also when considering high-dimensional data : dictionary atoms are picked one by one, while in FOCUSS and BP the whole set of representing functions is chosen at one time. Vectors are iteratively selected from the dictionary by optimizing the signal approximation (in terms of energy) at each step. Even though the expansion is linear, MP is a highly non-linear decomposition algorithm.

MP iteratively picks up the function belonging to $\mathcal{D}$ that best approximates the signal f. The first step of the MP algorithm decomposes f as

$$f = R^0 f = \langle f, \phi_0 \rangle \phi_0 + R^1 f \,, \tag{3.2}$$

where $R^1 f$ is the residual component after projecting f in the subspace described by ϕ_0. Since all elements in $\mathcal{D}$ have by definition a unit norm, it is easy to see from (3.2) that ϕ_0 is orthogonal to $R^1 f$, and this leads to

$$\|f\|^2 = |\langle f, \phi_0 \rangle|^2 + \|R^1 f\|^2 \,. \tag{3.3}$$

To minimize $\|R^1 f\|$ one has to select ϕ_0 such that the coefficient $|\langle f, \phi_0 \rangle|$ is maximum. At the next step, the same procedure is applied to $R^1 f$, which yields :

$$R^1 f = \langle R^1 f, \phi_1 \rangle \phi_1 + R^2 f \,. \tag{3.4}$$

This operation is recursively applied, and after N iterations we decompose f as

$$f = \sum_{n=0}^{N-1} \langle R^n f, \phi_n \rangle \phi_n + R^N f \,. \tag{3.5}$$

Similarly, the energy $\|f\|^2$ is decomposed into :

$$\|f\|^2 = \sum_{n=0}^{N-1} |\langle R^n f, \phi_n \rangle|^2 + \|R^N f\|^2 \,. \tag{3.6}$$

The MP algorithm has been shown to converge, i.e. $\|R^N f\|^2 \to 0$ when $N \to \infty$, and its approximation error decay rate has been shown to be bounded in finite dimension by an exponential [30]. Using the MP technique we can thus approximate f using N terms as

$$f \approx \sum_{n=0}^{N-1} c_n \phi_n \,, \tag{3.7}$$

where $c_n = \langle R^n f, \phi_n \rangle$.

The MP algorithm results practically usable also when dealing with signals of high dimensionality thanks to its iterative structure. Moreover the selected functions are ranked according to their contribution to the approximation of the signal, which makes the algorithm *scalable*. This means that one can choose the degree of accuracy of the representation by choosing the number N of considered functions, allowing flexible and parsimonious data representation.

3.5 Audio-Video Representation and Fusion

In the following sections we will briefly describe the techniques used to represent the audio signal and the video representation algorithm of Divorra, letting the interested reader refer to [35]. Based on such representations, in section 3.5.3 we will define meaningful audiovisual events.

3.5.1 Representation of the Video Signal

We would like to represent a video signal with a set of video components that are able to express the signal in terms of salient, meaningful structures. In this context such components are oriented edges that evolve through time [35]. Considering the sparse signal model expressed by (3.1) and indicating a discrete video signal as $\mathbf{V}(x_1, x_2, t)$, with (x_1, x_2) pixel coordinates and t the frame index, we can write :

$$\mathbf{V}(x_1, x_2, t) \approx \sum_n c_n \phi_n^{(v)}(x_1, x_2, t) \,, \tag{3.8}$$

where c_n is the coefficient for every atom $\phi_n^{(v)}(x_1, x_2, t) : \mathbb{R}^3 \to \mathbb{R}$. The limitation of this type of formulation is that we have to define a redundant dictionary of 3D atoms $\mathcal{D}^{(v)} = \{\phi_n^{(v)}\}_n$ that represent visual structures and their possible evolution through time. It is intuitive to understand that such a dictionary would be amazingly huge even considering a limited set of possible structures and transformations. Thus the computation of a video signal approximation in the form of (3.8) results basically untractable.

However, natural image sequences can be seen as a succession of 2D projected snapshots of 3D objects. Considering these objects to describe smooth trajectories through time, one usually assumes that image sequences are well modelled by smooth transformations of a reference frame [115]. A video sequence can thus be considered as a series of frames represented by a mixture of homogeneous regions and regular contours, where the motion is represented by smooth local deformations of these regions. Regular geometric structures can be represented using parametric over-complete dictionaries of geometric atoms and local deformations are then propagated along the sequence by updating the atoms' parameter field in order to approximate the succession of frames. Applying this

principle and the model (3.1), we assume that a discrete 2D image $I(x_1, x_2)$ can be approximated with a linear combination of atoms retrieved from a redundant dictionary $\mathcal{D}^{(i)} = \{\phi_n^{(i)}\}_n$ of 2D atoms, and we can write :

$$I(x_1, x_2) \approx \sum_n c_n \phi_n^{(i)}(x_1, x_2)\,, \tag{3.9}$$

where n is the summation index and c_n is the coefficient associated to every atom $\phi_n^{(i)}(x_1, x_2)$: $\mathbb{R}^2 \to \mathbb{R}$.

Each video frame is decomposed into a low-pass part, that takes into account the smooth components of images, and a high-pass part, where most of the energy of edge discontinuities lays. The low frequency component is obtained by low-pass filtering and down-sampling the images in the sequence, using the Laplacian-pyramid scheme [17]. We employ here the FIR low-pass filter proposed in [89]. The high-pass frames are obtained by subtracting the low frequency parts from the original frames. These high frequency residual frames which contain the geometric structures of images are represented using MP as

$$I \approx \sum_{n=0}^{N-1} c_n \phi_n^{(i)}\,, \tag{3.10}$$

where $c_n = \langle R^n I, \phi_n^{(i)} \rangle$.

The dictionary $\mathcal{D}^{(i)}$ is built by varying the parameters of a mother function, in such a way that it generates an overcomplete set of functions spanning the input image space. The choice of the *generating function* $g^{(i)}(x_1, x_2)$ is driven by the observation that it should be able to represent well edges on the 2D plane. Thus, it should behave like a smooth scaling function in the direction of the contour and it should approximate the edge transition along the orthogonal one [90, 109]. We use here an edge-detector atom with odd symmetry employed in [90], that is a Gaussian along one axis and the first derivative of a Gaussian along the perpendicular one (see Fig. 3.3). The generating function $g^{(i)}(x_1, x_2)$ is thus expressed as :

$$g^{(i)}(x_1, x_2) = 2x_1 \cdot e^{-(x_1^2 + x_2^2)}\,. \tag{3.11}$$

Each atom $\phi_n^{(i)} = U_n g^{(i)}$ is built by applying a set of geometrical transformations U_n to the mother function $g^{(i)}(x_1, x_2)$. Basically, this set has to contain three transformations :

- Translations $\vec{t} = (t_1, t_2)$ all over the image plane;
- Anisotropic scaling $\vec{s} = (s_1, s_2)$ to adapt the atom to the considered image structure;
- Rotations θ to locally orient the function along the edge.

Any atom $\phi^{(i)}$ in the dictionary rotated by θ, translated by t_1 and t_2, and anisotropically scaled by s_1 and s_2 can thus be written as :

$$\phi^{(i)}(x_1, x_2) = \frac{C}{\sqrt{s_1 s_2}} \cdot 2u \cdot e^{-(u^2 + v^2)}\,, \tag{3.12}$$

where C is a normalization constant an

$$u = \frac{\cos\theta(x_1 - t_1) + \sin\theta(x_2 - t_2)}{s_1}\,, \tag{3.13}$$

and

$$v = \frac{-\sin\theta(x_1 - t_1) + \cos\theta(x_2 - t_2)}{s_2}\,. \tag{3.14}$$

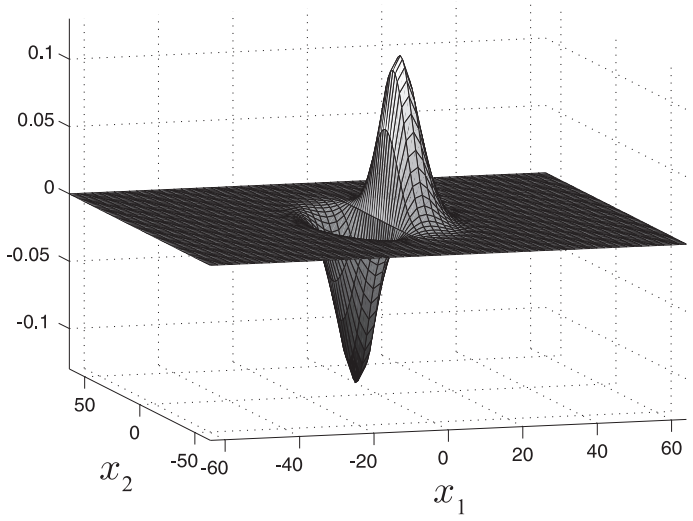


Figure 3.3 – *The generating function* $g^{(i)}(x_1, x_2)$ *expressed by* (3.11).

In order to obtain a dictionary $\mathcal{D}^{(i)}$ that has finite dimension all the atoms parameters are discretized. The translations $\vec{t}$ are kept over a uniform grid that covers the pixels locations. The two scaling factors are quantized in a dyadic way. The range of the scaling factor s_2 is twice as big as s_1 since we want the atoms to be edge detector kernels. The rotation step θ is uniformly quantized.

We consider an approach where 2D spatial primitives obtained in the expansion of a reference frame of the form of (3.10) are tracked from frame to frame. Given a set of images belonging to a sequence, the changes suffered from a frame I_t to I_{t+1} are modelled as the application of an operator F_t to the image I_t such that $I_{t+1} = F_t(I_t)$ and

$$I_{t+1} \approx \sum_{n=0}^{N-1} F_{n_t} \left(c_{n_t} \phi_{n_t}^{(i)} \right) , \tag{3.15}$$

where F_t represents the set of transformations F_{n_t} of all individual atoms that approximate frame t, and the subscript t indicates the time index. The transformations F_{n_t} act on the parameters associated to each atom $\phi_{n_t}^{(i)}$, i.e. the coefficient c_{n_t} and the position, scale and orientation parameters of $\phi_n^{(i)}$ ant frame t. A MP-like approach similar to that used for the first frame is applied to retrieve the new set of atoms $\phi_{n_{t+1}}^{(i)}$ (and the associated parametric transformation F_{n_t}). However, at every greedy decomposition iteration some new criteria have to be considered in order to establish the relationship with the expansion of the reference frame. Only a subset of functions of the general dictionary is considered as candidate functions to represent each deformed atom. This subset is defined according to the past geometrical features of every particular atom in the previous frame, such that only a limited set of transformations (translation, scale and rotation) are possible. This imposes smoothness on the set of deformed primitives, following the assumption of smooth transformation. The formulation of the MP approach to video representation is complex and is treated in detail in [35], to which the interested readers are referred.

To summarize, we want to point out here that the original sparse signal model expressed by (3.8)

has been simplified to :

$$\mathbf{V}(x_1, x_2, t) \approx \sum_{n=0}^{N-1} c_{n(t)} \phi_n^{(v)}(x_1, x_2, t)\,, \tag{3.16}$$

where the coefficients $c_{n(t)}$ vary through time and where each video atom $\phi_n^{(v)}$ is obtained by changing from frame to frame the parameters $(t_{1_n}, t_{2_n}, s_{1_n}, s_{2_n}, \theta_n)$ of a reference 2D atom $\phi_n^{(i)}(x_1, x_2)$ (see also Figure 3.4) :

$$\phi_n^{(v)}(x_1, x_2, t) = \phi_{n(t)}^{(i)}(x_1, x_2)\,. \tag{3.17}$$

A cartoon example of the used approach can be seen in Fig. 3.4(a), where the approximation of a simple synthetic object by means of a single atom is performed. The first and third row of pictures show the original sequence and the second and fourth rows provide the approximation composed of a single geometric term. Figure 3.4(b) shows the parametric representation of the sequence. We see the temporal evolution of the coefficient $c_{n(t)}$, and of the position, scale and orientation parameters. The MP decomposition of the video sequence provides a parametrization of the signal which represents the image geometrical structures *and* their evolution through time. In this way we can track the movements of relevant image features, getting an accurate description of the scene content. Besides, it is important to underline that the stream of video atoms that we consider is absolutely generic. It could be generated using different approximation techniques and it can be used to encode video sequences, as it is shown in [37].

3.5.2 Representation of the Audio Signal

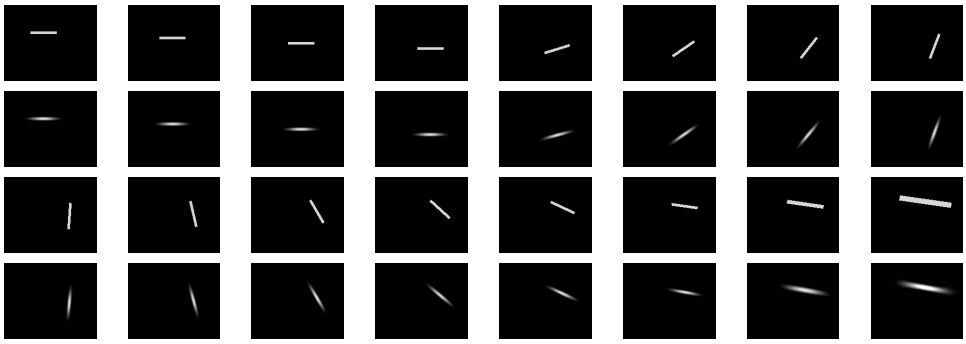
Audio signals have a rich variety of components that the human auditive system is able to perceive (Fig. 3.5). Correlations of the wide diversity of sounds with the also large variety of geometric configurations of the visual stimulus of a mouth are possible. Indeed, this is the main basis for *lip reading*. A positional model of lips may be assigned to each sound and transitional models between sounds can be established.

We consider here a much simpler and generic approach. As already stated, we look for synchrony between audio-video events. An interesting audio event, from our point of view, is the presence of a sound. Therefore, we need an audio feature that simply allows to assess the presence or not of an acoustic event. Finer audio features are unnecessary in this setting, but can be considered to perform more complex tasks.

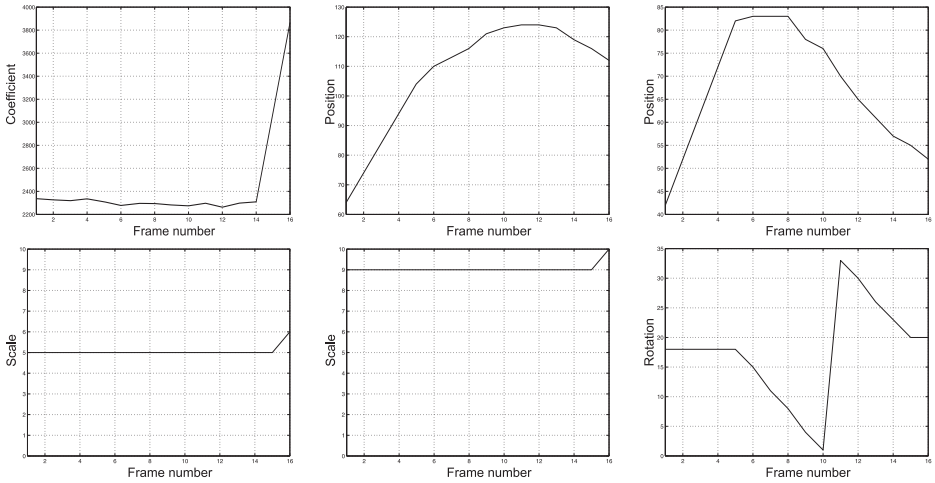
Typical features used to represent audio signals are based on MFCC analysis [93], mainly used in the speech recognition field, and employed in [13, 27, 82, 102]. Simple audio descriptors based on the average acoustic energy are used in [52, 54, 62]. In [18] the audio feature is obtained from the spectrogram of the audio track as the linear combination of the power spectrum coefficients exhibiting the biggest entropy. Fisher and Darrell [42] propose a similar feature that maximizes the mutual information with the video. In all cases, the final feature is a 1D function that is down-sampled in order to obtain the same temporal resolution for audio and video features.

Here, an estimate of audio energy contained per frame is considered. To compute such an estimate, we exploit the properties of sparse signal representations over redundant dictionaries using MP, that point out the most relevant signal structures. The audio signal $a(t)$ is decomposed over a redundant dictionary $\mathcal{D}^{(a)}$, composed of unit norm atoms. The family of atoms that compose $\mathcal{D}^{(a)}$ is generated by scaling by s, translating in time by u and modulating in frequency by ξ a 1D generating function $g^{(a)}(t)$. An atom belonging to $\mathcal{D}^{(a)} = \{\phi_k^{(a)}\}_k$ can thus be expressed as

$$\phi_k^{(a)}(t) = \frac{1}{\sqrt{s}} g^{(a)}\Big(\frac{t-u}{s}\Big) e^{i\xi t}\,. \tag{3.18}$$



(a) Synthetic sequence approximated by 1 atom : first and third row show the original sequence made by a simple moving object. Second and fourth row depict the different slices that form a 3D geometric atom.



(b) Parameter evolution of the approximated object; from left to right and from up down, we find : coefficient c, horizontal position t_1, vertical position t_2, short axis scale s_1, long axis scale s_2, rotation θ.

Figure 3.4 – *Approximation of a synthetic scene by means of a 2D time-evolving atom.*

In our case, we consider a dictionary of Gabor atoms, that is, the generating function $g^{(a)}(t)$ is a normalized Gaussian window. The choice of a Gabor dictionary is motivated by the optimal time-frequency localization of the Gaussian core [51].

As in the case of images, we can express a K-terms approximation of the signal $a(t)$ as

$$a(t) \approx \sum_{k=0}^{K-1} c_k \phi_k^{(a)} , \tag{3.19}$$

where $c_k = \langle R^k a, \phi_k^{(a)} \rangle$.

An estimate of the time-frequency energy distribution of the real function $a(t)$ can be derived from its MP decomposition by summing the Wigner-Ville distributions $W\phi^{(a)}(t,\omega)$ of the obtained atoms [73] :

$$E_a(t,\omega) \approx \sum_{k=0}^{K-1} |\langle R^k a, \phi_k^{(a)} \rangle|^2 \cdot W\phi_k^{(a)}(t,\omega) . \tag{3.20}$$

If $g^{(a)}(t)$ is, as in this case, the Gaussian window, its Wigner-Ville distribution is

$$Wg^{(a)}(t,\omega) = 2e^{-2\pi(t^2+(\omega/2\pi)^2)} . \tag{3.21}$$

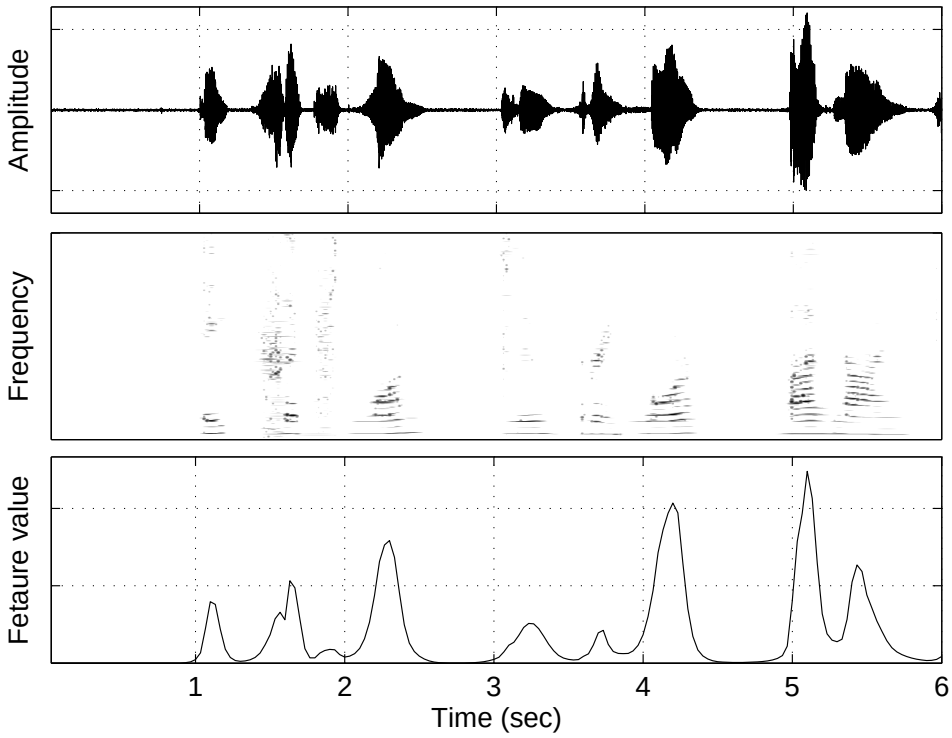


Figure 3.5 – *Audio signal of a subject uttering digits in English sampled at a frequency of 8 kHz [Top], its time-frequency energy distribution $E_a(t,\omega)$ [Middle], and the estimated audio feature $f_a(t)$ [Bottom]. The signal is decomposed using 1000 Gabor atoms. The color map of the time-frequency plane image goes from white to black, and the darkness of a pixel represents the value of the energy at each time-frequency location.*

The time-frequency energy distribution $E_a(t,\omega)$ is a sum of 2D Gaussian functions, whose positions and variances along time and frequency axes depend on the parameters s_k, u_k and ξ_k. One of the analyzed signals and its time-frequency energy distribution are shown in Fig. 3.5. On the top picture the audio signal of a person uttering digits in English sampled at 8 kHz is shown, while the plot of its energy distribution is on the picture in the middle.

Construction of the Audio Feature

The audio representation that we obtain from the MP decomposition is not directly exploitable as it is and it has to be further processed in order to obtain a function that is comparable with the evolution of the video parameters. We require audio features composed of the same number T of samples as the MP video features. Moreover, we would like to depict the audio signal with only one time-evolving feature, in order to speed-up the computation and to simplify the problem formulation.

Our audio feature $f_a(t)$ is obtained by estimating the energy present at each time instant, where the time-frequency energy distribution of the audio signal is found by decomposing it with the MP algorithm according to (3.20) :

$$f_a(t) = \sum_{k=0}^{K-1} |\langle R^k a, \phi_k^{(a)} \rangle|^2 \cdot \int_{-\infty}^{+\infty} W\phi_k^{(a)}(t,\omega) d\omega \,. \tag{3.22}$$

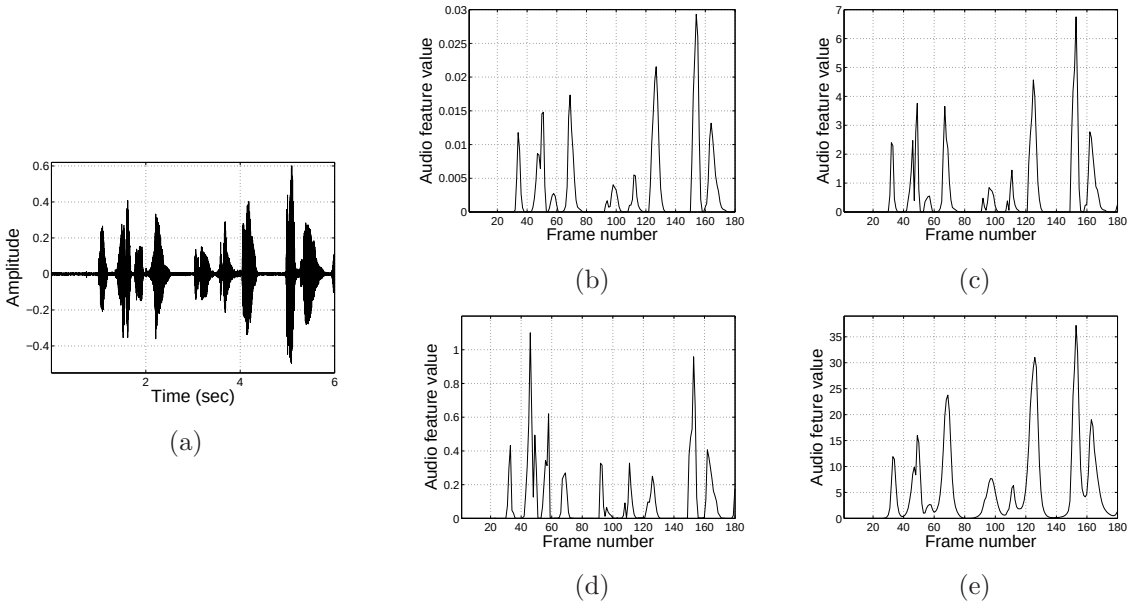


Figure 3.6 – *The signal of Fig. 3.5 is drawn in (a). The average, over a time window of two video frames, of the squared modulus of the audio signal is shown in (b), the average over frequencies of the energy spectrogram of the audio signal in (c), the mean over frequencies of the energy spectrogram after MFCC processing in (d) and the per-frame audio energy estimated from the MP decomposition in (e).*

Note that now the Wigner-Ville distributions are projected over the time axis. The so-obtained estimate of the audio energy per time instant is down-sampled to the temporal frequency of the video, in order to get a convenient number T of time samples. In fact, our feature is similar to those described in [18, 42], with the difference that we attribute to each frequency component the same importance, while in [18, 42] frequency bands are weighted optimizing some audio-video coherence criteria.

The audio characteristic $f_a(t)$ that has been just described has been compared with three other audio features. In Figure 3.6, the signal of Figure 3.5 and these four audio features associated to it are depicted :

- We draw in Fig. 3.6(b) a feature based on the average, over a time window spanning two video frames (in this case 534 audio samples considering a soundtrack at 8 kHz and the associated video at 29.97 frames/sec), of the squared modulus of the audio signal;
- Figure 3.6(c) shows another audio feature computed from the average over frequencies of the energy spectrogram of the signal. The spectrogram is computed as the magnitude of the windowed discrete-time Fourier transform of the signal using a sliding window. The energy distribution is given by the squared absolute value of such time-frequency function;
- Figure 3.6(d) shows a third feature based on the mean over frequencies of the energy spectrogram of the audio signal after MFCC processing [93]. In this case, the spectrogram is reconstructed after processing it using a Mel filter bank composed of 40 filters and taking the $\log_{10}$ of the output. The energy distribution is the squared absolute value of the time-frequency function;
- We draw in Fig. 3.6(e) the audio feature $f_a(t)$ obtained by estimating the per-frame audio energy using (3.22);

The four features behave similarly, and we have chosen the fourth one since it exhibits a smoother and more regular profile (see Fig. 3.6). This is due to the sparseness and the fine time-frequency resolution of the dictionary decomposition, that allow to obtain a description that captures nicely the evolution of the audio track, filtering out most of the signal noise. Moreover, informal tests on a set of real-world sequences have confirmed our intuition, showing that slightly better audiovisual fusion results are obtained when the audio feature (3.22) is used in our proposed framework.

3.5.3 Audiovisual Fusion

The audio feature $f_a(t)$ basically estimates the average energy present in the audio signal $a(t)$. The output of the MP video algorithm, instead, is a set of atom parameters describing the temporal evolution of the video features. From the positions, we can compute the displacement of each video atom and thus estimate the movement of important visual structures. For each video atom $\phi_n^{(v)}$ we compute the absolute value of the displacement as

$$d_n(t) = \sqrt{t_{1_n}^2(t) + t_{2_n}^2(t)}\,,$$

where $t_{1_n}(t)$ and $t_{2_n}(t)$ represent the evolution through time respectively of the horizontal and vertical positions of the atom. In order to be more easily compared to the audio feature and to filter out small spurious movements, we convolve the video feature $d_n(t)$ with a Gaussian kernel, obtaining a smooth function like the one depicted in Fig. 3.7(c).

We have now one audio feature and N video features describing the movement of relevant visual features, where N is the number of atoms used to represent the video. Each of these variables has the same number of samples T, since we down-sample $f_a(t)$ that has a higher temporal resolution. Peaks in these signals suggest the presence of an event. In the video case, it can be the movement with respect to a certain equilibrium position (e.g. lips opening and closing). For the audio, a peak indicates the presence of a sound. The temporal proximity of such audio and video peaks suggests the presence of a gestalt reflecting two expressions of the same phenomenon (production of a sound). Thus, for a given feature vector $\mathsf{x}(t)$ we build an *activation vector* $\mathsf{y}(t)$ which is based on the information about the peaks locations. First, we detect the peaks in the audio feature and in each of the N video features, obtaining vectors that equal 1 where peaks occur and 0 otherwise. Peaks are found by simply detecting positive signal slopes that are followed by negative slopes. Then, such vectors are filtered with a rectangular window of size W that models delays and uncertainty. An activation vector describes the presence of an event associated to the corresponding signal. It has value 1 when the feature is "active", and 0 otherwise.

We end up with one activation vector for the audio, $\mathsf{y}^{(a)}(t)$, and N activation vectors $\mathsf{y}_n^{(v)}(t)$, one for each video atom. By computing a logical AND between $\mathsf{y}^{(a)}(t)$ and all the video activation vectors constructed over a given observation time slot, we build N vectors, denoted as *synchronization vectors* $\mathbf{s}_n(t)$. The vectors $\mathbf{s}_n$ equal 1 at time instants at which both audio and video atoms are active and 0 otherwise. Thus, the number of 1 in the vector indicates the degree of synchronization between the audio-video pair. Figure 3.7 summarizes the construction of one synchronization vector $\mathbf{s}(t)$.

3.6 Detection of Audiovisual Meaningful Events

Once synchronization vectors are available, we need a method to select those vectors (and thus those audiovisual structures) associated to *meaningful* audio-video pairs. We want to do that in an automatic way, tuning as less parameters as possible. In the next section we will show how we can

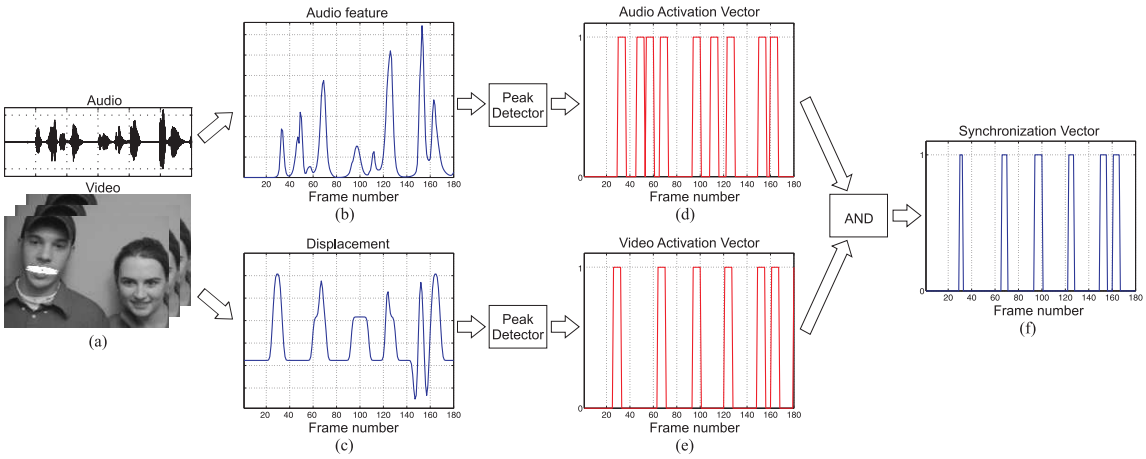


Figure 3.7 – *Scheme of the proposed audiovisual fusion criterion. Starting from the audiovisual sequence (a), we compute the audio feature $f_a(t)$ (b), and the displacement feature for a video atom representing the speaker's mouth (c). The two features exhibit a remarkable synchrony. From these signals we extract the audio energy peaks and the displacement peaks, and the activation vectors $y^{(a)}(t)$ and $y^{(v)}(t)$ are built (d–e). The synchronization vector $\mathbf{s}(t)$ is created computing the logical AND between the audio-video activation vectors (f).*

build a multi-modal event detector based on the Helmholtz grouping law presented in section 3.2. The parameters of the algorithm reduce to just one, from which the detection accuracy weakly depends.

