Redundancy Reduction for Computational Audition, a Unifying Approach

Paris Smaragdis

Bachelor of Music (magna cum laude)
Berklee College of Music,1995

Masters in Media Technology
Massachusetts Institute of Technology. 1997

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Ph.D. in Perceptual Computing
at the Massachusetts Institute of Technology

May 2001

Author

Paris Smaragdis
Program in Media Arts and Sciences
April 18, 2001

Certified by

Barry L. Vercoe
Professor, Program in Media Arts and Sciences
Thesis Supervisor

Accepted by

Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Redundancy Reduction for Computational Audition, a Unifying Approach

by

Paris Smaragdis

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on April 18, 2001
in partial fulfillment of the requirements for the degree of
Ph.D. in Perceptual Computing

Abstract

Computational audition has always been a subject of multiple theories. Unfortunately very few place audition in the grander scheme of perception, and even fewer facilitate formal and robust definitions as well as efficient implementations. In our work we set forth to address these issues.

We present mathematical principles that unify the objectives of lower level listening functions, in an attempt to formulate a global and plausible theory of computational audition. Using tools to perform redundancy reduction, and adhering to theories of its incorporation in a perceptual framework, we pursue results that support our approach. Our experiments focus on three major auditory functions, preprocessing, grouping and scene analysis. For auditory preprocessing, we prove that it is possible to evolve cochlear-like filters by adaptation to natural sounds. Following that and using the same principles as in preprocessing, we present a treatment that collapses the heuristic set of the gestalt auditory grouping rules, down to one efficient and formal rule. We successfully apply the same elements once again to form an auditory scene analysis foundation, capable of detection, autonomous feature extraction, and separation of sources in real-world complex scenes.

Our treatment was designed in such a manner so as to be independent of parameter estimations and data representations specific to the auditory domain. Some of our experiments have been replicated in other domains of perception, providing equally satisfying results, and a potential for defining global ground rules for computational perception, even outside the realm of our five senses.

Thesis Supervisor:
Barry Vercoe
Professor, Program in Media Arts and Sciences

Redundancy Reduction for Computational Audition, a Unifying Approach

Paris Smaragdis

Thesis Reader

Michael A. Casey
Research Scientist
Mitsubishi Electric Research Labs

Thesis Reader

Whitman Richards
Professor of Brain Sciences
Massachusetts Institute of Technology

Acknowledgements

Of course, taking full recognition for my thesis would be selfish! So, I'll keep the blame, but I'll spread the credit. Starting from my thesis committee, Barry Vercoe, also my advisor, has graciously let me work on whatever I desired throughout my time in MIT. This is probably the best gift an advisor can offer to a student, I can only hope my work has done some justice to this. Michael Casey, a long-time officemate and colleague, has been extra supportive of my ideas, even at times where I sounded like a nut, and he has been a great source of discussions, debates and vision. He is one of the few young people in my field to command my respect. Whitman Richards, for inspiring me to pursue my ideas and take a more formal approach to perception. He has had much more influence on me than he might think!

All the wonderful people in the machine listening group were also instrumental in developing this thesis. The 'elders', DAn Ellis, Bill Gardner, Mike Casey, Keith Martin, Eric Scheirer and Adam Lindsay, have all demonstrated extraordinary amounts of patience as I frequently knocked on their doors in search for enlightenment. My sincere apologies for bugging you all, and my deepest thanks for helping me start surviving at MIT. The 'young ones', Jonathan Feldman, Joe Pompei, Young-Moo Kim, Chai Wei, Ricardo Garcia and Rebecca Reich. Even though I frequently was the receiving end of their questions, I learned a lot more from this, than they think they did! Nyssim Lefford (aka mistress recycla), gets singled out from the latter group. She has been a great officemate and a source of really cool and challenging discussions (not to forget, the impetus of touching 80s pop music flashback moments!). I feel very lucky to have shared an office with her, I will sorely miss her as I move out. Judy Brown, has been my secret extra committee member. If it wasn't for her, my thesis would have about fifteen times as many mistakes! She has assumed the role of an audience type I was sure to oversee in my thesis, and for reminding me that I'm very grateful. Finally, Connie Van Rheenen for making sure we have something to eat and money to spend. One could not ask for a sweetest person to have around a group.

My work started taking off with the help of Malcolm Slaney and Tony Bell at Interval Research. I'm indebted to both of you, that summer helped me mature my thoughts and start working confidently towards my vision. Your technical and emotional assistance has been very valuable to me.

From the Greek mafia around MIT, Petros Boufounos and Elias Vyzas have been my trusty math advisors and impeccable friends. I hope I've managed to reciprocate. Kyratso Karahalios and Thodoras (aka Teddy D Dawg) Darmas, 'the couple at 310', provided many (endless) late night sessions playing 'age of empires', watching TV or going out, and sustained my sanity all this time. Carl (the Barbouni) Livadas and Maria Kartalou with their frequent mountain trips provided a much needed alternative to vegging out all weekend writing a thesis. Thanks guys!

