A MULTIMODAL PATTERN RECOGNITION FRAMEWORK FOR SPEAKER DETECTION

THÈSE N^{O} 3819 (2007)

PRÉSENTÉE LE 8 JUIN 2007

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
Laboratoire de traitement des signaux 1
SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Patricia BESSON

DESS en génie biomédical, Université Claude Bernard Lyon 1, France
et de nationalité française

acceptée sur proposition du jury:

Prof. A. Rufer, président du jury
Prof. M. Kunt, directeur de thèse
Dr A. Cavallaro, rapporteur
Dr Ph. Salembier, rapporteur
Prof. J.-Ph. Thiran, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Lausanne, EPFL
2007

A is the first letter of the alphabet. There are 25 more. This thesis contains all of them in very interesting combinations.*

* Free adaptation of an advertisement for the public library of Chicago.

Table of contents

List of figures

List of tables

Acknowledgements

I would not have succeed in this thesis without the support and advice of many people. I would like to thank them all.

My first thanks are for my thesis advisor Murat Kunt for proposing me and convincing me to do a thesis. This was not something I had in mind before applying at the Signal Processing Institute (ITS) and it was indeed a worthwhile proposition.

Further, I would like to thank all the members of my jury: the president, Alfred Rufer; Andrea Cavallaro for the very interesting questions he asked about my work, as well as for making me discover the via ferrate in Dolomiti when he was still a post-doc at ITS; Philippe Salembier for his very pertinent and constructive feedback on my thesis as well as for its interest and involvement in our common student project. Jean-Philippe Thiran is of course not the last person I would like to thank, not only for being part of my jury but also for his support all along this Ph.D. time, and especially for having welcomed me in his own research group.

I am also deeply grateful to the other professors in the institute, Pierre Vandergheynst and Pascal Frossard, for their advice and support.

The successful achievement of this work owes a lot to Jean-Marc Vesin and Vlad Popovici. Thank you for being still open to discussions, for your encouragements and new ideas. Also, I would like to specially thank Oscar Divorra Escoda, Rosa Maria Figueras i Ventura and Xavier Bresson who really pushed me ahead and whose discussions about work or life in general have been a great support.

The feedback I received from many colleagues, on my work and on this dissertation in particular, has been precious: thanks to Jonas Richiardi, Ulrich Hoffman, Gianluca Monaci, Merixell Bach and Effrosyni Koklopoulou. I do not forget Julien Meynet, who devoted a noticeable part of its time to provide me information and guidelines about his face detection software.

I also had fruitful discussions with people from outside the institute and I acknowledge Torsten Butz, Kevin O'Connor, Emma Frejinger, Michael Themans, Prof. Anthony Davison, Peter Berlin, and Samy Bengio for their availability and judicious comments.

Many thanks to the students who worked under my supervision (David Cuestas, Bertrand

Grandgeorge, Aurélien Mayoue) for the great job they achieved as well as for all I learned together with them.

I also acknowledge all the persons who accepted to be part of the audiovisual database I required for my tests.

People at ITS were more friends than colleagues: I already mentioned some of them but you have all been absolutely great and made going to work so much fun. I especially keep in mind how present you have all been when I had this bad leg injury which kept me off work for some weeks: thank you so much for that. More generally, we have shared together many good lunch times, after-work times at Sat, ski days, mountaineering activities and of course, BBQs lakeside. These are moments which have made these four years in Lausanne unforgettable. I address here a special thank to Lisa Falco, a very good friend who initiated me to the joy of climbing and was (and still is!) always ready for any of these activities. Thanks to the climber crew, to Elisa Drelie, Nawal Houhou and all the members of the "LTS girls" team, as well as to Lorenzo Peotta, Gianluca Antonini and Valérie Duay, my very first and very last officemates at the ITS.

My friends outside the lab have also been very important to keep my mind off work: thank you to Cristina Cellerai, Gaetano Parascandolo and all the other members of the Italian crew. Thanks as well to the Chaieb family, to Zina Charef and to Linda Achour, maybe far away but still close by mind.

Pour finir, je tiens à remercier ma famille, mes parents, mes frères et ma soeur, pour tout ce qu'ils m'ont apportés et continuent à m'offrir. Je remercie également la famille que j'ai ici en Suisse et qui m'a accueillie dès mon arrivée et s'est montrée présente durant tout mon séjour.

Abstract

Speaker detection is an important component of a speech-based user interface. Audio-visual speaker detection, speech and speaker recognition or speech synthesis for example find multiple applications in human-computer interaction, multimedia content indexing, biometrics, etc. Generally speaking, any interface which relies on speech for communication requires an estimate of the user's speaking state (i.e. whether or not he/she is speaking to the system) for its reliable functioning. One needs therefore to identify the speaker and discriminate from other users or background noise.

A human observer would perform such a task very easily, although this decision results from a complex cognitive process referred to as *decision-making*. Generally speaking, this process starts with the acquisition by the human being of information about the environment, through each of its five senses. The brain then integrates these multiple information. An amazing property of this multi-sensory integration by the brain, as pointed out by cognitive sciences, is the perception of stimuli of different modalities as originating from a single source, provided they are synchronized in space and time.

A speaker is a bimodal source emitting jointly an auditory signal and a visual signal (the motion of the articulators during speech production). The two signals are obviously co-occurring spatio-temporally. This interesting property allows us - as human observers - to discriminate between a speaking mouth and a mouth whose motion is not related with the auditory signal.

This dissertation deals with the modelling of such a complex decision-making, using a pattern recognition procedure. A pattern recognition process comprises all the stages of an investigation, from data acquisition to classification and assessment of the results. In the audiovisual speaker detection problem, tackled more specifically in this thesis, the data are acquired using only one microphone and camera. The pattern recognizer integrates and combines these two modalities to perform and is therefore denoted as "multimodal".

This multimodal approach is expected to increase the performance of the system. But it also raises many questions such as what should be fused, when in the decision process this fusion should take place, and how is it to be achieved.

This thesis provides answers to each of these issues through the proposition of detailed solutions for each step of the classification process. The basic principle is to evaluate

the synchrony between the audio and video features extracted from potentially speaking mouths, in order to classify each mouth as speaking or not. This synchrony is evaluated through a mutual information based function.

A key to success is the extraction of suitable features. The audiovisual data are then processed through an information theoretic feature extraction framework after having been acquired and represented in a tractable way. This feature extraction framework uses jointly the two modalities in a feature-level fusion scheme. This way, the information originating from the common source is recovered while the independent noise is discarded. This approach is shown to minimize the probability of committing an error on the source estimate. These optimal features are put as inputs of the classifier, defined through a hypothesis testing approach. Using jointly the two modalities, it outputs a single decision about the class label of each candidate mouth region ("speaker" or "non-speaker"). Therefore, the acoustic and visual information are combined at both the feature and the decision levels, so that we can talk about a hybrid fusion method. The hypothesis testing approach gives means for evaluating the performance of the classifier itself but also of the whole pattern recognition system. In particular, the added-value offered by the feature extraction step can be assessed.

The framework is applied in a first time with a particular emphasis on the audio modality: the information theoretic feature extraction addresses the optimization of the audio features using jointly the video information. As a result, audio features specific to speech production are produced. The system evaluation framework establishes that putting these features at input of the classifier increases its discrimination power with respect to equivalent non-optimized features.

Then the enhancement of the video content is addressed more specifically. The mouth motion is obviously the suitable video representation for handling a task such as speaker detection. However, only an estimate of this motion, the optical flow, can be obtained. This estimation relies on the intensity gradient of the image sequence. Graph theory is used to establishe a probabilistic model of the relationships between the audio, the motion and the image intensity gradient, in the particular case of a speaking mouth. The interpretation of this model leads back to the optimization function defined for the information theoretic feature extraction. As a result, a scale-space approach is proposed for estimating the optical flow, where the strength of the smoothness constraint is controlled via a mutual information based criterion involving both the audio and the video information. First results are promising even if more extensive tests should be carried out, in noisy conditions in particular.

As a conclusion, this thesis proposes a complete pattern recognition framework dedicated to audiovisual speaker detection and minimizing the probability of misclassifying a mouth as "speaker" or "non-speaker". The importance of fusing the audio and video content as soon as at the feature level is demonstrated through the system evaluation stage included in the pattern recognition process.

Keywords:

Multimodal signal, audiovisual speaker detection, feature extraction, information theory, hypothesis testing, pattern recognition, scale-space theory

Résumé

La détection de locuteur est une composante essentielle de nombreuses applications multimedias mettant en jeu la parole d'une façon ou d'une autre, telles les interfaces homme-machine, l'indexation de documents multimedias, les applications biométriques, etc. Toutes ces applications nécessitent de connaître l'état du locuteur, à savoir s'il est ou non en train de parler au système. Il est donc primordial de détecter dans la scène tout individu potentiellement à l'origine du signal de parole, et de distinguer parmi ceux-ci le locuteur effectif.

Pour un observateur humain, il s'agit-là d'une tâche fort simple, ou tout du moins, qui apparaît telle quelle car elle ne requiert en général nul effort particulier. En réalité, cette prise de décision met un jeu un processus cognitif complexe. D'une façon générale, ce procédé débute par l'acquisition d'informations sur le cadre du problème. Pour cela, les cinq sens de l'odorat, de l'ouïe, de la vision, du toucher, et du goût sont utilisés et les informations multiples qu'ils fournissent sont assimilées simultanément par le cerveau. Cette perception multi-sensorielle de l'information par le cerveau donne lieu à un phénomène intéressant: des stimuli issus de différentes modalités sont perçus comme provenant d'une seule et même source physique pour peu qu'ils soient synchrones tant temporellement que spatialement.

Un locuteur est une source bimodale émettant conjointement un signal auditif et un signal visuel (ce dernier étant lié à la mise en mouvement des articulateurs). Ces deux signaux présentent naturellement une co-occurence spatio-temporelle, ce qui permet à un observateur humain de faire la distinction entre une bouche prononçant effectivement les mots entendus et une bouche dont le mouvement n'est pas lié au signal auditif.

Cette thèse porte sur la modélisation du processus menant à la prise de décision par le biais d'un système de reconnaissance de formes. Un tel système comprend toutes les étapes d'investigation d'un problème depuis l'acquisition des données en passant par la classification et l'évaluation des résultats. Dans le cadre précis de cette thèse, le problème porte sur la détection du locuteur dans une séquence audiovisuelle. Les données sont acquises par une seule caméra et un seul microphone et le système traite donc des informations tant acoustiques que visuelles: il est par conséquent qualifié de "multimodal".

Une approche multimodale devrait permettre d'améliorer les performances du système mais elle soulève également de nombreuses questions, comme de définir ce qui doit être fusionné, à quelle étape du système et de quelle façon cette fusion doit être effectuée afin

d'être bénéfique. Cette thèse répond à chacun de ces points en proposant des solutions détaillées pour chacune des étapes du processus de classification.

Le principe de base repose sur l'évaluation de la synchronie entre des attributs audios et vidéos, ces derniers étant extraits de bouches potentiellement à l'origine du signal de parole. Cette synchronie est évaluée par le biais d'une fonction basée sur l'information mutuelle.

Le succès de la méthode réside dans l'extraction d'attributs appropriés. Une méthode basée sur la théorie de l'information est donc proposée dans cette optique et est appliquée aux données acquises et préalablement représentées de façon adéquate. Cette méthode utilise conjointement les deux modalités dans un schéma de fusion précoce afin d'extraire l'information liée à leur provenance commune tandis que le bruit est écarté. Il est démontré que la probabilité de commettre une erreur sur l'estimée de la source est ainsi minimisée. Ces attributs optimaux sont ensuite passés en entrée du classificateur défini au moyen de tests d'hypothèses et effectuant une nouvelle fusion des modalités à son niveau puisqu'une seule valeur est retournée en sortie. La méthode présentée est donc une méthode hybride dans laquelle deux fusions, précoce et tardive, sont effectuées. Formuler la fonction de classification au moyen de tests d'hypothèses permet de définir dans le même temps des outils pour évaluer tant le classificateur lui-même que le processus entier de classification, y compris le bénéfice apporté par l'étape d'extraction d'attributs.

Dans un premier temps, le système est appliqué en portant une attention plus particulière au signal audio: la méthode théorique d'extraction d'attributs est mise en oeuvre afin d'optimiser l'information auditive tout en utilisant bien entendu conjointement l'information visuelle pour ce faire. Des attributs audio spécifiques à la production de parole sont ainsi produits. La méthode d'évaluation montre que le pouvoir discriminant du classificateur est amélioré par l'emploi de ces attributs en lieu et place d'attributs audio équivalents, non-optimisés.

Dans un second temps, l'optimisation des attributs vidéo est abordée. L'information visuelle la plus évidemment rattachée à la production de parole est le mouvement de la bouche. Cependant, la séquence d'images à notre disposition nous permet uniquement d'estimer ce mouvement, en nous basant sur le gradient d'intensité. Cette estimation est connue sous le nom de flux optique. A l'aide de la théorie des graphes, un modèle probabiliste des relations entre l'audio, le mouvement et le gradient d'intensité est défini (dans le cas particulier où une bouche produit de la parole). L'interprétation de ce modèle nous permet de retrouver la fonction d'optimisation définie précédemment dans la méthode d'extraction d'attributs. En conséquence, une approche échelle-espace est proposée pour estimer le flux optique, en régulant la force de la contrainte de lissage au moyen d'un critère basé sur l'information mutuelle entre les contenus audio et vidéo. Des résultats prometteurs ont été obtenus, même si des tests plus importants sont encore à réaliser, dans des conditions bruitées en particulier.

En conclusion, cette thèse propose un système complet basé sur la reconnaissance de formes et destiné à détecter le locuteur courant dans une séquence audiovisuelle tout en minimisant la probabilité de classer une bouche dans la mauvaise catégorie. L'importance

de la fusion des informations acoustiques et visuelles à un niveau précoce du système, c'est-à-dire lors de l'extraction même des attributs, est démontrée par le biais de la méthode d'évaluation qui fait partie intégrante du système.

Mots-clefs:

Signaux multimodaux, détection audio-visuelle de locuteur, extraction d'attributs, théorie de l'information, tests d'hypothèses, reconnaissance de formes, théorie échelle-espace

Abbreviations and symbols

Abbreviations and acronyms

ANOVA	ANalysis Of VAriance
AUC	area under the curve
B&B	Branch & Bounds
CCA	Canonical Correlation Analysis
DE	Differential Evolution
Def.	definition
EA	Evolutionary Algorithm
EC	efficiency coefficient
Eq.	Equation
FCS	face coordinate system
Fig.	Figure
fps	frames per second
GA	Genetic Algorithm
GACS	Genetic Algorithm in Continuous Space
GMM	Gaussian Mixture Model
ICA	Independent Component Analysis
i.i.d.	identically independent distributed

LDA	Linear Discriminant Analysis
MC	Markov chain
MFCC	Mel-Frequency Cepstrum Coefficient
MI	mutual information
MO	multi-objective
ms	millisecond
N**D**	N-dimensional
1D	one-dimensional
OF	optical flow
PCA	Principal Component Analysis
pdf	probability density function
PR	pattern recognition
ROC	Receiver Operator Characteristic
rv	random variable
s	second
sec.	section
seq.	sequence
SO	single-objective
SOM	self-organizing map
TDOA	Time-difference of arrival
TS	Tabu Search

Symbols

$\mathbb{R}$	The set of real numbers
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}_+$	The set of positive real numbers
$\mathbb{N}^*$	The set of natural numbers, excluding the zero
X, x	Random variable and its specific value
$p_X(x)$ **or** $p(x)$	pdf of a discrete random variable X
log	Base-2 logarithm, $\log_2$
I	Shannon's mutual information
H	Shannon's entropy
e	Efficiency coefficient function
T	Temporal length of the observation window sampled at the video frame-rate
τ	Temporal length of the observation window sampled at the optical flow frame rate: $\tau = T - 1$
$X \perp\!\!\!\perp Y$	X is statistically independent of Y
§	paragraph

1 Introduction

1.1 Description of the problem

As digital multimedia applications integrate everyone's life, the development of dedicated tools for acquiring and processing the related information becomes critical. Human-computer interaction, multimedia content indexing, biometrics, etc. are examples of such applications. Often, speech is involved in some way, as it is a typical vector for human's communication. Generally speaking, any interface which relies on speech for communication requires an estimate of the user's speaking state (i.e. whether or not he/she is speaking to the system) for its reliable functioning.

This implies first of all the system to be able to discriminate the potential speakers from background noise, then to discriminate the "true" speaker from other users. This is quite challenging, as the scene may be pretty noisy, due to multi-speaker interference, natural sounds (coughing, outside activity), room reverberation, changes in lighting conditions, motion and partial occlusion of speakers. Moreover, this system should perform with simple material such as a single camera and microphone for being practical.

1.2 Motivations

Social interactions are an inherent component of any human society. These interactions would not be possible without communication, whose primary vector is speech. For a naive observer, speech consists only in an auditory signal produced by a speaker and catched by a listener through its sense of hearing. From this perspective, it is then a unimodal phenomenon, where by *modality*, we mean:

Definition 1 - *Modality [1]:*

1. *(formal) The particular way in which something exists, is experienced or is done.;*
2. *(biology) The kind of senses that the body uses to experience things.*

In reality, the production and the perception of speech are multimodal: articulators get into motion to modulate the emitted auditory wave. There are then two kinds of signals - or two modalities - emitted during speech and perceived by the listener in a face-to-face conversation. These are the auditory speech and the visual speech, the latter referring to the motion of the visible articulators (lips, jaws, ...).

There are number of empirical evidences demonstrating that the perception of speech is profoundly influenced by vision. Besides, this integration of the two modalities by the brain is not specific to speech. Cognitive sciences have put in evidence the early integration of multi-sensory information by the brain. In particular, stimuli of different modalities are perceived as originating from the same source if they are both spatially and temporally synchronous. This effect is experimented when watching a dubbed movie for example; or by an observer who try to guess who is the "true" speaker when two persons are standing side by side and moving both their lips whereas one of them only is speaking. By evaluating the synchrony between the auditory speech and the possible visual speeches, the observer will be able to take a decision about the current speaker. This approach is called *decision-making* and involves a complex cognitive process, especially in this specific case where an integration of multi-sensory information by the brain takes place.

Theoretical frameworks such as pattern recognition (PR), have been developed for machines to emulate this human's decision-making. As human beings start the decision-making process by acquiring knowledge about the environment, a pattern recognizer starts by the acquisition of the data through sensors. Different processing steps lead then from these raw data to a final decision. In problems dealing with source localization, such as the one just presented, the observation of human perception advocates a multimodal approach to PR. This rises some crucial questions however, about the integration of the modalities in the processing chain as well as about their interaction, for the system to really find an added-value in this complex approach.

1.3 Aim of the thesis

The aim of this thesis has been to design a whole pattern recognition system, dealing with signals of different modalities emitted by a common physical source - such as audiovisual speech - and trying to recover the spatio-temporal location of this physical source.

As a prior requirement, the system was asked to perform in simple acquisition conditions, i.e. with a single camera and microphone. Indeed, at the beginning of this thesis, most of the multimodal approaches to speaker detection were dealing with complex acquisition systems (multiple microphones, cameras, ...) and integrating the multimodal information

at a late stage of the pattern recognition process. This is because they were often combining simply the methods adjusted for years for each different modality individually. They were then not taking advantage of all the potential offered by a multimodal approach, as put in evidence by some exploratory studies. These were exploiting the temporal co-occurrence between auditory and visual speech for detecting the speaker with simpler material. Hence, they lent a fresh perspective on the problem, and asked the question of the level at which the fusion should take place in the detection process for its performance to improve. Indeed, a key to success in source localization problems, as in many others, is the extraction of suitable features. It is then natural to consider the added-value of a multimodal approach to feature extraction.

A global framework relying on a synchrony-based approach and integrating multimodal solutions for feature extraction, classification and system evaluation was missing. The main objective of this thesis is to propose such solutions. A statistical approach has been chosen for the flexibility it allows when dealing with signals of different physical nature. A postulate is that fusing the information at earlier stage of the detection process should improve the final performance of the system, i.e. minimize its probability of error. Therefore, an aspect we have concentrated on is the question of the validation of the results, in particular the evaluation of the performance of the whole multimodal process.

1.4 Main contributions

The main contributions of this thesis can be summarized as follows:

- Development of a full pattern recognition system for multimodal speaker detection, from the data acquisition step all the way to the system evaluation, such that the probability of error of the process is minimized.
- Development of an information theoretic feature extraction framework dealing with different modalities coming from a common physical phenomenon.
- A novel multimodal classifier relying on hypothesis testing which gives means for directly assessing its performance but also the performance of the whole PR process. In particular the added-value of the feature extraction step can be evaluated. Together with this evaluation framework, the precise definition of the experimental setup and of the ground truth of the dataset has been achieved, allowing future comparisons with the presented method.
- Application of the multimodal PR system to acoustic speech signal with development of an optimization framework which leads to extract audio features specific to speech production.
- Precise study of the optimization scheme and comparison of different optimization algorithms for best results.

- Development of a multimodal scale-space approach to optical flow estimation by applying the multimodal PR system to visual speech.

1.5 Thesis road map

This dissertation is organized as follows:

Chapter 2 The speaker detection problem is undertaken as a multimodal pattern recognition task in this work. But was is a pattern recognition system? What are the main components, or main processing steps, that are part of it? These questions are answered to in this chapter. The question of the integration of a multimodal perspective in a PR design is tackled then, through a discussion about what should be fused and when this fusion should occur in the recognition process. A presentation of seminal works in audiovisual speaker detection using different fusion schemes conclude this presentation of the problem.

Chapter 3 The undertaken approach relies on statistics and information theory, thus a review of some basic terms and concepts related to these fields is given.

Chapter 4 Presentation of the multimodal pattern recognition system. At each step of the process, errors may occur and pass on tho the next stages, resulting possibly in poor process performance. Thus solutions are proposed for the main three last steps of the PR (the first step being directly application dependent): feature extraction, classification, and system evaluation, in order to minimize this probability of error.

Chapter 5 The PR system is applied in the context of audiovisual speaker detection. The PR steps unaddressed previously (data acquisition and representation) are clarified. A particular emphasis is put on the auditory speech since the information theoretic feature extraction framework developed in the chapter 4 is applied to extract meaningful audio features using jointly the video content. This requires the precise definition of the optimization problem and a suitable optimization method to be chosen. Thus a precise study of different possible algorithms is carried out. The performance of the classifier and of the whole PR system are evaluated. For this purpose, an experimental framework is precisely described, including the ground truth labelling of the test database. This way, further comparisons are allowed.

Chapter 6 Application of the PR system to the audiovisual speaker detection task, focusing on the enhancement of the visual speech features through the information theoretic framework developed in chapter 4. A scale-space approach is proposed for the problem of OF estimation.

Chapter 7 This dissertation concludes with a resume of our achievements and a discussion of future research topics.

Speaker detection as a multimodal pattern recognition task

2

2.1 Introduction

Speaker detection is an important component of a speech-based user interface. Audio-visual speaker detection, speech and speaker recognition or speech synthesis for example find multiple applications in human-computer interaction, multimedia content indexing, biometrics, etc. Generally speaking, any interface which relies on speech for communication requires an estimate of the user's speaking state (i.e. whether or not he/she is speaking to the system) for its reliable functioning. One needs therefore to identify the speaker and discriminate from other users or background noise. Such a task can be handled as a PR problem, as will be shown in the next section.

After a presentation of the different components of a pattern recognizer in sec. 2.2, it will be shown how the multimodal perception of the environment, and of speech in particular, by human beings advocates a multimodal pattern recognition approach to the problem of speaker detection. This approach raises the question of knowing *what* should be fused, as well as *when* and *how* this fusion should take place in the PR process. These are critical issues, mainly depending on the problem at hand. A discussion about what should be fused for handling the best the speaker detection task is led in sec. 2.3. Once decided the modalities to be fused, they have to be combined at one of the steps of the PR process. The different possibilities are introduced in sec. 2.4 and discussed through the presentation of seminal works in audio-visual speaker detection in sec. 2.5. Of course, how to fuse the modalities is the main matter of this thesis and is considered in the next chapters.

2.2 Pattern recognition design

2.2.1 What is pattern recognition?

In their everyday life, human beings are facing situations where choices have to be made. These choices result from a complex cognitive process referred to as *decision making* and firstly theorized by Nobel Laureate Herbert Simon who created with Allen Newell the Logic Theory Machine [93]. This seminal work has been followed by many researches which tried to reproduce human's learning capabilities with computer programs. Pattern recognition is a particular field of artificial intelligence ensuing from these researches.

Like human beings use their senses to acquire knowledge about their environment in order to react to it, pattern recognizers observe the environment through sensors, learn to distinguish patterns of interest and make decision about their categories or classes. Thus, instead of using pre-defined rules, objects are classified through their observable information, which can be referred to as pattern.

A pattern is described through some features specified by the investigator and thought to be important for classification [141]. If patterns are generated by a probabilistic system, we talk about a statistical PR approach, in contrast to structural PR methods, based on the structural interrelationships between features [67].

A speaker detection task consists in taking a decision about the state - silent or speaking - of an individual, given some patterns characterizing this state. Formulated in these terms, it is obviously a pattern recognition problem.

2.2.2 Different stages involved in the design of a pattern recognizer

Decision making starts in human beings by acquiring knowledge about the environment. Similarly, a pattern recognizer starts by the acquisition of data through sensors. A tractable representation of the raw data is then generated and feeds the classifier which finally outputs a decision. The problem domain dictates the choice of sensor(s), preprocessing technique, representation scheme, and the decision making model [67]. A more complete pattern recognition system would also include a feature extraction and an evaluation stages. The structure of such a standard PR design is shown in Fig. 2.1.

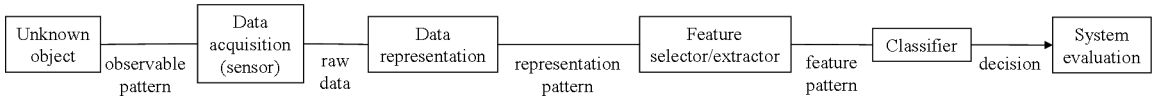


Figure 2.1 — Standard pattern recognition system.

The purpose of the feature extraction step is to produce a set of limited yet salient features. The redundant or useless information might degrade the classifier performance and should then be removed, while the valuable content should be preserved. Often the dimensionality of the features is also decreased to alleviate the so-called curse of dimension-

ality which states that the number of training data points must be an exponential function of the feature dimension [9].

Feature extraction methods must be differentiated from feature selection methods. The latter select the best subset of the input feature set, whereas the former create new features by transforming or combining the original feature set [67].

Many feature extraction methods exist, either linear or nonlinear. Presenting all of them is beyond the scope of this thesis. In a few words, linear algorithms such as Principal Component Analysis (PCA), factor analysis, or Linear Discriminant Analysis (LDA) are the most widely used, due to their simplicity and easy-of-use. However, they are likely to fail to capture a nonlinear relationship. Kernel PCA, multidimensional scaling, self-organizing map (SOM) or information theoretic approaches deal more efficiently with data that exhibit complex relationships. For a detailed coverage of the wide field of feature extraction methods, the reader is referred to [57], [141], [129], [22].

2.3 Multimodal pattern recognition

2.3.1 Data fusion

The Def. 1 given in the introduction of this dissertation gave the two following meanings for the term "modality":

1. (formal) The particular way in which something exists, is experienced or is done.;
2. (biology) The kind of senses that the body uses to experience things.

It has appeared as soon as in the 60's that using separate modalities in a complementary method should improve the desired results. A mathematical model referred to as "data fusion" has been developed for such a data manipulation [132], [138], [3]. "Data fusion is a formal framework in which are expressed the means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of 'greater quality' will depend upon the application." [137]. This framework allows a better description and formalization of the synergy potentials between the available sources of information, thus a better exploitation of the data can accordingly be performed.

In term of robustness and stability, the advantages in using data fusion are obvious: the system is operational even if one or several sources of information are missing or malfunctioning, and the coverage in space and time is extended. Hence it increases the quality of the deduced information and it reduces the vulnerability of the system, as well as its ambiguity.

Data fusion is a very general principle which has been applied to many different areas from military [58] to medical or robotics applications [71]. It is inspired from the capabilities humans and animals have evolved to improve their ability to survive. Indeed, when

humans perceive their environment, they make a coordinate use of their senses. In human computer interface problems, such as speaker detection, human perception system should be considered even more as a reference. Therefore, multimodal - or data fusion - approaches are advocated for such tasks. But then, the central question we face to is: "*What do we fuse, when and how do we perform the fusion*"? As mentioned in the introduction, these are critical issues in multimodal PR tasks that must be answered to. They imply choices to be made, which must be carefully considered for the multimodal approach to be benefit for the problem at hand. Let us firstly address the "what" point, i.e. the choice of the modalities to be fused. A close look to the new discoveries in cognitive sciences should help us to answer this matter.

2.3.2 Cross-modality effects and speech: the "what" point

Obviously, the world we are living in is multimodal: many natural events consists of co-occurring or correlated occurrences in different sensorial modalities [34]. Thus, human beings integrate information coming from each of its five senses to experience its environment and react to it.

Along the twentieth century, cognitive sciences have evolved from a unimodal model of perception (atomism theory) towards a perceptual grouping model (the Gestalt theory). Whereas atomism examined the parts, with the idea that these parts could then be put back together to make a whole, Gestalt theorists start from the whole, assuming that "the whole is something else than the sum of its parts" [75]. Instead of focusing on one sensory modality, with maybe a late integration, researchers have started to study multi-sensorial perception and the attached binding problem (the way this multimodal integration is achieved by the brain) [110] as a proper component of the perception process. Since the senses of sight and hearing are typical and important senses of human beings, most of the works focus on audio-visual stimuli, though multimodal experiments involving other modalities (e.g. the sense of touch) have also been carried out (see [37] for example).

Through new behavioral and brain imaging studies, it has been demonstrated that there exist cross-modality effects in human perception [135], [29], [10] [116] and that temporal co-occurrence, or synchrony, takes an important place in these effects [86], [108], [10].

Ventriloquism for example, is an audio-visual illusion resulting from this cross-modal processing of information by the brain. It is created by presenting synchronous auditory and visual cues in slightly separate locations. As a result, the sound source is shifted in the direction of the visual stimulus. Everyone experiences this effect when watching movies: the voice seems to emanate from the actor's lips rather than from the actual sound sources (the loudspeakers). A standard explanation to ventriloquism effect is that the perceptual system assume that a single event occurred when an auditory and a visual stimuli happen in close spatio-temporal proximity [136]. The preferential shift towards the direction of the visual event rather than the other way results from the superiority of the visual localization against auditory localization.

However, if the visual source is corrupted, auditory stimulus becomes dominant. Heron *et al.* have presented a series of experiments in [60] where an auditory cue is presented synchronously to a visual moving target whose spatial position uncertainty is modulated by varying its size. As the visual uncertainty increases (larger target sizes), the influence of the auditory signal in the visual localization increases. Shams has also shown in [114] that visual illusion can be induced by sound: a single flash accompanied by multiple auditory beeps is incorrectly perceived as multiple flashes.

These studies evidence that cues from multiple modalities are perceived neurally and perceptually as originating from the same source, if they are both spatially and temporally synchronous.

Speech is an obvious case of such a multimodal source, emitting jointly and synchronously an acoustic and a visual stimuli since the production of speech requires the speech articulators to move, inducing facial motion. The perception of speech is then subjected to cross-modal effect. Actually, human speech perception is profoundly influenced by vision. It is well-known that watching a speaker's mouth movements significantly improves comprehension, not only for hearing impaired but also for normal listeners in noisy environments [112], [49]. But the improvement offered by the bimodal speech perception exist even for non-recognition aspects. A recent set of experiments demonstrated that seeing the speaker's lips enhances sensitivity to acoustic information, decreasing the auditory detection threshold of speech embedded in noise [54], [53]. Finally, the McGurk effect is another well-know illusion which amazingly puts in evidence the multimodal perception of speech.In this experiment, a video of [ga] is seeing while listening to a synchronized audio of [ba]. In most case, the perceived results is [da] [85]. Several studies have investigated further this perceptual effect, trying to come to viable explanation of the phenomenon. However, how and why the auditory and visual information are integrated is still an issue [55], [109].

In other words, is speech perception just a special case of audio-visual perceptual grouping [36] or does it exist a multimodal speech-specific mode of perception [79], [130]? The Motor Theory of speech enounced by Liberman [78], [79] argues that speech is special as it originates from a pre-phonetic capacity to perform speech sounds and gestures, i.e. speech is not simply a string of syllable-based noises but a complex sequence of temporally overlapping, co-articulated units of movements. Since the Motor Theory postulates that speech is perceived by reference to how it is produced, thus in a non-unimodal way, it goes in favor of a special processing of audio-visual speech. However, as mentioned above, this question is still subject to debate. It may appear a bit beyond the scope of this thesis, though it is of relative importance for knowing at which step of the pattern recognition process the two modalities should be fused.

Whatever, the strong audio-visual perceptual grouping effects observed with speech make obvious the choice of a multimodal approach to speech-related problems. Speaker detection in particular is concerned with a spatial localization, at a given time, of the speech source (i.e. the speaker). Ideally, this detection should be performed with simple material such as a single microphone and camera. In such a scheme, neither the acoustic nor

the visual cues are discriminating enough in space for a robust localization to be possible. Visual cues are useful in deciding whether a user is facing the system and whether he is moving its lip. However, the distinction between a user and an active listener, who may be smiling or nodding without saying anything cannot be achieved on these visual cues alone. Audio cues, on the other hand, provide useful evidences for speech production but a mono signal does not permit to localize spatially the sound source. The synchronous occurrence of an acoustic and a visual speech events is then a valuable information which can be exploited to achieve such a localization.

Coming back to the question asked in the previous paragraph, we already have clues about the "what" point: audio and video modalities will be used in our multimodal approach to speaker detection. The question however is still open on how and at which step of the classification process the fusion of the two modalities should be performed to be the most efficient (the "when" point).

2.4 Fusion levels: the "when" point

The processing level at which the integration of the audio and video speech signals should take place is still an open issue. Classically, three levels of integration are defined [58], [38], [29]: *low-, middle- or high-level* fusion. *Hybrid* approaches, which combine fusions at different levels, also exist. Basically, each level corresponds to a stage in the pattern recognition process as described in Fig. 2.1.

- Low-level fusion or *raw data fusion* combines several sources of raw data (or observations) to produce a new raw data set.
- Middle-level fusion or *feature-level fusion* uses jointly the information in data of different modalities to extract representative features.
- High-level or *decision-level fusion* involves a multimodal decision function. Two different configurations can happen: the *information combination* and the *classification fusion*. Our definition of these two cases differs slightly from the definition given by Chibelushi in [29]. By the former, we denominate techniques where the scores output by single-modality classifiers are fused to produce a final output. The latter designates frameworks where the classification function takes at input features of each modality and outputs a decision.

Low-level fusion deals with commensurate data, thus data which are usually collected from similar sensors [58]. Therefore, this scheme does not occur in audiovisual speaker detection and the fusion level hierarchy resumes to:

- feature-level fusion;
- classification fusion;

- information combination.

The corresponding schematic representations are displayed in Figs. 2.2, showing at which steps of the pattern recognition process of Fig. 2.1 these fusion schemes take place.

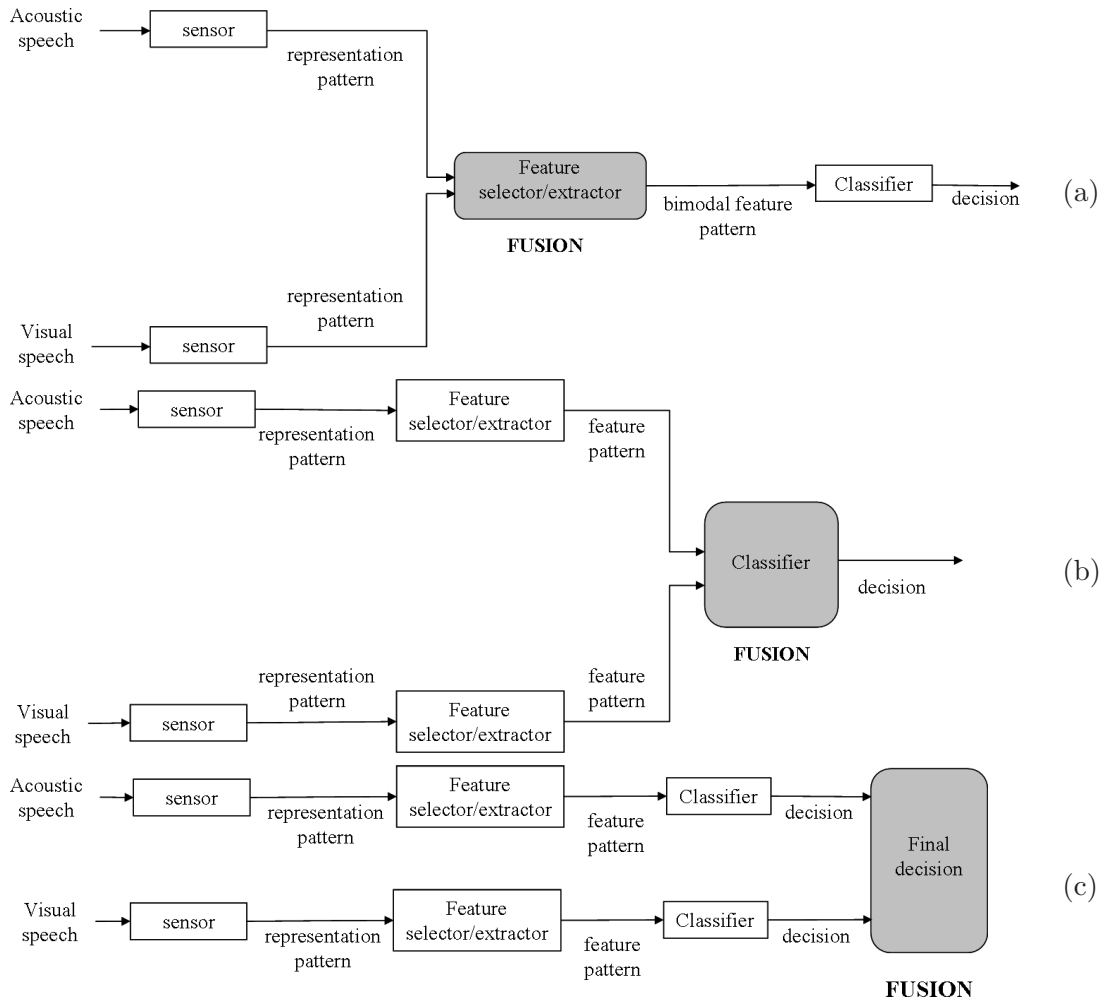


Figure 2.2 — (a) Feature-level fusion; (b) Classification fusion; (c) Information combination.

Seminal works in audio-visual speaker detection are presented in the next section, with reference to the fusion level scheme they have adopted.

2.5 Seminal work in multimodal speaker detection

2.5.1 Information combination

The first fusing approaches in speaker detection simply combined the classification approaches adjusted for years for each modality independently. The idea is that, by combining multiple modalities, they can compensate each other's drawbacks in tracking object and attain finally robust performance.

These methods require the use of several microphones (or microphone arrays) for spatially localizing the speech source from acoustic clues only. Usually, this localization is based on a time-difference of arrival (TDOA) followed by a triangulation.

Good examples of such approaches are given in [140] and [139], where audio and video signals are combined in a video-conference system to detect the current speaker. A coarse acoustic-based localization of the speech source is first performed. Thereby, the camera can be oriented towards the right direction. The processing of video data using different descriptors (motion estimation, contours extraction, color statistics, face analysis) allows then to detect the faces in the sequence. A lot of works using the same principle of detecting independently audio sources (generally TDOA-based) and faces (using color histograms, contours extraction, motion detection, etc.) have appeared over the last years (see for example [84], [65], [72], [83]). Information combination approach is also widely applied in the robotics field [92], [96]. For such applications, the combination of audio and video signals helps greatly the detection task: mobile robots capable of auditory perception usually have to stop before interacting with human because of the audio noise of the motor. The multimodal approach helps to alleviate this restriction.

2.5.2 Classification fusion

Classification fusion scheme is already a more elaborated approach to audiovisual fusion than information combination approach, the content of both signals being exploited more efficiently.

One of the first example of such a fusion scheme has been proposed by Takahashi and Yamasaki in [128], and a similar approach is taken in [52]. The speaker's face is detected by looking for colors of same tone than the skin. Of course, such a detection alone is not robust at all as background pixels can obviously present similar tone. Therefore, an audio localization (requiring multiple microphones) assigns heavier weights to pixels labelled as skin, and a joint probability considering both sound and color is estimated to take the final decision about the state - speaking or silent - of the speaker.

Generally speaking, statistical approaches predominate: the flexibility they offer allows to integrate easily the different modalities at the classification level. A graphical model for audiovisual tracking is for example proposed in [7]. In [100] and [47], a learning algorithm based on dynamic bayesian networks is used to compute the joint probability of having a speaker considering different acoustic and visual cues (skin color, texture, lips motion). Other methods use decentralized Kalman filter to perform tracking [123]. To avoid the formulation of implicit Gaussiannity and/or linearity restriction, non-parametric approaches such as particle filters are also in increasing use these years [134], [143], [48].

Some classification fusion approaches are particularly noticeable because they take advantage of the temporal co-occurrence between the audio and the video speech signals to get a spatial localization of the source. As a result, they only require one microphone and camera to perform.

In [61], Hershey and Movellan use a per-pixel measure of the correlation between the audio and video signals - considered as samples of independent Gaussian processes - to localize the bimodal source. They suggest that the synchrony is the perceptive effect of the causal relationship between the two signals.

A time-delayed neural network is used in [33] to learn the correlation between the audio and video signals. It is then used in a new audio-video sequence for searching correlated motion and audio indicative of a person talking. The method presented in [56] also require a training stage to learn the parameters of the Gaussian Mixture Model (GMM) estimating the joint probability density function (pdf) of the audio and video features. The audiovisual source is localized in a new sequence by estimating the joint pdf between the audio sample and video samples taken from a sliding image region analyzer. The region whose corresponding joint pdf maximizes the likelihood of having been generated by the learned parameters is labelled as the speaking mouth.

In [89], the audio and video signals are represented as sparse sums of few representative functions taken from a dictionary. The temporal displacement of visual relevant features, like the mouth is compared with the evolution of the audio feature to detect correspondences between the signals. These correspondences are found by evaluating the number of peaks falling jointly in a given time window.

The authors in [56] and [89] both compare their results with the ones presented by Nock *et al.* in [95], who carried out the first quantitative study about audiovisual speaker detection based on synchrony evaluation. They tested three methods relying on different statistical considerations on twelve sequences taken from the CUAVE database [98], and compared their performance. The so-called Gaussian mutual information (MI), where multivariate Gaussian distributions are assumed and the MI used as synchrony evaluation function, is shown to outperform the two other ones.

2.5.3 Feature-level fusion

Among the different methods that exploit the information contained in each modality, a few are performing the fusion directly at the feature level. It has though been pointed out in [28] and [44] for example, that such a fusion can greatly help the classification task: the richer and the more representative the features, the more efficient the classifier. Thus the purpose of the here-presented methods is to map the features into a subspace where their relationship is enhanced prior to perform the detection, in order to lead to a better detection performance. As the last methods presented in the previous paragraph, these approaches exploit the synchrony existing between the acoustic and visual speech signals. A single microphone and camera are then only required as well.

The various approaches differ in their way to process the data and the cost function to be optimized to perform the mapping or the synchrony evaluation, mainly mutual information, maximum likelihood, and entropy.

In [120] and [73], a Canonical Correlation Analysis (CCA) is performed (incidentally,

this is equivalent to maximum MI projection in the jointly Gaussian case).

Fisher *et al.* propose in [44] to find the linear combinations of low-level audiovisual representations that maximize the mutual information.

In [121], the authors perform a PCA followed by an Independent Component Analysis (ICA) on an audiovisual feature vector obtained by concatenating the audio and video feature sets. The maximally independent audio-video subspaces indicate the common source localization.

A methodology for learning meaningful synchronous structures from multimodal signals is proposed in [90]. Once the multimodal components have been learned in a training phase, they are used as filters on new audiovisual sequences. The temporal position (for the audio) and spatio-temporal position (for the video) of the maximal projections are obtained and can be used to recover the spatio-temporal location of the common source.