3.6.1 An Audiovisual Event Detector Based on the Helmholtz Principle

At this point of the reasoning for each video atom we have one synchronization vector $\mathbf{s}_n(t)$. Suppose that we observe a synchronization vector of length r (i.e. that is built over an observation window of r samples), and let the number of 1 in such vector be equal to q. We can ask ourselves : is the number q big enough, so that we can consider the corresponding video atom correlated with the audio signal? Or the co-occurrence of audio and video events is due only to chance? We can answer these questions using the Helmholtz principle.

We first have to define the background *a contrario* model which corresponds to the absence of correlated audiovisual events. In this case the observations $\mathbf{s}_n(t)$ are considered as independently identically distributed (i.i.d.) random variables. Since the general form of their distribution is unknown (anyway, it is not reasonable to assume that a single distribution could account for all the sequences), the empirical distribution is considered [32]. Integrating this distribution yields the function $P_{\mathbf{s}}(X)$, where X is a random variable distributed according to the empirical distribution of the observed values $\mathbf{s}_n(t)$ (with $n = 1, \ldots, N$).

Let A be a video atom with corresponding synchronization vector $\mathbf{s}_{\mathsf{A}}$ of length r, and let q be the number of points at which $\mathbf{s}_{\mathsf{A}}$ assumes value 1. Let us define the event $E =$ *"At least q points of a synchronization vector $\mathbf{s}_{\mathsf{A}}$ of size r keep a value equal to 1"*. Thus, according to the background model, the probability of the event E, $P(E)$,

$$P(E) = \mathcal{B}(q, r, P_{\mathbf{s}}(\mathbf{s}_{\mathsf{A}} = 1)) \,, \tag{3.23}$$

where $P_{\mathbf{s}}(\mathbf{s}_{\mathsf{A}} = 1)$ is directly deduced from $P_{\mathbf{s}}(X)$ and $\mathcal{B}(q, r, p)$ is the tail of a binomial distribution :

$$\mathcal{B}(q, r, p) = \sum_{i=q}^{r} \binom{r}{i} p^i (1-p)^{n-i} \,. \tag{3.24}$$

According to these notions, we can now define an ε-*meaningful* video atom. Let us stress that in this context, the meaningfulness of a video atom is referred to its correlation with the audio signal.

Definition 1. *For a given atom* A *with corresponding synchronization vector* $\mathbf{s}_{\mathsf{A}}$ *of size* r *and containing* q *matching points (i.e.* q *values equal to 1), we define the "number of false alarms" (NFA) as :*

$$NFA(\mathsf{A}) = N \cdot \mathcal{B}(q, r, P(\mathbf{s}_{\mathsf{A}} = 1)) \,, \tag{3.25}$$

where N *is the total number of candidate configurations to be tested. In this case* N *is the number of video atoms used for the decomposition of the sequence.*

An atom A *is said to be* ε*-meaningful if* $NFA(\mathsf{A}) \leq \varepsilon$*.*

It is easy to demonstrate that the expected number of ε-meaningful video atoms in a sequence, according to the *a contrario* model, is less then ε [19, 32]. Moreover, it is possible to show that the number q of matching points in a synchronization vector that are required to be significative depends on the logarithm of ε and N [19, 32]. This means that the detection results are robust to variations of these values.

Setting the *Meaningfulness Threshold* ε The value of ε controls the number of false detections. Setting ε equal to 1, as in [19], means that the expected number of false detections in a sequence distributed according to the background model is less than 1. However, the hypothesis of independence, especially for what concerns the video representation, is far from being realistic since the MP video algorithm exploits the correlation between neighboring atoms [35]. Because of that, some video atoms exhibit NFA smaller then $\varepsilon = 1$, even without being correlated with the audio. One solution is that of considering a smaller value of ε, as it is done in [32] where $\varepsilon = 1/10$.

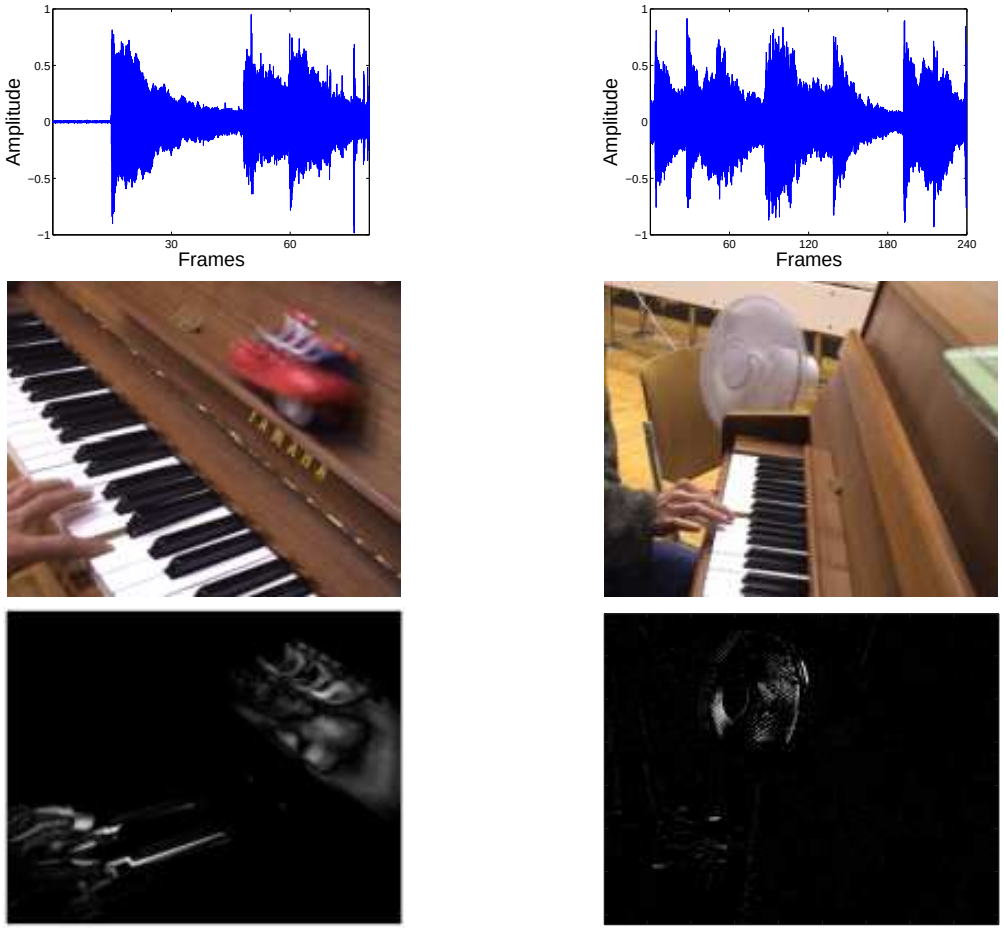
However, better results can be achieved by exploiting some additional knowledge about the scene. Here, we are implicitly assuming that a single audiovisual source is observed at each time instant. Thus, the solution we want to find should be well localized in the image plane. Following this reasoning, we can test multiple values of ε (smaller than 1), keeping the solution which is more localized in space. By doing that, we basically do not fix any detection threshold. Instead, we browse a set of interesting solutions and we chose the most suitable one.

In practice, what we will do is to consider a set of ε_i uniformly spaced in a logarithmic scale between ε_{MIN} and 1. For each value ε_i, we obtain a set of video atoms G_i for which $NFA(\mathsf{A}) \leq \varepsilon_i$, with $\mathsf{A} \in G_i$. For each group G_i, the variances along the horizontal (var_{x_1}) and vertical positions (var_{x_2}) are computed and the maximum value $V_{G_i} = \max\{\mathrm{var}_{x_1}(G_i), \mathrm{var}_{x_2}(G_i)\}$ is kept. Clearly, a set of video atoms can be composed of only one function : in that case the variance V_{G_i} is equal to zero. If a group is empty, its variance is set to a very high value (ideally infinite). This is done to avoid the algorithm to search for a very small threshold ε_i for which the corresponding group G_i is empty and has thus zero variance. Our considered solution G^* is the set of atoms that exhibits the smallest variance V_{G^*}.

3.7 Experiments

We show here how the proposed framework is used to locate the source of an audio signal in real video sequences.

Audio Source Localization The first test involves two clips, denoted as Piano 1 and Piano 2. They both show a hand playing piano while some distracting visual and acoustic noise is present.



(a) (b)

Figure 3.8 – *Test sequences Piano 1 (a) and Piano 2 (b). [Top] Audio tracks, [Middle] sample frames, [Bottom] corresponding dynamic pixels : gray-levels represent the absolute value of the difference between the luminance components of two successive frames. Black pixels indicate thus no motion.*

Sample raw frames of the sequences are shown in Fig. 3.8. In Piano 1 a toy car is passing through the scene, while in Piano 2 a ventilator is on and it is moving from left to right. These examples have been chosen to demonstrate the robustness of the proposed algorithm to audio distractors, thanks to the de-noising properties of the audio MP decomposition, and to video distractors both of constant velocity (Piano 1) and oscillating (Piano 2). The clips were recorded at 25 frames/sec (fps) at a resolution of 144×180 pixels and only their luminance components were considered. The soundtrack was collected at 44 kHz and sub-sampled to 8 kHz.

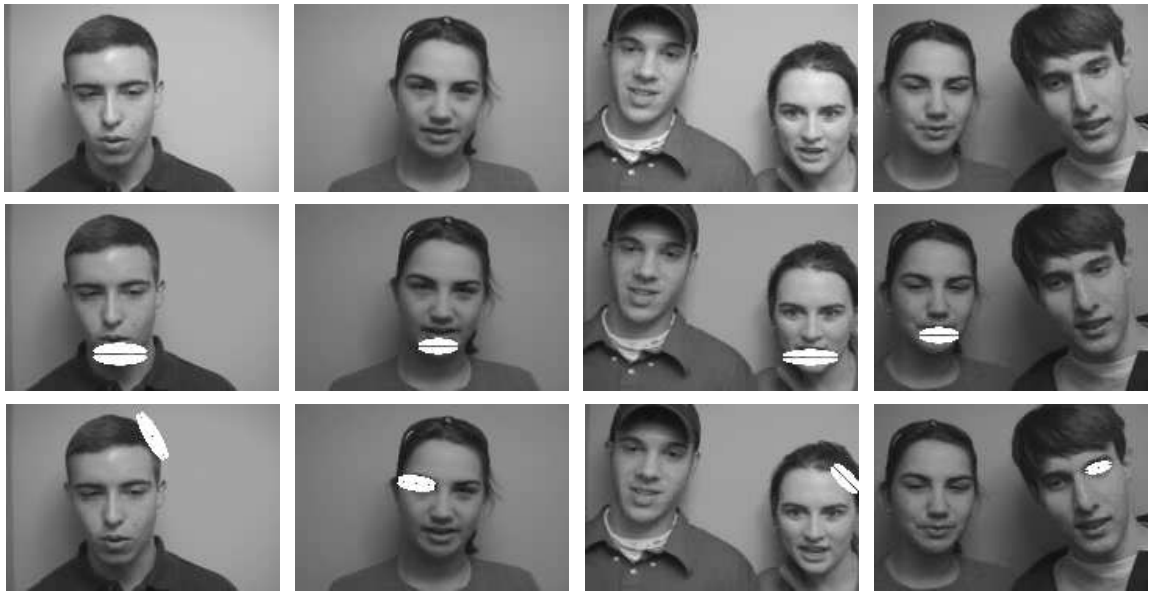
Image sequences are represented with 50 video atoms using the procedure described in section 3.5.1, while the audio track is decomposed using 1000 Gabor atoms whose window lengths range from 512 to 16384 time samples, using the implementation of MP for 1D signals of the *LastWave* software package [50]. The number of basis functions used for the decomposition of the image and audio sequences is heuristically chosen for these experiments, in order to get convenient representations. However, a distortion criteria can be easily set, to automatically determine the required number of atoms. Based on such decompositions, the audio and video features are extracted and the activation vectors are built using a window of size $W = 7$. The set of meaningful atoms G^* is selected using $\varepsilon_{MIN} = 10^{-5}$ and the thresholds $\varepsilon_i = \{10^{-5}, 10^{-4.5}, 10^{-4}, 10^{-3.5}, \ldots, 1\}$.

Figure 3.9 – *Results of the proposed algorithm run on the clip Piano 1. The most correlated atoms, highlighted in white, represent the player's fingers. The moving toy car is not detected.*

Figure 3.10 – *Results for the sequence Piano 2. The correlated atoms, highlighted in white, are on the player's fingers and the piano keys. The oscillating ventilator is not detected.*

In order to take into account the dynamics of the scene, a sliding observation window over which the synchronization vectors are computed has to be employed. A window of 60 frames length is used to detect the video atoms that are more correlated with the audio following the procedure described in section 3.5 and section 3.6. The observation window is then shifted by 20 samples and the procedure iterated. The values of window length and shift have been chosen considering a trade-off between the response time delay of the system and the robustness of the association. However, the algorithm is basically parameter-free since all the values that have to be set are fixed for all the experiments. Moreover, the choice of none of the parameters results to be critical.

Figure 3.9 shows resulting sample frames of the algorithm run on the sequence Piano 1. In white we highlight the footprints of the video atoms that are found to be more correlated with the soundtrack. The player's fingers are detected as sound sources. The moving toy car introduces a considerable distracting motion (see Fig. 3.8(a)) and a non-negligible acoustic noise. However, it is filtered out by the cross-modal localization algorithm. Figure 3.10 shows the same type of results for clip Piano 2. It is interesting to remark that in this case the visual distractor (the ventilator) does not have a constant velocity as in the previous case, but it is oscillating in the background. This results in peaks in the video activation vectors associated to the ventilator's edges. However, these oscillating structures are not detected as correlated with the audio, since they are not synchronous with the audio activation peaks. Both these clips can be downloaded from `http://lts2www.epfl.ch/~monaci/ag.html`.



(a) (b) (c) (d)

Figure 3.11 – *Results for* Experiment 1 *: the first row shows the original video frames, the second row shows the white footprints of the video atoms correlated with the corresponding audio signal. In all cases, the speaker's mouth is correctly detected. In the third row the more correlated video atoms for an incongruous audio source are plotted.*

Speaker Localization A second set of experiments has been carried out to test the proposed algorithm in a multi-modal speaker localization task. To this end we have used real-world video streams representing one or two persons speaking and moving in front of a camera. The test clips are taken from the *individuals* and *groups* sections of the CUAVE database [88]*. The video data was recorded at 29.97 fps and at a resolution of 480×720 pixels. The size of the clips has been then reduced to 120×176 pixels to be more easily and quickly processed. The soundtrack was collected at 44 kHz and sub-sampled to 8 kHz. The setting of the experiments is the same described above and all the parameters keep the same values. All the test video clips involved in the speaker localization task can be linked through `http://lts2www.epfl.ch/~monaci/multimodal.html`.

In a first series of experiments, called *Experiment 1*, we consider sequences involving only one active speaker. We have used clips consisting of one person standing in front of a camera reading digit strings, and videos involving two persons, only one of which is speaking. Each sequence lasts about 6–8 seconds. We want to point out that in this experiment, since only one source is present, no sliding analysis window is used (i.e. audio-video correlation is computed accumulating evidence from the whole sequence). Snapshots of some of the analyzed clips are shown in the first row of Fig. 3.11. We show here four non-trivial cases : speakers in sequence (a) and (b) move left and right and back and forth while uttering, the left person in clip (c) clearly mouths the text which is being pronounced by the right speaker and finally, the right subject in (d) moves significantly while the left person is speaking. In the second row of Fig. 3.11, the image structures that are more correlated with the corresponding soundtrack are highlighted in white. The audiovisual correspondence is assessed following the methodology described above and using the entire length of the sequence. The third row of Fig. 3.11 illustrates the video components that are more correlated with the audio signal of a different video sequence. Image sequences involving only one person are represented using 30 video atoms, while sequences with two subjects are represented with 50 functions. All the audio

*Only the luminance component of the video sequences has been considered.

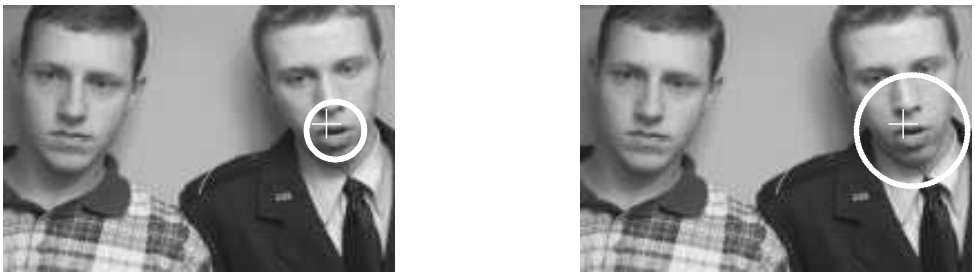
Figure 3.12 – *Results for* Experiment 2 *: in the first two samples the left person is speaking, while in the third the right one it is. The most correlated 3D atoms are highlighted in white. The mouth and the chin of the correct speaker are detected.*

tracks are decomposed with MP using 1000 Gabor functions.

The experimental results show that the proposed methodology allows to clearly locate and track the speakers mouth. In all the tested sequences, the algorithm chooses those visual primitives that constitute the mouth and/or chin structures of the speaker. Even when the active speaker moves, as in Fig. 3.11(a) and (b), or in presence of distracting motion (Fig. 3.11(c), (d)), the source of the sound signal is detected. On the contrary, when the video sequence is dubbed with an incongruous audio track, visual primitives which do not represent the speaker's mouth are typically detected (Fig. 3.11, third row). We expected such a behavior, since the proposed methodology does not simply extract moving structures, but it detects those geometric features that evolve synchronously with the audio. Finally, it is interesting to remark how video atoms adapt their orientation and shape according to the geometric characteristics of the structures they represent. Such information can be exploited in a successive stage of processing, in order to estimate the size, orientation and position of the speaker in the scene.

In a second series of experiments, *Experiment 2*, we have analyzed clips involving two active speakers arranged as in Fig. 3.12. The videos show two persons taking turn in reading series of digits and last about 20 seconds. The test clips are referred to with the names they have on the CUAVE dataset, i.e. g01, g04,.... The image sequences are represented with 50 video atoms and the audio signals are decomposed using 2000 or 3000 Gabor atoms, depending on the length of the clip. Figure 3.12 shows results for sequence g22. In the first two sample frames the left person is speaking, while in the third one the right person is speaking. The sequence is non-trivial, since the left person mouths the digits which are being uttered by the right speaker. The algorithm is able to correctly localize the mouth and the chin of the current speaker. It is interesting to remark how video atoms correlated with the sound shift from one speaker's mouth to the other, handling the dynamics of the scene.

In order to quantify the accuracy of the proposed algorithm, we have manually labelled the center of the speaker's mouth in the test sequences. The active speaker's mouth is considered to be correctly detected if the position of the most correlated video atom falls within a circle of diameter D centered in the labelled mouth center. If more than one atom is chosen, an atoms' centroid is estimated whose position on the image plane is given by the average of the single atoms coordinates. Since correlated atoms are detected every 20 frames, mouth labels are placed with this same frequency throughout each sequence, and performances are thus evaluated at test points distant 20 samples one from the other. In total, we have analyzed 273 test points. The values of the diameter D that are considered are 25 and 50 pixels. Figure 3.13 shows the regions of correct mouth detection on a frame of the test sequence g04 for the two values of D. The white markers indicate the position of the video atoms that are found to be more correlated with the audio. The values of D have been chosen so that we can compare the results with those presented in [82].



(a) (b)

Figure 3.13 – *Regions of correct mouth detection for $D = 25$ (a) and $D = 50$ (b). The white crosses indicate the position of the most correlated video atom.*

Nock and colleagues [82] propose a method to detect the mouth of the speaker founding the image zone over which the mutual information between audio and video features is maximized. As in our algorithm, in [82] mutual information values are estimated using a sliding time window of 60 frames that is shifted in time with steps of 30 frames. The goodness of the detection is assessed using the criterion that we use here, with the only difference that in [82] the speaker's mouth is considered to be correctly located if it is placed within a *square* of $L \times L$ pixels centered on the manually labelled mouth center. The considered values of L are 100 and 200 pixel. Thus, taking into account a down-sampling factor of 4 that we have applied to the video sequences, the areas of correct mouth detection are comparable. However, we must note that the test clips used in [82] could not exactly coincide with those used in these experiments, since the original sequences have been cropped in both cases.

Table 3.1 summarizes the results obtained for the two methods in term of percentage of test points at which the speaker's mouth is correctly detected. With the only exception of sequence `g11`, the proposed scheme seems to outperform Nock's method, considerably improving the detection accuracy. Our proposed method compares particularly favorably with Nock's one when the smaller region of correct mouth detection is considered and for challenging sequences where some distracting motion is present. To be fair, we recall that the considered test sets do not completely coincide, even if we have analyzed a larger number of test points (273 in our case, 252 in the cited paper). Results denote a superiority of the proposed algorithm, also considering that our correct mouth detection area is $4/\pi$ times smaller than in [82] because of the circular shape of the window. Moreover, a large fraction of errors is due to the delay introduced by the sliding observation window that causes an incorrect detection when the speaker changes. Such errors are practically imperceptible for a human observer, as can be checked observing the complete resulting sequences, that are available at `http://lts2www.epfl.ch/~monaci/multimodal.html`. We want to underline again that in contrast to previous methods, we do not simply seek for the video region that maximizes the correlation with the audio, but more generally we look for image zones whose synchrony with the audio are above a saliency threshold. This threshold does not require to be tuned, since a set of meaningful thresholds is fixed in advance and the one giving the most suitable solution is adopted.

The audio-video gestalts that are detected have a high semantic meaning. This allows to extract and manipulate these structures in a simple and intuitive way. For example, it is possible to reconstruct the scene using only those video atoms that are consistent with the audio track by simply encoding the video sequence with 3D atoms that are close to the detected sound source. Figure 3.14 shows sample raw frames of clip `g20` and their reconstruction obtained by summing to the low-pass images those video atoms that are closer than 80 pixels to the estimated sound source.

Sequence	Nock[82]*		Proposed	
	$L = 100$	$L = 200$	$D = 25$	$D = 50$
g01	-	-	86	*95*
g04	-	-	95	*86*
g11	44	*69*	46	*54*
g12	46	*68*	75	*82*
g13	19	*25*	82	*82*
g15	65	*70*	83	*83*
g19	41	*41*	87	*87*
g20	89	*93*	90	*93*
g21	75	*79*	78	*81*
g22	74	*79*	87	*87*

Table 3.1 – *Audiovisual source localization results expressed in percentage of correct detections. Results in the second column should be compared with those in the fourth one (in roman), while the third column should be compared with the fifth (in italic). * These values should be considered as indicative (see text).*

Figure 3.14 – *Sample raw frames of clip g20 [Top] and their reconstruction using only video atoms close to the estimated sound source [Bottom]. On the first sample the left person is speaking while on the second one the right person is speaking. The resulting video sequence can be linked through* `http://lts2www.epfl.ch/~monaci/ag.html`.

The reconstructed images can be seen as *audiovisual key frames* that focus on the sound source at a given time instant. Moreover, in a compression application scenario, a sequence can be selectively encoded using only video atoms associated with the soundtrack, saving bits for the coding while keeping the salient information about the scene.

3.8 Discussion

In this chapter we have presented a novel framework for the cross-modal fusion of audiovisual signals. The proposed audiovisual events detection method features several interesting properties :

- *The algorithm exploits the inherent physical structures of the observed phenomenon.* This allows the design of intuitive and effective audiovisual fusion criteria and demonstrates that temporal proximity between audiovisual events is a key ingredient for cross-modal integration of information. The proposed method exhibits robustness to significant audio-video distractors. In addition, the considered audiovisual structures have a high semantic role and can be easily extracted and manipulated;
- *The algorithm naturally deals with dynamic scenes*;
- *There is no parameter to tune.* All parameters are fixed and from informal tests the algorithm performances turn out to be robust to significant variations of their values;
- *Visual information is described in a very concise fashion.* For example, instead of processing $144 \times 180 = 25960$ time-evolving variables (pixel intensities), we consider only 50 variables (atoms displacements);
- *The atoms streams employed here are completely general*, could be generated by algorithms other than MP and can be used to encode the audio and video sequences;
- *The description of the scene is extremely rich.* The audio and video atomic decompositions carry a large amount of information (e.g. size and orientation of video structures or time-frequency characteristics of audio entities) that can be exploited at successive processing stages, as we will see in the following of this thesis.

The core of our approach are the employed signal representation methods that decompose multimodal signals over redundant dictionaries of atoms, obtaining concise representations that describe the structural properties of the data. This allows to define meaningful audio-video events (*gestalts*) that can be detected using a simple rule, the Helmholtz principle.

In the next chapters we will analyze more in details these audio and video representation techniques, studying their characteristics and trying to relieve their flaws and to exploit their strengths. We will start with the video approximation method in next chapter and we will continue with the audio decomposition in Chapter 5.

Tracking Atoms with Particles

4

One of the key ingredients of the audiovisual fusion framework introduced in Chapter 3 is the video representation technique, which allows to express a complex, high-dimensional video signal as a sparse sum of salient geometric terms that are easy to manipulate. The video decomposition is obtained using the video MP algorithm of Divorra [35]. Although effective for audiovisual source localization [77–79], the 3D MP algorithm is formally and computationally complex. Here we want to formalize the atom tracking problem in a more agile and well grounded fashion, in order to allow an easier and more intuitive understanding of the results. This should allow as well to improve and extend in a natural and elegant fashion the proposed algorithm, as we will discuss in the last section of the chapter.

4.1 Tracking Visual Features

The ability of tracking relevant structures of moving images provides spatio-temporal information that is intrinsically meaningful for the representation of the video signal. In the video MP algorithm this is achieved representing a reference frame as a sparse sum of geometric atoms taken from a redundant dictionary. These structures are then tracked through time, decomposing the subsequent frames with a modified MP algorithm that uses *a priori* information inherited from previous frames [35, 36]. In our work we are interested in the ability of the algorithm to track moving edges, as they represent the motion of relevant video structures, the key information we want to obtain. However, the video MP method is not designed as a tracking algorithm, but a coding algorithm which implicitly has some tracking skills. This poses several problems from the tracking point of view :

- The parameters of the video atoms are coarsely quantized to achieve better compression performances, which introduces tracking errors;
- Atoms are followed from one frame to the other using a search window of limited size, since, as in most video coding schemes, it is less expensive to code a new object than to encode the difference between two very different entities. This limits the robustness and flexibility of the

tracker;

- The algorithm is not formalized as a tracking method, which makes it difficult to understand how the variation of the tracking performances depends on the variation of the parameters.

In addition to that, as already mentioned, the video MP method is computationally complex since an MP decomposition of each frame has to be computed.

Therefore, in this chapter we formalize the atom tracking problem to enable a more intuitive interpretation of the decomposition results and to reduce the computational complexity of the atom tracking scheme. Object tracking is usually performed based on an appropriate description of the appearance of a target, either at a global or local level. Examples of global descriptions are simple templates [118], color histograms [23], or active appearance models [40]. Examples of local analysis are the methods developed to independently track and match feature points. The seminal work in this field is the KLT tracker [69] where stable corners are detected and then their appearance is represented by an affine invariant template computed on a small region around the point. The points detected at subsequent frames are matched based on the appearance. More advanced feature point detectors have been proposed to account for rotation, scale changes of the underlying object structures [68]. All the above mentioned methods are designed from a tracking-centric point of view :

- Stable structures are used to facilitate tracking;
- The representation is designed to reduce ambiguity between feature points [76].

The interpretation of the information obtained after tracking in the context of the considered signal is postponed to a subsequent analysis stage. But are stable structures also relevant from a signal representation point of view? We argue that a signal-centric (as opposed to a tracking-centric) representation can extend the application of a feature tracking system by fusing analysis and tracking in a single general framework.

In this chapter we introduce such a framework and we define an algorithm to follow across time important video structures like oriented edges. The tracker is automatically initialized by representing the first frame of a sequence as a combination of edge-like functions. These functions are retrieved and ranked from a redundant dictionary of atoms using MP. In contrast to classical tracking algorithms, the structures to be tracked are implicitly defined by MP that picks the most relevant image contours. Such visual features are then tracked using one of the most popular tracking algorithm, Particle Filter (PF) [6, 83, 121]. In this way we put the video atom tracking problem in the well grounded and understood framework of PF, which moreover ensures robustness, flexibility and lower computational complexity than the video MP method. The proposed scheme is integrated in the audiovisual fusion algorithm presented in the previous chapter and it is employed for an audiovisual source localization task.

4.2 Tracking Geometric Video Structures

4.2.1 Video Representation

The video approximation framework we consider here is the same introduced in the previous chapter. Thus we represent a video sequence as a sum of 2D geometric primitives obtained in the expansion of a reference frame $I_1(x_1, x_2)$ that are tracked from frame to frame. However, in this chapter the notation will be slightly changed in order to integrate the classical PF notation with those used in

the video MP framework. Thus, a 2D atom will be denoted with $\phi^{(i)}_{\mathbf{x}[n]}(x_1, x_2)$, where the index $\mathbf{x}[n]$ indicates the set of transformations associated to the n-th atom, i.e. $\mathbf{x}[n] = (t_{1_n}, t_{2_n}, s_{1_n}, s_{2_n}, \theta_n)$. We use this notation since in the classical PF formulation, $\mathbf{x}[n]$ is the *state vector* (i.e. the set of values that describe the considered target) associated to the n-th target to track. This target in our case is the n-th atom of the decomposition.

Therefore, the reference frame I_1 is approximated with a linear combination of N functions $\phi^{(i)}_{\mathbf{x}[n]}$ retrieved from a redundant dictionary $\mathcal{D}^{(i)}$ of edge-like atoms using MP as

$$I_1 \approx \sum_{n=0}^{N-1} c_n \phi^{(i)}_{\mathbf{x}[n]}\,, \tag{4.1}$$

where n is the summation index, $c_n = \langle R^n I_1, \phi^{(i)}_{\mathbf{x}[n]} \rangle$, $R^0 I_1 = I_1$ and $R^n I_1$ is the residual after n iterations. The codebook $\mathcal{D}^{(i)}$ is composed of oriented edge-detector functions defined in (3.12) and it is the same used in Chapter 3. Following this procedure the reference frame I_1 is decomposed into N atoms and the first Q of them are tracked through time.

4.2.2 Tracking Video Atoms with Particle Filter

The tracking is performed using Particle Filter (PF), a parametric method that solves nonlinear and non-Gaussian state estimation problems using a Bayesian approach [6, 83, 121]. Its robustness and flexibility makes of PF one of the most used tracking algorithm.

The reference image is represented with N atoms and the first Q atoms are *independently* tracked, i.e. each video structure is tracked without considering the interactions of such structures with the neighboring ones. This is mainly motivated by the fact that we are interested in the main structures present in the video (i.e., the first functions of the MP decomposition). If few atoms are considered, then their interactions are likely to be weak. One can measure such interactions by computing the scalar products between the atoms. If two atoms exhibit a large scalar product (the atoms have unit norm, thus the maximum scalar product is 1), their interaction is strong, while if it is small (i.e. close to 0), their interaction is weak. Figure 4.1 shows the sum of the scalar products between the atoms representing the first frame of a sequence [Left], and the average scalar product between atoms [Right], plotted as a function of the number of considered functions. The total scalar product clearly increases with the number of atoms, since there are more interactions between the structures. The average scalar product increases rapidly until when the atoms added to the decomposition become very small since they represent small image details, giving low scalar products with the other functions. In our experiments we will consider the first $Q = 30$ atoms selected by MP : as a first approximation, it seems reasonable to consider the atoms independently since the interactions between them are still limited. However, as highlighted in [37], neighboring functions can mutually influence each other and one of the main future research directions will be the design of a method that can account for the interactions between atoms.