Richard Boulanger, who transformed me from a hairy bass player at Berklee to a, umm ..., less hairy MIT geek (although women don't quite fall as much for the new persona, I still consider this to be some sort of improvement). Richard's faith to me and his true desire to help his students has made him one of the best people I've ever had the privilege of knowing.

Finally, an extra special thanks to my family for believing in me, and unconditionally helping me realize my dreams no matter how nebulous and premium they were. I dedicate this thesis to you.

Table of Contents

Chapter 1. Introduction 5

1 . 1 Introduction 5

1 . 2 Computational Perception 7

1 . 2 . 1 General Perception 7

1 . 2 . 2 Auditory Perception 9

1 . 3 Statistics for Perception 10

1 . 3 . 1 Decorrelation and Statistical Independence 10

1 . 3 . 2 Principal Components Analysis 13

1 . 3 . 3 Independent Components Analysis 18

1 . 3 . 4 Applications of ICA 22

1 . 4 Putting It All Together 24

Chapter 2. Auditory Preprocessing and Basis Selection 25

2 . 1 Introduction 25

2 . 2 Computation Background 25

2 . 2 . 1 Fixed Bases 26

2 . 2 . 2 Data-dependent Bases 33

2 . 2 . 3 Auditory Perception and Basis Decompositions 37

2 . 3 Environmental Statistics 40

2 . 4 Measuring Bases from Real Sounds 41

2 . 5 Conclusions 46

Chapter 3. Perceptual Grouping 49

3 . 1 Introduction 49

3 . 2 Perceptual Grouping 50
3 . 2 . 1 The Gestalt School 50
3 . 2 . 2 Auditory Grouping 51
a. Common Frequency/Amplitude Modulation 51
b. Common Onsets 53
c. Harmonic Relationship 53
d. Higher Level Cues 53
e. Streaming Cues 54
3 . 3 Grouping as redundancy reduction 55
3 . 3 . 1 The Theory 55
3 . 3 . 2 The Method 56
3 . 4 Harmonicity - Frequency Proximity 57
3 . 5 Common Fate 60
3 . 5 . 1 Frequency and Amplitude Modulation 60
3 . 5 . 2 Common onsets/offsets 62
3 . 5 . 3 Time Proximity 64
3 . 6 Prägnanz and Higher-order principles 65
3 . 7 Putting it all together 65
3 . 8 Conclusions 68

Chapter 4. Scene Analysis 71
4 . 1 Introduction 71
4 . 2 Source Separation and Scene Analysis 72
4 . 2 . 1 Psychoacoustic Separation 72
4 . 2 . 2 Multichannel Blind Source Separation 73
4 . 3 Decompositions of Magnitude Transforms 75
4 . 3 . 1 Object Detection 75

4 . 3 . 2 Applications in Music Processing 85
4 . 3 . 3 Source Separation 87
4 . 4 Conclusions 88

Chapter 5. Conclusions 89
5 . 1 Overview 89
5 . 2 Discussion 89
5 . 3 Future Directions 89

Bibliography 91

Chapter 1. Introduction

1.1 Introduction

Human perception has always been an exciting and mysterious field of study. Although other fields pertaining to human behavior are easy to rationalize and mimic with computers, perception is covered by the veil of subconsciousness. Trying to define what it means to perceive an image, getting used to background noise, or reminiscing a smell is not an easy task. Unlike higher-level thinking where individual steps can be rationalized and described, perceptual functions just happen without our active intervention. This very inability to describe our perception has made it a fruitful subject for debates and development of theories.

This thesis deals with a specific perceptual domain, the auditory domain. Although this is not a domain in the forefront of perceptual research, we believe it is a most appropriate one with which to study the foundations of perception. It is a domain that incorporates hard perceptual issues such as time, simultaneous processing and a strong aesthetic sense. It is also convenient for research since it takes place on a one dimensional space, thereby simplifying the computational and conceptual models. Most importantly it is not intimately linked to our way of thinking. Unlike the other prominent domain, vision, for which it is relatively clear to us what happens (we do tend to think in pictures after all!), in audition we lack this convenience. We have no grasp of what sounds are, and what it means to perceive them separately, we don't know how we listen, and only those trained in listening can start to vaguely describe sounds. Our state of ignorance when it comes to listening, makes it a fresh new ground on which to perform research without preconceptions and hasty judgements.

The goals of this thesis are multiple and highly intertwined. First and foremost is to bring a touch of mathematical formalization to the field of computational auditory perception and place it inside the grander picture of perception. Most auditory work has been quite isolated from general perception, yielding highly specialized theories, many of them based on a bootstrapped heuristic foundation. The resulting complexity from such poor descriptions has been a prohibiting factor in computational implementations. We will try to avoid this approach and instead deal with more solid and abstract formulations that are not audio specific, and provide a clear definition of our ideas. This will be done in hopes of providing a deeper insight about perception in general, and of providing something that can be easily implemented on a computer.