An estimation of the features' probability density functions is generally required. Gaussian distributions are often explicitly or implicitly assumed. However, such an a priori assumption is not necessarily valid. Fisher in [44], as well as Butz in [28], estimate the probability density functions directly from the available samples during the feature extraction process through Parzen windowing [97].

2.6 Discussion

This chapter has put the grounds of the approach tackled in this thesis. Let us recall here that the target application is the detection of the active speaker on audiovisual sequences, using a single camera and microphone. This task is typically related to human behavior. Thus a methodology inspired from how humans perceive their environment in order to react to it, such as pattern recognition approaches, is naturally advocated. Similarly, the study of humans' behavior and the discoveries in cognitive science support the choice of a multimodal approach to pattern recognition task.

Stating the problem this way opens the question of knowing "what do we fuse, when and how do we perform the fusion"?

The answer to the "what" point is straightforward, giving the specificity of the task at hand: the acoustic and visual signals have to be the two used modalities.

To know when in the PR process the fusion should occur is already a biggest issue. The main reasons to that, beside the fact that multimodal approach to speaker detection is a quite recent field of research, are that: 1) The binding problem, or integration of multiple sensory information by the brain, is still unsolved; 2) It is not sure that a computer approach should follows the humans way of fusing the information to be the most efficient.

However, it has been shown, through the presentation of different works in the domain in particular, that three main levels of fusion can be undertaken, each one corresponding broadly to a stage of the pattern recognition process. A feature-level approach looks the most promising since the sooner you combine the data information in the classification

process, the better you take advantage of their redundant and complementary content. All the more, hybrid fusion is also possible. Then a feature-level combination does not exclude fusion at further levels.

Finally, the speaker detection task should be undertaken through a bimodal PR framework combining the acoustic and visual content from the feature-level. A solution for all the stages of a standard pattern recognizer as shown in Fig. 2.1 has to be proposed so that the final process matches these requirements. The final evaluation step is of primary importance to establish the performance of the global system and of the multimodal feature extraction step in particular. For as far as we know, no audiovisual speaker detection method applying a feature-level fusion did propose such an evaluation stage. As stated in sec. 2.5, most of the works have been primarily concerned with studies investigating the feasibility of synchrony-based approaches to speaker detection. It is only recently that some evaluation works did appear [95], [89], [56], but this concerned classification fusion schemes only.

The precise description of the solutions offered for each step of the pattern recognizer constitutes the "how" point of the question previously asked. It will be answered to throughout the remaining of this thesis, and in particular in chapter 4 where a theoretical framework is proposed for the key stages of the process: the feature extraction, the classification and the evaluation stages. This chapter is preceded by a short introduction to some basic concepts of statistics and information theory. Indeed, a statistical approach to pattern recognition is undertaken due to the flexibility it offers.

Review of some basic concepts and notations

3

3.1 Probability and statistics: review and notation

3.1.1 Probability space

In this thesis, the problem of speaker detection is considered from a statistical point of view. This chapter is meant to briefly review some terms and concepts that will be widely used in the following chapters. However, for a more formal and exhaustive coverage of probability and statistic theories, the reader is referred to [25] and [101].

The probability theory is an attempt to work mathematically with the relative uncertainties of random events. Let us start by defining a set S as a collection of objects, or elements, s. An *outcome* is the result of a single experiment, or trial and a *sample space* Ω is defined as the set of all possible outcomes in any given experiments. A set of outcomes is often considered rather than a single outcome. Such a set of outcomes is called an *event* and forms a subset of the sample space. To each event on the sample space Ω, a *probability* P is assigned: $P(E)$ is then a function of the event E.

These concepts having been introduced, it is now possible to define the *probability space* which describes completely a random experiment:

Definition 2 ***(Probability space)*** *- A probability space $(\Omega, \mathcal{B}, P)$ is a triple consisting of a sample space Ω, a family $\mathcal{B}$ of subsets of Ω, thought as the family of observable events, and a function P called the probability measure $P : \mathcal{B} \rightarrow [0, 1]$. This function gives the probability $P(E)$ of occurrence of each observable event $E \in \mathcal{B}$.*

3.1.2 Random variable

In order to define events in a more consistent manner, let us now introduce the concept of random variable.

Definition 3 ***(random variable)*** *- A random variable (rv) X is a real-valued* function defined on a probability space. It maps the outcomes λ of a sample space Ω onto a set of real numbers. The properties associated to this function are:*

1. *The set $\{\lambda : X(\lambda) \leq x\}$ is an event for any real number x.*
2. *The probability of the events $\{\lambda : X(\lambda) = \infty\}$ and $\{\lambda : X(\lambda) = -\infty\}$ equals zero, that is, $P(X = \infty) = P(X = -\infty) = 0$.*

A random variable can be discrete or continuous. A *discrete rv* has only discrete values. However, its sample space can be discrete, continuous, or a mixture of discrete and continuous points. A *continuous rv* has continuous range of values and its sample space is continuous as well.

A *multivariate random variable*, or *random vector*, is a vector whose components are scalar-valued random variables on the same probability space. Scalar-valued random variables take values on $\mathbb{R}$, while vector-valued random variables are defined on $\mathbb{R}^d$ with $d > 1$.

3.1.3 Probability distribution and density function

A random variable is completely specified by the knowledge of its *distribution function*. The distribution function of the rv X is defined as:

$$F_X(x) = P(X \leq x), \tag{3.1}$$

where $P(X \leq x)$ is the probability of the event $\{\lambda : X(\lambda) \leq x\}$, thus a function of x.

The properties attached to a distribution function will not be given here. More details can be found in [101] for example.

The behavior of a random variable can also be characterized by the *probability density function* (pdf).

The probability density function, or simpler, density function, is defined as the derivative of the distribution function:

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{3.2}$$

It must satisfy the two following requirements:

$$f_X(x) \geq 0 \text{ for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(x)dx = 1. \tag{3.3}$$

*Complex values of random variables can also be considered.

For a discrete rv, the distribution function is not derivable. Introducing the Kronecker delta function $\delta(x)$:

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{else }, \end{cases} \tag{3.4}$$

and using the shortened notation:

$$P(x) = P(X = x), \tag{3.5}$$

the probability density function for a discrete rv is defined as:

$$f_X(x) = \sum_{x_i \in \Omega_X} P(x_i)\delta(x - x_i), \tag{3.6}$$

where Ω_X is the sample space of X. It is then a sum of Kronecker delta functions weighted by the appropriate probabilities.

3.1.4 Joint probability

Let X and Y be two random variables defined on a sample space Ω. Their specific values are denoted by x and y respectively. The probability of the joint event $\{X \leq x, Y \leq y\}$, function of both x and y, is called the joint distribution function $F_{X,Y}$ and is defined as:

$$F_{X,Y}(x, y) = P[(X \leq x) \cap (Y \leq y)]. \tag{3.7}$$

It is also possible to consider X and Y as two random variables defined on two sample spaces Ω_X and Ω_Y, with specific values still denoted by x and y respectively. Then a new sample space, called the combined, or joint, sample space Ω, can be defined as the Cartesian product of Ω_X and Ω_Y: $\Omega = \Omega_X \times \Omega_Y$. The elements of Ω are all the ordered pairs (x, y). These pairs (x, y) can be viewed as the specific values of the random vector $\vec{Z} = X, Y$. The joint distribution given by Eq. (3.7) becomes the joint distribution function $f(\vec{Z})$ of the two-dimensional random vector $\vec{Z}$. At each point $\vec{z} = (x, y) \in \mathbb{R}$, the value of $F(\vec{z})$ is specified by Eq. (3.7).

In other words, it is possible to define two or more random variables on the sample space of a single random experiment, or on the combined sample space of many random experiments.

The previous lines of reasoning, as well as Eq. (3.7), generalize easily to n rvs $X_1, \ldots, X_n$ or to a n-dimensional random vector $\vec{X} = X_1, \ldots, X_n$. The corresponding joint distribution function is given by:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P[(X_1 \leq x_1) \ldots (X_n \leq x_n)]. \tag{3.8}$$

For n random variables $X_1, \ldots, X_n$ (or for a n-dimensional random vector $\vec{X} \in \mathbb{R}^n$), the joint probability density function, denoted $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$, is defined by the nth derivative of the joint distribution function wherever it exists.

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}. \tag{3.9}$$

For discrete random variables, the joint distribution function possesses some step discontinuities where the derivatives are normally undefined. However, by admitting Kronecker delta functions, it is possible to define the joint density function at these points.

3.1.5 Marginal distribution function

Let X and Y be two random variables defined on $\Omega = \Omega_X \times \Omega_Y$ and taking values x and y. The knowledge of the joint distribution function $F_{X,Y}(x, y)$ allows to recover the distribution function of X, respectively Y, by simply setting Y, respectively X, to infinity. The resulting functions $F_X(x)$ or $F_Y(y)$ are called marginal distribution functions. For a continuous rv:

$$F_X(x) = \int_{\Omega_Y} F_{XY}(x, y)dy, \tag{3.10}$$

$$F_Y(y) = \int_{\Omega_X} F_{XY}(x, y)dx. \tag{3.11}$$

For discrete rvs, the Riemann integrations of Eqs. (3.10) and (3.6) is replaced by a simple summation over $y \in \Omega_Y$ or $x \in \Omega_X$.

Of course, marginal density functions can be inferred from these marginal distribution functions by applying Eq. (3.2) to $F_X(x)$ or $F_Y(y)$ (for the discrete case, Eq. (3.6) should be used).

3.1.6 Statistical independence

Two events E_1 and E_2 are statistically independent if and only if:

$$P(E_1 \bigcap E_2) = P(E_1) \cdot P(E_2). \tag{3.12}$$

The independence of two events states that the knowledge of one of the event gives no information about the other event. This condition can be generalized to two random variables X and Y by defining the events $E_1 = \{X \leq x\}$ and $E_2 = \{Y \leq y\}$ for two real numbers x and y. Then, X and Y are statistically independent random variables - in shorthand notation, $X \perp\!\!\!\perp Y$ - if and only if:

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y). \tag{3.13}$$

An equivalent definition can be given using the joint distribution function, or the joint density function of X and Y:

- two rvs X and Y are independent if their joint distribution can be expressed as the product of their marginal distribution functions:

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y). \tag{3.14}$$

- Two rvs X and Y are independent if their joint probability density function can be expressed as the product of their marginal density functions:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y). \tag{3.15}$$

These definitions generalize to n random variables in a straightforward way.

3.1.7 Some additional remarks about the notation

In the remainder of this thesis, we will deal with discrete rvs. A rv is denoted by a capital letter, and its specific values by the same letter in lower case. The pdf of a discrete rv X defined on a sample space Ω_X is noted $p_X(x)$ rather than $f_X(x)$ to follow standard notation in information theory. Moreover, this notation is shortened for convenience as $p(x)$, when no confusion is possible. Thus $p(x)$ and $p(y)$ refer to two different rvs and are different pdfs: $p_X(x)$ and $p_Y(y)$ respectively.

3.2 Information theory: basic concepts and definitions

3.2.1 The concept of information theory in a few words

Information theory refers to a mathematical framework introduced by Shannon in the 50's and devoted, in its initial form, to communication systems [115]. Such systems have the task of transmitting information from one place to another as efficiently as possible, given certain costs and constraints.

The basic idea behind information theory is that it is a "measuring theory": it gives the means for evaluating how much measuring one thing tells us about another thing, previously unknown. This general formulation intuitively shows that there is no reason to restrict information theory to communication systems. On the contrary, it can find applications in many fields of physics. In particular, it has been applied with success to signal processing during these last years. An interesting presentation of the link between information theory and signal processing can be found in [26]. Recently, information theory has proved particularly useful in the development of unsupervised learning algorithms. Linsker for example has used this theory in neural networks with considerable success (see for example [82]).

This thesis deals with a pattern recognition task. It largely uses information theory to achieve its end. Therefore, some basic definitions and concepts widely used afterwards will be introduced now. A signal processing point of view is used rather than a communication system perspective.

3.2.2 Entropy

The two central concepts of information theory are those of entropy and information [115]. As intuitively expressed in the previous subsection, the idea behind information theory

is to know how the uncertainty about the state of the world has decreased after some measurements have been made. Of course, this depends on how uncertainty is measured: for this, information theory uses *entropy*. The entropy is a measure of the average uncertainty in the rv X [32].

For a discrete random variable X with pdf $p(x)$ defined for all possible outcomes x taken on the sample space Ω_X, the Shannon's entropy H is defined as:

$$H(x) = -\sum_{x \in \Omega_X} p(x) \log p(x), \tag{3.16}$$

where the convention $0 \cdot \log 0 = 0$ is used. If the logarithm is taken to base 2, this quantity is expressed in bits. In the following of this thesis, the notation log is used for $\log_2$.

$H(X)$ is also called the marginal entropy of X and satisfies the two following properties:

- $H(X) \geq 0$ with equality if and only if $p(x) = 1$ for one x.
- If Ω_X counts N elements, then the entropy is maximized when all the probabilities are equal to $1/N$ (uniform distribution). In this case, $H(X) = \log N$.

If some outcomes y of a discrete rv Y related to X are observed, the uncertainty about X, thus its entropy, is decreased. The conditional entropy $H(X|Y)$ measures the uncertainty remaining in X after having observed Y. If Y is defined on the sample space Ω_Y, with outcomes y and a pdf $p(y)$, the conditional entropy of X given Y is defined as:

$$H(X|Y) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log p(x|y), \tag{3.17}$$

where $p(x, y)$ is the joint pdf of X and Y and $p(x|y)$ the conditional pdf of X and Y. Conditioning then reduces entropy: $H(X|Y) \leq H(X)$ with equality if and only if X and Y are independent.

The chain rule for entropy relates the conditional and the marginal entropies as follows:

$$H(X) + H(X|Y) = H(Y) + H(Y|X). \tag{3.18}$$

The two summations in Eq. (3.18) are equal to the joint entropy $H(X, Y)$ between the two rvs X and Y. The joint entropy is defined as:

$$H(X, Y) = -\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log p(x, y). \tag{3.19}$$

3.2.3 Mutual information

As mentioned previously, the uncertainty about the rv X is decreased by the knowledge of another rv Y. This uncertainty reduction results obviously in a gain of information. How much information is gained depends on how much X and Y are dependent. This leads to the concept of MI:

$$I(X, Y) = H(X) - H(X|Y), \tag{3.20}$$

with $I(X,Y) = I(Y,X)$ (symmetry of MI) and $I(X,Y) \geq 0$.

Using the expression of $H(X)$ and $H(X|Y)$ as stated by Eqs. (3.16) and (3.17), the mutual information can be defined as (see [32] for proof):

$$I(X,Y) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{3.21}$$

Mutual information thus measures the average amount of information that Y conveys about X. It can equivalently be understood as a measure of the dependence between X and Y.

3.2.4 Data processing inequality

Most of the systems require the data to be preprocessed before to be analyzed. A fundamental theorem of information theory, the data processing theorem, states that such processing on the data cannot increases their information.

Theorem - *(*Data processing inequality) - If X, Y, and Z are three rv forming a Markov chain $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Corollary: In particular, if Z is a function g of Y: $Z = g(Y)$, $X \to Y \to g(Y)$ forms a Markov chain, thus $I(X;Y) \geq I(X;g(Y))$.

Multimodal feature extraction and speaker detection framework

4

4.1 Introduction

In this thesis, the detection of the current speaker in an audio-visual sequence is understood as a pattern recognition problem, cast in a statistical framework. As mentioned in the previous chapter, a pattern recognition process counts five basic stages, from the data acquisition to the system evaluation. The schematic representation of such a pattern recognition chain, presented in chapter 2, is reminded in Fig. 4.1.

At each step, errors may occur which pass on to the next stages, resulting possibly in poor process performance. The ultimate goal when designing a pattern recognizer is obviously to minimize its probability of error. In this chapter, we will propose and discuss a framework which considers all the stages of the chain presented in Fig. 4.1, from the feature extraction to the evaluation stages. The data acquisition and initial data representation depend directly on the task at hand and will thus be tackled when addressing the concrete application of the framework, in the next chapters. The presented system aims at minimizing the overall error probability of the process, as assessed by a performance evaluation step at the end of the recognition chain. For that purpose, a multimodal approach is taken from the feature level.

This chapter starts by discussing the concept of error associated to a statistical clas-

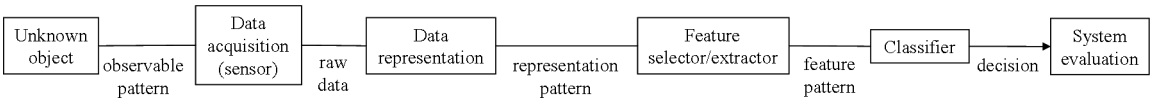


Figure 4.1 — Pattern classifier.

sification process. In the next section, we develop an information theoretic framework for performing feature extraction using a multimodal approach. Such an approach is shown to efficiently decrease the probability of committing an error at this processing stage. It implies the problem to be cast in a statistical framework. Therefore, density estimation methods, with a particular focus on non-parametric approaches, are briefly introduced in section 4.3. In section 4.4, a classifier is defined and justified using decision theory. Specifically, the problem is expressed through hypothesis tests. This formulation gives means for evaluating the performance of the classifier as well as of the whole pattern recognition system. In particular, it allows to validate the gain offered by the multimodal feature extraction step.

4.2 Information theoretic framework for pattern classification

4.2.1 Unimodal classification process

Let us consider a general unimodal classification task. The original class C as well as the signal, or source, S are not directly observable. Using some sensors, some features X are generated. In a next processing stage, some particular variables F_X are selected or extracted from the initial data space. This transformation is necessary as the original space of values is usually not suited for classification. From these features, an estimate $\hat{S}$ of the original signal S is inferred and a decision about the class label, $\hat{C}$ is finally taken.

From a Bayesian perspective, C, S, X, F_X, $\hat{S}$, and $\hat{C}$ are random variables. C and $\hat{C}$ take value on Ω_C, the set of possible class labels; S and $\hat{S}$ are defined on the sample space Ω_S; X and F_X on sample spaces Ω_X and Ω_{F_X}. The classification procedure leading to an estimate $\hat{C}$ is then described by a first order Markov chain (MC) shown in Fig. 4.2. The non-observable states C and S are hidden states. Using the chain rule of conditional

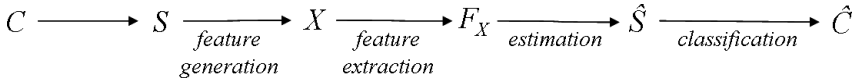


Figure 4.2 — First order Markov chain describing the classification procedure.

probability, the joint probability characterizing this Markov chain is:

$$P(C, S, X, F_X, \hat{S}, \hat{C}) = P(C)P(S|C)P(X|S)P(F_X|X)P(\hat{S}|F_X)P(\hat{C}|\hat{S}). \tag{4.1}$$

4.2.2 Probability of error on the unimodal classification process

As previously mentioned, the ultimate goal when designing a classifier is to minimize the probability of assigning the wrong class to the signal. The probability of committing an error during the classification process, $P_E = P(\hat{C} \neq C)$, depends on the classifier itself, i.e. on the discrimination power of the decision rule estimating $\hat{C}$ from $\hat{S}$. Actually, given an observation $\hat{s}$ of $\hat{S}$, a decision has to be taken about the value of C. This decision

can be modelled as a function $g : \Omega_S \rightarrow \Omega_C$, leading to an estimate $\hat{C} = g(\hat{S})$ [35]. The ultimate goal is to design the classifier such that the probability of error $P_E = P(g(\hat{S}) \neq C)$ is minimized. However, the Markov chain of Fig. 4.2 clearly shows that whatever the classifier, its performance will be poor if the feature F_X extracted from X are not suited, resulting in a poor estimate $\hat{S}$. In fact, the estimation of S from F_X is a function, as is the estimation of C from $\hat{S}$. Since there are two hidden variables, C and S, there are also two estimation processes whose probability of error must be minimized. From here onwards, we call "estimation process" the steps leading from S to $\hat{S}$ on the MC of Fig. 4.2 while "classification process" refers to the whole chain. Their error probabilities are designed by P_e and P_E respectively.

Let us consider firstly the estimation process, where the hidden state S is estimate from some collected data. A probability of error can be defined on this process: $P_e = P(\hat{S} \neq S)$. Let us introduce a binary random variable Γ taking value on $\Omega_\Gamma = \{0, 1\}$ and used as a predicate for the wrong estimation of S:

$$\Gamma = \begin{cases} 1 & \text{if } \hat{S} \neq S, \\ 0 & \text{if } \hat{S} = S. \end{cases} \tag{4.2}$$

The probability of error on the estimation process is then defined as $P_e = P(\Gamma = 1|\hat{S}, S)$. Introducing this error function into the estimation process, the Markov chain of Fig. 4.2 becomes the Bayesian network shown in Fig. 4.3. It is not a first order Markov chain

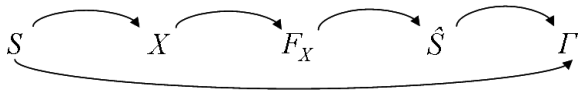


Figure 4.3 — Bayesian network describing the estimation process, with a predication step on the error probability included.

anymore since the error probability is conditioned by both the input S of the process and its final output $\hat{S}$ [28]. The joint probability characterizing this Bayesian network is:

$$P(S, X, F_X, \hat{S}, \Gamma) = P(S)P(X|S)P(F_X|X)P(\hat{S}|F_X)P(\Gamma|S, \hat{S}). \tag{4.3}$$

The expected error associated to the process is given by the expectation E[Γ] of Γ [32]:

$$P_e = P(\hat{S} \neq S) \tag{4.4}$$

$$= \int_{s\in\Omega_S} \int_{x\in\Omega_X} \int_{f_x\in\Omega_{F_X}} \int_{\hat{s}\in\Omega_S} p(s)p(x|s)p(f_x|x)p(\hat{s}|f_x)P(\Gamma = 1|\hat{s}, s). \tag{4.5}$$

Assuming the rvs to be discrete as it is the case when working with digital signals, Eq. (4.6) becomes:

$$P_e = \sum_{s\in\Omega_S} \sum_{x\in\Omega_X} \sum_{f_x\in\Omega_{F_X}} \sum_{\hat{s}\in\Omega_S} p(s)p(x|s)p(f_x|x)p(\hat{s}|f_x)P(\Gamma = 1|\hat{s}, s). \tag{4.6}$$

Let us now consider the whole classification chain leading from C to $\hat{C}$ as shown in Fig. 4.2. An indicator function for the misclassification of $\hat{C}$ can be introduced as was done for the estimation process. Let E be a binary rv defined on $\Omega_E = \{0, 1\}$ and such that:

$$E = \begin{cases} 1 & \text{if } \hat{C} \neq C, \\ 0 & \text{if } \hat{C} = C. \end{cases} \tag{4.7}$$

Introducing the two error indicators in the model, this classification process is now modelled by the Bayesian network shown in Fig. 4.4. The associated probability of classification error

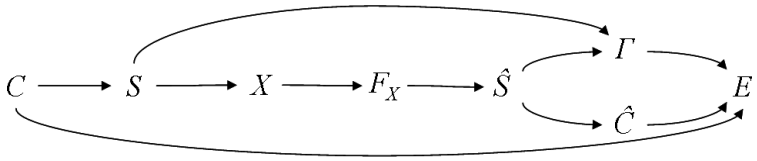


Figure 4.4 — Bayesian network describing the classification process, when the predicates for the estimation and the classification error probabilities are introduced.

is given by:

$$P_E = P(\hat{C} \neq C), \tag{4.8}$$

$$= \sum_{c,\hat{c}} \sum_{s,\hat{s}} \sum_{x} \sum_{f_x} \sum_{\gamma} p(c)p(s|c)p(x|s)p(f_x|x)p(\hat{s}|f_x)p(\gamma|\hat{s},s)p(\hat{c}|\hat{s})p(E=1|\hat{c},c,\gamma). \tag{4.9}$$

The migration of the summations to the left of each conditional probability they do reference leads to:

$$P_E = \sum_{c\in\Omega_C} p(c) \sum_{s\in\Omega_S} p(s|c) \sum_{x\in\Omega_X} p(x|s) \sum_{f_x\in\Omega_{F_X}} P(f_x|x) \sum_{\hat{s}\in\Omega_S} p(\hat{s}|f_x) \sum_{\gamma\in\Omega_\Gamma} p(\gamma|\hat{s},s) \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s})p(E=1|\hat{c},c,\gamma). \tag{4.10}$$

Using the Bayes rule, the following relationships hold:

$$\sum_{c\in\Omega_C} p(c) \sum_{s\in\Omega_S} p(s|c) = \sum_{c\in\Omega_C} \frac{p(c)}{p(c)} \sum_{s\in\Omega_s} p(s,c) \tag{4.11}$$

$$= \sum_{c\in\Omega_c} \sum_{s\in\Omega_s} p(s,c) \tag{4.12}$$

$$= \sum_{s\in\Omega_s} p(s), \tag{4.13}$$

where the step from Eq. (4.12) to Eq. (4.13) uses the definition of marginal distribution (sec. 3.2). Eq. (4.10) is then re-written as:

$$P_E = \sum_{s} p(s) \sum_{x} p(x|s) \sum_{f_x} p(f_x|x) \sum_{\hat{s}} p(\hat{s}|f_x) \sum_{\gamma} p(\gamma|\hat{s},s) \sum_{\hat{c}} p(\hat{c}|\hat{s})p(E=1|\hat{c},c,\gamma). \tag{4.14}$$

Or, by developing the summation term over γ:

$$P_E = \sum_{s\in\Omega_S} p(s) \sum_{x\in\Omega_X} p(x|s) \sum_{f_x\in\Omega_{F_X}} p(f_x|x) \sum_{\hat{s}\in\Omega_S} p(\hat{s}|f_x) \sum_{\gamma\in\Omega_\Gamma} p(\gamma=1|\hat{s},s) \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s}) p(E=1|\hat{c},c,\gamma)$$
$$+ \sum_{s\in\Omega_S} p(s) \sum_{x\in\Omega_X} p(x|s) \sum_{f_x\in\Omega_{F_X}} p(f_x|x) \sum_{\hat{s}\in\Omega_S} p(\hat{s}|f_x) \sum_{\gamma\in\Omega_\Gamma} p(\gamma=0|\hat{s},s) \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s}) p(E=1|\hat{c},c,\gamma). \quad (4.15)$$

Finally, we end up with a formulation of the error probability of the classification process which involves the probability of estimation error:

$$P_E = P(\hat{S}\neq S) \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s}) P(E=1|\hat{c},c,\Gamma=1) + (1-P(\hat{S}\neq S) \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s}) p(E=1|\hat{c},c,\gamma=0),$$
$$= P_e \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s}) p(E=1|\hat{c},c,\gamma=1) + (1-P_e) \sum_{\hat{c}\in\Omega_C} p(\hat{c}|\hat{s}) p(E=1|\hat{c},c,\gamma=0). \quad (4.16)$$

This relationship could be further simplify of course. But the objective of these developments was to come to a formulation which shows how errors caused by each estimation step (steps leading to $\hat{S}$ and to $\hat{C}$) impact on the overall error of the classification process.

On one hand, if the estimation step leading from $\hat{S}$ to S is optimal ($P_e \approx 0$), Eq. (4.16) reduces to its right-most hand term. It shows that in this case the minimization of P_E relies on the choice of an accurate classifier which minimizes $p(E=1|\hat{c},c,\gamma=0)$. On the other hand, if the most accurate classifier is found for the problem at hand, neglecting to consider the steps leading from S to $\hat{S}$, the process error probability might still be high due to P_e.

4.2.3 Information theoretic feature extraction

Letting aside for now the question of choosing the most suitable classifier for the problem at hand (thus assuming this classifier exists and has been picked up), we will focus in this part on minimizing the error probability P_e ensuing from the estimation process.

The estimation of one rv from another can be understood as a feature extraction step where some specific information must be recovered from the initial rv. Therefore, the information theoretic framework developed in [45] for extracting features can be applied. Using Fano's inequality, the probability of committing an error when getting an estimate $\hat{S}$ of the discrete rv S from another rv F_X can be related to the conditional entropy $H(S|F_X)$ [32]:

$$P_e \geqslant \frac{H(S|F_X)-1}{\log|\Omega_S|} = \frac{H(S)-I(S,F_X)-1}{\log|\Omega_S|}, \quad (4.17)$$

where $H(S)$ and $I(S,F_X)$ are the entropy of S and the mutual information between the rvs S and F_X respectively, and $|\Omega_S|$ is the cardinality of the domain of S.

The inequality (4.17) does not help us directly to minimize the error probability P_e. It does indicate however that an efficient minimization of P_e is conditioned by the minimization of the right hand side of the inequality, i.e. by a maximization of the mutual information

between S and F_X. In other words, a constraint is introduced on the feature extraction step that might improve this stage. Indeed, the conditional entropy $H(S|F_X)$ corresponds to the information present in S but missing from $\hat{S}$, which may be required for a correct classification of $\hat{S}$. If the mapping is deterministic, this conditional entropy has its minimal possible value.

4.2.4 Extension to the multi-modal case

The framework is presented here in the context of speaker detection but it can of course apply in any situation where two signals of different modalities are jointly emitted by a hidden multimodal source. In the particular problem of speaker detection, audio and video signals are jointly emitted during the speech production process and these two modalities can be used to constrain the feature extraction step.

Let C be a binary random variable which models the membership to the "speaker" or "non-speaker" class with respect to an audio-visual source modelled by the random variable S, defined on Ω_S. Notice that the probability for any prototype to belong to each class is the same, and is equals to $1/|\Omega_C|$, where $|\Omega_C|$ is the cardinality of Ω_C ($|\Omega_C| = 2$). The bimodal source S is not directly accessible but yields two observed signals of different physical nature: the audio and video signals A and V. For each of those signals, the unimodal classification process leading from the measurements A - respectively V - to an estimate $\hat{C}_1$ - respectively $\hat{C}_2$ - of the class, can be described through a first order Markov chain (Fig. 4.5.(a)), as previously discussed. Two probabilities of classification error, P_{E1} and P_{E2}, with the corresponding two lower bounds can be derived for each Markov chain.

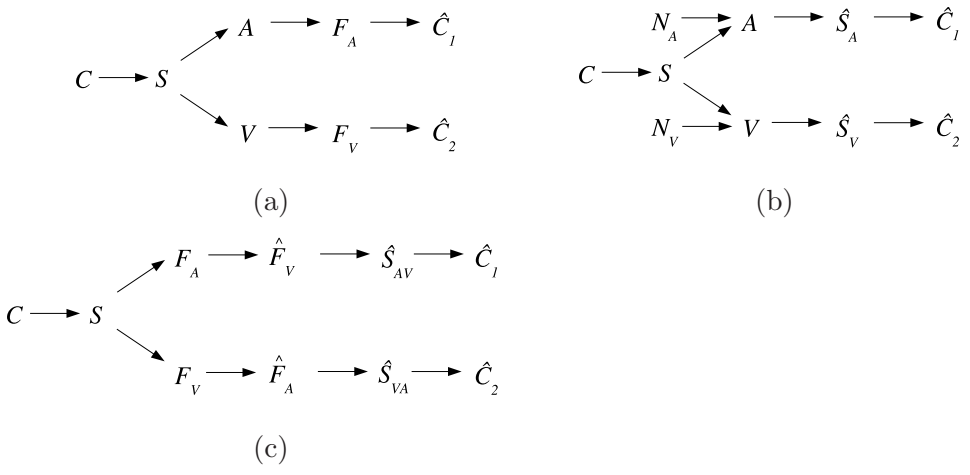


Figure 4.5 — (a) Graphical representation of the audio and video Markov chains (leading to $\hat{C}_1$ and $\hat{C}_2$ respectively) modelling the two unimodal classification processes associated to each modality; (b) Graphical representation of the Bayesian networks modelling the two unimodal classification processes associated to each modality, including the interfering sources; (c) Graphical representation of the related Markov chains modelling the multimodal classification process.

A possibility is to firstly obtain some estimates $\hat{S}_A$ and $\hat{S}_V$. Then a fusion at the decision or at the classification level can be performed in order to get a unique estimate $\hat{C}$ of the class from both unimodal processes. However, such an approach would not take advantage of the discriminant information offered by the bimodal nature of the source S. A better approach would be to use an extension of the previously described unimodal feature extraction framework to the multimodal case. Such an extension has been proposed by Butz *et al.* in [28] and applied more particularly to image registration. A similar multimodal feature extraction framework is proposed. It is extended to speaker detection and the various processing steps are completely justified.

As already stated, the original source S is accessible only through the measurements A and V. But, as mentioned in [44], these two measurements are corrupted by independent interference sources N_A and N_V. The signals coming from these sources account here for noise since they do not contain any information shared by both modalities. The classification process is then described through two Bayesian networks as shown on Fig. 4.5.(b). A good estimate of the source should include a feature extraction step which discards this noisy information present in each modality and recovers the information coming from the common source S, thus shared by both modalities. Obviously, such a goal can only be reached by considering the two modalities jointly. Let F_A and F_V be rvs modelling such audio and video features that contain only the information coming from the source S. Since they specifically describe this common source, they are related by their joint pdf $p(F_A, F_V)$. Thus an estimate of the feature related to one modality can be inferred from the other modality with transition probability $p(\hat{F}_V|F_A)$ or $p(\hat{F}_A|F_V)$. These transition probabilities can be obtained by joint probability estimation since $p(\hat{F}_V|F_A) = p(\hat{F}_V, F_A)/p(F_A)$ and $p(\hat{F}_A|F_V) = p(\hat{F}_A, F_V)/p(F_V)$, and if $\hat{F}_V$ and $\hat{F}_A$ are correctly estimated the approximation $p(\hat{F}_A, F_V) \approx p(F_A, \hat{F}_V) \approx p(F_A, F_V)$ can be assumed. These considerations lead to the definition of the classification problem with the two Markov chains shown in Fig. 4.5.(c). Notice that the source estimates associated to each chain are indexed by AV or VA, to stress that they have been obtained using information present in both modalities, in contrast with the previous case (Fig. 4.5.(a)).

For each Markov chain leading from S to either S_{AV} (audio MC) or S_{VA} (video MC), the probability of estimation errors P_{e_1} or P_{e_2} and their associated lower bounds are still defined according to inequality (4.17). For the audio MC for example, the inequality is stated as:

$$P_{e_1} = P(\hat{S}_{AV} \neq S), \tag{4.18}$$

$$P_{e_1} \geqslant \frac{H(S) - I(S, \hat{F}_V) - 1}{\log |\Omega_S|}. \tag{4.19}$$

From the data processing inequality, we have:

$$I(S, \hat{F}_V) \leqslant I(F_A, \hat{F}_V) \quad \text{and} \quad I(S, \hat{F}_A) \leqslant I(F_V, \hat{F}_A). \tag{4.20}$$

As a result, the bounds on the error probabilities can be weakened:

$$P_{e_1} \geqslant \frac{H(S) - I(F_A, \hat{F}_V) - 1}{\log |\Omega_S|}, \tag{4.21}$$

$$P_{e_2} \geqslant \frac{H(S) - I(F_V, \hat{F}_A) - 1}{\log |\Omega_S|}. \tag{4.22}$$

As previously said, $p(\hat{F}_A, F_V) \approx p(F_A, \hat{F}_V) \approx p(F_A, F_V)$. Similarly $I(F_A, \hat{F}_V) \approx I(\hat{F}_A, F_V) \lessapprox I(F_A, F_V)$. Introducing this approximation in Eqs. (4.21) and (4.22), and because of the symmetry property of MI, a joint lower bound can finally be defined:

$$P_{\{e1,e2\}} \geqslant \frac{H(S) - I(F_A, F_V) - 1}{\log |\Omega_S|}. \tag{4.23}$$

Minimizing the lower bound on $P_{\{e_1,e_2\}}$ then amounts to maximizing the mutual information between the extracted features F_A and F_V corresponding to each modality. The feature sets resulting from the maximization of the MI involved in these equations are expected to compactly describe the relationship between the two modalities. In that sense, the extraction stage produces optimized features.

To get a source estimate with a probability of estimation error close to this bound, a suitable estimator must be found. If F_A and F_V are correctly estimated, then they compactly describe the source S. However, for this last statement to be true, not only the mutual information $I(F_A, F_V)$ between features extracted from each modality must be increased, but also the conditional entropies $H(F_V|F_A)$ and $H(F_A|F_V)$ must be minimized. Indeed, if the entropies increase, they reduce the inter-feature dependencies. Dividing Eq. (4.23) by the joint entropy $H(F_A, F_V)$, a feature efficiency coefficient (EC) [28] can be defined:

$$e(F_A, F_V) = \frac{I(F_A, F_V)}{H(F_A, F_V)} \in [0, 1]. \tag{4.24}$$

This coefficient defines our estimator. Since $I(F_A, F_V) = H(F_A) + H(F_V) - H(F_A, F_V)$, maximizing $e(F_A, F_V)$ still minimizes the lower bound on the error probability defined in Eq. (4.23) while constraining inter-feature independence. In other words, the extracted features F_A and F_V will tend to capture just the information related to the common origin of A and V, while discarding the unrelated interferences coming from N_A and N_V: they lead to an estimate of the source S.

The feature extraction framework presented in this section has been published in [18] and [19].

4.2.5 Optimization problem

The multimodal feature extraction framework just presented comes up with the definition of an estimator named efficiency coefficient (Eq. (4.24)). This coefficient has to be maximized for the probability of estimation error to be decreased.

The precise definition of the objective function and the optimization method leading to the solution will be presented in the next two chapters, where the framework is applied to the audio and video signals. Let us assume for now that the features have been extracted optimally.

4.2.6 Classifier definition

Applying this framework to extract features, the estimation error probability comes closer to its minimum. However, the classification error probability P_E must still be minimized: this depends on the choice of a suitable classifier. A classifier derived from decision theory (hypothesis testing in particular) will be defined later on. It will be shown how this specific approach allows us to evaluate the classification process performance as well. In particular, the added value of the feature extraction step will be appraised.

4.3 Density estimation

4.3.1 Presentation of the parametric *versus* non-parametric approaches

The estimator of Eq. 4.24 which has to be maximized, involves the mutual information between the rvs F_A and F_V, modelling the extracted audio and video features. This mutual information is then defined as:

$$I(F_A, F_V) = \sum_{f_a \in \Omega_{F_A}} \sum_{f_v \in \Omega_{F_V}} p(f_a, f_v) \log \frac{p(f_a, f_v)}{p(f_a)p(f_v)}, \tag{4.25}$$

where f_a and f_v denote the outcomes of the rvs F_A and F_V, defined on Ω_{F_A} and Ω_{F_V}. It follows from Eq. (4.25) that an estimation of the marginal and joint pdfs of F_A and F_V is required for solving the optimization problem. These pdfs are to be inferred from a given sample, using either a parametric or non-parametric approach (a random sample is defined as the collection of independent observed values of a random variable [2]).

Parametric methods assume that the current sample is drawn from a probability density belonging to a specific family. The estimation of the pdf then reduces to the estimation of the parameters of this particular density. Various statistical methods allows to find these parameters, such as the method of moments or the maximum likelihood [2]. If the true form of the density function is known for sure, a parametric approach is obviously the most efficient and reliable method. However, in most case, the underlying density of real data rarely fits common density models so that a bias persists.

Non-parametric techniques, on the contrary, do not make any assumption about the family of the distribution from which the sample is drawn. If some parameters have to be defined, their specification does not uniquely determine a member of a family. Of course, they are less effective if the density of the data can be reasonably approximated by a parametric model. But they allow to approximate a large class of unknown densities and can reveal skewness in distribution, or the presence of multiple nodes in the data.

In the problem at hand, the pdfs to be estimated are those of *a priori* unknown features (since they have to be extracted solving the optimization problem). Their distribution is, then *a fortiori*, unknown. These considerations lead to the choice of a non-parametric approach for estimating the densities.

In the sequel, non-parametric techniques used in this work will be presented through the analysis of a d-variate random sample of n outcomes, $(\vec{x_1}, \ldots, \vec{x}_n)$, drawn from an unknown density function, $f(\vec{x})$, where $\vec{x} \in \mathbb{R}^d$. In a first time, $d = 1$.

4.3.2 Histogram estimator

There exist many non-parametric techniques to estimate a pdf. For a comprehensive coverage of the subject, the reader is referred to [118] for example.

Let us start with the definition of the histogram estimator, which is the most natural density estimator when basing the estimation on the solely knowledge of the data set. The sample space Ω_X is first divided into a number of intervals, or bins, of width h. Then the density estimate at point x is defined as:

$$\hat{p}(x) = \frac{s_k}{n \cdot h}, \tag{4.26}$$

where s_k denotes the number of outcomes falling into bin B_k.

The advantages of the histogram estimator are the simplicity of its concept and its computational efficiency. However, some limitations can be pointed out as well [23]:

1. A loss of information results from the replacement of the data points x_i by the central point of the interval in which it falls.

2. It is not a smooth estimator due to the sharp edges of the intervals from which it is build.

3. The bin width and the bin origin affect the behavior of the estimator.

Some improved histogram estimator techniques have been developed to alleviate these limitations. For example, variable bandwidths can be used to limit the influence of the bin width and bin origin on the estimate. It is also possible to convolves the histogram with a Gaussian to get smoother density estimate. The average shifted histogram (ASH) introduced by Scott *et al.* in [113], has been developed to reduce the influence of the bin origin on the estimation. ASH basically consists in estimating a density by averaging m histograms shifted by $\delta = h/m$ from the previous mesh.

As the number of shifts $m \to \infty$, ASH approximate the kernel estimator, or Parzen-window estimator [97]. The latter also alleviates the first two limitations associated to the histogram estimator and is then detailed in the next subsection.

4.3.3 Kernel estimator

In the kernel, or Parzen-window, estimator, the bins are replaced by smoothing functions directly centered over each observation. In other words, it can be viewed as a convolution on all the n points. More precisely, the Parzen-window density estimator is defined as:

$$\hat{p}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i), \tag{4.27}$$

where $K_h(x) = h^{-1}K(h^{-1}x)$. $K(x)$ is itself a probability density called kernel or Parzen-window, whose variance is controlled by the parameter h. Because of its role in determining how the probability associated with each observation is spread over the surrounding sample space, h is called the smoothing parameter or bandwidth [23].

The choice of the kernel is not too important and is more motivated by practicality consideration. Hence, it is often convenient to use for K a centered normal density function $K \sim \mathcal{N}(0, \Sigma)$:

$$\hat{p}(x) = \frac{1}{\sqrt{2\pi h}} \exp^{-\frac{(x-x_i)^2}{2h}}. \tag{4.28}$$

In such a case, the extension of the kernel estimator to vector-valued data $\vec{x} \in \mathbb{R}^d$, with $d > 1$, is straightforward. Instead of using a multivariate kernel, the density estimate can indeed be written as the product of univariate components [113]:

$$\hat{p}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{h_k}(\vec{x} - \vec{x}_i), \tag{4.29}$$

$$\hat{p}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{k=1}^{d} K_{h_k}(x^{(k)-x_i^{(k)}})\right], \tag{4.30}$$

where $K_h(\vec{x}) = h^{-d}K_h(\vec{x}/h)$ and h_k denotes the smoothing parameter in the $k-$th direction.

The Kernel estimator has been chosen and is used throughout this thesis: as it derives naturally from the histogram, it presents the same advantages (in particular, the simplicity of its concept and the straightforward computation) while alleviating the major limitations of this estimator. One difficulty however is common to both histogram and kernel estimators: it is the choice of the bandwidth - or smoothing parameter - h. The role played by this parameter is very similar for the two approaches and will now be discussed.

4.3.4 Role of the smoothing parameter

Non-parametric density estimations are often confronted to sparse data and this can limit the estimation accuracy. This problem is often referred to as the curse of dimensionality [9]: the number of data points required for a correct estimation of the underlying density structure increases exponentially with the feature dimensionality d.

As it fixes the width of the function kernel, the smoothing parameter h controls how each observation is spread over the surrounding sample space [23]. This allows the kernel

estimator to be less sensitive to data sparsity for revealing the underlying structure of the distribution, under the condition that h is correctly set up. If h is too large, the resulting density is oversmoothed and multiple modes might be missed. If h is too small, fine structures corresponding to individual observations will appear instead of the underlying structure of the whole sample.

