SAPIENZA
UNIVERSITÀ DI ROMA

DOTTORATO DI RICERCA IN INGEGNERIA INFORMATICA

XXIV CICLO – 2011

Top-Down Attention Modelling in a Cocktail Party Scenario

Letizia Marchegiani

SAPIENZA
UNIVERSITÀ DI ROMA

DOTTORATO DI RICERCA IN INGEGNERIA INFORMATICA

XXIV CICLO - 2011

Letizia Marchegiani

Top-Down Attention Modelling in a Cocktail Party Scenario

Thesis Committee

Prof. Fiora Pirri (Advisor)
Prof. Lars Kai Hansen
Prof. Francesco Quaglia

Reviewers

Prof. Bernard Gosselin
Prof. Ricardo Sanz

ISBN:

AUTHOR'S ADDRESS:
Letizia Marchegiani
Dipartimento di Informatica e Sistemistica "Antonio Ruberti"
"Sapienza",Università di Roma
Via Ariosto 25, I-00185 Roma, Italy.
****
IMM - Department of Informatics and Mathematical Modelling
DTU - Technical University of Denmark
Richard Petersens Plads, DK-2800 Lyngby, Denmark
E-MAIL: letizia.marchegiani@dis.uniroma1.it;
malet@imm.dtu.dk
http://www.dis.uniroma1.it/~marchegiani/

Ad Alessio
tenerezza infinita e ostinato orgoglio della mia giovinezza,
geloso custode di un indomabile ed irriverente spirito di giustizia...
perché il suo coraggio, la sua generosità
e la sua straordinaria purezza
sono stati il fine, il mezzo e l'opportunità
per tutta la vera bellezza che ha attraversato i miei giorni,
i miei pensieri e le mie convinzioni;
e perché l'inferno della sua delusione smetta di torturarlo
e la vita possa, finalmente, riconoscergli e concedergli
tutto quanto gli spetta.

Al mio piccolo e bellissimo angelo tutto nero,
perché lontano dalla prepotenza, dalla cattiveria,
dall'arroganza e dalla viltà degli uomini,
possa finalmente smettere di avere paura...
e dormicchiare in pace.

Acknowledgements

I would like to thank my supervisor, Prof. Fiora Pirri for having strongly exhorted me to start me on my way, for the passionate tutoring, the audacious scientific suggestions, for her assistance, and for having given me the chance to have my first teaching experience (which I think has been the most amusing and, at the same time, the most instructive thing I have ever done in my life! Thanks also to the students who made it so.) and in making this informal joint PhD real, leaving me the freedom to manage it. I would also like to thank my "unofficial supervisor", Prof. Lars Kai Hansen, for having enthusiastically accepted me into his group, for his always encouraging and cheering words (especially during our first meetings, in which, due to my great English and some fear, my way of coming across probably did not seem that reassuring to him, i guess!:D), for the precious supervision and the economical support which made me able to take part in the conferences where we presented our works.
A special thanks also to Prof. Bernard Gosselin and Prof. Ricardo Sanz, the reviewers of this thesis, and to Prof. Francesco Quaglia, my tutor, for their time and help and for the different points of view and perspectives they showed me: it was very useful to have a different approach to my work and improve it.
Thank you also to the rest of my working groups for their fundamental contribution, the stimulating discussions and the enjoyable meetings. Thanks, then, to Prof. Jan Larsen, Prof. Tobias Andersen, Matia and Seliz. Matia, the one who cleaned my desk on my first PhD day and who made my first months in the lab much easier than they might have been; Seliz, who was my free English teacher and translator (Seliz translate), my office mate, my neighbour and my partner in.. almost whatever I did (as we were always saying, "we are a great couple, even if we are not sure about the great part":D). She was, for sure, my lucky lottery ticket when I came to Denmark. Thanks for.. all the "shared things" she can "bring by herself": clever theories and genius plans are included, of course!
Thanks to Silvana for all her inestimable assistance during all these last three years, for her incredibly contagious energetic spirit, for her helpfulness and kindness: her door has always been open for me and she has always been smiling, giving me a warm welcome, even when I did not warn her before, even at the last minute, even when it was too late.

Thanks to Dimitri, my “from different parents twin”, for being a mirror to myself, for his imposing power as a really strict reviewer, for all our conversations and discussions around the world, for the thousands of suggestions (which I never followed! :D) and the millions of times I guffawed, for all the citations I did not get, for his time and help... and because I survived... literally speaking (“domani, che mo c’ho sonno!”). I am grateful and... happy. Actually, I am also sorry for all the times I made him pay, because I was not able to understand the mechanism for having a free call (chuckle).
Thanks to Matteo, my “arms mate” and my PhD information office... for his extremely useful advice about... everything, for listening and laughing to all my strange stories, for all his support at Dagstuhl and again.. for all the things I already listed and wrote to him (ho ritrovato il file per esserne certa.. ho fatto veramente la lista!:D) and the others he can imagine by himself. I am sure he will soon reach his “’beautiful place’.
Thanks to Massimo (and Valette, obviously), for reminding me of where I come from, for speaking my same real language and because sharing our weird experiences helped me keep my faith. In the end.. it seems we were right and they were wrong: stubbornness won! 8-)
Thanks to everyone who helped us with the experiments: all the participants and, in particular, Enis and Davide. The first one for having been our “Enispedia” in many occasions and for his code, which was crucial to make our job easier and quicker. Davide, because, even if by chance, was able to show where we made mistakes. Moreover, I need to thank him for having been, for a while, my bizarre Thursday mate, my confessor when I really needed to talk and my driver, parker and mechanic when it was necessary... and for a long list of funny things, events and people that crossed our way.
Thanks to everyone who helped me during my thesis writing period: above all to Laura (BI) and to the niño. Thanks to Laura for her great incurable optimism and encouragement, for being the only one who trusts my odd “social theories” (see “Pagafantas” and “little nerds in action” among others :D), for sharing a so nice illness, for the weird drinks (cfr. “CultMocho” and Japanese beer among others :D), and for having rescued me each time I needed to be. Thanks to the niño who even offered to read and review all my thesis the day before the deadline. After his plan to come to Prague, that was another big sign of his crazy and nice generosity.
Thanks to all IMM crew and to some honoured members, because they have been the best therapy ever in one of the hardest periods of my life. Yes, a great crew with a great Captain: thank you, club mate... for being one of last real Barcolosophers in the world, for remembering what I drink and what I do not, for having prayed with me for the coffee (that was incredible!), for your immense tolerance, meant in a wide sense, and because you like and enjoy all my so nice stories about my family and my queer relatives looking for water using a pendulum (uhm.. or maybe you don’t.. you are just nice and I have been torturing you for more than 1 year!!:D).

I also want to thank my family for their huge economical support, because it was not very easy for them and because living in Denmark, going to PhD schools would have been quite impossible without their intervention... and, of course, for the packages which each month make this house a “little Italy in Lyngby”, as Laura said.
Thanks to Stefania, for having told me about Denmark and helped when I moved, for the rice and the broccoli, of course, for having watched with me right until the end the best movie ever about the second world war and for my great sculpted abdominal muscles :D .
Thanks to Matteo and Giulia, two little wonders, for the awesome time we spend together and for being the living proof that my hopes for the future are definitely sensible.

But all my gratitude is to Alessio, for being the only one who, during the bad and good times, truly believed in me and in motivating me to carry on, who drove my car for three days to take me to Copenhagen in the middle of the craziest winter in the last 20 years, and who always did whatever he could, and sometimes even more, to help me or just to make things easier for me, sacrificing himself. And the best is that I never needed to ask for anything.
This PhD, as well as my Master and my Bachelor, is a common step, because, without him, it could have been just... just... a dream.

Thanks to my sweet little brother and to my beautiful grandmother for existing.

Contents

List of Figures

List of Tables

This thesis is the result of an informal joint Ph.D. between Sapienza, University of Rome and DTU, Technical University of Denmark. I spent the first half of my Ph.D. period in Italy, working under the supervision of Prof. Fiora Pirri and the other half in Denmark, working under the supervision of Prof. Lars Kai Hansen.
Unfortunately, because of Danish legislation at the time I started, it was not possible to have the formal agreement between the two universities.

Introduction

Computational auditory scene analysis (CASA) focuses on the problem of building machines able to understand and interpret complex acoustic scenarios and react, after a brief period, in an opportune way. A complex acoustic scenario can be characterized by several sounds of various origin and nature, coming from different sources. One of the main challenges is the contemporaneous elaboration of all this information with limited computational resources. Consequently, a preliminary selection between all the various signals, able to reduce the amount of data to manage simultaneously, is crucial.
Colin Cherry (1953) investigated human behaviour in the same circumstances, which he called "*cocktail party problem*", drawing inspiration from the human ability to pay attention, at a cocktail party, where there are various voices and noises, to the speech of the neighbour, ignoring the rest. He performed several experiments proving that this human ability is due to the use of attentive mechanisms. Attentive mechanisms, in fact, operating as a filter between all the incoming stimuli, allow the brain to focus on signals that are necessary to follow and ignore others that can be discarded (*selective attention*).
Therefore, the implementation of a computational model, that is able to mimic human attentive processes operating in these situations, represents an interesting and innovative approach to the analysis of different and variegated acoustic scenes and, specifically, to finding a solution to the so called "cocktail party problem".

Many applications could benefit from the use of such a model. Surveillance systems could select the interesting stimuli according to the perception of potential dangers associated with sounds in the scene and, thus, could address microphones and cameras in the direction of these stimuli, focussing more on these (see, e.g. López *et al.* (2006)). In this way, it could be possible to have more details and more knowledge about salient parts of the scene, instead of wasting resources paying attention to the useless signals. In the tracking and transcription of multi-speaker conferences, meetings, seminars, being able to recognize the speaker and identify only what he/she is saying, ignoring other sounds or other voices around is a fundamental condition for allowing a speech recognition procedure to work correctly. In Cognitive Robotics, modelling human attention can drive the development of spoken dialogue systems

and help in approaching human-robot interaction issues (see, e.g. Lang *et al.* (2003) and Marchegiani *et al.* (2009a)).

State of the Art

According to human attention studies, the identification of "interesting" sounds can be driven by many factors; depending on the nature of these factors, it is possible to distinguish between a bottom-up and top-down perspective. In the first case, sounds of interest are those which stand out from the scene, without involving a real attentive processing, but only a pre-attentive one. In the second case, the goal, a task, previous knowledge, acquired predispositions, emotions and motivations guide the subjects' attention. The combination of these two approaches indicates to the brain what is salient and what can be ignored.
Different models have been proposed in literature to analyse both these perspectives (see, e.g. Frintrop *et al.* (2010) and Itti and Koch (2001) for a review about visual attention). While visual attention has been deeply investigated as bottom-up and top-down, very few works have been carried out so far to implement top-down auditory attention (see, e.g. Kalinli and Narayanan (2008)). However, these works are largely domain and representation dependent and they do not provide a general model which could be valid and applicable in several circumstances.

Moreover, an analysis of the influence of top-down cues on human auditory perception is still missing. Aspects like the presence of a task, previous knowledge, emotional states or acquired predispositions have been not widely investigated yet. Nevertheless, those represent important components in the development of any attentive model which aims at extensively imitating human behaviour to perform various tasks in an efficient and effective way.

Aim and Contributions

The main goal of this work is to model human auditory top-down attention in a cocktail party scenario and investigate the role of some of the cues involved in the attentive selection procedure.

We proposed a generative model, representing top-down attention as a sequential decision making process, driven by a task. In particular, the decision process is implemented in a Gaussian mixture model and tested in a classification problem with missing information. The model has access just to a random subset of features (the so called *gist* of the scene), but there is the possibility to gather additional features among the ones that are not available.
The saliency of the missing features is given as the estimated difference in classi-

fication confusion (entropy) with and without the given feature. The difference in confusion is computed conditioned on the available set of features.
Thus, the top-down attentive procedure is reduced to find the answer to the question: *what to measure next to improve the decision process?*

We make our attention model able to operate in more realistic and complex scenarios by also allowing the initial training phase to take place with incomplete data. The missing data problem, in fact, is a really common issue in many applications. We investigate the behaviour of some missing data techniques, giving a new definition of efficiency and comparing the methods accordingly. Then, the algorithm which shows to be the most efficient is used for the training of our attention model.

Moreover, we test the attentive system proposed in a cocktail party simulation. We analyse its performance, mimicking the action of different levels of attention, in a classification problem in which the training phase is characterized by the presence of counfounders.
We confirm the results obtained thanks to this simulation, carrying out behavioural experiments, inspired by the ones of Cherry (1953). The aim is to investigate the role of a task in the distribution of attention in the cocktail party. To do so, we examine the effect of temporal and spectral overlaps for human speech intelligibility, and how it is influenced by the presence of the task.
We make subjects listen to a monaural mixture of two different narratives uttered by a speech synthesizer using the same virtual speaker, so that cues related to the voices or to the spatial position of the sound sources could be reduced, as also suggested by Cherry. Our intent is to check the ability of the subjects to hear the narratives, in spite of the overlaps. In order to do so, we present them a list of words and ask for selecting the ones they heard (some of the words in the list are really present in the narratives, others are not, but they are related to the content).

We also carry out initial studies about the role of priming in a cocktail party problem. We check if having some information about the content of what people are hearing could improve the segregation procedure between all the signals and attention could be more attracted by the speech containing a particular known topic (pre-defined or not), rather than another.
We perform some preliminary experiments to examine, then, the influence of priming on attention. The design of the experiments is close to the one of the experiments about the effect of a task on human speech intelligibility we have just described. But, in this case, before making the subjects listen to the narratives combination, we give them some indications about the content of one of the stories. The initial results we got seem to confirm that a known topic grabs human attention more.

With the aim of generating biologically plausible perception models, we also in-

vestigate the multimodal combination of stimuli of a different nature. We implement a multimodal speaker recognition procedure in a conversation scenario. A robot actively follows the conversation, localizing its current interlocutor, turning towards him/her and estimating his/her identity thanks to a semi parametric model, which combines visual and auditory features of the speakers.

Summary

This work is organized in four parts.

Part I : A general introduction about attentive mechanisms is given and computational approaches proposed over the years to model these mechanisms are discussed;

Part II : A different approach to top-down attention model as an *active decision* process is proposed and its performance is evaluated. An analysis of the efficiency of missing data techniques is carried out and the method providing the highest efficiency is used to test the top-down attention model in case of incompleteness in the data exploited to train the model itself;

Part III : A remake of the experiments performed by Cherry almost sixty years ago is described and the relative results are discussed. The effect of top-down cues like the presence of a task or priming on speech intelligibility is analysed. Moreover, we compare human conduct with the behaviour shown by the top-down attention model (introduced in the previous Part) in a cocktail party simulation, in the case of a present task;

Part IV : A human-robot interaction scenario is described in which the machine is able to follow a conversation involving different speakers and in which the current robot's interlocutor is recognized combining visual and auditory features.

Part I

Background and Related Works

Chapter 1

Human attentive mechanisms

In this Chapter, we mention some of the definitions of *attention* proposed over time and analyse the various aspects involved in human attentive mechanisms, citing and illustrating experiments and studies performed in different fields over years.

1.1 The cognitive process called *attention*

What does attention mean? The term "attention" was introduced in Roman times and several different interpretations have been proposed in the last centuries to explain the characteristics of this cognitive process. Malebranche (1764) provided the first thorough work, believing that "*attention is necessary for conserving evidence in our knowledge*" and suggesting methods to develop it, like the study of geometry. Locke (1689) described attention as the processing of ideas that are already in the memory:"*when the ideas that offer themselves (for, as I have observed in another place, whilst we are awake, there will always be a train of ideas succeeding one another in our minds) are taken notice of, and, as it were, registered in the memory, it is attention*". Wolff (1732), some years later, was the first one to dedicate an entire chapter of his textbook to attention. In this way, finally, attention started to be considered a new phenomenon needing its independent theory. Consequently, various hypothesis about its nature and the processes involved began to be studied and proposed.
James (1890) gave a definition of attention, close to its everyday meaning:"*Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought*". Since that time, the concept of attention has been investigated in many fields, using different instruments and many interpretative models have been presented. However, a definition on which everyone agrees and a synthesis able to elucidate all the mechanisms involved and their relations is still missing, as Pashler (1999) noticed more than one century after James:"*no one knows what attention is, and that there may even not be an "it" there to be known about (although of course there might*

be)". For a more complete review about history of attention, see Itti *et al.* (2005).

1.2 Attention as a selection process

"To behave adaptively in a complex world, an animal must select, from the wealth of information available to it, the information that is most relevant at any point in time. This information is then evaluated in working memory, where it can be analyzed in detail, decisions about that information can be made, and plans for action can be elaborated. The mechanisms of attention are responsible for selecting the information that gains access to working memory." (Knudsen (2007)). Attention theories often disagree about the nature of this selection. The most part of them affirms that this selection takes place because of the limited capacity of the brain and its consequent inability to analyse multiple complex perceptions. Others claim that the selection is necessary to handle several simultaneous thoughts.

Some of the earliest experiments about the selective mechanisms acting during an attentive process were performed by Cherry (1953). He studied the so-called *Cocktail Party Effect*: the human ability to pay attention, in a cocktail party, where there are various voices and noises, to the speech of the neighbour, ignoring the other sounds around. In particular, he performed two different kinds of tests. In the first set of experiments, subjects were made to listen to a monaural mixture of two different narratives uttered by the same speaker and asked to repeat what they were hearing (*shadowing*), to verify if they were able to separate the two streams. In the second set, instead, two distinct messages were played to different ears of the subjects (dichotic listening). The subjects selected one ear, keeping in mind and understanding completely the relative message, while the second one was neglected: the only information that the subject noted was relative to some physical characteristics like the gender of the voice, high or low tones, speech or noise; they did not comprehend the meaning and they were not able to notice if the speaker had changed the language.
Broadbent (1958) explained the results obtained by Cherry claiming that humans can not elaborate more than one stimulus at the same time. Consequently, he hypothesized the existence of a *premature filter*, operating before any kind of semantic interpretation takes place. The simultaneous signals are put in a sensorial buffer before being analysed and they overcome the filter one by one to avoid overloading in the analysis. The order in the buffer is established on the basis of the physical characteristics of the signals, like tone of voice, pitch, spatial location. However, the signal hanging on, stays in the buffer just for a short time and, because of the structural limitations of the system beyond the filter (the so called *short-term memory*), it will be analysed only in a superficial way. The *short-term memory*, in fact, is a memory able to register just a small amount of information and just for a really short time.
This theory is somehow consistent with what Panum (1858) claimed a long time

Left Ear	Right Ear
mice	3
5	eat
cheese	4

Left Ear	Right Ear
2	who
goes	3
9	there

Table 1.1: Two different examples of the experiments performed by Gray and Wedderburn (1960). The words in the same row are played at the same time. The general temporal order respects the order of the rows.

before. He was one of the first researchers to investigate the idea of attention as a selector within different signals in competition. He believed that if one of these signals takes enough consideration, then the others could not. So, if we are concentrated on a weak signal and a stronger one comes, it is hard to still be focussed on the first one. Hamilton (1859) criticized this view excluding that just one signal could grab all the attention; he rather proposed to examine how many stimuli can be processed at the same time as the key issue.
Following tests supported this theory. It was possible to demonstrate, in fact, that subjects could catch some words from the other stimulus: for example, they overheard their name, as Moray (1959) proved. Moreover, the more they were trained on shadowing, the better they could perform a good elaboration of the content of the other channel as well. This shows that a sort of semantic analysis is executed even before the filter, suggested by Broadbent, could operate. Gray and Wedderburn (1960) and MacKay (1973) carried out experiments to investigate this phenomenon, finding that some words apparently neglected could influence the interpretation of ambiguous messages. Gray and Wedderburn made people listen to a stereo signal, putting in both channels a mixture of digits and words. The words, if joined altogether, could give a meaningful sentence (see Table 1.1 for some examples). One group of people knew only that they were going to hear a list of digits and words; the other one was aware there could be sentences in the audio. In both cases, they prefered grouping the signals by meaning than by ear (the second group gave the best performance in this sense). Also MacKay made subjects listen to a stereo signal, but, for each trial, he put in one channel an ambiguous sentence and in the other one a particular word. The subjects assumed they had not heard those particular words; instead, it seems they were using them to interpret the sentence they were listening to in the other ear. For example, one of the sentences was “They threw stones at the bank”, in which “bank” could mean the credit institution or the shore of a lake or so. The word in the other channel, instead, could be “’river’ or “money”. Accordingly to which one of them they heard, they gave a different meaning to “bank” and, consequently, to the entire sentence.
Treisman (1960) tried to justify the same phenomenon, keeping the idea of an *early*

selection of the stimuli and hypothesizing the existence of a *delayed attenuate filter* which does not eliminate the information of the ignored message, but solely attenuates and delays it. The filter performs a hierarchical process: from the analysis of the physical features of the signal perceived to the elaboration of its grammatical structure and of the meaning of its content. Also in this case, not necessarily all the stimuli are totally examined. The most part of the resources are given to one channel, but a smaller part of them is still dedicated to the other one. This is why, Treisman believed, the signal in the supposed neglected ear could be, in some occasions, perceived.

In complete opposition to Broadbent, Deutsch and Deutsch (1963) asserted that all the stimuli were analysed at a high level and semantically: a sort of *response limitations*, instead of the *percpetual limitations*, suggested by *early selection* theories. Norman (1969, 1976), recovered the results of MacKay, Gray and Wedderburn, suggesting that all the acquired stimuli are analysed on the base of their relevance and pertinence. Thus, there is a deep semantic elaboration in any case and the selection take places just afterwards *(late selection)*.

The interpretation of Johnston and Heinz (1978, 1979) and Johnston and Wilson (1980) can be considered a sort of compromise between the idea of a premature attentive filter and a delayed one. They proposed the *flexible filter theory* which admits the possibility that the selection among various stimuli could happen through different elaboration steps, on the basis of the circumstances and the specific tasks. In this way the congestions are not managed by a unique filter and a signal can be processed at a lower level or at a higher one according to the situation. Of course, a premature analysis and selection is less complicated and wasteful than a delayed one, which has to deal with semantic elaborations: thus, the filter tries to operate as early as possible, in order to optimize the computation time and the resources to use. Before elaborating this theory, Johnston and Wilson (1980) performed experiments similar to the ones MacKay did, making subjects listen to a series of word pairs, asking them to detect the words that could be considered instances of a particular category. These words were called *Homonymic Target*, the others are called *Nontarget words*. The nontarget ones could be somehow related to the category, neutral or even misleading respect to the meaning of the category (some examples of the words used by Johnston and Wilson for the category "animals" is shown in Table 1.2). They proved that the Nontarget words influence the ability to detect the target words and this influence changes according to the type of Nontarget: appropriate, neutral or inappropriate. This suggetsts that there is a semantic processing of the supposed ignored stimili (as Mackay showed as well), and the amount of this processing is not fixed, but changes according to the circumstances (different types of Nontarget).

Neuroimaging methodologies studies seem to support the *flexible filter theory*, as Luck and Hillyard (1999) explained. There is not, then, a unique bottleneck operating as a filter for all the incoming stimuli, but various filters working at different levels on the base of the situations and the task. In particular, they demonstrated that the sup-

Target Category	Homonymic Target	Type of Nontarget		
		Appropriate	Neutral	Inappropriate
Animals	duck	quacking	movie	dodge
	ant	crawling	straw	uncle
	swallow	nesting	spoon	drink
	turkey	gobble	ash	india
	fly	buzz	wax	pilot
	badger	caged	violin	harass
	bear	hibernate	luck	naked
	steer	roundup	jam	guide
	deer	antler	amber	sweetheart

Table 1.2: Some examples of the word pairs used in the experiments of Johnston and Wilson (1980).

posed non attended information is not thrown out of the system, but just elaborated in different regions of the brain. Ruz *et al.* (2005) carried out an fMRI experiment to investigate brain activations in two different attentional conditions. Overlapping drawings and sequences of letters were shown to the subjects: some of these strings were real words, some did not have any meaning. The subjects were asked to direct their attention to either the drawings or the letters. The results suggested the use of different pathways relative to word processing, depending on the focus of attention, and that these different pathways actually generated different behavioural and neural responses.
These new ways of exploring selective attentive mechanisms and the findings they had provided changed completely the approach to this kind of issue, closing the dispute between the early and late selection theories.

Visual selective attentive theories followed an analogue evolution. Norman (1968) introduced the metaphor of visual attention as a spotlight in a dark environment, which illuminates the intended stimulus, leaving the rest in some kind of dimness. James (1890) provided a view of visual attention as a mechanism characterized by a focus, a margin and a fringe (see Figure 1.1). Resources are mainly concentrated on the focus and this concentration (and, consequently, the clearness with which it is possible to see) gets lower in a inverse proportion to the distance from the focus. Posner *et al.* (1980), thanks to their experiments, demonstrated that actually attention seems to have a central focus and the dimension of it depends on the task and the particular situation. Attention is generally correlated to eye movement (*overt attention*), but this is not necessarily true. They also proved, in fact, that it is possible to attend to something of interest without any eye movement (*covert attention*), which is much slower than a change in the focus. LaBerge (1983) confirmed these results

and he also showed that there is, in any case, a processing of stimuli out of the focus, even if minimal. But, unlike what seems to happen in the auditory processing, in this case, there is not a semantic analysis of the neglected signals. Johnston and Dark (1986) performed some experiments making subjects identify words on a screen. The words were blurred in the beginning and became clearer in the end. They put similar words or words with a close meaning to the ones to identify in other parts of the screen, (which were supposed to be unattended) and they noticed that these ones did not influence at all the identification of the test ones. Johnston and Dark (1986) proposed, then, the *zoom-lens* model: like the zoom of a camera, when we want to pay a particular attention to something, we can see more details about it (we zoom in on it). But human resources for doing it are limited, so we can focus on little things having many details about them or we can look at big things, like an entire landscape for example, with less details and precision. In this way, the total amount of information we are actually processing can still be in the physical existing limit. Both *spotlight* and *zoom-lens* theories support, however, the idea of attention as a *early selection* filter operating on the physical features of the signals (Laarni (1999)). Treisman and Gelade (1980) proposed the *Feature Integration Theory*, which sug-

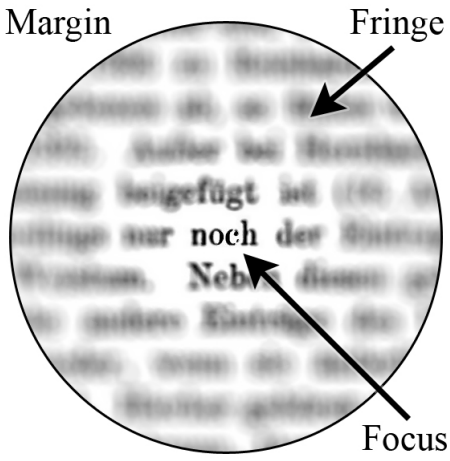


Figure 1.1: Illustration of the description of visual attention as a *spotlight*, as given by James (1890), who suggested attention as having a focus, a margin and a fringe.

gested that the basic features like color, orientation, depth, curvature and motion were elaborated in a preliminary parallel way and, only afterwards, these features were integrated. *Late selection* theories, instead, were investigated and proposed by Duncan and Humphreys (1989, 1992) and Duncan *et al.* (1997) among others, who hypothesized a parallel processing of the incoming signals without any kind of filter and a selection executed only after a semantic analysis. Each stimulus is totally examined, but just the attended ones are stored in the working memory.

Neurophysiological studies showed that the locus of visual selection seems to depend on the amount of perceptual load. Lavie (1995) demonstrated that attentive filtering takes place only if the perceptual load is so high to make a selection necessary, as claimed by Luck and Ford (1998): "*there is no point in suppressing the identification of irrelevant objects unless the visual system is so overloaded that the irrelevant objects interfere with the identification of relevant objects.*" This suggests that both kinds of selection are possible in different situations.

1.3 Divided Attention

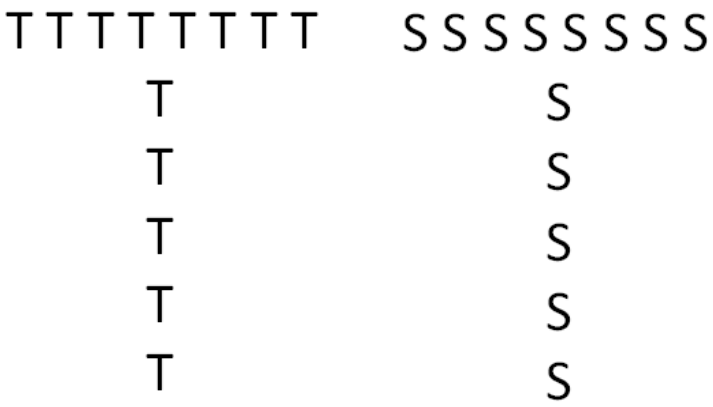
The collocation of the bottleneck in the processing path of the incoming stimuli is the main point in the analysis of selective attention. However, it does not give almost any kind of information about its so called *energetic side* of attention, which regards the way in which the limited resources of the brain are distributed among different tasks involving attention. The analysis of the *energetic side* can be considered another school of thought, exposed by Kahneman (1973) and Navon and Gopher (1979), dealing with attention not as a filter, but as a resource with limited capacity, shared by the stimuli, in proportion to their necessities.
Some of the first contributions in this direction were given by Spelke *et al.* (1976) and Hirst *et al.* (1980). In their first work, Hirst et al. carried out experiments in which the subjects were trained to execute two different tasks at the same time to check their attentive behaviour in this kind of situation. In particular, they were asked to read a passage and, simultaneously, to write down words someone was dictating to them. The subjects in the beginning found it really difficult, but their performances improved considerably in time. In the second work, instead of dictating words to the subjects, they used sentences, asking the subjects to recognize these sentences afterwards. In this way, they could understand if the subjects were really paying attention to the meaning of the sentences or if they were just mechanically writing these down. The results showed that actually both assignments were executed using attention. Thus, it seems that attention can be divided among different tasks and human ability in doing so gets better with practice. But practice is not the only factor influencing human performance in accomplishing more than one job at the same time: the nature of the jobs themselves has a huge impact on the way of facing this kind of interference. Generally, people can talk while driving, but are not able to read a book, for example. Of course, if they have just started learning to drive, even having a conversation could be hard, whereas after they get used to it, the resources necessary to drive will be much less and both tasks could be carried out contemporaneously without any effort. The difference between the two couples of simultaneous operations: $\{driving, talking\}$ and $\{driving, reading\ a\ book\}$ is that in the first case, the interference is, essentially, a resources matter; in the second one, it is structural, so it can not be removed: even being the best and the most expert driver, looking

at the street is still necessary, so there is no possibility to look at the book. This is because, the two tasks involve the same channel, being really similar, so they can not be parallelized. Instead, a visual task like checking the street and an auditory one like talking or listening to music are quite different. So, if they do not require too many resources, (which, as we said above, can depend on the experience and or on the previous knowledge or on aurosal), these tasks can be carried out easily enough. Kahneman (1973) was one of the first researchers analysing this aspect. He provided the concept of restricted resources using the metaphor of a tank, to indicate the finite capacity of the means that humans can use. He differentiated tasks relative to the mental effort they require to be performed, without a direct correlation to the information load. Wickens (1980) improved this theory, introducing the concept of structural interference and showing, that, even if two tasks are generally easy to be achieved separately without demanding a big amount of computation resources, if they need the same kind of resources, it could be not possible to work such an interference out. Even automatic processes can create this kind of interference. Stroop (1992) made subjects see a screen like the one in Figure 1.2 and ask them to do two different things: to say the color of each line and to read the same lines. The first task was accomplished with much more difficulty than the second one (this effect was later called *Stroop Effect*). The reason is that, reading is such a natural action for people (if able to do it, of course) that is automatically executed, even if the task does not require so. Consequently, interference is created. The same paradigm of this experiment was also used to test the validity of *late selection* theories, using the lower reaction of the responses of the subjects, because of interference, as the demonstration that unattended stimuli were also processed.

A similar result, known as *Navon effect*, was obtained by Navon (1977), who made

Figure 1.2: Example of the screen Stroop (1992) shown to subjects for some of his experiments about divided attention.

subjects look at larger characters made by small ones. The larger and the small character could be the same or different (such as in Figure 1.3). Subjects were asked to identify just the small ones or just the larger ones. Navon noticed that, in case of *incongruent characters*, they spent more time in the identification of the small letters, while the identification of the larger ones was not affected. Again the incongruence created and automatic interference and, consequently, a division of attentive resources. Thus, the *energetic side* of attention can be considered



(a) Congruent Characters (b) Incongruent Characters

Figure 1.3: Example of the screen Navon (1977) shown to subjects for some of his experiments about divided attention.

1.4 Attention and consciousness

Wolff (1734) and Leibniz (1765) elaborated, for the first time, a connection between consciousness and attention, which was further investigated by Wundt (1874). Wundt claimed that ideas are present in the consciousness on different levels, depending on attention and the ability of being concentrated on something increases or decreases on the basis of these levels of consciousness. Attentive mechanisms deal with all the information which has to be cognitively managed, while the consciousness deals with just the part of that information we are aware of. In opposition to Locke, Kames (1732) claimed:"*Attention is that state of mind which prepares one to receive impressions. According to the degree of attention objects make a strong or weak impression. Attention is requisite even to the simple act of seeing*". Some of the descriptions proposed were supported by primal behavioural experiments, like the ones carried out by Von Tschisch (1885) and later by Angell and Pierce (1892) to investigate how different simultaneous stimuli are perceived and analysed in the consciousness.
Shiffrin and Schneider (1977) postulated the existence of automatic and controlled

processes: in particular, they showed the qualitative difference between two modalities of information processing: *controlled search* and *automatic detection*. Shiffrin and Schneider supposed that automatic processes were really quick, because, even if they could attract attention, they were executed without needing an attentive effort or a precise intention. Moreover, they could not be controlled by the subject and they did not use resources like short term memory: thus they are particularly efficient. The controlled ones, instead, were slower, because based on the consciousness, on the will to reach a particular aim and, consequently, demanding planning and control and resources. On the other hand, these processes are extremely flexible compared to the automatic ones, which are rigid and not able to adapt to new conditions or changes in the operative environment. However, controlled processes could become automatic through practice: as we said before about the *Stroop effect*, reading becomes in time so natural as to be automatic even when we do not want to do it, because we are trying to use our attention to do something else. A similar pattern takes place when we learn a new language: after enough time and experience, we start thinking directly in the new language, instead of preliminarily thinking in our mother language before and then translating everything in the new one. While the process is becoming automatic, it requires less and less resources and attention, which can be used to perform other tasks in parallel. This is why, as we said before about divided attention, when someone starts learning to drive, it is hard do to other things at the same time; after a while, the practice is enough and driving becomes automatic: there is no need anymore to think about what to do to make the car move, to stop and so on, thus other tasks can be executed simultaneously.

Norman and Shallice (1986) and Shallice (1988) proposed a new model which assumes that actions could be controlled in two different ways. The first way takes place in case of consolidates activities, that, thanks to learning, becomes automatic in time. These kinds of activities can be executed in parallel without the generation of so much interference. But, if a conflict is created, it will be necessary to define which activity has the priority to be executed. The priority is decided according to a sort of catalogue of decisions, which defines the rules to determine the importance of each activity. The second way to control actions, according to Norman and Shallice, is a system they called *Supervisory Attentional System (SAS)*. This system is able to interrupt, voluntarily, some of the activities, thus, it provides a flexible reaction to new conditions.