Each atom $\phi^{(i)}_{\mathbf{x}[n]}(x_1, x_2)$ is fully characterized by the set of five parameters $\mathbf{x}[n]$, i.e. the position, scale and rotation parameters that describe its shape. PF solves the tracking problem considering each target object as a dynamic system that is described by the five-dimensional state vector $\mathbf{x}[n]$. The evolution of the characteristics of a target is then described by the state equation

$$\mathbf{x}_t[n] = \mathbf{f}_t(\mathbf{x}_{t-1}[n], \mathbf{v}_t)\,, \tag{4.2}$$

where $\mathbf{f}_t$ is a possibly non-linear and time-varying function of the state, $\{\mathbf{v}_t\}_{t=1,\ldots}$ is assumed to be an independent and identically distributed (i.i.d.) stochastic process and the subscript t indicates the frame index. The random process $\{\mathbf{v}_t\}_{t=1,\ldots}$ is added to the model to simulate the effect of

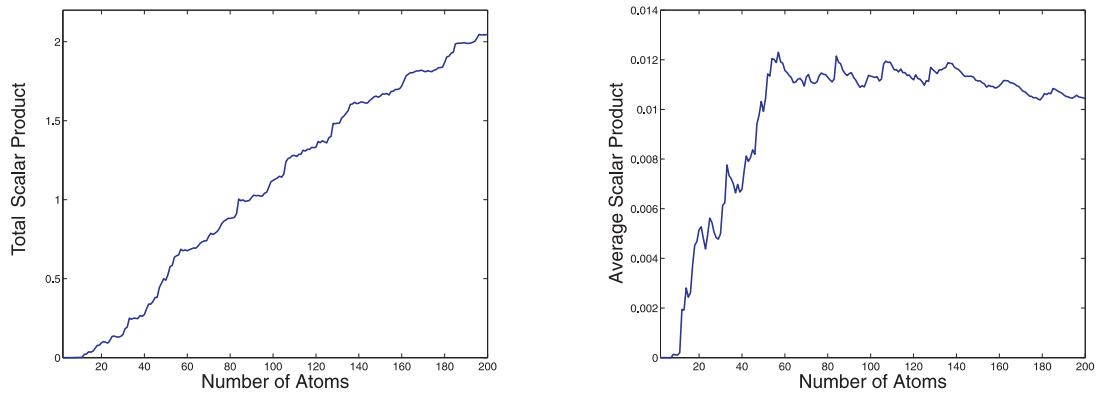


Figure 4.1 – *Sum of scalar products between the atoms representing the first frame of a sequence [Left], and average scalar product [Right], plotted as a function of the number of considered functions.*

noise on the system. In the present case it is considered as a zero mean Gaussian random variable with variance $\sigma_{\mathbf{v}}$. The function $\mathbf{f}_t$ defines the motion model, i.e. the *a priori* information about the evolution of the system that one introduces into the model. For example a certain continuity on the motion or on the velocity of the target can be imposed. In our case we consider a simple *zero-order model*, and thus (4.2) can be rewritten as

$$\mathbf{x}_t[n] = \mathbf{x}_{t-1}[n] + \mathbf{v}_t \,. \tag{4.3}$$

The state variable $\mathbf{x}_t[n]$ describes the characteristics of target number n at time t, and thus it defines the n-th atom at frame t. The goal of the tracking is to estimate the state $\mathbf{x}_t[n]$ based on a series of measurements related to the state vector

$$\mathbf{z}_t[n] = \mathbf{h}_t(\mathbf{x}_t[n], \mathbf{n}_t) \,, \tag{4.4}$$

where $\mathbf{h}_t$ can be a non-linear and time-varying function and $\{\mathbf{n}_t\}_{t=1,\dots}$ is an i.i.d. process that models the measurement noise. The relationship between the measurement and the state vector can be very complex and it is typically difficult to express. Fortunately, there is no need to explicitly define this function; instead, as will be shown soon (equation (4.10)), it is only necessary to define the *likelihood* of a measurement given a state vector. Such likelihood function can be more easily and intuitively designed (see (4.11)). The idea here is to find an estimate of $\mathbf{x}_t[n]$ based on all the available measurements up to time t, $\mathbf{z}_{1:t}[n] = \{\mathbf{z}_j[n]\}_{j=1,\dots,t}$. To simplify the notation, from now on the atom index n will be omitted, since anyway the atoms are tracked independently.

Considering a Bayesian point of view, the tracking problem consists in recursively estimating a certain confidence on the state $\mathbf{x}_t$ at time t given the set of available measurements up to time t, $\mathbf{z}_{1:t}$. Thus, the objective is to estimate the *pdf* $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ at each time instant t. Assuming that the initial *pdf* of the state $p(\mathbf{x}_0|\mathbf{z}_0) := p(\mathbf{x}_0)$ is available, $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ can be obtained recursively in two steps, namely prediction and update. The *prediction step* uses the state equation (4.2) to obtain the prior *pdf* as

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1} \,, \tag{4.5}$$

with $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ known from the previous iteration and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ determined by (4.2). When the measurement $\mathbf{z}_t$ is available, it is possible to perform the *update step* using Bayes' rule

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{\int p(\mathbf{z}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) d\mathbf{x}_t} \,. \tag{4.6}$$

PF approximates the densities $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ with a sum of N_P Dirac functions (*particles*) centered in $\left\{\mathbf{x}_t^i\right\}_{i=1,\ldots,N_P}$ as

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} w_t^i \delta\left(\mathbf{x}_t - \mathbf{x}_t^i\right) , \tag{4.7}$$

where w_t^i are the weights associated to the particles and they are calculated as

$$w_t^i \propto w_{t-1}^i \frac{p(\mathbf{z}_t|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i, \mathbf{z}_t)} . \tag{4.8}$$

The function $q(\cdot)$ is the importance density function which is often chosen to be $p(\mathbf{x}_t|\mathbf{x}_{t-1}^i)$, as it is done here. This leads to

$$w_t^i \propto w_{t-1}^i p(\mathbf{z}_t|\mathbf{x}_t^i) . \tag{4.9}$$

A re-sampling algorithm is then applied to avoid the degeneracy problem, i.e. the fact that after a few iterations all particles except one have negligible weight [6]. The basic idea of re-sampling is to eliminate particles that have small weights and to concentrate on particles with large weights (see also Fig. 4.3). In this case the weights are set to $w_{t-1}^i = 1/N_P \ \forall\, i$, and therefore

$$w_t^i \propto p(\mathbf{z}_t|\mathbf{x}_t^i) . \tag{4.10}$$

The weights are thus proportional to the *likelihood* of the measurement $\mathbf{z}_t$ given the particles. Here the natural choice for the likelihood function is the projection of the candidate atom over the image, since we want to track important video structures, i.e. video atoms exhibiting high projection on the frame. This is also coherent with the representational framework formulated in the previous section. The likelihood of a candidate particle is defined as the absolute value of the scalar product between the residual frame and the atom represented by the particle. In order to favor candidates with high likelihood, this quantity is filtered with a Gaussian kernel centered in the maximum likelihood value and with variance $\sigma_{\mathcal{L}}$, obtaining :

$$\mathcal{L}(\mathbf{x}_t^i[n]) = \exp\left(-\frac{(\mathcal{L}_t^M[n] - |\langle R^n I_t, \phi^{(i)}_{\mathbf{x}_t^i[n]}\rangle|)^2}{2 \cdot (\sigma_{\mathcal{L}} \mathcal{L}_t^M[n])^2}\right) , \tag{4.11}$$

with

$$\mathcal{L}_t^M[n] = \max\left(|\langle R^n I_t, \phi^{(i)}_{\mathbf{x}_t^i[n]}\rangle|\right) , \ i = 1, \ldots, N_P .$$

We want to underline that the atom $\phi^{(i)}_{\mathbf{x}_t^i[n]}$ is not projected over the frame I_t but over the residual at step n of the decomposition, $R^n I_t$ (see (4.1)). It is at this step that the interactions between atoms come into play : even though each atom is tracked independently, the values of the weights used to estimate its *pdf* depend on the projection $|\langle R^n I_t, \phi^{(i)}_{\mathbf{x}_t^i[n]}\rangle|$, i.e. on the $n-1$ atoms that precede the n-th one and contribute to the residual $R^n I_t$. We will use the function $\mathcal{L}$ to compute the weights w_t^i. Figure 4.2 shows the likelihood function of a candidate atom computed on a region extracted from one of the analyzed clips. The re-sampling step derives the particles depending on the weights of the previous step, then all the new particles receive a starting weight equal to $1/N_P$ which will be updated by the next frame likelihood function.

The best state at time t, $\hat{\mathbf{x}}_t$, is the particle $\mathbf{x}_t^i$ with biggest weight, weighted by a factor that takes into account the similarity of the particle with the corresponding best state at time $t-1$:

$$\hat{\mathbf{x}}_t = \mathbf{x}_t^M \ \text{ s.t. } \ w_t^M = \max(S(\mathbf{x}_t^i, \hat{\mathbf{x}}_{t-1}) \cdot w_t^i) . \tag{4.12}$$

The function S is a Gaussian in the 5D parameter space. The value of $S(\mathbf{x}[l], \mathbf{x}[m])$ is maximum when the particles $\mathbf{x}[l]$ and $\mathbf{x}[m]$ coincide and it decreases exponentially as the distance between $\mathbf{x}[l]$ and $\mathbf{x}[m]$ in the parameters space increases.

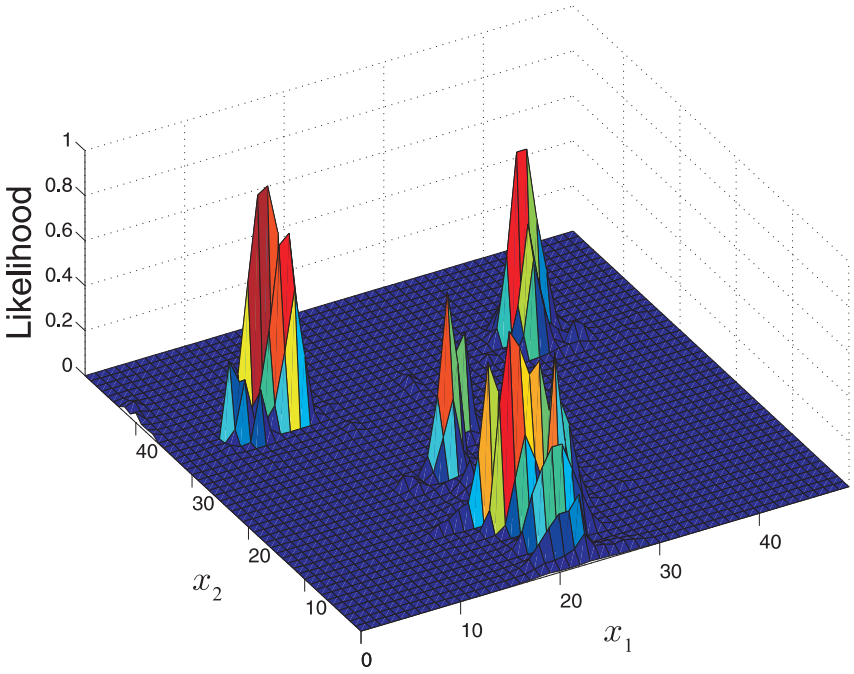


Figure 4.2 – *Likelihood function of a candidate atom computed on a region extracted from one of the analyzed clips. The function is clearly multi-modal, exhibiting peaks that have similar amplitude and that are spatially close.*

Alternative strategies to compute the best state would be to take the particle with highest weight or to consider the Monte Carlo approximation of equation (4.7) consisting in estimating the best state as the weighted sum of the particles, as in [6]. However, it was observed that unstable, noisy atom trajectories were generated considering simply the particles with largest weights, due to the multi-modality of the posterior *pdf*s, as can be seen in Fig. 4.2. The Monte Carlo solution would produce more stable atom trajectories. However, in this case there is no guarantee that the best state corresponds to an atom that matches a *real* visual structure, since several local maxima can be present in the likelihood function (Fig. 4.2). This causes errors due to the fact that when the n-th atom is found, it is subtracted, multiplied by its coefficient, from the residual image $R^{n-1}I_t$ to generate the new residual R^nI_t which is used to calculate the successive atoms (see (4.1)). If the n-th atom is not matching an image structure, its coefficient (i.e. its projection over the residual image) will be very small and thus its contribution to the MP decomposition will not be taken into account, inducing errors in the computation of the successive atoms.

The use of the weighting factor $S(\mathbf{x}[l], \mathbf{x}[m])$ results in a stabilization of the atoms tracks since the algorithm tends to prefer states that are as similar as possible to the previous ones, except if relevant modifications of the structures occur. At the same time, the representation of the scene is kept coherent. An example of PF with re-sampling is shown in Fig. 4.3.

4.3 Experiments

In this section we present the results of the atoms tracking algorithm with PF. We will term this method MP-PF and the video MP algorithm of Divorra 3D-MP. We test the MP-PF tracker on

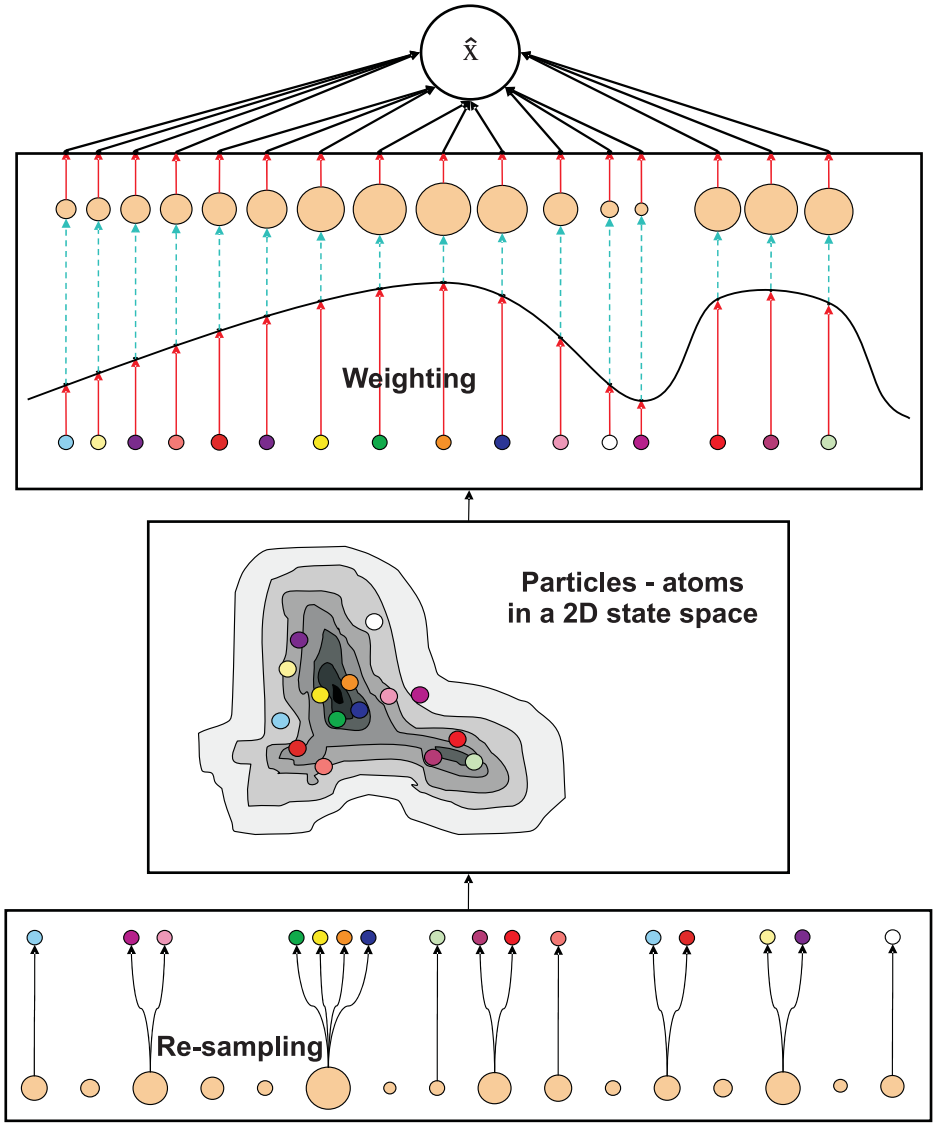


Figure 4.3 – *Schematic representation of the Particle Filter algorithm. At frame t, N_P particles (represented by the dots) with the corresponding weights are available from frame $t-1$ [Bottom]. In this sketch the radius of the dot is proportional to the likelihood of the particle. The re-sampling step then eliminates particles with a small weight and it introduces some new ones from those having a large weight. Based on the measurement available at frame t, the likelihood and thus the weight of each particle is computed and the best state at time t, $\hat{\mathbf{x}}_t$, is estimated [Top].*

sequences representing one or two persons speaking and moving in front of a camera. The clips used for the tests come from the *individuals* and *groups* sections of the CUAVE database [88] (only the luminance component of the clips has been considered). The video data was recorded at 29.97 fps and at a resolution of 480×720 pixels. The size of the clips has been then reduced to 120×176 pixels. We use a 5-dimensional state model for PF composed of the target position, (t_1, t_2), the target scale s_1 and s_2 and the orientation θ. In all experiments a zero-order motion model with fixed $\sigma_{\mathbf{v}}$ is used. Since we are considering a 5D state vector, $\sigma_{\mathbf{v}}$ as well has five components that are

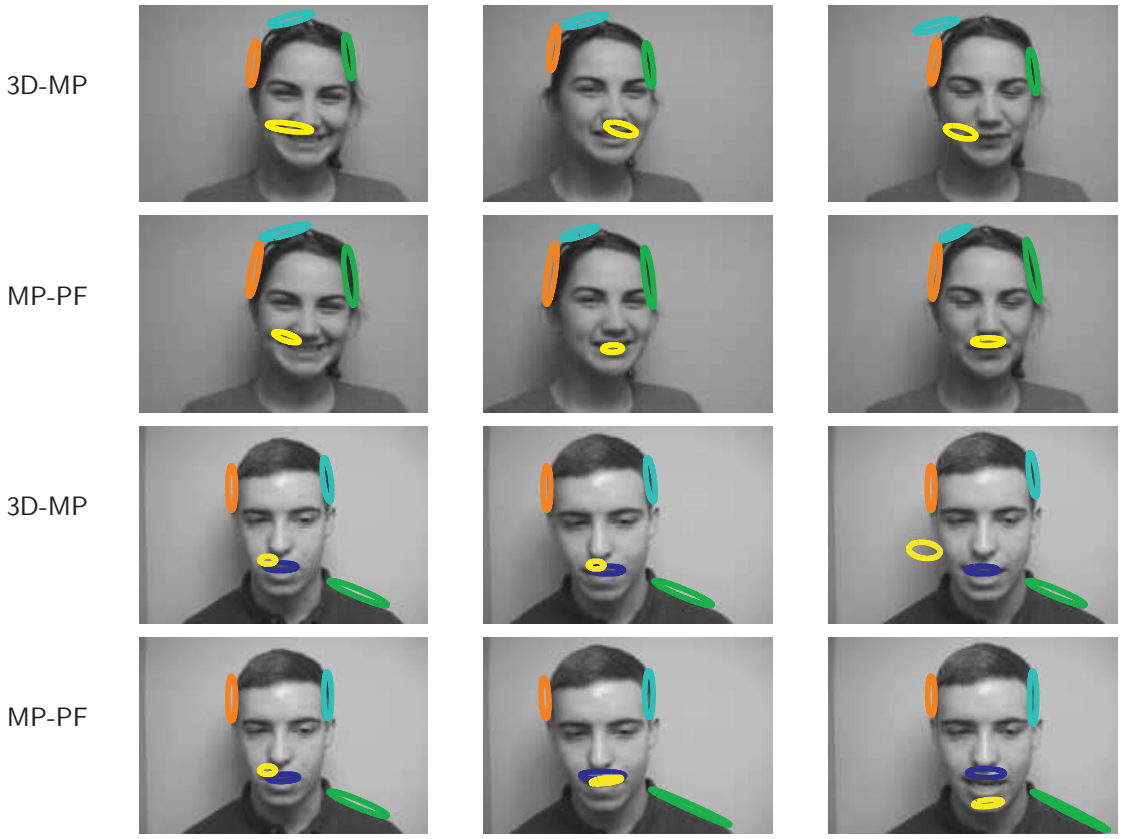


Figure 4.4 – *Video atoms tracking. The footprints of different atoms are depicted with different colors. Results for the 3D-MP approach are on the first and third rows and those for the MP-PF method are on the second and fourth rows. From the second to the third frame the subjects rapidly move towards their left : the 3D-MP tracker looses the track of some edges, while the MP-PF tracker does not.*

$\sigma_{t_1} = \sigma_{t_2} = 2$, $\sigma_{s_1} = \sigma_{s_2} = 0.03$ and $\sigma_\theta = 3.5$. Note that the position change is in pixels while the scale is in percentage and the orientation in degrees. The Gaussian kernel filtering the likelihood function has $\sigma_{\mathcal{L}} = 0.05$. The PF tracker uses 150 samples (particles).

4.3.1 Video Atoms Tracking

In the first experiment, the proposed MP-PF approach is compared to the 3D-MP algorithm [35]. The two methods are tested on four sequences taken from the *individuals* partition of CUAVE that represent one person speaking and moving in front of the camera. Sample frames of two test sequences are shown in Fig. 4.4.

Both trackers are initialized with the same video atoms using MP. The edges are then tracked using a video MP approach in 3D-MP, while the MP-PF method tracks the video structures using PF, as detailed in section 4.2. In Fig. 4.4 the tracking results using the two algorithms are compared. Footprints of different atoms are depicted with different colors. The first and third rows show the results obtained with the 3D-MP approach and the second and fourth rows show the results for the proposed MP-PF method. In the second part of the sequence (second and third frames) the subjects rapidly move towards their left. The 3D-MP tracker loses the track of two edges in the first case and of one in the second, while the MP-PF tracker does not. The same behavior has been observed in the other test sequences. While the 3D-MP algorithm easily looses the track of fast moving edges,

Figure 4.5 – *Frames from clips g19 [Top] and g21 [Bottom]. The footprints of the most correlated atoms are highlighted. The mouths of the correct speakers are detected.*

the MP-PF approach is more robust, even if errors can be observed. In both sequences for example it happens that the yellow atom associated with the upper lip is temporarily associated with the lower lip or the chin. This problem seems to be caused by the interactions between nearby atoms and we believe that it could be eased by jointly tracking groups of video structures.

4.3.2 Audiovisual Source Localization

In the second experiment, the proposed MP-PF tracker is integrated in the audiovisual fusion algorithm presented in Chapter 3 to perform a cross-modal source localization task. The audio-video features that are considered here are the same used in the previous chapter. The video signal is represented using $Q = 30$ video atoms and each atom has a feature associated describing its displacement. The video atoms exhibiting the highest degree of correlation with the audio are detected using a simple relevance criterion and the sound source location over the image sequence is estimated. A sliding window of 70 frames length is used to compute the synchronization vectors and to detect the video atoms that are more correlated with the audio. The observation window is then shifted by 20 samples and the procedure iterated.

We have tested the algorithm on the last four sequences of the *group* partition of the CUAVE database (g19, g20, g21, g22). The video clips involve two persons taking turn in reading series of digits in English and arranged as in Fig. 4.5. This figure shows the results of the described approach detecting the mouth of the speaker in two sequences where two persons speak in turns in front of the camera. In white are highlighted the footprints of the video atoms found to be correlated with the soundtrack. The mouth of the correct speaker is detected.

The proposed method has been quantitatively evaluated using the same protocol presented in Chapter 3 and the performances have been compared to algorithm introduced in the previous chapter and to the one presented in [82]. We recall here that the method proposed by Nock and colleagues [82] detects the mouth of the speaker founding the image zone over which the mutual information between audio and video features is maximized. They use test clips that could not exactly coincide with those used here, since the original sequences have been cropped in both cases. In contrast, the results presented in the previous chapter are obtained using exactly the same test sequences. The main differences between the algorithm proposed here and the one in Chapter 3 basically consist in the video edge tracking approach (here we use MP-PF, while the algorithm presented in the

Clip	Nock[82]	3D-MP	MP-PF
g19	41	87	94
g20	93	93	93
g21	79	81	78
g22	79	87	80

Table 4.1 – *Audiovisual localization results expressed in percentage of correct detections.*

previous chapter uses the 3D-MP approach) and in the different number of atoms considered. We use 30 atoms and not 50 as before because, as underlined in the previous section, we track the atoms independently : the higher the number of atoms, the stronger are their interactions, as exemplified by Fig. 4.1. The 3D-MP approach takes into account interactions between atoms and thus this aspect is not an issue.

Table 4.1 summarizes the results obtained for the three methods in term of percentage of test points at which the speaker's mouth is correctly detected. The results shown are for the largest regions of correct mouth detection defined in the previous chapter, i.e. a 200×200 square for Nock's method and a circle of diameter 50 pixels for both the 3D-MP and the MP-PF algorithms. Note that there could be no perfect coincidence between the test sequences used in [82] and those used here, thus the results for Nock's algorithm should be considered only as indicative. As already shown in Chapter 3 and in [78, 79], the use of geometric video decompositions combined with an audio-video event detector in general improves the results obtained by Nock and colleagues. The proposed method obtains detection performances similar to those of the algorithm using 3D-MP, slightly improving previous results for sequence g19 but obtaining inferior performances on clip g22.

As shown by the results in Fig. 4.4 the MP-PF method improves the tracking abilities of the 3D-MP algorithm. This is indeed interesting considering that the 3D-MP tracker, even without jointly tracking groups of structures, takes into account atoms' interactions, which was demonstrated to increase the accuracy of the 3D-MP approach [37]. We argue that an MP-PF algorithm that considers atoms' dependencies would correct tracking errors due to atoms' interactions (Figure 4.4) and would allow to improve the audiovisual localization results, that by now are essentially equivalent to those obtained using 3D-MP (Table 4.1). Concerning the computational complexity, we have tested the two methods on a video sequence whose 30 principal video atoms were tracked through time. The MP-PF algorithm clearly outperforms the 3D-MP approach, resulting approximately 7 times faster.

4.4 Discussion

In this chapter we have presented a new framework and an algorithm to represent and track relevant video structures. The proposed method improves the 3D-MP video representation algorithm presented in [35], which is designed as a coding algorithm and poses problems from the tracking point of view. These limitations are overcame by defining the video atom tracking problem in the well grounded and understood framework of Particle Filter, which ensures robustness, flexibility and lower computational complexity than the 3D-MP algorithm.

Experiments show that the proposed tracker is more robust and accurate than the 3D-MP one, while being considerably less time consuming. The audiovisual source localization algorithm, however, does not improve accordingly. This is mainly due to the fact that while the 3D-MP

algorithm takes into account atoms' interactions, the current MP-PF method does not. However these results show that there is room for further improvements by designing a mechanism that accounts for the interactions between video atoms. The tracking framework developed in this chapter seems to be appropriate to continue the evolution of the system, considering for example a multi-object tracking approach as in [65]. The MP algorithm in fact ranks dictionary functions that are analytically defined. These two characteristics (the atoms ordering and their analytical formulation) make the interactions between video structures easy to define, making the inference of higher level signal configurations intuitive.

At this point we have to turn the attention to the audio representation technique as well. In the next chapter we will focus on this issue and we will build a bridge between two signal processing fields that were basically separated : audiovisual fusion and one-microphone blind source separation.

Blind Audiovisual Source Separation

5

In Chapters 3 and 4 we have explored the capabilities of redundant parametric decompositions to represent audiovisual sequences. These techniques allow to interpret signals in terms of their salient structures, preserving good representational properties thanks to the use of redundant, well designed, dictionaries. Considering the signal of one single microphone and the video associated, we have shown that we can accurately spatially localize the active sound source in an audiovisual sequence. This is done correlating high-level video features (movements of relevant visual edges) with a simple audio feature (the average acoustic energy).

On the other hand, the time-frequency representation of the audio signal contains a great amount of information (see section 3.5) that is discarded considering a basic audio representation based on the mean energy. In this chapter we will show that, considering a more detailed representation of the audio modality, it is possible to jointly localize and extract video *and audio* sources. The key idea here is to exploit the time-frequency information provided by the MP audio representation and to perform a joint Audiovisual Source Separation correlating audio-video structures and clustering them into sources.

5.1 From Audio to Audiovisual Source Separation

This chapter introduces a new concept, Audiovisual Source Separation, which is achieved by exploiting the information contained in the *mono* audio signal and in the video sequence to separate and extract correlated sources in each one of these modalities.

Few methods exist that exploit audiovisual coherence to separate *stereo* audio mixtures [28, 94, 97, 104, 114]. All the existing algorithms consider the problem from an *audio source separation point of view*, i.e. they use the audio-video synchrony as side information to improve and overcome limitations of classical Blind Audio Source Separation (BASS) techniques. In the next section we will briefly introduce the BASS problem and its terminology, that will be used as well in the following of this chapter. After that we will review the main contributions in the two fields over which the presented approach is grounded, *Stereo* Audiovisual Source Separation and Single-Channel Blind Source Separation.

5.1.1 Blind Audio Source Separation

The Blind Audio Source Separation problem consists in recovering the set of signals $\{s_j(t)\}_{j=1,\dots,N_S}$, also called *sources*, from mixtures of them $\{y_i(t)\}_{i=1,\dots,P}$, typically signals recorded by a sensor array. Considering for simplicity the case of time-invariant linear mixing systems, the signals $\{y_i(t)\}_{i=1,\dots,P}$ can be expressed as

$$y_i(t) = \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{+\infty} m_{ij}(\tau) s_j(t-\tau)\,, \tag{5.1}$$

where $\{m_{ij}(\tau)\}_{i=1,\dots,P\,,j=1,\dots,N_S}$ is a set of mixing filters. This $N_S \times P$ system can be represented in a maybe more familiar matrix notation as

$$\mathrm{y}(t) = \mathrm{M}(t) * \mathrm{s}(t)\,, \tag{5.2}$$

where $*$ indicates the convolution, $\mathrm{y}(t) = [y_1(t), \dots, y_P(t)]^T$, $\mathrm{s}(t) = [s_1(t), \dots, s_{N_S}(t)]^T$ and the element in position (i, j) of the mixing filter matrix $\mathrm{M}(t)$ is $m_{ij}(t)$ (the operator $\cdot^T$ is the transposition). In a completely blind setting the sources and the mixing filters are unknown. In this case the BASS problem admits an infinite number of solutions [111], and thus assumptions must be made about the sources or/and the mixing process to obtain a unique solution.

Concerning the mixing process, the BASS literature typically classifies mixtures depending on the number of sources and sensors and on the characteristics of the mixing filters $m_{ij}(\tau)$ involved in the process [110]. Thus a mixture can be termed as *over-determined*, *determined* or *under-determined* if the number of sensors P is greater, equal or smaller than the number of sources N_S respectively. Concerning the mixing filters, a mixture is called *instantaneous* if the filters are simply scalar gains, i.e. $m_{ij}(\tau) = m_{ij}$ and (5.1) becomes

$$y_i(t) = \sum_{j=1}^{N_S} m_{ij} s_j(t).$$

A mixture is termed *anechoic* if the mixing filters are scalar and sources are delayed by a fixed shift, i.e. (5.1) can be written as

$$y_i(t) = \sum_{j=1}^{N_S} m_{ij} s_j(t-\delta_{ij})\,,$$

with δ_{ij} the time delay associated to the path between the j-th source and the i-th sensor. Finally in the most general case a mixing system is on the form of (5.1) and it is termed *convolutive*.