Since the pdf estimates appear in the objective function, h impacts finally on the objective function smoothness. From the optimization point of view, for too small values of h, the objective function is likely to be highly irregular, with a negative impact on the optimization algorithm. On the other hand, too large values of h may result in a loss of information. In particular, the loss of discrimination between the densities can be dramatic and lead to a wrong solution.

4.4 Classification through hypothesis tests

4.4.1 Classifier function based on audio-visual synchrony evaluation

Maximizing the efficiency coefficient defined by Eq. (4.24) produces features optimized for the classification problem at hand, i.e. for the speaker detection task. As mentioned in sec. 4.2, these features are such that the probability of estimation error is decreased. A classifier able to classify them as correctly as possible must now be defined. This way, the probability of classification error is minized, as shown by Eq. (4.16).

The classifier has to assign a "speaker" or "non-speaker" label to pairs of audio and video features. Previous works in the domain have established that evaluating the synchrony, or temporal co-occurrence, between the audio and video measurements is a good way of classifying them as originating from a common audio-visual source or not [120], [88]. In some other works, mutual information shows good performance in detecting synchronized audio-visual sources such as speakers [44], [27], [95]. Finally, Hershey *et al.* in [61] interpret synchrony as the degree of mutual information between audio and video signals. Consequently, all these studies point out the MI as a good measure for classifying candidate audio and video features as "speaker" or "non-speaker". The choice of an MI-based classifier function is also very coherent with the previous feature extraction step which involves an MI-based function as well.

Therefore, the classifier evaluates the MI between the audio features extracted from the audio signal, and the video features extracted from each mouth region presents in the image. The mouth region whose features lead to the largest MI is labelled as "speaker". Only one "speaker" class label is authorized per estimation so that the other mouth regions are labelled as "non-speaker". Such a classifier has also the advantage of fusing the bimodal information at the classification level resulting in a unique class estimate $\hat{O}$.

In the case where two candidate mouth regions, or speakers, are present, the classifier function previously defined can be expressed equivalently as an evaluation of the difference of MI between the audio features and each mouth region video features. The sign of the

difference indicates the video speech source.

The mutual information is a metric evaluating the degree of dependence between two rvs. Its use as a decision function has been justified in classification problems that can be formulated as hypothesis tests which ask about the statistical dependence or independence between the features [66], [44]. The previous classifier function will now be defined and justified using a hypothesis testing scheme, as we proposed in [12] and [13]. The objective however is to exploit the potential offered by hypothesis tests for performance evaluation. In particular, the benefit of performing a feature extraction step previous to the classification step can be assessed.

4.4.2 Hypothesis testing and the Neyman-Pearson lemma

Hypothesis tests are used in detection problems in order to take the most appropriate decision given an observation x of a rv X. In current context, the decision function has to choose which of m mouth regions extracted from a video sequence is the source of the simultaneously recorded acoustic speech signal.

Let set m to 1 in a first time. As previously stated, the decision can be taken based on the evaluation of the synchrony, or dependence relationship, that exists between the measured audio sample f_a associated to the rv F_A and the video sample f_v associated to the rv F_V. F_A models the audio features while F_V models the features associated to the mouth region. This statement can be formulated through two mutually exclusive statistical hypothesis:

$$\begin{aligned} H_0 &: f_a, f_v \sim p_{H_0}(f_a, f_v) = p(f_a) \cdot p(f_v), \\ H_1 &: f_a, f_v \sim p_{H_1}(f_a, f_v) = p(f_a, f_v). \end{aligned} \tag{4.31}$$

Obviously, the null hypothesis H_0 postulates the data to be governed by a pdf stating the independence between the audio and the video features: the mouth is not the visual speech source. The statistical dependence is stated by the alternative hypothesis H_1: the mouth is speaking and can be associated to the acoustic speech.

The Neyman-Pearson approach to hypothesis tests consists in formulating certain probabilities associated with a binary hypothesis test [91]. The false-alarm probability P_{FA}, or size α of the test*, is defined as:

$$\alpha = P(\hat{H} = H_1 | H = H_0), \tag{4.32}$$

while the detection probability P_D, or power β of the test, is given by:

$$\beta = P(\hat{H} = H_1 | H = H_1). \tag{4.33}$$

The probability of missed detection† is equal to $1 - \beta$.

*Also called type I error.

†Also called the type II error. Notice that β refers often to the probability of missed detection instead of P_D, given by $1 - \beta$ in such a case.

The Neyman-Pearson criterion selects the most powerful test of size α: the decision rule should be constructed so that the probability of detection β is maximal while the probability of false-alarm α does not exceed a given value. Then the Neyman-Pearson lemma states that the best test of size α for testing H_0 against H_1 is:

$$\phi(f_a, f_v) = \begin{cases} 1 & \text{if } p_{H_1}(f_a, f_v) > \nu \cdot p_{H_0}(f_a, f_v), \\ \gamma & \text{if } p_{H_1}(f_a, f_v) = \nu \cdot p_{H_0}(f_a, f_v), \\ 0 & \text{if } p_{H_1}(f_a, f_v) < \nu \cdot p_{H_0}(f_a, f_v), \end{cases} \tag{4.34}$$

for some $0 \leq \gamma \leq 1$ and a threshold $\nu > 0$. The test can be equivalently expressed using the likelihood ratio test:

$$\lambda(f_a, f_v) = \frac{p_{H_1}(f_a, f_v)}{p_{H_0}(f_a, f_v)} \gtreqless \nu. \tag{4.35}$$

Using the log-likelihood ratio, the Neyman-Pearson test can be expressed as:

$$\Lambda(f_a, f_v) = \log \left[\frac{p_{H_1}(f_a, f_v)}{p_{H_0}(f_a, f_v)} \right] \gtreqless \eta. \tag{4.36}$$

The test function must then decide which of the hypothesis H_1 or H_0 is the most likely to describe the probability density functions of the observations (f_a, f_v), by finding the threshold η that will give the best test of size α.

As the number of observations f_a and f_v grows, the normalized log-likelihood ratio approaches its expected value and becomes equal to the mutual information between the random variable F_A and F_V [66]. Indeed, as previously said, the mutual information is a metric evaluating the distance between two joint distributions: one stating the dependence of the variables and the other, the independence of those same variables.

$$E \left\{ \frac{1}{|\Omega_{F_A}| \cdot |\Omega_{F_V}|} \sum_{f_a \in \Omega_{F_A}} \sum_{f_v \in \Omega_{F_V}} \Lambda(f_a, f_v) \right\} = I(F_A, F_V). \tag{4.37}$$

Thus the test function becomes a simple evaluation of the mutual information between audio and video random variables, with respect to a threshold η:

$$I(F_A, F_V) = \sum_{f_a \in \Omega_{F_A}} \sum_{f_v \in \Omega_{F_V}} \left[p(f_a, f_v) \log \left(\frac{p(f_a, f_v)}{p(f_a) \cdot p(f_v)} \right) \right] \gtreqless \eta. \tag{4.38}$$

To sum up, the Neyman-Pearson lemma usually applies as follows:

1. Define the null and the alternative hypothesis of the binary test.
2. Select the relevant size α of the test for the problem at hand.
3. Calculate the threshold η, which is a function of α.

The threshold is selected in order to fix the desired size for the test. It is a big issue and often, this cannot be done analytically.

Increasing the threshold increases the probability that an individual identified as "speaker" really is the speaker (β is increased) and decreases the probability that an individual identified as "speaker" is not really the speaker (α is decreased), and vice versa. Or at least, as η increases, β should increase quicker than α for the MI-based classifier to perform well.

A precise analysis of the classifier performance should however consider more advanced criterion such as, for example, how quicker β increases compared to α. The Receiver Operator Characteristic (ROC) curve is a powerful means of characterizing and visualizing the performance of a two-classes discrimination rule in order to select a suitable decision threshold [141]. A short introduction to ROC curve is now given.

4.4.3 Performance evaluation using ROC analysis

A ROC graph is obtained by plotting the detection probability β against the false-alarm probability α as the threshold η is varied. This crossplot allows to evaluate the ability of a classifier to produce good relative instance scores [42]. It gives a visual representation of the fundamental trade-off in hypothesis testing and decision theory: to increase β, α must also be allowed to increase.

Let us firstly introduce some terms specific to decision theory. The rejection of the null hypothesis H_0 in favor of H_1 corresponds to a classification of the outcome pairs (f_a, f_v) in the positive class (the "speaker" class), while the acceptance of H_0 means classifying (f_a, f_v) in the negative ("non-speaker") class. Given these definitions, some particular points attached to the ROC space can be noted:

- The upper left corner point $(0, 1)$ corresponds to the ideal classifier.
- The points $(0, 0)$ and $(1, 1)$ correspond to classifiers that classify all examples as negative and positive respectively.

The closest a classifier to the point $(0, 1)$, the best it is. Classifiers appearing on the bottom left hand-side of a ROC graph are said to be conservatives. Those appearing on the upper right-hand side are liberal [42]. A random classifier lies on the diagonal line $\alpha = \beta$. A classifier is potentially optimal under some cost model if and only if it lies on the northwest boundary (i.e. above the line $\alpha = \beta$) of the convex hull of the set of points in ROC space [105]. More specifically, since the Neyman-Pearson strategy is to maximize the hit rate β for a fixed false-alarm rate, the more vertical the slope in the conservative part, the best the classifier.

Another interesting quantity to define is the Area Under the Curve (AUC $\in [0, 1]$), as it allows to reduce the ROC performance to a simple scalar value. It is a measure of the class discrimination capability of the classifier: the greater it is, the best the classifier. However, if the purpose is to compare two classifiers, this measure must be handled with care. It is only when one ROC curve clearly dominates another over the entire performance space that the corresponding classifier can be said to be better. When the ROC curves cross,

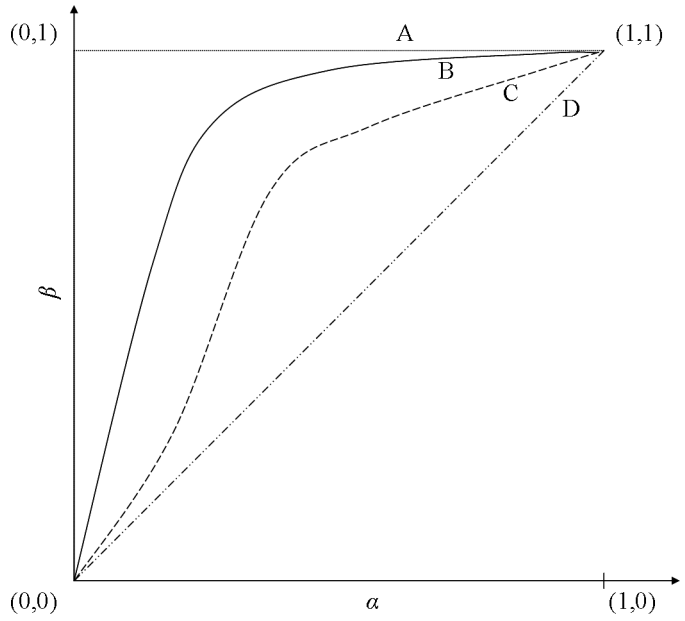


Figure 4.6 — Example of ROC curves: (A) ideal classifier; (B) corresponds to a better classifier than (C); (D) useless or random classifier.

the superiority of a classifier on the other may change for different thresholds. Thus the ultimate choice of the classifier depends on the problem at hand.

In our case, only one classifier has been defined (based on the mutual information between F_A and F_V). Its associated ROC curve is for sure interesting to analyze in order to get an idea about its performance. In particular, if it lies on the line $\alpha = \beta$, it can be ruled out since its performance are the same than those of a random classifier.

But an even more interesting use of the ROC analysis is for evaluating the whole classification chain performance rather than the performance of the classifier alone. We propose to evaluate the gain offered by the introduction of the feature optimization step prior to the classification step. To this end, two kind of features will be used in turn to estimate the mutual information in each mouth region: features extracted using the information theoretic framework defined in sec. 4.2, and equivalent non-optimized features. The comparative analysis of the performance of the classifier fed with each kind of features will be conveyed using ROC graphs.

4.4.4 Definition of the classifier for the two speaker case

Let us consider now the case where $m = 2$. That is, two mouth regions are present in the audio-video sequence. F_{V1} and F_{V2} are now the rvs modelling the video features associated to the mouth regions M_1 and M_2. Their samples are denoted f_{v1} and f_{v2} respectively. The extension of the classifier defined through the hypothesis test (4.31) to the two speaker case is straightforward: two hypothesis tests similar to the test (4.31) are simply associated to each mouth region M_1 and M_2. Four different cases can occur (a possible equality with the

threshold is solved by attributing randomly a class to the random variable pair):

1. $I_1(F_A, F_{V_1}) > \eta_1$ and $I_1(F_A, F_{V_2}) < \eta_2$: speaker 1 is speaking and speaker 2 is not;
2. $I_1(F_A, F_{V_1}) < \eta_1$ and $I_1(F_A, F_{V_2}) > \eta_2$: speaker 2 is speaking and speaker 1 is not;
3. $I_1(F_A, F_{V_1}) < \eta_1$ and $I_1(F_A, F_{V_2}) < \eta_2$: none of the speaker is speaking;
4. $I_1(F_A, F_{V_1}) > \eta_1$ and $I_1(F_A, F_{V_2}) > \eta_2$: both speakers are speaking.

A special case occurs if the experimental conditions are defined so as to eliminate the possibilities 3 and 4: the test set is composed of sequences where speakers 1 and 2 are speaking each in turn, without silent states. Therefore, if a speaker is silent, it implies that the other one is actually speaking. For this particular case, another hypothesis test can be stated:

$$\begin{aligned} H_0 &: f_a, f_{v1}, f_{v2} \sim p_{H_0} = p_{F_A,F_{V1},F_{V2}}(f_a, f_{v1}, f_{v2}) = p_{F_{V1}}(f_{v1}) \cdot p_{F_A,F_{V2}}(f_a, f_{v2}), \\ H_1 &: f_a, f_{v1}, f_{v2} \sim p_{H_1} = p_{F_A,F_{V1},F_{V2}}(f_a, f_{v1}, f_{v2}) = p_{F_A,F_{V1}}(f_a, f_{v1}) \cdot p_{F_{V2}}(f_{v2}). \end{aligned} \quad (4.39)$$

H_0 postulates the data to be governed by a pdf that states the independence of the video source 1 with respect to the audio and the video source 2. These last two are considered as being dependent: mouth 2 - or speaker 2 - is then classified as the speaker. The inverse dependency relationship is stated by the alternative hypothesis H_1, where speaker 1 - or mouth 1 - is then considered as responsible for the measured speech signal.

The log-likelihood ratio between the two hypothesis is defined in a manner similar to Eq. (4.36). Introducing the factor $p_{F_A}(f_a)/p_{F_A}(f_a)$, some equation manipulations lead to:

$$\Lambda(f_a, f_{v1}, f_{v2}) = \log\left[\frac{p_{F_A,F_{V1}}(f_a, f_{v1}) \cdot (p_{F_A}(f_a) p_{F_{V2}}(f_{v2}))}{p_{F_A}(f_a) \cdot p_{F_{V1}}(f_{v1}) \cdot p_{F_A,F_{V2}}(f_a, f_{v2})}\right] \gtrless \eta, \quad (4.40)$$

$$= \log\left[\frac{p_{F_A,F_{V1}}(f_a, f_{v1})}{p_{F_A}(f_a) p_{F_{V1}}(f_{v1})} \cdot \frac{p_{F_A,F_{V2}}(f_a, f_{v2})}{p_{F_A}(f_a) p_{F_{V2}}(f_{v2})}\right] \gtrless \eta, \quad (4.41)$$

$$= \log\left[\frac{p_{F_A,F_{V1}}(f_a, f_{v1})}{p_{F_A}(f_a) p_{F_{V1}}(f_{v1})}\right] - \log\left[\frac{p_{F_A,F_{V2}}(f_a, f_{v2})}{p_{F_A}(f_a) p_{F_{V2}}(f_{v2})}\right] \gtrless \eta. \quad (4.42)$$

The expectation of this normalized log-likelihood ratio gives rise to a difference of mutual information [44]:

$$\Lambda(f_a, f_{v1}, f_{v2}) = I(F_A, F_{V1}) - I(F_A, F_{V2}) \gtrless \eta. \quad (4.43)$$

Setting η to zero, the classifier function defined in paragraph 4.4.1 is found back.

One could think that the analysis of performance of this MI difference classifier could be done as well by varying the threshold η. It does not really make sense however since Eq. (4.43) can be equivalentally written as:

$$I(F_A, F_{V1}) \gtrless \eta + I(F_A, F_{V2}), \quad (4.44)$$

$$I(F_A, F_{V1}) \gtrless \eta' \cdot I(F_A, F_{V2}). \quad (4.45)$$

Thus, the mutual information between F_A and F_{V2}, or the expectation of the log-likelihood ration between F_A and F_V is compared to a threshold which is itself function of another log-likelihood ratio: such a comparison does not make any sense in decision theory. For performance analysis, only the scheme described in the previous paragraph and involving one hypothesis test per mouth will be used.

4.5 Discussion

In this chapter, a multimodal pattern recognition system has been proposed. It is more specifically dedicated to speaker detection, however, it can also apply to any similar detection task where a hidden source yields two signals of different modalities. The purpose is to take advantage of the multimodal specificity of the problem to increase the detector performance.

For each step of a standard pattern recognizer (Fig. 4.1), except for the two first stages, a solution has been presented in order to minimize the error probability of the whole system. The two first steps, data acquisition and representation will be discussed in the next two chapters, since they directly depend on the characteristics of the concrete application.

The first stage tackled here deals with feature extraction: a multimodal framework has been developed, where the two modalities are used jointly in a feature-level fusion scheme in order to recover the information originating from the common source while the independent noise is discarded. This approach is shown to minimize the probability of committing an error on the source estimate $\hat{S}$.

These optimal features feed in the classifier, which comes at the next processing step. This classifier is defined through an hypothesis testing approach. It fuses the two modalities to output a single decision about the label of each candidate mouth region ("speaker" or "non-speaker"). The hypothesis testing approach gives means for evaluating the performance of the classifier itself but also of the system. In particular, it is possible to appraise the added value offered by the feature extraction step.

As a final remark, let us recall here a question, central to this work, asked in chapter 2: *"What do we fuse, how and when?"*. Throughout this chapter, these three points, but more particularly the "how" and the "when" points , have been answered to:

- What: The acoustic and visual speech signals are the dedicated modalities. More precisions about the concrete representations used will be given in the next two chapters.
- How: A multimodal framework based on information theory has been proposed for performing the feature extraction as well as for the classification itself.
- When: The presented approach combines a fusion at both feature- and decision-levels (hybrid approach). The information present in each modality is used jointly to extract optimized features, but there are still two sets of features which input the

classifier. The latter performs a direct fusion of the information since it outputs a single outcome.

In the next two chapters, the designed pattern recognizer will be applied to the problem of speaker detection. A multimodal optimization of the audio features (chapter 5), then of the video features (chapter 6) will be carried out using the information theoretic feature extraction framework, and the performance of the resulting system will be analyzed through the evaluation method.

5 Audio feature extraction

5.1 Introduction

The pattern recognition framework defined in chapter 4 is now applied in the context of speaker detection. As previously mentioned, audio and video are the two cues for speech that are the most evident and the most easily measured. Therefore, these are the two modalities that are processed through all the stages of the classification chain shown in Fig. 4.1. However, the central part of the recognizer, namely, the information theoretic feature extraction framework, is specifically applied here for optimizing the audio information with respect to the video. The video content undergoes a feature extraction stage as well but it does not call to this multimodal framework.

The detector is asked as a preliminary requirement to perform without the use of a complex and numerous material: a single camera and microphone have to meet the needs. This already clarify the first step - data acquisition - of the pattern recognizer scheme. All the other steps, and in particular the points that were remaining unaddressed, such as data representation, or the definition of the optimization framework, are precisely discussed in the following of the chapter.

The chapter's structure roughly follows the recognition chain. Once acquired the raw data, an efficient representation must be chosen for both modalities, as tackled in section 5.2. Video features specific to speech are then extracted in section 5.3 by restricting specifically the video content to mouth regions present in the sequence. As far as audio features are concerned, they are optimized with respect to the video using the information theoretic feature extraction framework developed in chapter 4. This is developed in section 5.4, along with the precise definition of optimization criteria. The specific statistical

considerations associated with the optimized audio-visual feature sets are discussed in section 5.5. In section 5.6, the optimization problem is stated. The cost function is based on mutual information, which leads to a highly nonlinear optimization problem. Moreover, an analytical formulation of the gradient of the cost function is difficult to obtain without any parametric approximation of the pdfs. However, our purpose here is to avoid such an approximation and to directly solve our optimization problem using a suitable optimization method. Different optimization methods are tested. Their performances are presented and analyzed in section 5.7. In section 5.8, the results obtained by the classifier are presented and discussed, prior to a finer analysis involving the hypothesis testing approach. This analysis is carried out in section 5.9 where the performance of the whole pattern recognition system are evaluated.

5.2 Signal representation

5.2.1 Video representation

Physiologic evidences point out the motion in the mouth region as a visual cue for speech. The motion is a three-dimensional phenomenon. However, the only information available is a two-dimensional (2D) image sequence, i.e. a spatio-temporal map of the light intensity variations. The so-called *optical flow* (OF) gives an approximation of the projected 2D motion (the motion field or image velocity [119]). As defined by Horn and Schunck in [64], optical flow is "the apparent motion of the brightness pattern". It rests on the fundamental assumption of brightness constancy:

$$E_x(x, y, t) \cdot u + E_y(x, y, t) \cdot v + E_t(x, y, t) = 0, \tag{5.1}$$

where $E(x, y, t)$ denotes the image intensity as a function of position and time, E_x, E_y, E_t are the spatial and temporal derivatives of E, while u and v denote the horizontal and vertical velocity components. This equation states that the change of intensity in images is only due to motion.

The large amount of literature about optical flow computation proves that estimating optical flow from Eq. (5.1) is not an easy problem to solve. As a matter of fact, it is an ill-posed problem: many different vector fields can explain the same data (images) [76]. This problem is known as the aperture problem. Moreover, the brightness constancy assumption is often violated in the real life. Different attempts have been made in order to find a satisfactory solution to the problem of estimating OF. Beside high-level computer vision algorithms, which rely on image analysis to extract high-level features of data such as edges to solve the correspondence problem [76], several low-level algorithms exist. A good introduction to the main low-level optical flow techniques is given in [8]. A performance evaluation of several widely cited algorithms can also be found in [5].

Since no method can be said to outperform any other whatever the problem specificities, the latter have to be analyzed in order to pick up the optimal method. The estimation of

the mouth motion is a complex task as it is a non-rigid structure subject to self occlusions (the tongue and the teeth are appearing and disappearing structures), resulting in multiple motions and violations of the brightness constancy assumption. Moreover, it exhibits large velocities relative to standard video frame rate [46]. Pixel-based (or dense) motion representation given by gradient-based approaches is the least constrained approach to optical flow estimation. It allows to represent a very large number of motion fields [125], resulting in smooth and dense optical flow fields, thought maybe at the cost of reduced precision. These methods also exhibit relative computational efficiency. For these reasons, a gradient-based approach has been preferred to other low-level methods.

In the method of Horn and Schunck [64], a global spatial coherence constraint is introduced to regularize the problem stated by the fundamental equation of optical flow (Eq. (5.1)). An iterative scheme, given below, is then established and produces estimates of the vertical and horizontal components of the velocity at each pixel location, between two consecutive frames:

$$u^{n+1} = \bar{u}^n - I_x \frac{[I_x \bar{u}^n + I_y \bar{v}^n + I_t]}{\lambda + I_x^2 + I_y^2}, \tag{5.2}$$

$$v^{n+1} = \bar{v}^n - I_y \frac{[I_x \bar{u}^n + I_y \bar{v}^n + I_t]}{\lambda + I_x^2 + I_y^2}, \tag{5.3}$$

where λ is a regularization constant.

In order to get a video representation as related as possible to speech, the motion estimation is restricted to a rectangular region of $N \times M$ pixels including the lips and the chin of each speaker candidate. These regions are referred to as mouth regions. Notice that this restriction can be understood as a video feature optimization as presented in sec. 4.2, even if the theoretical feature extraction framework is not directly used: the motion related to speech production is kept, while the unrelated motion (interfering noise) is discarded.

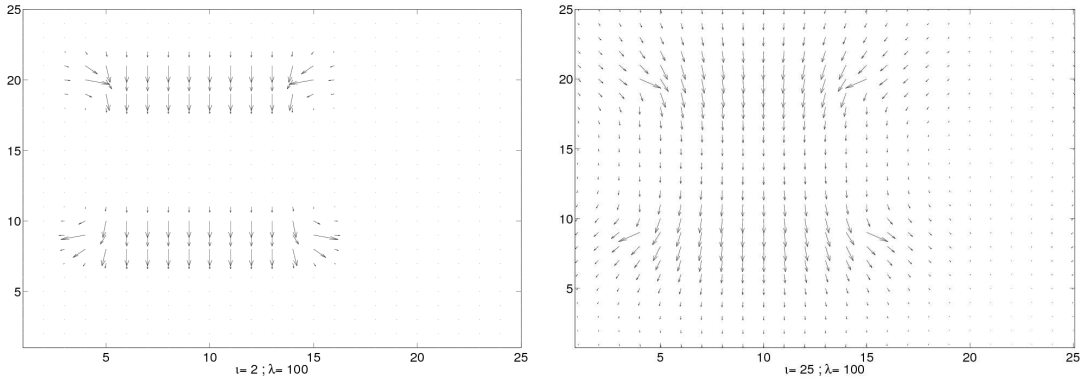
The method is implemented in a two-frame simple forward difference scheme so that the temporal resolution is large enough to capture complex and quickly varying mouth motions. First, a median pre-filtering is used on the raw intensity images to reduce the noise level.

Our purpose is to use the optical flow feature as a random variable F_V describing the visual speech production process. Ideally, the two components of the velocity at each pixel location - direction and norm - should be considered. However, in order to get reliable pdf estimates without using a very large sample, only the magnitude of the optical flow and the sign of the vertical component are kept, so that the video features are one-dimensional (1D). This signed norm has been preferred to the simple vertical component because it is less sensitive to head pose. The precise definition of the video features as random variables is discussed later on.

5.2.2 Determination of the optical flow parameters

The estimation of the optical flow field with Horn and Schunck's method requires to fix two parameters: the number of iterations ι and the value of the regularization term λ. According

to the authors, the flow field diffuses from the contours, or textured regions, toward the inside of the structures. Therefore, the number of iterations regulate the diffusion of the constraints and the parameter λ modulates directly the strength of these constraints. As a consequence, the number of iterations should be great enough for the objects of interest to be filled up, but small enough for preserving the structural information. Fig 5.1 shows how the number of iterations influences the estimation of the optical flow on two synthetic motion images where a black square 6×6 pixels large moves 2 pixels down.



(a) (b)

Figure 5.1 — Computation of the optical flow with Horn and Schunck's algorithm on two synthetic images where a black square 6×6 pixels large moves 2 pixels down. (a) optical flow estimated with $\lambda = 100$ and $\iota = 2$; (b) $\lambda = 100$ and $\iota = 25$. This is a very caricatural example since no texture information is present in the background or in the moving object. It simply intends to show how the flow diffuses from the edges as the number of iterations ι increases.

The size of the structure of interest in the images must then be known in order to fix the parameter ι. In the present case, the lips and the chin are the structures of interest. They are roughly equal to half the size of the mouth region. The number of iterations ι is then set up to $M/2$, where M is the height of the mouth region.

The regularization term, λ, must be set up so as to preserve a flow coherence. It also depends of the level of noise present in the image sequence.

5.2.3 Audio representation

As far as the audio signal is concerned, its representation should describe salient aspects of the speech signal, while being robust to variations in speaker or acquisition conditions. Mel-cepstrum analysis is one of the methods that fits best these requirements and as such, is widely used in speech-processing research [51], [102].

The human speech producing system is basically modelled as a system of tubes excited at one end by the vibrations of the vocal chords and the glottal pressure variations. The whole process amounts then, at a first approximation, as a convolution of the source (the excitation signal) with time varying filters (the time varying vocal tract). This is known as

the linear source filter model of Fant [40]. The cepstrum* of a speech signal is obtained by taking the logarithm of the power spectrum, before coming back in the temporal domain by an inverse Fourier Transform. Since a convolution in the time domain corresponds to a multiplication in the frequency domain and to a summation in the logarithmic power domain, cepstrum analysis permits to separate the spectral envelope from the spectral fine structures. Indeed, the low frequency spectral envelope is mostly characterized by the first cepstrum coefficients, whereas the finer structures appear in higher coefficients. The Figure 5.2 illustrates the different stages of the cepstral analysis leading to this separation at the end.

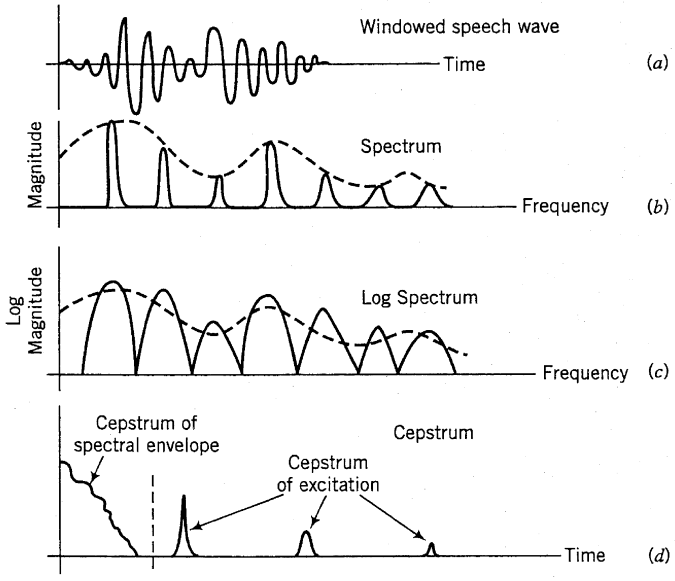


Figure 5.2 — Operations conducting from the acoustic speech signal $x(t)$ to the cepstrum coefficients. From [51].

In *mel-cepstrum* analysis, an integration stage over a filter bank is performed in the spectral domain before taking the logarithm. Fig. 5.3 summarizes the different steps of the process leading to the Mel-scaled Frequency Cepstral Coefficients (MFCCs) (or simply, the mel-cepstrum coefficients). The inserted filter bank models the way human beings perceive sounds. A lot of investigations have been performed in order to better understand human hearing. These works have shown that auditory neurons are tuned to specific characteristic frequencies with a critical band associated to. The auditory system act thus as a filter bank of varying bandwidths on the received audio signal. Moreover, the perceived pitch is related to the frequencies in a logarithmic way, leading to the following definition of the mel scale [124], relating the frequency in Hertz to its mel counterpart:

$$M_{mels} = x \cdot \ln\left(1 + \frac{f_{Hz}}{y}\right). \tag{5.4}$$

*Also called power cepstrum sometimes to avoid any confusion with the complex cepstrum [30]

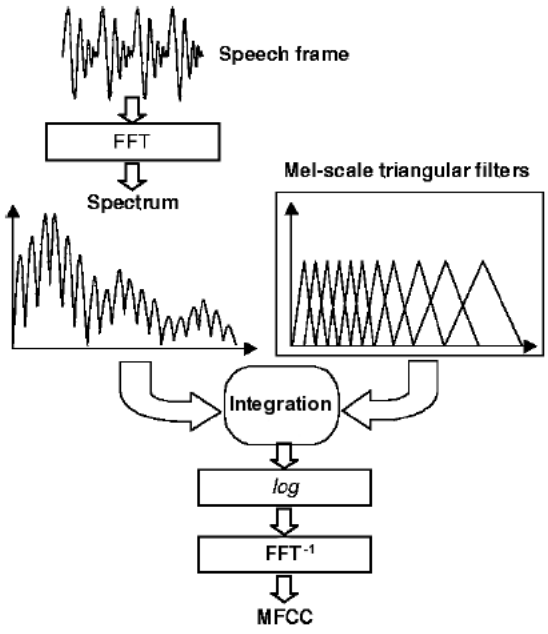


Figure 5.3 — Operations conducting from the acoustic speech signal $x(t)$ to the Mel-scaled Frequency Cepstral Coefficients. Based on [29].

Different values are possible for x and y but the most widely used are $x = 2595$ and $y = 700$ [133]. The filter bank using in the mel-cepstrum analysis consists then in overlapping triangular filters (typically between 20 and 30) uniformly positioned on the mel scale (then logarithmic-positioned on the frequency scale) over a predefined frequency range.

The last step of the mel-cepstrum computation consists in a spectral smoothing. To reduce the effects of noise, a cepstral truncation is performed and the highest coefficients are removed. Usually, the first coefficient is removed as well as it pertains for the average energy of the audio signal. As a result, mel-cepstrum analysis provides a representation corresponding to a smoothed short-term spectrum that has been compressed and equalized much as is done in human hearing [51].

As mentioned in sec. 5.2, the audio and video representations are observed on the same temporal window $[1, \ldots, T]$. The video features derived from the optical flow. Since this one is computed with a two-frame difference scheme, there are $\tau = T - 1$ frames of optical flow for the observation window. The OF frame rate differs then slightly from the initial video frame rate. Taking advantage of the properties of temporal continuities inherent to the mel-cepstrum coefficients, the mel-cepstrogram is downsampled to this OF frame rate using a bilinear interpolation, so that audio and video signal representations are temporally synchronized.

Finally, the speech signal is represented as a set of $\tau = T - 1$ vectors $\vec{C}$, each containing P mel-cepstrum coefficients $\{C_i(t)\}_{i=1,\ldots,P}$ with $t = 1, \ldots, \tau$.

5.3 Semi-automatic mouth region extraction

5.3.1 Motivation

As previously mentioned, the motion vectors must be as much as possible related to the production of speech for the audio feature optimization to be effective. For this reason, the motion estimation is limited to the mouth region. Incidentally, this removes the global head motion (which accounts for noise since it is not directly related to the acoustic speech) and decreases the amount of data to process, thus the computational time. The mouth region itself does not need to be accurately defined: it is just a rectangular region encompassing the lips and the chin. Since mouth extraction is only a first step towards a more general purpose, we are looking for a method as simple as possible.

5.3.2 Face detection

A simple approach consists in starting with the detection of the faces in the image. The mouth position is then found in each face coordinate system (FCS) using face and head anthropometric measures [41]. A face is indeed easier to detect than a mouth since it exhibits more steady features.

The face detector of Meynet *et al.* [87] is used. Basically, this face detector first constructs a number of classifiers by boosting Gabor-like discriminative local features. Then, a parallel mixture of these classifiers is defined using probability rules to take the final decision. A particle filter tracking helps also in taking a decision past the initial frames.

For the detector to achieve good performance, face views should be near-frontal and only small displacements should occur on the axis (Oz) (where the axis (Oz) is defined as the axis orthogonal on the image plane). This last requirement improves the detector performance as it can assume the face size to remain roughly unchanged all along the sequence. The detector eventually outputs the center coordinates of a rectangular boxes of size $h \times w$ pixels centered on each detected face (see Fig. 5.4 for an example).

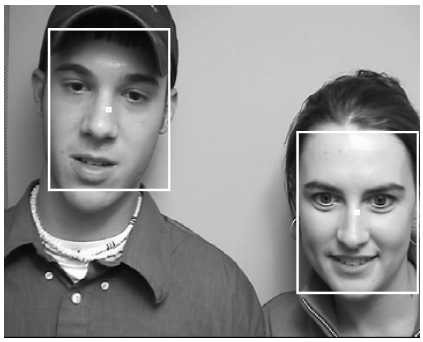


Figure 5.4 — Examples of detected faces. Sequence taken from [98].

5.3.3 Anthropometric measurements

In the face regions given by the detector, the recognition of meaningful features is easier than in the whole image. It would seem natural to look for the exact mouth position in the face box by extracting specific mouth features such as the corner of the lips for example. However, the large and complex motions of the lips during speech as well as the resulting illumination changes in the mouth region make this task arduous. Other features such as pupils, eyebrows or nostrils for example, present much more stability in their shape and position. Once one of these features has been extracted, the mouth is easily localized using face anthropometric measures [41]. The nostrils detection might be sensitive to person characteristics (compare for example the left and right persons on Fig. 5.4) as well as to the head rotation about the horizontal axis. The eyebrow shape is sensitive to small face rotations about the vertical axis and can be difficult to detect if the person is wearing a hat shadowing the eyebrow region (left person on Fig. 5.4). Specific haircuts may also mask this region. Therefore, the pupils are the most suitable features to extract and have been preferred as landmarks.

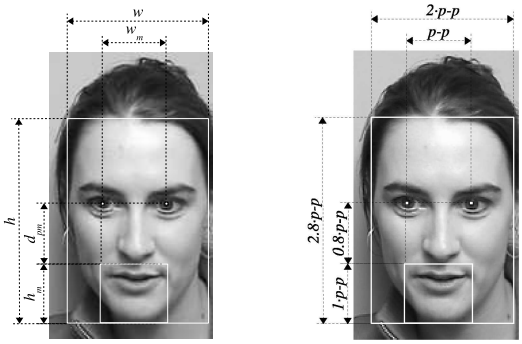
Anthropometric measures of interest can be expressed as a multiple of the inter-pupils distance, `p-p` (craniofacial norms used are those defined in [41] for North American Caucasians young adults). Table 5.1 and Figs. 5.5 present these measures. Notice that the ratio are the same for both male and female categories except for the height of the mouth (h_m).

Name	Distance	Male	Female
Width of the face	w	$2.1 \cdot$ p-p	$2.1 \cdot$ p-p
Height of the face	h	$2.8 \cdot$ p-p	$2.8 \cdot$ p-p
Width of the mouth region	w_m	$1.0 \cdot$ p-p	$1.0 \cdot$ p-p
Height of the mouth region	h_m	$1.1 \cdot$ p-p	$1.0 \cdot$ p-p
Distance pupil to top of mouth region	d_{pm}	$0.8 \cdot$ p-p	$0.8 \cdot$ p-p

Table 5.1 — Definition of some craniofacial norms in North American Caucasians young adults as a multiple of the distance between the pupil centers. (From [41]).

From the knowledge of `p-p`, it is possible to get the position and size of the $h_m \times w_m$ box cropping the mouth. This position is defined by the position of the upper-left corner of the mouth box: its upper corners lie at a vertical distance d_{pm} from each pupil. Some ϵ constants can be added or subtracted from the different distances to potentially adapt to specific heads or sequences. For example, if the individuals are not exactly facing the camera, or for non-Caucasian subjects, the measurements might slightly differ. Tilted faces may also impose to extract larger mouth regions.

The knowledge of the distance `p-p` gives an indication about the size $h \times w$ of the box cropping the face. It can be given as an indicative guess to the face detector so that the probability of false positives is decreased. However, for the detector to deal with slight changes in face size (due for example to change in the head orientation), it cannot be treated as a strict size value.



(a) (b)

Figure 5.5 — (a) Anthropometric measurements allowing to find the mouth region from the pupil locations; (b) Anthropometric measurements expressed as a multiple of the inter-pupils distance `p-p`.

Recall also that knowing the height of the mouth region allows to automatically set up the number of iterations ι required in the optical flow estimation algorithm (§ 5.2.1).

5.3.4 Mouth extraction

Given the face detector output and the inter-pupil distance, the mouth bounding box is easily extracted. A question remains however about how to get this distance `p-p`. Actually, if the detector gives a good global localization of the faces, this one is not sharply accurate: the detection is not always centered on the same exact point of the face, i.e. it does not have the same coordinates in the FCS. If the pupils, thus the mouth region, are extracted with respect to this detected face center, the mouth extractor might not be always centered on the mouth. As a result, artificial global motion can be introduced in the mouth sequence, or even worse, some part of the mouth might be missing. Figs 5.6 show an example of such changes in the detected center position and the resulting extracted faces and mouths.

Figure 5.6 — Examples of consecutive extracted faces (top) and mouths (bottom) when the center of the face detection box moves in the face coordinate system.

If the pupils might be detected by themselves into the face box, the mouth extraction would not be dependent of such slight changes in the face center detection. In some first experiments, segmentation-based or contour-based methods to extract the pupils have been

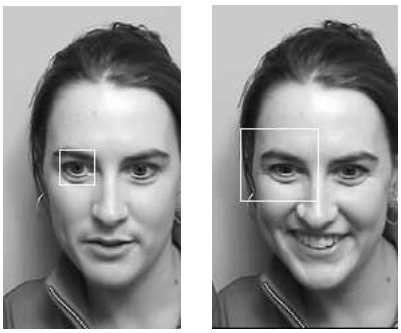
investigated. However, such automatic detection methods are too sensitive to illumination to be easily used. Moreover, the characteristics of the extracted regions are dependent of the head pose and different corresponding models should be established to get reliable results. A simpler, yet robust semi-automatic approach was preferred. Starting from a user-defined pupil positions pointed on the first frame, the correlation between templates extracted from frame to frame around the pupils is evaluated in order to track the pupil centers along the sequence.

Let P_{t_0} be the position of a pupil - the left one for example - on the first frame for a given frame. The distance between the pupil and the center C_{t_0} of the detection box is defined as follows:

$$\Delta PC_0 = C_{t_0} - P_{t_0}. \tag{5.5}$$

This distance is assumed to be constant along the sequence (no face translation in the *(Oz)* direction) and is taken as a referential distance.

A template $\mathcal{T}$ of size $\mathcal{T}_n \times \mathcal{T}_n$ (with $\mathcal{T}_n = |p - p|/2$ pixels), to be matched with the next frame, is extracted around the pupil. Experiments have shown that choosing a biggest region would increase the computational time without increasing the matching reliability. An example of a template region is presented on Fig. 5.7.(a).



(a) (b)

Figure 5.7 — (a): Template region $\mathcal{T}$ (white box) on frame $t-1$; (b): Search region $\mathcal{R}$ (white box) on frame t.

On a new frame t, the detector extract the face and outputs the detection center C_t. A hypothesized pupil position P_t is deduced from C_t and the referential distance ΔPC_0 defined in Eq. (5.5). In the same time, the correlation between the pupil template $\mathcal{T}$ specified at frame $t-1$ and a region $\mathcal{R}$ around P_t is evaluated. Example of such a search region is displayed on Fig. 5.7.(b). The point of maximal correlation indicates the position of the pupils P_t^c which might or might not coincide with P_t. If this maximal correlation value is above a pre-defined threshold, the true pupil position P_t^{true} is set to P_t^c, else $P_t^{true} = P_t$. The mouth is finally extracted given this true pupil position. Fig. 5.8 shows the result obtained using this correlation tracking on the same face sequence than the one previously presented on Fig. 5.6. Note that the face positions have also be corrected using $C_t^{true} = P_t^{true} + \Delta PC_0$.

Figure 5.8 — Faces centered back and corresponding extracted mouth regions for the three consecutive frames presented on Fig. 5.6.

Let us stress that the template is extracted from the previous frame and not from the first one for the whole sequence to deal with change in illumination, as well as change in the eye shape due to opening or closing. Also, the whole face is not taken as the region $\mathcal{R}$ because experiments have shown that non-eye's regions might correlating higher with the template region. There is a risk of propagating an error if the precedent pupil position is not right. However, the numerous tests have prove the method to be reliable enough for our purposes. The tests carried out on sequence $g22$ taken from the CUAVE database [98] (the sequence from which the here-presented examples are extracted) were particularly satisfactory even if the right person (the woman on Fig. 5.4) is moving in a noticeable way. In such a case, where the centers outputted by the face detector are rarely in a stable position within the FCS, the correlation method allows to limit the artificial global mouth motion introduced in the mouth sequence.

5.4 Audio feature extraction

5.4.1 Application of the feature extraction framework

The visual information related to speech production has been extracted from the initial representation without using the information theoretic framework discussed in sec. 4.2, but simply by limiting the optical flow estimation to the mouth region.