Recent experiments investigated the crucial role of attention on the registration of perceptual objects and events in consciousness. The so called *Change Blindness*, introduced first by McConkie and Currie (1996), is just an example of it. The results were quite similar and showed that, in particular circumstances, very large changes in a picture could be ignored by the observers, not able to notice these changes at all. Generally, these modifications are made simultaneously with other things able to attract the attention of the observers and to distract them from the changes. But, as Itti (2003) claimed, “this does not necessarily mean that there is no vision other

than through the attention bottleneck". Some experiments (*dual-task psychophysical experiments*), in fact, have been performed in which subjects were able to identify two different objects at two distant and not close positions (see Lee *et al.* (1999)). In conclusion, it is just possible to say is that attention influences the accuracy of the observations, and that, as Koch (2004) wrote, only stimuli that are voluntarily reported, can be considered "attended".

1.4.1 Bottom-up versus Top-Down

As we mentioned many times so far, humans have limited cognitive resources; attention is the process which allows answering quickly to complex stimuli using these circumscribed means, selecting the most salient ones and suggesting to the brain what to ignore. The selection is influenced by different factors operating at the same time and at different levels: in relation to these factors, it is possible to distinguish between *bottom-up* and *top-down* filters. According to a bottom-up perspective, stimuli of interest are those which do not involve a real attentive processing, but simply attract subjects' attention during a fast pre-attentive step, independently of a particular task or aim. Generally, a bottom-up approach considers salient stimuli which are not so habitual or which, somehow, break the balance of the environment: like a siren coming close or a flash of a light or a red object in a totally green background. These kinds of signals stand out from the scene, because of the characteristics of the brain's receptive fields. But there also signal which become salient in time, such as the voice of a parent, for example. According to a top-down perspective, instead, the goal, the previous decisions, emotions, desires, knowledge about the environment where to operate or about the target to look for and the acquired models drive the subjects' attention. As Connor *et al.* (2004) claimed, in their dispatch about visual attention: "*Top-down mechanisms implement our longer-term cognitive strategies*". The fusion of these two modalities of analysis suggests to the brain what is salient and what can be attenuated or deleted. The bottom-up approach catches what could be potentially important, in accordance with instinctive human reactions; the top-down filter which adjusts the results considering tasks and objectives. They somehow cooperate: the top-down attention can give a different weight to the saliency suggested by the bottom-up approach, in order to execute a task, but, still, the same saliency can change the focus of attention suddenly thanks to particular stimuli. The experiments of Moray (1959) showed this: when subjects heard their name, they immediately switch the focus, without considering a task, eventually assigned.
Many studies over years, proposed the idea of this attention duality. Nakayama and Mackeben (1989) demonstrated the existence of, at least, two different attentional mechanisms: one operating in an automatic way and without any kind of control, which is fast and momentary; the other which, instead, works slowly and involves the voluntary control of the subject. Same results were shown by Braun (1994, 1998) and Braun and Sagi (1990), who presented the idea of a *binary* nature of attention,

using the concept of a bottom and and a top-down attentive filters operating simultaneously: the first one based on the concept of saliency; the second one linked to voluntary control to reach some aim. More recently, in their investigation about the way in which attention is actually reflected in the activity of the brain, Baluch and Itti (2011) claimed that:"*Although the exact mechanisms of top-down attention have yet to be completely delineated, there are sufficient data available to demonstrate that attention is mediated by the merging of top-down and bottom-up information.*".

1.5 Units of Attention

A long debate took place, in the last decades, about the nature of units of visual attention. Some theories suggest that humans attend to features (*feature-based attention*), some that humans attend to spatial locations (*location-based attention*), some others to objects (*objects-based attention*). Treisman and Gelade (1980) supported the idea of *feature-based attention*, with their *Features Integration Theory*; more recently, the same school of thought was sustained by Liu *et al.* (2003), among others. *Location-based attention*, was, instead, exposed by Posner (1980), Eriksen and St. James (1986) with their *zoom-lens model* and, in the last years, by Bisley and Goldberg (2003). More recent theories support the idea of attention as a selection within *perceptual objects*. Nowadays, the dominant opinion about visual attention is that the different schools of thoughts (features,location,object-based attention) could be combined, because the units of attention can vary according to the situation. Arrington *et al.* (2000) gave an example of this, proposing a theory, which tries to link together objects-based and location-based approaches, according to which, a region of space attracts attention without the presence of eventual objects could influence somehow the choice about the region.
It is quite easy to understand what a visual perceptual object is: edges, other geometric characteristics, colour and textures can be linked to form objects in a scene, as proposed by Feldman (2003). In acoustic context, instead, it is harder to define precisely what acoustic objects are, but, generally, an audio stream coming from a physical source can be considered an object. Through a sort of clustering process, the audio samples which present the highest similarity can be grouped in the same cluster; the iteration of this procedure over time provides the formation of objects. Human auditory attention, in fact, seems to be composed of three different steps: a preliminary process of acoustic objects formation, the organization of the sequential sounds into auditory streams and a successive process of selection among these, as shown by Carlyon (2004) and Shinn-Cunningham (2008) among others. Several physiological studies, supported by brain imaging investigations, such as the ones of Mondor *et al.* (1998), Sussman *et al.* (1999), and Zatorre *et al.* (1999), were the basis of these intuitions, proving that attention is directed at streams and that the different signals are grouped to form different perceptual elements according to both low-level

and high-level analysis. This means humans use not only physical characteristics like frequency, location, pitch, timbre, (low-level examination) but also learned concepts like the word identity, the grammar structure, semantics (high-level examination). The identification of each stream within the complex sound received, then, depends on the frequency, the relative spectral features and the spatial location of the different sound sources.
For a complete review about units of attention, see Yantis (2000).

Chapter 2

Computational Modelling of Attention

In this Chapter, we expose the different computational approaches proposed over years to imitate human attentive behaviour. Bottom-up and top-down models are shown and some possible applications of the so called *machine attention* are presented.

2.1 Motivations

The possibility to reduce the amount of data to manage simultaneously is crucial also in many computational applications. Implementing attentive mechanisms on a machine allows the machine to be able to analyse the environment around and to react, quickly, in an opportune way. Computer vision and computational auditory scene analysis (CASA) would benefit very much from a filter discarding the pixels or the sounds which should not necessarily be taken into account, sparing resources and time. Frintrop *et al.* (2010) in their review listed some of possible applications of computational visual attention systems. They explained, for example, how saliency could be useful for image segmentation, compression or matching, as shown by Itti (2004) and Ouerhani (2003). In the first case, the saliency could be used to determine the starting point of the process; in the second one, the general procedure consists of giving more importance to the salient parts, using a smaller compression factor for them and a greater one for the others. In the third case, instead of comparing completely the images, it is possible to check only the salient regions and to make decision on the basis of this.
Also for robotics applications, the implementation of attentive mechanisms could be convenient and advantageous. Frintrop *et al.* (2010) mentioned, for example, how it is possible to exploit attention to find landmarks in a scene. These landmarks, then, can be used for localization, either having an initial map of the environment or needing

to build it (*simultaneous localization and mapping (SLAM)*). Similar considerations can be applied to object recognition problems, both in visual and auditory fields. Being able to discriminate in a complex scenario between the different audio streams and to focus and analyse just one of them is fundamental in applications involving transcriptions of conversations, multi-speaker conferences, meetings, seminars. The automatic speech recognition procedure needs to understand what is necessary to analyse and what can be ignored: it necessitates to isolate the voice of interest within the other sounds and voices around and to track it, to be able to recognize the words pronounced (Choi *et al.* (2002)). In this perspective, human-robot interaction can be considered, as well, an important area in which attention systems could be employed: robots following conversations, robots exchanging information with humans, trying to reach some aim and so on.

2.2 Bottom-up approaches

The concept of saliency is strongly related to bottom-up attentive mechanisms. According to Oxford Dictionary, in fact, *saliency* means prominence, conspicuousness and, as we already said in section 1.4, in a bottom-up approach, what pops out of the scene is considered interesting. Thus, these approaches use saliency as a measure to identify what have to be considered in the analysis and what can be ignored, building, generally a so-called *saliency map*, indicating for each "point" in the environment its level of "importance". The difference between them consists, basically, in the way of estimating saliency.
The allocation of the bottom up attention in different sensory systems involves similar mechanisms. For this reason, even if the features used to assign the salience values are very different, similar models have been developed to obtain visual and auditory saliency maps. The classical visual models work on spatial images only, while the auditory models consider also the temporal domain; still the maps obtained are structurally equal.

The most part of the computational models which have been proposed till now are based on the Feature Integration Theory of Treisman and Gelade (1980), asserting that some features are parallel processed in some kind of pre-attentive step, while combinations of features have to be searched for in a serial way. This procedure generates separate maps, which are, then, fused to obtain a unique final saliency map. Thus, the choice about which features to be selected for generating those maps becomes particularly important.
The model of Koch and Ullman (1985) can be considered the first visual model in this direction, regarding visual attention. It is essentially based on a two-dimension topographical saliency map, obtained as a combination of feature maps, computed according to the sensitivity to colours, intensity, orientations, motion and so on. The

most salient region is selected using a *Winner Takes All (WTA) network*, introduced by Feldman and Ballard (1982). The dynamic evolution of saliency over time is assured by a mechanism of Inhibition Of Return (IOR), which is crucial to avoid the attentive system to be focussed on the same region perpetually, as later pointed out by Itti and Koch (2001). This mechanism decreases the salience of the current attended target, updating the map and allowing, so, the focus of attention to shift to the next most salient target. The WTA network was criticised by Tsotsos (1993), who proposed a new version of it, which admits the possibility of having multiple winners, using an inhibitory beam. This net was afterwards used by Tsostos and colleagues in Tsotsos *et al.* (1995) in their *selective tuning (ST)* model, which exploites luminance, orientation as some of the discriminative features. Later, in Tsotsos *et al.* (2005), motion was also considered a feature. An other bottom-up model was presented by Milanese (1993), who used colour, orientation and edge magnitude as a feature to compute the feature maps and a relaxation rule to generate the final saliency map. Wolfe (1994a) proposed a list of so called *guiding attribute* that can be used to direct attention. In particular he believed that there is an independent map for each kind of feature and that, in each of this map, then, the feature can be separated in its components.
Itti *et al.* (1998) and Itti and Koch (2001) developed and improved the original model of Koch and Ullman (1985), introducing an easier and faster way, compared to the relaxation rule of Milanese, to combine the single maps in the ultimate one. This model has been taken as a reference in several following works and still developed and considered as one of the most efficient and biologically plausible. Its performance has been tested, in fact, in the analysis of natural color scenes and compared with human behaviour and ability in the same circumstances, like in Itti and Koch (2000), Itti (2006) and Ouerhani *et al.* (2004).

The system proposed by Kayser *et al.* (2005) operates in the auditory field, similarly to the ones presented in Itti and Koch (2001) and Itti *et al.* (1998) to implement bottom-up visual visual attention. Features relative to spectral or temporal modulation (such as intensity, frequency structure and temporal structure) are extracted in parallel at different scale through different sets of filters and compared thanks to a center-surround mechanism. Then, the maps, obtained from each feature separately, are normalized and finally combined to obtain the saliency representation of the scene. In this way, short and long tones in a noisy background are salient, long tones have more salience than short tones, temporally modulated tones are more salient than stationary tones and also the *forward masking* (in a sequence of two closely spaced tones, the second is less salient) holds true.
The model proposed by Duangudom and Anderson (2007) uses, instead, features obtained from the analysis of the behaviour of auditory spectro-temporal receptive fields (STRFs), extracting the spectro and temporal components of the auditory spectrogram of the sound input. For each of these features, a saliency map has been built and grouped according to four broad feature classes. Then, an *inhibition stage*

promotes those relative to the features with prominent peaks, deleting or inhibiting the others. So, for each group, the resulting maps are combined, again subjected to inhibition and again combined to obtain the final saliency map.

2.3 Top-down approaches

Over years, bottom-up mechanisms have been investigated and studied much more than the top-down ones: basically because they are much easier to be analysed and controlled. Moreover, the way in which the two attentive modalities influence each other and their correlation are not well known, as pointed out in Frintrop *et al.* (2010). As we already mentioned, attention can be allocated in an automatic way (bottom-up cues) or trying to accomplish a particular task (top-down cues). In this case, attention needs to be switched in spite of the instinctive natural reactions and this implies an effort (Itti and Koch (2001)). Baluch and Itti (2011) in their review, differentiated between two different kinds of top-down mechanisms: *volitional top-down process* and *mandatory top-down process*. The first one is linked to the desire to act, while the second one is more automatic somehow and it could be hard to eliminate, because it is developed and learned by experience.

Many psychologically and neuro-biologically motivated visual models propose a very similar architecture in which information from bottom-up and top-down sources combines in a saliency map. Mozer (1991) proposed, maybe, the first model in this perspective, but more known is the one elaborated by Wolfe (1994b). Wolfe believed that the discriminant features that have to be used in the analysis could change according to the circumstances. Thus, they could be selected according to top-down criteria. The saliency maps obtainable through the application of his model comes, then, from a bottom-up approach, filtered by a top-down one. Milanese *et al.* (1994) exploited top-down information to integrate in a bottom-up object recognition system. Basically, the bottom-up analysis selects some regions where the objects could be and, afterwards, the top-down analysis returns a map exalting just the regions with recognized objects. In the attentive model proposed by Tsotsos *et al.* (1995) and already cited in the previous section, the focus of attention was determined, biasing the analysis with some more information about location or features, which could help to reduce the space of search, on the basis of the particular situation. Navalpakkam and Itti (2006) included in a bottom-up model based on the one proposed by Koch and Ullman (1985), top-down cues to facilitate visual search.
In all these models, the bottom-up was used just as a filter to bias all the incoming stimuli, in the model of Hamker (2006), instead, the bottom-up information influences also the representations of objects. Hamker tested the model, for an object detection task in natural scenes. Frintrop (2006) implemented a top-down mechanism to recognize target objects. An original view is given by Ognibene *et al.* (2010),

who investigated if and how bottom-up attention could affect top-down cues and the manner in which these cues are analysed. They managed to demonstrate that bottom-up attention can actually influence top-down attention control proficiency.

As we mentioned in Chapter 1.5, many experiments have been executed to investigate the human behaviour in scenarios characterized by different voices and sounds at the same time, with or without the presence of preliminary information or specific tasks. But computational approaches able to cover the effect of the top-down factors on the scene analysis have not been investigated yet as much as for visual attention, even if, as we already mentioned, similar mechanisms and principles are involved. Models dealing with a singular problem have been shown, but not generalized yet. Kalinli and Narayanan (2008), for example, faced the problem with prominent syllable detection task in speech. A classifier has been trained to distinguish between different categories, (in particular: prominent and non-prominent syllables) on the basis of the information extracted thanks to a procedure similar to those used in a bottom-up approach. A more complete model was proposed by De Coensel and Botteldooren (2008) for a soundscape case. They tried to obtain a well-adjusted combination between top-down and bottom-up mechanisms, allowing the execution of the concept of divided attention, thanks to the possibility of having more than one environmental sound noticed simultaneously. A different perspective is given by Wrigley and Brown (2004), who examined the way in which attention influences the formation of the various audio streams and elaborated a framework based on neural oscillators to implement the separation of the incoming stimuli in the different auditory flows components and elaborating an attentive model as a Gaussian mixture in frequency.

The so called *Bayesian definition of surprise*, proposed by Itti and Baldi (2006), could be considered another different approach to the studies about the influence of the top-down factors in the attentive analysis of a scene. Surprise, in fact, measures how data affects an observer, in terms of their expectations, as the difference between posterior and prior beliefs about the world. Previous knowledge and tasks eventually assigned can be included in the prior beliefs so that they can influence the computation of the final saliency.
Other works, such as the one performed by Taylor and Fragopanagos (2005), started analysing the way in which also cues like emotions, wills and so on could impact on an attentive analysis of a scene, but the field is still unripe and computational systems dealing with it seem still missing.

Part II

Attention as Active Decision

Chapter 3

What to measure next?

In this Chapter, we investigate a different interpretation of top-down attention as a *sequential measurement problem*, in which only some information is initially available and a sort of *active decision* procedure is implemented to evaluate, step by step, the best measure to perform for accomplishing a general task. An attentive model is, thus, proposed and its behaviour tested and discussed.

The work has been already presented and published in Hansen *et al.* (2011).

3.1 Sequential measurement problem

The sequential measurement problem has been investigated in different application fields, such as petro-physics and medical diagnosis by Wiegerinck *et al.* (2010), geophysics by Van Den Berg *et al.* (2003), robot navigation by Burgard *et al.* (1997) and active vision by Zhou *et al.* (2003) and Kappen *et al.* (1998).
Kappen *et al.* (1998) described the concept of active vision for an object recognition process in which just some observations about the objects are known and others need to be discovered to accomplish the task. However, thanks to the available ones and to experience, it is possible to make probabilistic predictions about the nature of the objects. In such a scenario the choice about which observation should be performed next to improve the recognition phase becomes crucial.
According to Kappen et al., also *features selection* and *next-view planning* problems can be considered as examples of active vision. In the first case, the aim is to find the subset of features able to best represent the objects; in the second one, the aim is to understand the 3D structure of an object from some of its 2D views. In both situations, the problem consists of, using just partial information, finding out which other features or which other views could maximally improve recognition.
The idea of a limited initial knowledge able to address more specific investigations is close to the *Contextual Guidance model*, proposed by Torralba *et al.* (2006). They

used the concept of *gist* of the scene to implement a top-down control on object search tasks. The *gist* could be defined as the *global features* of a scene, giving information about the context.
This is in line with different studies demonstrating that humans could get the semantic of a scene without identifying the objects in it; in several occasions, the spatial layout could be sufficient, as showed in Schyns and Oliva (1994), Oliva and Schyns (2000) and Rousselet *et al.* (2005) among others. Attneave (1954), in particular, in an earlier work, illustrated with intuitive examples, supported by experiments, that there is a big amount of redundancy in natural images and that humans seem to have a good inner predisposition to manage this redundancy. Torralba et al. used as contextual features, the ones able to represent "*the entire image holistically by extracting global statistics from the image*". In their work there is no explicit model of the 'task', but rather the top-down control guides focal attention, suggesting how the global features should be used to select the regions which seem to be more significant for the exploration.
These analyses have been exploited to implement a different approach to top-down attention modelling, exposed in Hansen *et al.* (2011). The problem of considering the way in which tasks, experience and knowledge could influence the attentive behaviour is managed as a *sequential measurement problem*. A sequential measurement problem in which just some contextual information is given and we want to know what it is the best measurement to perform next, in order to achieve some goal, maximizing, step by step, the acquirable knowledge.

3.2 Information Maximization

Shannon (1948) analysed the notion of information in a message, introducing the idea of information as a statistical entity. The aim was to understand how messages carrying an amount of data higher than the capacity of the channel could be still transmitted without making errors, taking advantage of the redundancy of the message itself. The idea is close to the one of Attneave (1954) in vision field and fundamental for all the procedures involving compression algorithms, cryptography and for communication systems.
Shannon considered a set of possible events, characterized by known probabilities of occurrence $p_l, p_2, ..., p_n$ and established a quantity H of the form given in Equation (3.1), as a measure of information and uncertainty called *entropy*.

$$H = -c \sum p_i \log p_i \tag{3.1}$$

where c is a positive constant. Entropy, according to Shannon, is "*a measure of how much choice is involved in the selection of the event or of how uncertain we are of the outcome*".
Lindley (1956), drawing inspiration from Shannon, adapted the concept of entropy to measure the information of an experiment, rather than a message. He claimed that,

even without a clear definition of the task, experiments could be carried out even just to gain knowledge about the state of nature. Thus, the objective is to select, given a set of experiments, the best one to execute, according to the amount of information the experiments could provide. The acquirable knowledge is computed using the Shannon's equation (Eq. 3.1) and the average amount of information brought by each experiment is computed as the average of the difference between the knowledge before and after performing it.
Bruce and Tsotsos (2006) proposed a new way of estimating saliency, based on an information maximization approach. In particular, they use *self-information* as a measure of local contrast to generate a bottom-up saliency map, taking into account that several studies, like the ones performed by Nothdurft (1990), showed that saliency is more related to the contrast, than to the effective intensity of a feature. The difference between *self-information* and *entropy*, *"which is closer to a measure of local activity"* was investigated through some elucidative examples [1].
Later, as we already mentioned in the previous section, Kappen *et al.* (1998) analysed an object recognition procedure, as an active vision task. The values of some features are known and others, instead, have to be found. The computation of the values of the remaining features is implemented as a sequential decision making process, in which the next feature to measure is selected with respect to the entropy value associated to the choice of the feature itself. Consistent with the work of Lindley (1956), then, this feature is the one giving the lowest value of entropy and, consequently, the highest amount of information. Considering the recognition process as the task, gaining knowledge represents, in fact, what also humans would prefer to do in order to achieve their goal.

3.3 A top-down task driven model

As we already mentioned in Section 3.1, we are interested in modelling top-down attention as a sequential measurement problem. Like in Torralba *et al.* (2006), the task is not exactly defined and we operate in a scenario in which only vague information is available, *the gist of the scene*, and we want to infer which one is the best measurement to perform next to improve our knowledge. According to Lindley (1956), we consider as the best measurement, the one able to provide the highest amount of information and, consequently, relative to the lowest level of entropy. The aim is to take advantage of optimization strategies to increase our knowledge and, consequently, improve the decision making process.
We highlight the fact that our attention model, which is based on a top-down driven feature selection mechanism among missing ones should not be confused with feature selection methods which aim to reduce number of redundant features especially

[1] If an event has probability p_i to occur, then *self-information* is a function of p_i and it is defined as $-\log p_i$

for large databases.
The model mimics human behaviour in similar circumstances. Many studies, like the one of Yarbus (1967) in the visual field, show the influence exercised by top-down factors, like the presence of a task, on human way of addressing attention. Some examples in this context were provided by Kalinli and Narayanan (2008). They argued that if we had to estimate the age of someone, we would tend to look at the face; if we had to guess, instead, people's material conditions, our attention would be more attracted by the clothes or other particular worn accessories. In both cases, there is a clear tendency to shift the gaze towards the aspects giving more information about the task. A similar pattern could be observed in an acoustic scenario: on the basis of our aim, in a complex scenario characterized by different voices and sounds at the same time, our attention would be grabbed by the speech signal, in case we needed to know what someone is talking about; we would pay attention to the music if, instead, the task was to identify the instrument which is playing.

We implement the sequential measurement problem in a *generic* Gaussian Discrete mixture model. The main novelty of our idea consists of proving the computational feasibility of a decision process, if carried out using this kind of model. Generally, in fact, the other works, proposing to use an informational theoretic approach, are specialized and applicable only in particular domains.
The model was tested on a missing data classification problem. The values of some features are known, others can be obtained; thus, in each step, the decision about the next features to take into account, which is supposed to be the most salient one, is made. In order to do it, the input feature saliency of the missing features, conditional on the available ones, is computed. The saliency is given as the expected reduction in entropy of the classification model over the output probabilities. The missing data scenario can be considered close to the one analysed by Ahmad and Tresp (1993). The measurement procedure is articulated in two different phases, as performed by Kappen *et al.* (1998). In the first phase, we train a generative classification model based on complete training data. In the second one, we apply the trained model to test data potentially with missing features. To test the quality of the top-down saliency estimator we classify the test data with and without the additional feature, as well as with a randomly selected additional feature.
The classification problem is implemented over a set of C classes indexed by the discrete variable y, $(y = 1, ..., C)$. The initial available information is represented by vectorial observation $\boldsymbol{x}$ with components x_i, $i = 1, ..., I$. Attending to a specific channel j, an additional measurement z_j can be performed. The measurement is chosen among the set of missing features $\boldsymbol{z}$ with components $z_1, ..., z_J$. Let the joint probability of the classes and all features observed and missing be denoted $p(y, \boldsymbol{x}, \boldsymbol{z})$. The aim is to make decision about y, minimizing the error rate. Hence, we invoke Bayesian decision theory and choose y according to the posterior distribution $p(y|...)$. The condition depends on the stage in the sequential measurement process. Initially,

the information available is $\boldsymbol{x}$, thus the relevant probability is

$$\begin{aligned} p(y|\boldsymbol{x}) &= \int p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \frac{\int p(y, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}{\sum_{y=1}^{C} \int p(y, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}. \end{aligned} \tag{3.2}$$

Using the top down attention mechanism we will select an additional feature $\boldsymbol{z}_j$, which will result in the distribution

$$\begin{aligned} p(y|\boldsymbol{x}, z_j) &= \sum_{y=1}^{C} \int p(y, \boldsymbol{z}|\boldsymbol{x}) \prod_{i \neq j} dz_i \\ &= \frac{\int p(y, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i}{\sum_{y=1}^{C} \int p(y, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i} \end{aligned} \tag{3.3}$$

The information value of this choice is given as the difference in confusion (entropy) before and after the second measurement, which will depend on the particular outcome of the sequential measurement, z_j,

$$\begin{aligned} \Delta S_j(\boldsymbol{x}, z_j) &= \sum_{y=1}^{C} \log p(y|\boldsymbol{x}, z_j) p(y|\boldsymbol{x}, z_j) \\ &- \sum_{y=1}^{C} \int \log p(y, \boldsymbol{z}|\boldsymbol{x}) p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \end{aligned} \tag{3.4}$$

As z_j is unknown at this stage in the process we are forced to average $\Delta S_j(\boldsymbol{x}, z_j)$ with respect to this variable given the information we have access to, i.e., with respect to the distribution of z_j conditioned on the initial measurement $\boldsymbol{x}$. This procedure provides us with the *expected information gain* of measuring the value of feature j [2]:

$$\begin{aligned} G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\ &= \sum_{y=1}^{C} \int \log p(y|\boldsymbol{x}, z_j) p(y, z_j|\boldsymbol{x}) dz_j \\ &\quad - \sum_{y=1}^{C} \int \log p(y, \boldsymbol{z}|\boldsymbol{x}) p(y, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}. \end{aligned} \tag{3.5}$$

The information gain can be used to rank features in importance.

[2] The second term does not depend on j, hence, can be neglected in the saliency estimate.

3.3.1 Gaussian Discrete mixture model and Information Gain

The Gaussian Discrete mixture model (GDMM) is a generative model of the joint distribution, see e.g., Hansen *et al.* (2000) and Larsen *et al.* (2002),

$$p(y, \boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^{K} p(k)p(y|k)p(\boldsymbol{x}, \boldsymbol{z}|k) \tag{3.6}$$

where K is the number of components, $p(k)$ are component probabilities, $p(y|k)$ is a $C \times K$ probability table, and $p(\boldsymbol{x}, \boldsymbol{z}|k)$ are K Gaussian pdfs. We choose a generative representation to allow for modeling of input dependencies which is necessary in order to make inference about missing features. Maximum likelihood parameter estimation in the GDMM leads to a straightforward generalization of expectation maximization algorithm for conventional mixtures.
Introducing the generative model of Eq. (3.6) in the information gain and using

$$p(k|\boldsymbol{x}) = p(k)p(\boldsymbol{x}|k)/p(\boldsymbol{x})$$

we obtain

$$\begin{aligned} G_j(\boldsymbol{x}) &= \sum_{y=1}^{C}\sum_{k=1}^{K} p(y|k)p(k|\boldsymbol{x}) \times \\ &\qquad \int \log\left[p(y, \boldsymbol{x}, z_j)\right] p(z_j|\boldsymbol{x}, k)dz_j \\ &- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log\left[p(\boldsymbol{x}, z_j)\right] p(z_j|\boldsymbol{x}, k)dz_j \\ &+ \text{const.} \end{aligned} \tag{3.7}$$

where

$$p(y, \boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k)p(y|k)p(\boldsymbol{x}, z_j|k)$$

and

$$p(\boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}, z_j|k)$$

Thus, computing G for all I features amounts to calculate $Q = I * (C + 1) * K$ one-dimensional integrals over Gaussian measures

$$p(z_j|\boldsymbol{x}, k) = \mathcal{N}(\mu_j(\boldsymbol{x}, k), \sigma_j^2(\boldsymbol{x}, k))$$

with

$$\begin{aligned} \mu_j(\boldsymbol{x}, k) &= \mu_{j,k} - \boldsymbol{\Sigma}_{z_j,\boldsymbol{x},k}\boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x},k}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x},k}) \\ \sigma_j^2(\boldsymbol{x}, k) &= \sigma_{j,k}^2 - \boldsymbol{\Sigma}_{z_j,\boldsymbol{x},k}\boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x},k}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},z_j,k}. \end{aligned} \tag{3.8}$$

In these expressions $\mu_{j,k}, \sigma^2_{j,k}$ are the mean and variance of the jth feature in the kth component, while $\mathbf{\Sigma}_{a,b,k}$ is the part of the covariance matrix of the k component corresponding to variable sets a, b.

3.3.2 Experimental Evaluation

In order to investigate the behaviour of the model, we checked the error rate in the classification problem for four well-known benchmark problems of the UCI depository (see Frank and Asuncion (2010)), which have heterogeneous input feature sets:

- the Pima Indian diabetes data
- the Liver Disorders data set
- the Abalone data
- the Yeast data

We analyse the results given choosing the next feature to measure according to the saliency estimate and randomly, comparing with a classifier that has access to all features and one that only has access to the original two features of the incomplete measurement. The Gaussian Discrete mixture model is generated on a training set of N_{train} data points; the test step is implemented on the remaining N_{test} data points. We simulate an incomplete measurement situation in which only $D_0 < D$ features are available for the estimation of saliency, where D is the number of features in the classification problem. The Gaussian Discrete model is trained with a variable number of components (K). Component covariance matrices are estimated with a simple Wishart prior with a diagonal mean covariance matrix of unit variance. Prior to training all variables are normalized to zero mean and unit variance. The initial training phase is carried out using an EM procedure, which is straightforward to generalize to incorporate the table structure of the Gaussian Discrete model. The training phase involves a multi-start procedure with 10 random initializations and further 1000 EM iterations are carried out on the initialization that leads to the lowest training error rate in classification.
Classification is done according to the posterior distribution

$$\begin{aligned} p(y|\boldsymbol{x},\boldsymbol{z}) &= \frac{\sum_{k=1}^{K} p(k)p(y|k)p(\boldsymbol{x},\boldsymbol{z}|k)}{\sum_{k=1}^{K} p(k)p(\boldsymbol{x},\boldsymbol{z}|k)} \\ &= \sum_{k=1}^{K} p(k)p(y|k)p(k|\boldsymbol{x},\boldsymbol{z}) \end{aligned} \tag{3.9}$$

to minimize the miss-classification rate.

The Pima Indian diabetes data

This data set concerns prediction of Diabetes Mellitus. The included patients are females at least 21 years old of Pima Indian heritage (see Smith *et al.* (1988)). The data set is split in training and test sets of sizes $N_{\text{train}} = 200$ and $N_{\text{test}} = 332$ respectively. Initial pilot runs indicates that $K = 5$ was a good bias-variance trade-off.
The initial training phase on complete data set provides a model with training error rate $E_{\text{train}} = 18.0\%$ and when evaluated on the complete test set shows a test error rate of $E_{\text{test}} = 21.6\%$, at par with test error rates reported elsewhere, see e.g. Ripley (1996).
To emulate an incomplete data scenario we test the top-down saliency estimate, represented in Equation (3.7) using an initial feature vector $\boldsymbol{x}$ comprising variables $(1, 7)$ representing the '*number of times pregnant*' and '*age*' features within the list of a total of seven measures:

1. number of times pregnant
2. plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. diastolic blood pressure *(mm Hg)*
4. triceps skin fold thickness *(mm)*
5. body mass index *(weight in kg/(height in m)^2)*
6. diabetes pedigree function
7. age *(years)*.

The input feature 2 '*Plasma glucose concentration at 2 hours in an oral glucose tolerance test*' is interesting for its known diagnostic value (see Smith *et al.* (1988)) and also because it is both more expensive to obtain and delayed relative to the other features. The *diabetes pedigree function* represents the Diabetes Mellitus history in relatives and the genetic relationship of relatives to the subject. It is based on information about parents, grandparents, full and half siblings, full and half aunts and uncles, and first cousins (see Smith *et al.* (1988)), and thus is also complex and time consuming to obtain.
To have an initial expression of the input feature relevance we measured the mutual information between each input feature and the class label. The mutual information is tested against a null hypothesis of no mutual information using a simple permutation test ($N_{\text{resamples}} = 200$), the null was rejected if $p > 0.01$, and in this case the mutual information is reported. We found that features $(1, 2, 7)$ were significantly informative on a single feature basis, as seen in panel (e) of Figure 3.1, with feature 2 as the most informative as expected. We observe that the frequencies are not simply given by the mutual information. This result underlines the need for an attention model.
Going through all test examples we note both the saliency allocated to features

$(2, 3, 4, 5, 6)$, as well as the rate at which they are chosen for measurement as being the most salient. The resulting distributions are shown in panels (b) and (c) of Figure 3.1. Clearly, feature 2 is the most important in terms of saliency and is attended to in more than 80% of the test data. The choice of features is broken down within the two classes in panel (d), and interestingly we find that global result of panel (b) appears from a somewhat different distributions in the two classes indicating the importance of the interaction between the initial input $\boldsymbol{x}$ and the top down mechanism in the saliency estimate.
The classification performances of the various schemes are shown in panel (a). The saliency based feature choice leads to an error rate of 24%, while classification based on the initial features $(1, 7)$ only leads to poor performance close to baseline random guessing. Classification based on $(1, 7)$ combined with a *randomly selected feature* leads to a somewhat higher error rate (29%).
In conclusion, we would recommend to acquire the plasma glucose concentration, while the Diabetes Pedigree Function seems irrelevant in the present sample.

The Liver Disorders data set

This classification task concerns prediction of liver disorder based on blood tests and alcohol consumption of male individuals. The task is proposed and data donated by BUPA Medical Research Ltd. (see Frank and Asuncion (2010)). The list of features for this problem comprises:

1. mean corpuscular volume,
2. alkaline phosphotase,
3. alamine aminotransferase,
4. aspartate aminotransferase
5. gamma-glutamyl transpeptidase,
6. drinks *(as the number of half-pint equivalents of alcoholic beverages drunk per day)*.

The first 5 variables are blood tests features believed to be sensitive to liver disorders that might arise from alcohol consumption. Each data sample is based on the record of a single male individual.
We tested the single feature information content as shown in Figure 3.2 panel (e) and found that only features (5,6) are informative on their own $(p < 0.01)$. This data set is smaller with $N_{\text{train}} = 200$ and $N_{\text{test}} = 95$. We train the Gaussian Discrete model with $K = 10$. The baseline error rate on the test data is 42%.
We emulate a severely missing data situation by letting the initial incomplete measurement be given as two features $(3, 4)$ which both are deemed uninformative by the permutation test. As in the diabetes data we run through all test examples and

note both the saliency allocated to features $(1, 2, 5, 6)$ as well as the rate at which they are chosen as the most salient. The resulting distributions are shown in panels (b) and (c) of Figure 3.2. The behavioral feature 6 (number of drinks) obtains some attention but is rarely chosen as the additional most salient feature for classification. The most salient and most often chosen features are 1 and 5, that are chosen for attention in 55% and 40% of the test respectively. The choice of features is broken down in classes in panel (d), and again we see that the classes require somewhat different choice of feature showing the interaction between values of the initial input $(3, 4)$ and the top down influence on the saliency estimate.
We find it interesting that even in the case of such a relatively 'uninformative' initial measurement as we have access to in this experiment, the top down saliency estimate is successful in locating features that lead to a classification performance on par with that obtained when using all features ($\sim 30\%$).

The Abalone data

The task is to predict the age of Abalone from physical measurements (see Nash and Laboratories (1994) and Waugh (1995)). Here we convert the problem to a binary decision problem (young vs. old). The variables used as features are:

1. gender *(M, F)*,
2. length, longest shell measurement,
3. diameter, perpendicular to length,
4. height, with meat in shell,
5. whole weight, whole abalone,
6. shucked weight, weight of meat,
7. viscera weight, gut weight *(after bleeding)*,
8. shell weight, after being dried.