Even considering the simplest scenario, the case of over-determined instantaneous mixtures, the BASS problem does not have a unique solution [111]. To overcome this limitation, strong assumptions on the characteristics of the sources have to be made. One typical assumption is the statistical independence between the sources, that leads to a long series of separation methods based on Independent Component Analysis (ICA) [9, 20, 56]. ICA-based methods have been shown to be effective in separating over-determined and determined instantaneous and convolutive mixtures. However the independence assumption is insufficient in the under-determined scenario, and additional information has to be exploited to separate under-determined mixtures. One characteristic that is often exploited is the sparsity of the audio signal in the spectral domain [5, 119, 122]. The sparsity assumption basically states that only one source is present at any time-frequency point. Using this premise and the spatial information available from a stereo signal, instantaneous under-determined mixtures can be effectively separated.

Clearly several other interesting research directions and algorithms have been proposed in the prolific field of source separation. However an exhaustive review of BASS theory and methods

is out of the scope of this thesis (the interested reader can refer to numerous interesting survey articles like [57, 110, 111]). The goal of this section instead is to introduce the BASS formalization and nomenclature and to underline the complexity of the separation task in realistic conditions. Researchers have tried to solve source separation problem by formulating assumptions of different nature and exploiting additional available information. Recently, algorithms have been proposed that face the BASS problem using not only the information contained in the audio signal, but also the associated video information. We will review the most representative of these techniques in the next section.

5.1.2 Audiovisual Source Separation

It is well known from every-day experience that visual information strongly contributes to the interpretation of acoustic stimuli. This is particularly evident if we think to speech signals : speaker's lips movements are correlated with the produced sound and the listener can exploit this correspondence to better understand speech, especially in adverse environments [106, 107]. The multi-modal nature of speech is exploited since at least two decades to design speech enhancement [31, 45, 46] and speech recognition algorithms [70, 92] in noisy environments. Lately, this paradigm has been adopted also in the speech separation field to increase the performances of audio-only methods.

In [104] the authors propose to estimate the de-mixing process using a criterion based on audiovisual coherence : one speech source of interest is extracted using the visual information simultaneously recorded from the speakers face by video processing. The coherence between audio and video data is modelled by a joint audiovisual probability estimated as a mixture of Gaussian kernels whose parameters are learned from a large training set. Video information consists of geometric parameters describing the speaker's lips height and width that are extracted using a chroma-key process on lips under controlled head position and light conditions [64]. The system shows to be able to estimate the un-mixing matrix in the case of instantaneous mixing systems. A very similar approach, but for convolutive mixtures, has been developed in [97]. Another method inspired by [97, 104] is presented in [114]. In this case video features are deduced using active appearance model [40] and the algorithm is tested on a limited set of 2×2 (i.e. determined) instantaneous mixtures.

Dansereau [28] also proposes an audiovisual speech source separation system plugging the visual information, representing again the speaker's lip height and width, in a de-correlation system with first-order filters. Visual cues are mapped to word structures with a continuous HMM that is trained on a corpus of visual speech. The method was tested simulating a 2×2 speech separation problem by mixing one audio source recorded with one microphone and one speaker captured with one camera and one microphone. Rajaram and colleagues [94] suggest instead a Bayesian framework for 2×2 instantaneous mixtures of audio-video sources. In this case the video feature employed is quite simple and it basically provides a binary weight that indicates the activation of a source, and the mixing model parameters are estimated on-line.

The approach we consider in this chapter is very different from existing ones. First, we localize the "centroid" of visual sources using audiovisual synchrony in a manner similar to what we have done in Chapters 3 and 4 but employing a robust clustering algorithm. Once we have located the sources centers on the image sequence, we reconstruct the video sources by simply assuming that the structures close to a source belong to it. We obtain thus several groups of video structures, each group corresponding to a detected source. It is important to underline that sources in the video domain, e.g. people speaking in front of a camera, are typically well separated in space. This information will help us in separating the audio mixture as well, exploiting the correlations established between audio and video entities. Since only a one-microphone signal is considered, the separation of an unknown number of unknown sources is in fact extremely challenging.

We want to stress three important differences between our proposed approach and state-of-the-art audiovisual separation methods :

- In all the above mentioned methods the BASS problem is solved for stereo audio signals using more or less classic separation techniques helped by visual information. In contrast, the audio signal we consider here comes from only *one microphone*, which makes the source separation task considerably more challenging, as we will comment in the next paragraph;
- Existing methods simplify the task of associating audio and video information. Either the audio-video association is given *a priori*, i.e. it is known which audio signal corresponds to which video signal [94, 114], or one audiovisual source is mixed with an *audio-only* source [28, 97, 104]. In this second case the separation problem basically turns into the following : separate two mixed speech signals, one of which has a corresponding video counterpart. Here in contrast we simultaneously separate audio-video sources, automatically building correlations between acoustic and visual entities. Instead, the hypothesis that we make is that each video source detected in the scene has one and only one corresponding audio component in the audio mixture;
- Existing audiovisual separation methods, with the only exception of [94], require an off-line training step to build the audiovisual source model. This is mainly due to the fact that the algorithms proposed in [28, 97, 104, 114] try to map video information into the audio feature space using techniques similar to lip-reading (requiring moreover accurate mouth parameters that are difficult to acquire). In contrast, in the proposed method no training will be required to associate simple audio-video features.

To summarize we essentially want to solve a blind Single-Channel BASS problem, but aided by the video. Since no hypothesis is made on the relationships between audio and video structures, video sources have to be localized and separated at the same time, exploiting the information contained in the audio channel. The steps of our *Blind Audiovisual Source Separation* (BAVSS) algorithm will be detailed in the following of this chapter, while in the next section we describe the most representative approaches to Single-Channel BASS, pointing out their salient characteristics and limitations.

5.1.3 Single-Channel Blind Source Separation : A Difficult Problem

As underlined in section 5.1.1, solutions to the BASS problem typically require microphone arrays or stereo microphones [5, 20, 111, 119, 122]. However here we have at our disposal only the signal from one microphone. On the other hand, we can exploit the correlation with the video signal associated to separate the audio sources.

Single-Channel Source Separation is a relatively recent, hard and still open problem, faced for the first time by Roweis in [99]. When only the input signal of one microphone is available, simple generic assumptions do not suffice. For the Single-Channel Source Separation it is necessary to model different characteristics of the speech signal, such as the spectral envelope, the fundamental frequency or the temporal continuity. These known cues for speech separation [14, 16] have to be taken into account in order to build models that face this problem.

The existing research works relative to Single-Channel Source Separation can be divided into two main groups according to their blindness :

Generative - Approaches in this group build their models according to the speakers present in the mixture, i.e. for each mixed speaker the algorithm is trained on sequences where only he or

she is speaking. Thus, these works are situated in a *non blind* context. Early approaches in this field belong to this group [59, 95, 99]. The problem is that if the model is too simple it is not able to discriminate different sources, while, on the other hand, if the algorithm is too complex an inference problem of huge dimensionality has to be addressed.

Discriminative - These approaches focus on the spectral separation task instead of building complex models for each speaker. They try to exploit the sparsity of speech signals in the time-frequency domain, and do not assume any prior knowledge about the speakers present in the mixture. Examples of algorithms in this *blind* group are [7, 96].

Roweis [99] first challenged the Single-Channel Source Separation problem using a factorial Hidden Markov Model (FHMM) trained on sequences where the speakers present in the mixture are recorded alone. Through HMM, *binary mask functions* are computed for each frequency sub-band and applied to the mixture in order to extract the original signal of each speaker.

Jang and Lee propose in [59] a technique that utilizes the time-domain ICA basis functions previously learned from a training database consisting in sequences where the speakers in the mixture speak alone. This method recovers original signals through gradient-ascendent adaptation steps to find the maximum likelihood estimate of the sources.

In [95], the authors reduce the dimensions of the problem raised in [99] by dividing the spectral representation of the source signals into *multiple sub-bands*, i.e. multiple parallel horizontal sections of the spectrogram. Then, this approach computes a separate HMM to model each band, requiring few states per model and, for comparable computation expense, can achieve more accurate signal separation than *full-band* models. This model also presents an interesting basis for learning source models directly from mixed signals, since there are more opportunities to find a time-frequency slot with the energy of only one speaker. This characteristic is exploited by the same authors in [96]. This approach captures local deformations of the time-frequency energy distribution and describes each time-frequency region with a linear transformation applied to its predecessor. The spectrum is analyzed as the addition of harmonics and formant structures and no prior models about the present speakers are necessary.

A different approach is proposed by Bach and Jordan [7]. This algorithm builds *affinity* matrices combining classical cues from speech psychophysics [14, 16]. These matrices are employed to define a spectral segmenter that, applied to the mixture, performs the speech separation on the one-channel signal without prior knowledge about the speakers. The algorithm achieves interesting separation results, but it is computationally extremely complex.

In the next section, we introduce a new algorithm to challenge the Single-Channel Source Separation problem in a completely *blind* setting. The proposed approach does not require an off-line training procedure neither the inference of complex models of acoustic sources, but it exploits the video information associated with the audio signal.

5.2 Blind Audiovisual Source Separation (BAVSS)

The proposed method can be divided in four different parts. In order to separate audiovisual sources, first we localize the video sources in the image using the information present in the soundtrack, then we reconstruct them separately. After that, the relationships established between features in both modalities are used to define periods during which only one audiovisual source is active. Finally, exploiting such information, the audio source separation on the time-frequency plane is performed.

There are two main assumptions that we make on the type of sequences that we can analyze using the proposed algorithm. First, we assume that for each detected video source there is one and

Figure 5.1 – *Example of a sequence analyzed with the BAVSS algorithm. The sample frame [Left] shows the two speakers; as highlighted on the spectrogram of the audio [Right], in the first part of the clip the girl on the left speaks alone, then the boy on the right starts to speak as well, and finally the girl stops speaking and the boy speaks alone.*

only one associated source in the audio mixture. This means that if there is an audio "distractor" in the sequence (e.g. a person speaking out of the camera's field of view), it is considered as noise and its contribution to the mixture is associated to the sources found in the video. This assumption clearly simplifies the analysis, since we know in advance that a one-to-one relationship between audio and video entities exists. Moreover, we consider the video sources approximately static, i.e. their positions over the image plane do not change too much. This assumption is less stringent in our opinion and it is formulated only not to have to worry about dynamic aspects of the scene. However it can be removed for example by analyzing the sequences using shifting time windows, as it is done in the previous chapters for the localization algorithms.

One typical sequence that we consider in this work, taken from the *groups* section of the CUAVE database [88], is shown in Fig. 5.1. It involves two speakers arranged as in Fig. 5.1 [Left] that utter digits in English. As highlighted in Fig. 5.1 [Right], in the first part of the clip the girl on the left speaks alone, then the boy on the right starts to speak as well, and finally the girl stops speaking and the boy speaks alone. All the considered test sequences have similar characteristics, with two speakers well separated in space and without significant scaling differences. Thus, the different parameters of the proposed algorithm, that basically depend on the analyzed scene, are empirically set according to the considered scenario. This can be seen as a relaxation of the blindness of the method. However, features related to the geometry of the scene, like the size of the candidate speakers or the distance between them can be easily deduced analyzing the sequence with a face detector/tracker, allowing thus the automatic setting of the algorithm's parameters. For simplicity we skip here this analysis step and we focus our attention on the modelling and separation of audiovisual sources.

The interested reader will find the details of the proposed audiovisual source separation algorithm in the following sections. Those instead who want to get the principal ideas over which the presented separation method is based can refer to Fig. 5.2 and read the description of the main steps of the algorithm here in the following :

1. ***Spatial* Video Source Localization -** In this first step, we use the information contained in the audio signal to localize the video sources in the image using a method similar to that presented in Chapter 3, but now correlating audio and video atoms. This step is schematized in Fig. 5.2(a) : audio entities (the green dot on the right spectrogram) are correlated with video atoms (green and yellow footprints of video atoms are highlighted on the left image). Some correlations are correctly built (green footprints), but some errors can occur as well at this

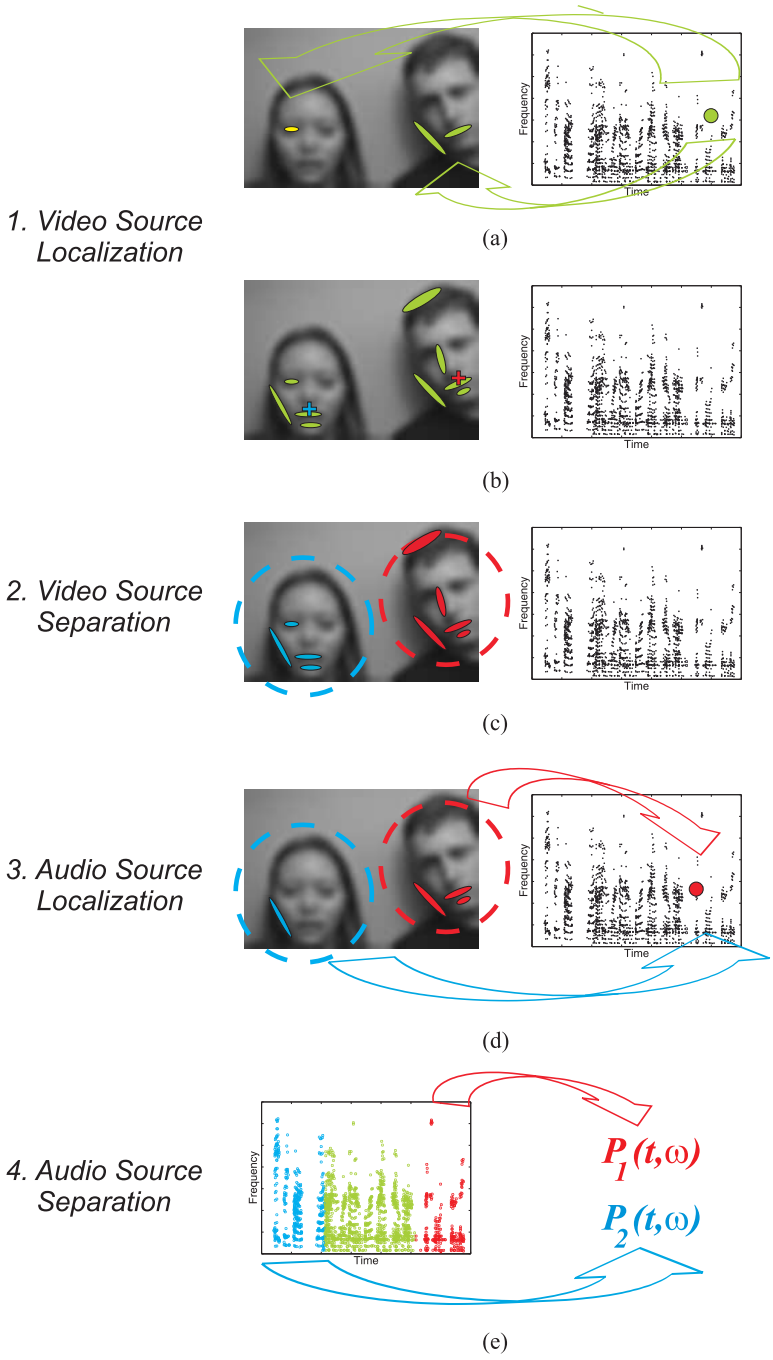


Figure 5.2 – *Schema of the audiovisual source separation algorithm.* Phase 1 *: in (a) audio entities (green dot on the spectrogram) are correlated with video atoms (green and yellow footprints are highlighted on the left image) and exploiting this information, in (b) video sources are localized (blue and red crosses).* Phase 2 *: video atoms are classified into the corresponding video sources (c), as highlighted by their footprints colors (blue for the left speaker and red for the right one).* Phase 3 *: audio atoms (red dot on the right) are classified into the corresponding audio sources using the audiovisual association information (d). Periods with only one audiovisual active source are detected.* Phase 4 *: in temporal periods when a single source is active (blue and red markers) the probability for each frequency to belong to one source is estimated (e). These probabilities are used to separate the sources in mixed periods (green markers).*

stage (yellow footprint). Exploiting this correlation information, video sources are localized using a robust clustering algorithm, as shown in Fig. 5.2(b) (blue and red crosses). This step is very important since it provides a measure of synchrony between audiovisual features, the *correlation scores*, that will be used in the following of the separation process.

2. **Video Source Separation -** The objective of the second part is to classify the video atoms into the detected video sources. This assignation is carried out using a simple spatial proximity criterium : video atoms are associated with the closest detected source, as highlighted by the colors of their footprints in Fig. 5.2(c) (blue for the left speaker and red for the right one).

3. ***Temporal* Audio Source Localization -** At the end of step 2, we have the lists of video atoms classified into the sources and their respective correlations with the audio features. Audio atoms are classified into one of the sources using these relationships. In the toy example shown in Fig. 5.2(d), the audio atom on the spectrogram is correlated with four video atoms, one in the blue cluster and three in the red one and thus it is assigned to this second source. Following this procedure, periods with only one audiovisual source active are clearly detected.

4. **Audio Source Separation -** The last and more ambitious objective (the One Microphone BASS) is pursued using the information present in the temporal periods when a single source is active. The idea is to determine, in these time slots (blue and red markers on the spectrogram on Fig. 5.2(e)), a probability for each frequency to belong to one source. Then, based on this information, we try to predict their behavior in those periods during which more than one source contributes to the mixture (green markers).

5.2.1 Phase 1 : Spatial Localization of Video Sources

The correct localization of the video sources in the spatial domain is the first part of the BAVSS process and it provides the relationship between audio-video atoms, a necessary step to separate the audio sources as well. As already mentioned, we use here a more sophisticated and rich audio representation with respect to the previous chapters. Instead of considering the average acoustic energy, here we will process each atom of the audio MP decomposition separately, attributing an audio feature to each one of them. Therefore a new method to detect meaningful events in audiovisual signals is required, since the dimensionality of the problem increases, as well as the available amount of information.

Audiovisual Association

As a first step, correlations between audio and video have to be established. First audiovisual features are extracted and then a simple audiovisual correlation criterion is designed.

Audio Representation - The audio signal is decomposed using MP over a dictionary of Gabor atoms $\mathcal{D}^{(a)}$, as described in section 3.5.2. Thus, according to (3.19) an audio signal $a(t)$ is approximated using K atoms as

$$a(t) \approx \sum_{k=0}^{K-1} c_k \, \phi_k^{(a)}(t) \, ,$$

where k is the summation index and c_k corresponds to the coefficient for every atom $\phi_k^{(a)}(t)$ from dictionary $\mathcal{D}^{(a)}$. In all the experiments performed in this chapter the audio signals are approximated using $K = 2000$ Gabor atoms selected by MP.

Video Representation - The video signal is represented using the same procedure presented in the previous chapters. Sequences are decomposed into time-evolving visual edges. We have presented two methods to do that, the 3D-MP algorithm of Divorra and the MP-PF tracker introduced in Chapter 4. When employed in an audiovisual fusion task, the two methods exhibit similar performances : here we use the 3D-MP algorithm. Thus, according to (3.16) the video signal is decomposed into N video atoms $\phi_n^{(v)}$ as

$$\mathbf{V}(x_1, x_2, t) \approx \sum_{n=0}^{N-1} c_{n(t)} \phi_n^{(v)}(x_1, x_2, t)\,,$$

where n is the summation index and $c_{n(t)}$ are the coefficients corresponding to each video atom. In all experiments, sequences are represented using $N = 100$ video atoms.

Audio-Video Atoms Association - The decomposition of the audio signal into atoms provides a clear representation of its energy distribution in the time-frequency plane. Thus, the temporal position of an audio atom indicates the presence of a sound in this time period. In a similar manner, the displacement of video atoms reflects the movement of relevant image structures and a peak in the displacement suggests the presence of an event. A temporal analysis is performed that takes into account the temporal co-occurrence of relevant events in both modalities : the temporal location of acoustic energy and the position of video displacement peaks.

For each of the K audio atoms we build a feature that indicates the temporal concentration of acoustic energy. As already mentioned in section 3.5.2, the time-frequency energy distribution of an atom $\phi_k^{(a)}$ can be derived from its Wigner-Ville distribution $W\phi_k^{(a)}(t, \omega)$ [73], that in the case of Gabor atoms is a 2D Gaussian function whose position and variance depend on the atoms parameters. The audio feature $\mathsf{f}_k(t)$ that we consider in this case is the projection over the temporal axis of the Wigner-Ville distribution of every audio atom,

$$\mathsf{f}_k(t) = \int_{-\infty}^{+\infty} W\phi_k^{(a)}(t, \omega) d\omega.$$

For each video atom instead we build an *activation vector* $\mathsf{y}_n(t)$ as described in section 3.5.3. The peaks in each of the N video features are detected, obtaining vectors that equal 1 where peaks occur and 0 otherwise. Then, such vectors are filtered with a rectangular window of size $W = 13$ which models delays and uncertainty. Here the window length is bigger ($W = 13$ frames instead of $W = 7$ as in the previous chapters) because we want to assign all the audio atoms to at least one video atom. If an important audio atom is not correlated to any video atom, it would be lost and it could not be used for the successive reconstruction of the sources. Thus we prefer to be conservative at this stage, since anyway eventual errors can be recovered at successive processing steps. The shapes of the final audio and video features are sketched in Fig. 5.3.

At this point the *correlation scores* $\chi_{k,n}$ between every audio atom $\phi_k^{(a)}$ and every video atom $\phi_n^{(v)}$ can be computed as the scalar product between each audio and video feature :

$$\chi_{k,n} = \langle \mathsf{f}_k(t), \mathsf{y}_n(t) \rangle, \ \forall\, k, n\,. \tag{5.3}$$

This value is high if the audio atom and the peak of displacement of the video atom have a big temporal overlap. In other words, a high correlation score means high probability for the video structures of having generated the sound.

At the end of this step we have built a list of correlations between acoustic and visual structures. The strength of such correlations is indicated by the magnitude of the correlation score. This information is extremely precious and it will be exploited to jointly separate audio and video sources.

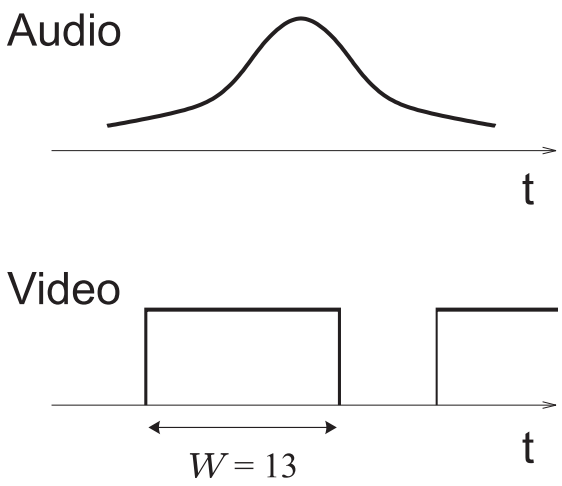


Figure 5.3 – *Sketches of the audio and video features.*

Clustering

Our objective is to detect and localize the sources of an audio signal on the video. We know the temporal relationship between audio and video atoms in the sequence, but how can we localize the signals? The video atoms that are more frequently related to the audio atoms are chosen in the proposed model as possible sound sources, locating them in the image. One visual source can be made up of several video atoms whose movements are coherent with the soundtrack evolution, like the lips, the chin and even the eyes. We propose thus to cluster video structures that are correlated with audio atoms and that are spatially close, to form a source. In order to easily define a measure of proximity between video atoms, we associate to each atom $\phi_n^{(v)}$ one fixed location over the image plane, (t_{1_n}, t_{2_n}). It is here that the hypothesis of having a quasi-static scene comes into play : the position of video atoms can in fact change from frame to frame, but if movements are limited we can reasonably assign one fixed position to each video atom throughout the sequence. This is what we do here and we assign to each video structure its position over the image plane on the first frame of the sequence.

In this section, we define an empirical *confidence value* κ_n of the n-th video atom as the sum of the MP coefficients c_k of all the audio atoms associated to it in the whole sequence :

$$\kappa_n = \sum_k c_k \quad \text{with } k \text{ s.t. } \chi_{k,n} \neq 0\,. \tag{5.4}$$

Thus, this confidence value is a measure of the number of audio atoms related to it and their weight in the MP decomposition of the audio track. Each video atom thus is characterized by its position over the image plane and by its confidence value, i.e. $((t_{1_n}, t_{2_n}), \kappa_n)$.

Looking at Fig. 5.4, the idea of a clustering is very intuitive. The picture shows the position of video atoms with confidence value different from zero for the test sequence shown in Fig. 5.1. The height of the peaks indicate the confidence of each atom. Comparing this picture with Fig. 5.1 [Left] it is clear that atoms with high confidence are grouped around the speakers mouths, one on the left and the other on the right of the image. Atoms with higher confidence value form two different and well separated groups pointing out the sources, while those lying far away from these regions have considerably smaller confidence. It seems thus that the audio-video atoms association has been successful, pointing out visual features that are close to the actual sound sources.

The clustering algorithm that we propose groups the video atoms without any assumption about

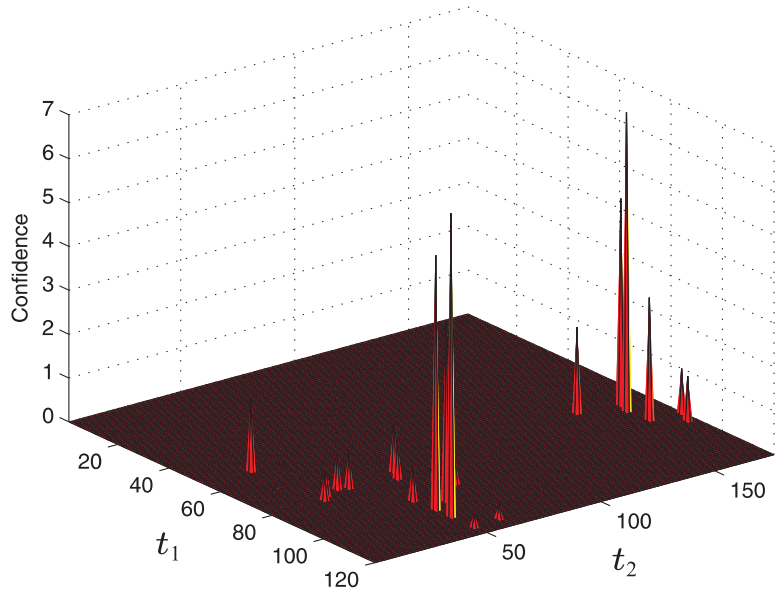


Figure 5.4 – *Video atoms location over the image plane. Their confidence value is represented in the third dimension.*

the number of sources present in the sequence. This characteristic makes the algorithm robust and it guarantees the blindness of the framework. The clustering is divided into three main steps : the first step consists in iteratively creating clusters by selecting the video atoms with highest confidence value and aggregating sufficiently close points around them. This leads to the creation of Z clusters. The second step of the algorithm estimates the centroid of each cluster. Finally, we use a simple criterium to eliminate non significant clusters and keep $N_S \leq Z$ clusters whose centroids provide the estimated positions of the sound sources. Below we give more details about each step of the algorithm.

Clusters Creation First the algorithm creates Z clusters $C_i \subset P$ where $P = \{((t_{1_n}, t_{2_n}), \kappa_n)\}_n$ is the set of all points to be classified, i.e. all video atoms with confidence value different from zero. The clusters are created with the following iterative algorithm :

1. Initialization : $Z = 0$, $P_Z = P_0 = P$;
2. Find the point $((\tilde{t}_{1_n}, \tilde{t}_{2_n}), \tilde{\kappa}_n) \in P_Z$ with highest confidence value. It has the most important audio atoms associated, and consequently this video atom is the most probable to be the center of a source;
3. Create a new cluster C_Z aggregating all the video atoms that are closer than a spatial maximum distance to $(\tilde{t}_{1_n}, \tilde{t}_{2_n})$ (*cluster size* defined in pixels);
4. Remove all the video atoms assigned to this cluster from the set of points to be classified, i.e. $P_{Z+1} = P_Z \setminus C_Z$;
5. Stop the algorithm if all the points with confidence over the mean are already classified, otherwise increment $Z \leftarrow Z + 1$ and go back to step 2. Only video atoms with significant confidence value can be the center of a new cluster.

Concerning the clusters creation, the most important parameter to fix is the cluster size. This characteristic determines the number of clusters created by the algorithm, and, consequently, the number of sources detected in the first stage of the clustering. However, as we will see in the next paragraphs, the setting of this parameter does not affect significantly the final result. Thus, in the third step of the algorithm, a radius around the main video atom between 30 and 60 pixels (width of the image : 176 pixels) is appropriate for the case we are analyzing. The database we are using in fact contains sequences with two speakers significantly separated as in Fig. 5.1.

The decision bound in step 5 is used to force the algorithm to form clusters around atoms with high confidence, that are thus more likely to be part of a source. As we can see in Fig. 5.4, most of the considered video atoms have a small confidence value (they are sometimes related to only one audio atom) and only few atoms exhibit high confidence. Therefore the threshold applied in step 5 is quite "conservative", and empirically we have seen that it is basically impossible to ignore video atoms belonging to real sources.

Estimation of the Centroids This step computes the center of mass of the video atoms belonging to the clusters. In order to perform it, the confidence value of every atom is taken as the mass, and it weights its contribution to the calculation of the centroid position over the image. The previous step of the algorithm has created Z clusters, $\{C_i\}_{i=1}^{Z}$. We calculate the centroid of each cluster C_i, $(\hat{t}_{1_i}, \hat{t}_{2_i})$, as :

$$(\hat{t}_{1_i}, \hat{t}_{2_i}) = \left(\frac{\sum_{j \in C_i} \kappa_j \cdot t_{1j}}{\sum_{j \in C_i} \kappa_j}, \frac{\sum_{j \in C_i} \kappa_j \cdot t_{2j}}{\sum_{j \in C_i} \kappa_j} \right) , \tag{5.5}$$

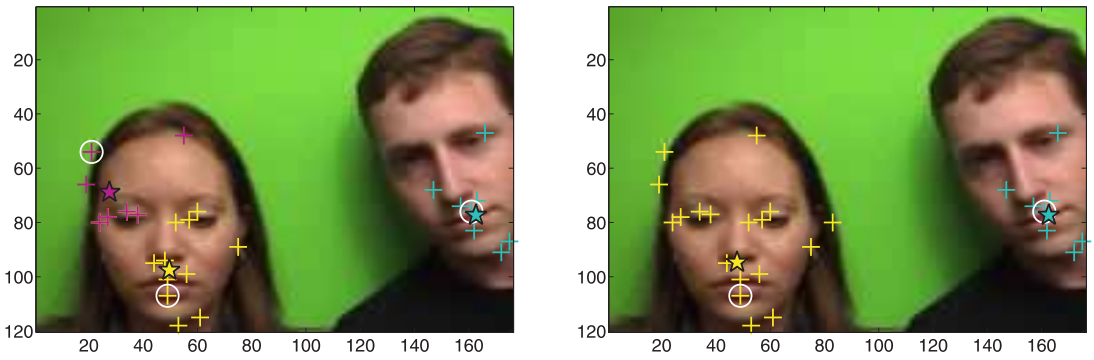
where (t_{1j}, t_{2j}) are the coordinates of the video atoms and κ_j their confidence values. These centroids are the coordinates in the image where the algorithm locates the audio sources. In this kind of sequences with several speakers, the centroids should be close to their mouths. Examples of the created clusters and their calculated centroids are shown in Fig. 5.5, where the test clip of Fig. 5.1 is used. We can see that some of the clusters are, as expected, close to the speakers mouth, while others do not represent a source (*magenta* cluster, the less important and the last one created, with cluster size 40 pixels). In the next step the proposed clustering algorithm takes into account these *unreliable clusters* and eliminates them.