To extract the acoustic information related to speech production from the initial audio representation (the P-dimensional mel-cepstrogram), the feature extraction framework is now used. Its application is done so as to both emphasizing the information related to the video content and reducing the feature dimensionality from P to one. Indeed, as for the video case, we want to consider the audio feature as a random variable F_A. This rv would be here P-dimensional and might as such requires a too large sample for its pdf estimate to be correct. Consequently, the 1D audio features $\{f_{a_t}(\vec{\alpha})\}_{t=1,\dots,\tau}$ composing the sample of the rv, are built as the following linear combination of the P Mel-Frequency Cepstral Coefficients:

$$f_{a_t}(\vec{\alpha}) = \sum_{i=1}^{P} \vec{\alpha}(i) \cdot C_t(i) \quad \forall t = 1, \dots, \tau, \tag{5.6}$$

where the weights $\vec{\alpha}(i)$ are chosen such that $\sum_{i=1}^{P} \vec{\alpha}(i) = 1$ and $\vec{\alpha}(i) \in \mathbb{R}^+ \quad \forall i = 1, \dots, P$.

Thus, the τ P-dimensional audio observations are reduced to τ 1D observations of the random variable $F_A(\vec{\alpha})$.

It must be clear that the i^{th} weight $\vec{\alpha}(i)$ is a scalar associated to the set of values of the i^{th} mel-cepstrum coefficient $\{C_t(i)\}_{t=1,\ldots,\tau}$ and that $F_A(\vec{\alpha})$ is 1D.

5.4.2 Optimization criteria

As exposed in sec.. 4.2, minimizing the lower bound on the estimation error is equivalent to maximizing the efficiency coefficient considering the audio and video features over a mouth region. Thus, the minimization of the estimation error given by Eq. (4.23) leads to the optimized vector $\vec{\alpha}$. To be exact, the set of weights $\alpha(i)$, with $i = 1, \ldots, P$ to be optimized with respect to the Efficiency Coefficient Criterion is defined as:

$$\begin{aligned}\vec{\alpha}_{opt} &= \arg\max_{\vec{\alpha}}\{I(F_A(\vec{\alpha}), F_V)/H(F_A(\vec{\alpha}))\},\\ &= \arg\max_{\vec{\alpha}}\{e(F_A(\vec{\alpha}), F_V)\}.\end{aligned} \tag{5.7}$$

The normalization term for the mutual information, $H(F_A(\vec{\alpha}))$ involves the marginal audio feature entropy instead of the joint entropy, since the video features remain unchanged during the optimization process.

To verify the necessity of normalizing the mutual information by the entropy during the optimization, ECC will be compared with a "simple" Mutual Information Criterion (MIC). The set of weights to be optimized is then defined as:

$$\vec{\alpha}_{opt} = \arg\max_{\vec{\alpha}}\{I(F_A(\vec{\alpha}), F_V)\}. \tag{5.8}$$

Finally, a more constraining criterion is introduced for the two speaker case. This criterion, referred to as ΔECC, takes into account a pair of mouth regions. It is the squared difference between the efficiency coefficient computed in each mouth region (referred to as M_1 and M_2). This way, the differences between the marginal densities of the video features in each region are taken into account. Moreover, only one optimization is performed for two mouths. If F_{V1} and F_{V2} denote the random variables associated to regions M_1 and M_2 respectively, then the optimization problem becomes:

$$\vec{\alpha}_{opt} = \arg\max_{\vec{\alpha}}\left\{\left[e(F_A(\vec{\alpha}), F_{V1}) - e(F_A(\vec{\alpha}), F_{V2})\right]^2\right\}. \tag{5.9}$$

5.5 Statistical considerations

5.5.1 Casting the problem in a probabilistic framework

The optimization criteria defined in the previous section involve the mutual information. Therefore, the problem must be cast in a probabilistic framework. F_A and F_V have already been defined as the one-dimensional random variables defined over the sample spaces Ω_{F_A}

and Ω_{F_V} and associated to the audio and video features respectively. Let us recall here that the audio features are the linear combination of P MFCCs and the video features, the signed magnitude of the optical flow values estimated in the mouth regions (sec.s. 5.4 and 5.2).

As stated in sec. 4.3, a sample is defined as the collection of independent observed values of a rv. Different definitions of a random variable and of an associated sample state then different statistical relationships.

The video sample can be defined in at least three different ways, leading to three different probability density functions of F_V. These different possibilities as well as the statistical considerations associated to the audio features are detailed now.

5.5.2 Audio random variable

Since the audio features have been defined as the linear combination of the P MFCCs, a sample of the audio rv F_A is the set of τ one-dimensional observations $\{f_{a_t}\}_{t=1,\dots,\tau}$, where f_a is given by Eq. (5.6).

As stated in sec. 4.3, the densities are estimated using a Parzen window estimator. At a given point $f_a(i)$ of the audio sample space, the marginal density of F_A is defined as:

$$p_{F_A}(f_a(i)) = \frac{1}{\tau}\sum_{t=1}^{\tau} K_{h_a}(f_a(i), f_a(t)) \quad \forall\, f_a(i) \in \Omega_{F_A}, \tag{5.10}$$

where K_{h_a} is a kernel function with smoothing parameter h_a.

5.5.3 *Case I*: Definition of a 1D video rv at each point of the mouth region

The random variable F_V modelling the video features can be defined firstly as a 1D rv whose associated sample comprises τ observations $\{f_{v_t}\}_{t=1,\dots,\tau}$ associated to a given location in the mouth region. Then, an observation at a given instant t belongs to $\mathbb{R}^1$ and if the mouth region counts N pixels, N random variables $F_V^{(n)}$, $n = 1, \dots, N$, are defined. In other words, for a given observation window, N different video samples of τ observations are then drawn from N different pdfs.

Such a statistical consideration means that we do not consider any statistical relationship between the video observations located at each pixel of the mouth region, since one random variable with its subsequent pdf is defined at each of these pixel locations.

For the rv $F_V^{(n)}$ associated to the n^{th} pixel of the mouth region, the pdf obtained using the Parzen window estimator is:

$$p_{F_V}^{(n)}(f_v(i)) = \frac{1}{\tau}\sum_{t=1}^{\tau} K_{h_v}(f_v(i), f_v(t)), \quad \forall\, f_v(i) \in \Omega_{F_V}, \quad n = [1 \dots N], \tag{5.11}$$

where K_{h_v} is a kernel function with smoothing parameter h_v.

If normal kernels are considered, the joint probability density function between the audio and video rv is given by (sec. 4.3):

$$p^{(n)}_{F_A,F_V}(f_a(i),f_v(j))=\frac{1}{\tau}\sum_{t=1}^{\tau}K_{h_a}(f_a(i),f_a(t))\cdot K_{h_v}(f_v(j),f_v(t)). \tag{5.12}$$

5.5.4 *Case II*: Definition of a ND video rv for the whole mouth region

A second way of defining $\vec{F}_V$ is as a N-dimensional random vector (where N again stands for the number of pixels in the mouth region). The sample associated to $\vec{F}_V$ comprises the set of observations $\{\vec{f}_{v_t}\}_{t=1,\dots,\tau}$, with $\vec{f}_v \in \mathbb{R}^N$.

This basically means that a statistical relationship is considered between the (signed) norm of the velocity at each point of the mouth region. The time-indexed vectors of observations are identically independently distributed (i.i.d.). Such a statistical consideration matches better the reality. However it requires a very large sample size for avoiding the pdf estimation to be affected by the curse of dimensionality.

The N-dimensional pdf of the random vector $\vec{F}_V$ is defined as:

$$p_{\vec{F}_V}(\vec{f}_v(i)) = \frac{1}{\tau}\sum_{t=1}^{\tau}K_{h_v}(\vec{f}_v(i),\vec{f}_v(t)), \quad \forall \vec{f}_v(i) \in \Omega_{F_V}. \tag{5.13}$$

Using normal kernels with different smoothing parameters $h_v^{(q)}$ in the q^th direction, $q = 1\dots N$, Eq. (5.13) becomes equivalent to [113]:

$$p_{\vec{F}_V}(\vec{f}_v(i)) = \frac{1}{\tau}\sum_{t=1}^{\tau}\prod_{q=1}^{N}K_{h_v^{(q)}}(f_v^{(q)}(i),f_v^{(q)}(t)). \tag{5.14}$$

The joint pdf between the audio and video random variables is:

$$p_{F_A,F_V}(f_a(i),\vec{f}_v(j))=\frac{1}{\tau}\sum_{t=1}^{\tau}K_{h_a}(f_a(i),f_a(t))\prod_{q=1}^{N}K_{h_v^{(q)}}(f_v^{(q)}(j),f_v^{(q)}(t)). \tag{5.15}$$

5.5.5 *Case III*: Definition of a 1D video rv for the whole mouth region

The third and last case can be seen as a trade-off between the two previous statistical representations. The first case indeed does not consider any statistical relationship between neighboring pixels in the mouth region while the second one requires too large sample sizes to be efficient. In this third case, the co-occurrence of an audio observation with the N observations of the velocity in a video frame is considered.

Thus, the video random variable F_V is 1D and one of its associated video sample comprises the set of observations $\{f_v(k)\}_{k=1,\dots,\tau\cdot N}$ with $f_v \in \mathbb{R}^1$. The marginal pdf of the video

random variable is given by:

$$p_{F_V}(f_v(i)) = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{N}\sum_{n=1}^{N}K_{h_v}(f_v(i), f_v(t\cdot n)), \tag{5.16}$$

$$= \frac{1}{\tau\cdot N}\sum_{k=1}^{\tau\cdot N}K_{h_v}(f_v(i), f_v(k)). \tag{5.17}$$

The observations being i.i.d., this statistical consideration is obviously a simplification of the real world. Indeed, neighboring pixels are correlated and cannot be truly independent. This simplification is somehow compensated by estimating the pdf with the Parzen window approach where each observation has an effect on its neighbors in the sample space (sec. 4.3).

The joint probability between the audio and the video random variables is given by:

$$p_{F_A,F_V}(f_a(i), f_v(j)) = \frac{1}{\tau\cdot N}\sum_{t=1}^{\tau}K_{h_a}(f_a(i), f_a(t))\cdot\sum_{n=1}^{N}K_{h_v}(f_v(j), f_v(t\cdot n)). \tag{5.18}$$

Notice that for each of the three considered cases, the kernels K_{h_a} and K_{h_v} are centered on pair of audio and video observations $\{f_a(t), f_v(t)\}$. Since these pairs are formed by picking up observations in the samples using a jointly varying index t, a temporal co-occurrence of the audio and video observations is considered.

The *Case III* is chosen to statistically model the video features as it corresponds to a good trade-off between a representative and efficient model.

5.5.6 Mutual information between audio and video random variables

The equations for the mutual information corresponding to each of the three previous statistical cases are given now. In the *case I*, the mutual information between the random variables F_A and F_V can be written as:

$$I^{(n)}(F_A, F_V) = \sum_{i\in\Omega_{F_A}}\sum_{j\in\Omega_{F_V}}p(f_a(i), f_v(j))\cdot\log\frac{p(f_a(i), f_v(j))}{p(f_a(i))\cdot p(f_v(j)}. \tag{5.19}$$

To get a single mutual information value between the audio and video signals for the whole mouth region, the mean of the MI per pixel is computed:

$$I(F_A, F_V) = \frac{1}{N}\sum_{n=1}^{N}I^{(n)}(F_A, F_V). \tag{5.20}$$

In the *case II*, the mutual information between the audio and video rvs for the current observation window and for the whole mouth region is given by:

$$I(F_A, \vec{F}_V) = \sum_{i\in\Omega_{F_A}}\sum_{j\in\Omega_{F_V}}p(f_a(i), \vec{f}_v(j))\cdot\log\frac{p(f_a(i), \vec{f}_v(j))}{p(f_a(i))\cdot p(\vec{f}_v(j))}. \tag{5.21}$$

Finally, the mutual information corresponding to the third statistical consideration case (the retained one) is defined as:

$$I(F_A, F_V) = \sum_{i \in \Omega_{F_A}} \sum_{j \in \Omega_{F_V}} p(f_a(i), f_v(j)) \cdot \log \frac{p(f_a(i), f_v(j))}{p(f_a(i)) \cdot p(f_v(j))}. \tag{5.22}$$

5.5.7 Smoothing parameter

A Gaussian kernel is chosen for its widespread validity an because it simplifies the estimation of joint probability density functions. For joint pdfs, this kernel must be 2D: $G(\mu_A, \mu_V, h_a, h_v)$. Zero means and diagonal covariance matrix diag(h_a;h_v) are chosen. These variances h_a and h_v are estimated from the audio and video data respectively, in a robust way, as described in [23]:

$$h = \left(\frac{4}{3k}\right)^{1/5} \cdot \frac{\text{median } |y_i - \tilde{\nu}|}{0.6745}, \tag{5.23}$$

where $\tilde{\nu}$ denotes the median of the audio or video data points, y_i the point where the pdf is estimated, and k is the total number of sampling data points. Since the video data remain the same during the optimization of the audio data, the value for h_v remains constant for a given set of video features, while h_a will adapt to the audio features during the optimization process. The choice of such an adaptive smoothing parameter induces a multi-resolution approach to the optimization problem resolution, as will be shown in the next section.

5.5.8 Feature normalization

The mouth sizes may vary from one speaker to the other. As indicated in Table 5.1, the height of the mouth is given by a different ratio of `p-p` for male and female; depending on the distance speaker-camera, the size of the mouth also differs. The resulting feature values lie in different dynamic ranges. As a result, features with large value may have a larger influence in the cost function than features with small values, although this does not necessarily reflect their respective significance. As advocated in [129], features must be normalized so that their values lie within a similar range .

To allow a general comparison of the features, the definition of a global normalization factor has been preferred to a scaling based simply on the maximal and minimal values, or on the mean and variance values over the current temporal window. Looking to a reference sequence, a maximal lips motion between two consecutive frames is estimated. This maximal velocity V_{max}^{ref} defines the normalization factor. It is found to be equal to 8 pixels on the reference sequence (shown in Fig. 5.9.(a)). A reference inter-pupils distance $\|\text{p-p}\|^{ref}$ is estimated as well on this sequence: $\|\text{p-p}\|^{ref} = 33$ pixels. For a new sequence, where the inter-pupil distance is equal to $\|\text{p-p}\|$, the normalization factor is defined as: $V_{max} = V_{max}^{ref} \cdot m_r$, where $m_r = \|\text{p-p}\|/\|\text{p-p}\|^{ref}$ is the magnification ratio between the new and the reference sequence. Notice that we have describe this method in [11]. An illustrative example is given on Figs. 5.9.

(a) (b)

Figure 5.9 — Determination of the magnification ratio between a new sequence (a) and the reference sequence (b). The inter-pupils distance equals $\|\text{p-p}\| = 28.55$ pixels on the new sequence, instead of $\|\text{p-p}\|^{ref} = 33$ pixels on the reference sequence. Then the normalization factor for the velocities is given by $V_{max} = V_{max}^{ref} \cdot m_r = 6.9$ pixels.

This method can produce outliers: it is not excluded that features greater than our maximal estimated displacement appear. If this maximal value has been correctly determined, these outliers should correspond to non-lips motions (e.g. if a hand appears in the mouth region) or to motions not related to speech production. These are then of non interest for our purpose and should be discarded anyway.

5.6 Optimization framework

The extraction of optimized audio features with respect to our classification task requires to find the real-valued vector $\vec{\alpha} \in \mathbb{R}^P$ that minimizes* the defined cost function $f(\vec{\alpha})$. This function is defined as the negative value of one of the optimization criteria defined in Eqs. (5.7), (5.8), or (5.9). Moreover, to restrain the set of possible solutions, the P weighting coefficients $\{\alpha_i\}_{i=1,\dots,P}$ must fulfill the following conditions:

$$0 \leq \vec{\alpha}(i) \leq 1 \quad \forall i = 1, 2, \dots, P \,, \tag{5.24}$$

$$\sum_{i=1}^{P} \vec{\alpha}(i) = 1. \tag{5.25}$$

This optimization problem is highly nonlinear and gradient-free. Indeed, an analytical formulation of the gradient of the cost function is difficult to obtain due to the unknown form of the pdf of the extracted audio features. In [44], Fisher and Darell use a second order Taylor approximation of the mutual information and the Parzen estimator to cast

*To be coherent with standard optimization problem formulations, the maximization problem is turned into a minimization problem: the objective function is defined as the negative value of the optimization criterion.

the optimization problem into a convex one and to derive the gradient in an analytical way. However, our purpose here is to avoid such an approximation and to directly solve our optimization problem using a proper optimization method.

Optimization methods can be classified as either local or global. The first category includes steepest gradient descent and gradient descent-based methods such as the Powell's direction set method. They mainly rely on the use of an exact or estimated formulation of the gradient of the cost function to find an optimum. They present the advantage to be fast and easy to use but are very likely to fail to reach the global optimum of the cost function if the latter is not convex.

The second category refers to algorithms which aim at finding the globally best solution, in the possible presence of multiple local minima. We find in this category stochastic and heuristic methods such as Simulated Annealing (SA) [74], Tabu Search (TS) [50], or Evolutionary Algorithms (EAs). These have proven their ability to approach the global optimum of highly nonlinear problems, possibly at a high computational cost. Both SA and TS are more dedicated to solve combinatorial problems. EAs, which include Genetic Algorithms (GAs), look more suitable for our problem. Such optimization procedures, first introduced by Holland in 1962 [63], are based on natural evolution principles: starting from an initial candidate *population* of *chromosomes* (or sets of parameters to be optimized), operators mimicking the biological ones of *crossover* and *mutation* are used to *select* and *reproduce* fittest solutions, the fitness of a solution being given by a scoring function. Basically, mutation enables the algorithm to explore new regions of the search space by randomly altering some or all *genes* (components) of some chromosomes in the population. On the other hand, crossover reinforce prior successes by recombining parent-chromosomes in order to produce fittest offsprings.

Although the underlying principles are relatively simple, EA algorithms have proven to be robust and powerful search tools, owing to their remarkable flexibility and adaptability to a given task [122]. As a matter of fact, their tuning relies on a proper selection of values for only a few parameters: this makes them very attractive and easy-to-use. Furthermore, EAs do not try to provide an exact match but an approximation of the optimal solution within an acceptable tolerance, which improves their effectiveness.

5.6.1 Multi-resolution approach

Whatever the optimization method, a pre-processing of the cost function can be introduced to improve the efficiency of the optimization. Indeed, the MI-based cost functions are *a priori* non-convex and are very likely to present rugged surfaces. To limit the risk of getting trapped in a local minimum, it is common to smooth the cost function. A trade-off has to be found however between smoothness and loss of information so there is still no guarantee of finding the global optimum.

The cost functions require the estimation of the pdf: using the non-parametric Parzen windowing approach, fine estimates of the distributions are obtained with a small number

of observations, but also the cost functions are smoother than what could be expected with histograms. The smoothness of the density estimates and thus the smoothness of the cost functions is controlled by the parameter h (see sec. 4.3). This parameter must therefore be carefully chosen: if it is too small, the cost functions are likely to be highly irregular, with a negative impact on the optimization algorithm. On the other hand, if it is too large, the loss of information and in particular, the loss of discrimination between the densities can be dramatic and may lead to a wrong solution. The smoothing parameter defined in Eq. (5.23) is a function of the data points y. Therefore it varies along the optimization process as the audio feature data points vary. These audio feature data points tend to evolve so that their distribution gets away from a uniform distribution. Indeed, the optimization process looks for features which maximize mutual information, while possibly minimizing the joint entropy between the audio and video features, and the entropy is maximal for rv with uniform density. Roughly speaking, the smoothing parameter evolves as follows: at the beginning of the optimization, the audio features are scattered in the space thus the smoothing parameter is large: the pdf, and implicitly the objective function, is largely smoothed. As the optimization proceeds, the distribution of the data points tends to concentrate in the sample space and the smoothing parameter decreases: fine structures of the pdf, thus of the objective function, appear. The use of an adaptive smoothing parameter as defined by Eq. (5.23) induces then a multi-resolution approach for solving the optimization problem. Multi-resolution schemes have been shown to perform better in the context of optimization problems involving mutual information, notably, in image registration problems (see for example [31]).

5.6.2 Local optimization: Powell's direction set method

In a first set of experiments (presented in [14] and [11]), we have used the deterministic Powell's direction set method [103]. To reduce the optimization problem as well as to deal with the constraints given by Eqs. (5.24) and (5.25), the objective function is reformulated through trigonometric relations. Namely, instead of directly looking for the set of $\{\alpha_i\}_{i=1,...,P}$ that maximizes the objective function, a set of $\{w_j\}_{j=1,...,\log P}$ weights is defined. Taking advantage of the trigonometric property of Eq. (5.26), these $\log P$ weights are then combined to define the P coefficients α. If $\log P$ is not an integer, the power of two immediately superior is considered and the weights α are normalized afterwards.

$$\sin^2(w) + \cos^2(w) = \cos^2(\frac{\pi}{2} - w) + \cos^2(w) = 1 \tag{5.26}$$

$$\alpha_i = \prod_{k_1=0}^{1} \dots \prod_{k_j=0}^{1} \left[\cos^2(k_1\frac{\pi}{2} - w_1) \dots \cos^2(k_j\frac{\pi}{2} - w_j)\right] \quad \text{with } j = 1, 2, ..., \log(P). \tag{5.27}$$

Thus, the $\vec{\alpha}$ coefficients still constrain the objective function but the number of parameters to optimize is reduced in a logarithmic way.

The Powell's algorithm is well-suited for problems where no analytical formulation of the gradient is available. It finds the minimum of a multidimensional cost function by solving

sequences of one-dimensional minimizations (using for example the one-dimensional Brent's optimization method) along N linearly independent, mutually conjugate set of directions. This method belongs however to the category of local optimization methods: if the surface of the cost function is not smooth and exhibits several local optima, the ability of the algorithm to reach the global optimum relies on a judicious initial guess of the solution.

Combining both smoothing and different initial trials, we obtained good results, showing that our framework was able to extract audio features specific to speech production. The mutual information measured thereafter between the extracted audio features and the video features of different mouth regions indicated the current speaking mouth in simple audio-video sequences [14].

However, the solutions found by this method were strongly dependent on the initial conditions, showing that the objective function still exhibited too many local optima. Therefore the method was not performing at its best level. To ensure the global optimum to be reached, an exhaustive trial of all possible initial points should be performed; an approach which is, obviously, unfeasible. Consequently, a global optimization strategy turned out to be preferable. Moreover, to be efficient, this global optimization method should fulfill the following requirements:

1. Efficiency for highly nonlinear problems without requiring the cost function to be differentiable or even continuous over the search space;
2. Efficiency with objective functions that present a flat, rough error surface;
3. Ability to deal with real-valued parameters;
4. Ability to handle the two constraints defined by Eqs. (5.24, 5.25) in the most efficient way.

5.6.3 Global optimization: Genetic Algorithm in Continuous Space (GACS)

An evolutionary approach such as genetic algorithm (GA) answers the two first demands previously defined while presenting flexibility and simplicity of use in a challenging context. Conventional GAs however have difficulties to handle the third and fourth requirements because they encode the solutions under the form of quantized and binarized representations (the *chromosomes*). Hence, working with real-valued parameters requires additional bits in chromosome representation to improve the precision, increasing the computational cost. Moreover, the crossover is likely to produce out-of-range values. Thus a validity test is required, decreasing the efficiency of the process. Finally, possible links between different solution parameters are ignored during crossover, slowing down once again the convergence process [59].

The genetic algorithm in continuous space (GACS) is an extension of the original GA scheme first described in [107] and [106] that alleviates these limitations by using the real valued parameter vectors instead of bit strings of chromosomes. This floating point representation presents the obvious advantage of retaining the proximity between two points

in both the representation and the problem spaces. The fourth requirement still has to be fulfilled, namely, efficient handling of the constraints defined by Eqs. (5.24) and (5.25).

The adaptation of GACS developed in [111] and [131], relates the genetic operators to the constraints on the solution parameters. It also speeds up the convergence of the algorithm by requiring the solution domain, or acceptance domain ($[0, 1]$ for each $\vec{\alpha}_i$ in our case, as indicated by Eq. 5.24), to be convex. At generation $t + 1$, a mutated vector $\vec{\alpha}_{t+1,k}$ (with $k = 1, \ldots, N$) is generated from a chromosome $\vec{\alpha}_{t,k}$ selected from the old population at generation t, by performing the following addition:

$$\vec{\alpha}_{t+1,k}(i) = \vec{\alpha}_{t,k}(i) + \epsilon, \tag{5.28}$$

where ϵ, the increment, is a zero-mean Gaussian perturbation which is applied to one element i of the chromosome vector that will mutate, with i randomly selected in the set $\{1, \ldots, P\}$. This scheme has shown to be more efficient in our case than mutating all the elements of the given chromosome vector at once.

For the mutation to be effective, that is, to possibly lead to improvement in the future populations by permitting the exploration of new regions of the search space, the variance of the Gaussian perturbation must be adequately chosen. A suitable value can be defined based on the above-defined acceptance domain for each element, as a certain fraction of this range. Note that it is necessary to check if the mutated gene still belongs to its acceptance domain. If it is not the case, the mutation is rejected.

The role of crossover is to reinforce the prior successes by merging the good characteristics of two chromosomes using a linear combination of candidates. To ensure that a recombined chromosome $\vec{\alpha}_{t+1,k_3}$ belongs to the acceptance domain, the crossover operator is defined as follows:

$$\vec{\alpha}_{t+1,k_3}(i) = \lambda \cdot \vec{\alpha}_{t,k_1}(i) + (1 - \lambda) \cdot \vec{\alpha}_{t,k_2}(i), \tag{5.29}$$

where $\vec{\alpha}_{t,k_1}$ and $\vec{\alpha}_{t,k_2}$ refer to two parent chromosome vectors at generation t, λ and i are randomly selected in the set $\{1, \ldots, P\}$. Since λ remains fix for each crossover operation, the search space is convex. Then the new chromosome vector $\vec{\alpha}_{t+1,k_3}$ is guaranteed to be valid if $\vec{\alpha}_{t,k_1}$ and $\vec{\alpha}_{t,k_2}$ are valid as well.

Finally, to ensure that the constraint defined by Eq. (5.25) is satisfied, all the chromosomes of the new generation are normalized. This implies de facto that each gene of each chromosome (excluding the replicated best one) in the new population is finally modified at the end of the iteration.

The specific evolution strategy implemented for the application addressed here is an extension of the scheme given in [111] and [131]. It is presented in Fig. 5.10 and can be summarized as follow:

1. Generate an initial population of N chromosomes (with N odd number) within the convex acceptance domain. Instead of randomly distribute the initial chromosome vectors in the search space, they are regularly placed in the acceptance domain according to a user-defined number of quantization levels Q [77].

2. Rank the chromosomes according to the evaluation (fitness) function, given by one of Eqs. (5.7), (5.8), (5.9). Reproduction is performed by keeping unchanged the best one for the next generation.

3. The remaining chromosomes then compete in pairs. Local pair-competitions for crossover are performed between a mutated and a crossovered chromosome of the previous generation. Crossover, using Eq. (5.29) is then applied to the winners of these local competitions until $(N-1)/2$ new chromosomes are generated and included in the next generation. Contrary to global competition, these local competitions allow the algorithm to preserve genetic diversity in the succeeding generations.

4. Complete the next generation by mutation of the best ranked chromosome $(N-1)/2$ times, using Eq. (5.28). These chromosomes combined with $(N-1)/2$ new chromosomes produced by crossover and the best ranked chromosome form the N chromosomes for the next generation. If the new chromosomes do not lie in the acceptance domain, reject the mutation.

5. Normalize the new parameters vectors such that the sum of the vector elements equals 1.

6. A stagnation of the best (reproduced) chromosome over a certain number of generations (typically 10 in our case) may indicate that the algorithm has reached a local extremum. To avoid such a situation, all chromosomes but the best one are in this case reset to random values.

7. Steps 2 to 6 are reiterated until the pre-defined maximum number of generations has been reached.

This evolution strategy is guaranteed not to diverge since the best chromosome is retained for the succeeding generations. Thus the GACS behaves at least like a random search process in a bounded search space. Note that unlike conventional optimization methods where the decrease of cost function over successive iterations can be used as the criterion to terminate the process, it is more difficult to assess the convergence in GACS since the stagnation in the cost function does not necessarily mean that the optimum is reached.

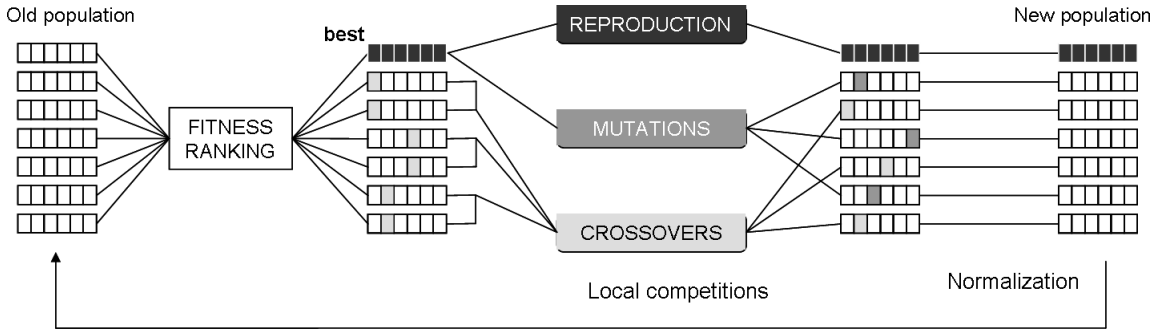


Figure 5.10 — Population renewal policy in GACS: reproduction, mutations and crossover (based on [131]).

Good results have been obtained using GACS. In particular, the optima reached were quite better than those obtained with Powell's method. However, the choice of an appropriate value for the parameters, especially the number of generations and the variance of the Gaussian perturbation still is a weak point. The latter has to be relatively high for the algorithm not to get stucked in local minima (i.e. to efficiently explore the search space). However, the highest the variance, the more likely a mutated parameter to fall outside the acceptance domain. As a result, the number of rejected mutations is too high for the population to preserve its diversity along the generations. Therefore, the mutation operator is not much more efficient with a high perturbation variance than with a small one. On some runs, only crossover maintains the evolution process active. Moreover, our solutions have proven to be sometimes very close to the boundaries of the search space. However, it is unlikely to approach the boundaries of the acceptance domain. As a result, a lost of the population diversity is observed which caused a noticeable difference between optima reached from one run to another.

What is needed is a scheme where the mutations applied are small for some parameters, and larger for others, allowing a better exploration of all the search space, including the region close to the boundaries, i.e. the perturbations need to adapt to the population evolution.

5.6.4 Differential Evolution (DE)

To overcome the problems encountered with GACS, the Differential Evolution approach (DE) introduced in 1997 by Storn and Price [127] has been used. As an evolutionary algorithm, it presents the same advantages than GACS and operates according to the same general scenario. The core difference between the two methods lies in the way the perturbation is generated. Rather than applying a perturbation generated by an *a priori* defined distribution as in the case of GACS, the perturbation in DE corresponds to the difference of chromosomes (rather called *vectors* in this context) randomly selected from the population. This way, the distribution of the perturbation is determined by the distribution of the vectors themselves. Since this distribution depends primarily on the response of the population vectors to the objective function topography, the biases introduced by DE in the random walk towards the solution match those implicit in the function it is optimizing [104]. In other words, the requirement for an efficient mutation scheme is more closely met: the generated increments move the existing vectors with both suitable displacement value and direction for the given generation.

The exact algorithm we used is based on the so-called *DE/rand/1/bin* algorithm [104]. Its pseudo-code, including the modifications for handling the constraints, is given in Algorithm 1. Let us describe here more in detail the different steps of this algorithm. An initial population of N vectors is first generated to lie within the convex acceptance domain, as was done with GACS, by dividing the search space in Q predefined quantization levels [77].

A perturbed vector $\vec{\alpha'}_{G,i}$, $i = 1, \ldots, N$ is then generated as a counterpart for each vector

$\vec{\alpha}_{G,i}$ of the current population N_G, where G refers to the current generation. This perturbed vector, or child vector, results form the linear combination of three parent vectors $\vec{\alpha}_{G,r_1}$, $\vec{\alpha}_{G,r_2}$, $\vec{\alpha}_{G,r_3}$ randomly picked up from the population N_G with $r_1 \neq r_2 \neq r_3 \neq i$ (these conditions ensure the DE mutation to be effective and not to simplify towards a classical crossover scheme [104])

$$\vec{\alpha}'_{G+1,i}(j) = \vec{\alpha}_{G,r_3}(j) + F \cdot (\vec{\alpha}_{G,r_1}(j) - \vec{\alpha}_{G,r_2}(j)), \tag{5.30}$$

where F is a scaling factor taking value on $[0, 2]$. A user-defined crossover probability CR controls the number of child vector element indices j subject to perturbation: P random numbers belonging to $[0, 1]$ are generated (i.e. one for each element of the vector under consideration); each time one of these random number is smaller than CR the corresponding vector element is subject to a perturbation. As a result, the child vector differs from its parent by at least one element ($CR = 0$) and at most, by all of its elements ($CR = 1$). Lines 2 to 11 of the Algorithm 1 sum up these operations.

Both the perturbed and the original populations are evaluated by the objective function and pair competitions are performed between child and parent vectors (so the population size remains constant). At the end of one iteration, a new population eventually emerges, composed by the winners of each local competition. The decision process is described in lines 8 to 11 of the Algorithm 1.

The constraints defined in Eqs. (5.24, 5.25) still hold. Therefore, the validity of each vector of the perturbed, or child, population has to be verified before starting the decision process. If the element j of a child vector i does not belong to the acceptance domain, it is replaced by the mean between its pre-mutation value and the bound that is violated [104] (lines 12 to 19 of the algorithm, where $\alpha^{(lo)}(j)$ and $\alpha^{(hi)}(j)$ refer respectively to the lowest and highest bounds defined for the j^{th} parameter - that is, 0 and 1 in our case). This scheme is more efficient than the simple rejection adopted with GACS. Indeed, it allows the bounds to be asymptotically approached, thus to cover efficiently the whole search space. To handle the second constraint (Eq. 5.25), a simple normalization is performed on each child vector, as it was done with GACS (lines 20-21 of the algorithm).

A good introduction to DE as well as some rules to tune the parameters properly can be found in [70] and [104].

Both the generation of the perturbation increment using the population itself instead of a predefined probability density and the handling of the out-of-range values allow the DE algorithm to achieve outstanding performance in the context of our problem.

5.7 Comparison of the optimization methods

The performances of the three different optimization methods are compared, while using them to minimize the objective function corresponding to *ECC* (Eq. (5.7)). For these tests, an audio-video sequence involving a single speaker - thus a single mouth region - has been used. A frame of this test sequence is shown as an example in Fig. 5.11. More details about

Algorithm 1: DE/rand/1/bin with modification for handling the constraints given by Eqs. (5.24, 5.25). Based on [104].

Input: $P, G_{max}, N \geq 4, F$ (scaling factor) $\in [0,2], CR \in [0,1], \vec{\alpha}^{(lo)}, \vec{\alpha}^{(hi)}$.
Initialize: initialization of the population;
$i = \{1, 2, \ldots, N\}$, $j = \{1, 2, \ldots, P\}$, $G = 0$;
while $G < G_{max}$ **do**
  for $i = 1, \ldots, N$ **do**
    Mutate and recombine:
    randomly select r_1,r_2,$r_3 \in \{1, 2, \ldots, N\}$, subject to $r_1 \neq r_2 \neq r_3 \neq i$;
    $j_{rand} \in \{1, 2, \ldots, P\}$, randomly selected once each i;
    for $j = 1, \ldots, P$ **do**
      s=rand([0,1))
      if $s < CR \vee j = j_{rand}$ **then**
        $\vec{\alpha}'_{G+1,i}(j) = \vec{\alpha}_{G,r_3}(j) + F \cdot (\vec{\alpha}_{G,r_1}(j) - \vec{\alpha}_{G,r_2}(j))$
      else
        $\vec{\alpha}'_{G+1,i}(j) = \vec{\alpha}_{G,i}(j)$
      Check validity:
      if $\vec{\alpha}'_{G,i}(j) < \vec{\alpha}^{(lo)}(j)$ **then**
        $\vec{\alpha}'_{G+1,i}(j) = (\vec{\alpha}_{G,i}(j) + \vec{\alpha}^{(lo)}(j))/2$
      else
        if $\vec{\alpha}'_{G,i}(j) > \vec{\alpha}^{(hi)}(j)$ **then**
          $\vec{\alpha}'_{G+1,i}(j) = (\vec{\alpha}_{G,i}(j) + \vec{\alpha}^{(hi)}(j))/2$
        else
          $\vec{\alpha}'_{i,G+1}(j)$
    Normalize:
    $\vec{\alpha}'_{i,G+1} = \vec{\alpha}'_{G+1,i} / \sum_{k=1}^{P} \vec{\alpha}'_{G+1,i}(k)$
    Select:
    if $f(\vec{\alpha}'_{G+1,i}) \leq f(\vec{\alpha}_{G,i})$ **then**
      $\vec{\alpha}_{G+1,i} = \vec{\alpha}'_{G+1,i}$
    else
      $\vec{\alpha}_{G+1,i} = \vec{\alpha}_{G,i}$
  $G = G + 1$

the sequence are given in the next section, where the main results on speaker detection are presented (this one-speaker sequence presents the same characteristic than the two-speaker ones presented next).

For both GACS and DE algorithms, different tests have first been performed in order to tune the parameters properly. As far as GACS is concerned, a choice of $Q = 5$ quantization

Figure 5.11 — Frame example of the test sequence used to perform the comparison between the different optimization methods. The white rectangular box delimits the extracted mouth region.

levels (resulting in a population of 125 chromosomes) combined with 400 generations and a perturbation variance σ fixed to 0.1, gave good results. The implementation of the DE algorithm has been based on Storn's public domain version software [126]. It achieved good performance with $Q = 5$ quantization levels, 500 generations, a scaling factor $F = 0.5$ and a crossover probability CR equals to 1.

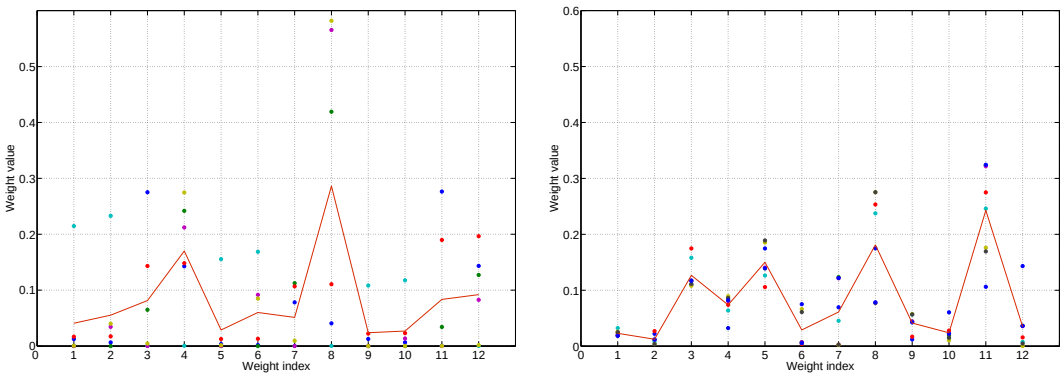
Once determined these optimal parameters, different runs have been performed with GACS and DE algorithms, whereas different initial conditions (i.e. different initial solution guesses) have been tried for the Powell's method. Table 5.2 summarizes the results obtained with each method. Obviously, much better minimization is obtained using the global optimization schemes instead of the local one (Powell's). A finer analysis of the results in Table 5.2 reveals that DE reaches the overall best solution and in a more stable way. Indeed, the standard deviation of the solutions is much smaller in the case of DE than in the case of the other two methods, giving us more confidence in the results.

	Best Value	Mean Value	Standard Deviation
Powell	-0.0213	-0.0183	0.0047
GACS	-0.0695	-0.0619	0.0052
DE	-0.0788	-0.0774	0.0017

Table 5.2 — Values of the objective function corresponding to *ECC* for different runs using Powell's, GACS, and DE approaches. All the runs were performed under the same conditions (except for Powell where different initial conditions were tried) on the same audio-video sequence.

Fig. 5.12 illustrates another aspect of the better behavior of DE algorithm with respect to Powell's: the weight values obtained at the end of each runs are more scattered in the parameter space with the latter. This indicates that the objective function is highly irregular and exhibits plenty of local minima. which makes Powell's method inadequate

While the high variation of the solutions found with Powell's method is not a surprise (as it is very sensitive to initial conditions), the instability of GACS solution seems intriguing.



(a) (b)

Figure 5.12 — Values of the MFCCs linear combination weights $\vec{\alpha}$ obtained on a given sequence with: (a) Powell's optimization algorithm with different initial guesses; (b) Different runs of DE. The continuous line connects the mean values of the obtained weights.

Let us look at the evolution of GACS and DE towards the solution over different runs. These are plotted on Figs. 5.13.(a) and 5.13.(b), for different runs.

As discussed previously, it has been noticed that the population losses its diversity during the GACS evolution. Also, the solution space is less systematically explored (especially the boundaries). As a result, GACS does not manage to reach solutions close to each others from run to run since it stops its evolution before reaching the global optimum. The solution reached can differ of 22% against 6% for DE. Fig. 5.14 displays the best runs for DE and GACS: DE reaches the solution found by GACS after more or less the same number of generations. Instead of stopping, it continues until a better minimum is found.

Another issue in using GACS and DE algorithms is the stopping criterion. One simple way consists in running the algorithm for an a priori defined number of iterations. However, the number of iterations needed to reach a good approximation of the global minimum depends on the data and on the inherent randomness of the algorithm. Thus this approach is unsteady. A more suitable criterion should be based on the analysis of the algorithm's evolution towards the global optimum. One may choose to stop if, during a number of iterations, the solution is not improved significantly. Even from this perspective, DE seems more convenient: from Figs. 5.12.(a) and 5.12.(b), it is clear that GACS exhibits long generations with no changes in the solution, possibly followed by slight improvements. This means that it is very hard to find a suitable stopping criterion for GACS, as we may always get an improvement after a long period of stagnation of the solution. So an early termination has the chance to leave the solution far from the best achievable one.

Definitely, the behavior of DE is preferable as we have steeper changes in the current solution and an early stop is not so dramatic from the perspective of the quality of the solution. All these considerations justify our choice of optimization algorithm for all subsequent experiments: we will use DE in its form given by Algorithm 1 for our study of different speaker detection criteria.

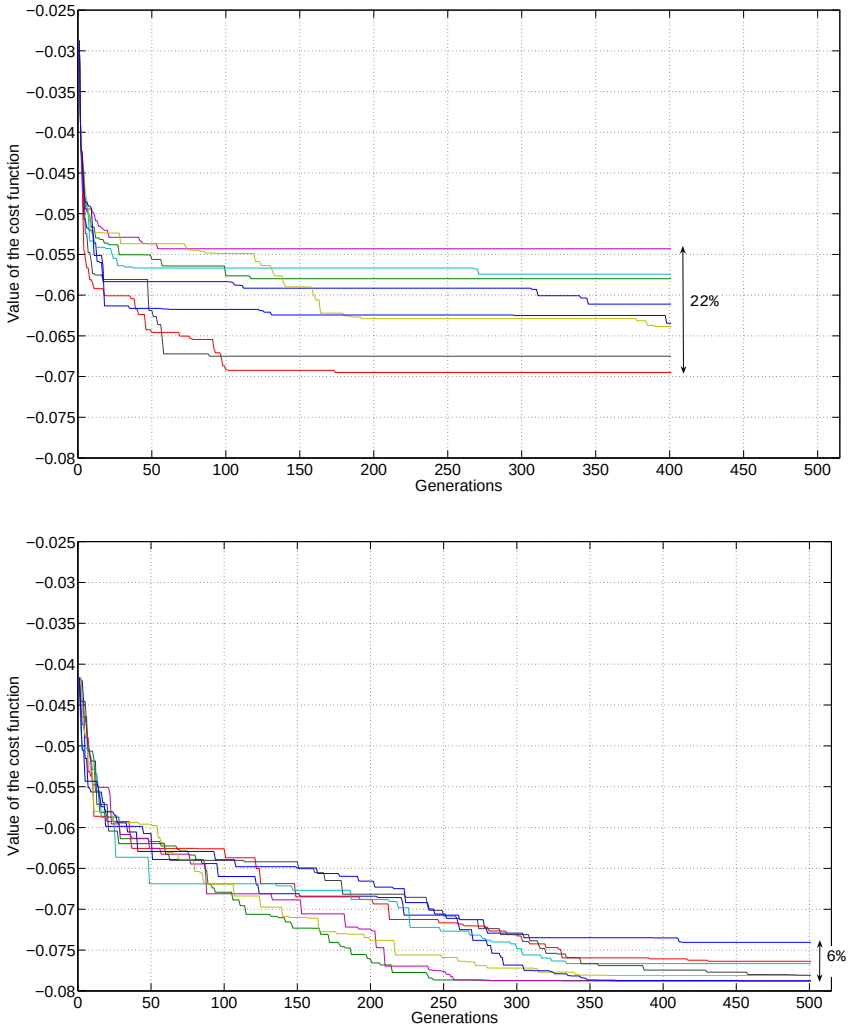


Figure 5.13 — Evolution of the cost function -*ECC* towards the solution for different runs using GACS (top) and DE (bottom) on a given audio-video sequence.