This data set is larger with $N_{\text{train}} = 3500$ and $N_{\text{test}} = 677$. We train the Gaussian Discrete model with $K = 17$.
The baseline error rate is 50% and the trained model obtains a test error below 20% based on complete measurements, thus this is a well calibrated modeling problem. We further tested the single feature information content as shown in Figure 3.3, panel (e) and find that all features are informative $(p < 0.01)$, with feature 8 as the most informative with almost 0.4 bits of information. We choose in this case to provide features $(1, 2)$ and evaluate the top down saliency for features $(3, 4, 5, 6, 7, 8)$.
As expected, we find feature 8 is allocated the most saliency and is most often attended to, being proposed in 75% of the test cases. The resulting classifier is significantly improved over random attention (22% vs. 27%).

The Yeast data

This last data set used for illustration of the top down saliency estimate concerns determination of protein cellular localization sites (see Horton and Nakai (1996)). The variables used as features include:

1. McGeoch's method for signal sequence recognition,
2. Von Heijne's method for signal sequence recognition,
3. score of the ALOM membrane spanning region prediction program,
4. score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins,
5. presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen),
6. peroxisomal targeting signal in the C-terminus,
7. score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins,
8. score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

The classification becomes a binary decision process by selecting a subset associated with two most frequent sequence types *CYT* (cytosolic/cytoskeletal 463 examples) and *NUC* (nuclear, 429 examples) in SWISS-PROT database.
This data set comprises of a training set with $N_{\text{train}} = 650$) and a test set with $N_{\text{test}} = 242$ samples. We train the Gaussian Discrete model with $K = 11$. The baseline error rate is 45% and the trained models obtain test and training error rates around 30% based on complete measurements, thus signifying a noisier decision problem than the previous Abalone case.
We further tested the single feature information content. As in the Abalone data set all features are informative (see Figure 3.4, panel (e)). To emulate incomplete data we provide the classifier with features $(1, 5)$ that are both significantly informative, but at a somewhat lower value than in the Abalone example (less than 0.08 bits of information). As in the previous examples we note an improvement in performance of the top-down saliency strategy relative to classifying based on both the initial features $(1, 5)$, as well as compared to providing an extra feature chosen at random. Inspecting the feature selection in Figure 3.4, panel (c) we see that the process almost exclusively chooses to add feature 8 to initially given features $(1, 5)$.

3.3.3 General discussion and conclusion

The experiments carried out for these data sets, show that our attention mechanism provides for improved classification over simple random attention. Improvements are

found even if the initial incomplete feature set was of limited information itself.
In the tests discussed above, the initial available features are randomly chosen to show the behaviour of the model in different situations. In the Pima indian diabetes diagnoses problem, for example, the starting scenario is given by features $(1, 7)$ which seem to be informative enough according to mutual information; unlike features $(3, 4)$ for Liver disorder problem; two more intermediate situations are proposed and analysed in the case of Yeast and Abalone data.
Further experiments were executed, considering also mutual information as one of the possible criteria to use for choosing the next feature to measure: the subset of features initially available are randomly generated. The averaged error rates are analysed and the choice driven by the attentive model proves to be the one providing the best classification performance. In Figure 3.5, 3.6, 3.7 and 3.8, the results obtained, respectively, for Pima indian diabetes diagnoses data sets, Liver Disorders, Abalone and Yeast data are presented.

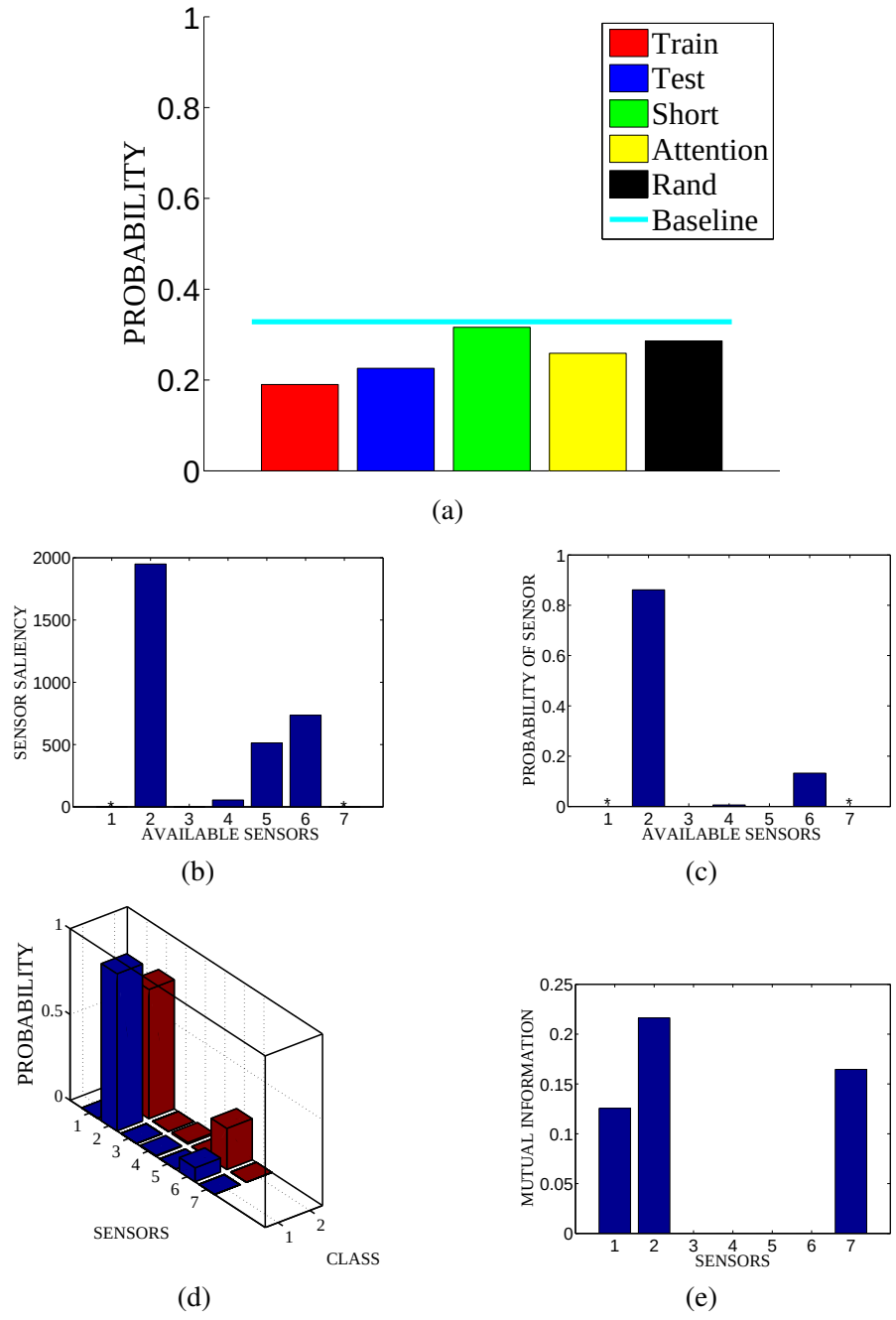


Figure 3.1: Pima indian diabetes diagnoses problem. Seven input features are considered. Here we simulate incomplete measurement, in which features $(1, 7) = (\#\text{pregnancies}, \text{age})$ are given. The panels show (a) Error rates in the training set ($N_{\text{train}} = 200$) and the test set ($N_{\text{test}} = 332$) for complete data, test error using only the initial feature set $(1, 7)$, and test set using $(1, 7)$, and the feature chosen among $2 - 6$ by the top down saliency estimate, and finally the test error obtained using features (1,7) and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector (1,7); (c) Frequency of selection of features 2-6; (d) Frequency of selection in test cases within the two classes;(e) The $\log_2$ mutual information between features and class label.

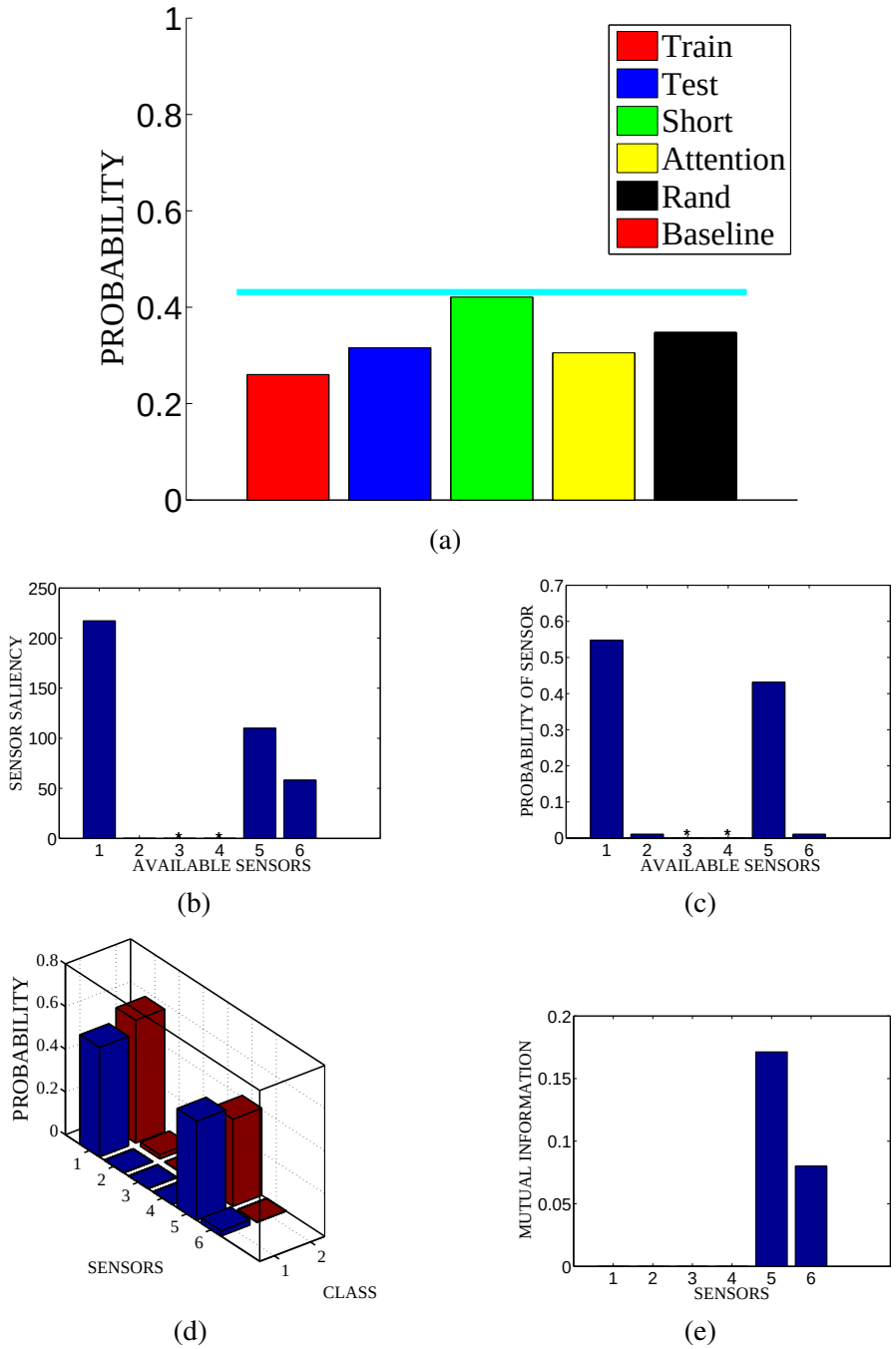


Figure 3.2: Liver disorder problem. Seven input features are considered. Here we simulate incomplete measurement, in which features $(3, 4)$ only are provided. (a) Error rates in the training set of complete data ($N_{\text{train}} = 200$) and the test set ($N_{\text{test}} = 95$) using complete data, the test error using the initial feature set $(3, 4)$, and the test error using $(3, 4)$ and the feature chosen by the top down saliency estimate, and finally the test error using features (3,4) and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector (3,4); (c) Frequency of selection of the additional features; (d) Frequency of selection of features in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

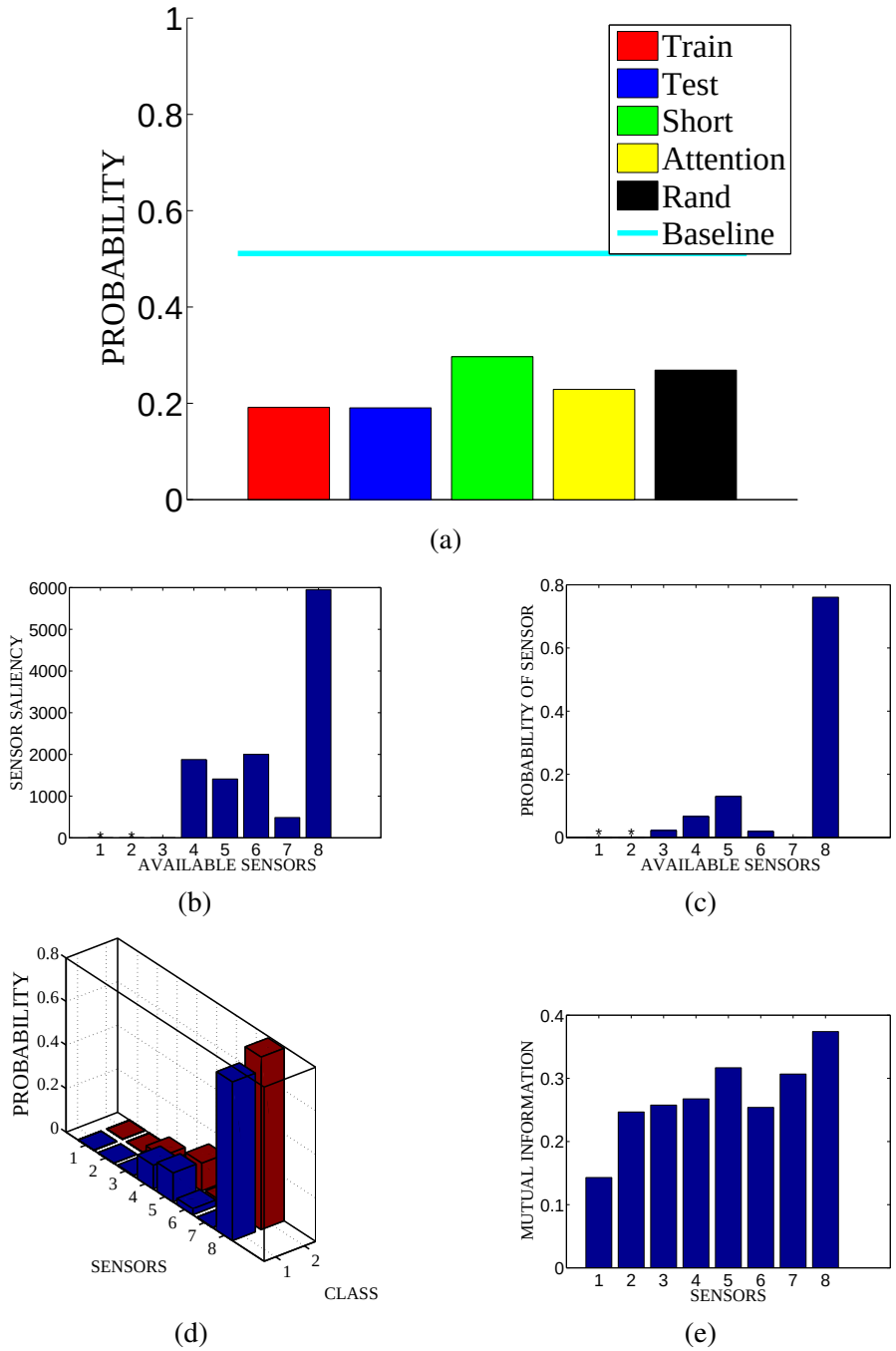


Figure 3.3: Abalone data converted to a classification problem (old/young). Eight input features are considered. We simulate incomplete test measurement, in which only features $(1, 2)$ are included. The panels show: (a) Error rates in the training set of complete data ($N_{\text{train}} = 2500$) and the error on the test set ($N_{\text{test}} = 677$) using complete data, test error rate when using the initial feature set $(1, 2)$, test error using $(1, 2)$ and the feature chosen by the top down saliency estimate, and finally the test error obtained using (1,2) and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector (1,2); (c) Frequency of selection of the additional features; (d) Frequency of selection of features in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

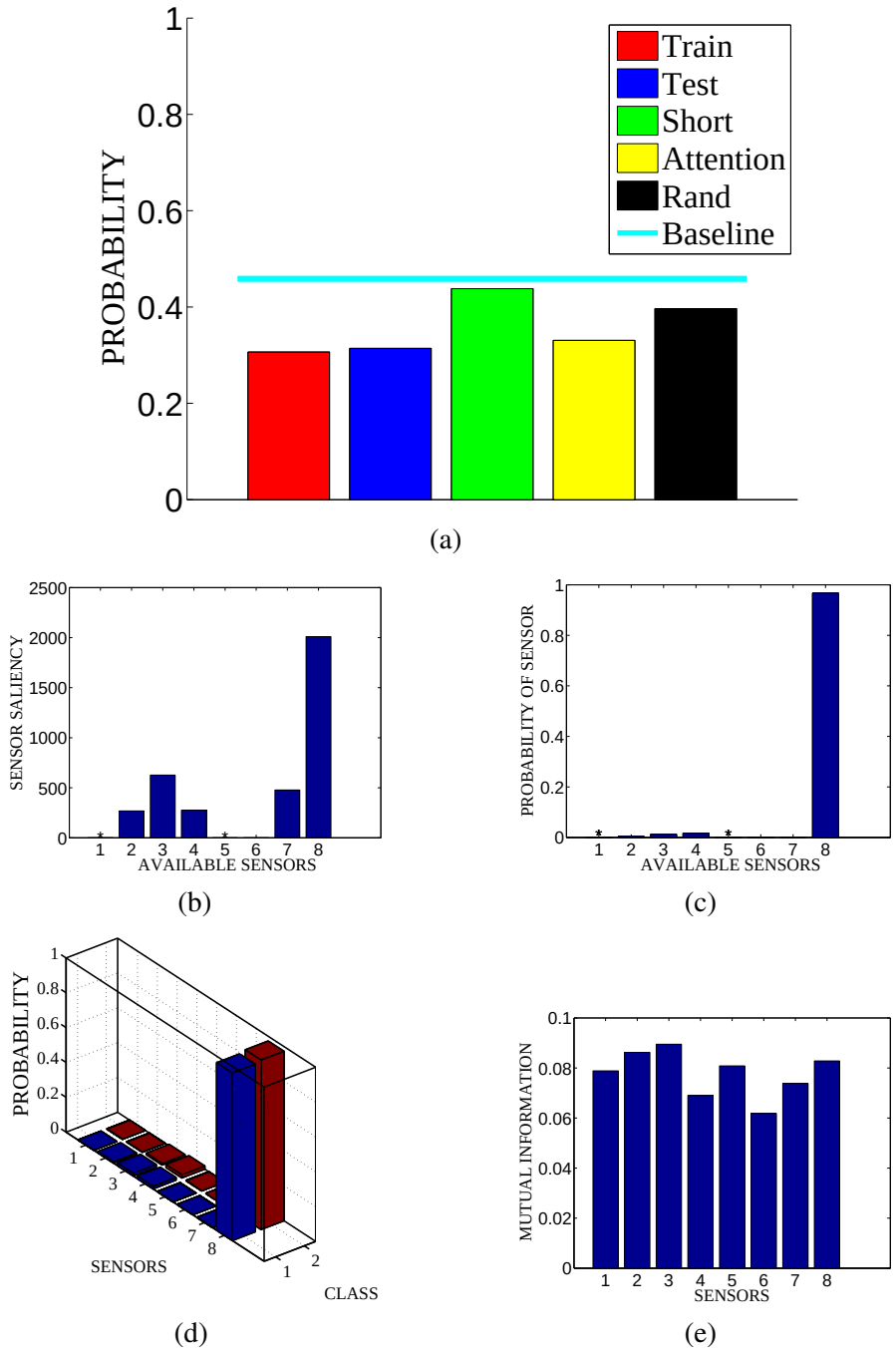


Figure 3.4: Yeast data. Eight input features are considered. We simulate incomplete test measurement, in which the two features $(1, 5)$ are given. The panels show: (a) Error rates in the training set of complete data ($N_{\text{train}} = 650$) and in the test set ($N_{\text{test}} = 242$) using complete data, test error rate using only the initial feature set $(1, 5)$, and test error using $(1, 5)$, and the feature chosen by the top down saliency estimate, and finally the test error obtained using $(1, 5)$ and a randomly chosen additional feature; (b) Estimated information saliency obtained on the test data, given the incomplete feature vector $(1, 5)$; (c) Frequency of selection of the additional features; (d) Frequency of selection of features in test cases within the two classes; (e) The $\log_2$ mutual information between features and class label.

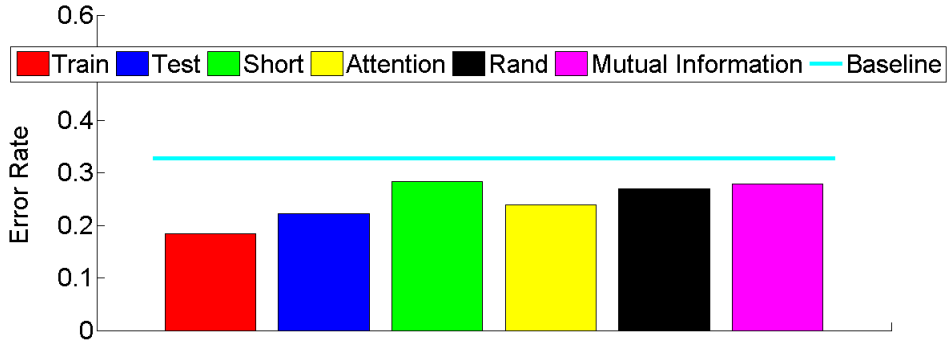


Figure 3.5: Pima indian diabetes diagnoses problem. Seven input features are considered. Here we simulate an incomplete measurement, in which only 2 of the features are given. 200 random subsets of features initially available are considered and the average error rates obtained are shown.. In particular, from the left to the right: error rates in the training set ($N_{\text{train}} = 200$) and the test set ($N_{\text{test}} = 332$) for complete data, test error using only the initial feature set, test set using the initial subset and the feature chosen among the remaining ones by the top down saliency estimate, the test error obtained using the initially available features and a randomly chosen additional feature and, finally, the test error using the initial features and as an additional feature, the most informative one according to mutual information.

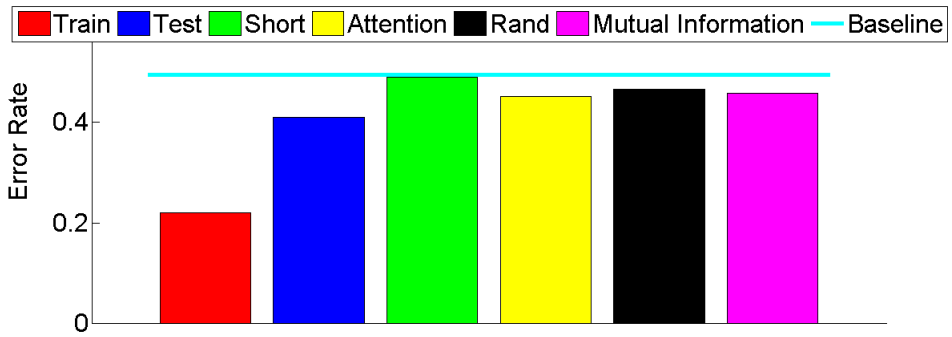


Figure 3.6: Liver disorder problem. Seven input features are considered. Here we simulate an incomplete measurement, in which only 2 of the features are given. 200 random subsets of features initially available are considered and the average error rates obtained are shown. In particular, from the left to the right: error rates in the training set of complete data ($N_{\text{train}} = 200$) and the test set ($N_{\text{test}} = 95$) using complete data, the test error using the initial feature set, the test error using the initial features set and the feature chosen by the top down saliency estimate, the test error using the initial features and a randomly chosen additional feature and, finally, the test error using the initial features and as an additional feature, the most informative one according to mutual information.

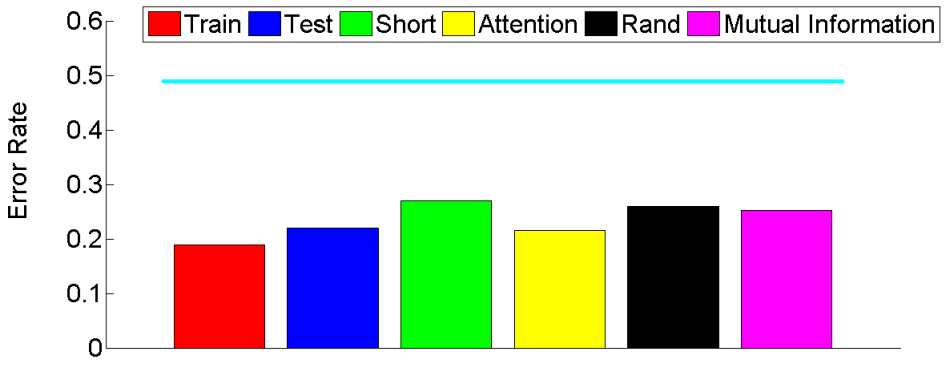


Figure 3.7: Abalone data converted to a classification problem (old/young). Eight input features are considered. Here we simulate an incomplete measurement, in which only 2 of the features are given. 200 random subsets of features initially available are considered and the average error rates obtained are shown. In particular, from the left to the right: error rates in the training set of complete data ($N_{\text{train}} = 2500$) and the test set ($N_{\text{test}} = 677$) using complete data, the test error using the initial feature set, the test error using the initial features set and the feature chosen by the top down saliency estimate, the test error using the initial features and a randomly chosen additional feature and, finally, the test error using the initial features and as an additional feature, the most informative one according to mutual information.

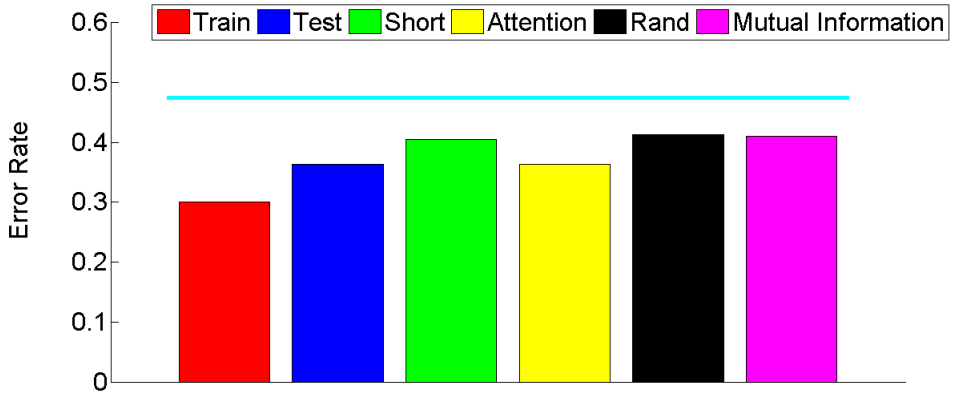


Figure 3.8: Yeast data. Eight input features are considered. Here we simulate an incomplete measurement, in which only 2 of the features are given. 200 random subsets of features initially available are considered and the average error rates obtained are shown. In particular, from the left to the right: error rates in the training set of complete data ($N_{\text{train}} = 650$) and the test set ($N_{\text{test}} = 242$) using complete data, the test error using the initial feature set, the test error using the initial features set and the feature chosen by the top down saliency estimate, the test error using the initial features and a randomly chosen additional feature and, finally, the test error using the initial features and as an additional feature, the most informative one according to mutual information.

Chapter 4

Missing Features Problem

In this Chapter, we analyse the so called *missing data problem*. We present the main techniques proposed over years to face this kind of issue, comparing their behaviour in classification problems. In order to do this, we introduce and exploit a new definition of efficiency.

The work has been already presented and published in Karadogan *et al.* (2011a).

4.1 Missing data problem

As we show in the Chapter 3, we model top-down attention as a sequential measurement problem. The top-down attention model is implemented using a *generic* Gaussian Discrete mixture model and tested on a missing data classification problem. Initially, just some feature values are available, but it may be possible to obtain the others. The job of the model is to suggest the best feature to measure next. In this case, the dataset used to train the model is complete. But, considering a more realistic and general situation, holes could be present in the training set as well.
This is what is called in literature *missing data problem*.
The necessity to deal with this kind of scenario is really common in various studies and several applications using statistical approaches, such as: psychological and psychometric analyses dealing with surveys without all the requested answers, market researches exploiting incomplete interviews or medical diagnoses based on partial accessible information. Audio and video processing, as well, have often to face these conditions of incompleteness: for example in the reconstruction of degraded audio and video sequences, in the analysis of images with missing pixels or occlusions (like in Ahmad and Tresp (1993)), in the manipulation of distorted signals because of a sensor failure or outliers and so on.

Rubin (1976) introduced a classification of missingness schemes. Following the ex-

emplification carried out by Schafer and Graham (2002), let us define M as the missingness of the data set, which can be expressed in different forms (see Schafer and Graham (2002) for more details). Let us indicate the complete data set as X_{full}, the missing part as X_{miss} and the present one as X_{pres}, so that $X_{full} = (X_{pres}, X_{miss})$. According to Rubin definition, it is possible, then, to distinguish, basically, between three big classes:

missing at random (MAR) : if the probability of missingness can depend on any of the observed variables, but not on the missing ones. This is the most common case and it also called *ignorable non-response*.

$$P(M|X_{full}) = P(M|X_{pres})$$

missing completely at random (MCAR) : if the probability of missingness can not depend on any of the observed variable or on the missing ones. It is a particular case of the missing at random scheme.

$$P(M|X_{full}) = P(M)$$

It is equivalent to flipping a coin to determine whether an observation is missing;

missing not at random (MNAR) : if the probability of missingness can depend on the missing variables themselves. It is also called *non-ignorable non-response*;

$P(M|X_{full})$ cannot be simplified

Schafer and Graham (2002) proposed a simple example to elucidate the meaning of the previous definitions. They consider a scenario in which the systolic blood pressures of some participants are recorded in January and some of the people are again called in February for a second reading. The February group could be chosen randomly: in this case, the data which are not present in February (the pressure of the subjects who did not participate in the second analysis) are missing completely at random. In case, instead, the people to be called for performing a new blood test are chosen on the basis of the previous values of their pressure (maybe just the ones with a pressure value in the hypertensive range), the data are missing at random, but not completely at random. There is, in fact, dependence on the previous observed data. The data are not missing at random, instead, if, for instance, all the subjects are called to take a new pressure measure, but just the ones who are in the hypertensive range are registered.

4.2 Missing data techniques

Roth (1994) and Allison (2001) provided deep reviews about missing data problem, analysing also weaknesses and strength points of the procedures presented over years to deal with it. Basically, it is possible to group these techniques in three main categories:

- *deletion* methods
- *imputation* methods
- *model-based* methods

Deletion methods are so called, because they consider in the analysis only the present data. The data with missing features are faced following, generally, two possible strategies: *listwise deletion* and *pairwise deletion*. In the first case, the entire data samples containing the missing input features are ignored (deleted); in the second one, only the missing features are removed, the other present features in the same sample are kept and used. As Allison (2001) pointed out, the use of *listwise deletion* could discard a huge amount of potentially usable data, but provide good estimates of standard errors; *pairwise deletion*, instead, exploits all the available information, but it makes pretty difficult to obtain accurate standard error estimates.

Imputation methods insert a new value in place of a missing variable, according to different criteria. The new value could be derived from the values of other variables in the dataset (*regression imputation*) or from the estimated distribution of the missing variable. The reason of this last replacement is founded on *"the idea that any subject in a study sample can be replaced by a new randomly chosen subject from the same source population"*, as reported by Donders *et al.* (2006). The simplest imputation method is the *mean imputation*. It consists of substituting the missing value of a variable with the mean value of the same variable. This method, in spite of its simplicity, tends to produce biased estimates, as shown by Haitovsky (1968). In the *hot-deck imputation*, instead, a missing value is replaced with a value extracted by a similar observed sample in the data set. Close to this idea is the principle of *similar response pattern imputation*, which consists of identifying the most similar unit without missing information and replace the missing part with the correspondent values of this unit.
If each missing instance is replaced with just another value, the procedure is called *single imputation*, otherwise, if more values are taken into account, *multiple imputation*. This procedure, that was firstly introduced by Rubin (1987), makes use of a Monte Carlo technique, in which many complete imputed data sets are generated and analysed by standard complete data methods. The results obtained give information about possible estimates of missing variables.

In all the imputation methods, as in the pairwise deletion, all the available information is considered.

The *model-based* methods are able to perform directly their analysis on the incomplete set, without changing or ignoring part of the available information. Maximum likelihood (ML) approaches are the most representative in this category. The general idea is to model the data distribution by the observed values and then to use this distribution for estimating the missing ones. Expectation Maximization procedures (EM) are often used in this perspective. The underlying principle is to iteratively estimate the distribution and the missing data, until there convergence in the estimation of the distribution is reached. Ghahramani and Jordan (1994) and Dempster *et al.* (1977) presented two of the most known algorithms for this kind of approach.

4.3 Missing data techniques evaluation

All these techniques have been analysed and tested over the years. Their behaviour has been checked in different circumstances, dealing with different missingness schemes and the relative performance evaluated. Roth (1994) provides a qualitative evaluation of the most common missing data approaches considering scenarios in applied psychology. Allison (2001) analyses advantages and disadvantages of the same methods, on the basis of three criteria: the capability to minimize bias, maximize the use of available information and yield good estimates of uncertainty. Different principles are, instead, used by Schafer and Graham (2002) to investigate and compare the functioning of the methods. They followed Neyman and Pearson (1933) and Neyman (1937), considering bias and mean square error to evaluate the model estimation accuracy and the trend of the standard error to estimate the margin of the uncertainty. Myrtveit *et al.* (2001) studied the behaviour of missing data techniques for software cost modelling. In this context, listwise deletion is considered the most frequently utilized. Their work focuses, then, on the possible benefits that could be obtained, using, rather than it, maximum likelihood, multiple imputation and similar response pattern imputation approaches. The strategies are compared operating on an ERP [1](Enterprise Resource Planning) dataset.

However, to our knowledge, a standard and general strategy to compare different missing data techniques and to evaluate their performance have not been proposed yet. In order to fill the gap, we proposed a specific definition of efficiency which can be used to analyse how an algorithm operates on data sets in which not all the information might be present. Specifically, we tested the behaviour of the maximum likelihood method in Ghahramani and Jordan (1994) (Complete EM), pairwise dele-

[1]Enterprise Resource Planning (ERP): IT system dedicated for companies and enterprises. For more information see http://www.systemerp.net/

tion and mean imputation in a classification problem, using the Gaussian Mixture Model proposed in Larsen *et al.* (2002), with different percentage of missing information in the training set.
As data have more missing values, the resultant error rate (ER) gets higher due to lack of information, but the resultant curves are different for different missing data techniques. Thus, we provide a new definition of efficiency, which is based on a comparative analysis of the error rate curve generated by each algorithm. In particular, the behaviour of each method is evaluated with respect to the performance shown by listwise deletion. In other words, we calculate how efficient a technique makes use of data with missing values instead of simply ignoring them.
As seen in Figure 4.1, the definition of efficiency (Eff) we introduce is given by calculating the area under the reference and actual curves (curves of the technique investigated), as in Equation (4.1).

$$\text{Eff } \% = \frac{A_R - A_A}{A_R} \times 100 \tag{4.1}$$

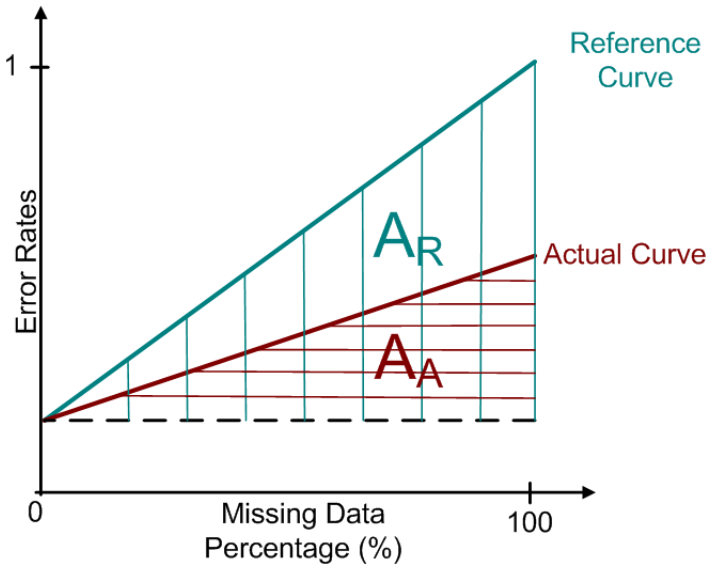


Figure 4.1: The illustration for the efficiency calculation method used.