Elimination of *Unreliable Clusters* We define the *cluster confidence value* K_{C_i} as the addition of the confidence values κ_j of the atoms belonging to the cluster indexed by C_i :

$$K_{C_i} = \sum_{j \in C_i} \kappa_j \, .$$

Based on this measure, *unreliable clusters*, i.e. clusters with small confidence value K_{C_i}, are removed and their elements are assigned to the closer reliable cluster. In this way we obtain the final set of clusters $\{C'_i\}_{i=1}^{N_S}$, with $N_S \leq Z$, whose centroids indicate the spatial location of the N_S detected sources.

A group of atoms is considered to be an unreliable cluster if its confidence value is 0.2 times the maximum value of K_{C_i} found. There are two main factors that influence the choice of this parameter. On the one hand, this threshold has to be high enough to eliminate the clusters that do not represent a speaker. Sometimes, a small cluster size involves the appearance of more than one cluster per source (e.g. the *magenta* cluster in Fig. 5.5(a)). On the other hand, if this value is too high the algorithm can remove clusters indicating real sources. This would be the case if one of the sources is active for a much longer time than the others. As a result, the video atoms belonging to these speakers would have many more correlated audio atoms and their cluster confidence value



(a) Clusters creation with radius 40 pixels

(b) Clusters creation with radius 60 pixels

Figure 5.5 – *Clusters created using different cluster sizes in step 4 of the algorithm. The atom represented with a white circle (∘) is the one with higher* confidence value *that builds the cluster in step 2 of the algorithm. Crosses (+) represent the coordinates of the video atoms aggregated to the cluster in step 3. Finally, the centroids of each cluster are indicated by a star (⋆). Each cluster is represented with a different color, from the first to last created (descendent importance of the cluster) : yellow, cyan and the last one, magenta, which is present only on picture (a). Actually, the magenta cluster will be classified as* unreliable *and eliminated at the next step of the processing.*

will be considerably bigger. Empirically, we observed that the threshold value we have applied fulfills the explained requisites. In the case shown in Fig. 5.5(a) for example, the *magenta* cluster is erroneously detected because of the small cluster size used. However its confidence value is small since it is made up of video atoms with very low confidence (see Fig. 5.4), and it is thus removed.

Using the proposed approach, a good speaker localization is achieved creating audiovisual synchronous structures and spatially grouping them into sources using a robust clustering algorithm. The number of clusters does not have to be specified in advance since a confidence measure is introduced to automatically eliminate unreliable clusters. The algorithm is robust and the localization results do not critically depend on the choice of the cluster size parameter nor on the confidence threshold.

5.2.2 Phase 2 : Separation and Reconstruction of Video Sources

Once the sources locations are estimated, the next step to carry out is to extract all the visual structures, separate them and associate them to the detected video sources. The characteristic to use at this point is the spatial distance between elements, since video sources are typically well separated on the image plane. The fundamental goal of this step regarding the audio separation objective is to classify the video atoms into the detected sources. Thus, we define a *maximum distance* in pixels from the centroid. All the points that are closer than such distance from a centroid $(\hat{t}_{1_i}, \hat{t}_{2_i})$ are assigned to the corresponding source. With this procedure, we end up with a set of N_S clusters, $\{S_i\}_{i=1}^{N_S}$. Each group of video atoms S_i describes the video modality of an audiovisual source. To set the *maximum distance* parameter, we have to take into account several conditions :

- We do not want to assign one video atom to more than one source. In this case, we would not be separating, and there would be errors in this classification and in the posterior audio separation;
- At the same time, the radius has to be big enough to contain the maximum number of atoms

Figure 5.6 – *Example of the video sources reconstruction. On the left picture the left person is speaking while on the right picture the right person is speaking.*

belonging to each source. It is important not to lose all the video atoms related to an audio atom (in that case it would not be possible to posteriorly assign it to a source and reconstruct completely the sequence without severe energy losses);

- It is important not to assign to one source structures belonging to the other sources.

Empirically, we have noticed that a radius around the centroids of 60 pixels (width of the image : 176 pixels) is appropriate. As shown in Fig. 5.1 in fact the database we consider is made up of sequences with two speakers significantly separated.

At the end of this phase, video sources are detected and reconstructed : the video separation is satisfactorily performed. Figure 5.6 shows an example of the reconstruction of the current speaker detected by the algorithm. For each frame, only video atoms close to the sources estimated by the presented technique are considered. Thus, to carry out the reconstruction, the algorithm adds their energy and the effect is a highlight of the speaker's face. In both frames, the correct speaker is detected.

5.2.3 Phase 3 : Temporal Localization of Audio Sources

At this point of the processing, we know the location of the video source on the image plane, the video atoms belonging to each one of the sources and the temporal relation between audio and video atoms represented by the correlation scores $\chi_{k,n}$ calculated with (5.3). Since we assume a one-to-one correspondence between audio and video sources, we also know the number N_S of audio-video sources present in the sequence. What we want to do now is to assign each audio atom to a source and in particular to detect time periods during which the different sources are active alone.

For every audio atom we take into account all the video atoms related to it, their correlation scores and their classification into a source. According to this, the audio atom is assigned to the source with higher number of video atoms belonging to it, but also rewarding the temporal synchrony between these video atoms and the analyzed audio structure. Therefore, for each audio entity $\phi_k^{(a)}$ the assignation to a source can be done in the following way :

1. Take all the video atoms $\phi_n^{(v)}$ correlated with the audio atom $\phi_k^{(a)}$, i.e. for which $\chi_{k,n} \neq 0$;
2. Each of these video atoms is associated to an audiovisual source S_i ; for each source S_i compute a value H_{S_i} that is the sum of the correlation scores between the audio atom $\phi_k^{(a)}$ and the video atoms $\phi_j^{(v)}$ s.t. $j \in S_i$:

$$H_{S_i} = \sum_{j \in S_i} \chi_{k,j} \, ;$$

3. Classify the audio atom into the source S_i if the value H_{S_i} is "big enough" : here we require H_{S_i} to be twice as big as any other value H_{S_h} for the other sources. Thus we attribute $\phi_k^{(a)}$ to S_i if

$$H_{S_i} > 2 \cdot H_{S_h} \quad \text{with } h = 1, \ldots, N_S \,,\; h \neq i \,.$$

If this condition is not fulfilled, this audio atom can belong to several sources and further processing is required.

The decision bound in step 3 is introduced because, at this point of the processing, not all audio atoms can be clearly classified into one of the sources. Some of them are in an intermediate position and we cannot base the decision only on a small difference of the sources scores H_{S_i}. These atoms may belong to more than one source, or we could be making a mistake choosing one source instead of another one. This is typically the case when several speakers are simultaneously active. For these atoms additional processing is required, as it will be shown in the next section.

As an example, let us consider the situation shown in Table 5.1. Here one audio atom has six video atoms associated (i.e. with correlation scores different from 0). Four of them belong to source S_1, and two to source S_2, with the correlation scores shown in in the table. Then, the sum of the scores are 13.88776 and 1.71717 for sources S_1 and S_2 respectively. The score for the first source is much bigger (approximately eight times bigger than the other) and thus this audio atom will be assigned to source S_1.

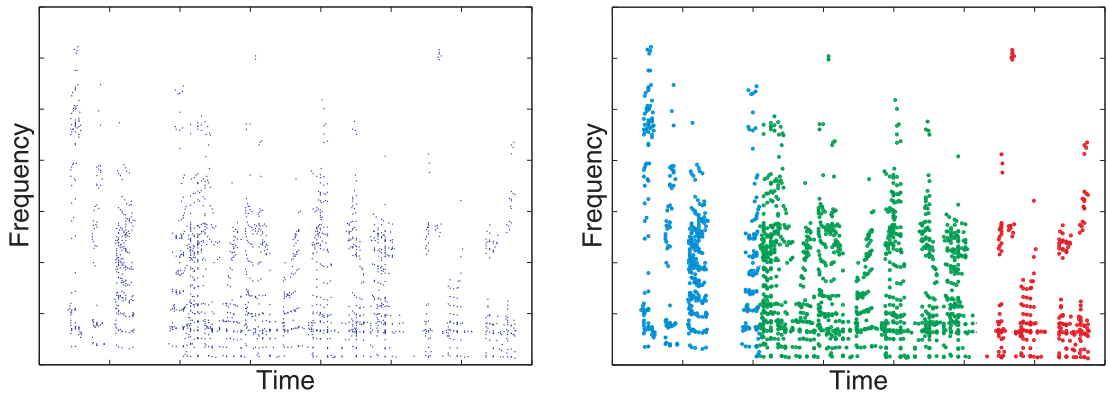
Source S_1	**Source** S_2
6.9348	1.1146
5.8186	0.60257
0.809	
0.32536	
13.88776	1.71717

Table 5.1 – *Example of the list of correlation values between one audio atom and the correlated video atoms. Four of them belong to source 1 and two to source 2.*

Using this labelling of audio atoms, time periods during which only one source is active are clearly determined. This is done using a very simple criterion : if in a continuous time slot longer than T seconds all audio atoms are assigned to source S_i, then during this period only source S_i is active. In the examples that we provide in this chapter, the value of T is set to 1 second.

The classification of the audio atoms representing the test soundtrack shown in Fig. 5.1 is depicted in Fig. 5.7. The points in the pictures represent the position over the time-frequency plane of the audio atoms centers. The atoms locations in the original mixture are shown in picture (a), while the atoms classification is in (b). The sequence involves two speakers : at the beginning only the girl talks, then both persons speak together and finally the boy only talks. This partitioning of the signal is reflected by the proposed audio source classification method : atoms assigned to the girl and the boy are highlighted in blue and red respectively, while *ambiguous* atoms are indicated with green markers.

When several sources are present, temporal information alone is not sufficient to discriminate different audio sources in the mixture. To overcome this limitation, in these *ambiguous* time slots a time-frequency analysis is performed, which is presented in details in the next section.



(a) Time-frequency representation (b) Audio segmentation

Figure 5.7 – *Example of the classification of audio atoms into the corresponding sources. The points represent the time-frequency position of audio atoms. The atoms of the original mixture are in (a), while the atoms classification is in (b). The speech evolution on the sequence is reflected by the proposed classification method : at the beginning only the girl talks (blue markers), then the two persons speak (green markers) and finally only the boy speaks (red markers).*

5.2.4 Phase 4 : Blind Audio Source Separation Aided by Video

In order to perform the Audio Source Separation task, we have to separate the audio atoms of the sequence both in time and in frequency. What we expect is that the frequency information will aid us to obtain better separation results when the sources are temporally overlapping, since this additional dimension can offer a new possibility of discrimination.

An audio atom $\phi_k^{(a)}$ is characterized by its position on the time-frequency plane, (u_k, ξ_k), and by a set of correlation scores $\{\chi_{k,n}\}_n$ that quantify its degree of correlation with the video atoms. Thus the audio atoms of the MP decomposition constitute the set of K points $A = \{(u_k, \xi_k), \{\chi_{k,n}\}_n\}_{k=0}^{K-1}$. Our aim is to associate each one of these points to one of the N_S audiovisual sources.

We want to underline that the MP algorithm considers a discretized version the time-frequency plane. In that way the atoms of the decomposition can be centered only in a discrete set of points placed on a uniform time-frequency grid. The idea here is basically to estimate for each point of such grid the probabilities to belong to each one of the detected sources, and thus classify the audio atoms according to this information. The starting point is to use the temporal periods during which sources are active alone to compute a probability for each frequency point $\hat{\omega}$ to belong to one source. The proposed method is based on the hypothesis that different sources have different frequency content, since otherwise the frequency assignation would be similar and this analysis vain.

In order to assign all the audio atoms to one of the sources, we have to consider in which of the following cases we are :

Time period with only one active source - We use the temporal analysis result to classify this atom. We already know which source is active at this moment and so it is not necessary to use the frequency information.

Time period with several active sources - There is a mixture in this period, and also frequency analysis is required. Each audio atom in this period is classified into a source according to the probabilities of its coordinates in the time-frequency plane. An audio atom centered in

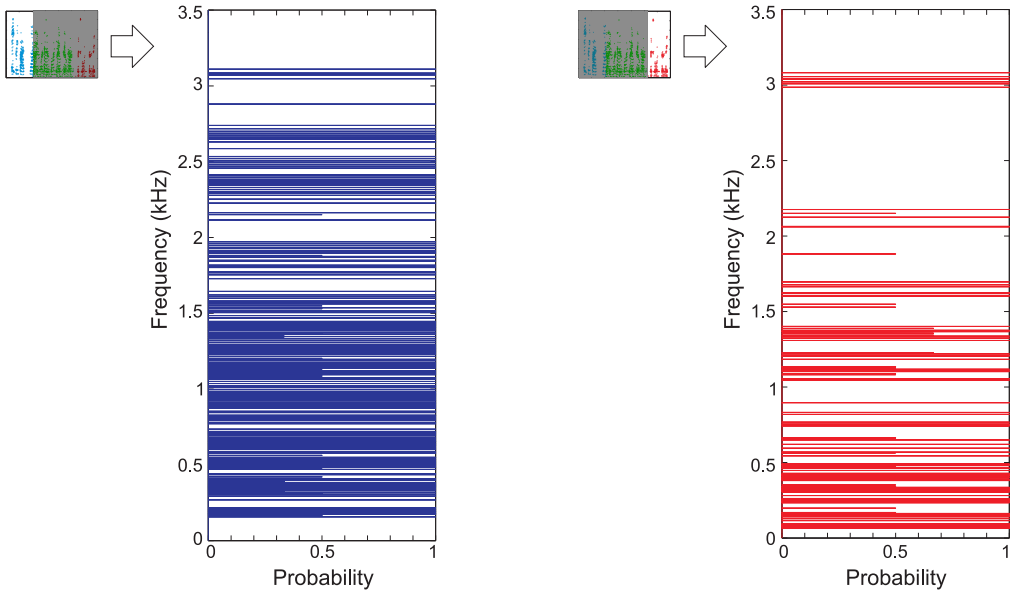


Figure 5.8 – *Estimated frequency probabilities for the female [Left] and male [Right] speakers involved in the test sequence. The two probabilities are estimated on parts of the test sequence during which the subjects speak alone (indicated by blue and red dots in the spectrogram of Fig. 5.7(b) that is reproduced on the upper left corners of the figures).*

coordinates $(\hat{t}, \hat{\omega})$ will be associated to source S_i if

$$P_{S_i}(\hat{t}, \hat{\omega}) = \max\{P_{S_j}(\hat{t}, \hat{\omega})\} \text{ , with } j = 1, \ldots, N_S \,. \tag{5.6}$$

These maps of probabilities are built computing the product between time and frequency probabilities as :

$$P_{S_i}(\hat{t}, \hat{\omega}) = P^T_{S_i}(\hat{t}) \cdot P^\Omega_{S_i}(\hat{\omega}) \tag{5.7}$$

where $P^T_{S_i}(\hat{t})$ is the probability of an audio atom with time index $\hat{t}$ to belong to source S_i, and $P^\Omega_{S_i}(\hat{\omega})$ is the probability for an audio atom with frequency index $\hat{\omega}$ to belong to source S_i. The frequency probabilities $P^\Omega_{S_i}(\hat{\omega})$ are computed considering temporal slots during which the sources are active alone (e.g. the blue and red portions of the spectrogram in Fig. 5.7(b)), so that a reliable association between audio atoms and sources can be established. In these signal slots we keep for every value of $\hat{\omega}$ the set of atoms $A_{\hat{\omega},k,n} = \{(u_k, \xi_k = \hat{\omega}), \{\chi_{k,n}\}_n\}_k$ that have frequency index $\xi_k = \hat{\omega}$. The probability $P^\Omega_{S_i}(\hat{\omega})$ of the frequency value $\hat{\omega}$ to be associated to source S_i is estimated as the number of atoms with frequency index $\hat{\omega}$ that we know to belong to source S_i (e.g. in the red *or* blue region in Fig. 5.7(b)) divided by the total number of atoms with frequency index $\hat{\omega}$ in the considered signal slots (e.g. the red *and* blue regions in Fig. 5.7(b)). Thus we can write :

$$P^\Omega_{S_i}(\hat{\omega}) = \frac{\text{card}\,(A_{\hat{\omega},k\in S_i,n})}{\text{card}\,(A_{\hat{\omega},k,n})} \,, \tag{5.8}$$

where card$(\cdot)$ is the cardinality (number of elements) of a set of points. The probability of each frequency value is normalized to one, i.e. $\sum_{i=1}^{N_S} P^\Omega_{S_i}(\hat{\omega}) = 1$.

Figure 5.8 shows the estimated probabilities for every frequency point for the two speakers of the sequence shown in Fig 5.1. Fig. 5.8 [Left] represents the frequencies probabilities for the girl and Fig. 5.8 [Right] shows the frequencies probabilities for the boy. As expected, lower frequencies are more likely to belong to the boy while higher ones are associated with the girl. This characteristic

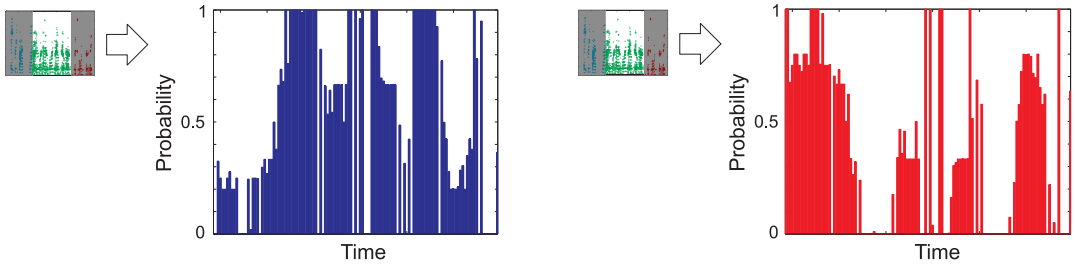


Figure 5.9 – *Estimated temporal probabilities for the female [Left] and male [Right] speakers involved in the test sequence. The two probabilities are estimated on the part of the test sequence during which both persons speak together (indicated by green dots in the spectrogram of Fig. 5.7(b) that is reproduced on the upper left corners of the figures).*

remarks the pitch frequency for both speakers. It is also possible to see the formant periodicity for both speakers, with their frequency components situated in pitch multiples. All these observations reinforce the belief that the temporal separation is well performed, correctly discriminating periods with only one active source. This allows to characterize clearly every source in the frequency domain. Of course, sequences with speakers with closer pitches would mean more overlapping in the frequency classification and, consequently, the impossibility to perform a good separation in this domain using such a simple static approach.

The temporal probability $P^T_{S_i}(\hat{t})$ instead, is estimated in period during which both sources are supposed to be active (e.g. the green part of the spectrogram in Fig. 5.7(b)). Since no sure association between audio atoms and sources can be established in these mixed periods, temporal probabilities are estimated exploiting the correlation information between audio and video atoms (that have already been assigned to sources) given by the correlation scores $\{\chi_{k,n}\}_n$. For each time instant $\hat{t}$ we select the set of points $A_{\hat{t},k,n} = \{(u_k = \hat{t}, \xi_k), \{\chi_{k,n}\}_n\}_k$ and we compute the temporal probabilities $P^T_{S_i}(\hat{t})$ as :

$$P^T_{S_i}(\hat{t}) = \frac{\sum_{k \in A_{\hat{t},k,n \in S_i}} \chi_{k,n}}{\sum_{k \in A_{\hat{t},k,n}} \chi_{k,n}} \,. \tag{5.9}$$

This probability basically acts like a mask : when it is 0 means that no chance is given to source S_i to be active, since no correlated event between the video source S_i and the audio signal is detected at this time instant. Again the probability of each temporal value is normalized to one, i.e. $\sum_{i=1}^{N_S} P^T_{S_i}(\hat{t}) = 1$. Figure 5.9 shows the estimated probabilities for every time point for the two speakers of the considered test sequence. The two probabilities are computed on the section of the clip during which both persons speak together (indicated by green dots in the spectrogram of Fig. 5.7(b)). Fig. 5.8 [Left] represents the temporal probabilities for the girl while Fig. 5.8 [Right] shows the temporal probabilities for the boy.

For each time-frequency point $(\hat{t}, \hat{\omega})$ the probability $P_{S_i}(\hat{t}, \hat{\omega})$ in (5.7) is computed as a product between $P^T_{S_i}(\hat{t})$ and $P^\Omega_{S_i}(\hat{\omega})$. One aspect has to be taken into account : not all the frequency values necessarily have a probability associated. In this case, the closest frequency with a probability value associated is used in (5.7). Figure 5.10 shows the final time-frequency probabilities computed for the two speakers in the test sequence. The color map of the pictures goes from blue to red through green and yellow and the pixel intensities reflect the probability of the time-frequency point. For example a blue point has probability zero to belong to the considered source. The probabilities hold for the part of the test sequence during which both persons speak together (indicated with green dots in the spectrogram of Fig. 5.7(b)). The probability for the female speaker is depicted on the left picture and the one for the male speaker is on the right.

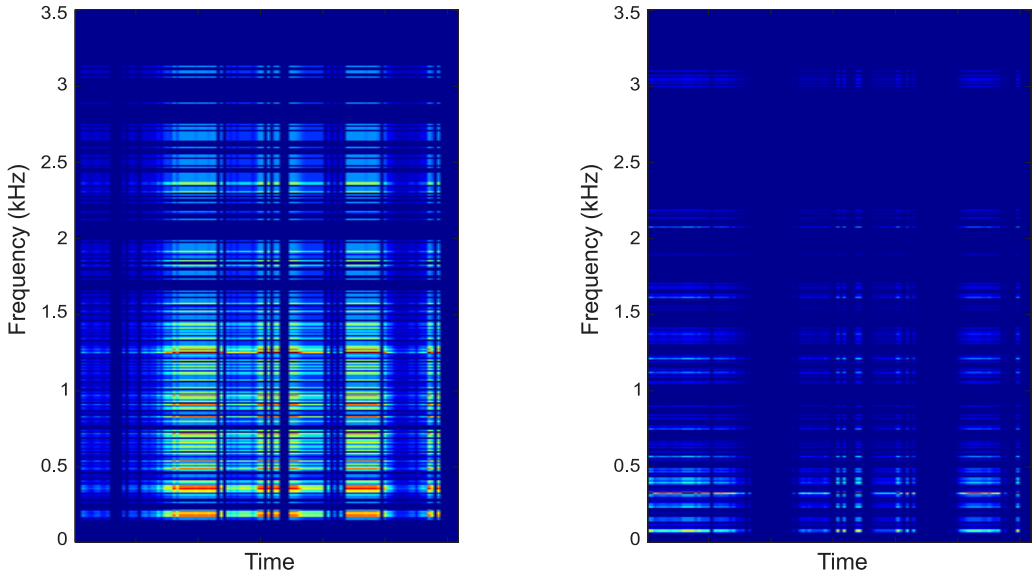


Figure 5.10 – *Estimated time-frequency probabilities for the female [Left] and male [Right] speakers present in the considered test sequence. The color map of the pictures goes from blue to red through green and yellow and the pixel intensities reflect the probability of the time-frequency point. The probabilities concern the part of the test sequence during which both persons speak together. Please note that the original probability maps have been low-passed and down-sampled by a factor of 100 in the frequency dimension in order to be visualized.*

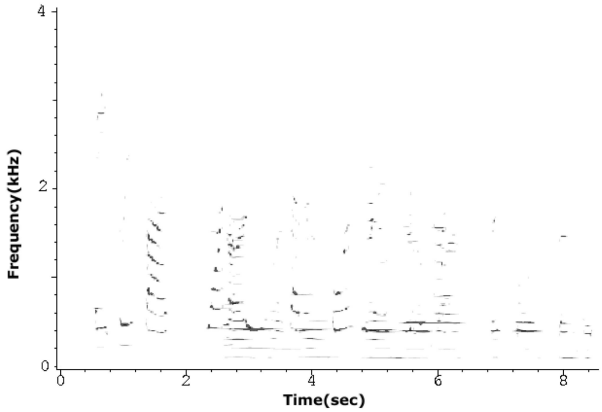
Figure 5.11 shows in (a) the spectrogram of an audio signal obtained with 2000 Gabor atoms selected by MP. The signal is the soundtrack of the test sequence shown in Fig. 5.1. In the sequence the girl starts to speak alone, then both persons speak at the same time and finally the girl stops talking and the boy speaks alone. On picture (b) the spectrogram corresponding to one separated source (the female speaker) is shown, while in (c) the spectrogram of the second detected source (the boy), is shown. The color map of the time-frequency plane images goes from black to red, through blue, green and yellow, and the pixel intensity represents the value of the energy at each time-frequency location, computed as in section 3.5.2.

Some considerations can be done observing these pictures. First of all, it seems that characteristic energy distributions of each speaker are correctly extracted. For example, in Fig. 5.11(b) we can see the separated signal for the girl. The first part of the soundtrack (2 initial seconds in the spectrogram) only contains her speech, so that it is possible to observe clearly the characteristic evolution of the frequency components of her voice. The same evolution is repeated in the period were the two persons are speaking at the same time. Another aspect to remark is that low frequencies, characteristic of male voices, are correctly assigned to the second source, the male speaker, and there is no presence of them in the spectrogram of the female speaker.

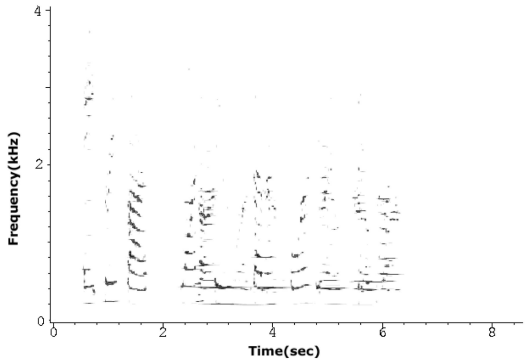
Reconstruction of the Separated Signals

The audio signal coming from a source is reconstructed by simply adding the audio atoms classified in this source, weighted by their energy coefficients. Therefore the i-th audio source, $\alpha_{S_i}(t)$, can be reconstructed as :

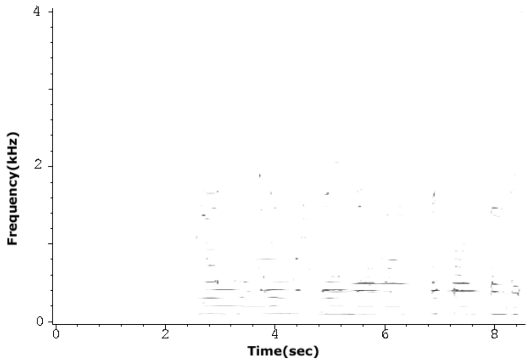
$$\alpha_{S_i}(t) \approx \sum_{k \in S_i} c_k \, \phi_k^{(a)}(t) \,, \tag{5.10}$$



(a) Original sequence



(b) Separated source 1 (girl)



(c) Separated source 2 (boy)

Figure 5.11 – *Source Separation of a real-world mixture representing a boy and a girl uttering digits simultaneously. The color map of the time-frequency plane images goes from white to black, and the pixel intensity represents the value of the energy at each time-frequency location.*

where c_k is the coefficient found by MP and corresponding to the Gabor atom $\phi_k^{(a)}(t)$ and S_i indexes the set of atoms attributed to the i-th source. The reconstructed sources $\alpha_{S_i}(t)$ are time-evolving waveforms that can be heard using a media-player. The reconstructed sources shown in Fig. 5.11, for example, result well audible and the digits uttered by the two speakers can be clearly distinguished. However, an objective, quantitative measure of the quality of the source separation and reconstruction is required, in order to asses the performances of the proposed algorithm.

5.3 Experiments

In this section the proposed BAVSS algorithm is evaluated on synthesized audiovisual mixtures. The interest of analyzing synthesized sequences resides in the fact that a ground truth can be assessed and thus an objective measure of the discrepancy between this ground truth and the reconstructed sources can be defined. The features used to evaluate the algorithm are the percentage of correctly classified atoms for each audio source and the percentage of acoustic energy of the source that these correctly classified atoms represent.

Synthesized sequences are generated using clips taken from the *groups* partition of the CUAVE database [88] with one girl and one boy uttering sequences of digits alternatively. The video data is at 29.97 fps with a resolution of 480×720 pixels, and the audio at 44 kHz. The video data have been

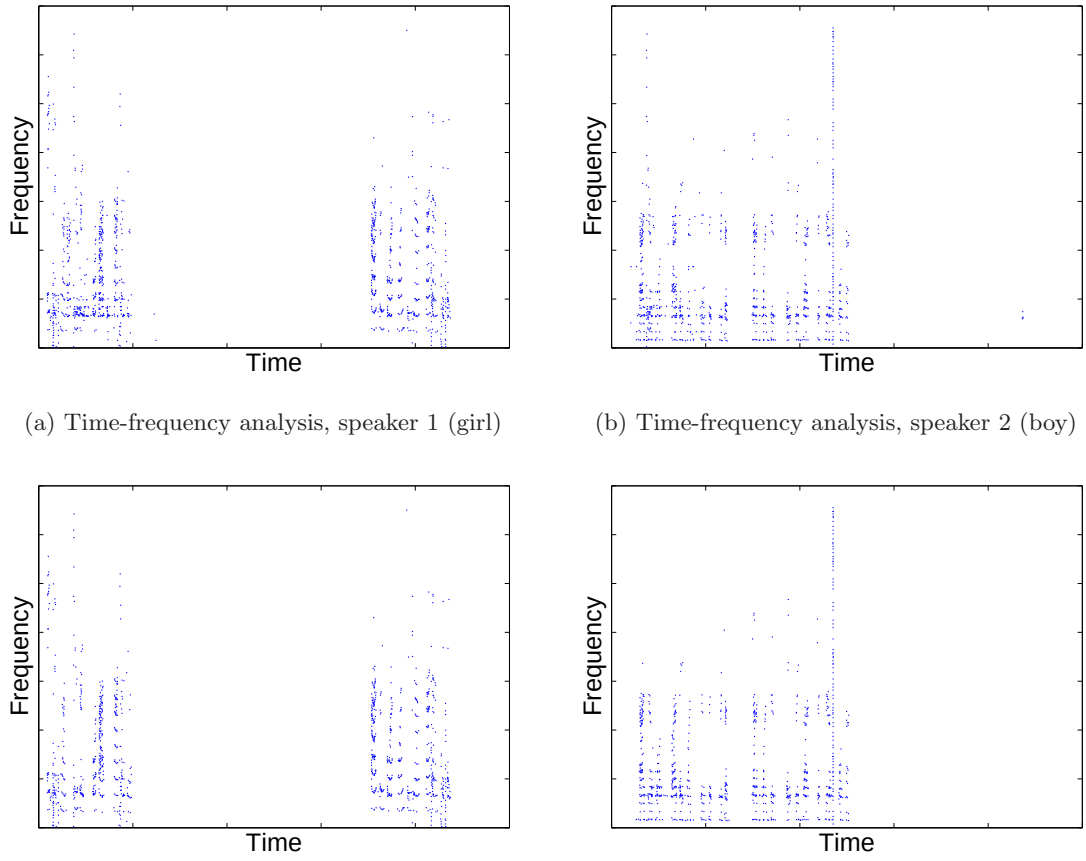
resized to a resolution of 120×176 pixels, while the audio signal has been sub-sampled to 8 kHz, with still a good audible quality. The video sequence is decomposed into 100 video atoms and the mixture soundtrack is decomposed into 2000 Gabor atoms. The audio and the video atoms of one speaker are then temporally shifted in order to obtain time slots with both speakers active. The steps carried out to synthesize the sequences employed in the experimental tests are the following :

1. Choose a clip of the *groups* section of the CUAVE database where two speakers (a boy and a girl) utter numbers in turns;
2. Shift the audio atoms of one speaker so that their voices are overlapped part of the time. The MP decomposition of the audio gives us the temporal position of the audio atoms belonging to each one of the speakers. Thus, we only need to take all the atoms of one speaker, which are temporally separated from those of the other one since they are speaking alternatively, and change their temporal index appropriately. The same quantity is added or subtracted from all the atoms;
3. The same procedure is applied to the video atoms. Once the video sequences decomposed into 2D time-evolving atoms, the feature to analyze is the evolution of the video atoms displacement through time. In the CUAVE database each speaker is located at one side of the image plane, so that video atoms belonging to one speaker have the abscissa value between pixels 1 and 88, and the other one between 89 and 176 (the resolution of the video being 120×176). Thus, the procedure consists in temporally shifting the video atoms corresponding to one speaker by the same value of the audio atoms belonging to the same speaker. Please note that the shift in the audio domain is in samples and we have to convert it in frames to apply the same temporal shift to the video atoms.