Notice that this comparative study has been described in [17].

5.8 Audiovisual speaker detection results

5.8.1 Experimental protocol

Two different sets of test sequences have been used*. The first set of sequences is part of an in-house data set containing five audio-video sequences of duration 4 seconds (labelled $1, 2, \ldots, 5$), each shot in PAL format (25 frames/second (fps), 44.1kHz stereo sound). In each sequence, two individuals are present, only one of them speaks during the entire sequence. Notice however that both are referred to as "speakers", since either one of them

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

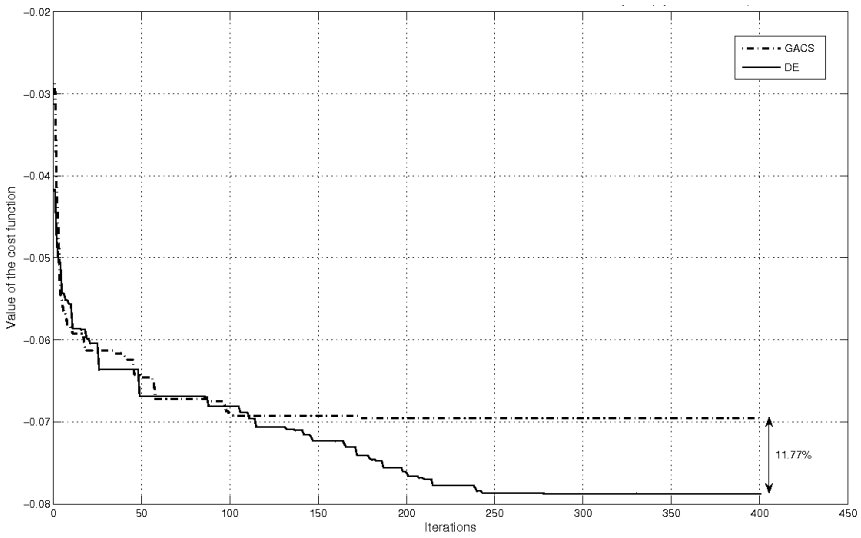


Figure 5.14 — Best run for GACS and DE.

may have uttered the recorded audio. These sequences are of increasing complexity, the fifth being the most challenging with the non-speaking individual moving randomly his head and lips. Frames extracted from two sequences are shown as an example in Fig. 5.15. These sequences, shot under controlled conditions, are used to assess the theoretical points developed in this chapter. For that purpose, the mouth regions are extracted based on manual localization initialized on the first frame. The duration of the sequences allows us to obtain a data sample set large enough to correctly estimate the pdfs. It also allows the speakers to remain still enough for the initial mouth localization to be valid throughout the sequence.

The second set of sequences is part of the CUAVE database [98]. The 11 two-speakers sequences *g11* to *g22* considered*, are shot in the NTSC standard (29.97fps, 44.1kHz stereo sound). On these sequences each speaker utters in turn two series of digits. The final seconds of the video clips, where both speakers read simultaneously different digit strings, have not been used since signal separation is not in our goal. The first seconds present challenging properties, making the detection task difficult: in some sequences the non-speaking person moves his lips and chin, sometimes even formulating the words without sounding them. A trade-off has to be found between the sample size required for correctly estimating the pdf, and a detection window offering enough flexibility to correctly deal with the speaker changes. Thus, the optimization is done over a $2s$ temporal window, shifted in one second steps over the whole sequence to make decisions once per second. The mouth regions are tracked along the sequence using the detector described in sec. 5.3.

For both sequence sets, the $N \times M$ mouth regions are extracted, with N and M varying between 22 and 57 pixels, depending on speakers' characteristics and acquisition conditions. Thus the video feature set (video sample) is composed of the $N \times M \times (T-1)$ values of the

*$g18$ has been discarded as it exhibits strong noise due to the compression

a b

Figure 5.15 — Typical frames extracted from the in-house test sequences. White rectangles delimits the extracted mouth regions. (a) Frame extracted from sequence 5; (b) frame extracted from the third sequence.

optical flow norm at each pixel location (T being the number of video frames within the analysis window, i.e. $T = 100$ for the in-house data set or $T = 60$ frames for the CUAVE database).

From the audio signal, $P = 12$ mel-cepstrum coefficients are computed using 30ms Hamming windows [51], [102].

Considering each mouth region and its associated video features, the MFCCs are projected on a new 1D subspace as defined in sec. 5.4. Let us point out here that no training set is used: the optimization and the detection are done for all the test sequences. As a result of the optimization, two sets of weights are obtained (one for each mouth region). They give the optimal linear combination of mel-cepstrum coefficients with respect to the optimization criterion (either *ECC* or *MIC*). Let us denote them $\vec{\alpha}_1^{opt}$ and $\vec{\alpha}_2^{opt}$, where the indices M_1 and M_2 indicate whether these weights result from the optimization performed on the first or second mouth region. Two corresponding audio feature sets derive from these weight sets: F_{A1}^{opt} and F_{A2}^{opt}.

Following the pattern recognition chain of Fig. 4.1, the features resulting from the extraction steps are now given as input of the MI-based classifier defined in § 4.4.1. Two pairs of mutual information values can be evaluated between the audio feature sets and the video feature of each mouth region. If F_{V1} denotes the video features of the first mouth region and F_{V2} those of the second, the two pairs of mutual information are given by:

$$\{I(F_{V1}, F_{A1}^{opt}) \, , \, I(F_{V2}, F_{A1}^{opt})\}, \tag{5.31}$$

$$\{I(F_{V1}, F_{A2}^{opt}) \, , \, I(F_{V2}, F_{A2}^{opt})\}. \tag{5.32}$$

First, a comparison of both *MIC* and *ECC* criteria is performed on the in-house sequences. As a result, *ECC* turned out to be indeed more discriminative than *MIC*. Therefore, *ECC* alone is then used on the same sequences to analyze the ability of the method to

Sequence	1	2	3	4	5
ΔI_{MIC}	73.54 %	76.18 %	91.67 %	69.64 %	52.13 %
ΔI_{ECC}	76.00 %	76.73 %	90.93 %	76.29 %	69.72 %

Table 5.3 — Normalized difference of mutual information measured in each mouth region for each of the 5 in-house test sequences, considering the audio features extracted with optimization criterion *MIC* or *ECC*, on the speaking mouth region.

extract audio features specific to speech production and to perform speaker detection. Finally, the discussion of the results leads to the definition of a more efficient criterion ΔECC stated by Eq. (5.9). Its performance on both sequence sets are presented and discussed in paragraph 5.8.4.

5.8.2 Comparison of optimization criteria *MIC* and *ECC*

The initial hypothesis is that *ECC* is more effective than the simpler *MIC*. The first set of experiments, carried out on the in-house sequence set, aims at testing this hypothesis. Therefore, the knowledge of the active mouth region is introduced *a priori* so that the optimization is only performed on this region, with each of the two optimization criteria successively. Using the resulting audio feature sets, the difference of mutual information between the speaking mouth region and the non-speaking one, normalized by the speaking mouth region MI (i.e. the *normalized difference of MI*), is measured for each of the five test sequences. Table 5.3 presents the results (ΔI_{MIC} and ΔI_{ECC} refer to the normalized difference of MI measured between the speaking and the non-speaking mouth regions when using optimization criterion *MIC* and *ECC* respectively).

Two observations can be made from these results. Firstly, the mutual information is always greater in the active mouth region, regardless the optimization criterion used, confirming that our scheme permits the detection of the current speaker. Secondly, we see that in 4 cases out of 5, the *ECC* criterion leads to larger difference between MI in the two regions. This indicates that the use of the *ECC* criterion gives rise to more discriminative features. Consequently, normalizing the mutual information by the entropy during the optimization allows to extract more specific information than using simply the mutual information alone, as stated in sec. 4.2.

5.8.3 Performance using *ECC*

From the first set of experiments we may conclude that *ECC* is more suitable as an optimization criterion for active speaker detection. This is why in the following we will focus only on its use and analyze its properties in detail. The purpose of the experiments described here is to assess the ability of our algorithm to extract audio features specific to speech production and to perform speaker detection. The tests are carried out on the in-house sequences.

Sequence	1	2	3	4	5
ΔI_{M_1}	76.00 %	76.73 %	90.93 %	76.29 %	69.72 %
ΔI_{M_2}	36.09%	-11.66	71.65 %	-0.66%	-17.28 %

Table 5.4 — Normalized difference of mutual information measured between the M_1 and M_2 mouth regions with the audio features obtained with optimization on mouth regions M_1 (I_{M_1}) and M_2 (I_{M_2}). The optimization criterion used in both case is *ECC*.

The capacity of the proposed method to act as a speaker detector is shown first. In contrast to the experiments described in § 5.8.2, no *a priori* knowledge of the active speaker is assumed. Then the technique described in § 5.8.1 is applied. Recall that the optimization is performed on each of the mouth regions (M_1 and M_2) and the mutual information between two pairs of audio and video features is measured as stated by Eqs. (5.31) and (5.32). If the approach is correct, the highest MI value should be measured between the video features of the speaking mouth and the audio features resulting from the optimization on the active speaker.

The values of MI are plotted in Fig. 5.16. We note that for all sequences (including the challenging seq. 5), the MI measured on mouth M_1 with $\vec{\alpha}_{opt}$ optimized on this same region is always strikingly greater than all the other three. Indeed, in all these sequences, M_1 is the speaking mouth, which gives 100% correct detections, a rather encouraging result.

Another issue that it is necessary to investigate is whether the features extracted from audio are specific to speech. For this, the difference between the normalized mutual information computed on mouth regions and the corresponding audio is measured as follows:

$$\Delta I_{M_1} = \frac{\max_i \left(I(F_{V_{M_i}}, F^{opt}_{A_{M_1}}) \right) - \min_i \left(I(F_{V_{M_i}}, F^{opt}_{A_{M_1}}) \right)}{\max_i \left(I(F_{V_{M_i}}, F^{opt}_{A_{M_1}}) \right)}, \quad \text{with } i = 1, 2, \tag{5.33}$$

$$\Delta I_{M_2} = \frac{\max_i \left(I(F_{V_{M_i}}, F^{opt}_{A_{M_2}}) \right) - \min_i \left(I(F_{V_{M_i}}, F^{opt}_{A_{M_2}}) \right)}{\max_i \left(I(F_{V_{M_i}}, F^{opt}_{A_{M_2}}) \right)}, \quad \text{with } i = 1, 2. \tag{5.34}$$

The results are listed in Table 5.4. It can be seen that $\Delta I_{M_1} > \Delta I_{M_2}$, as shown in Fig. 5.16. The new observation that can be made from these results is that $\Delta I_{M_1} > 0$ for all the sequences, whereas ΔI_{M_2} is sometimes negative. In other words, when the audio features are obtained by optimizing on the non-speaking mouth region, the difference of MI is sometimes favoring the non-speaking mouth (sequences 2, 4 and 5),and sometimes the speaking mouth (seq.. 1 and 3). So when optimizing on the non-speaking region, the features extracted cannot (and are not expected to) reflect any underlying relationship between audio and video. This result also appears in Fig. 5.16, since the mutual information measured between $F_{V_{M_1}}$ and $F_{A_{M_2}}$ is always smaller than the one measured between $F_{V_{M_1}}$ and $F_{A_{M_1}}$. Therefore the audio features can be said to be specific to speech production.

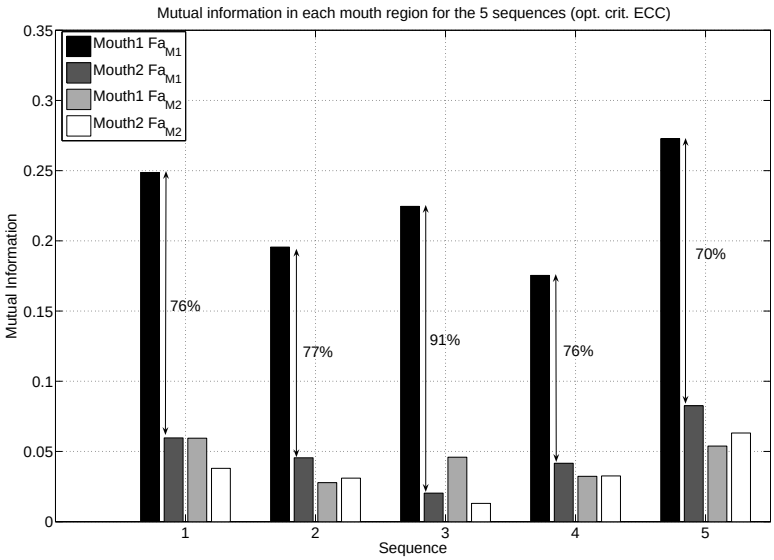


Figure 5.16 — Mutual information measured between the M_1 or M_2 mouth region features and the audio features obtained with optimization on mouth region M_1 or M_2 (Eqs. (5.31, 5.32)) for the 5 in-house sequences. The normalized difference of MI between the best value found and the best value found in the opposite mouth is indicated.

5.8.4 Results obtained with ΔECC on the in-house data set

Two optimizations were performed previously to decide who is the current speaker. They are now combined in a single optimization problem, which aims at maximizing the discrepancy between the two mouth regions. For this, the ΔECC, given by Eq. (5.9), is used. The result of the optimization is a vector $\vec{\alpha}_{opt}$ which generates a single audio feature vector. The latter is expected to maximize the mutual information with the video features of the active mouth region. This new detection approach has firstly been tested on the five in-house test sequences. Results are summarized in Table 5.5. The normalized difference of mutual information is always in favor of the active speaker, i.e. the correct speaking mouth region is always indicated. It is also interesting to note that the difference of mutual information here is greater than what was obtained with the previous ECC optimization scheme (Tab. 5.3). This stresses the benefit of using the video content related to each mouth region during the optimization.

Sequence	1	2	3	4	5
ΔI	84.23%	86.27%	95.55%	80.9%	76.15%

Table 5.5 — Normalized difference of mutual information measured between the speaking and the non-speaking mouth regions with the audio features obtained using ΔECC as cost function (tests performed on the in-house database).

5.8.5 Results obtained with ΔECC on the CUAVE database

To validate the results obtained with this simple detection scheme using ΔECC, experiments on the CUAVE database have been performed. Recall that a two second analysis window is shifted in one second steps over a given sequence. Due to the resulting overlap between the windows, the evaluation is restricted to the second half of each detection window, except for the very first window. The results are then evaluated based on the experimental framework proposed in appendix A: the ground truth for the evaluation window takes the label that mainly occurs over these 30 frames (*speaker 1* or *speaker 2*). Since our detector is not tuned to detect a silent state, the silent frames are not considered. Notice that the silent states can be easily detected prior to the detection by simply evaluating the mean energy of the audio signal. The results are listed in Table 5.6. The worst results are obtained for the sequence *g13*. The bad illumination conditions and the large movements of one of the non-speaker in the frontal plan may be an explanation of these bad results.

Sequence number	Correct detection rate (in %)
g11	69
g12	82
g13	53
g14	95
g15	89
g16	83
g17	100
g19	86
g20	92
g21	90
g22	95
Mean	85

Table 5.6 — Results on the CUAVE sequences, using the evaluation framework given in appendix A with evaluation on the last second of each detection window (silent windows are not considered).

As a comparison, the average rate of correct detections over the 11 sequences when using a simple motion-based detector (the mouth region with highest motion power value is labelled as the speaking mouth) is 60%. The detailed results are listed sequence by sequence in Table 5.7. The results are only slightly above a random speaker detection scheme. They indicate that the use of both audio and video information significantly improves the detection. Even in the case of seq.. *g13* where our results were not so good, the motion-based detection scheme ends up with worse results.

It is interesting to compare our results to those presented by Nock *et al.* in [95]. They compute the mutual information at each pixel location, considering the difference of pixel

Sequence number	Correct detection rate (in %)
g11	69
g12	53
g13	47
g14	68
g15	56
g16	89
g17	69
g19	21
g20	61
g21	75
g22	85
Mean	63

Table 5.7 — Results obtained if the mouth region with the highest motion power is labelled as the speaking mouth.

intensity as video features, and MFCC as audio features. In a first stage, the highest total MI value in the left and right of the image is assumed to indicate the current speaker (76% of correct detections in such a scheme). In a second experiment, the highest concentration of MI value in a $N \times M$ region indicates the speaking mouth. It is classified as correct if this region falls within a $K \times K$ pixels square centered around the true speaking mouth (K being equal to 200). They obtain 70% of correct detection with this detection scheme, which is the most directly comparable to ours since we also limit the MI measure to mouth regions.

As discussed in appendix A, the specific evaluation framework described in [95] differs from ours: the ground truth label for a detection window is given by the label of the central frame of the window. Such a ground truth labelling and the resulting result evaluation is not robust from our point of view since the consideration of another frame in the window as label can change drastically the ground truth. Moreover, choosing the central frame of the window implicitly means that the detection method needs the past and future frames to detect the current speaker. This discussion is detailed in appendix A. However, for a precise comparison of our results to those in [95], their evaluation framework has been used. As a result, our framework exhibits 77% of correct detection against 70% in [95]. Thus the optimization of the audio features as presented in this work leads to better speaker classification results. These results have appeared in [19] and [18].

5.9 Performance analysis through hypothesis tests

5.9.1 Performance of hypothesis testing as a classifier

The ability of hypothesis tests used as classifiers is discussed now. The evaluation of the possible gain offered by introducing a feature extraction step prior to the classification will be addressed in the next paragraph. Both of these performance evaluation tests have been published in [12] and [13].

The optimized audio features are put as input of the classifier, defined as the test function giving the best test of size α. The test sequences are those taken from the CUAVE database and already used in sec. 5.8. Obviously, the video features are also defined as in sec. 5.8.

The hypothesis test used as a classifier is the one defined by Test 4.31. Two potential speakers are present on each sequence, thus two tests can be defined as described in § 4.4.4, each test involving one speaker only. However, the experimental conditions are defined so as to eliminate the possibilities 3 and 4 where the two speakers are either both speaking or both silent (the silent window have been removed from the test set). For binary tests, a positive and a negative class have to be defined. We assume the positive class to be the class "speaker". The two possible positive classes are then "speaker1" and "speaker2". More precisely, since the experimental conditions imply that there is always one speaker speaking, the positive class is the label of the mouth region where the test is performed: i.e., "speaker1" for test1 (defined between the random variables F_A and F_{V1}), and "speaker2" for test2 (defined between F_A and F_{V2}). The two classes, "speaker1" (speaker on the left of the image) and "speaker2" (speaker on the right) are well balanced since their set sizes are both equal to 96. Table 5.8 compares the power of the tests for given sizes α.

	Test1			Test2		
β	37.9%	81.1%	90.5%	4.3%	24.7%	89.26%
α	5%	10%	20%	5%	10%	20%
η	0.41	0.25	0.16	0.55	0.45	0.25

Table 5.8 — Power of the tests "speaker1" versus "speaker2" (and vice versa) for different sizes α. The thresholds η defining the corresponding decision functions are also indicated.

Let us introduce now the accuracy of a test as the sum of the true positive and true negative rates divided by the total number of positive and negative instances [42]. Table 5.9 gives the classifier scores for the threshold corresponding to each test best accuracy: 87.0% and 85.42% for test1 and test2 respectively, obtained for thresholds $\eta_1 = 0.18$ and $\eta_2 = 0.19$.

These results indicate hypothesis test as a good method for assigning a speaker class to mouth regions, with a given α-β trade-off. The classifier produces better relative instance scores for test1. However, the thresholds giving the best accuracy values are about the same for the two tests. As mentioned in [105], accuracy is often used as the primary evaluation

	Test1		Test2	
	Positive class	Negative class	Positive class	Negative class
β	87.5%	86.5%	91.7%	79.2%
α	13.5%	12.5%	20.8%	8.3%

Table 5.9 — Detection probabilities β and false-alarm rates α for each class of each test at its best accuracy value.

metric when building classifiers, though it assumes that the class priors will be constant and relatively well-balanced which is rarely the case in the real world. In this case however, this can be assumed to be true for the two classes "speaker1" and "spaker2". Actually, the facts that the best accuracy values are obtained for a similar threshold in both case tends to indicate that this threshold is not speaker dependent. In other words, the performance of the classifier are not speaker dependent. As a result, the test set can be considered another way: the two possible classes become now "speaker" (positive class) and "non-speaker" (negative class), confounding speaker 1 and 2 in a single label. The test set counts then 192 test points instead of 96. As the size of the test set is increased, more robust conclusions can be drawn. Also, we evaluate now the capacity of the classifier to discriminate between a speaker and a non-speaker instead of between two speakers, which is much more our initial target. Table 5.10 compares the power of this test for given sizes α. Again, these

β	22%	40%	85%
α	5%	10%	15%
η	0.455	0.410	0.250

Table 5.10 — Power of the test "speaker" versus "non-speaker" for different sizes α. The thresholds η defining the corresponding decision functions are also indicated.

results indicate hypothesis test as a good method for discriminating between speaking and non-speaking mouths with given α-β trade-off (thus greater adaptability to changes of the target condition or the classification requirement). It is particularly interesting to see that accepting 15% of false-alarms instead of 10% leads to a great jump of the correct detection probability since it is almost doubled.

5.9.2 Evaluation of the classification chain performance

The advantage of using optimized audio features against simple ones at the input of the classifier is now discussed. The last test of the previous paragraph is retained: the positive and negative classes are respectively the "speaker" and the "non-speaker" classes.

The audio features are defined as the optimal linear combination of MFCCs. The ROC graph of this test is plotted on Fig. 5.17 (solid line). The good performance of the Neyman-Pearson classifier assessed in the previous paragraph can now be visually evaluated: the slope of the curve is high for false-alarm rates smaller than about 15%, meaning that the percentage of correct detections increases quite quicker than the false-alarm rate in the

beginning.

If now non-optimized audio features, defined as the mean value of the MFCCs are put as input of the classifier, another ROC curve indicating the performance of the classifier for different $\alpha - \beta$ trade-offs can be plotted. It is represented by the dashed line in Fig. 5.17. Except at the very beginning (α and β small), it is below the ROC curve of the classifier fed with optimized features. The later perform better in the conservative part of the ROC space in particular (northwest region). Table 5.11 sums up some interesting values

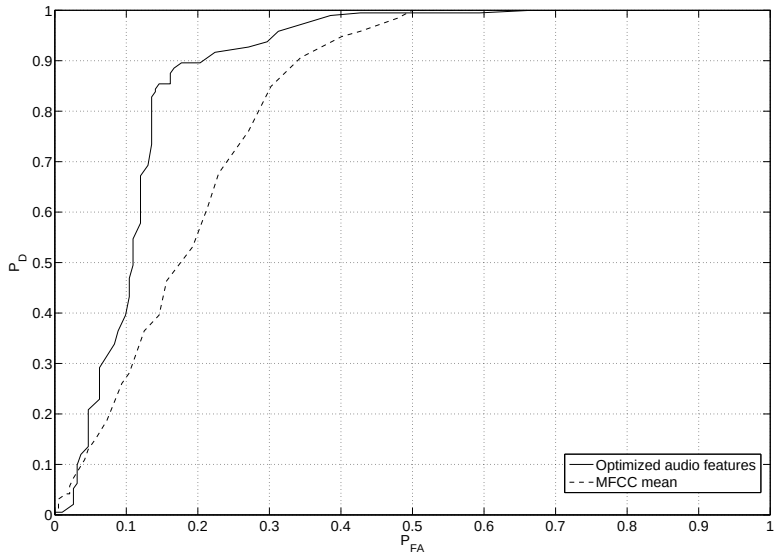


Figure 5.17 — ROC curves for the test "speaker" versus "non-speaker" when optimized (plain line) and non-optimized (dotted line) audio features are used as input of the classifier.

attached to the ROC curve such as the area under the curve (AUC), or the accuracy with corresponding thresholds. Once again, the AUC is higher when optimized audio features

Input features	MFCCs mean	Optimized audio features
AUC	0.81	0.88
Accuracy	78.1%	85.9%
β	90.6%	89.6%
α	34.4%	17.7%
η	0.13	0.18

Table 5.11 — Area under the curve, maximal accuracy and corresponding power β, size α and threshold η for each kind of input audio feature.

have been used. This means that in this case, the capacity of the classifier to discriminate between the "speaker" and "non-speaker" classes is increased. Actually, if looking at the distribution of the two classes shown in Fig. 5.18, the effect of the feature extraction step appears clearly: the two classes are much more separated after the feature extraction step (Fig. 5.18 (bottom)) than before (Fig. 5.18 (top)). As a result, the classifier can only be more accurate.

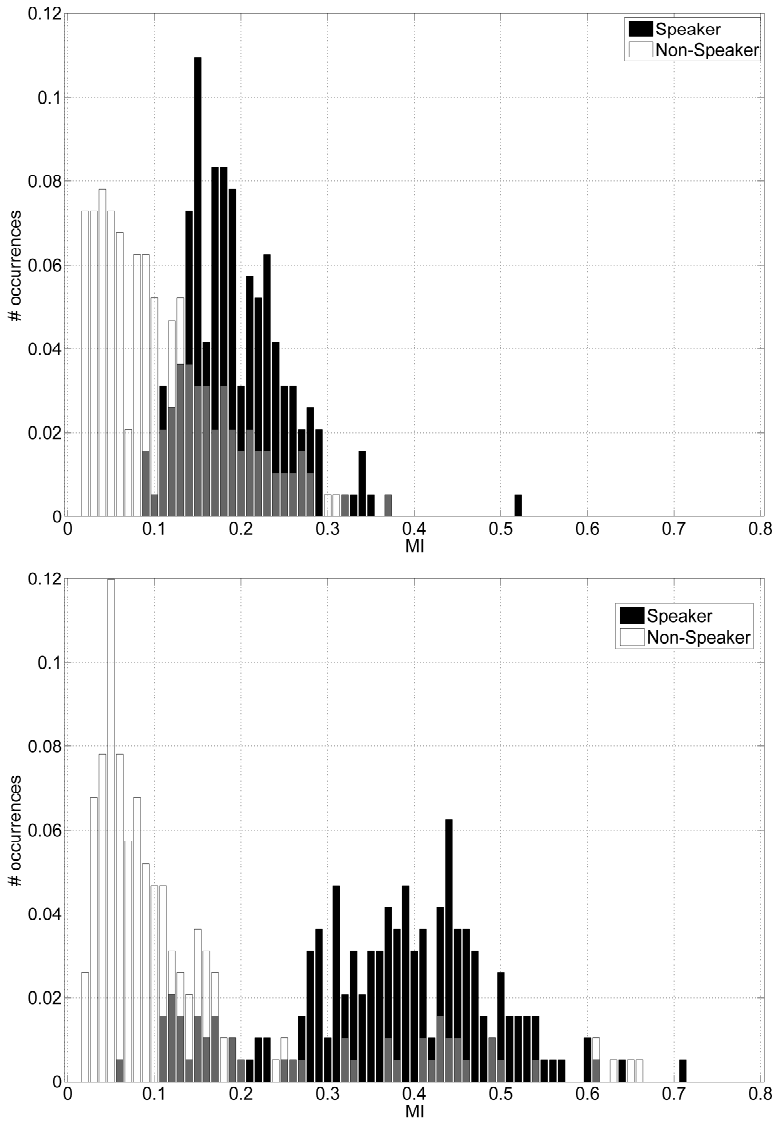


Figure 5.18 — Distribution of the "speaker" and the "non-speaker" classes for the classifier fed with non-optimized audio features (top) and optimized audio features (bottom). The bins colored in gray correspond to overlapping points between the two distributions.

Whatever the way of considering the problem, these results show that introducing a feature extraction step as described in sec. 4.2 prior to the classification increases the performance of the system.

As last remark, let us say a few words about extending the classification to the four cases described in § 4.4.4 (including the silent and the both-speaking cases). This scheme has not be extensively tested and has been let apart for future works. However, a short analysis based on a few number of silent windows tends to show that the cases 3 and 4 can not be detected if the audio features are optimized with optimization criterion ΔECC (Eq. (5.9)). Indeed, this criterion forces the algorithm to increase the efficiency coefficient in one mouth region while decreasing it in the other one. It is in contradiction with the statement that

both efficiency coefficients can be low or high. However, further investigation are required before coming to any conclusion. In particular, the use of the criterion ECC for performing an optimization in each mouth region could give rise to better results for such purposes.

5.10 Discussion

The pattern recognition system defined in chapter 4 has been applied to the problem of identifying the current speaker among several candidates in audio-video sequences. Each step of the scheme displayed in Fig. 4.1 has been considered and the corresponding choices have been discussed.

The data, acquired by a single camera and microphone, are firstly processed to get representations emphasizing their speech-related content: a specific energy representation for the audio and a motion-based representation for the video. A feature extraction step is then performed to decrease the feature dimensionality while increasing their specificity to speech. For the video, this is achieved by restricting the region of interest to the mouth regions. The information theoretic feature extraction framework developed in sec. 4.2 is applied to optimize the audio features using jointly the video information. The extracted audio features are made up of an optimal linear combination of P MFCCs. To this end, the optimization problem has been precisely defined, including the cost function and the optimization method. Actually, the MI-based cost function turned out to be complex and plagued by many local optima. For this reason, three optimization methods, one local and two global have been tested in turn and their performance compared. As a result, the so-called Differential Evolution algorithm [127] outperformed the two others.

A study of two optimization criteria that can be used in the multimodal feature extraction framework has been carried out. Results shown that the best performing criterion (namely, ECC) is able to extract audio features that are specifically related to the speaker video features. Using only these extracted features, the algorithm performs detection of the current speaker with 100% correct detection on 5 in-house test sequences.

In order to optimize the detection in the case of two-people sequences, a third optimization criterion (ΔECC) has been introduced and tested on the same sequence set as before as well as on the more widely used CUAVE database [98]. This criterion aims at simplifying the detection scheme, as well as improving the audio feature specificity by taking advantage of the video information related to both mouth regions. Indeed, the resulting audio features have been shown to be even more specific than with the previous optimization criterion. A number of experiments have therefore been carried out on 11 sequences of the CUAVE database to assess and compare the performance of this ΔECC-based detection method to the results presented in [95]. In the latter, MFCCs are used as audio features, without any optimization. The better results achieved by our method indicate that optimizing the features improves the classifier performance. The results are also significantly above those obtained by a simple motion-based detector, supporting the advantage of a multimodal approach.

Only two potential speakers are present in these test sequences but the method can easily be extended to sequences containing more speaker candidates using *ECC* as optimization criterion. These speakers should remain face to the camera. However, it is not a problem if they move, provided the mouth detector is able to deal with moving faces.

To better analyze the performance of the system, and in particular to evaluate the gain offered by the multimodal feature extraction step, the classifier is feed in turn with the optimal linear combination and with a simple average of the MFCCs. A ROC curve analysis is carried out. The use of the optimized audio features is shown to increase notably the classifier performance.

The next chapter handles the reverse issue of extracting optimized video feature using the information theoretic feature extraction framework.

6 Video feature extraction

6.1 Introduction

The problem at hand in this chapter is the same as the one discussed in the previous chapter. The PR framework is used for speaker detection. The detector still operates with a single camera and microphone in the data acquisition step. However, the focus is now put on the optimization of the video features instead of the audio ones, using the feature extraction framework developed in sec. 4.2.

The first processing step to be applied to the raw audiovisual data is the selection of a representation. This step is described in sec. 6.2. An energy-based representation is defined for the audio. The most suitable choice for representing the video content would have been the motion in the mouth region. However, only an estimation of this motion is available since this 3D phenomenon has to be measured from a sequence of 2D images. This estimation, the optical flow (OF), relies on the intensity gradient of the image. In sec. 6.3, a probabilistic model of the relationships between the audio, the motion and the image intensity gradient is proposed through graph theory, in the specific case of a speaking mouth. A link is found with the information theoretic estimator defined in sec. 4.2. Following this discussion, a framework for optimizing the video features, i.e. the OF, using the audio information, is proposed in sec. 6.4. Its potential is explored in sec. 6.5 on a simple one-speaker sequence. This study reveals the non-convergence of the cost function and the necessity of redefining the single-objective (SO) optimization problem as a multi-objective (MO) optimization task. Leaving this challenging task for future work, the chapter proceeds in sec. 6.6 with a simplified convergent version of the SO optimization problem in order to investigate the capacity of the method for speaker detection. The deterministic gradient-free approach chosen for solving the problem is described in the same section. In sec. 6.7,the approach is

tested on several audiovisual sequences and the results are compared to those obtained with non-optimized features. Finally, the performance of the entire pattern recognition system is analyzed in sec. 6.8 using the Neyman-Pearson hypothesis testing approach proposed in sec. 4.4.

6.2 Signals representation

6.2.1 Audio representation

In the PR system of Fig. 2.1, the acquisition of the raw data is followed by a data representation step. A mel-cepstrogram representation similar to the one described in § 5.2.3 is chosen for the audio information. The computation of the MFCCs leads to a P-dimensional audio feature, with $P > 1$. To reduce the dimensionality, only the mean value of the P MFCCs is used, similar to what has been done in sec. 5.9. Such features gave rise indeed to relatively good results for speaker detection. A 1D speech audio feature A is then finally obtained.

6.2.2 Video representation

As discussed in the previous of this thesis, psychophysiologic evidences point out the motion in the mouth region as the video clue related to the audio information: when speech is produced, the movements of the articulators (the lips in particular) induce facial motions.

The motion is a phenomenon occurring in a 3D space but only a sequence of 2D images is available. The true motion can not be directly measured but must be inferred from this sequence. This estimation, or apparent motion, is based on the information carried by the spatio-temporal variation of the image intensity. By assuming that the changes in the image brightness along time in the 2D sequence is only due to the 3D motion of the patterns *brightness constancy assumption*), an approximation of the motion, known as *optical flow*, is obtained. It is then only possible to speak about a probability of having a motion V at a given location given the spatio-temporal gradient value G of the intensity at that location. This conditional probability is denoted by $P(V|G)$.

For a classifier to perform at its best, it should be feed in with the most suitable features for the problem at hand. Here, only an approximation of the representative feature (the true mouth motion) is available. It is then necessary to discuss about the possible consequences on the classifier performance (the latter being defined according to sec. 4.4) as well as about a possible optimization of the optical flow, able to improve this performance. To this end, a better understanding of the relationships between the audio, the motion and the image intensity gradient is firstly required, in the special context of speech production.

6.3 Relationship between audio, motion and intensity gradient

6.3.1 Probabilistic model using graph theory

Let A, V and G be three random variables representing respectively the acoustic speech (represented by some tractable features), the motion and the intensity gradient (the region of support of V and G being limited to the mouth region extracted from the image sequence). Different probability models can be proposed, each one describing different relationships between these three rvs.

As previously stated, during speech, the acoustic signal A and the visual motion V of the mouth region are related so that some knowledge of V can offer some knowledge about A and vice et versa. Since the motion is actually estimated from the image intensity gradient G, a part of the information in V that is related to A is also present in G. Then, if V is unknown, the knowledge of G can help in recovering A. If V is known however, G becomes useless in learning something about A. A similar reasoning shows that the knowledge of A is useless to learn more about G when V is known. Speaking in probabilistic terms, A and G are said to be independent conditionally to V ($p(A|V,G) = p(A|V)$, or in a shorthand notation, $A \perp\!\!\!\perp G|V$).

A probabilistic graphical model can represent in a simple way the relationships just discussed between A, V, and G: A is both the child and parent* of V as is V for A. In the same way, the motion is estimated from the gradient of the image, thus the motion V is the child of G. However, the change in the image brightness, thus the gradient, is due to the motion (brightness constancy assumption). Therefore, the gradient is also a child for V.

The undirected graph, or Markov chain, shown on Fig. 6.1 is then the most appropriate for describing the relationships between the three random variables. According to this

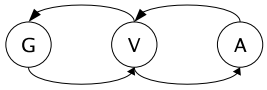


Figure 6.1 — Graphical representation of the probabilistic relationships between the random variables A, V, G. These random variables form a Markov chain where A and G are independent conditionally to V.

model, the joint probability of the three random variables is:

$$P(A,G,V) = P(G)P(V|G)P(A|V). \tag{6.1}$$

*The terms "child" and "parent" are used here in the graph theory sense [20], and do not stress necessarily a physic causal effect between the rvs.

Since it is an undirect graph, the following relationships can be equivalently stated:

$$P(A,G,V) = P(G)P(V|G)P(A|V) \tag{6.2}$$
$$= P(A)P(V|A)P(G|V) \tag{6.3}$$
$$= P(G|V)P(A|V)P(V). \tag{6.4}$$

6.3.2 Justification of the efficiency coefficient based estimator

Let us recall quickly the principle of the MI-based classifier defined in § 4.4.1: it rates the MI between acoustic and visual features extracted from different mouth regions. Then the "speaker" label is assigned to the mouth region with the highest MI.

As mentioned before, the most representative visual feature for speech would be the mouth motion V. But instead of working with A and V directly, the classifier defined in sec. 4.4 works in fact with the audio features and the optical flow. The data processing inequality (defined in sec. 3.2) applied to the Markov chain of Fig. 6.1 leads to the following equation:

$$I(A;V|G) \leq I(A;V), \tag{6.5}$$

where I is the mutual information. Eq. (6.5) indicates that the mutual information between the optical flow and the audio features is smaller than the mutual information between the true motion and the audio features. Hence the use of the optical flow instead of the motion as video feature makes the classifier less discriminative. However if $I(A;V|G)$ is maximized, it will tend towards $I(A;V)$. In other words, by estimating the optical flow so that the mutual information between this one and the audio feature is maximized, the performance of the classifier should increase. Indeed, a closer approximation of the ideal classifier function (the MI between the acoustic speech and the speech mouth motion) will be obtained. However, for this last statement to be true, the optimization of the optical flow should be done while keeping the conditional entropy $H(V|G)$ constant. Indeed, if this entropy increase, since $H(G)$ is constant, it reduces the dependency between the motion and the image intensity gradient. The Venn diagram relating the entropies and the MI of the three random variables A, V, and G, with $A \perp\!\!\!\perp G|V$, is drawn in Fig. 6.2.

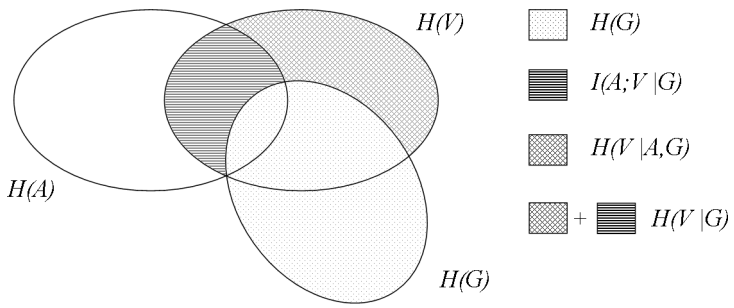


Figure 6.2 — Venn diagram representing the entropies and mutual information between the three random variables A, G and V, whose statistical relationship is stated by the graphical model of Fig. 6.1.

This result coincides with the information theoretic framework proposed in sec. 4.2. The latter stated that to decrease the probability of classification error, a multimodal feature extraction step should be carried out, where the EC, is maximized. We recall here the definition of this efficiency coefficient:

$$e(F_A, F_V) = \frac{I(F_A, F_V)}{H(F_A, F_V)} \in [0, 1], \tag{6.6}$$

where F_A and F_V are random variables modelling the features extracted respectively from the audio and the video signals. If only the video features are optimized, the efficiency coefficient becomes:

$$e(F_A, F_V) = \frac{I(A, F_V)}{H(A, F_V)}, \tag{6.7}$$

$$= \frac{I(A, F_V)}{H(F_V)} \in [0, 1]. \tag{6.8}$$

The transition from Eq. (6.7) to Eq. (6.8) is obtained by considering that A remaining constant during the extraction process, the variation of the joint entropy $H(A, F_V)$ is only due to the variation of the marginal entropy $H(F_V)$ which can be therefore directly used to constrain the extraction process.

If F_V is interpreted as the optical flow, Eq. (6.8) states that the mutual information between the audio feature and the optical flow must be increased, while minimizing the marginal entropy of the optical flow. It is then the mathematical formulation of the previous lines of reasoning.

6.4 Audio constrained optical flow

6.4.1 Standard optical flow estimation

Horn and Schunck have been among the first to formulate in [64] the optical flow as an approximation of the true motion from the image intensity gradient. The theoretical framework relies on the *brightness constancy assumption*:

$$E_x u + E_y v + E_t = 0, \tag{6.9}$$

where E_x, E_y, E_t denote the partial spatio-temporal derivatives of the image intensity E, $u = \partial x/\partial t$ and $v = \partial y/\partial t$ are the horizontal and vertical components of the optical flow, denoted then as a vector $\vec{F}_V$.

Motion estimation is hampered by the so-called aperture problem: the region for which the optical flow is estimated must be large enough for this motion to be catched. However, the larger the region, the more probable the brightness constancy assumption to be violated. Actually, the OF, relying on the brightness constancy assumption, leads to an ill-posed problem since the solution is not unique: the optical flow can only be determined in the direction parallel to the local intensity gradient. The problem is usually regularized by

assuming the flow to be locally smooth, i.e. to present a spatial coherence. In [64], this regularization is introduced through a quadratic smoothness constraint under the form of a Laplacian error term. Eq. (6.9) becomes then an error to be minimized:

$$\xi^2 = \int_{x\in\Omega_x}\int_{y\in\Omega_y}(\xi_b^2 + \lambda\xi_c^2)dxdy, \tag{6.10}$$

where ξ_b^2 is the error term related to the brightness constancy assumption and ξ_c^2 the error related to the Laplacian regularization term (departure from smoothness in the velocity flow), weighted by the regularization parameter λ.

$$\xi_b^2 = (E_x \cdot u + E_y \cdot v + E_t)^2, \tag{6.11}$$

$$\xi_c^2 = \nabla^2 u + \nabla^2 v. \tag{6.12}$$

Finally, the optical flow $\vec{F}_V = (u, v)$ on a given region Ω_{xy} can be estimated by solving the following minimization problem:

$$\vec{F}_V = \arg\min_{u,v} \xi^2, \quad \text{with } u, v, \in \mathbb{R}. \tag{6.13}$$

The solution to Eq. (6.13) should fit the best the brightness constancy assumption (error term ξ_b^2) while penalizing the large gradients (error term ξ_c^2) [4]. In [64], the authors propose an iterative solution to Eq. (6.13) through the calculus of variation:

$$u^{n+1} = \bar{u}^n - E_x\frac{[E_x \cdot \bar{u}^n + E_y \cdot \bar{v}^n + E_t]}{(\lambda + E_x^2 + E_y^2)}, \tag{6.14}$$

$$v^{n+1} = \bar{v}^n - E_y\frac{[E_x \cdot \bar{u}^n + E_y \cdot \bar{v}^n + E_t]}{(\lambda + E_x^2 + E_y^2)}. \tag{6.15}$$

6.4.2 Multimodal optimization of the optical flow parameters

Two parameters must be set up in order to find a solution to the Eq. (6.13), i.e. in order to correctly estimate the optical flow in the region of interest. These parameters are the number of iterations ι and the weight λ associated to the Laplacian regularization term.

The number of iterations allows the system to reach its equilibrium point. From the OF estimation point of view, as the number of iterations increases, the flow propagates from the edges in the images inwards the non-textured regions.

The regularization term forces the velocity vectors to have a coherent direction in a given neighborhood: the larger the weighting term, the stronger this constraint. More precisely, the weighting term, or regularization parameter, λ, defines a trade-off between the relative importance of the brightness term ξ_b^2 and the smoothness term ξ_c^2 in the cost function (6.13). The less reliable the brightness constancy assumption, the higher λ.