According to this definition, when the actual curve is the same as the reference curve, the efficiency is 0%, while it is 100%, when it is a horizontal line (the error rate is not effected as the percentage of missing data changes).
We calculate the efficiency where training data are missing completely at random.

4.3.1 Modelling Framework

The missing data techniques we want to evaluate are: the maximum likelihood method proposed in Ghahramani and Jordan (1994) (CEM), the pairwise deletion (PW) and the mean imputation (MI). In order to do it, we use for the efficiency computation the Gaussian mixture model (GMM) that is explained and implemented in Larsen *et al.* (2002).
Define $\boldsymbol{x}$ as the d-dimensional input feature vector and the associated output, $y \in \{1, 2, \cdots, C\}$, of class labels, assuming C mutually exclusive classes. The joint input/output density is modeled as the Gaussian mixture:

$$p(y, \boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} P(y|k)p(\boldsymbol{x}|k)P(k) \tag{4.2}$$

$$p(\boldsymbol{x}|k) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right) \tag{4.3}$$

where

- K is the number of components,
- $p(\boldsymbol{x}|k)$ are the component Gaussians mixed with the non-negative priors $P(k)$, such that $\sum_{k=1}^{K} P(k) = 1$
- the class-cluster posteriors $P(y|k)$, such that $\sum_{y=1}^{C} P(y|k) = 1$.

The kth Gaussian component is described by the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$. $\boldsymbol{\theta}$ is the vector of all model parameters,

$$\boldsymbol{\theta} \equiv \{P(y|k), \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$$

The joint input/output for each components is assumed to factorized

$$p(y, \boldsymbol{x}|k) = P(y|k)p(\boldsymbol{x}|k)$$

The input density associated with Equation (4.2) is given by

$$p(\boldsymbol{x}|\boldsymbol{\theta}_u) = \sum_{y=1}^{C} p(y, \boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|k)P(k),$$

where

$$\boldsymbol{\theta}_u \equiv \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(k) : \forall k, y\}$$

Assuming a 0/1 loss function, the optimal Bayes classification rule is

$$\widehat{y} = \max_{y} P(y|\boldsymbol{x})$$

where[2]

$$P(y|\boldsymbol{x}) = \frac{p(y,\boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{k=1}^{K} P(y|k)P(k|\boldsymbol{x})$$

with

$$P(k|\boldsymbol{x}) = p(\boldsymbol{x}|k)P(k)/p(\boldsymbol{x})$$

Let us define the data set of labeled examples as $\mathcal{D}_l = \{\boldsymbol{x}_n, y_n; n = 1, 2, \cdots, N_l\}$. The negative log-likelihood for the data sets, which are assumed to consist of independent examples, is given by

$$L = -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n\in\mathcal{D}_l} \log \sum_{k=1}^{K} P(y_n|k)p(\boldsymbol{x}_n|k)P(k)$$

The model parameters are estimated with the iterative modified EM algorithm in Larsen *et al.* (2002):

1. To initialize the mean ($\boldsymbol{\mu}_0$) and covariance ($\boldsymbol{\Sigma}_0$) matrices, all train data set is considered as one normal distribution. In the case of missing data, the calculations are done using only observed data and the $\boldsymbol{\Sigma}_0$ is regularized (see section 4.3.1). Then, since random points from the distribution can not be taken as cluster center points because of missing data, we draw L random samples using the $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, and get rid of outliers. Instead of taking random center points from the remaining samples, we use KKZ method assuming the clusters will be distant from each other, as suggested by Su and Dy (2007). The KKZ method is as the following:
   - The first center point is taken as the sample having the largest L_2 norm
   - Other center points are calculated as having the largest distance to the closest center points
2. Compute posterior component probability for all $n \in \mathcal{D}_l$:

$$p(k|y_n, \boldsymbol{x}_n) = \frac{P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}{\sum_k P(y_n|k)p(\boldsymbol{x}_n|k)P(k)}. \tag{4.4}$$

3. For all k, update mean vectors and covariance matrices:

$$\boldsymbol{\mu}_k = \frac{\sum_{n\in\mathcal{D}_l} \boldsymbol{x}_n P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}, \qquad \boldsymbol{\Sigma}_k = \frac{\sum_{n\in\mathcal{D}_l} \boldsymbol{S}_{kn} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n\in\mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}$$

   where $\boldsymbol{S}_{kn} = (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$.

[2]The dependence on $\boldsymbol{\theta}$ is omitted.

4. For all k, update cluster priors and class cluster posteriors:

$$P(k) = \frac{\sum_{n \in \mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}{N_l}, \qquad P(y|k) = \frac{\sum_{n \in \mathcal{D}_l} \delta_{y_n} P(k|y_n, \boldsymbol{x}_n)}{\sum_{n \in \mathcal{D}_l} P(k|y_n, \boldsymbol{x}_n)}$$

Pairwise Deletion

To implement the pairwise method, the only difference made on the model we just described, is the update of posterior input density $p(\boldsymbol{x}_n|k)$, the mean vector μ_k and covariance matrix $\boldsymbol{\Sigma}_k$. To update those, observed data for each variable or pair of variables are used.
However, the estimated covariance matrix is unbiased and is not guaranteed to be positive semi definite. We implement a commonly used approach, given by Schneider (2001) to regularize the covariance matrix. The approach consists of inflating the diagonal elements by the factor $(1 + h)$ as in Equation (4.5):

$$\boldsymbol{\Sigma}^{'} = \boldsymbol{\Sigma} + h\boldsymbol{I} \tag{4.5}$$

where $\boldsymbol{I}$ is the identity matrix and h is a regularization parameter. h is determined in the following way:

$$\boldsymbol{\Sigma}^{'} = \boldsymbol{\Sigma} + h\boldsymbol{I} = \boldsymbol{VUV^{-1}} + h\boldsymbol{VV^{-1}} = \boldsymbol{V(hI + U)V^{-1}} \tag{4.6}$$

where $\boldsymbol{VUV^{-1}}$ is the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{V}$ is the square matrix whose ith column is the eigenvector q_i of $\boldsymbol{\Sigma}$ and $\boldsymbol{U}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. Then, we choose h such that $(h + U) > 0$ to have nonnegative eigenvalues in regularized covariance matrix.

Mean Imputation

Mean imputation method is a replacement technique where a missing variable is replaced by the corresponding mean value. The model we use is not effected in this method, since we have complete data after imputation. This method keeps all data, and is easy to implement. However, the variance estimates are lessened as more means are added.

Complete Expectation Maximization

This method is a maximum likelihood missing data technique that is proposed in Ghahramani and Jordan (1994). EM is used both for the estimation of model components and for dealing with missing data. Posterior component probability, $p(k|y_n, \boldsymbol{x}_n)$

is again calculated as in Equation (4.4), but only on observed dimensions. To update the mean vector, $E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]$ is substituted for missing components of $\boldsymbol{x}_n$, and to update the covariance matrix, $E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^T}|\boldsymbol{x}_n^o]$ is substituted for outer product matrices containing missing components:

$$E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o] = \mu_n^m + \boldsymbol{\Sigma}_n^{mo}\boldsymbol{\Sigma}_n^{oo^{-1}}(\boldsymbol{x}_n^o - \mu_n^o),$$

$$E[\boldsymbol{x}_n^m \boldsymbol{x}_n^{m^\top}|\boldsymbol{x}_n^o] = \boldsymbol{\Sigma}_n^{mm} - \boldsymbol{\Sigma}_n^{mo}\boldsymbol{\Sigma}_n^{oo^{-1}}\boldsymbol{\Sigma}_n^{mo^T} + E[\boldsymbol{x}_n^m|x_n^o]E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]^\top.$$

4.3.2 Experimental Evaluations

As we already mentioned, we carried out some experiments, using MATLAB, for the efficiency computation of missing data techniques on synthetically generated data and two datasets from the UCI archive: Iris and Pima-Indian-Diabetes (Frank and Asuncion (2010)). The missing data percentage (MDP) is determined randomly (MCAR). The experiment is done in such that not all values can be missing in one observation (if all data in all directions are missing it would be equal to deleting it, so reducing training data as in our reference method).
We experiment how the misclassification rate (MR) changes with MDP and calculate the efficiency, according to Eq. (4.1), using those results for different MDP values. In particular, we made 100 iterations for each experiment, while changing MDP between 0% and 70%. We carry out experiments in two cases:

case 1 : the test dataset also have missing values with same MDP as for training. The aim is to check how robust the estimated model is to missing data.

case 2 : the test data are complete. The aim is to investigate how well the model is estimated, in spite of the missingness in the training.

Synthetic Dataset

The algorithms were tested on synthetic data. The multidimensional input data are generated by a Gaussian mixture model. The number of mixtures K, is 3. The difficulty of the problem is determined using the SNR calculation.
Let ds_{kl} be the distance between μ_k and μ_l, eig_k be a vector consisting of eigenvalues of Σ_k and mean() be the arithmetic mean operator, then

$$\mathrm{SNR_{dB}} = 10\log\left(\frac{(\mathrm{mean}(\sum_{1\le \mathrm{k}\le \mathrm{K}, \mathrm{k}\le \mathrm{l}\le \mathrm{K}} \mathrm{ds_{kl}}))^2}{\mathrm{mean}(\sum_{1\le \mathrm{k}\le \mathrm{K}} \mathrm{mean}(\mathrm{eig_k}))}\right)$$

We use SNR of 10 dB, for a 10 dimensional data. 150 observations are generated for both training and test sets. Figure 4.2 shows the first three principal components plotted against each other for data used for this work.

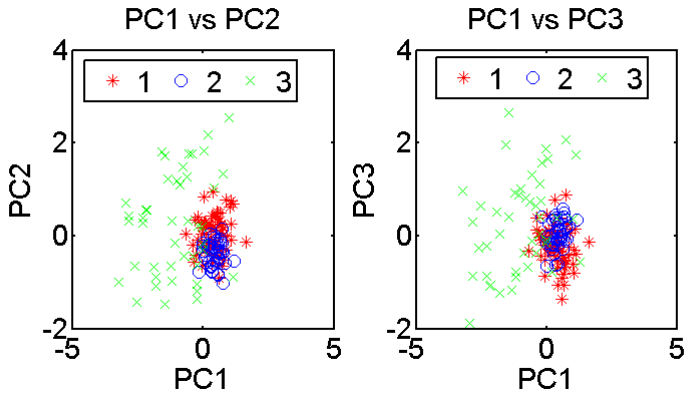


Figure 4.2: The principal components (PCs) plot for the data generated with 3 different classes.

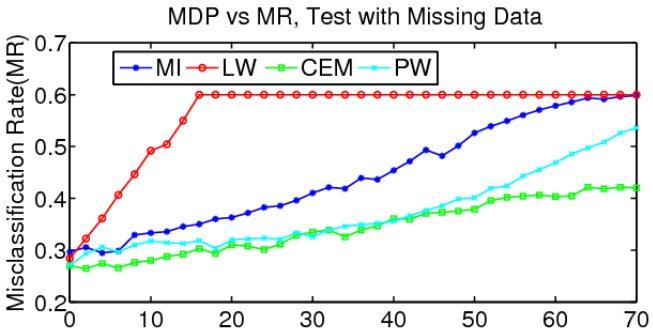


Figure 4.3: The results for synthetically generated data. Test set with missing (incomplete) data. MR plot against MDP.

The Figures 4.3, 4.4, 4.5 and 4.6 show the results for synthetic data generated. In case 1, CEM is the most efficient method, however, PW is competitive to it. CEM gives an efficiency of 40%, even at MDP of 70%. MI is clearly the worst method in terms of efficiency. The efficiency of MI decreases as MDP gets higher, while CEM and PW give more stable efficiency results. In case 2, results are similar and still CEM is the best. While MI performs better compared to case 1, CEM is slightly worse.

Iris Dataset

Iris dataset is one of the most commonly used datasets in machine learning literature. It consists of 3 classes of 50 instances each referring to a type of iris plant with 4 attributes. One class is linearly separable from the others; the other two are not linearly separable from each other.

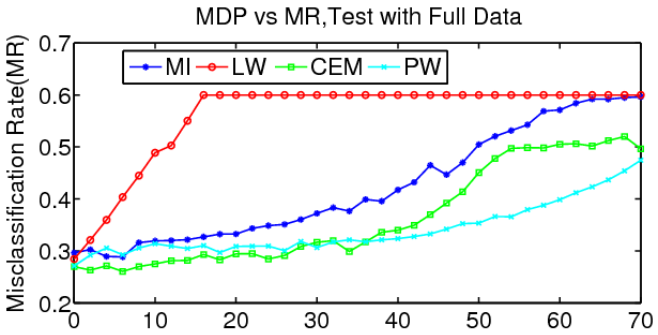


Figure 4.4: The results for synthetically generated data. Test set with full (complete) data. MR plot against MDP.

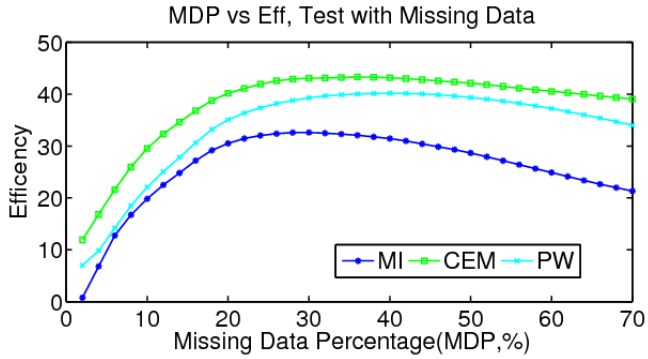


Figure 4.5: The results for synthetically generated data. Test set with missing (incomplete) data. Eff plot against MR

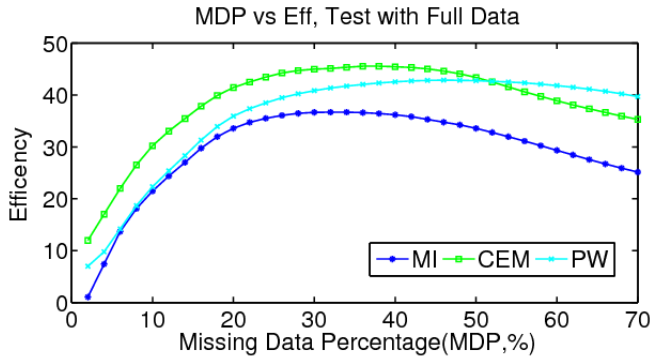


Figure 4.6: The results for synthetically generated data. Test set with full (complete) data. Eff plot against MR

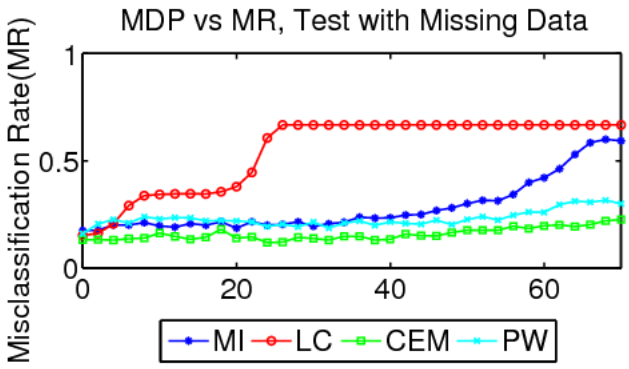


Figure 4.7: The results for Iris dataset. Test set with missing (incomplete) data. MR plot against MDP.

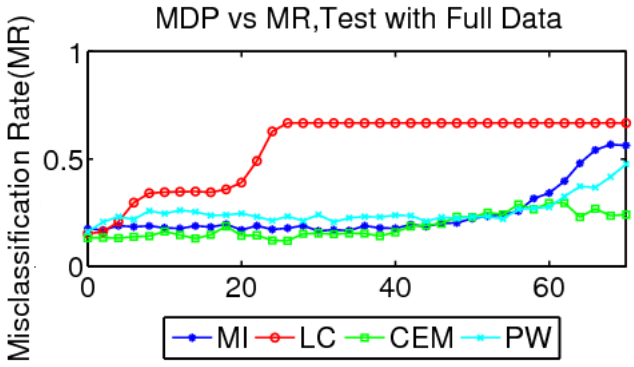


Figure 4.8: The results for Iris dataset. Test set with full (complete) data. MR plot against MDP.

We used 100 instances for train and 50 instances for test sets.
We show the results for this dataset in Figures 4.7, 4.8, 4.9 and 4.10. In case 1, CEM is still the most efficient method, MI and PW show a similar behaviour. CEM gives an efficiency of 70%, even at MDP of 70%. In case 2, PW is worse than MI and CEM is still the best method. Compared to case 1, the efficiency of CEM and PW is lower while the efficiency of MI is higher.

Pima Indians Diabetes Dataset

Pima Indians Diabetes Dataset contains 2 classes that are diabetes positive or negative with 7 attributes (for more details, see Section 3.3.2).
We use 200 instances for train and 200 instances for test sets.
The results are shown in Figures 4.11, 4.12, 4.13 and 4.14. Both in case 1 and case 2, CEM overcomes other two methods, whereas PW and MI give similar results.

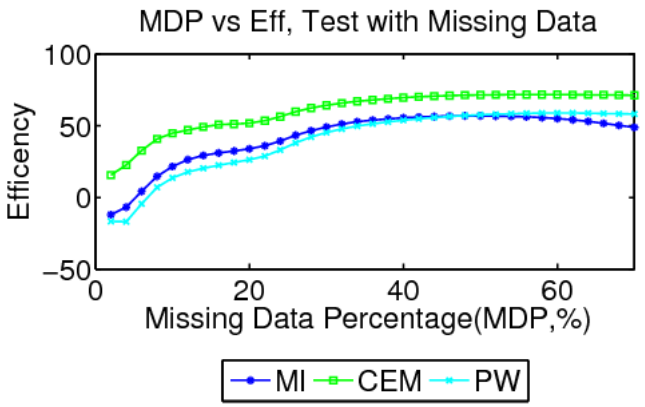


Figure 4.9: The results for Iris dataset. Test set with missing (incomplete) data. Eff plot against MR

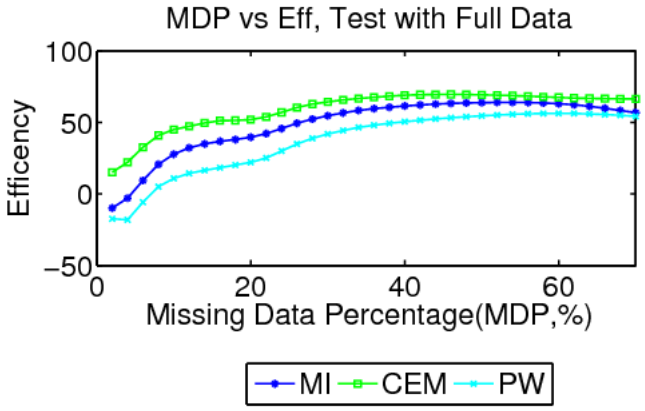


Figure 4.10: The results for Iris dataset. Test set with full (complete) data. Eff plot against MR

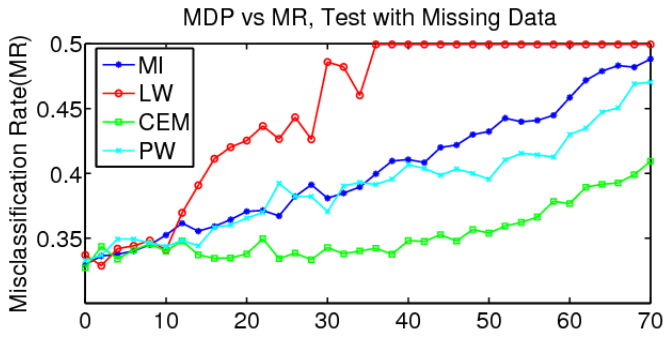


Figure 4.11: The results for Pima Indians Diabetes. Test set with missing (incomplete) data. MR plot against MDP.

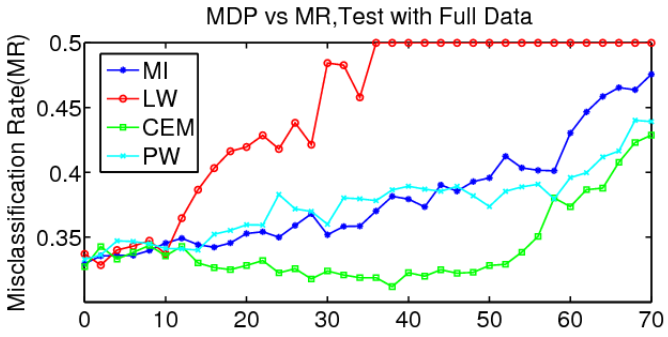


Figure 4.12: The results for Pima Indians Diabetes. Test set with full (complete) data. MR plot against MDP.

The efficiency of CEM at MDP of 70% is around 20%, not as high as other datasets, but still giving the highest performance.

General Discussion

We observe that, generally CEM is the most efficient missing data method, while PW is worse than CEM, but still slightly better than MI especially for high MDP values. The results coincide with previous work, such as the ones of Allison (2001)and Newman (2003). In Allison (2001), where, as we already mentioned in Section 4.3, they compare missing data methods using different criteria (the capability to minimize bias, maximize the use of available information and yield good estimates of uncertainty), ML methods are found to be the best. In Newman (2003), where they compare six different methods including PW and EM methods, the results again support ML approaches.
Although CEM and PW perform well for both cases we experimented, we observe that they are more efficient to use when test data set also has missing values. MI is

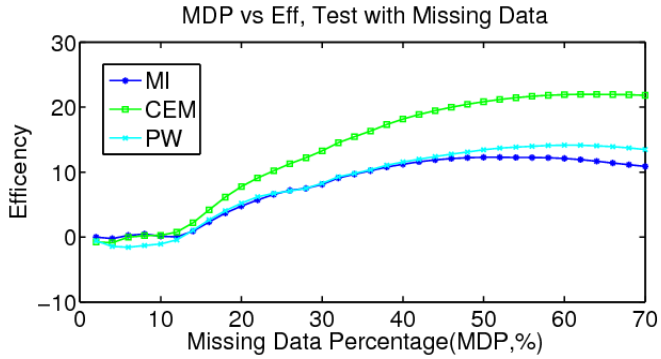


Figure 4.13: The results for Pima Indians Diabetes. Test set with missing (incomplete) data. Eff plot against MR

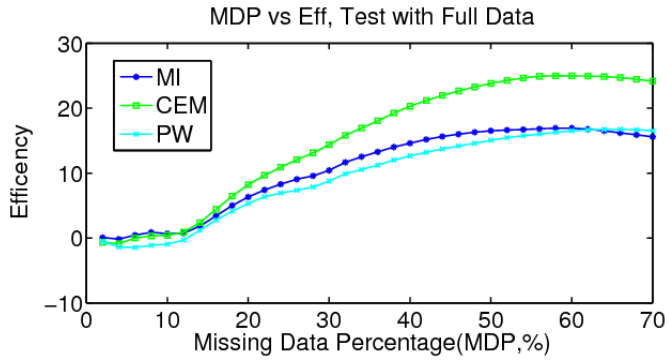


Figure 4.14: The results for Pima Indians Diabetes. Test set with full (complete) data. Eff plot against MR

more efficient to use when we have a complete test data set. Thus, MI is better at estimating the model, but the estimated model is not that robust to missing data in test set, and vice versa for CEM.
Another observation made from the results is that CEM and PW give more stable results for higher MDP values, so it would be more reliable to use them in situations where MDP for test set is undetermined. Although MI turned out to be the least efficient approach, it would still be acceptable to use it especially for low MDP values, since it is very easy to implement and clearly computationally less expensive.
In conclusion, even if CEM is the most efficient method, according to the definition of efficiency given in Eq. 4.1, there are situations in which the use of other techniques could more advantageous. The best algorithm has, then, to be chosen on the basis of the particular application and context.

Chapter 5

Top-Down Attention with Features Missing at Random

In this Chapter, we show the robustness of the top-down model exposed in Chapter 3 to missing data. In particular, we test the behaviour of the model on a classification problem in which features might be missing completely at random also in the training set to simulate a more realistic scenario.

The work has been already presented and published in Karadogan *et al.* (2011b).

5.1 Top-down attention model robustness to missing data

As we already mentioned in Chapter 3, the top-down model we propose works in an environment in which just some features are available and we want to know what is the next measurement to perform, among all those possible, for obtaining as much information as we can to execute a task. The situation is simulated on a classification problem, using a Gaussian mixture model in which the test dataset has some missing features. However, in that case, the training set we used for generating the model was complete. Now we test the behaviour of this attentive system in case of missing data also in the training set.
Considering the results obtained and discussed in the previous Chapter, we decided to use the maximum likelihood technique proposed in Ghahramani and Jordan (1994). This method, in fact, is the most efficient both in cases of complete and with missing data test sets.
The aim is to check the robustness of the model to missing data and, consequently, its behaviour in a more realistic scenario.
Since, we want to reduce the error rate of an ensuing decision classification problem making use of our attention model, we evaluate its robustness again by the misclassification rates on test data, after having trained the model with different amount of

incompleteness.
The analysis is performed on synthetic data with known distributional properties that conform with the model and the Yeast dataset (for more details, see section 3.3.2). The results obtained show that, exploiting our strategy, misclassification rate is still lower choosing the new measurement on the basis of the information value provided by the entropy computation, than randomly.

5.2 Experimental Evaluation

The evaluation of the model is implemented carrying out experiments in MATLAB. The Gaussian mixture model is trained and initialized as explained in Section 4.3.1 with a multi-start procedure with 5 different initializations, and 250 iterations. We first simulate an incomplete measurement situation on N_{train} and N_{test} data points. Data are missing completely at random with different missing data percentages. The randomization is done in such that not all values can be missing in one observation: one feature per observation is kept present randomly (see section 4.3.2). We test the performance either choosing the next feature to be measured with highest saliency or randomly and compare with classifiers, one trained with full data and the other with the original missing data we have.

5.2.1 Synthetic Dataset

Synthetic data are generated by a Gaussian mixture model. The number of mixtures K, is 3, and the number of dimensions D is 10. The difficulty of the problem is determined with the SNR calculation in Equation (4.3.2).
Figure 5.1 shows the illustration of our SNR calculation for a 2D data with 3 clusters. If we formulate that SNR calculation example according to the equation above,then we have:

$$\text{SNR}_{\text{dB}} = 10 \log \left(\frac{\text{ds}_{12}^2 + \text{ds}_{13}^2 + \text{ds}_{23}^2}{(\text{a}_1^2 + \text{b}_1^2)/2 + (\text{a}_2^2 + \text{b}_2^2)/2 + (\text{a}_3^2 + \text{b}_3^2)/2} \right)$$

We want to investigate how well the attention model performs for problems with different complexities. Thus, $N_{\text{train}} = 1500$ and $N_{\text{test}} = 300$ observations are generated for training and test sets respectively. For all situations, we train the model with missing data (MCAR).
We check the performance in the interval from 40% to 70%. The baseline error rate is 66% and the trained models with missing data have error rates between 2% and 15% based on complete test measurements. An SNR of 10 dB is used.
We want to investigate how well the attention model performs for problems with different complexities. Thus, we change the difficulty of the problem by generating data with different SNRs: the higher the SNR the easier the problem. Figure 5.2

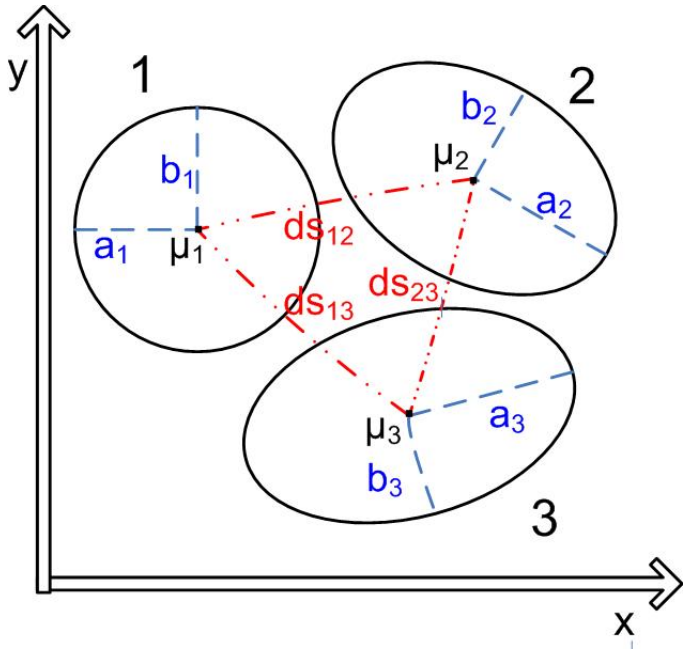


Figure 5.1: The illustration of SNR calculation for a 2D data with 3 clusters

shows the first three principal components plotted against each other for this dataset, for 15 dB and 5 dB of SNRs as an example.

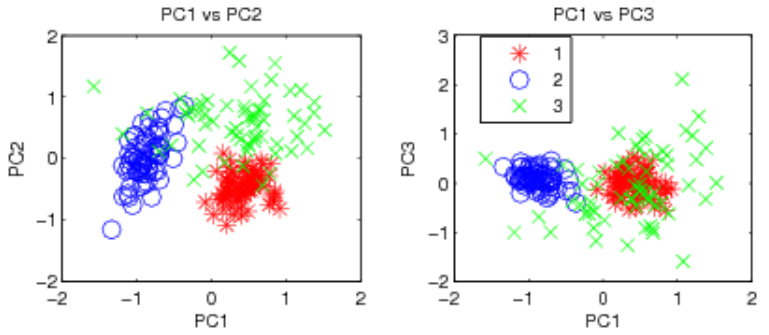
With the same goal, we design the data set such that all features are equally important. We measure the mutual information between each input feature and the class label. We useed a permutation test ($N_{\text{resamples}} = 200$) to test the mutual information against a null hypothesis of no mutual information. Mutual information is recorded if the null is rejected with $p > 0.01$. We found that all features are informative as expected (see Figure 5.3 (b)).

Figure 5.3 (c) shows the frequency of which features are selected with the attention model. Also in this case, characterized by having missing data in the training set, we observe that the frequencies are not simply given by the mutual information. This result underlines again the need for an attention model.

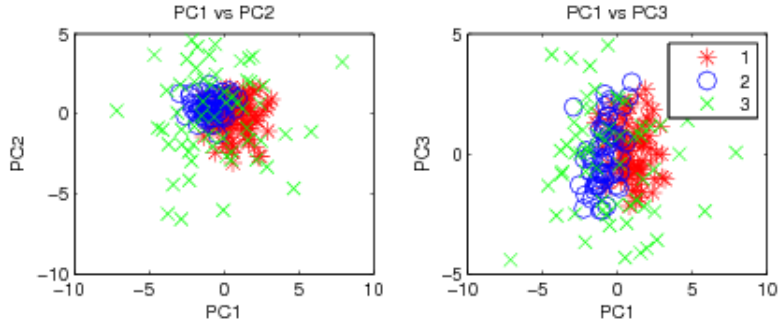
In addition, we observe that the feature saliency depends on the class label as well (see Figure 5.3 (d)).

Figure 5.3 (a) shows the error rates for the different methods for variable SNR. We tested with missing data and full data where all features are available. We compared the performances of those to the cases in which we test with missing data with one additional feature chosen randomly or using the attention model. As expected, we reduce the error rate adding one feature randomly or with attention model.

However, the top-down attention model outperforms the random saliency model for all MDPs being closer to the full data error rate.



(a)



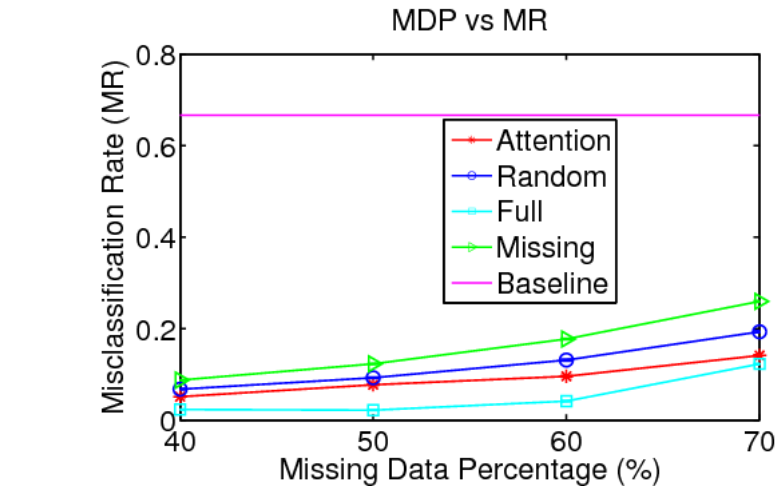
(b)

Figure 5.2: The principal components (PCs) plot for the data generated, with SNR of 15 dB (top) and 5 dB (bottom).

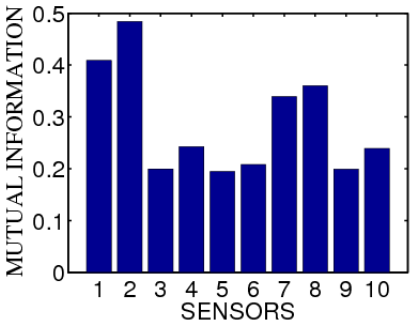
The results given by testing with different SNR values show that the attention model is less effective for more difficult problems, as illustrated in Figure 5.4.

5.2.2 The Yeast Dataset

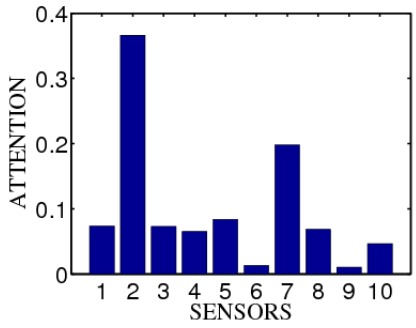
The Yeast data set used for the analysis of our top-down attention model concerns determination of protein cellular localization sites, as explained in Section 3.3.2.
We select a subset associated with two most frequent sequence types *CYT* (cytosolic/cytoskeletal 463 examples) and *NUC* (nuclear, 429 examples) in SWISS-PROT database reducing the classification problem to a binary decision. We have a training set with ($N_{\text{train}} = 630$) and a test set ($N_{\text{test}} = 262$) samples. We train the Gaussian Discrete model with $K = 11$.
The baseline error rate is 50% and the trained models with missing data give error rates around 40% based on complete test measurements. This is a noisier decision



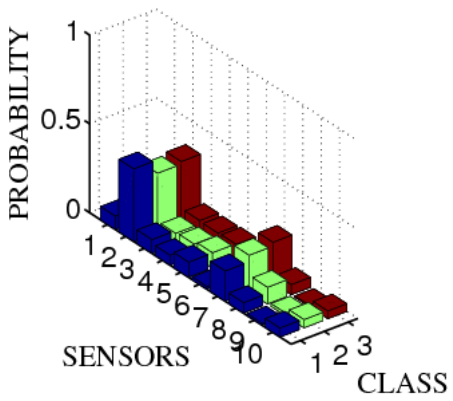
(a)



(b)



(c)



(d)

Figure 5.3: Synthetic data. Ten input features are considered. Train and test data are missing completely at random (MCAR). (a) For different missing data percentages, the misclassification error rates for the test set where data are MCAR (Missing), all missing features are added (Full), one random feature is added among missing (Random) and one feature is added among missing using the attention model (Attention). (b) The log_2 mutual information between features and class label. (c) Frequency of selection of additional features with attention model. (d) Frequency of selection of features within the three classes.