This procedure translates the whole part of the audiovisual sequence belonging to one speaker in order to have a synthetic mixture where both speakers are uttering different numbers at the same time. In the resultant synthetic clips, four cases are represented : both persons speak at the same time, only the boy or the girl speaks or silence.

First, the percentage of correct atoms is assessed. Figure 5.12 shows the sources extracted by the proposed algorithm [Top] and the real ones represented with 2000 Gabor atoms [Bottom], for a syntectic sequence generated by applying a shift of 150 frames to the sequence part with the male speaker in clip `g20` of the CUAVE database. Using the proposed technique on this sequence 92% of atoms for the girl and 90% for the boy are correctly classified. This is a good result, taking into account that it is at the atoms level that our algorithm is performed. Thus, on average our algorithm assigns 91% of the audio atoms to the correct source.

Another measure is employed in order to evaluate this method : the percentage of the original energy that these correctly classified atoms represent. This value gives us the information relative to the difference of the original and estimated soundtracks for each speaker after the reconstruction step. This measure is performed in order to discard the very improbable fact that the 9% of audio atoms that are misclassified contribute to the separated soundtracks with the main part of the energy, i.e. these audio atoms are the first in the MP decomposition of the original mixture. For each source, this percentage is computed as the sum of the coefficients of all the atoms correctly assigned by the algorithm to the source divided by the sum of the coefficients of all the atoms belonging to this source. Therefore, this percentage can be seen as the part of the estimated signal belonging to the original source. The remaining energy is due to the assignation of audio atoms to the incorrect speaker and constitutes the noise of the separated signal estimated by the algorithm. Figure 5.13 shows the original waveforms reconstructed with 2000 Gabor atoms on the right and those estimated by the proposed time-frequency analysis on the left.



(a) Time-frequency analysis, speaker 1 (girl)

(b) Time-frequency analysis, speaker 2 (boy)

(c) Real audio atoms, speaker 1 (girl)

(d) Real audio atoms, speaker 2 (boy)

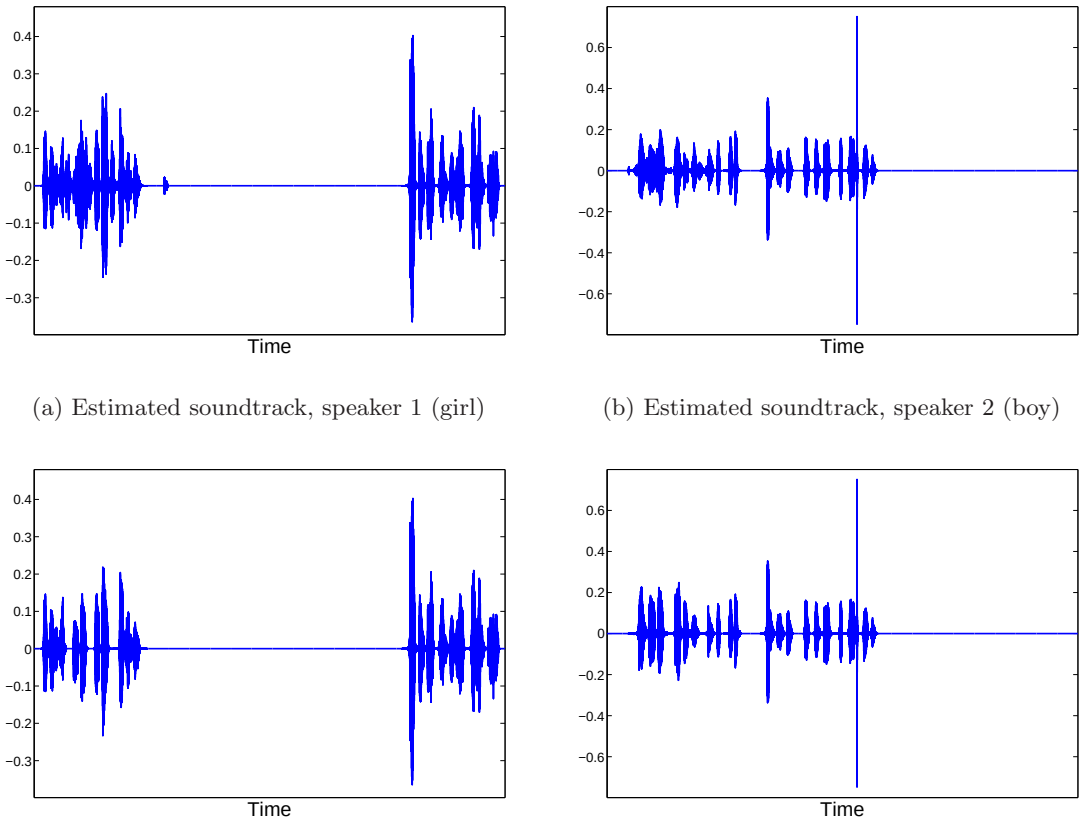
Figure 5.12 – *Comparison between audio atoms resulting of time-frequency analysis in a synthetic mixture [Top] and the original ones [Bottom]. The points are the centers of the audio atoms over the time-frequency plane. The sequence is generated by applying a shift of 150 frames to the male speaker in clip g20 of CUAVE database.*

Sequence	% correct atoms		% correct energy	
	girl	boy	girl	boy
g12 shift 100 frames	86	54	73	42
g20 shift 150 frames	92	90	92	86
g21 shift 130 frames	83	81	81	75
g21 shift 169 frames	82	78	84	73

Table 5.2 – *Results obtained with synthetic sequences generated for different clips of CUAVE database.*

Waveforms are very similar in the original and estimated sequences, and the percentages of the original energy that the correct atoms assigned to each source represent the 92% and 86% for the male and female speaker respectively. These percentages are high and similar to those obtained for the number of correct atoms assigned to each speaker (92% and 90%). It seems thus that correctly assigned audio atoms represent most of the energy of the speakers separated signals. Results obtained analyzing different sequences are summarized in Table 5.2.

The values obtained for the percentage of correct atoms and the percentage of energy that these



(a) Estimated soundtrack, speaker 1 (girl)

(b) Estimated soundtrack, speaker 2 (boy)

(c) Ground truth, speaker 1 (girl)

(d) Ground truth, speaker 2 (boy)

Figure 5.13 – *Comparison between estimated [Top] and real [Bottom] soundtracks for a synthetic sequence generated by applying a shift of 150 frames to the male speaker in clip g20 of the CUAVE database. Please note that the ground truth soundtracks are not the original ones but their reconstructions using 2000 atoms selected by MP.*

atoms represent are similar. We can thus argue that the algorithm distributes the errors over audio atoms of all sizes, and the percentage of correct atoms is already a good measure of the algorithm performance. Results are satisfactory, around 80–90% except for sequence g12 of CUAVE database, with a worse performance for the boy. Table 5.2 also shows that the results obtained are linked with the sequence to analyze and they are independent of the shift introduced. The performance for sequence g21 is around 80% with shifts of 130 or 169 frames, with a small difference in favor of the first case.

It is important to underline that lower performances in sequence g12 are mostly due to errors done in the sequence part during which both speakers are active and they are caused by the low discriminative power of the simple model based on the probability maps of the speakers. Actually, for all tested sequences the time periods during which the sources are active alone are correctly localized except for some minor error in sequence g12. The signals in these time slots are essentially perfectly reconstructed, with a Signal to Noise Ratio (SNR) between the ground-truth MP reconstructions and the separated sources of about 50 dB. In contrast, performances are much lower in mixed periods. Although the separated speech signals are still audible and the uttered digits can be clearly distinguished most of the time, we have measured SNR values ranging from 3 dB (for the first part of the signals shown in Fig. 5.13(b),(d)), down to -1dB. This shows that while the proposed

framework is able to localize the sources on the video and to detect time slots during which a speaker alone is present, improvements are needed in the time-frequency separation of audio mixtures. This can be done using more complex one-microphone source separation techniques that can be either discriminative [7, 96] or generative [59, 95, 99] (see section 5.1.3). An HMM-based generative model like the one proposed in [95] would probably match well our considered scenario, since we could still keep a completely blind setting and we could think of learning a model of the sources in time slots during which they are active alone. However this type of techniques typically require large training audio portions that can be unavailable in the presented scenario. Another interesting option could be then the use of a blind method to track the evolution of harmonics and resonances, like the one proposed in [96], but aided here by the information available in time periods presenting audiovisual sources active alone.

As a final remark, we have noticed that the quality of the reconstructed signals is considerably better for synthetic sequences than for real ones. This effect is caused by the change in the speakers fundamental frequency, and, consequently, spectral harmonics, when they speak simultaneously in real sequences. Humans tend to change their speech characteristics in order to differ more from the other speakers and to be, thus, more easily heard. This change in the sources frequency behavior causes a worse performance of the algorithm, since the speakers models are learned in temporal periods during which they are alone.

5.4 Discussion

In this chapter we have introduced a new algorithm to perform a Blind Audiovisual Source Separation task. We consider sequences made of one soundtrack and the video signal associated, without the stereo audio signal usually employed for the BASS task. The method builds correlation between acoustic and visual structures that are represented using atoms taken from redundant dictionaries. Video atoms that exhibit strong correlations with the audio track and that are spatially close are grouped together using a robust clustering algorithm that can confidently count and localize on the image plane audiovisual sources. Then, using such information and exploiting the coherence between audio and video signals, audio sources are localized as well and separated. The presented algorithm needs time periods with sources active alone to predict their behavior in the mixture. This condition is however not very restrictive, since it is rare that in real-world mixtures all the sources are active all the time.

Several tests are performed in real-world and synthetic sequences, and encouraging results are obtained for both of them. The speaker spatial localization is successfully performed in challenging sequences where two persons speak simultaneously. Concerning the audio source separation part, the audible quality of the separated audio signals is also reasonably good, with reconstructed waveforms close to the original ones. However, we believe that the proposed method can be improved using more sophisticated techniques for the separation of audio sources in time slots that present source mixtures. To this end, HMM-based models [95] or audio feature tracking techniques [96] could be plugged in the proposed framework. Moreover, a more systematic evaluation of the audio separation results should be performed, employing for example the performance evaluation protocol proposed in [110].

6 Learning Multi-Modal Dictionaries

We have shown throughout this thesis that it is possible to design intuitive and effective techniques to analyze multi-modal signal if proper representations of the considered signals are available. The presented algorithms exploit signal processing techniques capable of extracting meaningful structures from audio and video data, making it possible to define and handle in an efficient and relatively simple way correlated cross-modal structures. However, audio-video features are still extracted separately using general dictionaries of audio and video atoms, and then correlations between them are searched. We argue that a better strategy would be to jointly extract meaningful multi-modal structures, introducing cross-modal correlations at the model level.

This chapter introduces a model for multi-modal signals that represents multi-component data as a sparse sum of recurrent multi-modal structures. Such structures can be retrieved from a codebook of functions. Since however the definition of a multi-modal dictionary results extremely complex, we propose as well an algorithm that allows to learn dictionaries of such multi-modal functions. Signal patterns are learned using a recursive algorithm that enforces synchrony between the different modalities and de-correlation between the dictionary elements.

6.1 Modelling and Understanding

In this section we introduce a new model to represent multi-modal signals. Instead of separately decomposing each signal modality over a dictionary as it was done in previous chapters, here we propose to represent a multi-modal signal as a sparse sum of *multi-modal atoms*. This chapter features three main contributions :

- In Section 6.1.1 we define a general signal model to represent multi-modal data using sparse representations over dictionaries of multi-modal functions and in Section 6.1.2 we refine such model adding two properties that are useful in order to represent real-world multi-modal data, notably synchrony between the different components of multi-modal atoms and shift invariance of the basis functions;
- In Section 6.2 we propose an efficient algorithm to learn dictionaries of multi-modal, synchro-

nous, shift-invariant functions;

- In the experiments section we apply the proposed signal model and the learning method to audiovisual data. Results show that the proposed algorithm allows to learn meaningful audio-video signal patterns and that detecting such structures in challenging real-world audiovisual sequences it is possible to effectively detect and localize audio-video sources.

6.1.1 Sparse Approximations of Multi-Modal Signals

Multi-modal data are made up of M different modalities and they can be represented as M-tuples $s = (s^{(1)}, \ldots, s^{(M)})$ which are not necessarily homogenous in dimensionality : for example, audio-visual data consist of an audio signal $s^{(1)} = s^{(a)}(t)$ and a video sequence $s^{(2)} = s^{(v)}(\vec{x}, t)$ with $\vec{x} \in \mathbb{R}^2$ the pixel position. Other multi-modal data such as multi-spectral images or biomedical sequences could be made of images, time-series and video sequences at various resolutions.

To date, methods dealing with multi-modal fusion problems basically attempt to build general and complex statistical models to capture the relationships between the different signal modalities $s^{(m)}$. However, as already underlined in this thesis, the employed features are typically simple and barely connected with the physics of the problem. Efficient signal modelling and representation require the use of methods able to capture particular characteristics of each signal. Therefore, the idea is basically that of defining a proper model for signals, instead of defining a complex statistical fusion model that has to find correspondences between barely meaningful features.

Applications of this paradigm to audiovisual signals have been presented in previous chapters. A sound is assumed to be generated through the synchronous motion of important visual elements like edges. Audio and video signals are thus represented in terms of their most salient structures using redundant dictionaries of functions, making it possible to define acoustic and visual *events*. An audio event is the presence of an audio signal with high energy and a visual event is the motion of an important image edge. The synchrony between these events reflects the presence of a common source, which is effectively localized. The key idea of this approach is to use high-level features to represent signals, which are introduced by making use of codebooks of functions. The audio signal $a(t) = s^{(a)}(t)$ is approximated as a sparse sum of atoms from a Gabor dictionary $\{\phi_k^{(a)}\}_k$,

$$s^{(a)} \approx \sum_{k \in J_a} c_k^{(a)} \phi_k^{(a)} ,$$

while the video sequence $s^{(v)}(\vec{x}, t)$ is expressed as a sparse combination of edge-like functions that are tracked through time, $\{\phi_k^{(v)}\}_k$, as

$$s^{(v)} \approx \sum_{k \in J_v} c_k^{(v)} \phi_k^{(v)} .$$

Such audio and video representations are still quite general, and can be employed to represent any audiovisual sequence.

One of the main advantage of dictionary-based techniques is the freedom in designing the dictionary, which can be efficiently tailored to closely match signal structures. For multi-modal data, distinct dictionaries $\mathcal{D}^{(m)} = \{\phi_k^{(m)}\}_k$ for each modality do not necessarily reflect well the interplay between events in the different signals, since the sets of salient features J_m involved in the models of each modality are not necessarily related to one another. An interesting alternative consists in capturing truly multi-modal events by the means of an intrinsically *multi-modal dictionary* $\mathcal{D} = \{\phi_k\}_k$ made of *multi-modal atoms* $\phi_k = (\phi_k^{(1)}, \ldots, \phi_k^{(M)})$, yielding a multi-modal sparse signal model

$$s \approx \sum_{k \in J} \left(c_k^{(1)} \phi_k^{(1)}, \ldots, c_k^{(M)} \phi_k^{(M)} \right) . \tag{6.1}$$

Here, a common set J of salient multi-modal features forces *at the model level* some correlation between the different modalities.

Given the multi-modal dictionary $\mathcal{D} = \{\phi_k\}_k$ and the multi-modal signal s, the inference of the model parameters J and $\{c_k^{(m)}\}_{k,m}$ is not completely trivial : on the one hand, since the dictionary is often redundant, the are infinitely many possible representations of any signal; on the other hand, choosing the best approximation with a given number of atoms is known to be an NP-hard problem. Fortunately, several suboptimal algorithms such as multi-channel Matching Pursuit [49, 108], can provide generally good sparse approximations.

6.1.2 Synchrony and Shift Invariance in Multi-Modal Signals

Very often, the various modalities in a multi-modal signal will share synchrony of some sort. By synchrony, we usually refer to time-synchrony, i.e. events occurring in the same time slot. When multi-modal signals share a common time-dimension, synchrony is a very important feature, usually tightly linked to the physics of the problem. As explained above, synchrony is of particular importance in audio-visual sequences. Sound in the audio time series is usually linked to the occurrence of events in the video *at the same moment*. If for example the sequence contains a character talking, sound is synchronized with lips movements. More generally though, multi-modal signals could share higher-dimensions, and the notion of synchrony could refer to spatial co-localization, for example in multi-spectral images where localized features appear in several frequency bands at the same spatial position.

For the sake of simplicity, we will focus our discussion on time-synchrony and we now formalize this concept further. Let

$$\phi = \left(\phi^{(1)}(\vec{x}_1, t), \ldots, \phi^{(M)}(\vec{x}_M, t)\right), \vec{x}_m \in \mathbb{R}^{d_m}$$

be a multi-modal function whose modalities $\phi^{(m)}$, $m = 1, \ldots, M$ share a common temporal dimension $t \in \mathbb{R}$. A modality is temporally localized in the interval $\Delta \subset \mathbb{R}$ if $\phi^{(m)}(\vec{x}_m, t) = 0$, $\forall t \notin \Delta$. We will say that the modalities are synchronous whenever all $\phi^{(m)}$ are localized in the same time interval Δ.

Most natural signals exhibit characteristics that are time-invariant, meaning that they can occur at any instant in time. Think once again of an audio track : any particular frequency pattern can be repeated at arbitrary time instants. In order to account for this natural shift-invariance, we need to be able to shift patterns on modalities. Let ϕ be a multi-modal function localized in an interval centered on $t = 0$. The operator T_p shifts ϕ to time $p \in \mathbb{R}$ in a straightforward way :

$$T_p\phi = \left(\phi^{(1)}(\vec{x}_1, t-p), \ldots, \phi^{(M)}(\vec{x}_M, t-p)\right). \tag{6.2}$$

This temporal translation is homogeneous across channels and thus preserves synchrony. With these definitions, it becomes easy to express a signal as a superposition of synchronous multi-modal patterns ϕ_k, $k \in J$ occurring at various time instants $t_1, \ldots, t_k$:

$$s \approx \sum_{k \in J} c_k T_{t_k} \phi_k \,,$$

where the sum and weighting coefficients are understood as in (6.1). We often build a large subset of a dictionary by applying such synchronous translations to a single multi-modal function. In that case, we will often refer to this function as a *generating function* and we will indicate it with g_k.

In complex situations, it is sometimes difficult to manually design effective dictionaries because there is no good *a priori* knowledge about the generating functions g. In these cases, one typically

would want to learn a good dictionary from training data. Successful algorithms to learn dictionaries of basis functions have been proposed in the last years and applied to diverse classes of signal, including audio data [2, 60, 67], natural images [10, 25, 60, 63, 66, 85] and video sequences [84]. In the next section, we propose a learning strategy adapted to synchronous multi-modal signals.

6.2 Learning Multi-Modal Dictionaries

Our goal is to design an algorithm capable of learning sets of multi-modal synchronous functions adapted to particular classes of multi-modal signals. However, the design of an algorithm for learning dictionaries of multi-modal atoms is non-trivial and an extended literature survey showed that it has never been attempted so far. Two major challenges have to be considered :

- Learning algorithms are inherently time and memory consuming. When considering sets of multi-modal signals that involve huge arrays of data, the computational complexity of the algorithm becomes a challenging issue;
- Natural multi-modal signals often exhibit complex underlying structures that are difficult to explicitly define. Moreover, modalities have heterogeneous dimensions, which makes them complicated to handle. Audiovisual signals perfectly illustrate this challenge : the audio track is a 1D signal typically sampled at high frequency rate ($\mathcal{O}(10^4)$ samples/sec), while the video clip is a 3D signal sampled with considerably lower temporal resolution ($\mathcal{O}(10^1)$ frames/sec).

We will design a novel learning algorithm that captures the underlying structures of multi-modal signals overcoming both of these difficulties. We propose to learn *synchronous multi-modal generating functions* as introduced in the previous section using a generalization of the MoTIF algorithm [60]. Each such function defines a set of atoms corresponding to all its translations. This is notably motivated by the fact that natural signals typically exhibit statistical properties invariant to translation, and the use of generating functions allows to build huge dictionaries while using only few parameters. In order to make the computation feasible, the proposed algorithm learns the generating functions by alternatively localizing and learning interesting signal structures on the different signal components. As detailed in the following, this allows moreover to enforce synchrony between modal structures in an easy and intuitive fashion. Generating functions are learned successively and the procedure can be stopped when a sufficient number of atoms have been found. A constraint that imposes low correlation between the learned waveforms is also considered, such that no function is picked several times.

The goal of the learning algorithm is to build a set $\mathcal{G} = \{g_k\}_{k=1}^{K}$ of multi-modal generating functions g_k such that a very redundant dictionary $\mathcal{D}$ adapted to a class of signals can be created by applying all possible translations to the generating functions of $\mathcal{G}$. The function g_k can consist of an arbitrary number M of modalities. For simplicity, we will treat here the bimodal case $M = 2$; however, the extension to $M > 2$ is straightforward. To make it more concrete, we will write a bimodal function as $g_k = (g_k^{(a)}, g_k^{(v)})$ where one can think of $g_k^{(a)}$ as an audio modality and $g_k^{(v)}$ as a video modality of audiovisual data. More generally, the components do not have to be homogeneous in dimensionality; however, they have to share a common temporal dimension.

For the rest of the chapter, we denote discrete signals of infinite size by lower case letters. Real-world finite signals are made infinite by padding their borders with zeros. Finite size vectors and matrices are denoted with bold characters. We need to define the time-discrete version $\mathcal{T}_p$, $p \in \mathbb{R}$ of the synchronous translation operator (6.2). Since different modalities are in general sampled at different rates over time, the operator $\mathcal{T}_p$ must shift the signals on the two modalities by a

different integer number of samples, in order to preserve their temporal proximity. We define it as $\mathcal{T}_p = (\mathcal{T}_p^{(a)}, \mathcal{T}_p^{(v)}) := (T_{q^{(a)}}, T_{q^{(v)}})$, where $T_{q^{(a)}}$ translates an infinite (audio) signal by $q^{(a)} \in \mathbb{Z}$ samples and $T_{q^{(v)}}$ translates an infinite (video) signal by $q^{(v)}$ samples. In the experiments that we will conduct at the end of the chapter, typical values of the sampling rates are $\nu^{(a)} = 1/8000$ for audio signals sampled at 8 kHz and $\nu^{(v)} = 1/29.97$ for videos at 29.97 frames per second. Therefore the discrete-time version of the synchronous translation operator $\mathcal{T}_p$ with translation $p \in \mathbb{R}$ is defined with discrete translations $q^{(a)} := \text{nint}(p/\nu^{(a)}) \in \mathbb{Z}$ and $q^{(v)} := \text{nint}(p/\nu^{(v)}) \in \mathbb{Z}$ where $\text{nint}(\cdot)$ is the nearest integer function. Without loss of generality we may assume that $\nu^{(v)} \geq \nu^{(a)}$ and define a *re-sampling factor* $\text{RF} = \nu^{(v)}/\nu^{(a)}$.

For a given generating function g_k, the set $\{\mathcal{T}_p g_k\}_{p\in\mathbb{R}}$ contains all possible atoms generated by applying the translation operator to g_k. The dictionary generated by $\mathcal{G}$ is then $\mathcal{D} = \{\{\mathcal{T}_p g_k\}_p, k = 1, \ldots, K\}$. Learning is performed using a training set of N bimodal signals $\{(f_n^{(a)}, f_n^{(v)})\}_{n=1}^N$, where $f_n^{(a)}$ and $f_n^{(v)}$ are the components of the signal on the two modalities. The signals are assumed to be of infinite size but they are non zero only on their support of size $(S_f^{(a)}, S_f^{(v)})$. Similarly, the size of the support of the generating functions to learn is $(S_g^{(a)}, S_g^{(v)})$ such that $S_g^{(a)} < S_f^{(a)}$ and $S_g^{(v)} < S_f^{(v)}$. The proposed algorithm iteratively learns translation invariant filters. For the first one, the aim is to find $g_1 = (g_1^{(a)}, g_1^{(v)})$ such that the dictionary $\{(\mathcal{T}_p^{(a)} g_1^{(a)}, \mathcal{T}_p^{(v)} g_1^{(v)})\}_p$ is the most correlated in mean with the signals in the training set. Hence, it is equivalent to the following optimization problem :

$$\text{UP} : g_1 = \underset{\|g^{(a)}\|_2 = \|g^{(v)}\|_2 = 1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} \sum_{m} |\, \langle f_n^{(m)}, \mathcal{T}_{p_n}^{(m)} g^{(m)} \rangle \,|^2 \,, \tag{6.3}$$

which has to be solved simultaneously for the two modalities ($m = a, v$), i.e. we want to find a pair of synchronous filters $(g^{(a)}, g^{(v)})$ that minimize (6.3). There are two main differences with respect to classical learning methods, which make the present problem extremely challenging. First of all, we do not only want the learned function g_1 to represent well in average the training set (as expressed by the first maximization over g), but we want g_1 to be the best representing function up to an arbitrary time-translation on each training signal (as indicated by the second maximization over p_n) in order to achieve shift-invariance. In addition, we require these characteristics to hold for both modalities simultaneously, which implies an additional constraint on the synchrony of the couple of functions $(g_1^{(a)}, g_1^{(v)})$. Note that solving problem UP requires to compute simultaneous correlations across channels. In the audio-visual case, the dimension of the video channel makes this numerically prohibitive. To avoid this problem, we first solve UP restricted to the audio channel :

$$\text{UP}' : g_1^{(m)} = \underset{\|g^{(m)}\|_2 = 1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} |\, \langle f_n^{(m)}, \mathcal{T}_{p_n}^{(m)} g^{(m)} \rangle \,|^2 \,, \tag{6.4}$$

where $m = a$. We can then solve (6.4) for $m = v$ but limit the search for best translations around the time-shifts already obtained on the audio channel, thus avoiding the burden of long correlations between video streams.

For learning the successive generating functions, the problem can be slightly modified to include a constraint penalizing a generating function if a similar one has already been found. Assuming that $k-1$ generating functions have been learnt, the optimization problem to find g_k can be written as :

$$\text{CP} : g_k^{(m)} = \underset{\|g^{(m)}\|_2 = 1}{\arg\max} \frac{\sum_{n=1}^{N} \max_{p_n} |\, \langle f_n^{(m)}, \mathcal{T}_{p_n}^{(m)} g^{(m)} \rangle \,|^2}{\sum_{l=1}^{k-1} \sum_{q\in\mathbb{Z}} |\, \langle g_l^{(m)}, T_q g^{(m)} \rangle \,|^2} \,, \tag{6.5}$$

which again has to be solved simultaneously for the two modalities ($m = a, v$). In this case the optimization problem is similar to the unconstrained one in (6.4), with the only difference that a de-correlation constraint between the actual function $g_k^{(m)}$ and the previously learned ones is added. The

constraint is introduced as a term at the denominator that accounts for the correlation between the previously learned generating functions (the first summation over l) and the actual target function shifted at all possible positions (the second sum over q). By maximizing the fraction in (6.5) with respect to g, the algorithm has to find a balance between the goodness of the representation of the training set, which has to be maximized being expressed by the numerator, and the correlation between g_k and g_l ($l = 1, \ldots, k-1$), which has at the same time to be minimized, being represented by the denominator.

Finding the best solution to the unconstrained problem (UP$'$) or the constrained problem (CP) is indeed hard. However, the problem can be split into several simpler steps following a *localize and learn* paradigm [60]. Such a strategy is particularly suitable for this scenario, since we want to learn synchronous patterns that are localized in time and that represent well the signals. Thus, we propose to perform the learning by iteratively solving the following four steps :

1. ***Localize :*** for a given generating function $g_k^{(a)}[j-1]$ at iteration j, find the best translations $p_n^{(a)}[j] := \nu^{(a)} \cdot q_n^{(a)}[j]$ with
$$q_n^{(a)}[j] := \arg\max_{q \in \mathbb{Z}} \mid \langle f_n^{(a)}, T_q g_k^{(a)}[j-1] \rangle \mid ;$$
2. ***Learn :*** update $g_k^{(v)}[j]$ by solving UP$'$ (6.4) or CP (6.5) only for modality (v), with the translations fixed to the values $p_n = p_n^{(a)}[j]$ found at step 1, i.e. $q_n^{(v)} := \mathrm{nint}(\mathrm{RF} \times q_n^{(a)}[j])$;
3. ***Localize :*** find the best translations $p_n^{(v)}[j] := \nu^{(v)} \cdot q_n^{(v)}[j]$ using the function $g_k^{(v)}[j]$;
$$q_n^{(v)}[j] := \arg\max_{q \in \mathbb{Z}} \mid \langle f_n^{(v)}, T_q g_k^{(v)}[j] \rangle \mid ;$$
4. ***Learn :*** update $g_k^{(a)}[j]$ by solving UP$'$ (6.4) or CP (6.5) only for modality (a), with the translations fixed to the values $p_n = p_n^{(v)}[j]$ found at step 3 i.e. using $q_n^{(a)} = \mathrm{nint}(q_n^{(v)}[j]/\mathrm{RF})$.

A schematic representation of the four steps of the multi-modal learning algorithm is sketched in Fig. 6.1. The first and third steps consist in finding the location of the maximum correlation between one modality of each training signal $f_n^{(m)}$ and the corresponding generating function $g^{(m)}$. The temporal synchrony between generating functions on the two modalities is enforced at the learning steps (2 and 4), where the optimal translations p_n found for one modality are also kept for the other one.

We now consider in detail the second and fourth steps. We define $\mathbf{g}_k^{(m)} \in \mathbb{R}^{S_g^{(m)}}$ the restriction of the infinite size signal $g_k^{(m)}$ to its support. We will use the easily checked fact that for any translation p, any signal $f^{(m)}$ and any filter $g^{(m)}$ we have the equality $\langle f^{(m)}, \mathcal{T}_p^{(m)} g^{(m)} \rangle = \langle \mathcal{T}_{-p}^{(m)} f^{(m)}, g^{(m)} \rangle$, in other words the adjoint of the discrete translation operator $\mathcal{T}_p^{(m)}$ is $\mathcal{T}_{-p}^{(m)}$. Let $\mathbf{F}^{(m)}[j]$ be the matrix (with $S_f^{(m)}$ rows and N columns), whose columns are made of the signals $f_n^{(m)}$ shifted by $-p_n[j]$. More precisely, the n-th column of $\mathbf{F}^{(m)}[j]$ is $\mathbf{f}_{n,-p_n[j]}^{(m)}$, the restriction of $\mathcal{T}_{-p_n[j]}^{(m)} f_n^{(m)}$ to the support of $g_k^{(m)}$, of size $S_g^{(m)}$. We also denote

$$\mathbf{A}^{(m)}[j] = \mathbf{F}^{(m)}[j] \cdot \mathbf{F}^{(m)}[j]^T ,$$

where $\cdot^T$ indicates the transposition.