In this work, we propose to introduce a third constraint on the flow estimation, when speech mouth motions are concerned. In such a case, we argue that the audio information jointly emitted with the video signal can be used to constraint the optical flow estimation in

the mouth region. The estimation of the mouth motion is a challenging issue. The mouth is a non-rigid structure where complex deformations like bending or twisting occur. As pointed out by Black and al. in [21], the brightness constancy assumption is often violated: the reflected light areas change with variations of the mouth shape and occlusions are very common (apparition of disappearance of the tongue or the teeth for example). Some illustrative examples are depicted in Figs. 6.3. Obviously, these violations are inherent to

Figure 6.3 — Example of speech mouth motions, with violations of the brightness constancy assumption: the teeth are appearing switching the luminosity in the region from black to white. The shadowed region on the left-hand side of the mouth region, or the reflection of the incident light on the bottom lip are also changing. (Extracted from a CUAVE sequence [98]).

the speech production process. Therefore, the introduction of a constraint related to this speech production process should help in the OF estimation. For this purpose, the statistical dependence between some audio energy features and the optical flow vectors is taken into acount through a MI-based measure: the efficiency coefficient defined in Eq. (6.6).

More precisely, the audio related constraint is introduced in the optical flow estimation by evaluating the efficiency coefficient between fixed audio features and varying video features. The latter correspond to optical flow fields obtained with different parameters ι and λ. The parameters for which the efficiency coefficient between the audio features and the optical flow field is maximal are the optimal ones. It must be noticed that this constraint does not aim at improving necessarily the accuracy of the estimated flow. Rather, it intends to emphasize the information specific to speech that can be catched by the OF.

Let $\vec{F}_V$ denotes the optical flow estimated in the speaking mouth region and let A be some features extracted from the co-occurring acoustic speech signal. The objective function to be maximized* is given by:

$$f(\iota,\lambda) = -I(\vec{F}_V(\iota,\lambda), A)/H(\vec{F}_V(\iota,\lambda)), \tag{6.16}$$

$$= -e(\vec{F}_V(\iota,\lambda), A). \tag{6.17}$$

So the optimization problem consists in finding $(\iota,\lambda)_{opt}$ by solving:

$$(\iota,\lambda)_{opt} = \arg\min_{\iota,\lambda} f(\iota,\lambda) \quad \text{with } \iota \in \mathbb{N}^*, \lambda \in \mathbb{R}_+. \tag{6.18}$$

6.4.3 Statistical considerations

The optimization problem must be cast in a probabilistic framework since it deals with mutual information. Let A and $\vec{F}_V$ be viewed now as random variables associated to the

*To be coherent with the standard formulation of optimization problems, the maximization problem could be turned into a minimization problem by simply changing the sign of the objective function.

audio and video features and defined respectively over the sample spaces Ω_A and Ω_{F_V}. Let the acoustic and visual signals be observed on the same temporal window $[1 \dots T]$ and the mouth region being of size N pixels.

The video features associated to the random vector $\vec{F}_V \in \mathbb{R}^2$, are the optical flow values $\vec{f}_v = (u, v)$ estimated between two consecutive images of the mouth region. For a given analysis window, there are then $\tau = T-1$ observations. The video sample is defined according to the statistical *case III* presented in sec. 5.5: one video rv is associate to the whole mouth region. However, instead of a 1D random variable, we have now a 2D random vector whose sample comprises $N \times \tau$ observations: $\{\vec{f}_v(k)\}_{k=1,\dots,\tau \cdot N} = \{(u(k), v(k))\}_{k=1,\dots,\tau \cdot N}$ with $\vec{f}_v \in \mathbb{R}^2$, and u, $v \in \mathbb{R}$.

Using the Parzen estimator, a (discrete) estimate of the probability density function of the video random vector is obtained:

$$p_{\vec{F}_V}(\vec{f}_v(k)) = \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{N} \sum_{n=1}^{N} K_{h_{f_v}}(\vec{f}_v(k), \vec{f}_v(t \cdot n)) \quad \forall \vec{f}_v(k) \in \Omega_{F_V}, \tag{6.19}$$

where $K_{h_{f_v}}$ is a kernel function with smoothing parameter h_{f_v}. Distinguishing between the horizontal and vertical velocity components, u and v, and using normal kernels with smoothing parameters h_u and h_v (see sec. 4.3), Eq. (6.19) becomes:

$$p_{\vec{F}_V}(u(i), v(j)) = \frac{1}{\tau \cdot N} \sum_{q=1}^{\tau \cdot N} K_{h_u}(u(i), u(q)) \cdot K_{h_v}(v(j), v(q)), \quad \forall (u(i), v(j)) \in \Omega_{F_V}. \tag{6.20}$$

The audio features have been defined in sec. 6.2 as the mean value of the P MFCCs. The audio and the video signals are observed on the same temporal window and the audio features are down-sampled to the video frame rate. Thus the audio sample associated to the random variable A consists in the set of 1D observations, or outcomes, $\{a(t)\}_{t=1 \dots \tau}$, with $\tau = T - 1$. The marginal pdf of the audio random variable A is:

$$p_A(a(i)) = \frac{1}{\tau} \sum_{t=1}^{\tau} K_{h_a}(a(i), a(t)) \quad \forall\, a(i) \in \Omega_A, \tag{6.21}$$

where K_{h_a} is a kernel function with smoothing parameter h_a.

The joint probability between the audio and the video random variables is given by:

$$p_{A,\vec{F}_V}(a(i), \vec{f}_V(j)) = \frac{1}{\tau \cdot N} \sum_{t=1}^{\tau} K_{h_a}(a(i), a(t)) \cdot \sum_{n=1}^{N} K_{h_{f_v}}(\vec{f}_V(j), \vec{f}_V(t \cdot n)). \tag{6.22}$$

Equivalently:

$$\begin{aligned} p_{A,\vec{F}_V}(a(i), (u(j), v(k))) = & \frac{1}{\tau \cdot N} \sum_{t=1}^{\tau} K_{h_a}(a(i), a(t)) \cdot \\ & \sum_{n=1}^{N} K_{h_u}(u(j), u(t \cdot n)) K_{h_v}(v(k), v(t \cdot n)). \end{aligned} \tag{6.23}$$

The mutual information corresponding to these statistical considerations is:

$$I(A,\vec{F}_V)=\sum_{i\in\Omega_A}\sum_{j\in\Omega_V} p_{A,\vec{F}_V}(a(i),\vec{f}_v(j))\cdot\log\frac{p_{A,\vec{F}_V}(a(i),\vec{f}_v(j))}{p_A(a(i))\cdot p_{\vec{F}_V}(\vec{f}_V(j))}. \tag{6.24}$$

6.5 Feasibility study

6.5.1 Experimental framework

To investigate the validity of the previous development, a first set of experiments is carried out, where no optimization is performed. Simply, one parameter ι or λ being kept fix, the second one is varied. This results in different estimations of the optical flow for a given frame set. The efficiency coefficient between these resulting video features (the video samples being defined as stated in § 6.4.3) and the audio feature extracted from the co-occurring acoustic signal is computed. The results, presented and discussed in details in this section, show that the EC varies as the parameters used to estimate the flow vary.

The tests are performed on an audiovisual sequence of duration 25 seconds (500 frames) shot in the PAL standard (25 fps, 44.1kHz stereo sound) where one individual is speaking, face to the camera (seq_1). A frame taken from this sequence is shown as an example in Fig. 6.4.

Figure 6.4 — Frame taken from seq_1. The mouth region is indicated by the white rectangle.

Another 20 seconds record of the same individual is also used, where this person is remaining silent (seq_2). In this case, the optical flow in the mouth region estimates natural motions of the mouth, unrelated to speech (swaling, smiling) or intensity variations due to background noise.

The mouth region (a rectangular region encompassing the lips) is extracted and tracked along the two sequences following the methodology described in sec. 5.3.

Two sets of audio features (mean value of $P = 12$ MFCCs, computed with 30ms Hamming windows) are extracted from two kinds of audio signal: the first audio signal ($audio_1$) corresponds to the normal acoustic speech record, while the second one ($audio_2$) is obtained

by temporally reversing the first audio signal. This second audio signal and features extracted from it are then no more temporally related to the speech mouth motions of seq_1 but exhibit the same global energy level than $audio_1$.

The extracted mouth region is of size 50×38 pixels. For computational efficiency, it is downsampled to 25×19 pixels. According to the reasoning hold in sec. 5.2, the number of iterations should be fixed to half the height of the mouth region. However, the latter encompasses narrowly the lips, discarding the chin, contrary to the region extracted in chapter 5. Thus, there are not two but one region of interest (the lips but not the chin) and the number of iterations is fixed to the height of the mouth region: $\iota = 19$.

The range of the audio and video rvs must be defined in order to estimate the probabilities at given points of their respective sample spaces. As in chapter 5, the audio features are normalized by 127, which was shown empirically to be large enough for all the audio features to be in the interval $[-1, 1]$. The video feature values must lie between the largest negative and positive velocities authorized by the specifications of the mouth region. The method described in sec. 5.5, where the normalization factor was defined with respect to a reference sequence, is not used here since no comparison to other sequences is intended in this set of experiments. Moreover, the mouth region being more tightened to the lips, the maximum velocity V_{max} corresponds roughly to the maximal opening of the lips between two frames, i.e. to a displacement per frame of half the mouth region height. The optical flow values are then normalized by $V_{max} = 10$, to lie in $[-1, 1]$. Notice that the horizontal components of the optical flow, generally smaller than the vertical component values, are however normalized by the same vertical normalization factor V_{max} for simplicity reasons.

6.5.2 Influence of the λ parameter value on the EC

In a first experiment, λ is varied exponentially from 2 to 2^{13}, ι being kept fix to 19. The OF is estimated on the two test sequences, using in turn each of the couples of parameters (ι, λ). The EC is computed between the resulting optical flows and the audio features extracted from $audio_1$ and $audio_2$. The evolutions of the EC as a function of λ for each of these four cases are shown in Fig. 6.5.

As the weight of the regularization factor λ increases, the EC increases up to a maximum. Once this maximum has been reached, a further increase of λ makes the EC to decrease. The regularization parameter λ is a weight associated to the quadratic smoothness term ξ_c^2 appearing in Eq. (6.10). It allows one to trade-off between the relative importance of the brightness term (Eq. (6.11)) and of the smoothness constraint term (Eq. (6.12)). This smoothness constraint penalizes the high gradients and forces the velocity vectors to have a coherent direction in a given neighborhood. As λ increases, the optical flow (OF) becomes more accurate in a first time: the incoherences in the flow, due to violation of the brightness constancy assumption (i.e. to noise) are smoothed. Then the smoothness becomes too important and destroys the visual information.

It can also be observed that, contrary to what was awaited, the evolution of the EC is

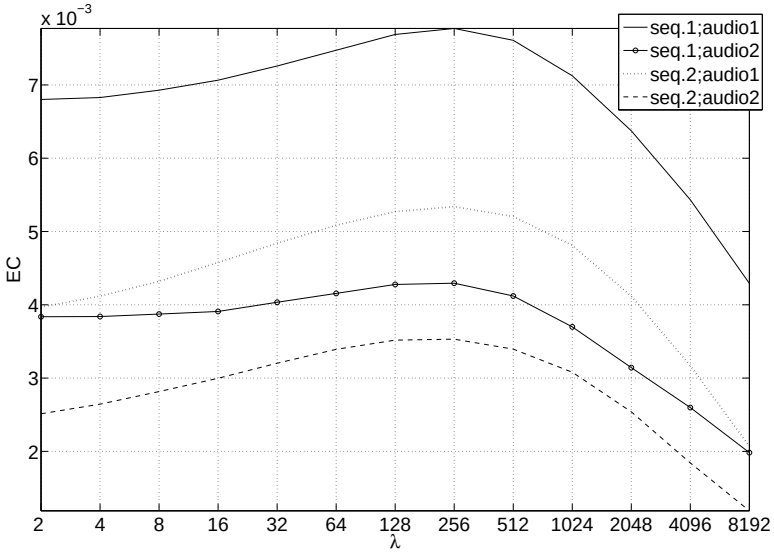


Figure 6.5 — Evolution of the efficiency coefficient as a function of λ, with ι kept fix to 19. The tests have been performed on the couples $(seq_1, audio_1)$, $(seq_1, audio_2)$, $(seq_2, audio_1)$, $(seq_2, audio_2)$. Only for $(seq_1, audio_1)$ are the acoustic and visual speech signals related.

similar when there exist a relationship between the audio and the video $(seq_1, audio_1)$ and when no relationship exist (other cases). The EC is however greater in the first case than in the others, as expected. In Fig. 6.6, the differences of the EC computed with features taken from $(seq_1, audio_1)$ and the EC computed with features extracted from $(seq_1, audio_2)$ $(\Delta EC1)$, $(seq_2, audio_1)$ $(\Delta EC2)$, $(seq_2, audio_2)$ $(\Delta EC3)$ are drawn. There is only one case - namely $\Delta EC1$ - for which the difference between the EC estimated with the related signals $(seq_1, audio_1)$ and the unrelated ones $(seq_1, audio_2)$ increases. Thus an EC-based detector would gain in discrimination power in the case where, given a speaking mouth region, it has to assign it the correct audio signal.

In Table. 6.1, the differences of MI (with $\Delta MI1$ equivalents to $\Delta EC1$, $\Delta MI2$ equivalents to $\Delta EC2$, etc.) are given for the MI evaluated at the points $\lambda = 2$ (first column of Table 6.1) and $\lambda = 256$ (second column). $\lambda = 2$ corresponds to a non-optimized scheme, whereas $\lambda = 256$ is the abscissa of the maximal EC value found in each test case. Therefore, the ΔMI obtained for this last λ corresponds to the classifier scheme: it evaluates the difference of MI between the feature optimized with respect to EC. From these results, it appears that increasing the EC by finding the suitable λ parameter might not increase the discrimination power of the MI-based classifier. However, it must be noticed that the maximal EC values might not be the global optima since the functions are sampled on some pre-defined points only. Further investigations, on a larger test set, are then required prior to draw firm conclusions.

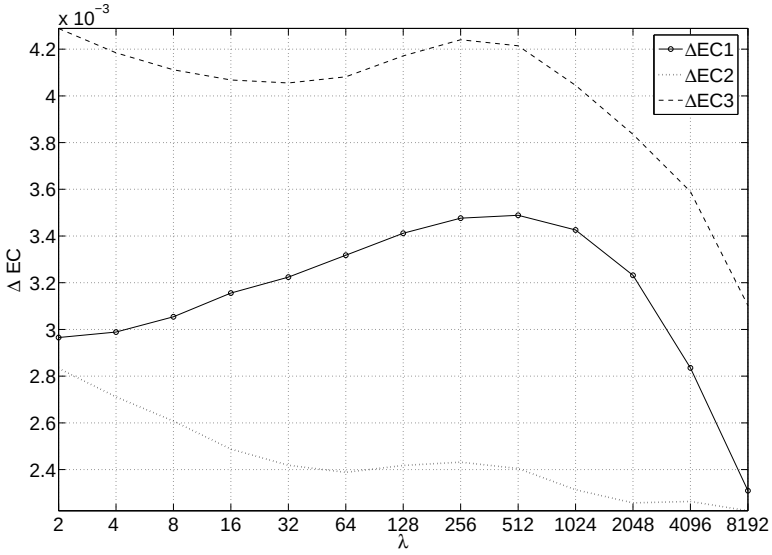


Figure 6.6 — Difference between the EC computed with features taken from $(seq_1, audio_1)$ and the EC computed with features extracted from $(seq_1, audio_2)$ for $\Delta EC1$, $(seq_2, audio_1)$ for $\Delta EC2$, $(seq_2, audio_2)$ for $\Delta EC3$. λ is varied while ι was fixed to 19.

λ	2	256
$\Delta MI1$	0.0202	0.0212
$\Delta MI2$	0.0287	0.0270
$\Delta MI3$	0.0352	0.0341

Table 6.1 — Difference between the MI computed with features taken from $(seq_1, audio_1)$ and the MI computed with features extracted from $(seq_1, audio_2)$ for $\Delta EC1$, $(seq_2, audio_1)$ for $\Delta EC2$, $(seq_2, audio_2)$ for $\Delta EC3$. λ The differences are computed for two values of λ: $\lambda = 2$ (non-optimized scheme) and $\lambda = 256$ (values for which EC is maximum in all the cases).

6.5.3 Influence of the ι parameter value on the EC

In a second experiment, λ is kept fix to 100 (empirical value which gave good results in the experiments of chapter 5) and the number of iterations ι is varied exponentially from 2 to 2^{10}. As in the precedent case, the EC is computed for the four cases corresponding to the couples (seq.1, audio1), (seq.2, audio1), (seq.1, audio2), (seq.2, audio2). The evolution of EC as a function of ι is displayed for each of these cases in Fig. 6.7.

When the number of iterations ι is increased, the efficiency coefficient increases up to a maximum. Further increases of ι let the EC at that maximum value. This result shows that, as the number of iterations increase, the optical flow is diffused from the textured regions of the image to the smooth gradient regions, until the system reaches an equilibrium.

As it was observed in the previous experiments (where λ was varying), the evolution of the EC is similar for the cases involving related audio and video signals $(seq_1, audio_1)$ or those involving unrelated signals, though EC is still higher in the first case.

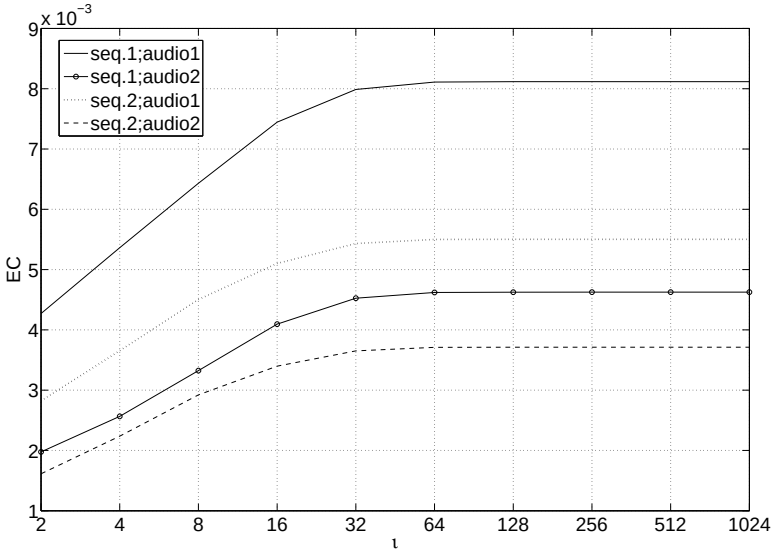


Figure 6.7 — Evolution of the efficiency coefficient as a function of ι, with λ kept fix to 100. The tests have been performed on the couples $(seq_1, audio_1)$, $(seq_1, audio_2)$, $(seq_2, audio_1)$, $(seq_2, audio_2)$. Only for $(seq_1, audio_1)$ are the acoustic and visual speech signals related.

6.5.4 Scale-space interpretation

The variation of the weight λ associated to the smoothness constraint in Eq. (6.10) accounts for a scale-space approach to the problem of OF estimation.

The scale-space theory states that any signal possesses a proper scale at which its content is best handled by a given operator [81], [80]. A scale-space framework provides a hierarchical representation of a signal at a continuum of scales [24] from the finer to the coarser, each scale corresponding to a smoother version of the previous scale representation. Linear scale-spaces are generated usually by solving the diffusion equation:

$$\partial_s E = \nabla^2 E, \quad \text{with } E(t=0) = E_0. \tag{6.25}$$

Eq. (6.25) states that the derivative to scale s equals the Laplacian of the intensity function E, i.e., equals the smoothness constraint ξ_c^2 of Eq. (6.12). Thus, increasing values of the weight λ associated to ξ_c^2 define coarser levels of scale.

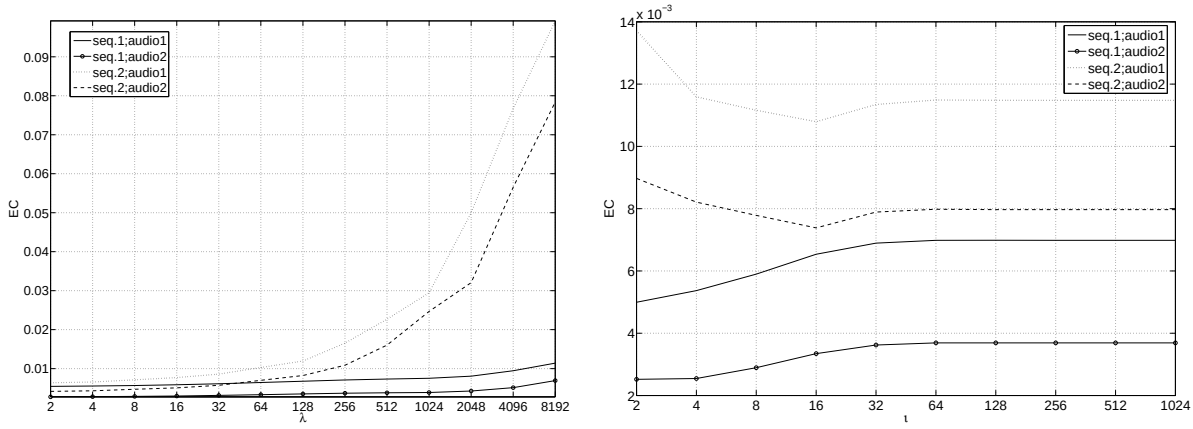
Existing methods have of course taken a multi-resolution approach to OF estimation (see for example [119], [6]). These approaches however introduce a multi-resolution scheme in order to capture large motions. They usually rely on OF estimations done at a coarser scale of the image, used then as initial guesses for OF estimation at a finer image scale.

Our purpose here is quite different. We are simply proposing an automatic way of finding the proper scale of the signal, for which the operator (namely, the EC and then the MI-based classifier) can best analyzed it. In other words, our intention is not to increase the accuracy of the flow but to enhance its audio related content.

Notice that in chapter 5, a similar scale-space concept had been introduced in the optimization scheme. The value of the smoothing parameter h required for estimating

the pdfs was varied along the optimization of the audio features as it was a function of these features. It was shown (see sec. 5.6) that estimating h this way accounted for a multi-resolution approach to the optimization. Let us now examine what happen if this variational estimation for h (called the kernel variance from here onwards to avoid any confusion with λ) is used jointly with the scale-space approach to optical flow. Indeed, in the two experiments just presented (paragraphs 6.5.2 and 6.5.3), the kernel variance was fixed to 0.005 for the audio (h_a) and to 0.004 for each video feature component, h_u and h_v. These settings were found by applying to the signals the formula defined in Eq. (5.23) for robustly estimating the kernel variance, the optical flow being computed with $\iota = 2$ and $\lambda = 2$.

If the kernel variance is estimated using Eq. (5.23) for each different computation of the optical flows, instead of remaining constant, the evolution of the EC as function of λ and ι is different from what was previously observed. Figs. 6.8 display the corresponding EC evolution curves.



(a) (b)

Figure 6.8 — Tests performed on the couples $(seq_1, audio_1)$, $(seq_1, audio_2)$, $(seq_2, audio_1)$, $(seq_2, audio_2)$, with varying values for the kernel variances h_u and h_v: they are estimated using Eq. (5.23) for each new couple (ι, λ). a): Evolution of the efficiency coefficient as a function of λ, with ι kept fix to 19; b): Evolution of the efficiency coefficient as a function of ι, with λ kept fix to 100.

When h_u and h_v vary, the EC evolution is not coherent as it was when these variances were kept fix. Moreover, the EC estimated using $(seq_1, audio_1)$ is not the highest anymore. If the system reaches an equilibrium after a given number of iterations, the EC increases continuously with λ. As this parameter increases, the OF values tend to evolve so that their distributions depart from the uniform one. As a result, the kernel variance, estimated from the data points, decreases with λ. The evolution of the variance h_v associated to the vertical component of the optical flow is plotted versus λ in Fig. 6.9. The evolution of the variance h_u associated to the horizontal component of the flow is not displayed but is very alike. This is similar to what was observed during the audio feature optimization in sec. 5.6, where this phenomenon was accounted for introducing a multi-resolution scheme

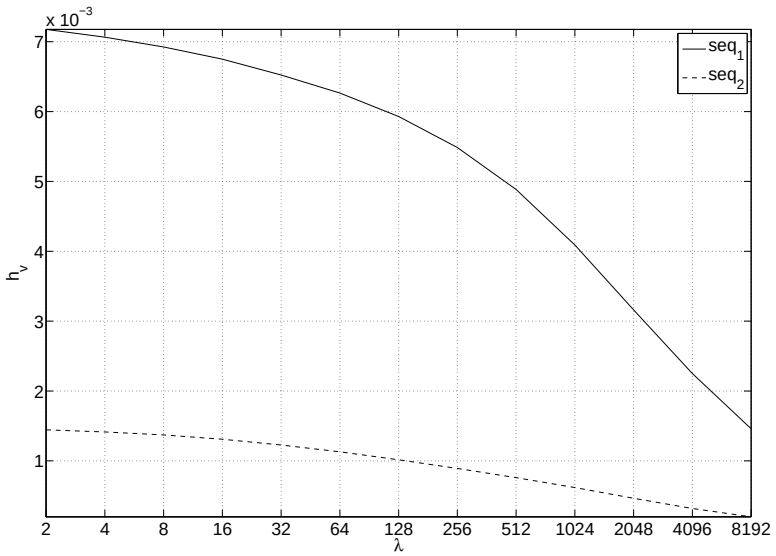


Figure 6.9 — Evolution of the variance associated to the vertical component v of the optical flow when λ is varied exponentially from 2 to 2^{13}.

in the optimization process. Associating the scale-space approach and the multi-resolution introduced by the varying kernel variance is useless, or even worse, counterproductive. At least, the benefits offered by each approach disappear.

6.5.5 Analysis of the 2D optimization problem

The previous experiments have shown that the value of the efficiency coefficient between the optical flow in a speaking mouth region and some energy features extracted from the corresponding speech audio signal is varying with the accuracy of the flow. In these tests, one parameter (λ or ι) was kept fix at an empirically estimated value, while the other one was varying. The question is now to know whether there exist an optimal pair of parameters ι and λ for estimating the OF in a speaker mouth. That is, an optimal couple (ι, λ) so that the EC between the resulting OF and the feature extracted from the co-occurring acoustic speech signal is maximized. In order to analyze this 2D optimization problem, ι and λ are now jointly varied exponentially from 2 to 2^7 and from 2 to 2^9 respectively.

The Figs. 6.10 display the EC evolution with respect to λ, respectively ι, for different values of ι, respectively λ, using the data (seq_1, $audio_1$).

It can be observed that these curves are similar to those obtained in the first experiments (Fig. 6.6 and Fig. 6.5). They show however that no global optimum exist to the optimization problem. Indeed, if a maximum is still reached as λ increases, this value if always higher for growing values of ι. The couples of parameters (λ, ι) define different equilibrium for the system. In other words, for a given number of iterations, the system possesses a proper best scale.

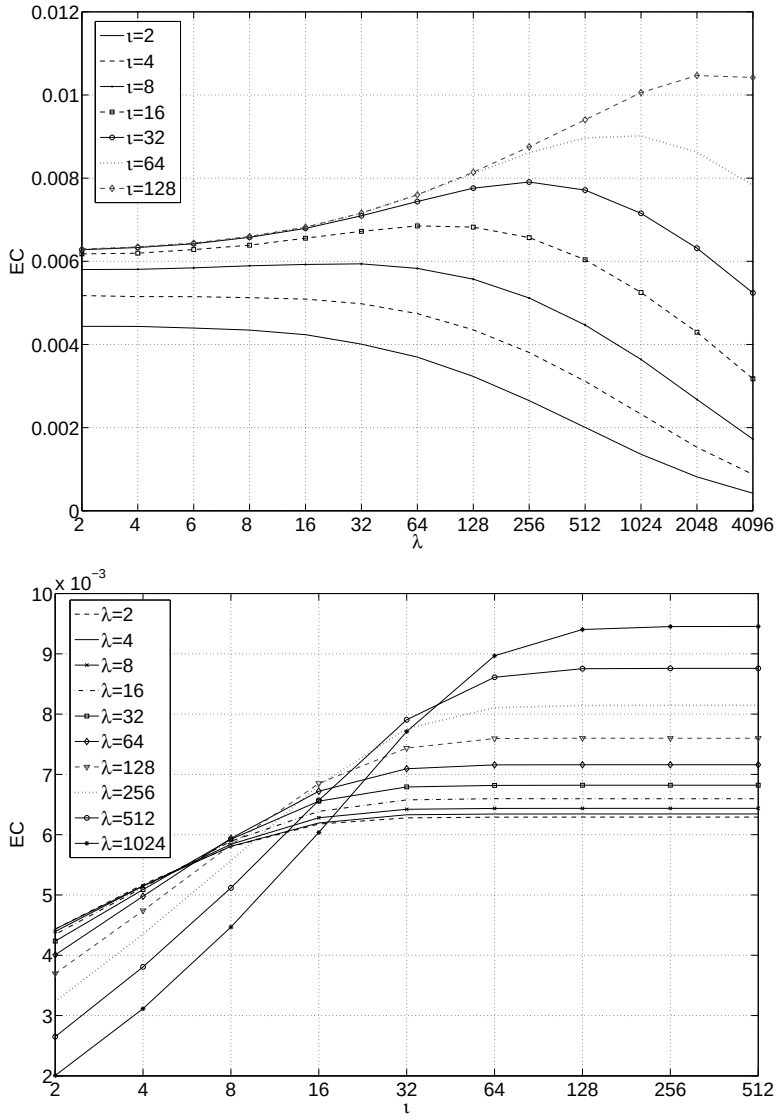


Figure 6.10 — Evolution of the EC with respect to λ (top), respectively ι (bottom), for different values of ι, respectively λ, using (seq_1, $audio_1$).

6.6 Optimization framework

6.6.1 Re-definition of the optimization problem

The first sets of experiments presented in paragraphes 6.5.2 and 6.5.3 have shown that the value of the efficiency coefficient between the optical flow in a speaking mouth region and some energy features extracted from the corresponding speech audio signal is varying with the accuracy of the flow.

The problem appears however to be more complex than expected. Indeed, it does not exist a globally optimal pair of parameters (ι, λ) for estimating the optical flow in a speaking mouth region, such that this OF would maximize the EC with the audio features extracted

from the co-occurring acoustic speech signal.

The problem can be considered as a MO optimization problem, where the goal is to maximize the objective function defined in Eq. (6.18) with an additional constraint on ι which must be kept as small as possible. Indeed, the higher the ι, the slower the OF estimation.

Classical approaches to solve MO problems convert it in a SO problem, usually by defining a new objective function in the form of a weighted sum of the different objectives. The global optimum (which may not exist since the new search space is not guaranteed to be convex) can then be found with standard methods designed for SO problems. The big issue in such approach however is about the weight to assign to each objective.

A better approach is offered by the so-called *Pareto optimality concept*: a solution belongs to the Pareto set if there is no other solution that can improve at least one of the objectives without degradation of any other objective [94]. This approach gives rise to a trade-off analysis of the problem. Since the individual objective functions are typically conflicting, there exist no single optimal solution as in SO problems. Rather, there exist a set of Pareto optimal solutions (the Pareto optimal set) from which the preferred solution is selected.

The definition of a framework for solving the MO problem is let for future work. In this thesis, we will address the simplified SO problem of determining an optimal value for λ, ι being fixed a priori. The objective is to investigate the potential offered by the approach for a speaker detection task.

The physical meaning of the number of iterations can be more easily apprehended, thus ι can be more easily set up, than the choice of the proper scale, thus of λ. As mentioned in sec. 5.2, the number of iterations fixes the diffusion of the flow from the textured regions of the images inwards the non-textured regions. Information extracted from the image can help in setting this parameter. In the case of motion related to mouth regions for example, we have some clues about the structures present, based on the characteristics of this mouth region. Moreover, it has been seen in the feasibility study that choosing a too high number of iterations could not reduce the information, contrary to a wrong choice of the regularization parameter.

As said previously, the regularization parameter λ determines how strongly the solution to Eq. (6.13) must matches the brightness constancy assumption. It is an important parameter, uneasy to set up correctly, especially for speech mouth motion estimation, were the brightness constancy assumption is often violated. It should correspond to the proper scale of the system.

Assuming then the number of iterations ι to be fixed at a given value, the problem is to find the regularization parameter λ_{opt} such that the OF field estimated in the mouth region maximizes the EC with the co-occurring acoustic speech signal. This optimization

problem is stated as:

$$\begin{aligned} &\text{Find } \lambda_{opt} \text{ such that} \\ &\quad f(\lambda) = \arg\max_{\lambda} e(A, \vec{F}_V(\lambda)), \quad \text{subject to } \lambda \in \mathbb{R}_+. \end{aligned} \tag{6.26}$$

In this case, the exploratory tests presented in sec. 6.5 have proven that the efficiency coefficient function, e, should possesses a single maximum. A method able to efficiently find this optimum has now to be chosen.

6.6.2 DIRECT algorithm for solving the optimization problem

The optimization problem is not as challenging as the one faced in chapter 5. The problem is 1D and should not exhibit local minima*. However, an analytical formulation of the gradient of the cost function is still difficult to obtain due to the unknown form of the pdfs (in particular, $p(F_A, \vec{F}_V)$ is 3D), and we would still like to avoid any restrictive assumptions. Thus the method should be gradient-free. Moreover, to limit the computational time allocated to the optimization, the chosen method should evaluate the function as few times as possible. The computational effort is indeed dominated by the cost of evaluating f, which is directly dependent on the number of iterations, a static parameter.

There is no need to choose a meta-heuristic optimization approach such as the EA strategy used in chapter 5. A simple deterministic gradient-free technique like the Powell's algorithm described in sec. 5.6 should perform well. However, some methods take directly advantage of the constraints imposed on the parameters to be optimized. In particular, in the problem at hand, the parameter λ is lower-bounded. Reasonable assumptions allow us to define an upper bound as well. Then space partitioning approaches like the Branch & Bounds (B&B) methods could be used with advantage to solve the problem. These are a variety of adaptive partition strategies which search a global optimum by exploring only a part of the search space. The derived bounds on the objective function guarantee that no optimal solutions exist on the pruned part of the search space [39]. The complexity of branching may exponentially increases with dimension. But the cost function being 1D, this is not a problem.

The so-called DIRECT algorithm (acronym for DIviding RECtangle) developed by Jones et al. in [68] is an approach for finding the global minimum in a bracketed region. It is based on the well-known Shubert's method [117] for Lipschitzian objective functions, which uses basically a lower envelope of the function to estimate the global minimum. However, DIRECT alleviates some restrictions associated to standard Lipschitzian methods by striking a balance between the global and the local search. Both the Shubert's and the Jones' algorithms are described in details in Appendix B.

*In this paragraph, and in this paragraph only, the maximization problem is turned into a minimization problem to fit the standard formulation of optimization tasks. To this end, the sign of the cost function defined in Eq. (6.26) is changed.

Using Jones' DIRECT optimization algorithm, the optimum of the objective function defined in Eq. (6.26) is expected to be reached efficiently. The resulting optimal regularization parameter allows us to estimate an optimal OF for each mouth region presents in the audio-visual sequence. The MI between these different video features and the audio feature extracted from the acoustic signal (see sec. 6.2) is then evaluated. The largest value indicates the speaking mouth (classifier defined in § 4.4.1). In the next section, this pattern recognition chain is tested on a set of audiovisual sequences to assess its performance.

6.7 Audiovisual speaker detection results

6.7.1 Experimental framework

The proposed PR framework, including the video feature optimization step, is now tested on the 11 CUAVE sequences $g11$ to $g22$* [98] previously used in the experiments carried out in chapter 5†. In a few words, these are two-speaker clips, shot in the NTSC standard (29.97fps, 44.1kHz stereo sound), where each speakers is face to the camera and utters in turn two series of digits.

The mouth regions are tracked along the sequence using the detector described in sec. 5.3. Only the regularization parameter λ is optimized: the number of iterations ι required for estimating the OF must be fixed manually by the user. As stated many times in the previous sections, the specificities of the problem indicate that ι should be set at half the mouth region high. Obviously, the size of the extracted mouth regions differs from one speaker to the other, depending on each individual characteristics, or on its distance from the camera. As a result, the number of iterations used to estimate the OF are also different. This can be a problem since the feasibility study of sec. 6.5 established that the EC optimum (then the MI one as well) corresponding to λ_{opt} increases with ι. For a fair comparison of the MI computed in each mouth region, the same number of iterations should be used to estimate the OF. On each sequence, the mouth region m_2 of the right-hand side speaker (*speaker2*) is resized for its size to match the size of the mouth region m_1 (i.e. the *speaker1*, or left-hand side speaker).

The optimization is done over a $2s$ temporal window, shifted in one second steps over the whole sequence to make decisions once per second. The Finkel's implementation of the DIRECT algorithm [43] is used. The ϵ parameter to be set up in the DIRECT method (see Appendix B) is set to 10^{-4} as advocated in [68]. The stopping criterion is given by a maximum number of iterations, fixed to 15. The smoothing parameter values, or kernel variances, required by the Parzen window estimation of the pdfs, are fixed to 0.005 for the audio features as well as for the vertical and horizontal components of the OF (let us recall that the features are defined on the interval $[-1, 1]$). This is an average value based

*The sequence $g18$ is discarded, as was done in chapter 5, because of the strong level of compression noise it presents.

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

on the results obtained when applying Eq. (5.23) to the data, with the optical flows being estimated with $\iota = 2$ and $\lambda = 2$. This is a similar approach to the one taken in sec. 6.5. The bounds on the λ parameter are defined as $[2,\ 3 \times 10^4]$. If the optimal value lies near to a bound limit, the optimization can be launched again using a highest or smaller value for the corresponding bound. This did not happen however.

For a given analysis window, the video feature set (video sample) is composed of the $N \times M \times \tau$ horizontal and vertical components of the OF at each pixel location of the mouth region (the latter being of size $N \times M$ pixels, where N and M vary between 22 and 57 pixels). The number of video frames T within an analysis window is equals to 60, thus $\tau = T - 1 = 59$.

From the audio signal, $P = 13$ mel-cepstrum coefficients are computed using 30ms Hamming windows. The first coefficient, which pertains to the energy, is removed. Thus, for a given analysis window, the audio feature set (audio sample) is composed of the τ^* mean values of the 12 MFCCs.

The same evaluation framework is used than in sec. 5.8: the evaluation is performed on the last second of each detection window (silent windows are not considered). More details are provided in Appendix A.

In a first step, the ability of the classifier to detect the speaker is evaluated in § 6.7.2. In this section, two other kinds of audio features, A_2 and A_3, are introduced in addition to A_1 (the mean value of the 12 MFCCs). A_2 denotes the value of the first mel-cepstrum coefficient, and A_3 the mean value of the 13 MFCCs. They are used in turn for both the optimization of the video features and the classification step itself.

In § 6.7.3, a comparative study is performed. To this end, different kinds of visual features are put in turn in the classifier: F_V^o, $F_V^{\lambda 100}$ and F_V^{no}. F_V^o denotes the OF obtained by applying the optimization framework proposed in this chapter, using jointly A_1, A_2 or A_3. $F_V^{\lambda 100}$ stands for the OF estimated with a regularization parameter pre-fixed to 100. Last but not least, the video features used for optimizing the audio features in chapter 5 are tested as well. These features, denoted by F_V^{no} are defined as the magnitude of the OF (estimated with $\lambda = 100$) signed as the vertical component. Table 6.2 summarizes the audio and video features used in the experiments

6.7.2 Results for speaker detection

Optimized video features F_V^o are obtained for each analysis window applying the optimization framework proposed in sec. 6.6. They are then put as input of the MI-based classifier (defined as in § 4.4.1) which outputs 77% of correct detections. These were already satisfactory results (let us recall that the motion-only scheme tested in sec. 5.8 gave rise to 63% of correct detections only). However, another set of experiments has been performed using different audio features: we wanted to check whether discarding the first mel-cepstrum coefficient was judicious or not. As already mentioned, this coefficient pertains to the en-

*The mel-cepstrograms are downsampled to the OF frame rate

Audio features	
A_1	Mean values of the 12 MFCCs
A_2	First MFCC
A_3	Mean value of the 13 MFCCs

Video features	
F_V^o	Optimized video features
$F_V^{\lambda 100}$	Non-optimized video features: the λ parameter is fixed a priori to 100.
F_V^{no}	Non-optimized video features: signed norm of the OF estimated with λ set to 100.

Table 6.2 — Name and description of the different audio and video features used in the experiments.

ergy of the signal. It can be thought that peaks of energy in the acoustic speech signal should co-occur with visual events such as opening or closing of the lips. Indeed, in the linear source filter model of Fant [40] presented in sec. 5.2, the lips are the last time-varying filters acting on the acoustic speech signal. When uttering the plosive /p/ for example, a movement of the lips clearly co-occur with a peak in the energy of the acoustic signal.

The audio features A_2 and A_3 previously described have then been introduced both in the optimization of the video features and in the detection step.

Table 6.3 sums up the results output by the MI-based classifier in each of these three cases. The worst results are obtained for the sequences $g13$ and $g22$, whatever the audio

	Correct detection rate (in %)		
Sequence number	F_A^1	F_A^2	F_A^3
g11	69	81	75
g12	82	88	71
g13	63	50	50
g14	89	95	95
g15	83	89	72
g16	68	84	99
g17	100	83	83
g19	86	86	71
g20	96	100	85
g21	80	90	100
g22	30	80	45
Mean	77.0	84.2	76.9

Table 6.3 — Speaker detection results on the CUAVE sequences with evaluation on the last second of each detection window (silent windows are not considered). A_1, A_2, and A_3 are used in turn to obtain the optimized video features F_V^o, and to perform the detection.

features used (even if the results obtained on $g22$ with A_2 are acceptable). The $g13$ clip presents challenging features. Speakers are moving, especially speaker2 who is getting back and forht with respect to the camera and tilting the head forward. This induces intensity changes in the mouth region, even when the speaker is silent. As a matter of fact, the mistakes occur when speaker1 is speaking. A frame taken from this sequence is displayed in Fig. 6.11. On seq. $g22$, the detection failures mainly happen with speaker2. A reason

Figure 6.11 — Frame extracted from the sequence $g13$.

might be that speaker1 is moving and uttering the digits without sounding them while speaker1 is speaking. The use of A_2 removes this black point.

Generally speaking, the use of A_2 leads to better achievements than the use of either A_1 or A_3. It should be noticed however that the gap reduced if $g22$ is accounted an outlier and removed from the test set. In this case, the scores are of 81.6%, 84.6% and 80.1% for A_1, A_2 and A_3 respectively. Due to the small size of the test set, these difference are less significant.

These experiments emphasize a point central to this thesis: the importance of choosing the right features. Choosing different audio features changes the detector performance. Actually, the fact that the optimization framework developed in sec. 4.2 is applied only to one modality, the other one being defined empirically, plugs of course the performance of the pattern recognizer. Future work should address the problem of the joint optimization of the audio and the video features using the feature extraction framework.

These results must draw our attention to a second point: do the best results obtained with A_2 demonstrate that the first MFCC is more related to visual speech than the two other audio features? The results obtained by using A_3, the mean value of the 13MFCCs - thus including the first one - put a doubt on that hypothesis. They lead indeed to the worse score among the three. The values of the three audio features for an analysis window of seq. $g20$ are plotted in Fig. 6.12. It can be observed that taking the mean value of the 12 or 13 MFCCs results in audio features with more variations of smaller amplitude difference compared to the audio feature made of the first coefficient solely. The latter exhibits sparse but salient amplitude variations. Thereafter, it appears that considering with an equal importance all of the MFCCs does not improve the information content but hides the significant events. Incidently, it would be interesting to carry out a study using

in turn each MFCC solely as an audio feature to evaluate how each of them is related to the video content.