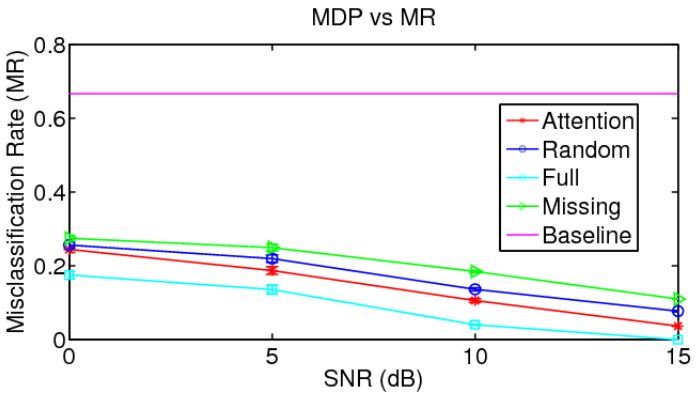
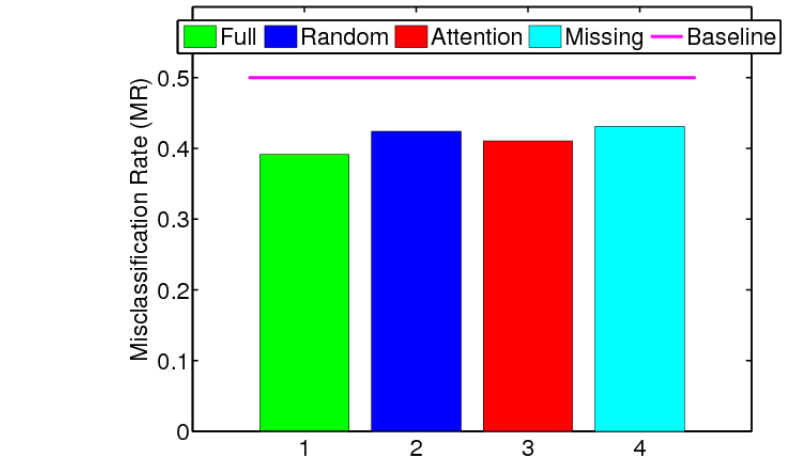


Figure 5.4: The result for syhthetic dataset created with different SNRs. The easier the problem (higher SNR), the stronger the effect of our attention modeling

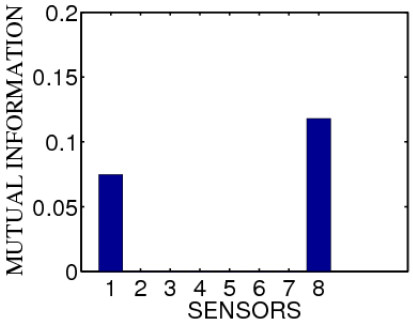
problem than the synthetic one.
Unlike the synthetic data, not all features are informative, while features (1,8) are the most informative ones (see Figure 5.5 (b)). Even if there are similarities between the mutual information and the selection of features with attention model (see Figure 5.5 (b) and (c)) (feature 8 is the most informative and the most frequently selected), we observe the features that are not informative are selected quite often as well. This result again underlines the need for an attention model that combine the available data and the task.
The saliency of the features depends on the class label like in synthetic data (see Figure 5.5 (d)). Figure 5.5 (a) shows the error rates for MDP of 50% (not all MDPs were convenient to use, too few or too much data problem) for the same situations as explained for synthetic data (data are MCAR, full, missing with added feature randomly or with attention model). The attention model performs better than choosing a random feature to be measured.

5.3 General Discussion

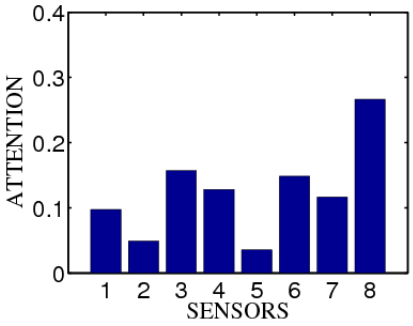
The results obtained showed that the top-down attention model we propose outperformes simple random attention, even with missing data in the training set. Moreover, the investigation about the dependency of the method to the complexity of the problem (problems with different SNRs and different missing data percentages), proved that the potential of the method is greatest for easy problems with higher missing data rate (the complexity of the problem should be analyzed to decide for the convenience of the use of the method).



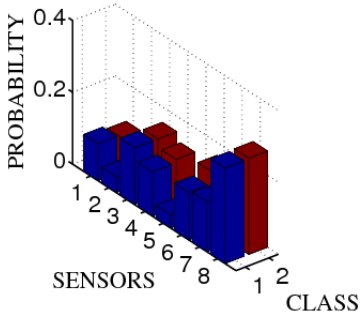
(a)



(b)



(c)



(d)

Figure 5.5: Yeast data. Eight input features are considered. Train and test data are missing completely at random (MCAR). (a) For 50% of missing data percentage , the misclassification error rates for the test set where data are MCAR (Missing), all missing features are added (Full), one random feature is added among missing (Random) and one feature is added among missing using the attention model (Attention). (b) The log_2 mutual information between features and class label. (c) Frequency of selection of additional features with attention model. (d) Frequency of selection of features within the two classes.

Part III

Cherry's Experiments Remake: the Role of Top-Down Attention

Chapter 6

Cocktail Party Problem

In this Chapter, we investigate the so called *Cocktail Party Problem*, analysing some of the computational techniques proposed so far to deal with it. We also face the human ability to operate in such a scenario and the various factors which can influence this ability.

The work has been already presented in Marchegiani *et al.* (2011).

6.1 Cocktail Party Problem and Source Separation

The cocktail party is an often used analogy in machine learning and signal processing, referring to the situation in which multiple signals are mixed and the aim is to separate these, or to recover at least one of the signals in the mixture. In many applications, in fact, the necessity to operate in environments without noise and in which it is possible to distinguish between the different signals present at the same time, represents, generally, one of the preliminary challenges to be confronted.
Many areas have to face this kind of problem: biomedical signals have to be extracted and refined before being used, the performance of telecommunication systems depend also on their ability of "cleaning" the transmitted information. Astronomical data and satellite images are investigated through a multispectral analysis, which considers each pixel value as a mixture of different sources (Nuzillard and Bijaoui (2000)). In the transcription of multi-speaker conferences, meetings, seminars and in dialogue systems of robots operating in complex acoustic scenes, the automatic speech recognition procedure needs to isolate the voice of interest within the other sounds and voices around and to track it, to be able to recognize the words pronounced, as shown e.g. by Choi *et al.* (2002).
In the last decades, several techniques have been introduced to handle these scenarios, but a complete and general solution is not available yet. Most of the general methods use low level statistical properties of the signals, such as independence, or other

simple distributional assumptions. Independent component analysis (ICA), which is one of most used methods in this context, for example, assumes the mixture to be a linear combination of the signals (Hyvarinen *et al.* (2001). Moreover, this method has other strong limitations, requiring particular constraints to be satisfied to execute the separation between the various sounds. For example, ICA needs as many recording channels as there are speakers (Stone (2004)).
For a more complete review about source separation methods, see Pedersen *et al.* (2007).

6.2 Masking and Human separation ability

Many studies show that, in the case of audio mixtures, human beings are very efficient cocktail party solvers, performing the source separation process quite easily, as demonstrated in the famous experiments carried out by Cherry (1953) almost sixty years ago.
Several investigations have been conducted over the years to discover which characteristics of the auditory scene could help the segregation process of a mixture of stimuli and in which way they can influence each other and the human ability to discriminate within the different signals. Almost all these experiments make use of pure tones (see Bregman (1994)), manipulated to test the subjects' reactions. But there are also cases in which human reactions are evaluated in situations with different people talking at the same time (see Bregman (1994), Moore and Gockel (2002) and Drullman and Bronkhorst (2000) among others).
In such a scenario, the separation between the sounds seems to be more complicated when the voices are of the same gender, as suggested by Drullman and Bronkhorst (2000), when they are emitted from close positions in the scene, when there is not enough difference in their fundamental frequency, in their phase spectrum or in their intensity, as described by Moore and Gockel (2002). Moreover, also the vocal tract size, accent or other prosodic features can change the complexity of the grouping of signals belonging to the same stream (coming from the same acoustic source), as demonstrated by Bregman (1994).
For a more complete review, see Bronkhorst (2000) and the more recent Bee and Micheyl (2008).

The studies performed by Bregman (1994), Cusack *et al.* (2004) and Carlyon *et al.* (2001) proved that the way in which stimuli are perceived as part of the same audio flow, and the proficiency in selecting just one of these flows and understanding its content, is widely affected by attention, both on a bottom-up and a top-down perspective. It should also be considered that, in fact, the segregation ability is a learned skill, that is improved by experience (top-down). But the same Bregman suggested that there are, in any case, physical factors, working on the more primitive auditory

processes, able, in particular conditions, strongly influence the behaviour of the attentive mechanisms (bottom-up).

Brungart (2001) investigated how masking [1] effects operate in the perception of multiple talkers and made a distinction between *energetic* and *informational masking*:*"Traditional energetic masking occurs when both utterances contain energy in the same critical bands at the same time and portions of one or both of the speech signals are rendered inaudible at the periphery. Higher-level informational masking occurs when the signal and masker are both audible but the listener is unable to disentangle the elements of the target signal from a similar-sounding distracter"*.
Both kinds of masking act simultaneously. Thanks to his experiments, he demonstrated that human ability to discriminate between the contemporaneous speech of two talkers is more influenced by informational masking than energetic masking. Thus, the voice characteristics are crucial for the segregation procedure.

[1] In psychoacoustics masking refers to the effect that one signal prohibits the other from being detected.

Chapter 7

The effect of a task in the Cocktail Party Problem

In this Chapter, we show the effect of a task in the Cocktail Party Problem. In order to do this, we test the behaviour of our top-down attention model, illustrated in Chapter 3, in such a scenario. Moreover, we discuss the results provided by behavioural experiments, which are designed revisiting the ones performed by Cherry (1953). The aim of the experiments is to investigate how human auditory perception and, in particular, speech intelligibility can be affected by the presence of a task in a cocktail party simulation.

The work has been already presented in Marchegiani *et al.* (2011).

7.1 Weak and Strong Top-Down Attention

In order to analyse the way in which the top-down attention model proposed in Chapter 3 works in presence of a task and the way in which its behaviour could be, consequently, influenced by the task itself, we introduce the concept of *weak* and *strong* top-down attention. By *weak* it is meant that the $p(y|k)$ table (see Equation (3.6)) has been smoothed and that the attention mechanism is stochastic determined by the top-down mechanism; the *strong* one corresponds to the original model. The first case represents the absence of the task, the second one the opposite situation.
To simulate strong and weak top-down attention we augment the model by smoothing the label-component table

$$p(y|k) \rightarrow p(y|k,\beta) = \frac{p(y|k)^{\beta}}{\sum_c p(y|k)^{\beta}}$$

and by letting the attention selection be stochastic based on the expected gains, i.e., we select attention using the induced probability distribution:

$$P(j) = \frac{\exp(\gamma G_j)}{\sum_j \exp(\gamma G_{j'})} \tag{7.1}$$

Task-driven top-down attention as in Hansen *et al.* (2011) is obtained when $\beta = 1, \gamma = \infty$. In this work we use $\beta = 0.2, \gamma = 0.33$.

The aim is to use our top-down attention model to make predictions about the influence of confounders on task labeling performance, in both strongly task dependent attention and under weakened task influence.

7.1.1 Experimental Evaluation

We challenge the strong and weak top-down attention models by a simulated cocktail party scenario by confounding the input of test data for a $C = 2$ environment. In other words, we simulate a situation in which a signal is masked by another overlapping signal.
In particular, we mix for each pattern a fraction f (mixing fraction) of the input features with a randomly chosen input feature vector from the opposite class (confounder):

$$overlapped\ signal = (1 - f) * input\ signal + f * confounder$$

The $C = 2$ simulated decision problem is based on a four component Gaussian mixture. The resulting configuration is first established in two dimensions and resembles the well-known XOR-problem, hence can not be separated linearly. The two dimensional input space is augmented by six noise dimensions $SNR \approx 1$, so that the total input dimension is eight. In the attention experiments one signal dimension and one noise dimension is provided as 'gist' (Torralba *et al.* (2006)).
For a range of mixing fractions $f \in [0.0\ 0.2]$, we measure the resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after having provided the 2-dimensional gist.
The strong and weak top-down attention response is shown in figure 7.1. The rates are represented as relative excess errors:

$$\frac{E(f) - E(0)}{B - E(0)}$$

where $B = 0.5$ is the baseline error rate. The error rate of the top-down attention model is $E_{\mathrm{DIR}}(0) = 0.08$ while the error rate of the weak attention is $E_{\mathrm{UNDIR}}(0) = 0.23$.
We use *DIR* to indicate the strong attention, because we refer to the case of *Directed*

attention, in which attention is addressed by a task. We use *UNDIR* to indicate the weak attention, referring to the opposite case: *Undirected* attention.
The experiment indicates that strong top-down attention model (DIR) is less sensitive to the confounding mixture than the weak attention model, hence it will make more informed decisions in the cocktail party.

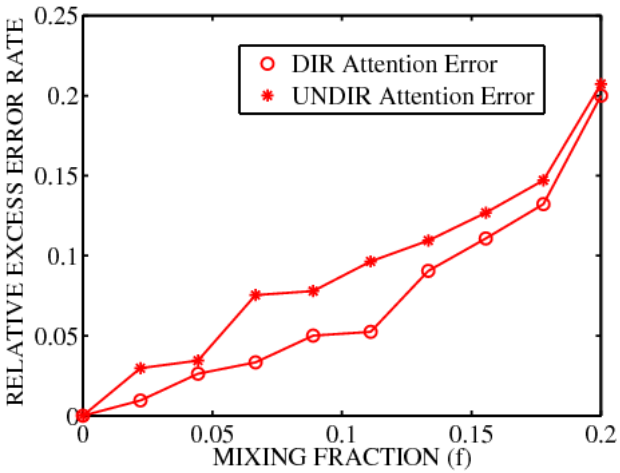


Figure 7.1: The resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after having provided the 2-dimensional gist for a range of mixing fractions $f \in [0.0\ 0.2]$.

7.2 Cherry's Experiments Remake

While it is not possible to directly read out the information flows in the human brain while solving a difficult speech separation task, some insight can be obtained by observing the macroscopic behaviour. Thus, in order to better understand the influence of top-down attention cues on human auditory perception in a cocktail party problem and how these cues may modulate informational masking effects, we designed behavioural experiments inspired by the pioneering work of Cherry (1953). We return to his basic experimental setup, i.e., a listening experiment in which we analyse participants' ability to hear individual naturally sounding speech signals in a mixture with reduced spatial and speaker cues.
Basically, we design a hard cocktail party problem by reducing conventional auditory cues, as described above, leaving only high-level cognitive cues, such as semantic and context representation in narratives. Cherry alluded to these high-level representations as what he called word *transition probabilities*.
Our hypothesis is that these representations precisely are subject to top-down atten-

tion, while the more basic cues could be more automatic and operate beyond conscious control.
In particular, in order to simulate a cocktail party scenario,we present the audio signal to the listener as a *monaural mixture* of two different narratives uttered by a *speech synthesizer* (TTS projects at AT&T Labs - Research, see Beutnagel *et al.* (1999)), using the same virtual speaker. In this way, there are no cues to separation due to a different spatial location of the sound sources, to a different accent or to a different gender of the speaker.
The speech synthesizer we used produces prosodic speech which also provides a cue to stream separation and tracking, as exposed in Syrdal and Kim (2008). Moreover, Jilka *et al.* (2003) illustrated that the introduction of this kind of voice modifications affects the quality of the TTS and the perception of the sound it generates. However, we expect these differences to be much less incisive and forceful compared to conventional human speech.
Same considerations can be made about the loudness of the two sounds. We checked the energies of the signals and, as expected, even if there are little changes word by word, totally there is not much difference between them and the effect can be disregarded.
For the sake of reducing basic semantic masking effects we opt for narratives with little expected interest to the listeners. Specifically, we chose excerpts from neutral texts used in preparation for the TOEFL (Test of English as a Foreign Language) test in Phillips (2006). These texts are more coherent and naturally sounding than the short command sentences exploited in Brungart (2001).
We use two different pairs of stories for each kind of experiment, making little changes (removing or adding sentences or words, switching the order of some sentences or words) to the original ones. This is necessary to have stories with the same length and in which pauses are synchronized as much as possible, to avoid the so called "*listening in the gaps*" effect, described by Bregman (1994) and by Bronkhorst (2000). If there is a pause just in one of the stories, in fact, the listener takes advantage of it, giving all his/her attention to the other one and then, even if he/she wanted to switch to the other story, it would need time and it could make the subject miss some words, in any case. This happens, because even short gaps or short changes of attention could influence the perceptual grouping of a set of signals, resetting all the process, as Cusack *et al.* (2004) proved in their investigations.

7.2.1 The effect of a task on speech intelligibility

Following the design illustrated in the previous section and in analogy to the simulation discussed in Section 7.1, we performed two different kinds of experiments to explore the impact of energetic masking on the speech intelligibility of a complex audio signal.
In the first kind of experiments (we call it *undirected attention experiments (UNDIR)*),

the subjects do not have any task (they can follow any of the stories), while, in the second case (*directed attention experiments (DIR)*), they are asked to follow just one of the two stories (the choice about which of them is left to the subject), so that it could be possible to understand how attentive mechanisms (paying attention to just one of the stories) can behave and interfere with the pure energetic masking.
We use two different narratives for each experiment. The relative texts are reported below:

First Story: *The giant panda still lives in the wild in only a few mountain ranges in the southwestern part of China because its survival has been threatened. both by hunters and by the destruction of the habitat it needs to survive. What has been noted and stressed in the last few decades is that the pandas survival is also threatened by the reproduction cycles of the bamboo where the pandas live. Here's what the problem is. It is the main source of food for the giant panda. However, when there's a massive reproduction of the bamboo, the one that has just reproduced dies, so there's a lag of quite a few years before the new, young seedlings grow enough to provide food. If the bamboo where the giant pandas living dies, then the giant panda needs to move to new areas to find food. The search for food has led the giant panda into areas, that are more settled and more full of danger.*

Second Story: *Conifers are hardy trees that have been able to live long, so, as a result, both the oldest and the biggest trees in the world belong to the conifer family. The oldest known tree is located in east California. That tree is a four thousand years old bristlecone pine. The giant redwoods, in California, are the largest and oldest trees; they can be several hundred feet tall with a weight of two thousand tons. An interesting note about the giant redwoods is that, even though the trees are so big and tall, they have small cones. They are evergreens with short and spiky leaves. The needle-like shape of Conifer leaves evolved as a reaction to drought and aridity. When compared with a flat leaf, a needle presents a much smaller surface area. Most conifers are evergreens. They lose and replace their needles throughout the year, rather than shedding all their leaves in one season.*

In both kinds of experiments, in order to understand and measure what and how much of the narratives subjects could hear, we present them a list of terms and ask to check which terms they have heard. The list contains words which are actually in the stories and words that are not, but that are related to the content of the stories (see Table 7.1). In this way, we can understand if they report what they really have heard or if they try to guess, after getting the topics.
The words in the list can appear various numbers of times in the narratives, but we tried to balance this number, in total, for both tracks. In particular, the total number

of occurrences of all the words we ask for is the same in both stories. Moreover, we aimed to balance the frequency of each word in the list; which means, for example, that if there are two words in the list, appearing respectively, three and five times in the first story, there are also two words appearing three and five times in the second one.
The list of words contains 48 words: 24 of these are truly present in the audio signal and each story contains the half of them. We do not put in the list words able to attract attention, due to the particular way in which the speech synthesizer pronounces them. In the UNDIR case, we analyse just 22 of the 24 words, since 2 are replaced by new words in different trials to improve the balance of the stories.

List of the words		
Agriculture	Chlorophyl	Food
Fiber	Species	Rare
Die	Oldest	Reproduce
Family	Wood	Fur
River	Zoo	Bamboo
Ecological	Threaten	Conifer
Niddle	Stem	Green
Destruction	Northern	Farming
Attack	Tall	Root
Survive	Shape	Big
Problem	Source	Extinct
Asia	Flower	Mountain
Pollen	Fungi	Tree
Hundred	Eyes	Bear
Plant	California	Vegetable
Black	China	Panda

Table 7.1: List of the words used to check which story grabs more subjects' attention at particular moments. The words in green are the words truly present in the story; the ones in red are not in any of the narratives.
The order according to which the words appear in the list is randomly generated.

Cherry made his participants listen as many times as they wanted; we performed three different trials (4 people for each trial). In the first and in the second one, we made our subjects listen just once and then we asked for the words and they repeatead the same process three times. In the third one, we made them listen twice and we asked for the words; afterwards, we made them listen once more and we asked for the words. However, we did not write in the instructions anything about what we would ask them to do, we just said that, in the end, there would be questions about what they had just listened to.

Our subjects were twelve people among whom were master students, PhD students and post-docs from the Technical University of Denmark. Two participants were not able to accomplish any of the experiments and were excluded.
The behavioural experiments were carried out using a GUI implemented in Java, while the results were analysed using MATLAB.

7.2.2 Temporal and Spectral Overlap

To test the relative influence of top-down and bottom-up information flow on attention and masking (and, consequently, speech intelligibility) we give a new definition of overlap, based on the so-called ideal binary mask (IBM), which has been attributed as the goal of computational auditory scene analysis (CASA) that studies perceptual audition (see Wang (2005)). An IBM consists of zeros and ones where ones represent the powerful parts of the target audio signal compared to an interference audio signal. Being attributed as the goal of CASA is not the only reason making IBMs a reasonable tool for us to compute overlaps. IBMs have also been shown to improve human speech intelligibility when applied to noisy speech signals. The subjects have been exposed to the resynthesized speech signals from the IBM gated (segregated) signal and they recognized words quite well even for a signal-to-noise-ratio (SNR) as low as -60 dB which corresponds to pure noise (see Wang *et al.* (2008) and Kjems *et al.* (2009)). In addition,the features obtained from IBMs have worked successfully for an automatic speech recognition (ASR) application by Karadogan *et al.* (2010).

An ideal binary mask is obtained by comparing the spectrogram of a target sound signal to that of the interference signal and to keep only the strong time-frequency regions of it. More specifically, its value is one where the target is stronger than the interference for a local criterion (LC), and zero elsewhere. The timefrequency (T-F) representation is obtained by using the model of the human cochlea as the basis for data representation (see Lyon (1982)).
If $T(t, f)$ and $I(t, f)$ denote the target and interference time-frequency magnitude, then the IBM is defined as

$$\mathrm{IBM}(t, f) = \begin{cases} 1, & \text{if } T(t, f) - I(t, f) > LC \\ 0, & \text{otherwise} \end{cases} \tag{7.2}$$

In Figure 7.2, we show an example of an IBM obtained with a sample sound from one of the stories (the sound relative to the word 'navigate' of the monaural mixture already described) compared to a speech shaped noise (SSN) as the interference signal. The most energetic parts of the target signal are kept.
We measure the spectral and temporal overlap between two sound signals, specifically between a word in one of the narratives pronounced by the speech synthesizer and the corresponding part in the second story.
We define the temporal overlap between them as a percentage of the whole duration

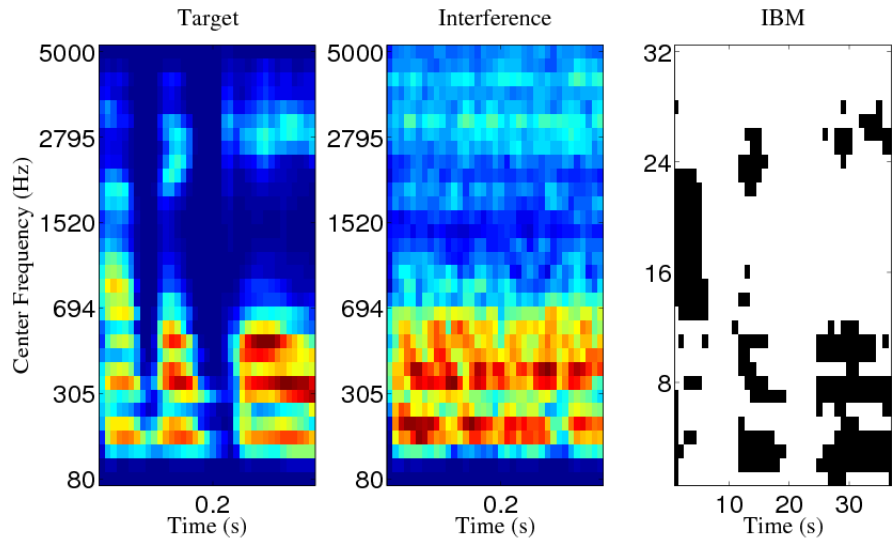


Figure 7.2: The spectrograms of a target sound signal, the interference signal (SSN) and the resultant IBM (SNR=0 dB, LC=-4 dB, windows length=20 ms (50% overlap), frequency channels=32, frequency bins are not equally spaced, gammatone filtering is used, 1 = black 0 = white).

without silence in the time domain. We use IBMs of the sound signals as mentioned before and we compress both IBMs over frequency. For a time slot, we assign one if there is at least one value one on the mask, otherwise zero where zeros are considered as silence. Then, the temporal overlap is simply the overlap of ones on the masks (see Figure 7.3).
The spectral overlap is defined similarly based on co-occurrence of signals in the time-frequency bins. Once we have IBMs for both sound signals, we simply compute the percentage of the overlapping ones on both masks over the whole time-frequency bins (see Figure 7.3).

7.2.3 Overlap and Speech Intelligibility

The influence of masking is measured as the correlation between the number of times specific words are heard (WOH) for both directed and undirected cases and the relative overlap. IBM control parameters were first chosen taking as references previous works on speech intelligibility, like the ones performed by Kjems *et al.* (2009)) and Karadogan *et al.* (2010), as a pre-analysis step. In the work of Kjems et al., subjects listened to IBM-gated (multiplying the spectrogram of a noisy-speech with IBM of it, and resynthesizing in time domain) and for a range of IBM parameter values, the best speech intelligibility results (recognizing which word is pronounced) were obtained. In the one of Karadogan et. al., the best performance for an ASR system was

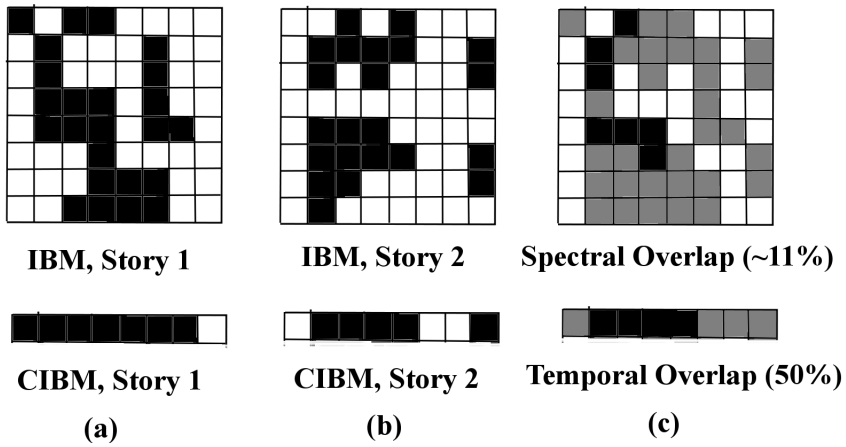


Figure 7.3: The illustrations of temporal and spectral overlap definitions, the bins represent time-frequency regions of an IBM (frequency bins are not equally spaced, gammatone filtering is used). Only black regions represent overlapped parts on (c).

obtained again with same range of values.
However, referenced to those studies, even if those parameters are expected to result in overlaps closer to what humans perceive, they are not necessarily optimal to investigate the correlation between overlap and word detection rates. Therefore we optimized the IBM parameters including the *local criteria (LC)* with fixed SNR, the *windows length* (winLength) and the *number of frequency channels* (numChan) to gain the most negative correlation.
We kept other two parameters constant while optimizing one.
With the optimized values, we applied a permutation test with 10000 resamples, at 5% significance level, to validate the results.

The word boundaries were determined manually to be more precise (The limited number of words enabled us to do so). The sampling frequency of the audio signals is 16 kHz. We use gammatone filters which is a commonly used model for auditory filters in the auditory system to obtain IBMs (Johannesma (1972) and De Boer and De Jongh (1978) and De Boer and Kruidenier (1990)).
Figures 7.4 and 7.5 show the temporal and spectral overlaps for each word, for UNDIR and DIR cases respectively, using non-optimized parameters from Kjems *et al.* (2009) and Karadogan *et al.* (2010). We observe that the correlation between overlaps (temporal and spectral) and rate of heard words are -0.35 and -0.31 respectively, in the UNDIR case. While, for DIR case, we find positive correlations of 0.23 for temporal and 0.34 for spectral.

We next optimized the three IBM parameters, LC, WinLength and NumChan, to

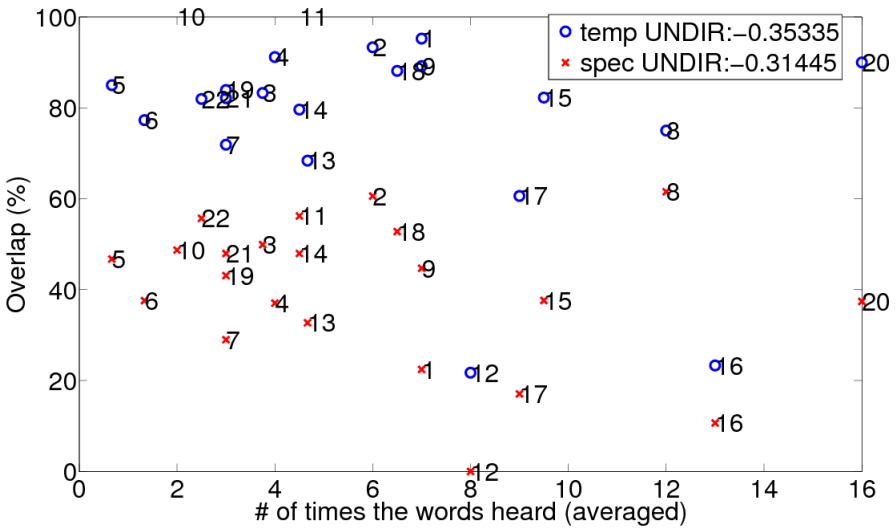


Figure 7.4: Temporal and spectral overlap versus the averaged number of times the words heard (WOH) for UNDIR case, and the correlation between them shown on the legend.

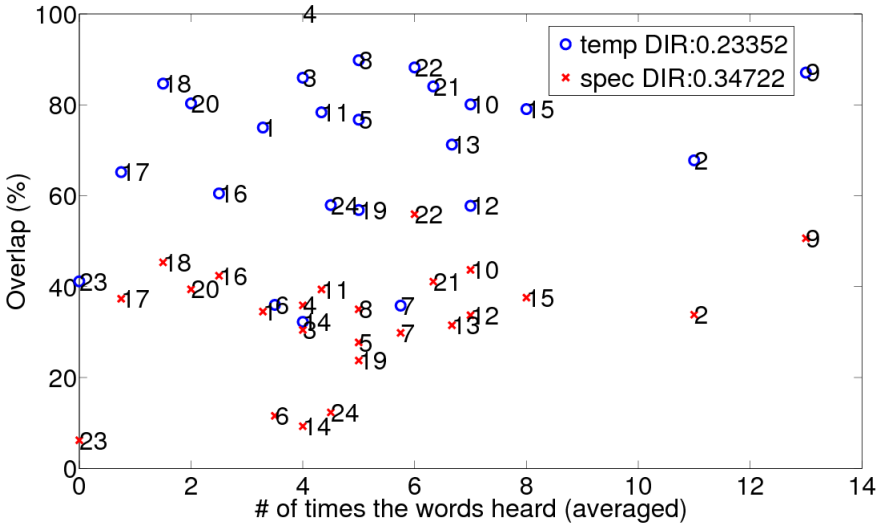


Figure 7.5: Temporal and spectral overlap versus the averaged number of times the words (WOH) heard for DIR case, and the correlation between them shown on the legend.

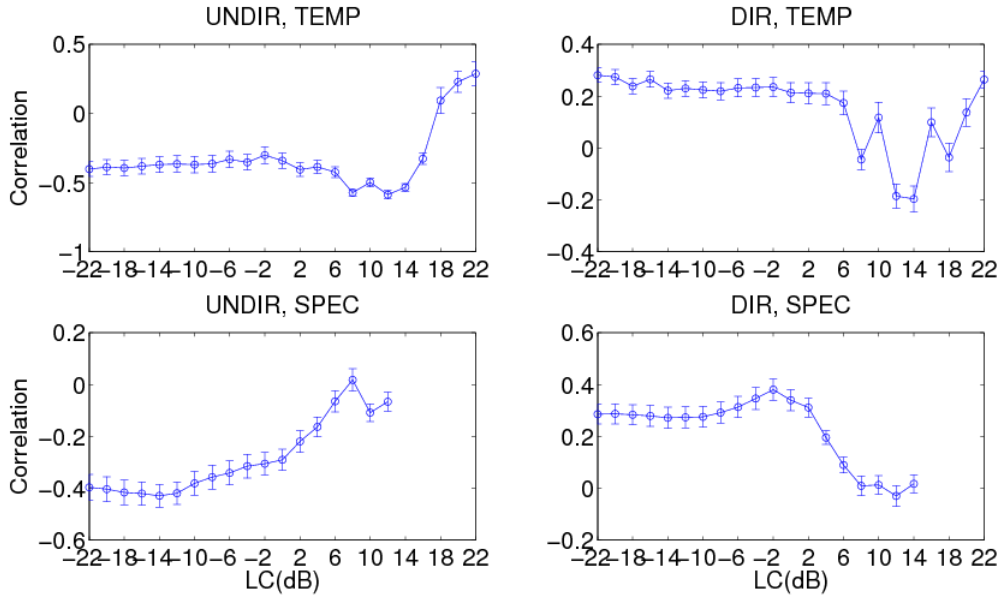


Figure 7.6: Correlation for different LC values, WinLength = 20ms and NumChan=32. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.

produce the most negative correlation between overlaps and words detected. The resulting figures show that optimal LC values are around -10dB for all cases except for the spectral overlap in the UNDIR case, which is -14dB (see Figure 7.6). We also conclude that 20ms is the optimal value for the windows length for all cases (see Figure 7.7). We see that for the spectral overlap in the DIR case, the correlation values for WinLength greater than 20ms are not present. This is due to the fact that with the high optimal value found previously, the resultant IBMs were all zeros (we did not try to play with the values, because it was already hard to find significant results for DIR case). Finally, we observe that the optimal values for number of frequency channels is 16 and 32 (see Figure 7.8).
Using optimal IBM parameters for each case (UNDIR,DIR, temporal and spectral) we obtain similar results. The correlations between overlaps (temporal and spectral) and WOH are -0.59 and -0.43 (more negative than not-optimized case) respectively for UNDIR case. However, for DIR case, they are -0.20 for temporal and -0.03 for spectral. Even if the results for DIR case also are more negative than not-optimized case , they are evidently less than those of UNDIR case (almost no correlation for DIR spectral case).