With these notations, the second step (respectively fourth step) of the *unconstrained* problem can be written as :

$$\mathbf{g}_k^{(m)}[j] = \arg\max_{\|\mathbf{g}^{(m)}\|_2 = 1} {\mathbf{g}^{(m)}}^T \mathbf{A}^{(m)}[j]\, \mathbf{g}^{(m)} . \tag{6.6}$$

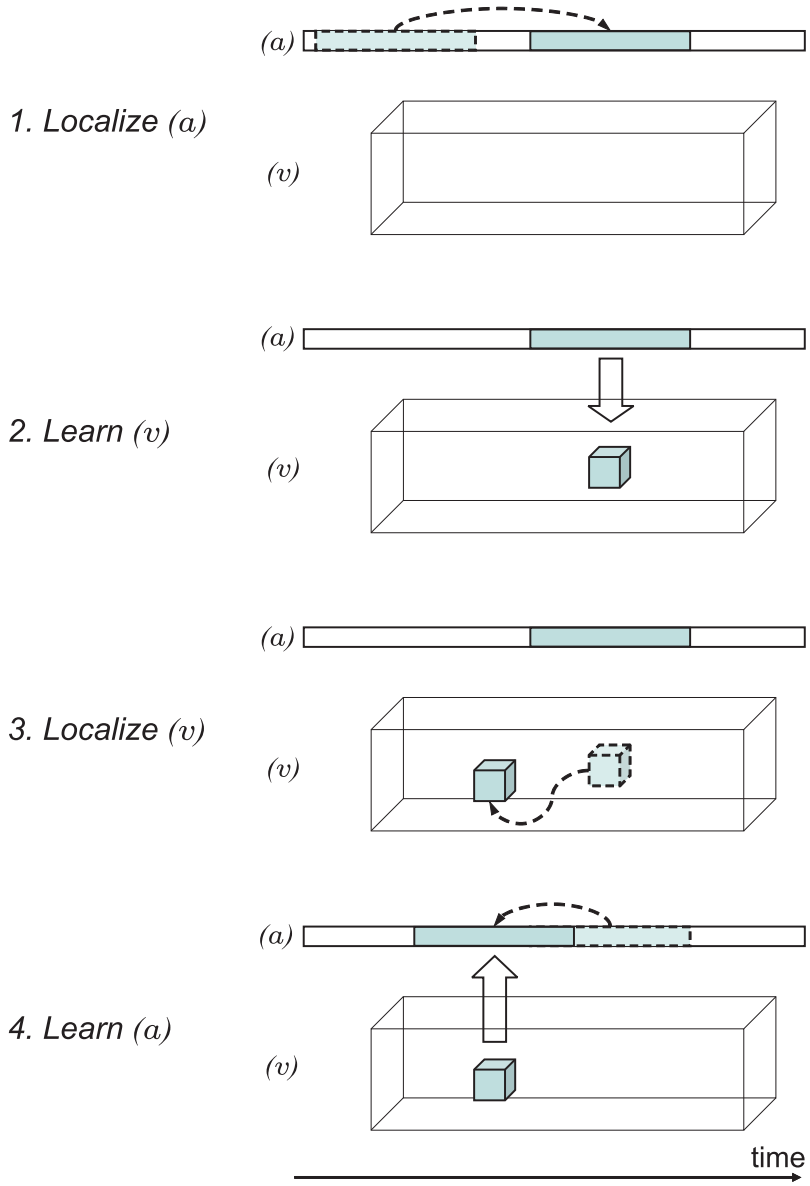


Figure 6.1 – *Schematic representation of the multi-modal learning algorithm.* Step 1 *: using the available generating function for modality* (a)*, find the best translations in* (a)*.* Step 2 *: using the found translations on* (a)*, update the generating function in* (v)*.* Step 3 *: using this generating function, find the best translations for modality* (v)*.* Step 4 *: using the translations found in modality* (v)*, update the generating function in* (a)*.*

with $m = v$ (respectively $m = a$). The best generating function $\mathbf{g}_k^{(m)}[j]$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}^{(m)}[j]$. Let us underline that in this case it is possible to easily solve the learning problem because of the particular form of the function to optimize. In fact, it is only because the objective function in (6.4) can be expressed as the quadratic form (6.6), given the translations p_n, that it is possible to turn the learning problem into an eigenvector problem.

For the *constrained* problem, we want to force $g_k^{(m)}[j]$ to be as de-correlated as possible from all the atoms in $\mathcal{D}_{k-1}$. This corresponds to minimizing

$$\sum_{l=1}^{k-1} \sum_{q \in \mathbb{Z}} | \langle T_{-q} g_l^{(m)}, g^{(m)} \rangle |^2 \tag{6.7}$$

or, denoting

$$\mathbf{B}_k^{(m)} = \sum_{l=1}^{k-1} \sum_{q \in \mathbb{Z}} \mathbf{g}_{l,-q}^{(m)} \, \mathbf{g}_{l,-q}^{(m)\,T}, \tag{6.8}$$

to minimizing $\mathbf{g}^{(i)T} \mathbf{B}_k^{(m)} \mathbf{g}^{(m)}$. With these notations, the constrained problem can be written as :

$$\mathbf{g}_k^{(m)}[j] = \underset{\|\mathbf{g}^{(m)}\|_2 = 1}{\arg\max} \frac{\mathbf{g}^{(m)T} \mathbf{A}^{(m)}[j] \, \mathbf{g}^{(m)}}{\mathbf{g}^{(m)T} \mathbf{B}_k^{(m)} \mathbf{g}^{(m)}}. \tag{6.9}$$

The best generating function $\mathbf{g}_k^{(m)}[j]$ is the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem defined in (6.9). Defining $\mathbf{B}_1^{(m)} = \mathbf{Id}$, we can use CP for learning the first generating function $\mathbf{g}_1$. Note again that the complex learning problem in (6.5) can be solved as the generalized eigenvector problem (6.9) because of the particular quadratic form imposed to the objective function to optimize, when the translations p_n are fixed.

The proposed multi-modal learning algorithm is summarized in **Algorithm 1**.

Algorithm 1 Principle of the multi-modal learning algorithm

$k = 0$, training set $\{(f_n^{(a)}, f_n^{(v)})\}$;
for $k = 1$ to K **do**
  $j \leftarrow 0$;
  random initialization of $\{(g_k^{(a)}[j], g_k^{(v)}[j])\}$;
  compute constraint matrices $\mathbf{B}_k^{(a)}$ and $\mathbf{B}_k^{(v)}$ as in (6.8);
  while no convergence reached **do**
    $j \leftarrow j + 1$;
    localize in modality (a):
    for each $f_n^{(a)}$, find the translation $p_n^{(a)}[j] \leftarrow \nu^{(a)} \cdot \arg\max_q | \langle f_n^{(a)}, T_q g^{(a)}[j-1] \rangle |$, maximally correlating $f_n^{(a)}$ and $g^{(a)}[j-1]$;
    learn modality (v):
    set $\mathbf{A}^{(v)}[j] \leftarrow \sum_{n=1}^{N} \mathbf{f}^{(v)}_{n,-p_n^{(a)}[j]} \mathbf{f}^{(v)\,T}_{n,-p_n^{(a)}[j]}$;
    find $\mathbf{g}_k^{(v)}[j]$, the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem $\mathbf{A}^{(v)}[j] \, \mathbf{g} = \lambda \mathbf{B}_k^{(v)} \mathbf{g}$, using (6.9);
    localize in modality (v):
    for each $f_n^{(v)}$, find the translation $p_n^{(v)}[j] \leftarrow \nu^{(v)} \cdot \arg\max_q | \langle f_n^{(v)}, T_q g^{(v)}[j] \rangle |$, maximally correlating $f_n^{(v)}$ and $g^{(v)}[j]$;
    learn modality (a):
    set $\mathbf{A}^{(a)}[j] \leftarrow \sum_{n=1}^{N} \mathbf{f}^{(a)}_{n,-p_n^{(v)}[j]} \mathbf{f}^{(a)\,T}_{n,-p_n^{(v)}[j]}$;
    find $\mathbf{g}_k^{(a)}[j]$, the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem $\mathbf{A}^{(a)}[j] \, \mathbf{g} = \lambda \mathbf{B}_k^{(a)} \mathbf{g}$, using (6.9);
  end while
end for

It is easy to demonstrate that the unconstrained single-modality algorithm converges in a finite number of iterations to a generating function locally maximizing the unconstrained problem [60]. It has been observed on numerous experiments that the constrained algorithm [60] and the multimodal constrained algorithm typically converge in few steps to a stable solution independently of the initialization.

6.3 Experiments

6.3.1 Audiovisual Dictionaries

The first experiment demonstrates the capability of the proposed learning algorithm to recover meaningful synchronous patterns from audiovisual signals. In this case the two modalities are audio and video, which share a common temporal axis, and the learned dictionaries are composed of generating functions $g_k = (g_k^{(a)}, g_k^{(v)})$, with $g_k^{(a)}$ and $g_k^{(v)}$ respectively audio and video component of g_k. Two joint audiovisual dictionaries are learned on two training sets. The first audiovisual dictionary, that we call *Dictionary 1* ($\mathcal{D}_1$), is learned on a set consisting of four audiovisual sequences representing the mouth of the same speaker uttering the digits from zero to nine in English. *Dictionary 2* ($\mathcal{D}_2$) is learned on a training set of four clips representing the mouth of four different persons pronouncing the digits from zero to nine in English. *Dictionary 1* should represent a collection of basis functions adapted to a particular speaker, while *Dictionary 2* aims at being a more "general" set of audio-video atoms.

For all sequences, the audio was recorded at 44 kHz and sub-sampled to 8 kHz, while the grayscale video was recorded at 29.97 fps and at a resolution of 70×110 pixels. The total length of the training sequences is 1060 video frames, i.e. approximately 35 seconds, for $\mathcal{D}_1$, and 1140 video frames, i.e. approximately 38 seconds, for $\mathcal{D}_2$. Note that the sampling frequencies along the time axis for the two modalities are different, thus when passing from one modality to the other a re-sampling factor RF equal to the ratio between the two frequencies has to be applied. In this case the value of the re-sampling factor is RF $= 8000/29.97 \approx 267$. Video sequences are filtered following the procedure suggested in [84], in order to speed up the training. The video component is thus "whitened" using a filter that equalizes the variance of the input sequences in all directions. Since the spatio-temporal amplitude spectrum of video signals roughly falls as $1/f$ along spatial and temporal axes [38, 85], whitening can be obtained applying a spherically symmetric filter $W(f) = f$ that produces an approximately flat amplitude spectrum at all spatio-temporal frequencies. The whitened sequences are then low-pass filtered to remove the high-frequency artifacts typical of digital video signals. We use a spherically symmetric low-pass filter $L(f) = e^{-(f/f_0)^4}$ with cut-off frequency f_0 at 80% of the Nyquist frequency in space and time. We thus end up with a filter $H(f) = W(f) \cdot L(f) = f \cdot e^{-(f/f_0)^4}$.

The learning is performed on audio-video patches $(f_n^{(a)}, f_n^{(v)})$ extracted from the original signals. The size of the audio patches $f_n^{(a)}$ is 6407 audio samples, while the size of the video patches $f_n^{(v)}$ is 31×31 pixels in space and 23 frames in time. We learn 20 generating functions g_k consisting of an audio component $g_k^{(a)}$ of 3204 samples and a video component $g_k^{(v)}$ of size 16×16 pixels in space and 12 frames in time. The 20 elements of $\mathcal{D}_2$ are shown in Fig. 6.2. The dictionary $\mathcal{D}_1$ has similar characteristics. The video component $g_k^{(v)}$ of each function is shown on the left, with time proceeding left to right, while the audio part $g_k^{(a)}$ is on the right, with time on the horizontal axis.

Concerning the video components, they are spatially localized and oriented edge detector functions that shift smoothly from frame to frame, describing typical movements of different parts of the mouth during the utterances. The audio parts of the generating functions contain almost all

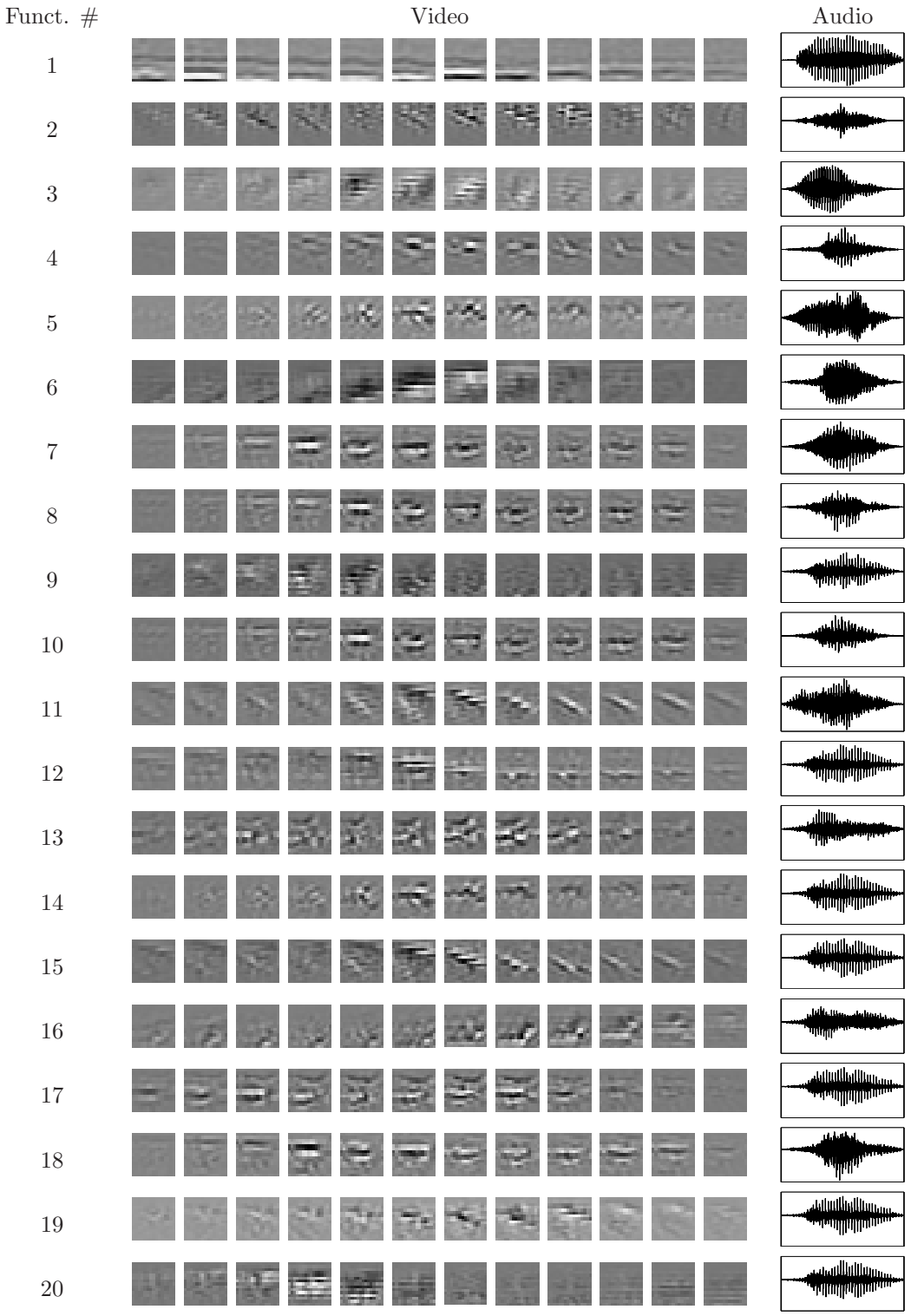


Figure 6.2 – *Audio-video generating functions of* Dictionary 2. *Twenty learned functions are shown, each consisting on an audio and a video component. Video components are on the left, with time proceeding left to right. Audio components are on the right, with time on the horizontal axis.*

the numbers present in the training sequences. In particular, when listening to the waveforms, one can distinguish the words *zero* (functions #11, #13, #16), *one* (#7, #9), *two* (#5, #6), *four* (#3), *five* (#1), *six* (#4), *seven* (#8, #18), *eight* (#10). Functions #12, #14, #15, #17, #19, #20 express the first two phonemes of the word *five* (i.e. /f/,/ay/), and they are also very similar to the word *nine* (i.e. /n/,/ay/). Typically, different instances of the same number have either different audio characteristics, like length or frequency content (e.g. compare audio functions #7 and #9), or different associated video components (e.g. functions #12, #14, #15, #17, #19, #20). As already observed in [60, 80], both components of generating function #2 are mainly high frequency due to the de-correlation constraint with the first atom.

The learning algorithm captures well high-level signal structures representing the synchronous presence of meaningful acoustic and visual patterns. All the learned multi-modal functions consist in couples of temporally close signals : a waveform expressing one digit when played, and a moving edge (horizontal, diagonal or curved) that follows the contour of the mouth during the utterances. The result is indeed interesting *per se*, considering that no prior on the shape of audio-video generating functions has been imposed.

6.3.2 Audiovisual Speaker Localization

In this experiment we want to test if the learned dictionaries are able to recover meaningful audiovisual patterns in real multimedia sequences. The dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$ are used to detect synchronous audio-video patterns revealing the presence of a meaningful event (the utterance of a sound) that we want to localize. We consider three test clips, Movie 1, Movie 2 and Movie 3, consisting in two persons placed in front of the camera arranged as in Fig. 6.3. One of the subjects is uttering digits in English, while the other one is mouthing *exactly the same words*. Test sequences consist in an audio track at 8 kHz and a video part at 29.97 fps and at a resolution of 480 $\times$ 720 pixels*. In all three sequences, the speaker is the same subject whose mouth was used to train $\mathcal{D}_1$; however, the training sequences are different from the test sequences. In contrast, none of the four speaking mouths used to train $\mathcal{D}_2$ belongs to the speaker in the test data set. We want to underline that the test sequences are particularly challenging to analyze, since both persons are mouthing the same words at the same time. The task of associating the sound with the "real" speaker is thus definitely non-trivial. The test clips used in this chapter can be downloaded through `http://lts2www.epfl.ch/~monaci/avlearn.html`.

With the experimental results that we will show in the following we want to demonstrate that :

- For both dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$, the positions of maximal projection between the dictionary atoms ϕ_k and the test sequences are localized on the actual location of the audiovisual source;
- The detection of the actual speaker using both $\mathcal{D}_1$ and $\mathcal{D}_2$ is robust to severe visual noise (the person mouthing the same words of the real speaker) as well as to acoustic noise. The mouth of the correct speaker is effectively localized also when strong acoustic noise (SNR=1dB) is summed to the audio track in the form of additive white gaussian noise or out-of-view talking people;
- The detection of the speaker's mouth is more robust and accurate using dictionary $\mathcal{D}_1$, which is adapted to the speaker, than using the general dictionary $\mathcal{D}_2$.

The audio tracks of the test clips are correlated with all time-shifted version of each audio component $g_k^{(a)}$ of the 20 learned generating functions g_k, which is efficiently done by filtering.

*Only the luminance component is considered, while the chromatic channels are discarded.

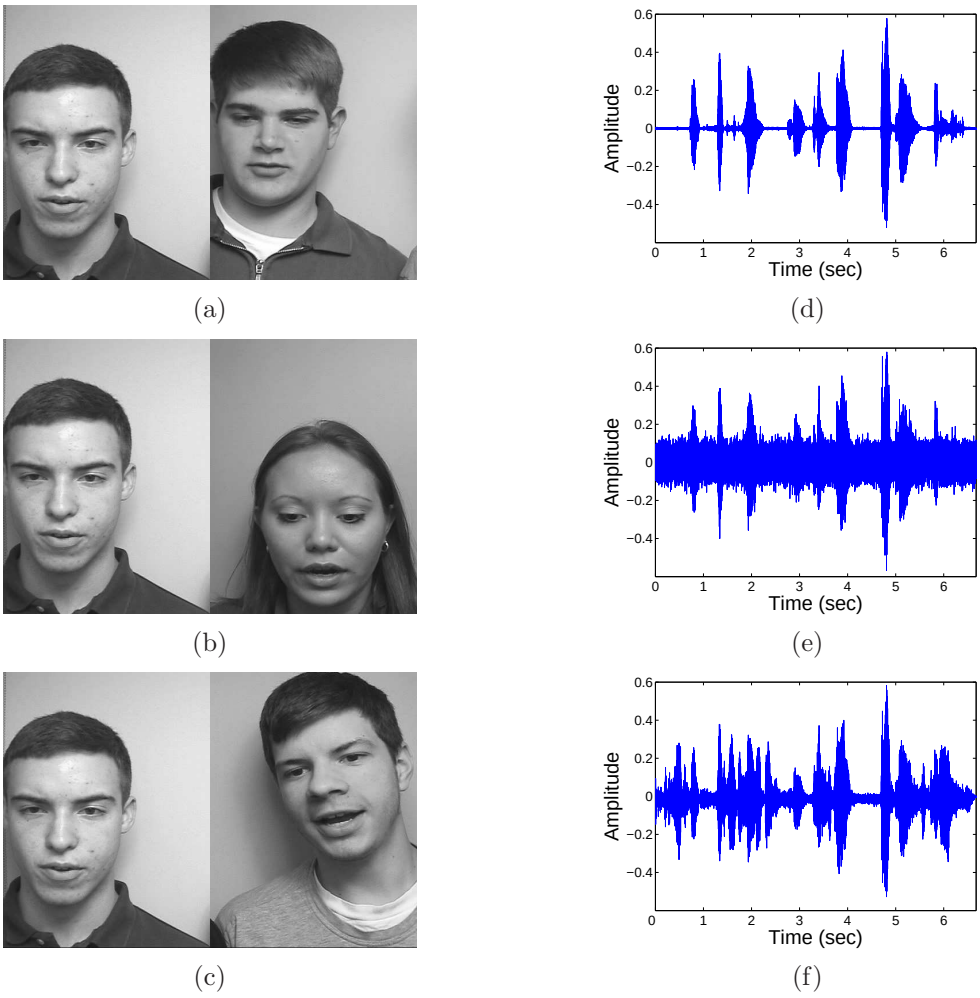


Figure 6.3 – *Test sequences. Sample frames of Movie 1 (a), Movie 2 (b) and Movie 3 (c) are shown on the left. The original audio track a (d), together with its noisy versions with additive gaussian noise a+AWGN (e) and added distracting speech and music a+speech (f) are plotted on the right. Test clips can be downloaded through* `http://lts2www.epfl.ch/~monaci/avlearn.html`.

For each audio function we find the the time position of maximum correlation, $\hat{p}_k^{(a)}$, and thus the audio atom $\phi_k^{(a)}$ with highest correlation. We consider a window of 31 frames around the time position in the video corresponding to $\hat{p}_k^{(a)}$, which is computed as $\tilde{p}_k^{(v)} = \text{nint}(\hat{p}_k^{(a)}/\text{RF})$. This restricted video patch consists of frames in the interval $[\tilde{p}_k^{(v)} - 15; \tilde{p}_k^{(v)} + 15]$ and we compute its correlation with all spatial and temporal shifts of the video component $g_k^{(v)}$ of g_k. The spatio-temporal position $\left(\widehat{\vec{x}}_k, \hat{p}_k^{(v)}\right)$ of maximum correlation between the restricted video patch and the learned video generating function yields the video atom $\phi_k^{(v)}$ with highest correlation. The positions of maximal projection of the learned atoms over the image plane $\widehat{\vec{x}}_k$, $k = 1, \ldots, 20$, are grouped into clusters using a hierarchical clustering algorithm*. The centroid of the cluster containing the largest number of points is kept as the estimated location of the sound source. We expect the estimated sound source position to be close to the speaker's mouth.

In Fig. 6.4 sample frames of the test sequences are shown. The white marker over each image

*The MATLAB function `clusterdata.m` was used. Clusters are formed when the distance between groups of points is larger than 50 pixels. According to several tests, the choice of the clustering threshold is non-critical.

Figure 6.4 – *Sample frames of Movie 1 [Left], Movie 2 [Center] and Movie 3 [Right]. The left person is the real speaker, the right subject mouths the same words pronounced by the speaker but his audio track has been removed. The white cross highlights the estimated position of the sound source, which is correctly placed over the speaker's mouth.*

indicates the estimated position of the sound source over the image plane, which coincides with the mouth of the actual speaker. The position of the mouth center of the correct speaker has been manually annotated for each test sequence. The sound source location is considered to be correctly detected if it falls in a circle of radius 100 pixels centered in the labelled mouth. The source position is correctly detected for all the tested sequences and using both dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$. Results are accurate when the original sound track a is used (signal in Fig. 6.3(d)), as well as when considerable acoustic noise (SNR=1dB) is present (signals a+AWGN and a+speech in Fig. 6.3(e-f)).

In order to assess the goodness of the estimation of the sound source position, a simple measure can be designed. We define the *reliability* of the source position estimation, r, as the ratio between the number of elements belonging to the biggest cluster, which is the one used to estimate the sound source location, and the total number of elements considered, N (i.e. the total number of functions used for the analysis of the sequence, in this case 20). The value of r ranges from $1/N$, when each point constitutes a one-element cluster, to 1, when all points belong to the same group. Clearly, if most of the maxima of the projections between the video basis functions and the sequence lie close to one another, and are thus clustered together, it is highly probable that such cluster indicates the real position of the sound source and the value of r is high in this case. On the other hand, if maxima locations are placed all over the image plane forming small clusters, even the biggest cluster will include a small fraction of the whole data. In this situation it seems reasonable to deduce that the estimated source position is less reliable, which is reflected by the value of r being smaller in this case.

As we have already observed, for all the test sequences the sound source position is correctly localized. Moreover, it is interesting to remark that in all cases, the detection of the speaker's mouth is more *reliable* using dictionary $\mathcal{D}_1$, which is adapted to the speaker, than using the general dictionary $\mathcal{D}_2$. An example of the described situation is depicted in Fig. 6.5. The images show sample frames of Movie 3. The positions of maximal projection between video functions belonging to dictionaries $\mathcal{D}_1$ (Left) and $\mathcal{D}_2$ (Right) and the test sequence are plotted on the image plane. Points belonging to the same cluster are indicated with the same marker. The centroid of the biggest cluster is indicated by the white cross, which is correctly placed over the speaker's mouth. In both cases *Cluster 1* is the group containing the largest number of points and it is thus the one used to estimate the sound source position. When using dictionary $\mathcal{D}_1$ (Left), the biggest cluster has 17 elements and thus the reliability of the source position is $r = 17/20 = 0.85$, while when using $\mathcal{D}_2$ (Right), the biggest cluster groups only 13 points and the reliability equals $r = 13/20 = 0.65$. This behavior is indeed interesting, since it suggests that the learning algorithm actually succeeds in its task. The algorithm appears to be able to learn general meaningful synchronous patterns in the data. Moreover, the fact that more reliable localization results are achieved using the dictionary

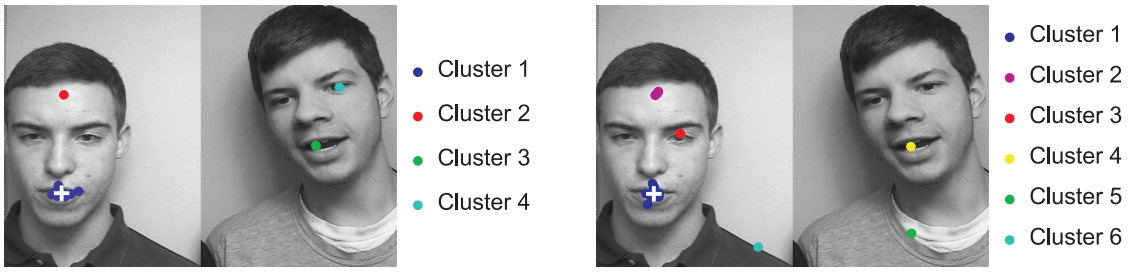


Figure 6.5 – *Sample frames of Movie 3. The positions of maximal projection between video functions and test sequence are plotted on the image plane. Points belonging to the same cluster are indicated with the same marker. The white cross indicates the centroid of the biggest cluster, that in both cases is Cluster 1; it contains 17 elements when $\mathcal{D}_1$ is used [Left] and 13 when $\mathcal{D}_2$ is used [Right].*

adapted to the speaker ($\mathcal{D}_1$) suggests that the proposed method allows to capture important signal structures typical of the considered training set.

At this point it is interesting to compare the localization performances achieved using the learned dictionaries with those obtained by the audiovisual gestalts detection method presented in Chapter 3. The interest of such a comparison is twofold. First, the cross-modal localization algorithm introduced in Chapter 3 relies on signal representation techniques that model *separately* audio and video modalities using sparse decompositions over *general* dictionaries of Gabor and edge-like functions respectively. This comparison is the occasion to check if a modelling of cross-modal correlations done at a level that is closer to the signals themselves (the model proposed here) than to the features (the model presented in Chapter 3) is advantageous or not. Second, the audiovisual gestalts localization algorithm exhibits state-of-the-art performances on the CUAVE audiovisual speech corpus [88], as discussed in [78, 79] and Chapter 3. The comparison thus is significant *per se*.

The test movie clips have thus been resized to a resolution of 120×176 pixels to be more quickly processed, and they have been decomposed using 50 video atoms with the 3D-MP algorithm. The audio tracks have been represented using 1000 Gabor atoms with MP. Audio-video features are extracted and meaningful gestalts are detected as described in Chapter 3. It is worth underlining that since a single speaker is assumed to be present in the sequence, audiovisual gestalts are built considering the entire movie (i.e. no sliding analysis window is used, see section 3.7). Mouth positions have been manually labelled in these resized clips as well and the region of correct source detection is defined as a circle of diameter 25 pixels centered in the "real" mouth.

Table 6.1 summarizes the experimental results for all tested sequences and both localization methods (denoted as *learning* and *gestalts*). The first column indicates the video clip used, the second one the audio track used and the third one the dictionary employed for the analysis. The fourth column shows the source localization result using the learned dictionaries and the fifth column indicates the reliability r of the localization. In all cases the audio source position is correctly found on the image plane, as indicated by the green ticks. Finally, the sixth column reports the localization results for the audiovisual gestalt detection method presented in Chapter 3. In this case the speaker's mouth is erroneously detected on four out of nine clips (red crosses).

These results highlight that detecting the learned multi-modal atoms, it is possible to effectively localize audiovisual sources in challenging real-world sequences. The algorithm proposed here outperforms the localization method presented in Chapter 3, which is more general (no specific assumption on the type of sequences is made and no training is required) but less robust to audio and video distractors. The audiovisual gestalt model relies on the assumption that in general audio-video synchronous events occur randomly, except if a meaningful audiovisual source is observed.