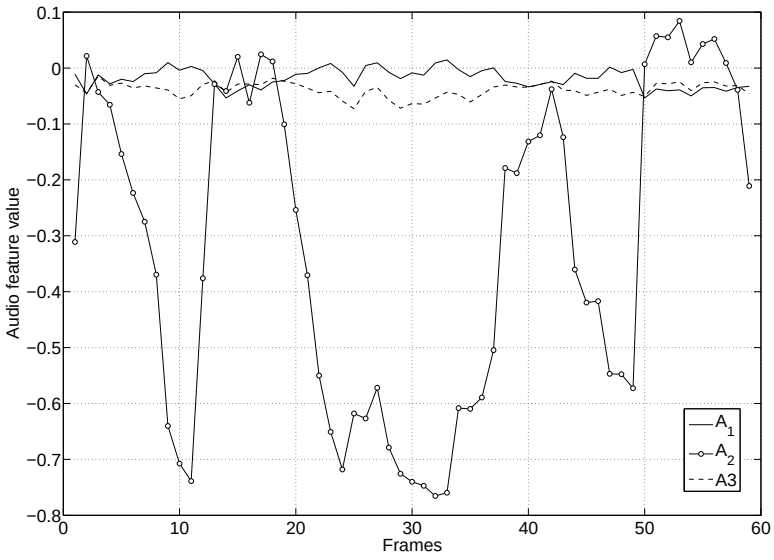


Figure 6.12 — Values of the audio features A_1, A_2 and A_3 for an analysis window of seq. $g20$.

An analysis of the values found for the regularization parameter at the end of the optimizations is carried out in the next paragraph. It demonstrates that it is more difficult for the algorithm to find a relationship between the audio and the video when the mean mel-cepstrogram value is used instead of a single MFCC.

6.7.3 Analysis of the optimized regularization parameters

Let λ_{A1} and λ_{A2} be the regularization parameters obtained by using A_1, respectively A_2, during the optimization process. The histogram of both λ_{A1} and λ_{A2} are shown in Figs. 6.13. Both the λ_{A1} and the λ_{A2} values mainly concentrates between 0 and 100. However, the tail of the distribution is noticeably longer for λ_{A1}. This effect is also present on the distribution of the λ values found using A_3 (histogram not plotted here). This indicates that finding a relationship between the audio and the video features is harder for the optimization algorithm when the mean value of the 12 MFCCs is used instead of the first coefficient solely. A smoother OF is required for a relationship to be found with the averaged audio features.

The values found for mouth1 (λ_1) and those found for mouth2 (λ_2) on seqs. $g20$ and $g17$ - using A_2 in both cases - are adjusted to $[-1, 1]$ and their difference is computed as $\Delta\lambda = \lambda_1 - \lambda_2$. The evolution of the difference of these normalized λ values is plotted in Figs. 6.14 and 6.15. This evolution is plotted jointly with the ground truth and the detector output for the corresponding sequence.

For seq. $g20$, where the detector hit is of 100%, the regularization parameters increase when the mouth they stand for is speaking. This seems coherent with the theory: when a

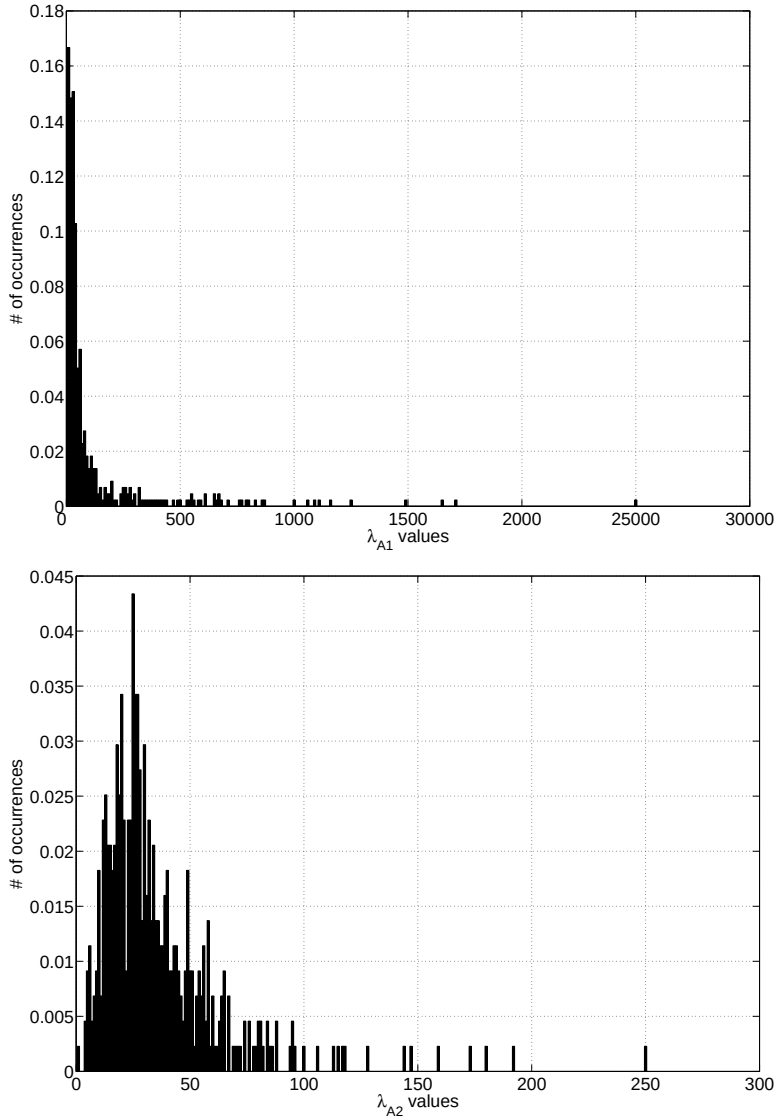


Figure 6.13 — Top: Histogram of the optimized λ values when A_1 has been used as audio feature; Bottom: Histogram of the optimized λ values when A_2 has been used as audio feature.

mouth is speaking, the irregularities in the flow are more important. As said in sec. 6.4, the mouth motions are complex and subject to occlusions, thus violating the brightness constancy assumption: a larger weight has to be put on the smoothness term. Unlikely, this behavior is not true anymore in other sequences, like in the seq. $g17$ for example (for which the detector rates is of 83%) but of course, violations of the brightness constancy assumption can also occur on the non-speaking mouth.

6.7.4 Limits and performance

In order to assess the benefit of using the proposed optimization approach to OF estimation, a comparative study is carried out, using in turn optimized or fixed λ values for estimating

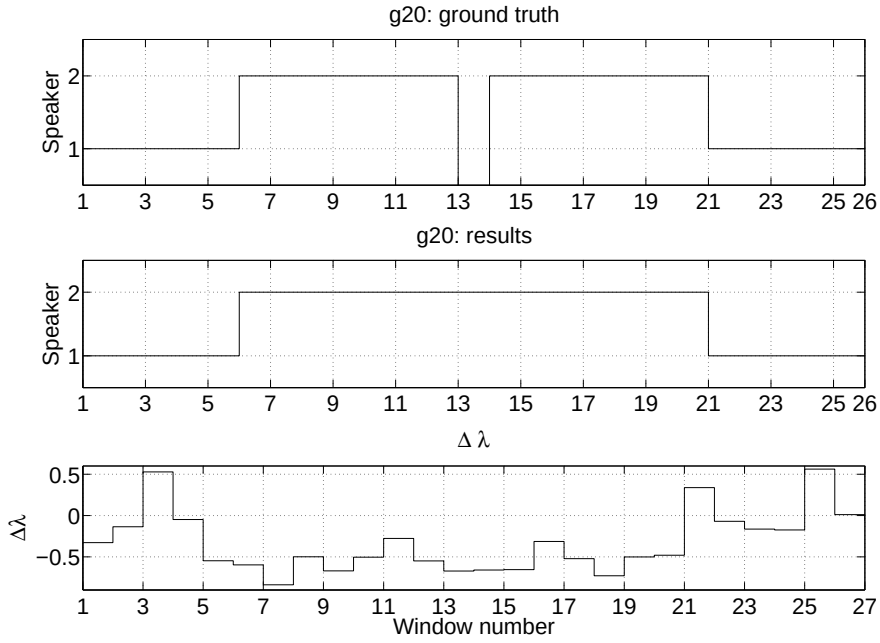


Figure 6.14 — Top and middle: Ground truth and detector output for sequence $g20$. Bottom: Difference of normalized λ values: $\Delta\lambda = \lambda_1 - \lambda_2$.

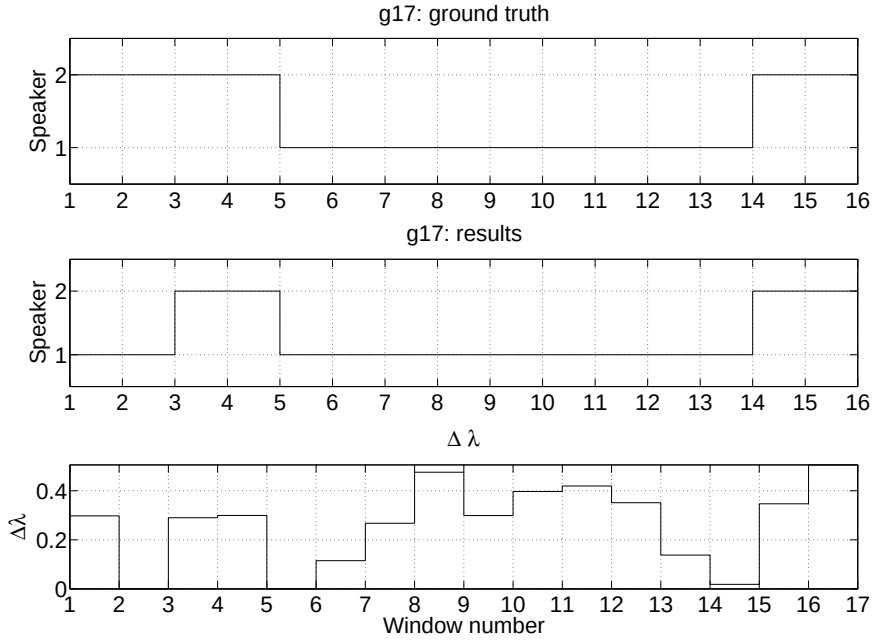


Figure 6.15 — Top and middle: Ground truth and detector output for sequence $g17$. Bottom: Difference of normalized λ values: $\Delta\lambda = \lambda_1 - \lambda_2$.

the flow. Three kinds of video features F_V^o, $F_V^{\lambda 100}$ and F_V^{no}, introduced in the experimental framework part, are then used jointly with the audio feature A_2 as input for the MI-based classifier. The results obtained in each cases are listed in Table. 6.4. We were of course

	Correct detection rate (in %)		
Sequence number	F_V^o	$F_V^{\lambda 100}$	F_V^{no}
g11	81	88	63
g12	88	88	82
g13	50	46	44
g14	95	95	95
g15	89	89	89
g16	84	84	100
g17	83	89	89
g19	86	86	86
g20	100	96	100
g21	90	90	85
g22	80	80	40
Mean	84.2	84.6	79.4

Table 6.4 — Speaker detection results obtained on the CUAVE sequences with evaluation on the last second of each detection window (silent windows are not considered). A^2 has been used as audio features with either the optimized video feature F_V^o, or the non-optimized ones, $F_V^{\lambda 100}$, F_V^{no}.

expecting a comparison in favor of the optimization scheme. It is the case when the features defined in chapter 5 are used. But it is unlikely not true when the 2D OF features $F_V^{\lambda 100}$ are used.

It should be noticed noted however that the difference is not really significative especially due to the small test size.

ANOVA (acronym for ANalysis Of VAriance) is an hypothesis test where the detection results found in each of the three cases are considered as samples drawn from populations with the same mean (null hypothesis). The p-value determines the confidence level associated to the null hypothesis. Both the inter-group and intra-group variability is analyzed through this method. It is then applied to the results shown in Table 6.4 to study the significance of the means. The Matlab implementation of the ANOVA test is used [62].

The results are plotted in Fig. 6.16: the boxes indicate the lower and upper quartile values. The central line stands for the median of the sample. The whiskers show the extend of the data. The circles denote values considered as outliers (the results obtained for seq. $g13$ with F_V^o and $F_V^{\lambda 100}$). The p-value of 0.7 means that the null hypothesis cannot be rejected. The means of the different samples are not significantly different. Or in other words, the inter-group differences are not meaningful.

It is interesting however to observe the intra-group variability. The analysis performed on the results shown in Table 6.4 are confirmed. The results obtained using F_V^{no} are the less relevant. Notice however that the results found for the seq. $g13$ must be accounted outliers when using F_V^o and $F_V^{\lambda 100}$. The latter case present the smaller intra-group variance.

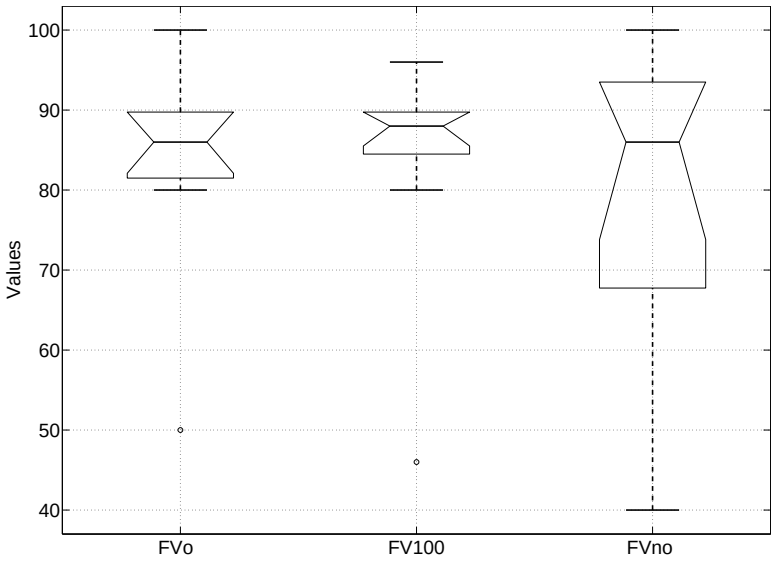


Figure 6.16 — Results of the ANOVA test performed on the results presented in Table 6.4. The boxes indicate the lower and upper quartile values. The central line stands for the median of the sample. The whiskers show the extend of the data. The circles denote values considered as outliers.

These observations have motivated another series of tests that has been performed using non-optimized video features. These are obtained by fixing the λ parameter to different a priori fixed values: 50, 100, 300 or 500. The results are summarized in Table 6.5. The

	Correct detection rate (in %)			
Sequence number	$F_V^{\lambda 50}$	$F_V^{\lambda 100}$	$F_V^{\lambda 300}$	$F_V^{\lambda 500}$
g11	75	88	81	81
g12	88	88	88	88
g13	46	46	46	46
g14	95	95	95	95
g15	89	89	89	83
g16	84	84	84	84
g17	100	89	89	83
g19	86	86	86	86
g20	96	96	100	100
g21	95	90	90	90
g22	80	90	80	80
Mean	84.9	84.6	84.4	84.3

Table 6.5 — Speaker detection results obtained on the CUAVE sequences with evaluation on the last second of each detection window (silent windows are not considered). A^2 has been used as audio features with non-optimized video features obtained by setting a priori the λ parameter to 50 $F_V^{\lambda 50}$, 100 for $F_V^{\lambda 100}$, 300 for $F_V^{\lambda 300}$ and 500 for $F_V^{\lambda 500}$.

mean percentage of correct detections tends to decrease as λ increases. But if the results are analyzed sequence by sequence, this is not always true (look at the results obtained for the sequence $g11$ for example). An ANOVA test has been performed on the data presented in Table 6.5. Each column of the table (i.e. the detection results for the set of 11 sequences and for a given value of λ) stands for a sample. The results are plotted in Fig. 6.17. The results obtained with the optimized video features (column 1 of Table 6.4) are also considered in the ANOVA test.

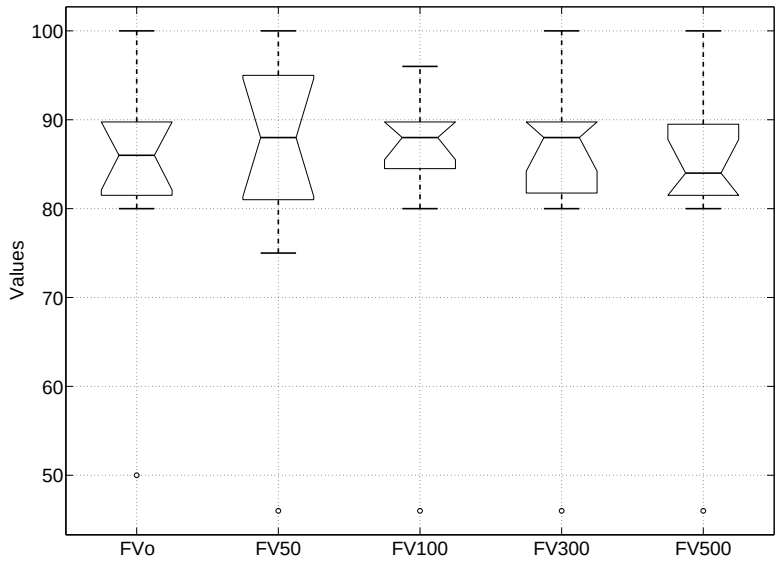


Figure 6.17 — Results of the ANOVA test performed on the detection results obtained when using in turn F_V^o, $F_V^{\lambda 50}$, $F_V^{\lambda 100}$, $F_V^{\lambda 300}$, $F_V^{\lambda 500}$ on the 11 sequences. The boxes indicate the lower and upper quartile values. The central line stands for the median of the sample. The whiskers show the extend of the data. The circles denote values considered as outliers.

Of course, the inter-group differences are not significant (p-value of 0.99). It is more interesting to analyze the intra-group variance. The smaller intra-group variability is observed for the results obtained using $F_V^{\lambda 100}$. Whereas the results obtained with λ fixed to 50, which exhibited the best average detection rate, present the highest variability. The analysis of the results per sequences for the different video features $F_V^{\lambda 50}$ to $F_V^{\lambda 100}$ also demonstrates the variability in the detection rates for certain sequences. For the sequence $g11$ for example, the best result is obtained using $F_V^{\lambda 100}$, whereas the score is 13% smaller with $F_V^{\lambda 50}$. These same features however achieve the best score for seq. $g17$. As far as the features $F_V^{\lambda 300}$ and $F_V^{\lambda 500}$ are concerned, theirs scores are usually worse than those reached by $F_V^{\lambda 50}$ and $F_V^{\lambda 100}$ but for the seq. $g20$ where they hit 100% of correct detection.

As a consequence, the empirical setting of an appropriate λ parameter turns out to be a tricky task. Depending on the user's choice, poor results can be obtained. In any case, the user has no guarantee that its choice is suitable. Actually, the setting we choose initially for comparing the performance of the λ optimization scheme versus the non-optimization approach seems to be one of the best among the four λ values tested (incidently, it was the chosen setting for the optimization of the audio features in chapter 5). But with respect

to the different λ-fix video features, the optimized video features stand the comparison. One can argue of course that the scores obtained with the optimized λ values are not the best for all the sequences. They are neither the best, nor the worse. If the optimization scheme is used, a certain confidence can be associated to the speaker detection results. The optimization approach avoids the user to waste time in different trials and to have to assume a certain unreliability on the performance of the system.

Also, it must be noticed that the test set is made of simple cases shot in very clean conditions. It is well known that the scale-space approach present an added-value when data are noisy. Further tests should be performed in poorer conditions in order to assess the gain offered by the optimization scheme.

6.8 Pattern recognition chain performance

The performance of the whole classification chain, and of the gain possibly offered by the feature extraction step, is now appraised using the hypothesis framework proposed in sec. 4.4. The test sequences are those taken from the CUAVE database and already used in sec. 6.7. Obviously, the video and audio features are also defined as in sec. 6.7 (see Table 6.2). The optimized video features are put as input of the classifier, defined as the test function giving the best test of size α (α being the probability of false-alarm, as defined in sec. 4.4).

The hypothesis test used as a classifier is the one defined by Eq. (4.31). Two potential speakers are present on each sequence, thus two tests can be defined as described in § 4.4.4, each test involving one speaker only. As was done in chapter 5 to evaluate the performance of the PR when audio features were optimized, the experimental conditions are defined so as to eliminate the possibilities 3 and 4 where the two speakers are either both speaking or both silent (the silent windows have been removed from the test set).

For binary tests, a positive and a negative class have to be defined. We assume the positive class (hypothesis H_1) to be the class "speaker", confounding speaker 1 and 2 in a single label. Thus the negative class (null hypothesis) is the "non-speaker" class. The test set counts then 192 test points.

In a first time, the Neyman-Pearson classifier receives at input the audio features A_1, A_2 and A_3 and the video features F_V^o optimized using each of these audio features. The threshold is varied and the ensuing evolution of the probabilities of false alarm P_{FA}, or α, is plotted against the detection probability P_D, or β, in the ROC curves of Fig. 6.18. The ROC curves are far above the line $\alpha = \beta$, thus give better results than a random scheme. For small probabilities of false-alarms, better performances are obtained using A_3. For detection probabilities above 50%, using A_2 becomes the best choice. The ROC curve corresponding to the audio features A_1 never cross noticeably the two others. It is the worse case. Its performance are notably worse in the liberal part of the ROC space.

The three pattern recognizers discriminate better the "non-speaker" class than the

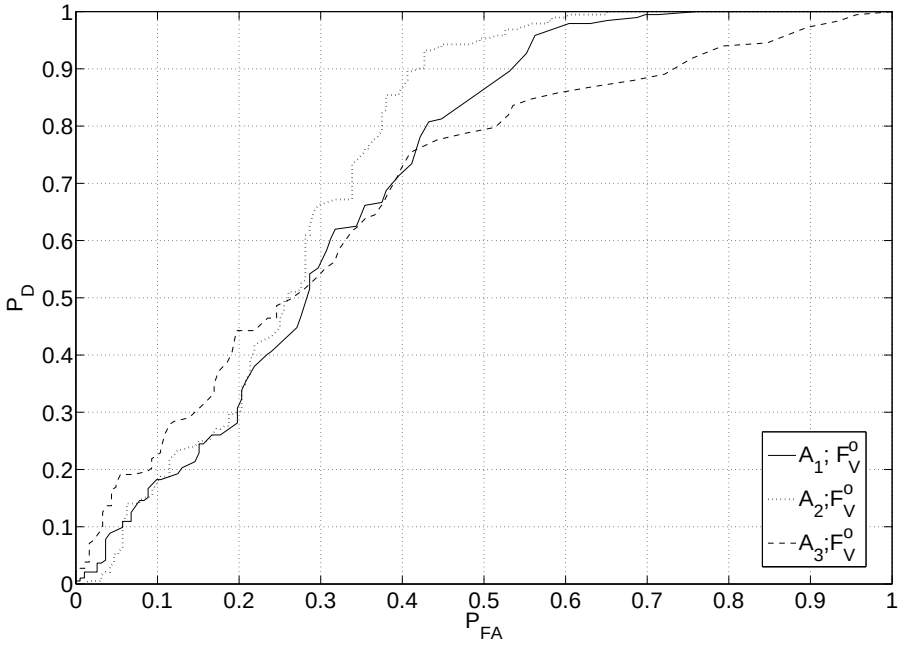


Figure 6.18 — ROC curves for class1 = "speaker". Pattern recognition performance when the three kinds of audio features A_1, A_2 and A_3 are given as input of the classifier, jointly with the video features (optimized with respect to these same audio features).

"speaker" one. For an easiest visualization of this statement, the classes 1 and 2 are switched (the "non-speaker" becomes the positive class) and the ROC curves are plotted in Fig. 6.19. The slopes of the curves in the conservative part are of about 70 deg against 60 deg for "speaker" as the positive class. Unlikely, these slopes decrease above a detection probability of about 60%.

Table 6.6, gives the AUC (a measure of the discrimination capability of the classifier). The highest value, thus the best, corresponds to the use of the A_2 audio features (the first MFCC) in the pattern recognition process.

Audio feature	A_1	A_2	A_3
AUC	0.71	0.75	0.69

Table 6.6 — Area under the curve for each of the ROC curves plotted in Fig. 6.18 (or in Fig. 6.19 equivalently).

In a second experiment, the classifier receives at input the audio features A_2 and the video features F_V^o optimized with these audio features in turn with the non-optimized video features $F_V^{\lambda 100}$ and F_V^{no}. The resulting ROC curves are displayed in Fig. 6.20 (the positive class is the "speaker" class). The three curves are very similar, especially if the small size of the test set is taken into consideration (curves are not smooth). Thus the small differences can not be considered as significant. This is confirmed by the AUC values, equal for each of the three schemes.

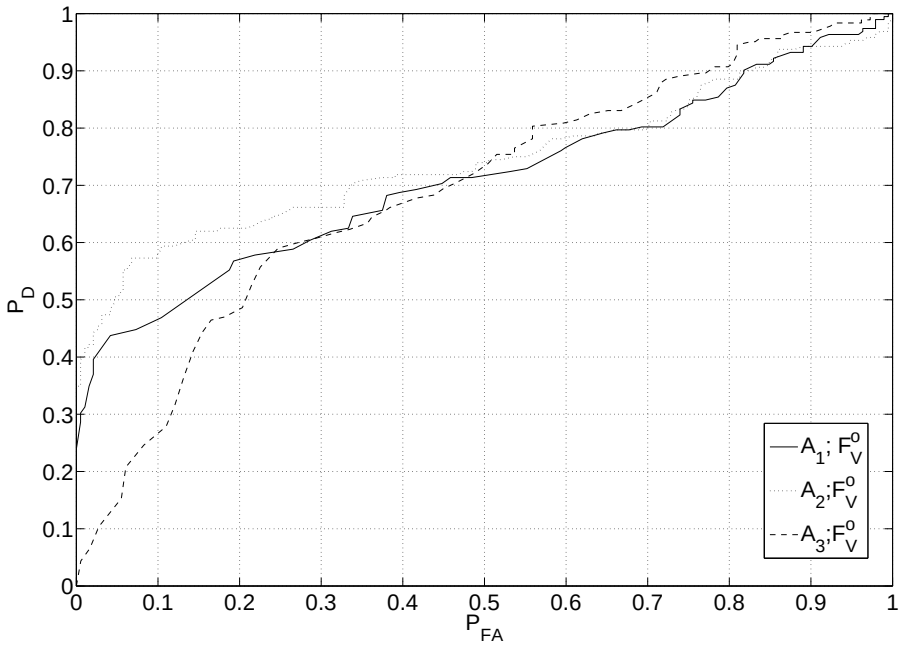


Figure 6.19 — ROC curves for class2 = "non-speaker". Pattern recognition performance when the three kinds of audio features A_1, A_2 and A_3 are given as input of the classifier, jointly with the video features (optimized with respect to these same audio features).

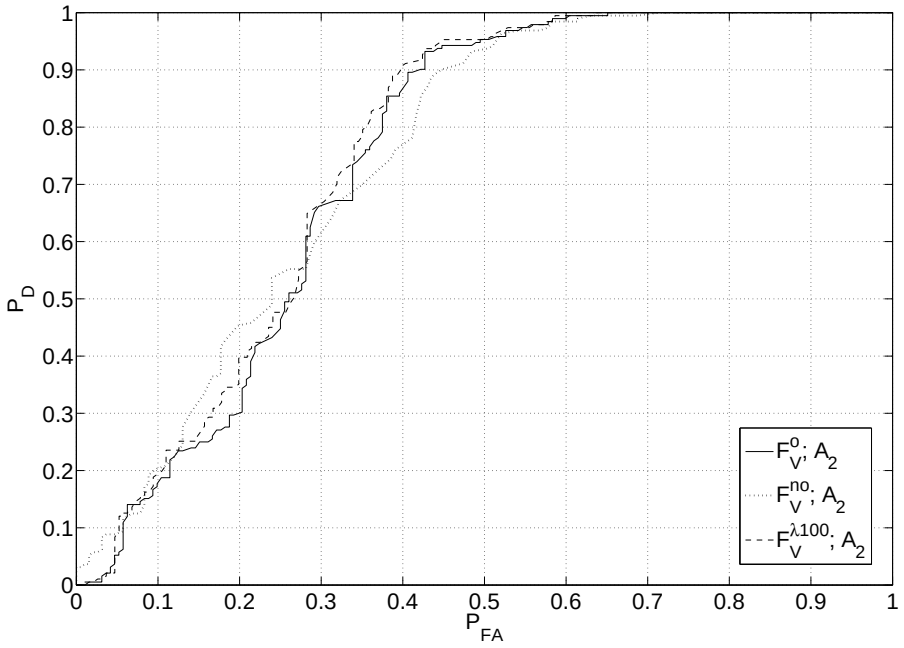


Figure 6.20 — ROC curves for "speaker" class as the positive class. Comparison of performance of the pattern recognition process when either the optimized video features F_V^o or the non-optimized video features F_V^{no} or $F_V^{\lambda 100}$ are put as input of the classifier, with the audio features A_2.

The introduction of the video feature optimization scheme does not improve (nor de-

Video feature	F_V^o	F_V^{no}	$F_V^{\lambda 100}$
AUC	0.75	0.75	0.75

Table 6.7 — Area under the curve for each of the ROC curves plotted in Fig. 6.20.

grade) the discrimination power of the pattern recognition design. During the feasibility study carried out in sec. 6.5, it had already been pointed out that this might be the case. However, a larger number of tests should be performed to draw a robust conclusion on that point. As mentioned in sec. 6.7, noisy sequences should be put in the test set as well. Also, it must not be forgotten that the optimization deals only with the regularization parameter λ, not with the iteration number. The entire MO optimization problem defined in Eq. (6.18) should be solved to assess the benefit of optimizing the OF estimation for speaker detection. For now, only a part of the OF estimation has been optimized.

6.9 Discussion

The pattern recognition system defined in chapter 4 has been applied to the problem of identifying the current speaker among several candidates, focusing specifically on the enhancement of the video features with respect to the content of the co-occurring audio signal.

Similarly to what was done in chapter 5, the data were acquired by a single camera and microphone. The motion has been retained as the visual information specific to speech production process. It is well-known that motion is a 3D phenomenon that is hardly estimated correctly by image processing algorithms. A well-known estimation of the apparent motion in image sequences is known as the optical flow, estimated from the image intensity gradient. A probabilistic model of the relationship between the audio, the video and the image intensity gradient has been proposed trough graph theory. A link has been made with the information theoretic estimator defined in sec. 4.2 (the efficiency coefficient, EC), demonstrating that an increase of the mutual information between the audio and the optical flow (the entropy of the OF being kept constant) should improve the performance of the MI-based classifier.

A corresponding optimization framework has been introduced. The two parameters required for gradient-based OF estimation are optimized with respect to a cost function assessing the improvement they lead to. This improvement is measured by evaluating the MI between the resulting OF and some energy-based features extracted from the co-occurring acoustic signal. The two parameters are the number of iterations ι required by the system to come to a solution, i.e. to reach its equilibrium, and the weight λ to assign to the Laplacian regularization parameter. This parameter allows us to trade-off between the brightness and the smoothness constraints OF estimation.

An exploratory study has been carried out on a one-speaker sequence. The results shown an evolution of the EC conformed to what was expected. For a fixed value of λ

and an increasing number of EC iterations, the EC between the resulting OFs and the audio feature reaches a maximum, with no changes for further increase of ι: the system has reached its equilibrium. When ι is kept fixed while λ is increased, the EC goes through a maximum before decreasing for higher values of λ. The introduction of a more important smoothness constraint in the OF equation improves the information in a first time, before degrading it. It was shown that this variation of the weight associated to the smoothness constraint accounts for a scale-space approach to the OF estimation. The proper scale of the audio-visual system is found for the λ value maximizing the EC. Therefore, there exist an optimal pair of (ι, λ) parameter, such that the resulting OF is, maybe not more accurate, but more related to the co-occurring acoustic speech signal.

However, the joint variation of the two parameters shown that increasing values of ι result in increased values of the λ-optimal EC value. The 2D optimization problem is likely to not converge towards a global solution. A possibility might be to turn the SO problem into a MO one by adding a constraint on the number of iterations. Indeed, the higher this parameter the slower the OF estimation. A Pareto approach could then be taken to solve this MO problem.

In this chapter, a simplified 1D version of the 2D optimization problem has been rather undertaken, letting the resolution of the MO approach for future work. The addressed 1D optimization problem looks for the optimal value of the λ parameter only, the number of iterations ι being fixed empirically. The deterministic gradient-free extension of the Shubert's algorithm developed by Jones *et al.* in [68] is used to reach efficiently the solution.

A set of experiments has been carried out on the 11 CUAVE sequences, previously used in chapter 5, to evaluate the performance of the system for audiovisual speaker detection. Only the video features are automatically extracted. Thus different choices of audio features have been tested in turn. Each choice led to different performance of the speaker detector and the best one was not achieved with the features initially defined. These results stresses how crucial is the extraction of specific features in the PR process and why an automatic approach is important to avoid multiple trials or unreliability in the results. Future works should try of course to apply the feature extraction framework proposed in this thesis to the audio and the video modalities jointly.

A comparative study has then been performed using two kinds of non-optimized video features in turn with the audio features leading to the best results in the previous experiments (namely, the first MFCC). It appeared that the optimized video features do not significantly improve the speaker detection results on this test set. However, this set is small and deals with clips shot in very clean and simple conditions. Thus the added-value of the scale-space approach might simply not shows up in such conditions: tests on noisier image sequences should be performed prior to draw definitive conclusions. Nevertheless, there are advantages in using this method for automatically set-up the regularization parameter. For different choices of λ might change noticeably the results on some sequences (once again, noisier conditions might even accentuate the difficulties of setting λ properly). The optimization scheme can alleviate this limitation.

Finally, the system evaluation step, defined in sec. 4.4 through the use of a Neyman-Pearson classifier, has been applied to appraise the performance of the whole pattern recognition chain, and in particular the added-value offered by the introduction of the video feature extraction step. Unlikely, the simplified version of the video feature optimization proposed and tested in this chapter does not improve the discrimination power of the classifier (at least on these simple sequences).

As a conclusion, the proposed video feature extraction framework has shown some advantages such as avoiding multiple trials for setting properly the λ parameter. However, further tests should be performed on noisier sequences to evaluate whether or not all the potentialities of the method have been established. A solution to the complete 2D optimization problem should also be developed, which should give rise to better performance of the PR system. It must be noticed also that a further step might be required in the extraction of suitable video features. This step would aim at reduce the feature set so that only the salient components of the optical flow would be kept. Indeed, the results obtained by now might also be explained by the fact that the irrelevant motions are not discarded in the presented framework.

7 Conclusions and perspectives

7.1 Discussed topics and achievements

Throughout this thesis, we have explored the possibilities provided by a multimodal approach to the problem of speaker detection. A multimodal pattern recognition system is proposed. If it is more specifically dedicated to speaker detection, it can also apply to any similar detection task where a hidden source yields two signals of different modalities. The purpose is to take advantage of the multimodal specificity of the problem to increase the detector performance.

For each step of a standard pattern recognizer a solution has been presented in order to minimize the probability of error of the whole system. Answers have been produced for each of the three points arose by the question, central to this work and ensuing from the choice of a multimodal approach: *"What do we fuse, how and when?"*.

- *What*: The acoustic and visual speech signals are the dedicated modalities in the context of a speaker detection task. They are acquired by a single camera and microphone.
- *How*: A multimodal framework based on information theory has been proposed for performing the feature extraction as well as for the classification itself.
- *When*: The presented approach combines a fusion at both feature- and decision-levels (hybrid approach). The information presents in each modality is used jointly to extract optimized features, but there are still two sets of features which input the classifier. The latter performs a direct fusion of the information since it outputs a single outcome.

The basic principle of the detection consists in evaluating the synchrony between audio and video features extracted from potentially speaking mouths, in order to classify each mouth as speaking or not. This synchrony is evaluated through a mutual information based function.

A key to success is the extraction of suitable features. The audiovisual data are then processed through an information feature extraction framework after having been acquired and represented in a tractable way. This feature extraction framework uses jointly the two modalities in a feature-level fusion scheme in order to recover the information originating from the common source while the independent noise is discarded. This approach is shown to minimize the probability of committing an error on the source estimate.

These optimal features feed in the classifier, which comes at the next processing step. This classifier is defined through an hypothesis testing approach. It fuses the two modalities to output a single decision about the label of each candidate mouth region ("speaker" or "non-speaker"). The hypothesis testing approach gives means for evaluating the performance of the classifier itself but also of the system. This is one of the major contribution of this thesis as it permits in particular to assess the added value offered by the feature extraction step.

The pattern recognition system has been applied to the problem of identifying the current speaker among several candidates in audio-video sequences with emphasis puts in turn on the audio and the video modalities.

As far as the audio modality is concerned, a new acoustic feature has been proposed. It consists in a linear combination of mel-cepstrum coefficients which optimizes the efficiency coefficient - defined in the feature extraction framework - with the video features. It gives rise to a challenging optimization problem with an objective function plagued by many local optima. For this reason, three optimization methods, one local and two global have been tested in turn and their performance compared for the most efficient to be finally retained. Results have shown that the optimized acoustic features where specific to speech production. The analysis of the system performance through the developed evaluation framework demonstrated the added-value of the feature extraction step as it increases the discrimination capacity of the classifier.

As far as the video is concerned, the mouth motion is obviously a visual information specifically related to the speech production process. It is well-known that motion is a 3D phenomenon which is hardly estimated correctly by image processing algorithms. A common estimation of the apparent motion in image sequences is known as the optical flow. It is estimated from the image intensity gradient. A probabilistic model of the relationships between the audio, the video and the image intensity gradient has been proposed trough graph theory, in the particular case of a speaking mouth. The analysis of this model led back to the information theoretic estimator defined in the feature extraction framework: the efficiency coefficient. Therefore, increasing this coefficient between the audio features and the optical flow is expected to improve the performance of the MI-based classifier.

A corresponding optimization framework has been introduced. In its initial version, the

two parameters required for gradient-based OF estimation - regularization parameter and number of iterations - are to be optimized with respect to a cost function assessing the improvement they lead to. This improvement is measured by evaluating the EC between the resulting optical flow and some energy-based features extracted from the co-occurring acoustic signal. The results of an exploratory study shown that this 2D optimization problem did not have a single global optimum. A simplified 1D version of the problem has been addressed in this thesis, looking for the optimal value of the regularization parameter only. Actually, this optimization scheme accounts for a scale-space approach to optical flow estimation. The evaluation of the classification chain through the evaluation process framework shown that this simplified problem did not improve (nor degrade) the discrimination capabilities of the classifier. It presents however the noticeable advantage of giving means for an automatic set up of the regularization parameter, being sure to not have the worse setting for a given sequence.

As a conclusion, a complete pattern recognition framework dedicated to audiovisual speaker detection has been proposed in this thesis, where the probability of misclassifying a mouth as "speaker" or "non-speaker" is minimized. The importance of fusing the audio and video content as soon as at the feature level has demonstrated through the system evaluation stage included in the pattern recognition process.

This framework can be applied with success in applications where real-time is not a priority, like multimedia content indexing for example.

7.2 Future research directions

There are many directions future researches can take. One of them concerns the multimodal approach to optical flow - in case where a speaking mouth is concerned - whose basis have been presented in chapter 6. Only a simplified 1D version of the resulting multi-objective optimization problem has been addressed in this thesis: only one of the two parameters required for OF optimization was optimized. However, some clues for undertaking the 2D optimization problem have been given in chapter 6: through a Pareto approach, a first analysis of the problem and of its solutions could be achieved. This could be performed by picking up simply different solutions on the Pareto front (i.e. solutions corresponding to different trade-offs between the different objectives) and analyzing the improvements they lead to. Based on these results, a further challenge should be to integrate the audio constraint directly in the optimization function, i.e. to re-define a mathematical framework leading straight away to a multimodal optical flow estimation.

Of course, a prior work to this challenge should be to test the actual method in noisy conditions, in order to assess the benefit of automatically setting the regularization parameter λ.

Furthermore, as pointed out in the conclusions of chapter 6, a further step might be required in the extraction of suitable video features. This step would aim at reduce the feature set so that only the salient components of the optical flow would be kept. Indeed, the

results (the discrimination power of the PR system in particular) might improve provided the irrelevant motions are discarded prior to the classification. The use of representative optical flow basis for mouth motions, as proposed in [21] for speech recognition, might be an interesting approach to the problem.

Other possible improvements have been pointed out when optimizing the audio modality thought they would also apply in the video optimization scheme. The main issue would be to extend the method to the detection of the four possible speaking cases (in two-speaker sequences), including the silent and both speaking cases, and not only the cases where one individual only is speaking.

Another interesting point to tackle would be to embed the method in a global mouth detector approach instead of performing the mouth extraction independently and prior to apply the speaker detection method. Despite of the pupil tracker, the mouth extraction as presented in sec. 5.3 introduced indeed noise into the mouth motion through the slight variations of the mouth position with respect to the face coordinate system. This effect should be compensate by a multimodal mouth extraction. Also, such an approach should possibly be able to address more complex situations (where faces would move towards the camera for example).

By introducing the multimodal speaker detection into the mouth detector, we come closer to a simultaneous optimization of the audio and the video modality (as mentioned in chapter 5, restricting the motion to the mouth region accounts for an optimization of the video features in some way). This leads us to the last main issue future work might address: the development of a method for simultaneously extract optimized audio and video features. The importance of using specific features for handling the detection task efficiently has been demonstrated throughout this thesis. In chapter 6 for example, it has been shown that the results were dependent to some extend on the empirical choice of the audio features. Optimizing jointly the two modalities would thus be certainly more efficient and lead to best results. A sequential optimization approach can be undertaken. But another interesting approach to consider might be to optimize a multimodal feature, defined for example as the concatenation of the acoustic and visual features, in a similar way to [121].

Experimental evaluation framework for speaker detection on the CUAVE database

A

A.1 Introduction

Multimodal speaker detection is a field which has recently risen up. As different methods are emerging, the definition of common evaluation frameworks for objectively evaluate their performance becomes a critical issue.

We have brought up a discussion on the subject in [16], leading to the proposition of such an evaluation framework. The latter aims at becoming a standard to validate speaker detection methods on audio-visual databases like the CUAVE one [98], [99]. It is dedicated to approaches that require a temporal analysis window on which the detection is performed. The duration of this analysis window is considered of t frames or seconds and is shifted by a given value of s frames or seconds.

In a first part, the CUAVE corpus is shortly presented. The second part briefly reviews two speaker detection evaluation frameworks taken from the literature. The main advantages and drawbacks of these experimental methods are exposed before discussing possible solutions in the third part. Finally, an evaluation methodology based on the previous argumentation is proposed.

A.2 Description of the CUAVE database

The CUAVE speech corpus [98], [99] is a moving-talker speaker-independent database, designed to aid researchers in multimodal speech processing. 36 individual speakers and


Figure A.1 — Two examples of sequences involving two persons from CUAVE.

22 speaker pairs utter continuous, connected and isolated digits. The sequences present two distinct parts of different complexity. In the first part, the speakers remain still, with only some small, natural motions. In the last part, the speaker moves around intentionally in the individual sequences and both persons are speaking simultaneously in the multiple speaker sequences.

The individual sequences are about 2 minutes long, and the group ones about 20–25 seconds long. The NTSC video standard was used (29.97fps) and the stereo audio signal was sampled at 44kHz.

A.3 Existing evaluation frameworks

To the best of our knowledge, the CUAVE database has been used to test speaker detection algorithms in two cases [95], [89]. This database has been more widely used in the context of speech recognition, but due to the difference between the two problems, the evaluation frameworks are not directly applicable to the speaker detection case.

The goal of speech recognition is, as the name suggests, that of recognizing the words uttered by a speaker using acoustic and/or visual information. Thus, the evaluation of speech recognition algorithms requires a precise ground truth of the sequences, including the words pronounced, as well as a precise evaluation procedure. On the other hand, the goal of a speaker detection algorithm is that of detecting the person who is currently speaking, in a multi-speaker environment or in adverse conditions. Thus, no information about the content of the speech is needed, and the evaluation framework can be more relaxed than in the previous case.

Since the framework we are introducing is devoted to the evaluation of speaker detection algorithms performances, we consider only the multi-speaker partition of the database, as it was done in [95], [89]. Each of the 22 clips involves two speakers arranged as in Fig. A.1, taking turn in reading series of digits. At the end of the clips, both subjects speak simultaneously reading different sequences of digits. The final part of each sequence has been discarded, and only the parts on which a single person is speaking are considered.

Nock *et al.* present in [95] an empirical study of speaker detection based on audio-visual

synchrony. Tests are carried out on the multi-speaker portion of the CUAVE database using the experimental protocol described in the following.

The speaker detection function is based on mutual information and therefore requires a static analysis on a given temporal window. This temporal window is $t = 2$ seconds long (i.e. 60 frames) and it is shifted by $s = 1$ second (i.e. 30 frames) along the sequence. The results are compared to the ground truth at the center point of each temporal window: the estimates are therefore scored at one second intervals through each clip. Thus, performing their tests on 12 sequences, they end up with a total of 252 test points.