Finally, we applied a permutation test to these data, as we already mentioned. In both spectral and temporal overlaps, for UNDIR experiments, under the 5% signifi-

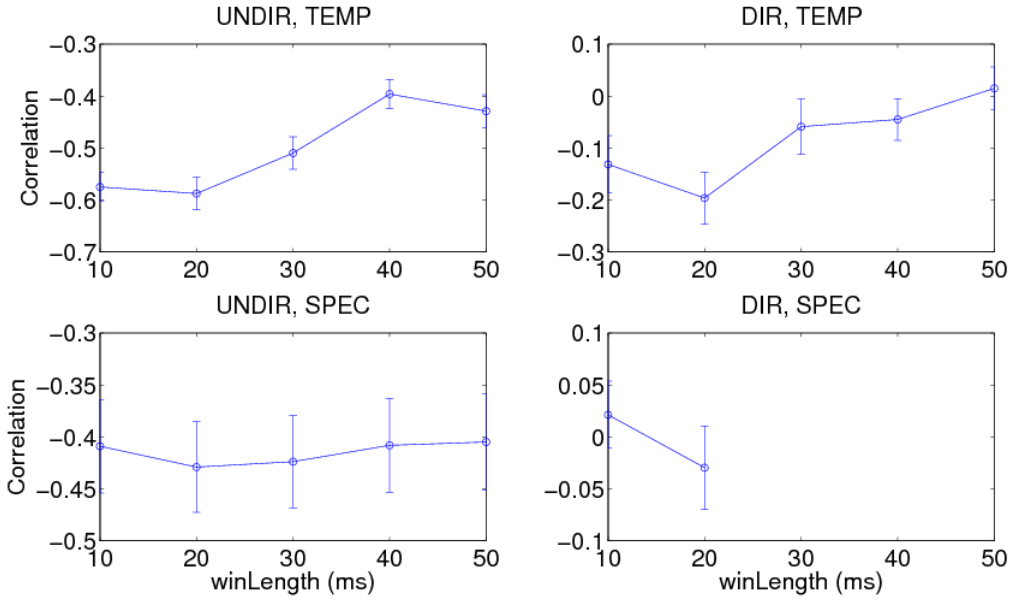


Figure 7.7: Correlation for different WinLength values, optimal LC for each case and NumChan = 32. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.)

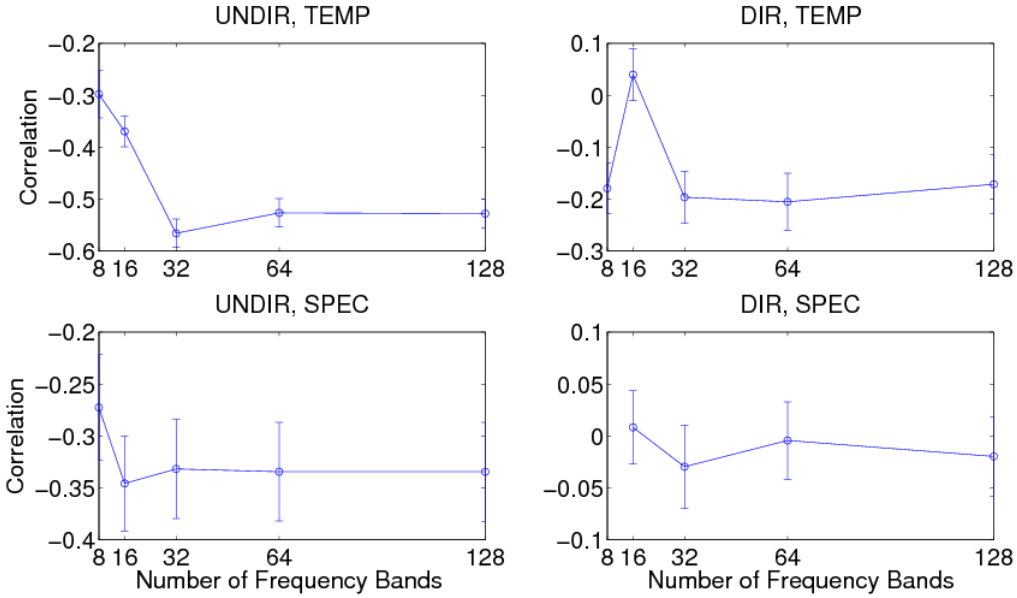


Figure 7.8: Correlation for different NumChan values, optimal LC and WinLength for each case. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.)

cance level, the null hypothesis, that the data is uncorrelated, is rejected (spectral = 3.1% and temporal = 0.07%). In the DIR experiments the null hypothesis is accepted, indicating the influence of masking is compensated by a more detailed model.

7.2.4 General Discussion

Based on our recent top-down attention model we can simulate the cocktail party effect. We found that the top-down attention model shows less sensitivity to the amount of the confounding overlap, than the weak attention model. This indicates that the top-down mechanism can assist to compensate for structured noise.
In the 'hard cocktail party' behavioural experiments, we found significant negative correlations between overlaps of two concurrent sounds and speech intelligibility for the data collected in the undirected attention experiments (UNDIR, no task). While in the directed attention experiments (DIR, task-driven) we accepted the no-correlation null hypothesis, even after careful optimization for correlation, a finding well-aligned with the simulation result.
We conclude that the relation between energetic masking and speech intelligibility is modulated by the presence of a task. Based on our top-down attention model, we expect this to be a special case of a more general phenomenon, namely that the top-down knowledge can enhance pattern recognition by compensating for noise and the presence of confounders.

Chapter 8

The effect of priming in the Cocktail Party Problem

In this Chapter, we illustrate some experiments, with a close design to the ones described in the previous Chapter. The experiments are carried out to analyse the role of priming on attentive mechanisms in a cocktail party scenario.

8.1 Priming

As shown in some of the experiments described in Chapter 3, a sensorial stimulus could influence the perception of other stimuli. In cognitive science, this effect is called *priming*.
More formally, if a stimulus (*prime*) precedes another one (*probe*) and the elaboration of the second one becomes easier thanks to the elaboration of the first one, this effect is called *positive priming*. If, instead, the elaboration of the first stimulus makes the second elaboration harder, this effect is called *negative priming*.
This phenomenon can take place even without being aware of it. For instance, in the investigations performed by MacKay (1973), the words in the supposed neglected ear determined the meaning of the sentence in the attended one. A similar experiment was exposed by Friederici *et al.* (1999), who proved that word recognition can be facilitated by the presence of a previous word with a close meaning. For example,the word "doctor" would be more easily identified if the word before this was "nurse" or something similar to this. In both these cases, the priming acts semantically.
The *Stroop effect* (Stroop (1992)), instead, can be considered an example of negative priming. The stimulus given by the meaning of the word, makes slower the identification of the color, even operating unconsciously.
Priming can work even between different modalities. Hernandez *et al.* (1996) analysed spanish-english bilinguals behaviour, using auditory text with visual target words.

8.2 Effect of Priming in a Cocktail Party Problem

We investigated the effect of priming, as a top-down cue, in a Cocktail Party scenario. We want to test if previous knowledge about one of the topics of the simultaneous speeches can affect the distribution of attention. Our hypothesis is that some indications about the content of one of the streams should help the separation between the different streams themselves, making it easy to follow the one characterized by a known topic and keeping the attention on it. In this way, in fact, even if attention could be grabbed by particular words coming from other sources, recovering and going back to attend the previous speech, should still be easier.
In order to do this, we carried out experiments close to the ones described in the Chapter 7.2.4, drawing inspiration from the ones executed by Cherry (1953). In particular, the design is the same as exposed in Section 7.2. We make subjects listen to a monaural mixture of two narratives, pronounced by a speech synthesizer (Beutnagel *et al.* (1999)), using the same virtual speaker to eliminate eventual speaker and spatial cues (including the energies of the two speeches). The stories are modified versions of texts coming from the same book (Phillips (2006)). Also for these experiments, changes are made to the narratives to avoid the "listening in the gap effect" (see section 7.2).

In this case, we performed two different kinds of experiments under the UNDIR conditions (cf. Section 7.2.1): *Unprimed experiments (UPex)* and *Primed Experiments (Pex)*.
Our subjects were twelve people among whom were master students, PhD students and post-docs from the Technical University of Denmark. We made half of them listen to the combination of two stories, without any priming (Unprimed Experiments) and we primed the other half by some preliminary information about one of the narratives (Primed Experiments). In particular, we made them read and then listened to a short introduction about one of the stories and only afterwards listen to the monaural mixture.
The text of narratives (used in both cases) and of the priming are reported below:

Priming for the first story: *This passage is about how animals like groundhogs and hedgehogs handle winter and low temperatures. Some animals move south to warmer weather, that is safer. Some animals increase their activity to stay warm, and other animals like groundhogs and hedgehogs hibernate during the cold weather.*

First Story: *Groundhogs, hedgehogs and bears go into a state of unconsciousness or semiconsciousness during the cold winter months. The groundhog is one of the best-known hibernators. It goes into its tunnel four or five feet underground and it does not come out until spring. A groundhog or a hedgehog stays in its underground tunnel for the entire cold winter. Because the groundhog hiber-*

nates so completely, it has achieved prominence in our folklore as the animal that's responsible for determining winter is over, warm weather and it is safe to come out of hibernation. If cold winter is over the groundhog will come out of its tunnel feeling safe, but if winter is going to last for more time the groundhog will run back into its tunnel. Other animals like bears and squirrels hibernate in a similar fashion.

Second Story: *The Nile river flows north from the equator to empty into the Mediterranean and irrigates more than a million acres of land. Asia also has a massive river system. The Yangtze River is Asia's longest at three thousand four hundred and thirty six miles of length. Because the mountains at its source are at such a high altitude, the Yangtze flows more rapidly than other major rivers for most of its length. The Amazon River in South America is the world's second longest river, but It carries more water than any other. It is slightly shorter than the Nile. The Mississippi River is the best-known river system in North America. It is the United States chief inland waterway. However, it is not the longest river in North America; because the Missouri River is slightly longer than the Mississipi.*

Also in this case, we present them a list of words asking them to report the ones they have heard. In this way, it is possible to understand if, after priming, people prefer the first story (the one they are primed on). Again the list consists of 48 words: one half of these are really present in the stories, the others are not, but they are related to the content (see Table 8.1). Aagain, we did not put in the list words characterized by a particular pronunciation. However, none of them are present in the priming.
Again we tried to balance the number of times the words appeared in the stories, following the same criteria as discussed in the previous Chapter (cf. section 7.2.1).
As in third trial of the experiments described in the previous Chapter, we made subjects listen twice and then asked for the words (first session); afterwards, they listened once more and we asked for the words (second session). Again, we did not write in the instructions anything about what we would ask them to do, we just say that, in the end, there would be questions about what they hade just listened to.
The behavioural experiments are carried out using a GUI implemented in Java, while the results are analysed using MATLAB.

8.2.1 Preliminary analysis of the results

So far we could just perform a preliminary analysis of the results, which seems to confirm our hypothesis.
In Figures 8.1 and 8.2 the number of times words from each story were heard by each subject is shown, in the first and in the second session, respectively, of Unprimed experiments. In Figures 8.3 and 8.4 the number of times words from each story were

List of the words		
Bear	Metabolism	Hydropower
Year	Mediterranean	Feet
System	Underground	Mountain
Warm-blooded	Folklore	Chief
Achieve	Awake	Asia
Month	Source	Spring
Sea	Lemur	Inland
Slightly	Summer	Japan
Marsupial	Forest	Success
Kilometer	Sleep	Tunnel
Long	Desert	Equator
State	Shelter	Altitude
Time	Precipitation	Channel
Hot	Direction	Flood
Responsible	Climate	Stream
Hundred	Fish	Consciounsness

Table 8.1: List of the words used to check which story grabbed more subjects' attention at particular moments. The words in green are the words truly present in the story; the ones in red are not in any of the narratives. The order according to which the words appear in the list is randomly generated.

heard by each subject, is shown, in the first and in the second session respectively of Primed Experiments.

The results show that in the first session of UPex, 40.5% of the words totally heard by the subjects come from the first story; in the Pex, instead, 50% of the words chosen come from the first story. Considering the words heard in both sessions, in the Upex 42.8% of the words come from the first story, while in the Pex 55.5%.
In both cases, subjects, provided for priming, followed more the first story. This proves that priming actually pushes people to attend the narrative they are primed on, which was our initial hypothesis.

However, these are just some indicative considerations. We plan to carry out other experiments exploiting more subjects and further investigations to analyse the results obtained. Moreover, we plan to represent human behaviour in this situation also through the top-down attention model discussed in Chapter 3.

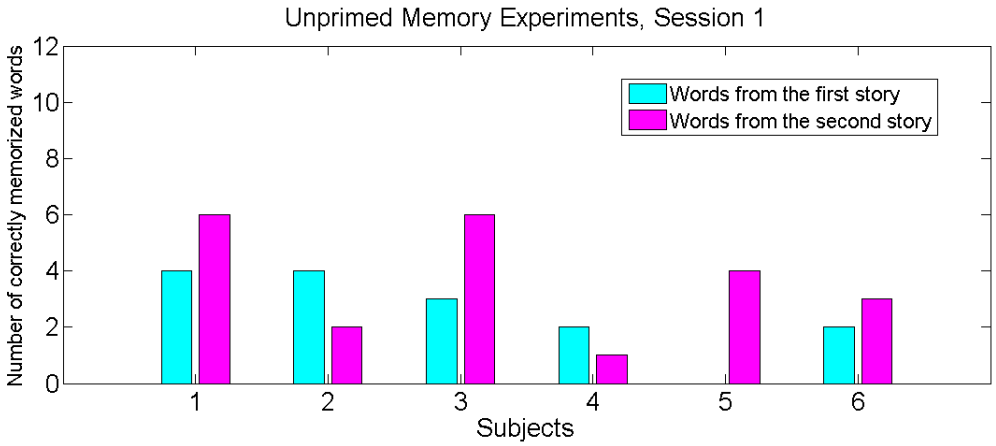


Figure 8.1: Results of Unprimed Experiments (Session 1): number of times words from each story are heard by each subject.

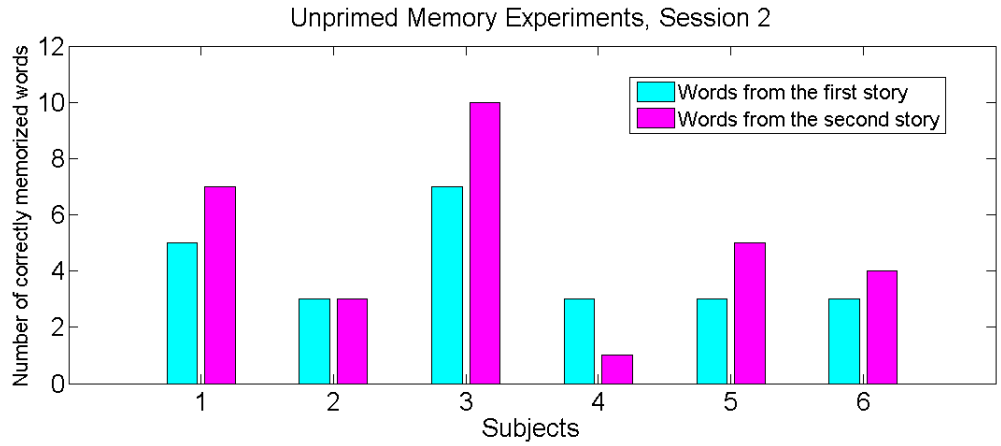


Figure 8.2: Results of Unprimed Experiments (Session 2): number of times words from each story are heard by each subject.

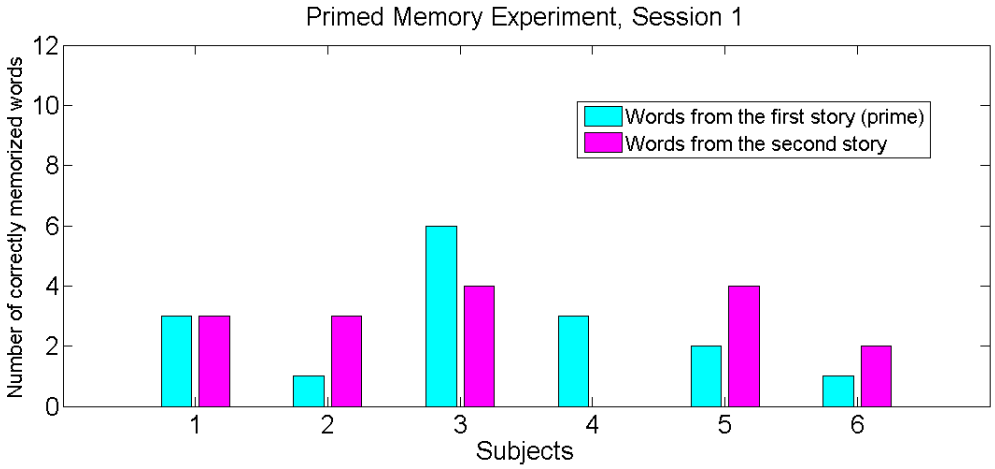


Figure 8.3: Results of Primed Experiments (Session 1): number of times words from each story are heard by each subject. The subjects are primed on the first story.

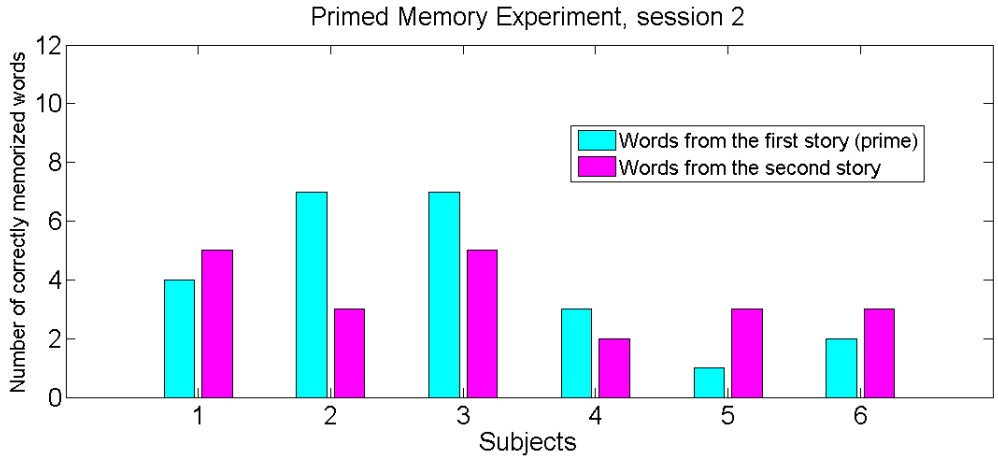


Figure 8.4: Results of Primed Experiments (Session 2): number of times words from each story are heard by each subject. The subjects are primed on the first story.

8.2.2 Counting Experiments: interaction between task and priming

We also tried to investigate the effect of priming in a cocktail party scenario using a different kind of experiment. The design is the same as the one we just discussed, but, instead of asking for words heard, we ask the subjects to push a button each time they heard the word "and" in one of the stories. Thus, again we make subjects listen to a monarual composition of two different narratives, which are extracts of texts of the same book (Phillips (2006)). Some changes are made, also in this case, not to have "listening in the gap effect"(Bregman (1994), Bronkhorst (2000)). The narratives are uttered by the same speaker of a speech synthesizer (Beutnagel *et al.* (1999)). Eventual spatial and speaker cues are removed (for more details, see section 7.2).

In this case, we used four different stories. We made subjects listen to the first two without giving any task and any priming. Then, we made them listen to other two giving them a priming about one of the stories.
The narratives we used for the unprimed part are reported below:

First Story: *The giant panda still lives in the wild in only a few mountain ranges in the southwestern part of China because its survival has been threatened both by hunters and by the destruction of the habitat it needs to survive. What has been noted and stressed in the last few decades is that the pandas survival is also threatened by the flowering and seeding cycles of the bamboo where the pandas live. Here's what the problem is. Bamboo is the main source of food for the giant panda. However, when there's a massive flowering or seeding of the bamboo, the bamboo that has just seeded dies, so there's a lag of quite a few years before the new, young seedlings grow enough to provide food for the giant panda. If the bamboo where the giant pandas living dies and then the giant panda needs to move to new areas to find food. The search for food has led the giant panda into areas, that are more settled and more full of danger for the giant panda.*

Second Story: *Conifers are hardy trees that have been able to survive well, so, as a result, both the oldest and the biggest trees in the world belong to the conifer family. The oldest known living tree is located in California. It is a four thousand years old bristlecone pine. The giant redwoods, found in California, are the largest trees; they can be several hundred feet tall with a weight of two thousand tons. A really interesting note about the giant redwoods is that, even though the trees are so big and tall, they have relatively small cones. They are evergreens with short and spiky leaves. The needle-like shape of Conifer leaves evolved as a reaction to drought and aridity. When compared with a flat leaf, a needle presents a much smaller surface area. Most conifers are evergreens. They lose and replace their needles throughout the year, rather*

than shedding all their leaves in one season.

The narratives and the priming used for the primed part are reported below:

First Story: *There are a few points to make about echolocation. Basically echolocation refers to the technique of using echoes to locate, range and identify whatever is in the surrounding area. The first point is that animals like whales and dolphins can use echolocation. Some whales have teeth, others don't, but it's only the toothed whales that have this capability and potential. A whale uses echolocation by sending out clicks, that are reflected back to the whale after bouncing off objects in the water. So echolocation is actually a series of clicks sent out by a whale that bounce off objects , then reflected back to the whale. A whale can learn from these clicks that bounce off an object just a bit, actually. A whale can learn the size and shape of objects that are out there. But from the reflected clicks a whale can learn even more, like how far away the object is, its movements and its speed.*

Priming for the second story: *Phyllotaxy is a scientific term that refers to the arrangement of leaves on the stem of a plant. On most plants, leaves are arranged in a definite pattern. It's very unusual for a plant to have randomly placed leaves. One of the main reasons why the leaves on a plant stern are arranged in an orderly way is to ensure that each leaf is exposed to the maximum amount of light with a minimum amount of interference from other leaves.*

Second Story: *There are three possible types of leaf arrangement: alternate arrangement, opposite arrangement and whorled arrangement. The first type of leaf arrangement is the alternate arrangement. In this type of leaf arrangement, there is only one leaf at each node. A node is the spot where the leaf and flowers are attached to the stem Magnolia. Prunus and Rubus are plants having this type of leaf arrangement. In the opposite arrangement, there are two leaves at each node, opposite each other on the stem. This type of leaf arrangement isn't as common as the alternate arrangement, with one leaf at each node. Examples are Acer and Buxus . The last possible type of leaf arrangement is called whorled arrangement. This type of arrangement is less common than the opposite and the alternate ones. In this type of arrangement, three or more leaves are attached to the stalk of the plant at the same node.*

Producing natural sounding speech, the speech synthesizer tends to put emphasis on the words "and" when they connect propositions or verbs. Thus, in order not have this emphasis, which could attract the attention, distracting the subject, in these experiments, the syntactical structure of the stories has been a bit modified with respect to the texts above.
The terms "and" are put in particular positions of the texts, distant enough from each other, so it could be possible, at least, in theory, to pay attention to all of them.

When subjects push the button, the exact time is registered and later used to determine which "and" they heard.
Subjects performed the same experiments four times (trials). Each subject executed both the primed and the unprimed experiments with the narratives reported above.
Some initial results are shown in Figures 8.5 and 8.5.

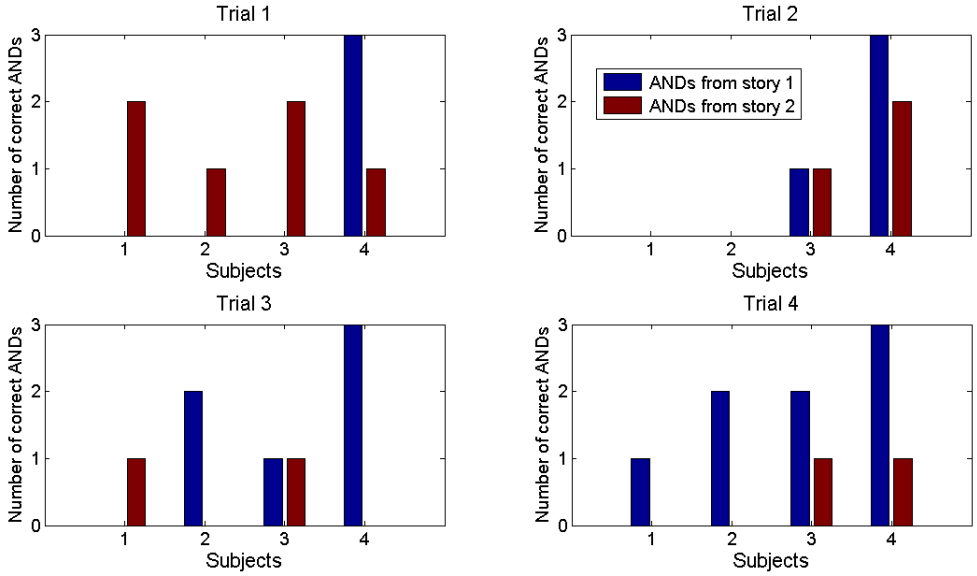


Figure 8.5: Results of Counting Experiments (Unprimed case): number of times ANDs from each story are heard by each subject.

The results show that, in the unprimed case, 38% of the identified ANDs comes from the second story, while, in the Primed case, 46% of the identified ANDs come from the second story, which is the story they are primed on.
It seems, then, that still priming helps people to pay attention to the story they have more information about. However, a deeper analysis is necessary exploiting more subjects and investigating the correlation existing between the effect of the task (looking for ANDs while listening) and the priming.
Also in this case, we want to represent the interaction between these two different top-down cues in human behaviour through the model discussed in Chapter 3.
Note, finally, that, in the counting experiments the same subjects performed both the primed and the unprimed experiments using two different pairs of stories, while in the ones described in section 8.2, half of the subjects performed the Unprimed ones, the other half the Primed ones and the stories were the same two in both occasions. The aim is to make the study as much as possible independent on the characteristics of the stories (content, words) and to the subjects' interest.

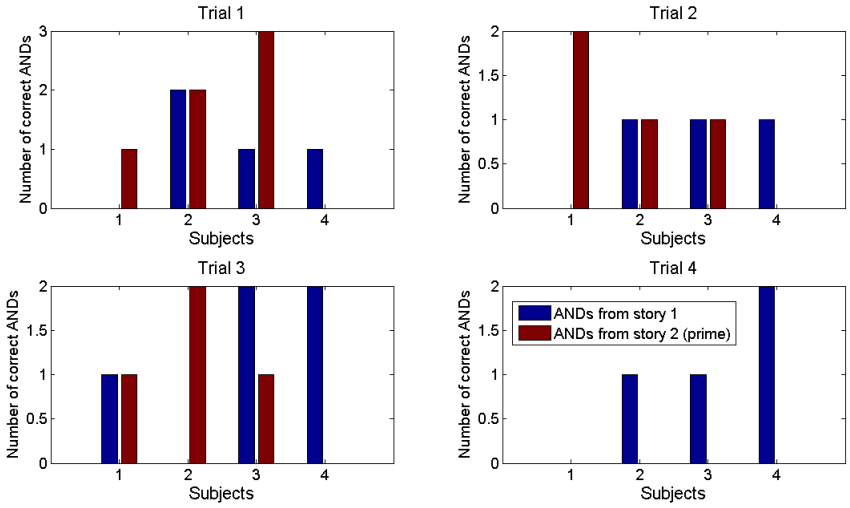


Figure 8.6: Results of Counting Experiments (Primed case): number of times ANDs from each story are heard by each subject. The subjects are primed on the second story.

Part IV

Multimodal perception and human-robot interaction

Chapter 9

Multimodal Speaker Recognition in a Conversation Scenario

In this Chapter, we investigate multi-modal perception for a human-robot interaction purpose. We propose a robotic system that, taking advantage of multiple perceptual capabilities, actively follows a conversation among several human subjects. The essential idea of our proposal is that the robotic system can dynamically change the focus of its attention according to visual or audio stimuli to track the current speaker throughout the conversation and infer his/her identity.

The work has been already presented and published in Marchegiani *et al.* (2009a), Marchegiani *et al.* (2009b) and Marchegiani and Pirri (2009).

9.1 Introduction

Multi-people, multi-modal detection and tracking scenarios have been modelled in the context of *smart rooms* (see Pentland (1995) among others), e-learning, meeting and teleconferencing support, but also in robot-human interface. See Waibel *et al.* (2003) for a complete review of technologies for intelligent rooms.
Recently, multi-modal features have been used in domestic environments in order to annotate people activities for event retrieval by Desilva *et al.* (2006), perform face and speech recognition, people tracking, gesture classification and event analysis (e.g. Reiter *et al.* (2005)). A conversation scenario has been addressed in Bennewitz *et al.* (2005); here a robot interacts with several people performing changes in focus and showing different emotions.

The mentioned works do no directly address the problem of multi-modal identity estimation. To fill the gap we introduce a completely new model of a conversation scenario. We propose a framework for real-time, multi-modal speaker recognition

combining acoustic and visual features to identify and track people taking part to a conversation. The robot is provided with acoustic and visual perception: a colour camera and a pair of microphones oriented by a pan tilt unit (see Figure 9.1). The robot follows a conversation among a number of people turning its head and focusing its attention on the current speaker, according to visual and audio stimuli.
It tracks the participants exploiting a learning phase to settle both visual-audio descriptors and integration parameters. In particular, four audio and visual descriptors of features are defined for both real time tracking and identification. Visual and audio identification is obtained by combining the outcome of these descriptors analysis with a generalised partial linear model (GPLM) (see Müller (2001) and Severini and Staniswalis (1994)). Finally, the process undergoes a dynamic updating.
People are then identified against a set of individuals whose audio and visual features are suitably structured in a knowledge base, as prior knowledge.
The proposed scenario is quite general and people can change position, leave or join the conversation, thus we cannot exploit the current location of the speaker to infer her identity.

Figure 9.1: On the left: an example of the robot following a conversation in the Lab. On the right: the robot head with a pair of microphones and a camera oriented by a pan-tilt unit.

9.2 Data acquisition

The knowledge base is a complex data structure that includes both the voice and visual features of R subjects, male and female, with $R = 30$. Each speaker's voice is modelled as a Gaussian mixture density (GMM). The models are trained with the first 18 Mel frequency cepstral coefficients (MFCC) (see section 9.3, Pols (1977) and Zheng *et al.* (2001)) of a particular set of part of speech, made up of very short word utterances of the English phonemes (a single word contains a single phoneme, such as: put, pet, do, etc.).
These particular utterances allow to collect only a small set of vocal samples per

speaker (two or three examples for phoneme), rather than a whole conversation. Furthermore experiments prove better performance on short words, in particular when the system works in real-time and the active speaker has to be recognised by a short observation sequence.
The j-th phoneme pronounced by the i-th speaker is described by a mono-dimensional vector of the audio signal, and its relative MFCC by a 18-dimensional matrix S_j^i for each utterance.
Given the number of phonemes N_f (in this case 44) and the number R of voice sampled, $\mathbf{S}^i = [S_1^i S_2^i ... S_j^i \dots S_{N_f}^i]$, with $i = 1, \dots, R$ and $j = 1, \dots, N_f$, indicates the complete features matrix of the speaker i.
Let χ be the number of Gaussian density in the mixture, let c_i be the weights components of the i-th model, with $\sum_{l=1}^{\chi} c_{il} = 1$, and let Σ_{il} and μ_{il}, $l = 1, \dots, \chi$ be the covariances and the means of each component of the i-th model, each voice model is completely specified by the parameters, i.e.

$$\lambda_i = (c_{i1}, \dots, c_{i\chi},\ \mu_{i1}, \dots, \mu_{i\chi},\ \Sigma_{i1}, \dots, \Sigma_{i\chi}) \tag{9.1}$$

The face appearance features of the R=30 people are coded in 2 coefficient matrices. Columns of both matrices encode the observations, namely the people features. In the first matrix, rows encode the Karhunen-Loève coefficient vectors of the *eigenfaces* (Turk and Pentland (1991)). In the second matrix, rows encode the values of the non-parametric 2D face colour density, taken at each bin.

The acquisition process performs, on a continuous loop, the following tasks:

1. tracking people in the field of view over the frame stream, shifting the focus to include the current speaker into the field of view according to the angle θ (see Figure 9.2) determined by voice analysis;

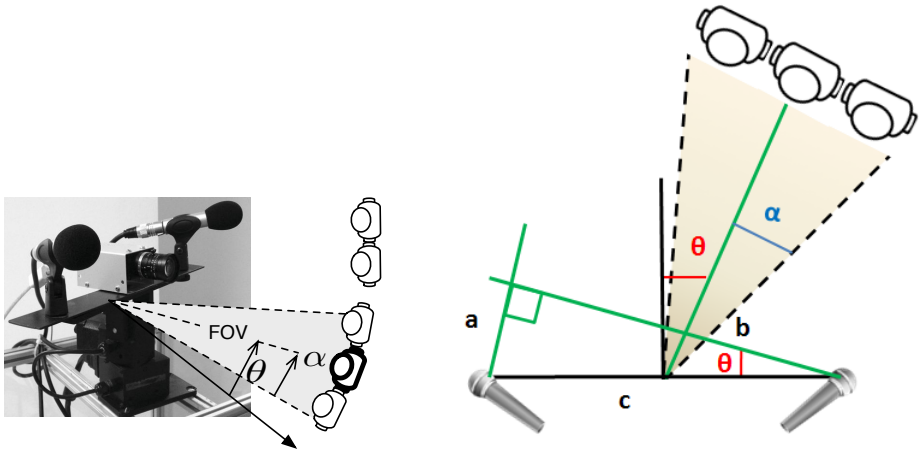
2. extracting appearance based features, given the number of people in the current field of view and an hypothesis on the current speaker, returned by the voice analysis;

3. collecting the visual and voice descriptors to feed the multi-people identification process.

By "field of view"(FOV) we mean the width and the height of the scene that can be observed with the camera lens. In the following, given our audio set-up allowing only for horizontal speaker localisation, we refer the term FOV to the interval

$$FOV = [\theta - \alpha, \theta + \alpha], \quad \text{with } \alpha = \tan^{-1}(w/2f) \tag{9.2}$$

here f is the focal length of the camera, w is the image width and θ is the current pan angle of the estimated voice source (see Figure 9.2).



(a) (b)

Figure 9.2: (a): The concept of the robotic head following the conversation is shown: pan angles are set according to the direction θ of the estimated voice source, relative to the zero pan position (the solid arrow). Besides the speaker (bold in the figure), other subjects are detected in the FOV, spanned by 2α. (b): An other view of the angles θ and 2α.

9.3 Acoustic scene modelling

In this section we present our approach to locate the active speaker and estimate the likelihood of the speaker features recorded during the conversation, with respect to the models created, as described in Section 9.2. The result is an ordered sequence of voice likelihoods that is suitably combined with the visual features, described in the next section, to produce the dataset further used to identify the speaker in the scene.
We adopt the real-time algorithm proposed by Murray *et al.* (2004), based on the time delay of arrival and the cross-correlation measure to compute, every 200 ms, the angle θ between the sound source and the robot on the horizon plane.