Video	Audio	Dict.	Localization – *learning*	r	Localization – *gestalts*
Movie 1	a	$\mathcal{D}_1$	✓	0.65	✓
		$\mathcal{D}_2$	✓	0.50	
	a+AWGN	$\mathcal{D}_1$	✓	0.65	×
		$\mathcal{D}_2$	✓	0.50	
	a+speech	$\mathcal{D}_1$	✓	0.65	✓
		$\mathcal{D}_2$	✓	0.50	
Movie 2	a	$\mathcal{D}_1$	✓	0.90	✓
		$\mathcal{D}_2$	✓	0.60	
	a+AWGN	$\mathcal{D}_1$	✓	0.90	✓
		$\mathcal{D}_2$	✓	0.65	
	a+speech	$\mathcal{D}_1$	✓	0.85	✓
		$\mathcal{D}_2$	✓	0.60	
Movie 3	a	$\mathcal{D}_1$	✓	0.85	×
		$\mathcal{D}_2$	✓	0.65	
	a+AWGN	$\mathcal{D}_1$	✓	0.80	×
		$\mathcal{D}_2$	✓	0.65	
	a+speech	$\mathcal{D}_1$	✓	0.85	×
		$\mathcal{D}_2$	✓	0.70	

Table 6.1 – *Summary of the source localization results for all the tested sequences. Green ticks indicate that the source is correctly localized while the red crosses denote a localization error. In all cases, using the learned dictionaries the audio source position is correctly determined on the image plane (fourth column). Employing the speaker-adapted dictionary $\mathcal{D}_1$ the localization results to be more reliable than using $\mathcal{D}_2$, as indicated by the values of r in the fifth column. The gestalt detection method in contrast fails in localizing the speaker's mouth on four out of nine clips (last column).*

The test sequences employed in this chapter do not satisfy this hypothesis : in this case in fact visual distractors exhibit some strong correlation with the audio signal since the characters on the right in the test clips utter the same words pronounced by the real speaker. However, it is worth underlining that for all the sequences the video atom exhibiting the highest degree of correlation with the audio signal (according to the synchrony criterion formulated in Chapter 3) is localized around the correct speaker's mouth. Errors are caused by several other video structures exhibiting similar high correlations with the audio and positioned on the "fake" speaker's mouth. The localization method proposed here overcomes these difficulties exploiting the temporal proximity between adapted audio and video patterns.

6.4 Discussion

In this chapter we have introduced a model for multi-modal signals that enforces sparsity and synchrony between modalities by making use of multi-component heterogeneous basis functions. We have proposed as well a new method to learn dictionaries of translation invariant multi-modal functions adapted to a class of multi-component signals. Generating functions are iteratively found using a *localize and learn* paradigm which enforces temporal synchrony between modalities. Thanks to the particular formulation of the objective function, the learning problem can be turned into a generalized eigenvector problem, which makes the algorithm fast and free of parameters to tune. A constraint in the objective function forces the learned waveforms to have low correlation, such that no function is picked several times. The main drawback of this method is that the few generating

functions following the first one are mainly due to the de-correlation constraint, more than to the correspondence with the signal. Despite that, the algorithm seems to capture well the underlying structures in the data. The learned functions have been used to analyze complex multi-modal sequences, obtaining encouraging results in localizing the sound sources in the test sequences.

7 Conclusion

7.1 Discussed Topics and Achievements

As it was stated in the introductory chapter, and as the title of the dissertation says, the main objective of this research work is to try to understand and model the complex relationships existing between multi-modal signals. In this sense we propose to adopt sparse signal representations based on redundant dictionaries of functions that allow to express the data in terms of few, relevant signal structures that can be easily and intuitively analyzed. Although some of the developed methodologies are completely general, in this thesis we target applications in the field of audiovisual signal fusion. Audiovisual data are in fact used as a paradigm of multi-modal signals and they represent the main application area for this research.

In order to put this thesis in the proper context, in Chapter 2 we propose a classification of multi-modal signals based on their characteristics. We have thus the opportunity to highlight that this work deals with complex heterogeneous multi-channel signals that exhibit correlations along time, just like audio-video sequences do. In the same chapter we present as well a literature survey in the main fields considered in this dissertation, that are audiovisual source localization and separation. This review highlights one main drawback of existing studies that motivates the proposed approach : the lack of structural modelling of multi-modal signals and of the correlations between modalities.

One of the major contributions of this thesis is the definition of a simple and intuitive model of audiovisual signals that explains the relationships between audio and video modalities as a co-occurrence of synchronous signal patterns, termed in Chapter 3 *audiovisual gestalts*. The definition of these patterns is possible because audio and video signals are concisely represented with salient structures that describe physically-related quantities like moving edges and audio atoms. The detection of multi-modal gestalts allows to localize and extract correlated audio-video structures, showing the effectiveness of the proposed approach. The core of the audiovisual localization method is the video representation technique. In Chapter 4 we analyze this issue more in details and we propose a new framework for the tracking of visual structures based on Particle Filtering, which ensures robustness and flexibility.

Structural properties of audiovisual signals are exploited as well in Chapter 5 to design a blind

audiovisual source separation algorithm. Up to our knowledge this is the first attempt to link two very different research fields, audiovisual signal fusion and single-channel blind source separation. We show in this chapter that the tools developed in the context of multi-modal source localization can be extremely useful also to challenge the source separation problem. Exploiting the coherence between acoustic and visual structures, audio-video sources are detected, localized and extracted. The localization of the sources on the image sequence results accurate and robust in challenging scenarios, as well as their temporal localization on the audio domain. Problems arise in the separation of audio sources in mixtures. The single-channel source separation problem is challenging and the simple static approach used can achieve only limited performances. However, the proposed framework seems to be appropriate to further develop the system, as we will discuss in the next section.

Finally, in Chapter 6 we have reconsidered the multi-modal signal model introduced in Chapter 3 in the light of the experience accumulated throughout the thesis. Instead of projecting each modality on a distinct dictionary, we propose to model a multi-modal signal as a sparse sum of recurrent synchronous multi-modal structures, using thus intrinsically multi-modal dictionaries. Since the manual design of such dictionaries results extremely complex, we propose an algorithm capable of learning multi-modal synchronous functions from training patches. Applied to audiovisual sequences, the algorithm demonstrates its capability of capturing real multi-modal patterns present in the data. Moreover, the detection of such patterns in challenging audiovisual sequences allows to localize the actual sound source on the video.

To conclude, we believe that it is extremely important to consider models of multi-modal signals that take into account the structural properties of the data. The experimental results presented in this dissertation point out that such an approach can be very advantageous and that it offers promising, interesting potentialities. Besides that, the information more or less "hidden" into the data is represented in a concise, meaningful and intuitive fashion, making it more accessible and easy to handle and analyze.

7.2 Future Research Directions

There are many different research directions future work can take. One issue that offers interesting possibilities for further developments is the design of dictionaries of multi-modal functions. The learning algorithm builds collections of shift-invariant generating functions : one straightforward extension, based on the properties of the inner product, is to add invariance to other transformations that admit a well defined adjoint, for example invariance to rotations, that can be particularly useful for image or video representation. Moreover, the presented algorithm learns generating functions iteratively, which makes the method few computationally and time consuming. However, as already underlined, a constraint has to be added to force de-correlation between the learned waveforms. This constraint introduces distortions that are evident especially in the second, third functions built. It would be interesting to remove the de-correlation penalization by learning a whole dictionary at once, for example using the K-SVD algorithm [3]. A useful test for the proposed algorithm would be its application to other types of multi-modal signals. We have in mind for example EEG-fMRI data : a couple of 1D-4D signals with strong correlations between patterns in the two modalities.

As already underlined in the text, the tracker of video structures based on Particle Filtering can be further developed in order to account for interactions between video structures. A convenient approach can be the one presented in [65], that could naturally extend the atom tracking method to a multi-atom tracking algorithm.

Finally, one topic that seems very appealing is Blind Audiovisual Source Separation. The

methodology presented in this thesis is very effective in localizing audio-video sources in space and time. In contrast, the de-mixing capabilities of the simple audio separation model used appear to be limited. However, the proposed algorithm can confidently provide several extremely valuable information : the association between audio and video sources and the detection of time periods during which audio sources are active alone. On these bases a natural extension to the presented method is the adoption of more sophisticated separation algorithms that can learn speakers characteristics in time slots during which they speak alone through HMM modelling [95, 99]. Another interesting option could be to track the evolution of acoustic features starting in time periods presenting a single speaker and then to continue the tracking in the mixtures, for example using the techniques developed in [96].

Bibliography

[1] (2003). Time-series econometrics: Cointegration and autoregressive conditional heteroskedasticity. *Advanced information on the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel.*

[2] S. Abdallah, M. Plumbley (2001). If edges are the independent components of natural images, what are the independent components of natural sounds? In *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation (ICA)*, pp. 534–539.

[3] M. Aharon, M. Elad, A. Bruckstein (2006). K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Proc.* **54**(11):4311–4322.

[4] T. W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, 2^{nd} edn.

[5] S. Arberet, R. Gribonval, F. Bimbot (2006). A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture. In *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation (ICA)*, pp. 536–543.

[6] M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. Signal Proc.* **50**(2):174–188.

[7] F. R. Bach, M. I. Jordan (2004). Blind one-microphone speech separation: A spectral learning approach. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 17, pp. 65–72.

[8] M. J. Beal, N. Jojic, H. Attias (2003). A graphical model for audiovisual object tracking. *IEEE Trans. Pattern Anal. Machine Intell.* **25**(7):828–836.

[9] A. Bell, T. Sejnowski (1995). An information-maximization approach to blind source separation and blind deconvolution. *Neural Computation* **7**:1129–1159.

[10] A. Bell, T. Sejnowski (1997). The "independent components" of natural scenes are edge filters. *Vision Research* **37**(23):3327–3338.

[11] P. Bertelson, M. Radeau (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics* **29**(6):578–584.

[12] P. Bertelson, J. Vroomen, G. Wiegeraad, B. deGelder (1994). Exploring the relation between McGurk interference and ventriloquism. In *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, pp. 559–562.

[13] P. Besson, M. Kunt, T. Butz, J.-P. Thiran (2005). A multimodal approach to extract optimized audio features for speaker detection. In *Proc. European Signal Proc. Conf. (EUSIPCO).*

[14] A. Bregman (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts.

[15] J.-P. Bresciani, F. Dammeier, M. Ernst (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision* **6**(5):554–564.

[16] G. J. Brown, M. P. Cooke (1994). Computational auditory scene analysis. *Computer Speech and Language* **8**(4):297–336.

[17] P. J. Burt, E. H. Adelson (1983). The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **31**(4):532–540.

[18] T. Butz, J.-P. Thiran (2005). From error probability to information theoretic (multi-modal) signal processing. *Signal Processing* **85**(5):875–902.

[19] F. Cao (2004). Application of the Gestalt principles to the detection of good continuations and corners in image level lines. *Computing and Visualization in Science* **7**:3–13.

[20] J.-F. Cardoso (1998). Blind signal separation: statistical principles. *Proc. IEEE* **90**(8):2009–2026.

[21] C. Carmona-Moreno et al. (2005). Characterizing inter-annual variations in global fire calendar using data from earth observing satellites. *Global Change Biology* **11**(9):1537–1555.

[22] S. S. Chen, D. L. Donoho, M. A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computing* **20**(1):33–61.

[23] D. Comaniciu, V. Ramesh, P. Meer (2003). Kernel-based object tracking. *IEEE Trans. Pattern Anal. Machine Intell.* **25**(5):564–577.

[24] E. Cosatto, J. Ostermann, H. Graf, J. Schroeter (2003). Lifelike talking faces for interactive services. *Proc. IEEE* **91**(9):1406–1429.

[25] S. Cotter, B. Rao (2002). Application of total least squares (TLS) to the design of sparse signal representation dictionaries. In *Proc. 36th Asilomar Conf. on Signals, Systems and Computers*, vol. 1, pp. 963–966.

[26] T. M. Cover, J. A. Thomas (1991). *Elements of information theory*. John Wiley & Sons, New York.

[27] R. Cutler, L. Davis (2000). Look who's talking: speaker detection using video and audio correlation. In *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, vol. 3, pp. 1589–1592.

[28] R. Dansereau (2004). Co-channel audiovisual speech separation using spectral matching constraints. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 645–648.

[29] I. Daubechies (1988). Time-frequency localization operators: A geometric phase space approach. *IEEE Trans. on Inform. Theory* **34**(4):605–612.

[30] G. Davis, S. Mallat, A. M. (1997). Adaptive greedy approximations. *Journal of Constructive Approximations* **13**(1):57–98.

[31] S. Deligne, G. Potamianos, C. Neti (2002). Audio-visual speech enhancement with AVCD-CN (Audiovisual Codebook Dependent Cepstral Normalization). In *Proc. Int. Conf. Spoken Language Proc. (ICSLP)*, pp. 1449–1452.

[32] A. Desolneux, L. Moisan, J.-M. Morel (2000). Meaningful alignments. *International Journal of Computer Vision* **40**(1):7–23.

[33] A. Desolneux, L. Moisan, J.-M. Morel (2001). Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision* **14**(3):271–284.

[34] A. Desolneux, L. Moisan, J.-M. Morel (2003). A grouping principle and four applications. *IEEE Trans. Pattern Anal. and Machine Intell.* **25**(4):508–513.

[35] O. Divorra Escoda (2005). *Toward Sparse and Geometry Adapted Video Approximations.* Ph.D. thesis, EPFL, Lausanne. [Online] Available: `http://lts2www.epfl.ch/`.

[36] O. Divorra Escoda, L. Granai, P. Vandergheynst (2006). On the use of a priori information for sparse signal approximations. *IEEE Trans. Signal Proc.* **54**(9):3468–3482.

[37] O. Divorra Escoda, P. Vandergheynst (2004). A Bayesian approach to video expansions on parametric over-complete 2-D dictionaries. In *Proc. IEEE Int. Workshop on Multimedia Signal Proc. (MMSP)*, pp. 490–493.

[38] D. Dong, J. Atick (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems* **6**(2):159–178.

[39] J. Driver (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* **381**(6577):66–68.

[40] G. Edwards, C. Taylor, T. Cootes (1998). Interpreting face images using active appearance models. In *Proc. 3^{rd} Int. Conf. on Automatic Face and Gesture Recognition*, pp. 300–305.

[41] I. R. Farah, M. B. Ahmed, M. R. Boussema (2003). Multispectral satellite image analysis based on the method of blind separation and fusion of sources. In *Proc. Int. Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 6, pp. 3638–3640.

[42] J. W. Fisher III, T. Darrell (2004). Speaker association with signal-level audiovisual fusion. *IEEE Trans. Multimedia* **6**(3):406–413.

[43] J. W. Fisher III, T. Darrell, W. T. Freeman, P. Viola (2000). Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 772–778.

[44] X. Gigandet et al. (2005). Region-based satellite image classification: Method and validation. In *Proc. Int. Conf. Image Proc. (ICIP)*, vol. 40, pp. 832–835.

[45] L. Girin, J.-L. Schwartz, G. Feng (2001). Audio-visual enhancement of speech in noise. *Journal of the Acoustical Society of America* **109**(6):3007–3020.

[46] R. Goecke, G. Potamianos, C. Neti (2002). Noisy audio feature enhancement using audio-visual speech data. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, pp. 2025–2028.

[47] I. F. Gorodnitsky, B. D. Rao (1997). Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Proc.* **45**(3):600–616.

[48] K. Green, P. Kuhl, A. Meltzoff, E. Stevens (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception & Psychophysics* **50**(6):524–536.

[49] R. Gribonval (2002). Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 3, pp. 3057–3060.

[50] R. Gribonval, E. Bacry, J. Abadia (2004). Matching Pursuit software and documentation. `http://www.cmap.polytechnique.fr/~bacry/LastWave/packages/mp/mp.html`.

[51] R. Gribonval et al. (1996). Analysis of sound signals with High Resolution Matching Pursuit. In *Proc. IEEE Int. Symposium on Time-Frequency and Time-Scale Analysis (TFTS)*, pp. 125–128.

[52] M. Gurban, J.-P. Thiran (2006). Multimodal speaker localization in a probabilistic framework. In *Proc. European Signal Proc. Conf. (EUSIPCO)*.

[53] F. Hayashi (2000). *Econometrics*. Princeton University Press, Princeton, NJ.

[54] J. Hershey, J. Movellan (1999). Audio-vision: Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 12, pp. 813–819.

[55] J. Hipwell et al. (2003). Intensity-based 2-D-3-D registration of cerebral angiograms. *IEEE Trans. Med. Imag.* **22**(11):1417–1426.

[56] A. Hyvärinen (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks* **10**(3):626–634.

[57] A. Hyvärinen, E. Oja (2000). Independent component analysis: Algorithms and applications. *Neural Networks* **13**(4–5):411–430.

[58] C. E. Jack, W. R. Thurlow (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills* **37**(3):967–979.

[59] G.-J. Jang, T.-W. Lee (2002). A probabilistic approach to single channel blind signal separation. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 15, pp. 1173–1180.

[60] P. Jost, P. Vandergheynst, S. Lesage, R. Gribonval (2006). MoTIF: an efficient algorithm for learning translation invariant dictionaries. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 857–860.

[61] G. Kanizsa (1980). *Grammatica del vedere. Saggi su percezione e gestalt*. Il Mulino, Bologna.

[62] E. Kidron, Y. Schechner, M. Elad (2007). Cross-modal localization via sparsity. *to appear in IEEE Trans. Signal Proc.* .

[63] K. Kreutz-Delgado et al. (2003). Dictionary learning algorithms for sparse representation. *Neural Computation* **15**(2):349–396.

[64] M. Lallouache (1991). *Un poste visage-parole couleur. Acquisition et traitement automatique des lèvres*. Ph.D. thesis, Institut National Polytechnique, Grenoble, France.

[65] O. Lanz (2006). Approximate bayesian multibody tracking. *IEEE Trans. Pattern Anal. Machine Intell.* **28**(9):1436–1449.

[66] M. Lewicki, B. Olshausen (1999). A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America* **16**(7):1587–1601.

[67] M. Lewicki, T. Sejnowski (2000). Learning overcomplete representations. *Neural computation* **12**(2):337–365.

[68] D. G. Lowe (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60**(2):91–110.

[69] B. Lucas, T. Kanade (1981). An iterative image registration technique with an application to stereo vision. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 674–679.

[70] S. Lucey, T. Chen, S. Sridharan, V. Chandran (2005). Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. *IEEE Trans. Multimedia* **7**(3):495–506.

[71] H. Lütkepohl, M. Krätzig (2004). *Applied Time Series Econometrics*. Cambridge University Press, Cambridge, UK.

[72] F. Maes et al. (1997). Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imag.* **16**(2):187–198.

[73] S. Mallat, Z. Zhang (1993). Matching Pursuits with time-frequency dictionaries. *IEEE Trans. Signal Proc.* **41**(12):3397–3415.

[74] E. Martínez-Montes et al. (2004). Concurrent EEG/fMRI analysis by multiway Partial Least Squares. *NeuroImage* **22**(3):1023–1034.

[75] H. McGurk, J. W. MacDonald (1976). Hearing lips and seeing voices. *Nature* **264**(5588):746–748.

[76] K. Mikolajczyk, C. Schmid (2005). A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Machine Intell.* **27**(10):1615–1630.

[77] G. Monaci, O. Divorra Escoda, P. Vandergheynst (2005). Analysis of multimodal signals using redundant representations. In *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, vol. 3, pp. 46–49.

[78] G. Monaci, O. Divorra Escoda, P. Vandergheynst (2006). Analysis of multimodal sequences using geometric video representations. *Signal Processing* **86**(12):3534–3548.

[79] G. Monaci, P. Vandergheynst (2006). Audiovisual gestalts. In *Proc. Computer Vision and Pattern Recognition (CVPR) Workshop on Perceptual Organization in Computer Vision*.

[80] G. Monaci et al. (2006). Learning multi-modal dictionaries: Application to audiovisual data. In *Proc. of Int. Workshop on Multimedia Content Representation, Classification and Security*, vol. 4105 of *Lecture Notes in Computer Science*, pp. 538–545.

[81] H. Nock, G. Iyengar, C. Neti (2002). Assessing face and speech consistency for monologue detection in video. In *Proc. 10^{th} ACM Int. Conf. on Multimedia*, pp. 303–306.

[82] H. J. Nock, G. Iyengar, C. Neti (2003). Speaker localisation using audio-visual synchrony: an empirical study. In *Proc. Int. Conf. on Image and Video Retrieval (CIVR)*, pp. 488–499.

[83] K. Nummiaro, E. Koller-Meier, L. Van Gool (2002). A color-based particle filter. In *Proc. 1^{st} Workshop on Generative-Model-Based Vision*, pp. 53–60.

[84] B. A. Olshausen (2003). Learning sparse, overcomplete representations of time-varying natural images. In *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, vol. 1, pp. 41–44.

[85] B. A. Olshausen, D. J. Field (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* **37**(23):3311–3325.

[86] K. C. Partington (2000). A data fusion algorithm for mapping sea-ice concentrations from Special Sensor Microwave/Imager data. *IEEE Trans. Geosci. Remote Sensing* **38**(4):1947–1958.

[87] E. Parzen (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**(3):1065–1076.

[88] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy (2002). Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus. *EURASIP Journal on Applied Signal Processing* **2002**(11):1189–1201.

[89] L. Peotta, L. Granai, P. Vandergheynst (2003). Very low bit rate image coding using redundant dictionaries. In *Proc. SPIE, Wavelets: Applications in Signal and Image Processing X*, vol. 5207, pp. 228–239.

[90] L. Peotta, L. Granai, P. Vandergheynst (2006). Image compression using an edge adapted redundant dictionary and wavelets. *Signal Processing* **86**(3):444–456.

[91] P. Pérez, J. Vermaak, A. Blake (2004). Data fusion for visual tracking with particles. *Proc. IEEE* **92**(3):495–513.

[92] G. Potamianos et al. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* **91**(9):1306–1326.

[93] L. Rabiner, B.-H. Juang (1993). *Fundamentals of speech recognition.* Prentice Hall, Englewood Cliffs, New Jersey.

[94] S. Rajaram, A. V. Nefian, T. Huang (2004). Bayesian separation of audio-visual speech sources. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 657–660.

[95] M. Reyes-Gomez, D. Ellis, N. Jojic (2004). Multiband audio modeling for single-channel acoustic source separation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 641–644.

[96] M. Reyes-Gomez, N. Jojic, D. Ellis (2004). Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation/tracking model. In *Research Workshop on Statistical and Perceptual Audio Proc.*

[97] B. Rivet, L. Girin, C. Jutten (2005). Solving the indeterminations of blind source separation of convolutive speech mixtures. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 533–536.

[98] L. D. Rosenblum, H. M. Saldaña (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* **22**(2):318–331.

[99] S. T. Roweis (2000). One microphone source separation. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 793–799.

[100] A. Sarkar et al. (2002). A MRF model-based segmentation approach to classification for multispectral imagery. *IEEE Trans. Geosci. Remote Sensing* **40**(5):1102–1113.

[101] L. Shams, Y. Kamitani, S. Shimojo (2000). What you see is what you hear. *Nature* **408**(6814):788.

[102] M. Slaney, M. Covell (2000). FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 814–820.

[103] P. Smaragdis, M. Casey (2003). Audio/visual independent components. In *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation (ICA)*, pp. 709–714.

[104] D. Sodoyer, L. Girin, C. Jutten, J.-L. Schwartz (2004). Developing an audio-visual speech source separation algorithm. *Speech Commununication* **44**(1-4):113–125.

[105] N. Sugamura, F. Itakura (1986). Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP. *Speech Commununication* **5**(2):199–215.

[106] W. H. Sumby, I. Pollack (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* **26**(2):212–215.

[107] Q. Summerfield (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd, R. Campbell (eds.), *Hearing by Eye: The Psychology of Lipreading*, pp. 3–51, Lawrence Erlbaum Associates.

[108] J. Tropp, A. Gilbert, M. J. Strauss (2005). Simultaneous sparse approximation via greedy pursuit. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 721–724.

[109] P. Vandergheynst, P. Frossard (2001). Efficient image representation by anisotropic refinement in Matching Pursuit. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 3, pp. 1757–1760.

[110] E. Vincent, R. Gribonval, C. Févotte (2006). Performance measurement in Blind Audio Source Separation. *IEEE Trans. Acoust., Speech, Signal Processing* **14**(4):1462–1469.

[111] E. Vincent et al. (2005). *Blind Audio Source Separation.* Tech. Rep. C4DM-TR-05-01, Centre for Digital Music, Queen Mary University of London.

[112] A. Violentyev, S. Shimojo, L. Shams (2005). Touch-induced visual illusion. *Neuroreport* **10**(16):1107–1110.

[113] M. T. Wallace et al. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research* **158**(2):252–258.

[114] W. Wang et al. (2005). Video assisted speech source separation. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, vol. 5, pp. 425–428.

[115] Y. Wang, J. Ostermann, Y.-Q. Zhang (2001). *Digital Video Processing and Communications.* Prentice Hall.

[116] S. Watkins et al. (2006). Sound alters activity in human V1 in association with illusory visual perception. *NeuroImage* **31**(3):1247–1256.

[117] M. Wertheimer (1923). Untersuchungen zur lehre der gestalt, II. *Psychologische Forschung* **4**:301–350. Translation published as "Laws of Organization in Perceptual Forms", in: W. Ellis, *A Source Book of Gestalt Psychology*, pp. 71-88, Routledge and Kegan Paul, London, 1938.

[118] Y.-S. Yao, R. Chellappa; (1995). Tracking a dynamic set of feature points. *IEEE Trans. Image Proc.* **4**(10):1382–1395.

[119] O. Yilmaz, S. Rickard (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Proc.* **52**(7):1830–1847.

[120] C. Zhang et al. (2006). Boosting-based multimodal speaker detection for distributed meetings. In *Proc. IEEE Int. Workshop on Multimedia Signal Proc. (MMSP).*

[121] S. Zhou, R. Chellappa, B. Moghaddam (2004). Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Trans. Image Proc.* **13**(11):1491–1506.

[122] M. Zibulevsky, B. Pearlmutter (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation* **13**(4):863–882.

Curriculum vitæ

Name:	Gianluca Monaci
Citizenship:	Italian
Birthdate:	July 26, 1976
Birthplace:	Siena, Italy
Marital status:	Single

Contact information

Address:	Rue du Lac 20 1020 Renens, Switzerland
Phone:	+41 21 693 26 57
Fax:	+41 21 693 76 00
Email:	gianluca.monaci@epfl.ch
Web page:	http://lts2www.epfl.ch/∼monaci

Work experience

- ***January 2003 – present***: Research assistant at the Signal Processing Institute, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
  - PhD Thesis research in the field of image, video and multi-modal signal processing.
  - Teaching: supervision of master thesis and responsible of exercises and laboratory activities for the Digital Signal Processing and Advanced Image Processing courses.
- ***2006***: Visiting Researcher at the Department of Electronic Engineering, Queen Mary, University of London, UK
  - Development of an algorithm for the tracking of visual structures.
- ***2002***: Research internship at the Audiovisual Communication Laboratory, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
  - Development of a model of human color perception.

Education

- ***October 2003 – present***: *Ph. D. student* in multi-modal signal processing. Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
- ***2002***: *Master Thesis* at the Audiovisual Communication Laboratory, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
- ***2000***: *Erasmus student* at the University of Technology of Belfort-Montbéliard, France.
- ***1995 – 2002***: *Student in Telecommunications Engineering.* University of Siena, Italy.

Awards

- *IBM Student Paper Award* of the 2005 IEEE Int. Conference on Image Processing (ICIP).
- *UCLA Grant* to attend the graduate summer school "Intelligent Extraction of Information from Graphs and High Dimensional Data" at IPAM, UCLA, Los Angeles, USA (July 2005).

Professional activities

- *Reviewer* for IEEE Transactions on Circuits and Systems II.
- *Member of the Scientific Committee* of the 2004 IEEE Int. Workshop on Multimedia Signal Processing (MMSP).

Skills

Languages

Italian:	mother tongue
English:	fluent oral and written
French:	fluent oral and written
Spanish:	intermediate
German:	basic

Computer literacy

Operating systems:	Linux, Windows
Programming languages:	C, C++, Fortran, MySQL
Other:	HTML, LaTeX, Matlab, CVS, WinVis and VSG toolboxes

Extra-curricular activities

- *Music* : I like to play trumpet; I used to study improvisation at "Siena Jazz" School
- *Sports* : Ski, Football, Basketball, Sailing
- *Reading*, *Cinema*, *Photography*

Personal publications

Journal papers

- G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage and R. Gribonval, *Learning Multi-Modal Dictionaries*, submitted to IEEE Trans. on Image Processing, 2006.
- G. Monaci, O. Divorra Escoda and P. Vandergheynst, *Analysis of Multimodal Sequences Using Geometric Video Representations*, Signal Processing, Vol. 86, Nr. 12, pp. 3534-3548, 2006.
- G. Monaci, G. Menegaz, S. Süsstrunk and K. Knoblauch, *Chromatic Contrast Detection in Spatial Chromatic Noise*, Visual Neuroscience, Vol. 21, Nr. 3, pp. 291-294, 2004.

Conference papers

- A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, *Blind Audiovisual Source Separation Using Sparse Representations*, submitted to IEEE Int. Conf. on Image Proc., 2007.
- G. Monaci, P. Vandergheynst, E. Maggio and A. Cavallaro, *Tracking Atoms with Particles for Audiovisual Source Localization*, Proc. of IEEE Int. Conf. on Acoustic, Speech, Signal Proc., 2007.
- G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage and R. Gribonval, *Learning Multi-Modal Dictionaries: Application to Audiovisual Data*, Proc. of Int. Workshop on Multimedia Content Representation, Classification and Security, in Springer-Verlag LNCS series, Vol. 4105, pp. 538-545, 2006.
- G. Monaci and P. Vandergheynst, *Audiovisual Gestalts*, CVPR Workshop on Perceptual Organization in Computer Vision, 2006.
- G. Monaci, O. Divorra Escoda and P. Vandergheynst, *Analysis of Multimodal Signals Using Redundant Representations*, Proc. of IEEE Int. Conf. on Image Proc., pp. 46–49, 2005.
- G. Monaci, P. Jost and P. Vandergheynst, *Image Compression with Learnt Tree-Structured Dictionaries*, Proc. of IEEE Int. Workshop on Multimedia Signal Proc., pp. 35–38, 2004.
- G. Monaci and P. Vandergheynst, *Learning Structured Dictionaries for Image Representation*, Proc. of IEEE Int. Conf. on Image Proc., pp. 2351–2354, 2004.
- G. Monaci, G. Menegaz, S. Süsstrunk and K. Knoblauch, *Color Contrast Detection in Spatial Chromatic Noise*, 17^{th} Symposium of the International Color Vision Society, 2003.

- G. Monaci, G. Menegaz, S. Süsstrunk and K. Knoblauch, *Spectral Bandwidths for the Detection of Colour within Random Colour Textures*, 26$^{\text{th}}$ European Conf. on Visual Perception, 2003.

Technical reports

- G. Monaci, P. Vandergheynst, E. Maggio and A. Cavallaro, *Tracking Atoms with Particles*, Technical Report TR-ITS-2006.11, 2006.
- P. Besson, G. Monaci, P. Vandergheynst and M. Kunt, *Experimental evaluation framework for speaker detection on the CUAVE database*, Technical Report TR-ITS-2006.03, 2006.
- G. Monaci and P. Vandergheynst, *Detection of Synchronous Audiovisual Events*, Technical Report TR-ITS-2005.36, 2005.
- G. Monaci, O. Divorra Escoda and P. Vandergheynst, *Analysis of Multimodal Sequences Using Geometric Video Representations*, Technical Report TR-ITS-2005.17, 2005.
- G. Monaci, O. Divorra Escoda and P. Vandergheynst, *Multimodal Analysis Using Redundant Parametric Decompositions*, Technical Report TR-ITS-2004.24, 2004.
- G. Monaci and P. Vandergheynst, *Learning Structured Dictionaries for Image Representation*, Technical Report TR-ITS-2004.10, 2004.