In [89], the authors also used the multi-speaker part of CUAVE to test their speaker detection method. Since their method also requires the use of temporal windows, they define an analysis window of length 60 frames, which is shifted by 20 frames. They eventually end up with 273 test points corresponding to the last frame of the analysis windows.

Clearly, the main advantage of such methods is that the pre-processing time is reduced. The ground truth points where the detection function must be evaluate are in fact easily and quickly established.

However, four main drawbacks can be identified for these two approaches:

- The evaluation function may not be accurate enough. Since the truth about the current speaker for a one second interval is given from a unique frame, this make evaluation not very reliable: what happens if speaker 1 is mostly speaking over the current temporal window, but speaker 2 is actually speaking at the observed time instant? Or if nobody is speaking at the considered time instant?

- From the detection algorithm point of view, it does not seem much reasonable to consider 2 seconds of information to perform the detection, and to compare the result to a single 1/60th of the information in the ground truth. Taking again the previous example, a good detection algorithm should indicate speaker 1 as the active speaker if it is the most active on the whole window. But it may happen that the ground truth will indicate such a result as false.

- The evaluation results are very sensitive to the choice of the ground truth and to the choice of the speaker detector's parameters. If, for example, the detection is evaluated not on the central frame of the analysis window but on the next one, the results may be very different.

- Notice that choosing the central frame of the analysis window for the evaluation, somewhere means that the method needs the past and future frames to perform the current speaker detection.

A.4 Possible solutions and limits of these solutions

The first point to consider is how to establish the ground truth to which the detector outputs should be compared, in order to assess the detector performances. Let us just recall here

that we only consider speaker detection methods where an analysis window is required.

Two ways of establishing the ground truth can be imagined:

1. **Frame level:** each frame is labelled with the current speaker label.
2. **Window level:** each of the t frames that constitute the analysis window are labelled with the label of the person speaking the majority of the time during this period.

The first point to stress is that establishing the ground truth is a tedious and uneasy task, whatever is the chosen approach. The audio and video signals are not perfectly aligned: movements appear in the mouth region before any sound can be heard (co-articulation effect). Thus the starting and stopping frames are difficult to be labelled.

From this point of view, the window level ground truth is much more sensitive to the choices made in the labelling step: let us just consider that a change of speaker occurs in the middle of a period of t frames. If each person is speaking approximately for the same amount of time, it is then a big issue to decide which one is the dominant speaker, and this is strongly dependant on how the labelling has been done. We might consider giving more weight to the speaker already labelled as the dominant one in the previous period.

On the other hand, it must be noticed that the detection algorithm is not performing at the frame level. Then, if choosing to establish the ground truth at the frame level, the evaluation of the method can be performed at a too high resolution.

The possible ways to constitute the ground truth have been discussed. Now we have to consider the evaluation criterion, i.e. the way the detector outputs are compared to the ground truth. Once again, different approaches may be considered.

If the window level ground truth is used, the comparison is straightforward: the output of the detector for a given analysis window is compared to the ground truth established for this set of frames. Detection and evaluation are performed on the same temporal window, for the same duration. However, the detector's analysis windows and the ground truth windows should perfectly overlap, in order to avoid further processing of the results.

When considering the frame level ground truth, three different approaches can be adopted:

1. Only the ground truth corresponding to the central frame, or any given frame, of the analysis window is compared to the detector output for the given analysis window. This is the approach adopted in [95] and [89]. As discussed previously, this approach make obviously the problem very simple: there is little chance to fall on a silent frame. The ground truth is easily established and we have little chance to face the tricky situation where a speaker is starting or stopping to talk on that very frame.
2. The output of the detector for a given analysis window is compared to each of the t corresponding frames in the ground truth. By using this approach, the number of test points increases compared to the previous case. Thus the evaluation is more accurate. However, if the analysis windows are overlapping, some frames will be scored more

than once in the final results. There might be contradictory labels at the output of the detector for a given frame, and additional processing steps are required.

3. A third, and more natural option in the case of overlapping analysis windows, might be to consider a trade-off between the two previous options, where a given fraction of the frame set belonging to the analysis window is compared to the output of the detector for the same set of frames. This last approach presents the advantage of being more accurate than the first one without the problem of multiple scoring frames, provided a judicious choice for the shifting window value and for the number of frames on which the algorithm performance is evaluated.

The advantage of using the frame level ground truth with approaches 2 or 3 rather than the window level ground truth is that it allows to cope with speaker turn points if these ones fall in the current analysis window. Let us consider the case where the two persons are speaking about the same amount of time in a given analysis window where a turn point occurs. If the window level ground truth has been chosen, the output of the detector will be judged as either 100% right, or 100% false. Whereas, if the frame level ground truth has been chosen, the performance of the algorithm will anyway be proportionate to the situation.

In addition, the case of the silent frames must be carefully studied in both cases, and especially using the frame level ground truth. In this case, a "silent state" has to be considered when the number of consecutive silent frames is above a certain number L.

The last point to be discussed concerns the choice of the analysis window length and of the shift parameter value. The length of the analysis window has to be a trade-off between the algorithm requirements, the computation time, and the "inertia" of the detection (i.e. the delay between two detections). The value of the shift parameter determines the resolution of the detection algorithm. The best approach in that sense would be to shift the window by one frame all along the sequence. It is also the most expensive from a computation time point of view. The less time consuming approach would be, on the contrary, to make use of non-overlapping windows. But then the resolution is much lower and the results may drop significantly. In particular, this approach is not accurate enough to cope with the speaker turn points. Therefore a trade-off has to be found between accuracy and computation time.

A.5 Proposed experimental evaluation framework

Since our aim is that of evaluating speaker detection algorithms, in here we consider the multi-speaker partition of the CUAVE database. This section includes 22 sequences exhibiting two persons taking turn in reading series of digits. We have decided to build the ground truth using a frame level approach. Each frame has a label (0, 1 or 2), which indicates if no one (0), the left person (1) or the right person (2) is speaking. A group of frames is labelled as silent (0) when it is composed of at least $L = 25$ frames. This value

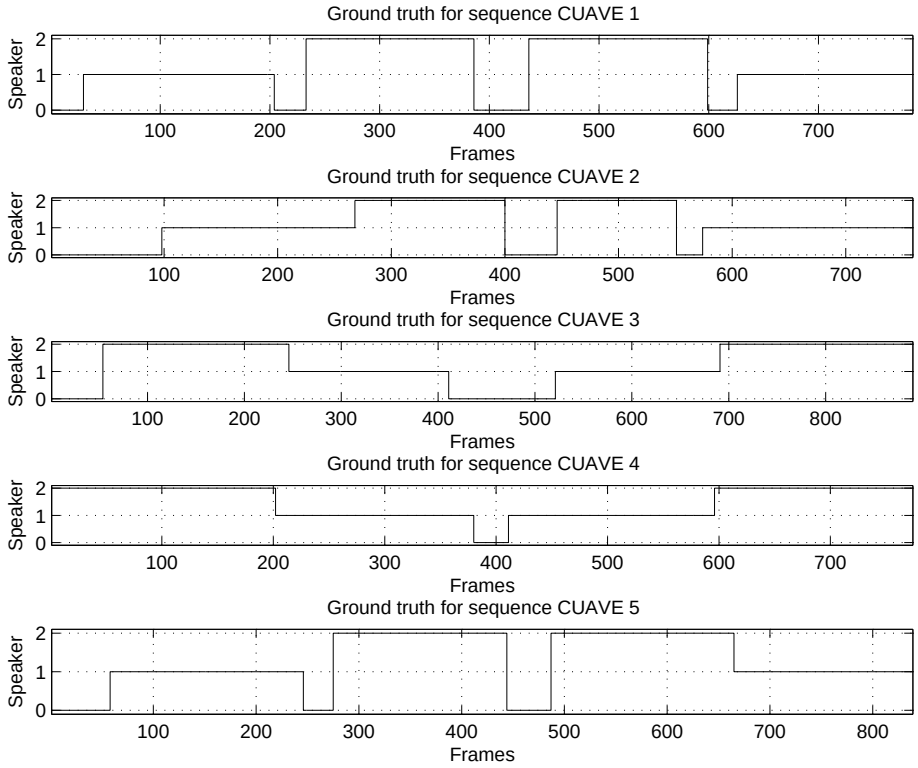


Figure A.2 — Ground truth labels for the first five sequences of the group partition of the CUAVE database.

corresponds to the perceived limit between an interruption in the speech flow and a "true" silence. An example of the obtained labels for the first five sequences of the group partition of the CUAVE database is shown in Fig. A.2. The complete set of labels for all the 22 sequences is available on the author's web page [15].

For what concerns the evaluation of the speaker detector's results, we propose to use a window-based evaluation method. Speaker detection algorithms typically output sets of frames denoted by a single speaker label. The detection is considered to be correct if the detector's output for a given window matches the most present label in the corresponding ground truth window (see Fig. A.3).

A.6 Conclusions

In this report, problems related to the evaluation of multimodal speaker detection methods have been discussed. An evaluation methodology has been proposed, in order to make possible the comparison between different algorithms, and the labelled ground truth for a multi-speaker audiovisual database, the CUAVE corpus, has been made available.

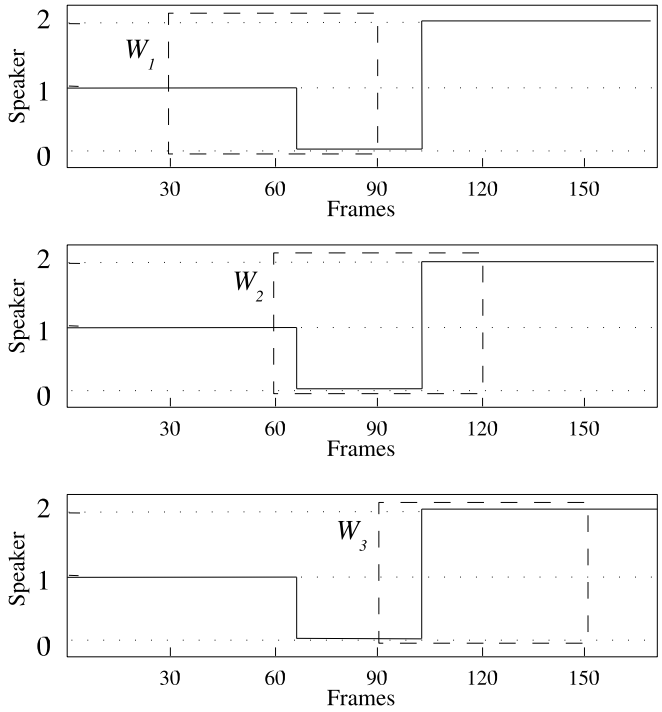


Figure A.3 — Schematic representation of a sliding detection window applied on the ground truth. The detector output for each of the three window will be compared to the corresponding window ground truths: 1 for W_1 (top), 0 for W_2 (middle) and 2 for W_3 (bottom).

Shubert's and DIRECT optimization algorithms

B

B.0.1 Shubert's algorithm for Lipschitzian minimization problems

The Shubert's algorithm [117] is a well-known general approach for finding the global minimum in a bracketed region. Basically, it uses a lower envelope of the function to estimate this global minimum. Let us now describe this method more in details.

In its initial form, this algorithm solves optimization problems involving one-dimensional Lipschitzian objective functions. A function is said to be Lipschitzian if a bound can be assigned on its rate-of-change. This bound is called the Lipschitz constant.

Definition 4 ***(Lipschitz constant)*** *- Let $M \subseteq \mathbb{R}$ and $f : M \to \mathbb{R}$. The function f is called Lipschitz continuous on M with Lipschitz constant K if:*

$$|f(x) - f(x')| \leq K|x - x'| \quad \forall x, x' \in M. \tag{B.1}$$

The Shubert's algorithm exploits Eq. (B.1) to iteratively reach the minimum of the objective function. For an interval $M = [a, b]$, the Eq. (B.1) leads to the two following upper-bounding inequalities:

$$f(x) \geq f(a) - K(x - a), \tag{B.2}$$

$$f(x) \geq f(b) + K(x - b). \tag{B.3}$$

These two bounding functions are two lines with slopes $-K$, $+K$. Their intersection forms an under-estimator for f and their crossing point (x_1, B_1), given below, provides a lower

bound on the minimum of f:

$$x_1 = \frac{(a+b)}{2} + \frac{[f(a) - f(b)]}{2K}, \tag{B.4}$$

$$B_1 = \frac{[f(a) + f(b)]}{2} - K(b-a). \tag{B.5}$$

The searching interval M is subdivided in two smaller regions $M_1 = [a, x_1]$, $M_2 = [x_1, b]$, and Eq. (B.1) is applied to each of these two regions. The region where a better lower bound B on the minimum value of f is found is then subdivided, while the other one is pruned. The process continues further until the stopping criterion is met. This stopping criterion consists usually in a given number of iterations, or in a pre-specified tolerance on the minimum of the approximation with respect to the current best solution. Fig. B.1 illustrates a few steps of the algorithm.

Three main restrictions to Lipschitzian algorithms can be pointed out [69]:

1. The need of specifying the Lipschitz constant.
2. A slow convergence of the process.
3. A curse of dimensionality that limits them to moderate-size problems.

The two first points are closely related. A Lipschitz constant K can be estimated using for example the derivative of the objective function. However, when this function is not differentiable, or when no analytical form of this derivative is known, the determination of K becomes a problem. Actually, determining the Lipschitz constant of a function is itself an optimization problem [142]. Often, K is then fixed at a quite high value so as to bound weakly the rate-of-change of the function. Since the Lipschitz constant remains fix during the optimization process, the algorithm convergence may becomes slow.

Let us closely look at Eq. (B.5) and explain the meaning of its two terms $[(f(a) + f(b))/2]$ and $[-K(b-a)]$. The first term leads us to select intervals where previous function evaluations have been good: it leads to do a local search [68]. The second term is lower, thus better, if the size of the search interval $[a, b]$ is large. In other words, large unexplore territories are favored which leads us to do global search. The Lipschitz constant K clearly appears as a weighting factor on the relative importance between the global versus the local search in the process, as stressed by Jones *et al.* in [68]. Therefore, if K is large, only global search is performed and the process is slowed down.

B.0.2 DIRECT algorithm for 1D minimization problems

Starting from this observation, Jones and al. have developed in [68] the so-called DIRECT algorithm (acronym for DIviding RECtangle) which alleviates the two first initial restrictions to standard Lipschitzian methods: the need to specify the Lipschitz constant is avoided by carrying out simultaneous searches using all possible constants. This way, a balance is stroke between the global and the local search: by using different constants, the

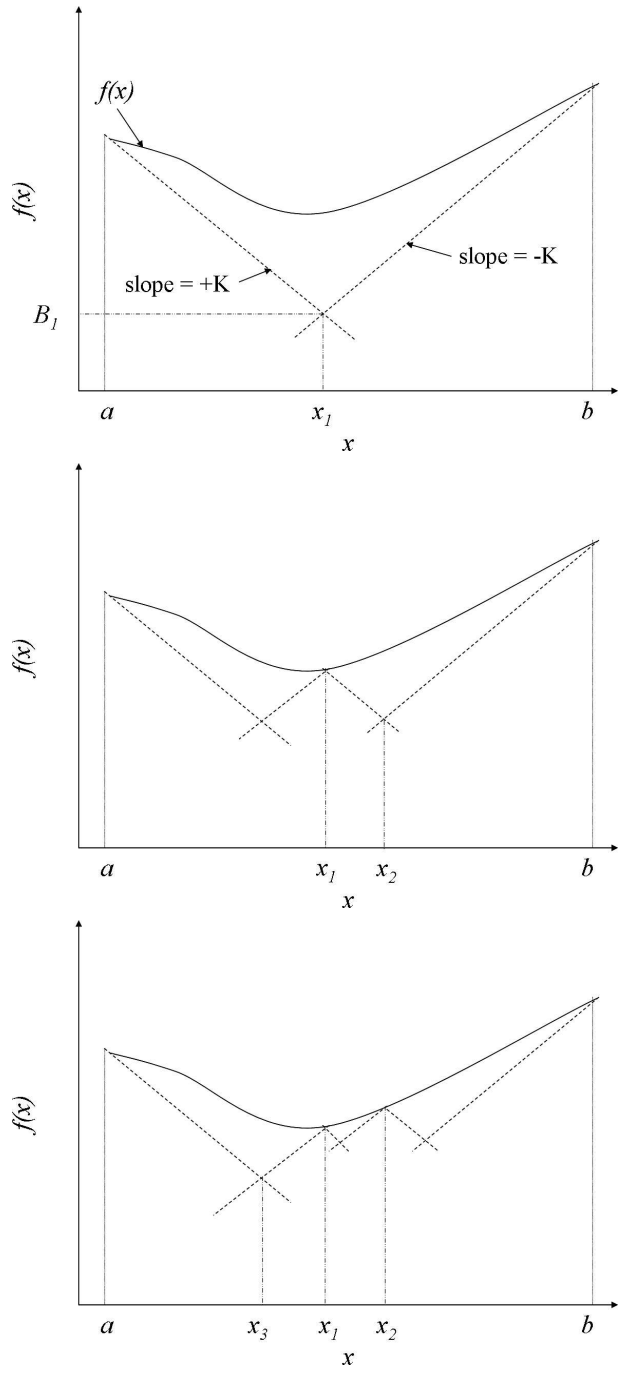


Figure B.1 — Shubert's algorithm for three iterations.

algorithm firstly works at the global level and once the basin of convergence is found, it can switches to the local level to be more efficient.

Incidently, Jones' method also deals efficiently with ND objective functions by changing the sampling point position (i.e. the value used to select the intervals). The center point c of the interval instead of the endpoints, then the number of function evaluations is reduced.

Thus, the third and last weak points previously mentioned are also improved.

The lower bound on the value of the 1D function for an interval $[a, b]$, previously denoted by B, is now given by:

$$B' = f(c) - K(b-a)/2, \tag{B.6}$$

where $f(c)$ stands for the objective function evaluated at the center point $c = (a+b)/2$ of the interval $[a, b]$. Also, instead of pruning the non-selected regions, DIRECT algorithm samples all the intervals that are labelled as potentially optimal. The assignment of this label is ruled by [68]:

Definition 5 ***(Potentially optimal interval)*** *- Suppose that the interval M has been partitioned into intervals $[a_i, b_i]$ with midpoints c_i, for $i = 1, \ldots, m$. Let $\epsilon > 0$ be a positive constant, and f_{min} be the current best function value. Interval j is said to be potentially optimal if there exist some rate-of-change constant $\tilde{K} > 0$ such that*

$$f(c_j) - \tilde{K}[(b_j, a_j)/2] \leq f(c_i) - \tilde{K}[(b_i - a_i)/2)], \quad \forall i = 1, \ldots, m, \tag{B.7}$$

$$f(c_j) - \tilde{K}[(b_j - a_j)/2] \leq f_{min} - \epsilon|f_{min}|. \tag{B.8}$$

The parameter ϵ indicates that $f(c_j)$ exceeds the current best solution by a non-trivial amount. According to Jones and al., it has a negligible effect provided it lies in $[10^{-2}, 10^{-7}]$. Note that the tilde in $\tilde{\text{K}}$ is used to strike that $\tilde{\text{K}}$ is not a Lipschitz constant as defined by Def.. 4 but a simple rate-of-change constant.

Bibliography

[1] (2007). *Oxford English Dictionary*. Oxford University Press.

[2] J. I. Ansell, P. M. J. (1994). *Practical Methods for Reliability Data Analysis*. Oxford University Press.

[3] R. Antony (1995). *Principles of data fusion automation*. Artech House.

[4] G. Aubert, R. Deriche, P. Kornprobst (1999). Computing optical flow via variational techniques. *SIAM Journal on Applied Mathematics* **60**(1):156–182.

[5] J. L. Barron, D. J. Fleet, S. S. Beauchemin (1994). Performance of optical flow techniques. *International Journal of Computer Vision* **12**(1):43–77.

[6] R. Battiti, E. Amaldi, C. Koch (1991). Computing optical flow across multiple scales: an adaptive coarse-to-fine strategy. *International Journal of Computer Vision* **6**(2):133–145.

[7] M. J. Beal, N. Jojic, H. Attias (2003). A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(7):828–836.

[8] S. S. Beauchemin, J. L. Barron (1995). The computation of optical flow. *ACM Computing Surveys* **27**(3):433–467.

[9] R. E. Bellman (1961). *Adaptive control process: A guided tour*. Princeton University Press.

[10] P. Bertelson, B. de Gelder (2004). *Crossmodal space and crossmodal attention*, chap. 7: The psychology of multimodal perception, pp. 141–177. Oxford University Press.

[11] P. Besson, M. Kunt (2005). *Information theoretic optimization of audio features for multimodal speaker detection*. EPFL-ITS Tech. Rep. 08/2005, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

[12] P. Besson, M. Kunt (2006). Hypothesis testing as a performance evaluation method for multimodal speaker detection. In *2nd International Workshop on Biosignal Processing and Classification (BPC2006), ICINCO*, pp. pp. 106–115, Setúbal, Portugal.

[13] P. Besson, M. Kunt (2007). Evaluation of multimodal speaker detection using hypothesis testing. *Submitted to Journal of NeuroEngineering and Rehabilitation (JNER)* (invited paper).

[14] P. Besson, M. Kunt, T. Butz, J.-P. Thiran (2005). A multimodal approach to extract optimized audio features for speaker detection. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey.

[15] P. Besson, G. Monaci, P. Vandergheynst, M. Kunt (2006). CUAVE database ground truth. Available: http://itswww.epfl .ch/~besson/cuave gt.html.

[16] P. Besson, G. Monaci, P. Vandergheynst, M. Kunt (2006). *Experimental evalutation framework for speaker detection on the CUAVE database*. Tech. Rep. TR-ITS-2006.003, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

[17] P. Besson, J.-M. Vesin, V. Popovici, M. Kunt (2006). Differential evolution applied to a multimodal information theoretic optimization problem. In *Proceedings of the 8th European Workshop on Evolutionary Computation in Image Analysis and Signal Processing (evoIASP)*, LNCS 3907, pp. 505–509, Budapest, Hungary.

[18] P. Besson et al. (2005). *Extraction of Audio Features Specific to Speech using Information Theory and Differential Evolution*. Tech. Rep. TR-ITS-2005.018, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

[19] P. Besson et al. (2007). Extraction of audio features specific to speech production for multimodal speaker detection. *To appear in IEEE Transactions on Multimedia.* .

[20] C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer.

[21] M. J. Black, Y. Yacoob, A. D. Jepson, D. J. Fleet (1997). Learning parameterized models of image motion. In *ICCV*, Puerto Rico.

[22] S.-T. Bow (2002). *Pattern Recognition and Image Preprocessing*. Signal Processing and Communications Series. Marcel Dekker, 2nd edn.

[23] A. W. Bowman, A. Azzalini (1997). *Applied smoothing techniques for data analysis*. Oxford science publications.

[24] X. Bresson, P. Vandergheynst, J.-P. Thiran (2006). Multiscale active contours. *International Journal of Computer Vision* **70**(3):197–211.

[25] H. Brunk (1975). *An Introduction to Mathematical Statistics*. Xerox college publishing, third edn.

[26] T. Butz (2003). *From error probability to information theoretic signal and image processing*. Ph.D. thesis, Ecole Polytechnique Fd́ŕale de Lausanne, EPFL, Lausanne, Switzerland.

[27] T. Butz, J.-P. Thiran (2002). Feature space mutual information in speech-video sequences. In *Proceedings of ICME*, vol. 2, pp. 361–364, Lausanne, Switzerland.

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

[29] C. C. Chibelushi, F. Deravi, J. S. Mason (2002). A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia* **4**(1):23–37.

[30] D. G. Childers, D. P. Skinner, R. C. Kemerait (1977). The cepstrum: A guide to processing. In *Procedings of the IEEE*, vol. 65, pp. 1428–1443.

[31] A. A. Cole-Rhodes, K. L. Johnson, J. LeMoigne, I. Zavorin (2003). Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient. *IEEE Transaction on Image Processing* **12**(12):1495–1511.

[32] T. M. Cover, J. A. Thomas (1991). *Elements of Information Theory*. John Wiley & Sons.

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

[34] B. de Gelder, P. Bertelson, J. Vroomen (1996). *Speechreading by humans and machines*, vol. 150 of *NATO ASI Series F*, chap. Aspects of modality in audio-visual processes, pp. 179–192. Springer-Verlag, Gmbh.

[35] L. Devroye, L. Györfi (1985). *Nonparametric Density Estimation*. Probability and mathematical statistics. John Wiley & Sons.

[36] L. Diehl, Randy, K. R. Kluender (1989). On the objects of speech perception. *Ecological Psychology* **1**:121–144.

[37] M. O. Ernst, M. S. Banks (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**:429–433.

[38] J. Esteban et al. (2005). A review of data fusion models and architectures: towards engineering guidelines. *Neural Computation and Applications* **14**:273–281.

[39] E. Fan (2002). *Global optimization of Lennard-Jones atomic clusters*. Master's thesis, McMaster University, Ontario, Canada.

[40] G. Fant (1960). *The acoustic theory of speech production*. Mouton de Gruyter, The Hague, Netherlands.

[41] L. Farkas (1994). *Anthropometry of the Head and Face*. Raven Press.

[42] T. Fawcett (2003). *ROC Graphs: Notes and practical considerations for researchers*. Tech. Rep. HPL-2003-4, HP Laboratories.

[43] D. E. Finkel (2003). *DIRECT Optimization Algorithm User Guide*. Tech. rep., Center for Research in Scientific Computation, North Carolina State University.

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

[45] J. W. Fisher III, J. C. Principe (1998). A methodology for information theoretic feature extraction. In *Proceedings of International Joint Conference on Neural Networks*, vol. 3 of *IEEE World Congress on Computational Intelligence*, pp. 1712 – 1716, Anchorage, Alaska.

[46] D. J. Fleet, M. J. Black, Y. Yacoob, A. D. Jepson (2000). Design and use of linear models for image motion analysis. *International Journal of Computer Vision* **36**(3):171–193.

[47] A. Garg, V. Pavlović, J. M. Rehg (2003). Boosted learning in dynamic bayesian networks for multimodal speaker detection. In *Proceedings of the IEEE*, vol. 91, pp. 1355–1369.

[48] D. Gatica-Perez et al. (2003). Audio-visual speaker tracking with importance particle fitlers. In *Proceedings of the International Conference on Image Processing (ICIP)*, Barcelona, Spain.

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

[50] F. Glover (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research* **13**(5):533–549.

[51] B. Gold, N. Morgan (2000). *Speech and audio signal processing*. John Wiley & sons, Inc.

[52] H. G. Goodridge, M. G. Kay (1996). Multimedia sensor fusion for intelligent camera control. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 655–662.

[53] K. W. Grant (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America* **109**(5):2272–2275.

[54] W. Grant, Ken, P.-F. Seitz (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America* **108**(3):1197–1208.

[55] K. P. Green (1996). Studies of the McGurk effect: implications for theories of speech perception. In *Proceedings of the 4th International Conference on Spoken Language*, vol. 3, pp. 1652–1655.

[56] M. Gurban, J.-P. Thiran (2006). Multimodal speaker localization in a probabilistic framework. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*.

[57] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (eds.) (2006). *Feature Extraction, Foundations and Applications*. Series Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer.

[58] D. Hall, J. Llinas (1997). An introduction to multisensor data fusion. In *Procedings of the IEEE*, vol. 85.

[59] R. He, P. A. Narayana (2002). Global optimization of mutual information: application to three-dimensional retrospective registration of magnetic resonance images. *Computerized Medical Imaging and Graphics* **26**:277–292.

[60] J. Heron, D. Whitake, P. V. McGraw (2004). Sensory uncertainty governs the extent of audio-visual interaction. *Vision Research* **44**:2875–2884.

[61] J. Hershey, J. Movellan (1999). Audio-vision: Using audio-visual synchrony to locate sounds. In *Proc. of NIPS*, vol. 12, pp. 813–819, Denver, CO, USA.

[62] R. Hogg, J. Ledolter (eds.) (1987). *Engineering Statistics*. MacMillan.

[63] J. H. Holland (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

[64] B. K. P. Horn, B. G. Schunck (1981). Determining optical flow. *Artificial Intelligence* **17**:185–203.

[65] J. Hu et al. (2002). A self-calibrated speaker tracking system using both audio and video data. In *Proceedings of the International Conference in Control Applications*, vol. 2, pp. 731–735.

[66] A. T. Ihler, J. W. Fisher III, A. S. Willsky (2004). Nonparametric hypothesis tests for statistical dependency. *IEEE Transactions on Signal Processing* **52**(8):2234–2249.

[67] A. K. Jain, R. P. Duin, J. Mao (200). Statistical pattern recogition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1):4–37.

[68] D. Jones, C. D. Perttunen, B. Stuckman (1993). Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications (JOTA)* **79**(1):157–181.

[69] D. R. Jones, C. D. Perttunen, B. E. Stuckman (1992). Global optimization: beyond the Lipschitzian model. In *IEEE International Conference on Systems, Man and Cybernetics*, pp. 565–570, Chicago, USA.

[70] R. Joshi, A. C. Sanderson (1999). Minimal representation multisensor fusion using differential evolution. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* **29**(1):63–76.

[71] M. Kam, X. Zhu, P. Kalata (1997). Sensor fusion for mobile robot navigation. In *Proceedings of the IEEE*.

[72] B. Kapralos, M. Jenkins, M. E (2003). Audio-visual localization of multiple speakers in a video teleconferencing settings. *International Journal of Imaging Systems and Technology* **13**:95–105.

[73] E. Kidron, Y. Schechner, E. M. (2005). Pixels that sound. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pp. 88–95.

[74] S. Kirkpatrick, C. D. Gelatt, J. M. P. Vecchi (1983). Optimization by Simulated Annealing. *Science* **220**(4598):671–680.

[75] K. Koffka (1935). *Principles of Gestalt Psychology*. Harcourt-Brace, New-York.

[76] J. Konrad, E. Dubois (1992). Bayesian estimation of motion vector fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(9):910–927.

[77] Y.-W. Leung, Y. Wang (2001). An orthogonal genetic algorithm with quantization for global numerical optimization. *IEEE Transactions on Evolutionary Computation* **5**(1):41–53.

[78] A. M. Liberman (1996). *Speech: A special code*. MIT Press.

[79] A. M. Liberman, I. G. Mattingly (1985). The motor theory of speech revised. *Cognition* **21**:1–36.

[80] T. Lindeberg (1994). Scale-space theory: a basis tool for analysing structures at different scales. *Journal of Applied Statistics* **21**(2):225–270.

[81] T. Lindeberg (1996). Scale-space: A framework for handling image structures at multiple scales. In E. aan Zee (ed.), *Proceedings of the CERN school of Computing*, The Netherlands.

[82] R. Linsker (1988). An application of the principle of maximum information preservation to linear systems. In D. Touretzky (ed.), *Proceedings of Advances in Neural Information Processing Systems*, vol. 1, pp. 186–194, MorganKaufmann.

[83] D. Lo et al. (2003). Robust joint audio-video localization in video conferencing using reliability information. In *Proceedings of Instrumentation and Measurement Technology Conference*, Vail, CO, USA.

[84] N. Matsuo, H. Kitagawa, S. Nagata (1999). Speaker position detection system using audio-visual information. *FUJITSU Scientific and Technical Journal* **35**:212–220.

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

[86] G. F. Meyer, S. M. Wuerger, F. Rhrbein, C. Zetszsche (2005). Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Exp. Brain Res.* **166**:538–547.

[87] J. Meynet, V. Popovici, J.-P. Thiran (2005). *Face Detection with Mixtures of Boosted Discriminant Features*. Tech. Rep. 2005-35, EPFL, 1015 Ecublens.

[88] G. Monaci, O. Divorra Escoda, P. Vandergheynst (2005). Analysis of multimodal signals using redundant representations. In *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, pp. 814–820, Genova, Italy.

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

[90] G. Monaci et al. (2006). Learning multi-modal dictionaries: application to audio-visual data. In Springer-Verlag (ed.), *Proceedings of the International Workshop on Multimedia Content Representation, Classification and Security*, vol. 4105 of *LNC*, pp. 538–545.

[91] T. K. Moon, W. C. Stirling (2000). *Mathematical Methods and Algorithms for Signal Processing*. Prentice hall.

[92] K. Nakadai, K. Hidai, H. G. Okuno, H. Kitano (2002). Real-time speaker localization and speech separation by audio-visual integration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

[93] A. Newell, H. A. Simon (1957). The logic theory machine. a complex information processing system. *The Journal of Symbolic Logic* **22**(3):331–332.

[94] P. Ngatchou, A. Zarei, M. El-Sharkawi (2005). Pareto multi objective optimization. In *Proceedings of the International Conference on Intelligent Systems Application to Power Systems (ISAP)*, pp. 84–91.

[95] H. J. Nock, G. Iyengar, C. Neti (2003). Speaker localisation using audio-visual synchrony: An empirical study. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR)*, pp. 488–499, Urbana, IL, USA.

[96] H. G. Okuno, K. Nakadai, T. Lourens, H. Kitano (2004). Sound and visual tracking for humanoid robot. *Applied Intelligence* **20**(3):253–266.

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

[98] E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy (2002). CUAVE: a new audio-visual database for multimodal human-computer interface research. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 2017–2020, Orlando.

[99] 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* **11**:1189–1201.

[100] V. Pavlovic, A. Garg, J. R. Rehg (2000). Multimodal speaker detectoin using input/ouput dynamic bayesian networks. In *International Conference on Multimodal Interfaces*, pp. 308–316.

[101] J. Peebles, Peyton Z. (1987). *Probability, random variables, and random signal principles*. Electrical Engineering. McGraw-Hill, 2nd edn.

[102] J. W. Picone (1993). Signal modeling techniques in speech recognition. In *Proceedings of the IEEE*, vol. 81.

[103] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery (1992). *Numerical Recipes in C*. Cambridge University Press, 2nd edn.

[104] K. V. Price (1999). *New Ideas in Optimization*, chap. 6: An Introduction to Differential Evolution, pp. 79–108. McGraw-Hill.

[105] F. Provost, T. Fawcett (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pp. 43–48, Huntington Beach, CA.

[106] X. Qi, F. Palmieri (1992). *General properties of genetic algorithms in the euclidean space with adaptive mutation and crossover.* Tech. Rep. CT 06269-3157, Department of Electrical and System Engineering U-157, University of Connecticut, Storrs.

[107] X. Qi, F. Palmieri (1994). Theoretical analysis of evolutionary algorithms with an infinite population size in continuous space, Part I: basic properties of selection and mutation. Part II: analysis of the diversification role of the crossover. *IEEE Transaction on Neural Networks* **5**(1):102–129.

[108] S. Ravulapalli, S. Sarkar (2006). Association of sound to motion in video using perceptual organization. In *Proceedings of the 18th International Conference on Pattern Recognition*.

[109] L. D. Rosenblum, M. A. Schmuckler, J. A. Johnson (1997). The McGurk effect in infants. *Perception and Psychophysics* **59**(3):347–357.

[110] A. L. Roskies (1999). The binding problem. *Neuron* **24**:7–9.

[111] P. Schroeter, J.-M. Vesin, T. Langenberger, R. Meuli (1998). Robust parameter estimation of intensity distributions for brain magnetic resonance images. *IEEE Transactions on Medical Imaging* **17**(2):172–186.

[112] J.-L. Schwartz, F. Berthommier, C. Savariauy (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition* **93**:B69–B78.

[113] D. W. Scott, S. R. Sain (2005). *Data Mining and Computational Statistics*, vol. 23 of *Handbook of Statistics*, chap. 9: Multi-dimensional Density Estimation, pp. 229–262. Elsevier, Amsterdam.

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

[115] C. E. Shannon (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**:379–423.

[116] S. Shimojo, L. Shams (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current Optinion in Neurobiology* **11**:505–509.

[117] B. O. Shubert (1972). A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis* **9**(3):379–388.

[118] B. Silverman (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

[119] E. P. Simoncelli (1999). *Handbook of Computer Vision and Applications*, vol. 2, chap. 14, pp. 297–422. Academic Press, Spring.

[120] M. Slaney, M. Covell (2001). FaceSync: A linear operator for measuring synchronisation of video facial images and audio tracks. In *Proc. of NIPS*, vol. 13.

[121] P. Smaragdis, M. Casey (2003). Audio/visual independent components. In *Proceedings of ICA*, pp. 709–714, Nara, Japan.

[122] T. Spalek, P. Pietrzyk, Z. Sojka (2005). Application of the genetic algorithm joint with the Powell method to nonlinear least-squares fitting of powder EPR spectra. *J. Chem. Inf. Model.* **45**:18–29.

[123] S. Spors, R. Rabenstein, N. Strobel (2001). Joint audio-video object tracking. In *Proceedings of the International Conference on Image Processing (ICIP)*.

[124] S. S. Stevens, J. Volkmann, E. B. Newman (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America* **8**(3):185–190.

[125] C. Stiller, J. Konrad (1999). Estimating motion in image sequences: A tutorial on modeling and computation of 2D motion. *IEEE Signal Processing* **16**(4):70–91.

[126] R. Storn (2003). Differential Evolution homepage [Online]. Available: http://www.icsi.berkeley.edu/~storn/code.html.

[127] R. Storn, K. Price (1997). Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces. *Journal of Global Optimization* **11**:341–359.

[128] K. Takahashi, H. Yamasaki (1994). Audio-visual sensor fusion system for intelligent sound sensing. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 493–500.

[129] S. Theodoridis, K. Koutroumbas (1999). *Pattern recognition*. Academic Press.

[130] J. Tuomainen, T. S. Andersen, K. Tiippana, M. Sams (2005). Audio-visual speech perception is special. *Cognition* **96**:B13–B22.

[131] V. Vaerman (1999). *Multi-dimensional Object Modeling with Application to Medical Image Coding*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

[132] Varshney (1996). *Distributed Detection and Data Fusion*. Springer.

[133] D. Vaufreydaz (2002). *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Ph.D. thesis in computer sciences, University Joseph Fourier, Grenoble (France).

[134] J. Vermaak, M. Gangnet, A. Blake, P. Pérez (2001). Sequential Monte Carlo fusion of sound and vision for speaker tracking. In *Proceedings of the International Conference on Computer Vision*.

[135] J. Vroomen, B. de Gelder (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance* **26**:1583–1590.

[136] J. Vroomen, B. de Gelder (To appear in). *Handbook of multisensory processes*, chap. Perceptual effects of cross-modal stimulation: ventriloquism and the freezing phenomenon. MIT Press.

[137] L. Wald (1999). Some terms of reference in data fusion. *IEEE Transactions on Geosciences and Remote Sensing* **37**(3):1190–1193.

[138] E. Waltz, J. Llinas (1990). *Multisensor Data Fusion*. Artech House.

[139] C. Wang, M. Brandstein (1999). Multi-source face tracking with audio and visual data. In *Proceedings of the 3rd Workshop on Multimedia Signal Processing*, pp. 169–174.

[140] C. Wang, S. Griebel, M. Brandstein (2000). Robust automatic video-conferencing with multiple cameras and microphones. In *Proceedings of the International Conference on Multimedia and Expo*, vol. 3.

[141] A. R. Webb (2002). *Statistical Pattern Recognition*. Johm Wiley & Sons, Ltd, 2nd edn.

[142] G. R. Wood, B. P. Zhang (1996). Estimation of the Lipschitz constant of a function. *Journal of Global Optimization* **8**:91–103.

[143] D. N. Zotkin, R. Duraiswami, L. S. Davis (2002). Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing* **11**:1154–1164.

Curriculum Vitae

Full name:	**Patricia Besson**
Degrees:	Master of Science in Biomedical Engineering, (D.E.S.S. en Génie Biomédical), Université Lyon I, France
Address:	Signal Processing Institute (ITS) School of Engineering (STI) Swiss Federal Institute of Technology (EPFL) CH-1015 Lausanne Switzerland
Contact numbers:	Tel. (+41 21) 693 56 46 Fax. (+41 21) 693 76 00 E-mail: patricia.besson@epfl.ch
Civil status:	Single
Date and place of birth:	July 11$^{\text{th}}$ 1977, France
Nationality:	French

Professional Experiences

Since 2003 Research Assistant at the Signal Processing Institute (ITS), Swiss Federal Institute of Technology at Lausanne (EPFL) - leading to a Ph.D. in multimodal signal processing. Under the direction of Prof. Murat Kunt

- Research activities: audio and video signal processing, statistical pattern recognition, information theoretic approach to signal processing, optimization problems, feature extraction.
- Teaching: supervision of several master thesis, responsible of laboratories (signal processing field).
- Other: submission and acceptation of a proposal for a grant to the Swiss National Found.

2002-2003 Research assistant, KBA Giori - Signal Processing Institute, EPFL. Research activities: development of a color separation algorithm.

Internships

2001
4 months Scoliosis Computer Laboratory L.I.S.3D, Ste Justine's Hospital (Canada)
Image processing project: automatic detection of calibrating markers for a 3D reconstruction of the spinal column from two orthogonal views.

2000
2 months Ultrasound Department, Siemens (France)
Market analysis of the echographic equipment in Paris hospitals.

1999
2 months O.R.L. research laboratory, Saint Antoine's Hospital (France)
Audio signal processing project: analysis of the wavelet decomposition applied on speech signals for cochlear implant application.

1998
2 months Biomedical Department, Guy's Hospital (United Kingdom)
Work on "year 2000" project: evaluation of the equipment for year 2000 bug issue.

Publications

Journal Papers

- P. Besson and M. Kunt. Evaluation of multimodal speaker detection using hypothesis testing. *Submitted to Journal of NeuroEngineering and Rehabilitation (JNER)* (Invited paper), 2007.
- P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran and M. Kunt. Extraction of audio features specific to speech production for multimodal speaker detection. *To appear in IEEE Transactions on Multimedia*, 2007.

Conference Papers

- P. Besson and M. Kunt. Hypothesis testing as a performance evaluation method for multimodal speaker detection. In *2nd International Workshop on Biosignal Processing and Classification, ICINCO*, pp. 106-115, Setúbal, Portugal, 2006.
- P. Besson, J.-M. Vesin, V. Popovici and M. Kunt. Differential evolution applied to a multimodal information theoretic optimization problem. In *Proceedings of the 8th European Workshop on Evolutionary Computation in Image Analysis and Signal Processing (evoIASP), LNCS 3907*, pp. 505-509, Budapest, Hungary, 2006.
- P. Besson, M. Kunt, T. Butz and J.-P. Thiran. A multimodal approach to extract optimized audio features for speaker detection. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, 2005.

Technical Reports

- P. Besson, G. Monaci, P. Vandergheynst and M. Kunt. Experimental evaluation framework for speaker detection on the CUAVE database. TR-ITS-2006.003, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, January 2006.
- P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran and M. Kunt. Extraction of audio features specific to speech using information theory. TR-ITS-2005.018, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, July 2005.
- P. Besson and M. Kunt. Information theoretic optimization of audio features for multimodal speaker detection. EPFL-ITS Technical Report 08/2005, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, February 2005.

Master Thesis

- D. C. Carrasco, P. Besson and M. Kunt. Detection of the mouth on audio-visual sequences using binary partition trees. Master thesis École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, November 2006.
- A. Mayoue, P. Besson and M. Kunt. Détection du locuteur courant dans une séquence audio-visuelle. Master thesis École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, June 2005.

Studies

Since 2002	Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland Signal Processing Institute (ITS) of the School of Engineering (STI) Working towards a Ph.D. thesis (Docteur ès Sciences)
1998-2001	University Lyon I, France - University of Montreal, Canada Master of Sciences in Biomedical Engineering
1995-1998	University of Marne-la-Vallée, France Bachelor in Physics

Post-graduate Education

2005 July	Pattern Recognition: Computer Society Summer School on Pattern Recognition Plymouth, U.K.
2003 September	Tutorial on Joint Audio-Visual Signal Processing, International Conference on Image Processing (ICIP) Barcelona, Spain
2002	C++ programming by Central Informatics Service at EPFL

Languages

French:	mother tongue
English:	fluent
German:	intermediate
Italian:	basics

Computer Skills

Programming languages:	Matlab , C/C++, HTML
Operating Systems:	Linux, Unix, Windows