Each speaker's voice i is modelled as a GMM λ_i. Since each voice feature set i corresponds to a mixture, we also indicate the speakers with their corresponding voice model λ_i.
The GMM are often selected for this kind of tasks, being able to describe a large class of sample distributions for acoustic classes relative to phonetic events, such as vowels, nasals or fricatives Reynolds and Rose (1995).
In order to obtain the models, we extract MFCC up to the 18^{th} order from the particular set of utterances described in the previous section. The mel-frequency cepstrum coefficients are robust against noise and defined as a cosine transform of the real log-

Variable	Meaning
R	Number of people whose voices and faces we have sampled. $R = 30$
N_f	Number of phonemes. $N_f = 44$.
S_j^i with $j = 1 \dots N_f$ and $i = 1 \dots R$	MFCC matrix of phoneme j pronounced by the speaker i, used to train the Gaussian Mixture model λ_i of the voice of each speaker
$\mathbf{S}^i = \left[S_1^i S_2^i \dots S_j^i \dots S_{N_f}^i\right]$ with $i = 1 \dots R$	MFCC matrix of all the utterances pronounced by the speaker i, used to train the model λ_i
λ_i with $i = 1...R$	GMM of the voice of speaker i
$\{ c_{il}, \vec{\mu_i}, \Sigma_i \}$	Weights, means and covariances of the model λ_i
θ	Angle between robot and its interlocutor
w	Width of the image
f	Focal length of the camera
$FOV = \gamma \in [\theta - \alpha, \theta + \alpha]$	Field of view where $\alpha = \arctan(w/2f)$

Table 9.1: Table of the variables used in the section 9.2.

arithm of the short term energy spectrum, reported on a mel-frequency scale, which closely approximates the human auditory system (Pols (1977)). In Figure 9.3 the general procedure to compute MFCC is shown. A comparison between different ways of implementing MFCC is provided by Zheng *et al.* (2001).
We use 18 coefficients as a trade off between complexity and robustness after 25 experiments. The parameters initialisation of the EM algorithm and the number of Gaussian components are provided by the the mean shift clustering technique (Comaniciu and Meer (2002)); we get a varying number of components χ, with $7 \leq \chi \leq 15$.
For each utterance x_t acquired during a conversation and the associated MFCC matrix $\mathcal{A}_t$, composed of N_A 18-dimensional coefficients, we obtain, through the complete features matrix and the GMM, a probability $p(x_t^j|\mathbf{S}^i)$ for all coefficient components $x_t^j,\ j = 1, \dots, N_A$. Thus, the expectation is

$$E(\lambda_i|\mathcal{A}_t, \mathbf{S}^i) = \sum_{k=1}^{N_{At}} p(\lambda_i|x_t^k, \mathbf{S}^i)\ p(x_t^k|\mathbf{S}^i) \tag{9.3}$$

The identification process, on the other hand, also involves clustering of speakers voice labels (see Section 9.6).

To prevent the robot from turning its head to follow noises or random sounds in the environment, we trained a linear classifier based on support vector machine (SVM), able to distinguish between speech and no-speech frames.
We use support vector machine, because this classification method has shown excellent performance in Voice Activity Detection (VAD) problems, providing more

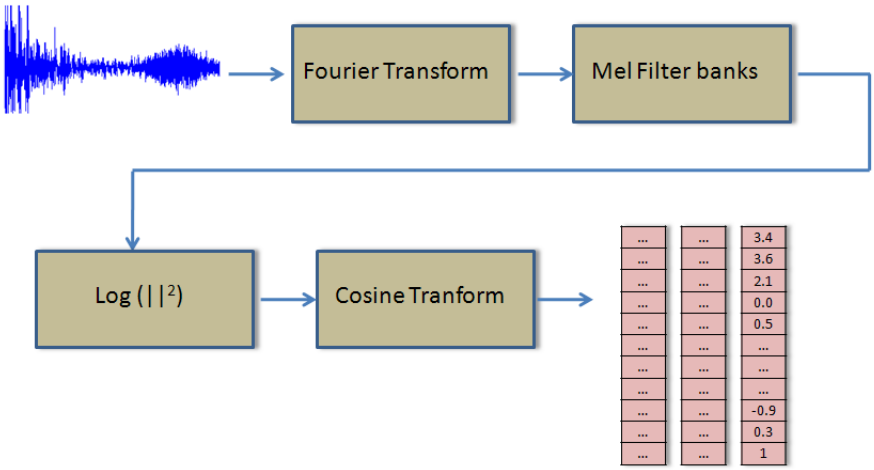


Figure 9.3: MFCC (Mel-frequency cepstrum coefficients) computation procedure.

accurate results than other techniques, as illustrated by Ramirez *et al.* (2006) and Enqing *et al.* (2002).
We consider the short term energy and the zero crossing rate of the signal received (Rabiner and Sambur (1975), Atal and Rabiner (1976), Childers *et al.* (1989)) as discriminant features for SVM classification. The short term energy is the sum of the squares of the amplitude of the fast Fourier transform of the samples in a frame and the zero crossing rate is the number of times the signal changes its sign within the same frame.
Other features, like MFCC or Linear Prediction Coding (LPC) Itakura and Saito (1968) have been used in literature for speech/sound discrimination (see *kinnune07* and Nemer *et al.* (2001)among others) but, in our scenario, which is characterized by a not that high presence of noise, short term energy and zero crossing rate are able to guarantee a robust classification, in spite of the low computational cost involved.
The set that we use to train the classifier is composed of 10 different speech frames for each speaker in the knowledge base and the same number of frames including silence or background noise (see Figure 9.4).

9.3.1 Acquisition and processing

Acquisition and processing of voice features is performed in real time: both the voice detection and the localisation procedure work with frames of length $\Delta = 200ms$, the identification process with frames of length $5\Delta = 1s$.
Specifically, given the signal $u(t)$, acquired by the microphones at time t, and the frame $u_\Delta(t)$, containing the values of $u(t)$ in the time interval $(t - \Delta,\ t)$, the SVM

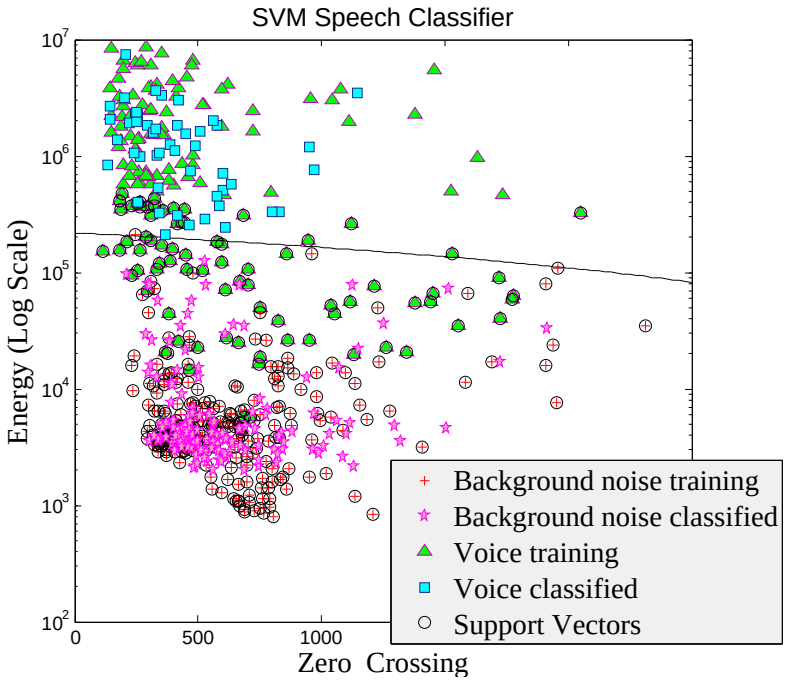


Figure 9.4: SVM classification of voice against background noise: training and testing.

classification implements a filter that provides a signal $\hat{u}_\Delta(t)$, defined as follows:

$$\hat{u}_\Delta(t) = \begin{cases} u_\Delta(t-\Delta) & \text{if } u_\Delta(t) \text{ does not include a human voice} \\ u_\Delta(t) & \text{otherwise} \end{cases} \tag{9.4}$$

This filtered frame is used for speaker localisation and identification.
The angle between the robot and its interlocutor is computed for each signal $\hat{u}_\Delta(t)$. On the other hand, the segment of conversation which we link to an identity, among the sampled voices, is represented by the utterance x_t corresponding to 5 consecutive frames $\hat{u}_\Delta(t)$.
On this premises, the acoustic scene modelling provides the list of the M most likely speakers. The first part $\hat{S}$ of the list includes the labels associated with the models λ_i, maximising $E(\lambda_i|\mathcal{A}_t, \mathbf{S}^i)$, given the utterance at time t and its MFCC matrix $\mathcal{A}_t$. The other values, modulo expectation, concern people indicated by the visual analysis, if not already in $\hat{S}$.

Variable	Meaning
χ	Number of components of the models λ
x_t	Audio signal captured by the microphones every $1s$
$\mathcal{A}_t$	MFCC matrix of the utterance in a conversation, the current relative speaker of which we have to estimate.
N_A	Number of coefficients in $\mathcal{A}_t$
$x_t^j,\ j = 1 \dots N_A$	j-th vector of the matrix $\mathcal{A}_t$
$u(t)$	Audio signal captured by the microphones at time t
Δ	Sampling interval of acoustic signal. $\Delta = 200$ms
$u_\Delta(t)$	Frame containing the values of $u(t)$ in the time interval $(t - \Delta, t)$
$\hat{u}_\Delta(t)$	Filtered signal by SVM classifier
M	Number of most probable speakers given by acoustic analysis
$\hat{S}$	List of the labels associated with the models λ_i maximising $E(\lambda_i \mid \mathcal{A}_t, \mathbf{S}^i)$

Table 9.2: Table of the variables used in section 9.3

9.4 Visual face descriptors

Visual scene analysis starts with a multi scale face detector. A cascade of classifiers is used to progressively discard areas that are not likely to include a face, by combining successively more complex and computationally expensive classifiers, to mimic the attention that selects areas deserving further processing.
Once a face is detected, the face area is divided in regions of interest on which different detectors are scanned to locate the eyes and the mouth. If these detections succeed over a number of different frames the computational process enters the tracking state in which the eye and mouth detectors are scanned across a predicted face region that is computed from the previous frame by a face tracker, based on mean shift.

The core of visual feature extraction is the integration of the detection and tracking facilities, by a finite state machine with detection and tracking states. Pre-processing involves equalisation, aligning and segmentation of face images. For the alignment, we rely on the position of eyes and mouth: assuming all faces are frontally detected, the images are rotated and scaled to compensate the angle γ formed by the eyes. Being d the computed distance between eyes, $\sigma = \bar{d}/d$ is the scale factor needed to obtain the desired distance $\bar{d}$, (x_c, y_c) is the centroid of the eyes and mouth, $a = \sigma \cos\gamma$ and $b = \sigma \sin\gamma$. The transformation H that maps the original to the transformed image is expressed by the 2×3 matrix

$$H = \begin{pmatrix} a & b & (1-a)x_c - by_c \\ -b & a & bx_c + (1-a)y_c \end{pmatrix} \tag{9.5}$$

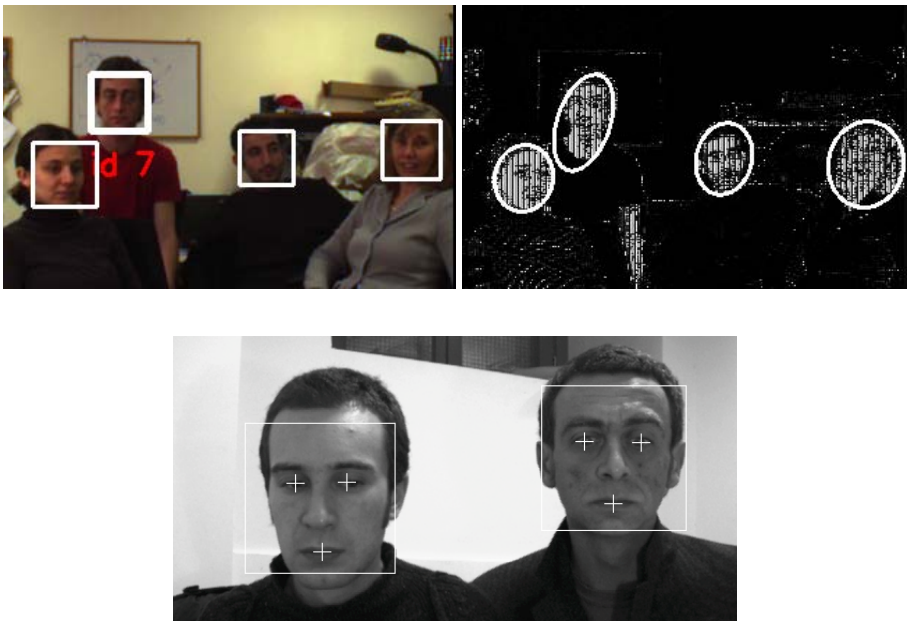


Figure 9.5: The figures illustrate the detection and tracking of facial features: faces detected in the field of view are tracked; eye positions are estimated and used for face normalisation, while mouth ROI is needed by the speaker detection process. Brighter pixels in the backprojected image indicate higher probability values; ellipses are centred and oriented according to the extracted eye and mouth positions.

A fixed size region of interest is centred on (x_c, y_c) to extract, from the transformed image, an aligned face image.

In the following we introduce a set of descriptors providing a compact representation of a person's appearance and are suitable in this identification problem. Namely, three kind of visual descriptors are defined for each detected face in the scene:

- a probability that the subject is currently speaking based on mouth movements;
- a compressed representation of the intensity image;
- a non-parametric colour distribution.

All descriptors refers to a specific region Q_i, $i = 1, \ldots, K_{FOV}$, with K_{FOV} the number of people visible in the camera FOV (see Table 9.4).

9.4.1 Visual speech descriptor

This descriptor sizes the significant mouth movements, in so contributing to the cross-relation between audio and visual features for the recognition of the speaker in the

scene. Indeed, the problem is to evaluate the amount of pixel changes needed to tell that the mouth is articulating a phrase.
To face this problem we define a binary mask M_B by thresholding differences of frames from subsequent time steps. Each pixel is treated as an i.i.d. observation of a binary random variable x representing the detected change.
Assuming a Binomial distribution for the binary variables within the mask, we estimate the parameter μ using a Beta prior distribution characterised by the hyperparameters α and β.
While the μ parameter accounts for the influence of the number of pixels that have changed over the all pixel set, the two binomial proportions α and β enforce or weaken the amount of changes according to the effective number of samples that seem to vary. The best values for our implementation are $\alpha > 0.7$ and $\beta < 0.2$.
Let N_B be the size of the observations, with ρ the number of pixels set to 1 by the mask (those changing).
The likelihood that the observations come from a windows in which the mouth has significantly moved, as for articulating speech utterances (of course also for smiling or yawning) is thus defined as

$$\sum_{x \in M_b} p(x|\mu_B, N_B - \rho + \beta, \rho + \alpha)\mu \tag{9.6}$$

Note here that μ is re-estimated from each detected M_B, and thus μ underlines the model which most likely is induced by mouth activity. In any case the M_Bs are chosen, among those having best expectation, also according to the chosen voice models.

9.4.2 Face appearance feature descriptor

Karhunen-Loève(KL) coefficients provide efficient compression and are suitable to encode frontal faces that have been previously aligned.
Compression is achieved by projecting D-dimensional face data into a D'-dimensional subspace spanned by the D' principal directions. Being $\mathbf{c}_i$ the D'-dimensional KL coefficient column vector representing the visual features of the i-th subject, we measure the similarity between i and j in face intensity images by computing the coefficient Mahalanobis distance $d_{\mathcal{M}}(\mathbf{c}_i, \mathbf{c}_j) = (\mathbf{c}_i^\top \mathbf{\Lambda}^{-1} \mathbf{c}_j)^{1/2}$, where $\mathbf{\Lambda}$ is the diagonal matrix, with the eigenvalues corresponding to each KL coefficient.

9.4.3 Face colour feature descriptor

The similarity in colour space is based on the Bhatthacharyya distance between non parametric densities of colour features. More precisely, given two histograms specified by the vectors of bin value $\mathbf{h}_j$ of the face colour features, $j = 1, \ldots, R$, the Bhatthacharyya distance is defined as $d_{\mathcal{B}}(\mathbf{h}_i, \mathbf{h}_j) = (1 - (\tilde{\mathbf{h}}_i^\top \mathbf{\Omega}^{-1} \tilde{\mathbf{h}}_j))^{1/2}$, here $\tilde{\mathbf{h}}$ is

the vector obtained by computing the square root for every element of $\tilde{\mathbf{h}}$ and $\mathbf{\Omega}$ is the diagonal matrix mentioning the normalisation coefficients, such that $0 \leq d_{\mathcal{B}} \leq 1$. These colour descriptors, although are not robust against changes in face illumination conditions, compensate degradation of shape cues caused by poor resolution or changes in head orientation.

Variable	Meaning
γ	Angle between the eyes
d	Distance between the eyes
$\bar{d}$	Desired distance between the eyes
σ	Scale factor to obtain the desired distance $\bar{d}$
(x_c, y_c)	Coordinates of the centroid of the detected faces.
$a = \sigma \cos\gamma$	Transformation equation
$b = \sigma \sin\gamma$	Transformation equation
H	Transformation matrix
K_{FOV}	Number of people visible in the camera FOV
$Q_i,\ i = 1 \ldots, K_{FOV}$	Regions currently in the camera FOV
M_B	Binary mask for visual speach detection
x	Random variable representing the changes of the mouth.
μ	Parameter estimated by a Beta prior distribution
α, β	Hyperparameters used for the estimate of μ
N_B	Size of the observations for visual speach detection
ρ	Number of pixels set to 1 by the mask
$\mathbf{c}_i$	KL coefficient column vector
$d_{\mathcal{M}}(\mathbf{c}_i, \mathbf{c}_j)$	Coefficient Mahalanobis distance
$\mathbf{\Lambda}$	Diagonal matrix with the eigenvalues D corresponding to KL coefficient
$\mathbf{h}$	Vector of bin value of the face colour features
$\tilde{\mathbf{h}}$	Vector obtained by computing the square root of every element of $\mathbf{h}$
$\mathbf{\Omega}$	Diagonal matrix mentioning the normalisation coefficients
$d_{\mathcal{B}}(\mathbf{h}_i, \mathbf{h}_j)$	Bhatthacharyya distance between ($\mathbf{h}_i$ and $\mathbf{h}_j$

Table 9.3: Table of the variables used in section 9.4

.

9.5 Discovering people identities

We recall the reader that the problem we have to solve is online identification of the speaker. We have discussed in the previous sections the different descriptors of voice, lips movements, and face features. We have shown that the voice and the lips movements are defined by a probability distribution (Gaussian Mixture for voice

Q_i= $region\ label$	X_1	X_2	X_3	X_4	Y	$Peop.\ label$
Q_1	0.6909	0.0013	0.4419	0.505	1	A
Q_2	0.3090	0.0922	0.4652	0.505	0	A
Q_1	0.6909	0.1529	0.9516	0.334	0	B
Q_2	0.3090	0.1237	0.3638	0.334	0	B
Q_1	0.6909	0.0014	0.3954	0.161	0	C
Q_2	0.3090	0.1897	0.5641	0.161	0	C

Table 9.4: The descriptors table corresponding to a trial with two regions (Q_1, Q_2) in the camera FOV and three people labels A, B and C with best descriptors classification. From left to right Q_i is the region label, X_1 is the lips movements descriptor, X_2 is the normalised Mahalanobis distance between the Karhunen-Loève coefficients for the current observed regions, labelled Q_i, and the analogous coefficients stored for each identities in the Knowledge base.
X_3 is the normalised Bhatthacharyya distance between the non parametric functions in colour space, sampled in the region Q_i and in the images recorded in the knowledge base.
Finally X_4 is the voice descriptor and Y will take value 1 in correspondence of the estimated speaker.
Here we assume that there are only two regions in the current camera FOV, labelled by Q_1 and Q_2 and that the people A, B, C chosen are those in the union of the voice set and the distance feature set with best classification. Data are repeated for each potential identity. The task is to identify which row is the correct one. The row will tell who is the current speaker and which is the region in the current camera FOV that corresponds to the speaker. This implies that the real speaker is identified by both the voice and the face. In this case the correct row is the first, thus the correct region label is Q_1 and the speaker is A.

and Beta-Binomial distribution for lips movements) while the other features are normalised distance measures with respect to data coded in the knowledge base.

In this section we discuss how to infer from these heterogeneous data the current speaker identity, presuming that the speaker changes in time. Now, data are collected during a time lapse $t : t + 1sec$ and descriptors are computed for each of these intervals. We define each time lapse a trial, hence from each trial a data structure is formed.
This data structure is peculiar because the descriptors have been generated from different sample spaces.

1. The voice descriptor computes a probability with respect to MFCC codes, hence it returns the likelihood that A or B or C, etc. are speaking. It is clear that if the MFCC of two people with very similar voices are stored in the database, say A and Z, even if Z is not in the room, any time A will speak there will be a good chance that the voice estimator will return a high likelihood also for Z.

2. The lips descriptor will tell who in the scene is moving the lips, so this feature needs to be combined with a voice and a face to be instantiated with a speaker. Indeed, people can articulate the mouth also, for example, for laughing and yawning.

3. The two normalised distances, on the other hand, tell who is plausibly in the camera FOV in the current trial, but cannot tell who is speaking.

A consistent data structure for these varied dataset is illustrated in Table 9.4.
According to the structure of the descriptors, trials will nominally include all the people enumerated in the knowledge base (see Section 9.2). The chosen labels are those in the union of the sets of descriptors with best classification.
Note that if B is in the set, because it has a good classification for the distance features, with respect to a region labelled Q_2, this does not imply that it keeps a good classification with respect to voice or to another region Q_j, $j \neq 2$. That is why we take the union of the sets, instead of the intersection, that might be the empty set.
Note also that distances are normalised with respect to the whole dataset and not with respect to the chosen ones (see Section 9.4).

In order to find the correct row, given a trial, we use regression, in particular we use a semi parametric regression model. The model will estimate the parameters $\beta = (\beta_1, \beta_2)$ and a function g which, when applied to a trial, will return the row that most plausibly indicate the current speaker.
To estimate these parameters, however, we had to resort to a training phase. Using the voices and the images in the knowledge base, and suitably adding errors for simulating environment influences, we have defined a set of 925 simulated trials 427 of

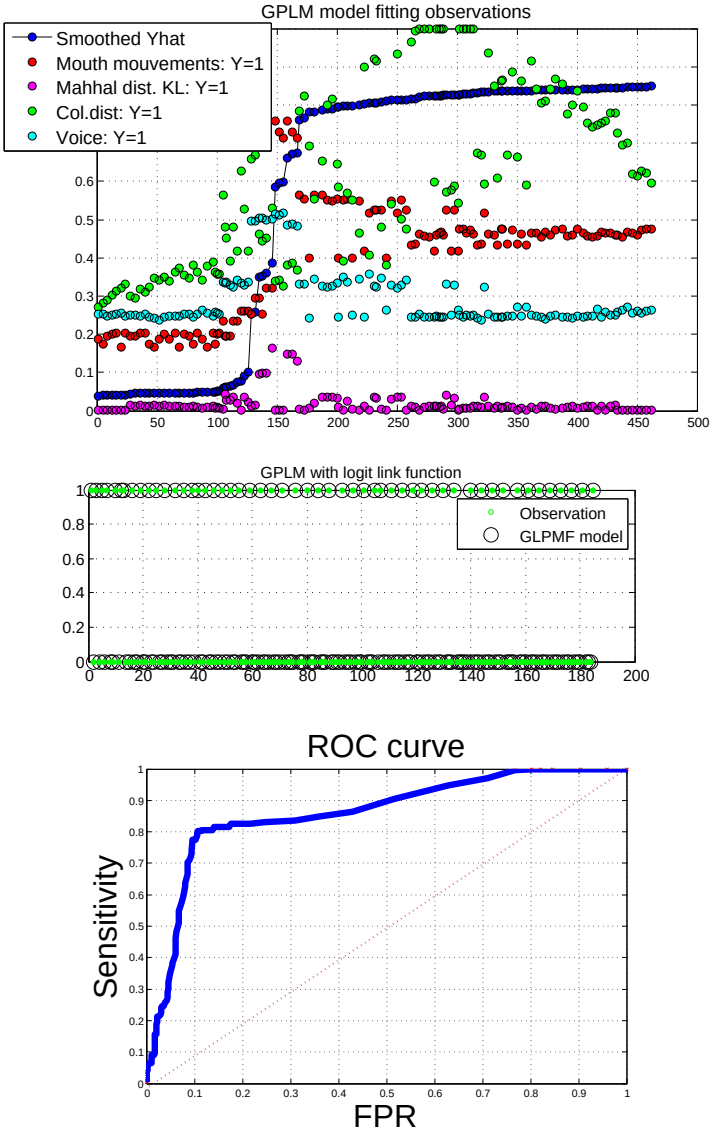


Figure 9.6: The left image illustrates the behaviour of each descriptor, as indicated in the label, taken at the value $\hat{Y}$, indicated $YHAT$, chosen for $Y = 1$. The table on the centre illustrates 187 of the 498 matches obtained during testing with the GPLM. On the right the ROC curve. The false positive rate and true positive rates are defined as $FPR = FP/(FP+TN), TPR = TP/(TP+FN)$ (here FP are false positive, TP true positive, FN false negative and TN true negative). Sensitivity (TPR) is plotted in function of the false positive rate for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold for $\hat{Y}$: if the decision threshold is chosen to be 0.5 then $FPR < 0.1$, while if it is 0.9 then $FPR = 0.5$.

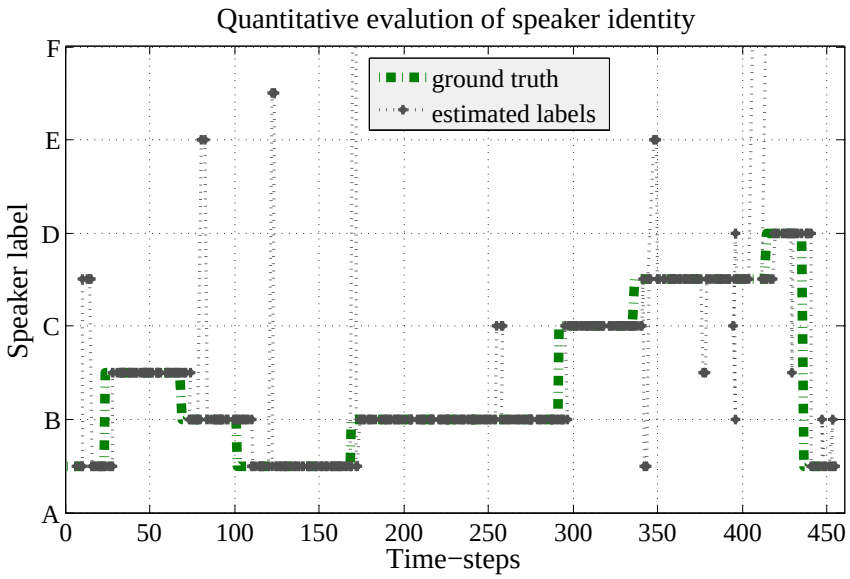


Figure 9.7: Qualitative evaluation of a speaker identity estimation performance over 450 time-steps. The Figure shows the system tracking the current speaker identity and recovering from failure.

which have been used for training and the remaining for testing.
We can now introduce the model. Given the descriptors $X_1, \ldots, X_4$ a semi parametric regression model for inferring the row corresponding to the speaker is defined as:

$$E(Y|\mathbf{UT}) = P(Y = 1 \mid \mathbf{U}, \mathbf{T}) + \epsilon = f(X_2\beta_1 + X_3\beta_2 + g(X_1, X_4)). \quad (9.7)$$

Here f is the standard logistic distribution $f(z) = (exp(z)/(1 + exp(z)))$, $\mathbf{U} = (X_2, X_3)^\top$, $\mathbf{T} = (X_1, X_4)^\top$, and β and g are the parameters and function to be estimated.
Note that we have grouped on one side the normalised distances $\mathbf{U} = (X_2, X_3)$, for which we want to estimate the parameters β_1 and β_2, and on the other side we have grouped the two probabilities $\mathbf{T} = (X_1, X_4)$.
Differently from other regression models, the general non parametric regression model (9.7) is optimal for capturing the combination of linear and non-linear characters of the descriptors.
Figure 9.6 illustrates the different behaviours of the features descriptors considering 427 trials. Here YHAT denotes the $\hat{Y}$ estimated by regression, that has been set to 1, with a decision threshold of 0.67. We estimate g and β according to the algorithm proposed by Müller (2001) and Severini and Staniswalis (1994), here for the logit case. The iterative steps of the algorithm are reported in Table 9.5.

Initialisation:
$\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$ are the descriptors for the 427(498) trials of the training(test) set
$\mathbf{Y}$ the regressed vector
$\mathbf{U} = (\mathbf{X}_2, \mathbf{X}_3), \mathbf{T} = (\mathbf{X}_1, \mathbf{X}_4)$, $Y_i = 1$ on all the correct Q_i, for each trial.
$\hat{\beta} \leftarrow 0$,
$\hat{g} \leftarrow f^{-1}((\mathbf{Y} + 0.5)/2)$
$\mu_{\hat{\beta}} \leftarrow \exp(\eta)/(1 - \exp(\eta))$ with $\eta = \mathbf{U}^\top \hat{\beta} + \hat{g}(\mathbf{T})$

Loglikelihood and derivatives for μ:
$\mathcal{L}(y, \mu)$: $y \log(\mu) + (1 - y) log(1 - \mu)$
$\mathcal{L}'(y, \mu)$: $((y - \mu)/(\mu(1 - \mu)))\mu'$
$\mathcal{L}''(y, \mu)$: $(y - \mu)(\ \mu''/\mu(1 - \mu) - (1 - 2\mu)\mu'^2/(\mu(1 - \mu))^2) - \mu'^2/(\mu(1 - \mu)))$

Repeat estimate $\hat{\beta}, \hat{g}$, using a smoothing matrix $\hat{\mathscr{M}}$, see eq. (9.8)
until $\mu_{\hat{\beta}}^{new} - \mu_{\hat{\beta}} < \epsilon$
here K is the number of trials, $\mathbf{1_K}$ is a vector of ones, $\otimes$ is the Kronecker product:
$\mathbf{W} = diag(\mathcal{L}''(\mathbf{Y}, \mu_{\hat{\beta}}))$
$\mathbf{Z} = \mathbf{U}^\top \hat{\beta} + \hat{g} - \mathbf{W}^{-1} \mathcal{L}'(\mathbf{Y}, \mu_{\hat{\beta}})$
$\mathscr{M} = \mathscr{M}_1 ./ \mathscr{M}_2$, with $\mathscr{M}_1 = (\mathcal{L}''(\mathbf{Y}, \mu_{\hat{\beta}}) \otimes \mathbf{1_K})^\top \hat{\mathscr{M}}$ and $\mathscr{M}_2 = \sum \mathscr{M}_1 \otimes \mathbf{1_K}$
$\hat{\mathbf{U}} = (\mathbf{I}_{K \times K} - \mathscr{M})\mathbf{U}$
$\hat{\mathbf{Z}} = (\mathbf{I}_{K \times K} - \mathscr{M})\mathbf{Z}$;
$\hat{\beta} = (\hat{\mathbf{U}}^\top \mathbf{W} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^\top \mathbf{W} \hat{\mathbf{Z}}$
$\hat{g} = \mathscr{M}(\mathbf{Z} - \mathbf{U}^\top \hat{\beta})$
$\mu_{\hat{\beta}}^{new} \leftarrow \exp(\eta)/(1 - \exp(\eta))$ with $\eta = \mathbf{U}^\top \hat{\beta} + \hat{g}(\mathbf{T})$

Table 9.5: Estimation of $\hat{\beta}$ and $\hat{g}$ for the regression model $f\{\mathbf{U}^\top \beta + g(\mathbf{T})\}$, using the training set of $K = 427$ simulated trials and a test set of $K = 498$ trials.

The goal of an empirical analysis of the data collected is to use the finite set of observations obtained for training, that is, $(\mathbf{X}_{ji}, Y_j), j = 1, \ldots, 427, i = 1, \ldots, 4$ to estimate β, g.
These values, together with the canonical logit f are used to finally predict a speaker identity. Estimation amounts to the following steps:

1. Analysis of predictors performance to prove their impact on Y. Estimation of the β and of the unknown function g using the training set of the 427 trials, from the 925 obtained by simulation (using the data collected in the knowledge base). Validation of g and β with the remaining 498 trials, for all the plausible $Q_i, i = 1, \ldots, m$ in the camera FOV (in our case $m = 2, 3, 4, 5$).

2. Prediction, in real time, of the speaker identity given the current observations and the knowledge of the current trial dimension (that is, $m^2 \times 5$, with m the number of identified regions in the camera FOV, Q_i, $i = 1, \ldots, m$), considering the whole dataset.

3. Convergence of the identification process after a burning period. Expectation of the features of each identified speaker can be used to track the conversation

dynamically and refine the probability of the identity of each speaker using a dynamical model, not described here.

We consider each descriptor $X_1, \ldots, X_4$ as an explanation or predictor of the speaker identity. By looking at the performance of each explanation (see Figure 9.6) and also because we have two probabilities and two distances, we have chosen to group the two probabilities, that is, lips movements (X_1) and MFCC (X_4) with the non-parametric g.
The iterative algorithm, with training data, starts with an initial approximation of $\hat{g} = f^{-1}((Y + 0.5)/2)$, with Y set to 1 on the correct regions labelled Q_i, and with initial values of $\hat{\beta}$ set to 0.

Now, to estimate μ a smoother matrix $\mathscr{M}$ is defined using kernel approximation.
We have used the Epanechnikov kernel, defined with respect to the elements of $\mathbf{T}$

$$K_{\mathbf{h}}(X_j - X_i) = \prod_{w=1,2} (1/h_w)(3/4)(1 - ((X_{jw} - X_{iw})/h_k)^2) \cdot (\|(X_{jw} - X_{iw})/h\| \leq 0.75) \tag{9.8}$$

Here $h_w = 0.4$ and $w = 1, 2$ because $\mathbf{T}$ is $k \times 2$, with $K = 427$ in the training phase and $K = 498$ in the testing phase. Note that the kernel is evaluated with respect to the matrix $\mathbf{T}$, mentioning all the trials both in the training and testing phases. Then the smooth matrix $\mathscr{M}$, according to Müller (2001), can be formed by the following elements κ_{ij} of $\mathscr{M}$:

$$\frac{(\mathcal{L}''(Y, \mu_j))K_{\mathbf{H}}(X_j - X_i)}{(1/n)\sum_{j=1}^{n}(\mathcal{L}''(Y, \mu_j))K_{\mathbf{H}}(X_j - X_i)} \tag{9.9}$$

Convergence is achieved when difference of likelihood and the estimates of β is below a certain threshold τ. We used $\tau = 0.1E\text{-}004$ and for our set of trials 48 iterations were needed to converge on the data train set with $K = 427$.
On data test the error is 0.4%. The error rate is, indeed, very low, as shown in the ROC curve displayed in Figure 9.6, reporting the behaviour of the estimator on data test.
Figure 9.7 shows, instead, a qualitative evaluation of a speaker identity estimation performance over 450 time-steps. The number of false alarms in the figure is pretty low; the implementation of a dynamic filter able to remove such false alarms could be, then, particularly useful for improving the general performance of the system. With this aim a future analysis can be executed about the nature of the wrong estimations obtained to design an adequate filter.

9.6 Updating

One main problem for the online application of the system is the knowledge base dimension. If the knowledge base is large, then online acquisition for the voice descriptors and the visual descriptors, concerning the two distances, is a quite hard task. Indeed, it requires a huge set of comparisons, since nothing is known about the people in the scene. So the question is: is there a time t at which the system knows who is in the scene and can rely on that for online identification?
Experiments show that a time t at which all people have spoken is difficult to predict, and if no constraint is put on the scene, some people can leave and new people can join. Thus there is not a fixed set that can be devised after a specified time.
To solve this problem and induce a partial knowledge of the people in the scene, we assume that changes are smooth: not all current people suddenly disappear nor are substituted altogether with new ones. So in a time lapse at most one person joins the conversation and one leaves, and partial updates can be inferred for voice and face similarities acquired up to time T.
More specifically, for the same effective speaker, the list $\hat{S}$ of the most probable relative labels, estimated via the acoustic analysis, tends to involve the same ones. After a specified time T (burning period), clusters of different cardinality, for each different list $\hat{S}$, are generated, with the associated frequency of occurrence.
Thus, if at time $t > T$ there are δK new people, with $\delta K \geq 2$, in the list, only the most probable labels $\hat{S}_{mp}$ are maintained, while the others are replaced with the labels mentioned in the most likely cluster, of the same cardinality. This includes $\hat{S}_{mp}$, according to the likelihood computed after the burning period.
These clustering on voices is integrated with an analogous clustering on visual distances and are thus used for setting a dynamic model in which known states and unknown new states are devised.

9.7 Future Improvements

The model proposed have some limitations, linked to the ability of the system to adapt itself to the dynamic evolution of the scene and to be, consequently, much less dependent on the characteristics of the scenario. At the state of the art, the system is able to combine different perception modalities and very different kinds of data, but it works only on a predefined set of people in the knowledge base and it does not sufficiently consider the experience after the first learning step. The system is able to manage people leaving or joining the conversation, but these people should be part of a predefined dataset, so that the robot could know their auditory and visual features of interest.
A solution could be to make the entire model Bayesian, so that it would be possible to deal with new subjects that would like to be included in the conversation and update

the current models using the information obtained during the previous steps.
An interesting cue, by this point of view, is provided in Richardson and Green (1997) with regards to the estimation of the parameters of the Gaussian Mixtures. They propose a fully Bayesian mixture modelling, analysing the number of components and the mixture components parameters jointly and base inference about these quantities on their posterior probabilities.
Moreover, they obtain posterior distribution of their object of inference, (model parameters and predictive densities) and not just best estimates.
However, this model can be improved and made more general and flexible, varying the structures of the priors and their dependences, the MCMC sampling method used, etc. In this way, it can be exploited to represent features with different types of distributions that are able to evolve on the basis of the new knowledge acquired by experiments.

It would also be interesting to add a speech recognition system to the actual platform, to implement the possibility for attention to be driven also by the content of the speech.

The various components of the system work simultaneously and real-time. Thus, it could be useful to take advantage of specific methodologies to arrange the design of the parallel elaboration executed by the system. For example, according to *PCAM* (see Foster (1995)) the design of parallel algorithms can be broken down into four major steps:

Partitioning : decomposition of the process into small tasks;

Communication : the necessary communication to coordinate task execution is determined;

Agglomeration : the decomposition and the communication structure defined in the previous two steps are evaluated on the basis of the implementation cost and other particular requirements;

Mapping : assignment of the various tasks to each processor, trying to maximize processor utilization and minimize communication costs.

But other strategies could be used in this direction, like *cloning*. This was first introduced by Von Neumann (1956) to generate fault-tolerance mechanisms and implemented over the years for different purposes and in different scenarios. *Simulation cloning* is particularly interesting. It allows the creation of different copies of a running simulation and each copy follows a different evolution of the process in a parallel way. Decision points are established to determine the moment in which the clones can be generated and start their execution (for more details about cloning simulations, see Hybinette and Fujimoto (1997)).

Finally, thanks to new tools and low-cost devices like, for example, the Microsoft Xbox Kinect Corp. (2011), the system could be improved and enriched with other functionalities, in order to make the robot a more realistic conversationalist. The visual aspects, above all, could benefit from the integration with these new instruments. Gestures recognition and pose estimation could be implemented using the Kinect (seeShotton *et al.* (2011); Ren *et al.* (2011) among others), so that the robot would be able to establish another way of communicating to humans and use it to obtain a better interpretation of the scene and a more complete human-like reaction. Hands, arms and, more generally, body movements could, in fact, show fast commands, for example, or even emotions, if supported by a model able to analyse body language and to associate some actions to the relative emotions those actions aspire to express (see, e.g., Bianchi-Berthouze and Kleinsmith (2003); Castellano *et al.* (2008)) .

Part V

Conclusions, future directions and References

Conclusions and future directions

Top-down auditory attention has been investigated since Colin Cherry, in 1953, performed his first experiments, but the computational models proposed to imitate human behaviour, largely, do not aim to be general representation of the mechanisms involved. Rather they are domain dependent and hard to adapt to other contexts, providing ad hoc solutions to particular problems.
We presented a new interpretation of top-down attention as an active decision problem, in which just some information is initially available and we want to know what the best measurement is to execute next to improve the performance of the decision making procedure. The main novelty of the idea is relative to the use of a generative model, given by a Gaussian mixture, which, thanks to its level of generality, can be used in several domains or applications.
The saliency is computed according to an informational theoretic approach, based on an information maximization intent.

The *cocktail party effect* that makes people, at a cocktail party, able to pay attention to the speech of the neighbour ignoring the other sounds and voices around inspired, over the years, researchers to execute different kinds of experiments. We performed behavioural experiments, rivisiting the ones carried out by Cherry in his pioneering work, to analyse the effect of cues like the presence of a task or priming. In particular, we checked how these cues affect speech intelligibility, simulating a cocktail party scenario as a monaural composition of two narratives, uttered by a speech synthesizer using the same virtual speaker. In this way, as Cherry suggested, other cues linked to the voices or so were deleted. The results of our experiments showed that both cues help the segregation between the two audio streams.
Moreover, we simulated a similar scenario to investigate the behaviour of the top-down model, inserting counfounders in the training set. The results are similar to the ones provided by the behavioural experiments.
We also investigated the missing data problem and the efficiency of some of the techniques mostly used in the literature to deal with this kind of issue. The most efficient method was, then, exploited to make the top-down attention model proposed able to operate in more realistic contexts, characterized by missing information also in the training set.

Finally, we explored multimodal perception as a fusion of auditory and visual data. Specifically, we combined these two different kinds of information to perform a speaker recognition process, in a conversation scenario. The robot, making use of this data combination, follows the conversation, estimating the position in the scene of its actual interlocutor and his/her identity. The results obtained seem to support the idea of multimodal combination of data as an encouraging and not so explored way to examine and model human perception.

Future Directions

Auditory top-down attention

Further studies could investigate human auditory top-down attention, carrying on the ideas discussed in Chapter 8.2.2 and taking care also of the aspects less analysed till now in the literature. In order to do this, behavioural experiments could be necessary and useful to understand patterns operating in human attentive comportment. In particular, it would be interesting to focus the research on the impact of factors like emotions, previous knowledge and particular acquired predispositions on auditory attention.
For example, human acoustic apparatus has a particular structure, due to the fact that the higher informative content of speech, the consonants, are characterized by specific frequencies. This proves the predisposition of the man paying attention to other human voices. However, over the years humans develop other particular predispositions. In time, people become more sensitive, for instance, to known voices, like a child being naturally more sensitive to the voice of the mother. This happens also following conferences, multi-speaker seminars and so on. In time, people are able to isolate the voice of the speakers also because they somehow "learn" these voices and the separation becomes easier.
In analogy with this, as we already mentioned and started to investigate, knowledge about the content of what they are hearing could help the segregation step, pushing subjects to pay attention to a speech containing a particular known topic (pre-defined or not), rather than another.
Also emotions play an important role in the decision about the saliency of each stimulus. Some words or sounds, for example, grab drastically human attention, due to awakening strong feelings, to alerting words and/or to specific emotions which characterize the sound.

Auditory bottom-up attention

It could be interesting as well to establish a measure of saliency for the different stimuli independently from the purpose of the application and of other top-down cues. In analogy to visual attention, saliency maps could be, then, generated and procedures proposed to deal with the interaction between all these aspects, in order to understand and represent whether and when shifting attention between an object to an other.
Consequently, studies about bottom-up attentive strategies could be useful to have a model able to combine both the attention components and provide a more general solution to the cocktail party problem, imitating human behaviour.

Multimodal Perception and Attention

Several models using sensing fusion techniques have been built for different tasks, whereof speaker or other sound source recognition and tracking are the most representative (Checka and Wilson (2002), Chen and Rui (2004), Zou and Bhanu (2005), Beal *et al.* (2002)). Integration between different attentive modalities is very much in its infancy, even if many studies suggest the existence of important similarities between auditory and visual perception in complex scene: both are based on the concept of perceptual objects as basic units of attention, both can modulate spatial and non-spatial feature processing and both operate in parallel through bottom-up, automatic, scene-based saliency examination and top-down, attentional, task-dependent examination (see Fritz *et al.* (2007)). Moreover, some experiments demonstrate that the mechanisms extracting conspicuous events from a sensory representation are similar in auditory and visual pathways (Kayser *et al.* (2005)).
However, there are some important differences. In Kawashima *et al.* (1999) functional areas of the brain modulating processing of selective auditory or visual attention toward utterances are analysed. Experiments prove that the visual task activates the visual association, inferior parietal, and prefrontal cortices. Both tasks activate the same area in the superior temporal sulcus. During the visual task, deactivation is observed in the auditory cortex. This behaviour indicates there exists a modality-dependent selective attention mechanism which activates or deactivates cortical areas in different ways.
All these results suggest that the two modalities act with similar procedures, but the analysis is executed separately. For this reason, it would be opportune to take advantage both of the affinity, creating multi-modal saliency maps, and their differences, building the complete map from the two maps generated by the auditory and visual analysis singularly, to increase the accuracy of the operations and to detect and delete the “false alarms”. The identification of false alarms is possible, since these two maps catch dissimilar characteristics of the scene. Visual maps proposed in the literature are generally three-dimensional and deal with the stimuli in the 3D space in the field

of view of the robot's camera. They do not consider all that is not included in this. Auditory saliency maps, instead, are bidimensional, but are able to capture information in the entire space: as it happens for human ears, it would be possible to hear anything in the scene, without considering the relative location (of course, it depends on the distance between the source and the listener). It is possible to assert that the nature of the acoustic devices allows a spherical perceptual field to be created around the robot, in which privileged directions or limitations in the hearing process do not exist.

For the sake of having an idea about the impact of these architectures would have on the development of real applications, it is sufficient to consider, for example, how much the surveillances systems would become more robust and easier to use. CASSANDRA, a smart surveillance system created by Zajdel *et al.* (2007) is one of the first examples in this direction.

But the saliency identification is not the only task that would benefit from a multimodal approach. The perceptual objects formation and the selection procedure would take advantage of it. Visual information could help to distinguish two near sources that the audio would be not able to opportunely separate (as shown in Nakagawa *et al.* (1999)) or, vice versa, the nature of sounds received could suggest the presence of another person in the scene not noticed by video analysis.

Moreover a so generated multi-modal map, indicating the saliency of the visual and auditory objects in the scene, can be used to discover which kind of features have to be used, depending on the particular applications, to improve the performance of recognition and identification systems that combine different perceptual modalities.

In this perspective, preliminary works investigating sensing fusion techniques (like the one described in Chapter 9) could be considered as a base for the study of multimodal perception and for the development of multimodal attentive models able to combine both visual and auditory data and to explain the mechanisms resultant from this integration.

Bibliography

S. Ahmad and V. Tresp. Some solutions to the missing feature problem in vision. *Advances in Neural Information Processing Systems*, pages 393–393, 1993.

P.D. Allison. *Missing Data*. Quantitative Applications in the Social Sciences. Sage Publications, 2001.

J.R. Angell and A.H. Pierce. Experimental research upon the phenomena of attention. *The American Journal of Psychology*, 4(4):528–541, 1892.

C.M. Arrington, T.H. Carr, A.R. Mayer, and S.M. Rao. Neural mechanisms of visual attention: object-based selection of a region in space. *Journal of Cognitive Neuroscience*, 12(Supplement 2):106–117, 2000.

B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Signal Processing*, 24(3):201 – 212, 1976.

F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

F. Baluch and L. Itti. Mechanisms of top-down attention. *Trends in neurosciences*, 2011.

M.J. Beal, H. Attias, and N. Jojic. Audio-video sensor fusion with probabilistic graphical models. In *in Procedeeng of ECCV*, 2002.

M.A. Bee and C. Micheyl. The cocktail party problem: What is it? how can it be solved? and why should animal behaviorists study it?. *Journal of Comparative Psychology*, 122(3):235, 2008.

M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. Multimodal conversation between a humanoid robot and multiple persons. In *Proceedings of the Workshop on Modular Construction of Humanlike Intelligence at the Twentieth National Conferences on Artificial Intelligence (AAAI)*, 2005.

M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The at&t next-gen tts system. In *Joint Meeting of ASA, EAA, and DAGA*, pages 18–24. Citeseer, 1999.

N. Bianchi-Berthouze and A. Kleinsmith. A categorical approach to affective gesture recognition. *Connection Science*, 15(4):259–269, 2003.

J.W. Bisley and M.E. Goldberg. Neuronal activity in the lateral intraparietal area and spatial attention. *Science*, 299(5603):81, 2003.

J. Braun and D. Sagi. Vision outside the focus of attention. *Attention, Perception, & Psychophysics*, 48(1):45–58, 1990.

J. Braun. Visual search among items of different salience: Removal of visual attention mimics a lesion in extrastriate area v4. *The Journal of Neuroscience*, 14(2):554, 1994.

J. Braun. Divided attention: Narrowing the gap between brain and behavior. *The attentive brain*, pages 327–351, 1998.

A.S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1994.

D. E. Broadbent. *Perception and Communication*. Pergamon Press, 1958.

A.W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128, 2000.

N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in neural information processing systems*, 18:155, 2006.

D.S. Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109:1101, 2001.

W. Burgard, D. Fox, and S. Thrun. Active mobile robot localization by entropy minimization. In *eurobot*, page 155. Published by the IEEE Computer Society, 1997.

R.P. Carlyon, R. Cusack, J.M. Foxton, and I.H. Robertson. Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1):115, 2001.

R.P. Carlyon. How the brain separates sounds. *TRENDS in Cognitive Sciences*, 8(10), 2004.

G. Castellano, L. Kessous, and G. Caridakis. Emotion recognition through multiple modalities: face, body gesture, speech. *Affect and emotion in human-computer interaction*, pages 92–103, 2008.

N. Checka and K. Wilson. Person tracking using audio-video sensor fusion. In *Proceedings of the MIT Project Oxygen Workshop*. Citeseer, 2002.

Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proceedings of the IEEE*, 92(3):485–494, 2004.

E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *Journal pf Acoustic Society of America*, 25:975–979, 1953.

D.G. Childers, M. Hand, and J.M. Larar. Silent and voiced/unvoied/ mixed excitation(four-way),classification of speech. *IEEE Transactions on Acoustics, Speech, and. Signal Processing*, 37(11):1771–1774, 1989.

S. Choi, H. Hong, H. Glotin, and F. Berthommier. Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network. *Neurocomputing*, 49(1-4):299–314, 2002.

D. Comaniciu and P. Meer. Robust analysis of feature spaces: color image segmentation. In *CVPR*, 1997.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. In *IEEE TPAMI*, 2002.

C.E. Connor, H.E. Egeth, and S. Yantis. Visual attention: bottom-up versus top-down. *Current Biology*, 14(19):R850–R852, 2004.

Microsoft Corp. Redmond wa. kinect for xbox 360, 2011.

R. Cusack, J. Decks, G. Aikman, and R.P. Carlyon. Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):643, 2004.

E. De Boer and HR De Jongh. On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *The Journal of the Acoustical Society of America*, 63:115, 1978.

E. De Boer and C. Kruidenier. On ringing limits of the auditory periphery. *Biological cybernetics*, 63(6):433–442, 1990.

B. De Coensel and D. Botteldooren. Modeling auditory attention focusing in multisource environments. In *Proceedings of the Acoustics' 08 Conference*, page 3255, 2008.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

G. C. Desilva, T. Yamasaki, and K. Aizawa. Interactive experience retrieval for a ubiquitous home. In *ACM CARPE*, 2006.

J.A. Deutsch and D. Deutsch. Attention: Some theoretical considerations. *Psychological review*, 70(1):80, 1963.

A.R.T. Donders, G.J.M.G. van der Heijden, T. Stijnen, and K.G.M. Moons. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006.

R. Drullman and A.W. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America*, 107:2224, 2000.

V. Duangudom and D.V. Anderson. Using auditory saliency to interpret complex auditory scenes. *Journal of the Acoustical Society of America*, 121(5):3119, 2007.

J. Duncan and G.W. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

J. Duncan and G.W. Humphreys. Beyond the search surface: Visual search and attentional engagement. 1992.

J. Duncan, G.W. Humphreys, and R. Ward. Competitive brain activity in visual attention. *Current Opinion in Neurobiology*, 7(2):255–261, 1997.

D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi. Applying support vector machines to voice activity detection. In *Signal Processing, 2002 6th International Conference on*, volume 2, pages 1124–1127. IEEE, 2002.

C.W. Eriksen and J.D. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Attention, Perception, & Psychophysics*, 40(4):225–240, 1986.

J.A. Feldman and D.H. Ballard. Connectionist models and their properties. *Cognitive science*, 6(3):205–254, 1982.

J. Feldman. What is a visual object? *Trends in Cognitive Sciences*, 7(6):252–256, 2003.

I. Foster. *Designing and building parallel programs: concepts and tools for parallel software engineering*. Addison-Wesley, 1995.

A. Frank and A. Asuncion. Uci machine learning repository, 2010.

A.D. Friederici, K. Steinhauer, and S. Frisch. Lexical integration: Sequential effects of syntactic and semantic information. *Memory & cognition*, 27(3):438–453, 1999.

S. Frintrop, E. Rome, and H.I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6, 2010.

S. Frintrop. *VOCUS: A visual attention system for object detection and goal-directed search*, volume 3899. Springer-Verlag New York Inc, 2006.

J.B. Fritz, M. Elhilali, S.V. David, and S.A. Shamma. Auditory attention focusing the searchlight on sound. *Current opinion in neurobiology*, 17(4):437–455, 2007.

Z. Ghahramani and M.I. Jordan. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems 6*. Citeseer, 1994.

JA Gray and AAI Wedderburn. Grouping strategies with simultaneous stimuli. *The Quarterly Journal of Experimental Psychology*, 1960.

Y. Haitovsky. Missing data in regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 67–82, 1968.

W. Hamilton. Lectures on metaphysics and logic, ed. *HL Mansel and J. Veitch, i–iv (Edinburgh: Blackwood)*, 1859.

F.H. Hamker. Modeling feature-based attention as an active top-down inference process. *Biosystems*, 86(1-3):91–99, 2006.

L.K. Hansen, S. Sigurdsson, T. Kolenda, F.A. Nielsen, U. Kjems, and J. Larsen. Modeling text with generalizable gaussian mixtures. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 3494–3497. IEEE, 2000.

L.K. Hansen, S. Karadogan, and L. Marchegiani. What to measure next to improve decision making? on top-down task driven feature saliency. In *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2011 IEEE Symposium on*, pages 1–7. IEEE, 2011.

A.E. Hernandez, E.A. Bates, and L.X. Avila. Processing across the language boundary: A cross-modal priming study of spanish-english bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(4):846, 1996.

W. Hirst, E.S. Spelke, C.C. Reaves, G. Caharack, and U. Neisser. Dividing attention without alternation or automaticity. *Journal of Experimental Psychology: General*, 109(1):98, 1980.

P. Horton and K. Nakai. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 109–115. AAAI Press, 1996.

M. Hybinette and R. Fujimoto. Cloning: a novel method for interactive parallel simulation. In *Proceedings of the 29th conference on Winter simulation*, pages 444–451. IEEE Computer Society, 1997.

A. Hyvarinen, J. Karhunen, and E. Oja. *Independent component analysis*. J. Wiley, 2001.

F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics*, volume 17. pp. C, 1968.

L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Advances in neural information processing systems*, 18:547, 2006.

L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

L. Itti, G. Rees, and J.K. Tsotsos. *Neurobiology of attention*. Academic Press, 2005.

L. Itti. Visual attention. *The Handbook of Brain Theory and Neural Networks*, pages 1196–1201, 2003.

L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10):1304–1318, 2004.

L. Itti. Quantitative modelling of perceptual salience at human eye position. *Visual cognition*, 14(4-8):959–984, 2006.

W. James. *The principles of psychology*. Henry Holt and Co, 1890.

M. Jilka, A.K. Syrdal, A.D. Conkie, and D.A. Kapilow. Effects on tts quality of methods of realizing natural prosodic variations. In *Proc. ICPhS*, 2003.

P.I.M. Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *IPO Symposium on Hearing Theory. IPO, Eindhoven, The Netherlands*, pages 58–69, 1972.

W.A. Johnston and V.J. Dark. Selective attention. *Annual review of psychology*, 37(1):43–75, 1986.

W. A. Johnston and S. P. Heinz. Flexibility and capacity demands of attention. *Journal of Experimental Psychology*, 1978.

W.A. Johnston and S.P. Heinz. Depth of nontarget processing in an attention task. *Journal of Experimental Psychology: Human Perception and Performance*, 5(1):168, 1979.

W.A. Johnston and J. Wilson. Perceptual processing of nontargets in an attention task. *Memory & Cognition*, 8(4):372–377, 1980.

D. Kahneman. Attention and effort. 1973.

O. Kalinli and S. Narayanan. A top-down auditory attention model for learning task dependent influences on prominence detection in speech. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3981–3984. IEEE, 2008.

H.H. Kames. *Elements of Criticism*. 1732.

H.J. Kappen, M.J. Nijman, and T. van Moorsel. Learning active vision. *Industrial applications of neural networks*, page 193, 1998.

S.G. Karadogan, J. Larsen, M.S. Pedersen, and J.B. Boldt. Robust isolated speech recognition using binary masks. In *European Signal Processing Conference, EUSIPCO*, 2010.

S.G. Karadogan, L. Marchegiani, L.K. Hansen, and J. Larsen. How efficient is estimation with missing data? In *International Conference on Acoustics, Speech and Signal Processing*. IEEE Press, 2011.

S.G. Karadogan, L. Marchegiani, J. Larsen, and L.K. Hansen. Top-down attention with features missing at random. In *International Workshop on Machine Learning for Signal Processing*, 2011.

R. Kawashima, S. Imaizumi, K. Mori, K. Okada, R. Goto, S. Kiritani, A. Ogawa, and H. Fukuda. Selective visual and auditory attention toward utterances–a pet study. *Neuroimage*, 10(2):209–215, 1999.

C. Kayser, C.I. Petkov, M. Lippert, and N.K. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li. Voice activity detection using mfcc features and support vector machine. In *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, volume 2, pages 556–561, 2007.

U. Kjems, J.B. Boldt, M.S. Pedersen, T. Lunner, and D. Wang. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, pages 1415–1426, 2009.

E.I. Knudsen. Fundamental components of attention. *Annu. Rev. Neurosci.*, 30:57–78, 2007.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985.

C. Koch. Selective visual attention and computational models. *CNS/Bi*, 186, 2004.

J. Laarni. Allocating attention in the visual field: the effects of cue type and target-distractor confusability. *Acta psychologica*, 103(3):281–294, 1999.

D. LaBerge. Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance*, 9(3):371, 1983.

S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G.A. Fink, and G. Sagerer. Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 28–35. ACM, 2003.

J. Larsen, A. Szymkowiak, and L.K. Hansen. Probabilistic hierarchical clustering with labeled and unlabeled data. *International Journal of Knowledge Based Intelligent Engineering Systems*, 6(1):56–63, 2002.

N. Lavie. Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):451, 1995.

D.K. Lee, C. Koch, and A.J. Braun. Attentional capacity is undifferentiated: concurrent discrimination of form, color, and motion. *Attention, Perception, & Psychophysics*, 61(7):1241–1255, 1999.

G.W. Leibniz. Nouveaux essais sur l'entendement humain, edition originale dans, oeuvres philosophiques de mr. de leibnitz, amsterdam et leipzig 1765. *Nouveaux Essais*, 1765.

D.V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

T. Liu, S.D. Slotnick, J.T. Serences, and S. Yantis. Cortical mechanisms of feature-based attentional control. *Cerebral Cortex*, 13(12):1334, 2003.

J. Locke. *An essay concerning human understanding*. 1689.

M.T. López, A. Fernández-Caballero, M.A. Fernández, J. Mira, and A.E. Delgado. Visual surveillance by dynamic visual attention method. *Pattern Recognition*, 39(11):2194–2211, 2006.

S.J. Luck and M.A. Ford. On the role of selective attention in visual perception. *Proceedings of the National Academy of Sciences*, 95(3):825, 1998.

SJ Luck and SA Hillyard. The operation of selective attention at multiple stages of processing: Evidence from human and monkey electrophysiology. *The new cognitive neurosciences*, pages 687–700, 1999.

R. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, volume 7, pages 1282–1285, 1982.

D. M. MacKay. Lateral interaction between neural channels sensitive to texture density. *Nature*, 245:159–161, 1973.

N. Malebranche. 1674. recherche de la verite, 1764.

L Marchegiani and F Pirri. Selective attention for voice matching and recognition. *Women in Machine Learning Workshop, Co-located with NIPS*, 2009.

M. Marchegiani, F. Pirri, and M. Pizzoli. Multimodal speaker recognition in a conversation scenario. *Computer Vision Systems*, pages 11–20, 2009.

M.L. Marchegiani, F. Pirri, and M. Pizzoli. Toward the design of a conversational robot. Technical Report 7, Dipartimento di Informatica e Sistemistica (DIS), Sapienza Università di Roma, 2009.

L. Marchegiani, S.G. Karadogan, T. Andersen, J. Larsen, and L.K. Hansen. The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years. In *The tenth International Conference on Machine Learning and Applications (ICMLA'11)*. to be published, 2011.

G.W. McConkie and C.B. Currie. Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):563, 1996.

R. Milanese, H. Wechsler, S. Gill, J.M. Bost, and T. Pun. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 781–785. IEEE, 1994.

R. Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation*. PhD thesis, University of Geneva, 1993.

T.A. Mondor, R.J. Zatorre, and N.A. Terrio. Constraints on the selection of auditory information. *Journal of Experimental Psychology: Human Perception and Performance*, 24(1):66–79, 1998.

B.C.J. Moore and H. Gockel. Factors influencing sequential stream segregation. *Acta Acustica United with Acustica*, 88(3):320–333, 2002.

N. Moray. Attention and dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11:56–60, 1959.

M.C. Mozer. *The perception of multiple objects: A connectionist approach.* The MIT Press, 1991.

M. Müller. Estimation and testing in generalized partial linear models - a comparative study. *Statistics and Computing*, 11:299–309, 2001.

John C. Murray, Harry Erwin, and Stefan Wermter. Robotics sound-source localization and tracking using interaural time difference and cross-correlation. In *AI Workshop on NeuroBotics*, 2004.

I. Myrtveit, E. Stensrud, and U.H. Olsson. Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering*, pages 999–1013, 2001.

Y. Nakagawa, H.G. Okuno, and H. Kitano. Using vision to improve sound source separation. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 768–777. JOHN WILEY & SONS LTD, 1999.

K. Nakayama and M. Mackeben. Sustained and transient components of focal visual attention. *Vision Research*, 29(11):1631–1647, 1989.

W.J. Nash and Tasmania. Marine Research Laboratories. The population biology of abalone (haliotis species) in tasmania: Blacklip abalone (h. rubra) from the north coast and the islands of bass strait, 1994.

V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2049–2056. IEEE, 2006.

D. Navon and D. Gopher. On the economy of the human-processing system. *Psychological review*, 86(3):214, 1979.

D. Navon. Forest before trees: The precedence of global features in visual perception* 1. *Cognitive psychology*, 9(3):353–383, 1977.

E. Nemer, R. Goubran, and S. Mahmoud. Robust voice activity detection using higher-order statistics in the lpc residual domain. *Speech and Audio Processing, IEEE Transactions on*, 9(3):217–231, 2001.

D.A. Newman. Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, 6(3):328, 2003.

J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289, 1933.

J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333, 1937.

D.A. Norman and T. Shallice. Attention to action. *Consciousness and self-regulation*, 4:1–18, 1986.

D.A. Norman. Toward a theory of memory and attention. *Psychological review*, 75(6):522, 1968.

D.A. Norman. Memory and attention: An introduction to human information processing. 1969.

D.A. Norman. *Memory and attention*. John Wiley and Sons, 1976.

HC Nothdurft. Texture discrimination by cells in the cat lateral geniculate nucleus. *Experimental Brain Research*, 82(1):48–66, 1990.

D. Nuzillard and A. Bijaoui. Blind source separation and analysis of multispectral astronomical images. *Astronomy and Astrophysics Supplement Series*, 147(1):129–138, 2000.

D. Ognibene, G. Pezzulo, and G. Baldassarre. How can bottom-up information shape learning of top-down attention-control skills? In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 231–237. IEEE, 2010.

A. Oliva and P.G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41(2):176–210, 2000.

N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 3(1):13–24, 2004.

N. Ouerhani. *Visual attention: from bio-inspired modeling to real-time implementation*. PhD thesis, Université de Neuchâtel, Faculté des sciences, 2003.

P.L. Panum. *Physiologische Untersuchungen "uber das Sehen mit zwei Augen*. Schwers, 1858.

H.E. Pashler. *The psychology of attention*. The MIT Press, 1999.

M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra. A survey of convolutive blind source separation methods. In *Springer Handbook of Speech Processing*. Springer Press, 2007.

A. P. Pentland. Machine understanding of human action. In *M.I.T. Media Laboratory*, 1995.

D. Phillips. *Preparation Course for the TOEFL: Next Generation IBT*. Longman, 2006.

L.C.W. Pols. Spectral analysis and identification of dutch vowels in monosyllabic words. 1977.

M.I. Posner, C.R. Snyder, and B.J. Davidson. Attention and the detection of signals. *Journal of experimental psychology: General*, 109(2):160, 1980.

M.I. Posner. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25, 1980.

L.R. Rabiner and M.R. Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 54(2):297–315, 1975.

J. Ramirez, P. Yélamos, JM Górriz, and JC Segura. Svm-based speech endpoint detection using contextual speech features. *Electronics letters*, 42(7):426–428, 2006.

S. Reiter, S. Schreiber, and G. Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *IEEE ICASSP*, pages 294–299, 2005.

Z. Ren, J. Meng, J. Yuan, and Z. Zhang. Robust hand gesture recognition with kinect sensor. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 759–760. ACM, 2011.

D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE TSAP*, 3(1), 1995.

S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

P.L. Roth. Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3):537–560, 1994.

G. Rousselet, O. Joubert, and M. Fabre-Thorpe. How long to get to the" gist" of real-world natural scenes? *Visual Cognition*, 12(6):852–877, 2005.

D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581, 1976.

D.B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 519. Wiley Online Library, 1987.

M. Ruz, M.E. Wolmetz, P. Tudela, and B.D. McCandliss. Two brain pathways for attended and ignored words. *Neuroimage*, 27(4):852–861, 2005.

J.L. Schafer and J.W. Graham. Missing data: Our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

P.G. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195, 1994.

T.A. Severini and J.G. Staniswalis. Quasi-likelihood estimation in semiparametric models. *J. Amer. Stat. Assoc.*, 89:501–511, 1994.

T. Shallice. *From neuropsychology to mental structure*. Cambridge Univ Pr, 1988.

C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(7):379–423, 1948.

R.M. Shiffrin and W. Schneider. Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological review*, 84(2):127, 1977.

B.G. Shinn-Cunningham. Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5):182–186, May 2008.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, volume 2, page 3, 2011.

J.W. Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

E. Spelke, W. Hirst, and U. Neisser. Skills of divided attention* 1. *Cognition*, 4(3):215–230, 1976.

J.V. Stone. *Independent Component Analysis. A tutorial introduction*. The MIT Press, 2004.

J.R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, 121(1):15, 1992.

T. Su and J.G. Dy. In search of deterministic methods for initializing k-means and gaussian mixture clustering. *Intelligent Data Analysis*, 11(4):319–338, 2007.

E. Sussman, W. Ritter, and H.G. Vaughan. An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology*, 36:22–34, 1999.

A.K. Syrdal and Y.J. Kim. Dialog speech acts and prosody: Considerations for tts. In *Proceedings of Speech Prosody*, pages 661–665, 2008.

J.G. Taylor and N.F. Fragopanagos. The interaction of attention and emotion. *Neural Networks*, 18(4):353–369, 2005.

A Torralba, M.S. Castelhano, A. Oliva, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113:2006, 2006.

A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

A. M. Treisman. Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12:242–248, 1960.

J.K. Tsotsos, S.M. Culhane, W.Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1-2):507–545, 1995.

J.K. Tsotsos, Y. Liu, J.C. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou. Attending to visual motion. *Computer Vision and Image Understanding*, 100(1-2):3–40, 2005.

J.K. Tsotsos. An inhibitory beam for attentional selection. In *Spatial vision in humans and robots: the proceedings of the 1991 York Conference on Spatial Vision in Humans and Robots*, page 313. Cambridge Univ Pr, 1993.

M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proc. of IEEE CVPR*, pages 586–591, 1991.

J. Van Den Berg, A. Curtis, and J. Trampert. Optimal nonlinear bayesian experimental design: an application to amplitude versus offset experiments. *Geophysical Journal International*, 155(2):411–421, 2003.

J. Von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, 34:43–98, 1956.

W. Von Tschisch. Ueber die zeitverhaltniss der apperception einfacher und zusammengesetzer vorstellungen. *Phil. Stud*, 2:603–634, 1885.

A. Waibel, T. Schultz, M. Bett, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang. Smart: the smart meeting room task at isl. In *Proc. of ICASSP*, pages 752–755, 2003.

D.L. Wang, U. Kjems, M.S. Pedersen, J.B. Boldt, and T. Lunner. Speech perception of noise with binary gains. *The Journal of the Acoustical Society of America*, 124:2303, 2008.

D. Wang. On ideal binary mask as the computational goal of auditory scene analysis. *Speech separation by humans and machines*, pages 181–197, 2005.

S. Waugh. *Extending and benchmarking cascade-correlation*. PhD thesis, Computer Science Department, University of Tasmania, 1995.

C.D. Wickens. The structure of attentional resources. *Attention and performance VIII*, 8, 1980.

W. Wiegerinck, B. Kappen, and W. Burgers. Bayesian networks for expert systems: Theory and practical applications. *Interactive Collaborative Information Systems*, pages 547–578, 2010.

J.M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

J.M. Wolfe. Visual search in continuous, naturalistic stimuli. *Vision Research*, 34(9):1187–1195, 1994.

C. Wolff. *sychologia empirica, methodo scientifica pertractata, qua ea, quae de anima humana indubia experientiae fide constant, continentur et ad solidam universae philosophiae practicae ac theologiae naturalis tractationem via sternitur (Empirical Psychology, Treated According to the Scientific Method)*. Renger, Frankfort & Leipzig, 1732.

C. Wolff. *Psychologia rationalis*. Renger, Frankfort & Leipzig, 1734.

S.N. Wrigley and G.J. Brown. A computational model of auditory selective attention. *Neural Networks, IEEE Transactions on*, 15(5):1151–1163, 2004.

W. Wundt. Grundz
"uge der physiologischen psychologie (engelmann, leipzig). *NAMEN-UND SACHREGISTER*, 1874.

S. Yantis. Goal-directed and stimulus-driven determinants of attentional control. *Control of cognitive processes: Attention and performance XVIII*, pages 73–103, 2000.

A.L. Yarbus. Eye movements during perception of complex objects. *Eye movements and vision*, 7:171–196, 1967.

W. Zajdel, JD Krijnders, T. Andringa, and DM Gavrila. Cassandra: audio-video sensor fusion for aggression detection. 2007.

R.J. Zatorre, T.A. Mondor, and A.C. Evans. Auditory attention to space and frequency activates similar cerebral systems. *NeuroImage*, 10:544–554, 1999.

F. Zheng, G. Zhang, and Z. Song. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16(6):582–589, 2001.

X.S. Zhou, D. Comaniciu, and A. Krishnan. Conditional feature sensitivity: A unifying view on active recognition and feature selection. 2003.

X. Zou and B. Bhanu. Tracking humans using multi-modal fusion. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005.