Mathematically modelling audition is certainly a daunting task which will have to address multiple processes and functions. To bypass this problem we will use a different approach to studying audition. We will take interest in finding the unifying basic principles of perception rather than examining different stages of perception with varying methodologies. To do so our approach will be to model the development of auditory perception. It is inarguable that perception did not always exist; it did evolve from some first principles, and it was somehow shaped through time. We will try to explore what it was that drove this evolution and hope to mimic it (in shorter time frames!). We are interested in the basis of perceptual evolution and its driving principles. This approach will hopefully shed more light on how perception can develop and give us a deeper, yet simpler, insight on what perception does. This evolutionary approach reveals an additional goal: to examine and stress the effect of the environment on our perception. Adopting the developmental approach we will see how it is that audition developed to be what it is, and how the statistics of its environment have shaped it. It is quite acceptable by now that the basic principles of perceptual development have been driven and latched onto the statistics of environmental stimuli. We will point out how these statistics have affected our development and, in some cases, speculate on some alternate scenaria.

With our experiments, we will construct simulations of the auditory functions, which will have evolved to be as such and not instructed to. We feel this is a very important point and we will stress it throughout this thesis. We do not wish to construct high performance systems that draw from our knowledge and experience. We want to make 'tabula rasa' systems that adapt to their stimuli and 'grow' to perform useful functions.

We will also not deal with a proper and accurate portrayal of human perception. We will draw inspiration from it and assume it is an example well worth copying. Our goal however is to construct artificial perception. We will investigate the suspected principles that shaped human perception, in hope of using this knowledge to construct its artificial counterparts. The motive for this investigation is the possible application of this knowledge, to the design of data processing machines. Machines that deal with data, such as cameras, microphones, text parsers or stock displays, machines that can benefit from their own intimate link of their environment, rather than our interpretation of it.

The theories we will comply with have long been developed in an abstract form, but have been never applied to listening, and are only recently starting to make a big impact on visual perception. Our work will address this gap in auditory research and demonstrate that listening functions that have been traditionally seen as unrelated do in fact perform the same kind of processing. We will cover three major fields of lower level listening, preprocessing, grouping and object segmentation, and provide a common underlying theory. In conclusion we will propose similar applications that span to higher-level listening processes that include music parsing and memory.

1.2 Computational Perception

1.2.1 General Perception

Measurements and explanations of perceptual functions are the focus of experimental psychology and psychophysics. Perceptual responses to a vast amount of stimuli have been documented and are our only objective window to this world. Based on this knowledge and with the use of some additional reasoning and theorizing, researchers in artificial intelligence have long tried to make computational implementations of perception. This type of work requires coming up with a mathematical formulation of particular perceptual functions and then translating it to computer code. The rigor of the mathematical formulation ranges from extremely involved modelling of neuronal responses to superficial binary decisions, and is a matter of research direction and goals. The entire field is too broad to adequately cover here, but we will examine a particular trend which is closely related to this thesis.

Of interest to our work are the writings of Horace Barlow (1959, 1961, 1989). His viewpoints epitomize a significant thread of perceptual research, which spans about a century. Starting from early researchers of perception, like Helmholtz and Mach, it was observed that the environment that we live in has been a major shaping factor to the development of our perception. Unfortunately at the time, there were neither the proper mathematical foundations, not the means to develop models of such observations. Upon the creation of information theory by Shannon in the 40's, a new language was developed with which perception researchers could express these ideas. Information theory provided the means to mathematically express the coding and transmission of data through channels. This work was soon intimately connected to the perception which was starting to be looked at as a data processing problem. One of the first to use information theory in the perceptual framework was Attneave (1954), who used the

principles of information, channel capacity and redundancy to express perceptual processes. His key observation was that the perceptually important information on natural images are the borders of the image. These are also the points where most information lies, rendering the rest of the image as redundant information. Barlow (1959, 1961, 1989), was expressing similar ideas and explored a possible structure of our neural system to perform such optimizations. He elaborated on Attneave's observation that sensory input redundancy is crucial to perception, an observation that it is a recurring principle throughout perception from the basis preprocessing stages to higher level cognitive functions. He suggested that the sensory mechanism strives for Mach's 'economy of thought' and achieves that by performing sparse or factorial coding. Sparse coding being a decomposition of a signal that yields as little energy as possible in the resulting coefficients, and factorial coding being a decomposition that strives for maximally independent output (these are two related problems that we will cover computationally in the succeeding sections). Barlow subsequently revised his idea and let go of the economy principle, he did however maintain the position that redundancy exploitation is very important. Following suit, Watanabe (1960) introduced the idea that inductive inference was related to the Minimum Description Length (MDL) principle. This is a principle closely related to Occam's razor and strongly linked to information theory. It states that among various possible models for describing a data set, the preferred one is the one that yields the shortest and most compact description.

All of these ideas were a central theme in the late 80's where research led by Linsker (1986a, 1986b, 1986c, 1988) and Field (1987), used statistics to understand the visual system at the neural level. From this work stemmed the opinion that the perceptual system is defined by adapting to its environment, an approach that was further elaborated by Atick (1991) by elegantly making use of information theory and ecological adaptation to explain sensory processing. Related experiments were conducted by Atick and Redlich (1990) and Redlich (1993), who experimentally explored the possibilities that our neural system exploits redundancy. Their simulations proved to be very successful, yielding feature maps that were in agreement with transfer functions of neural processing. Although convincing, these experiments were complicated due to limited development of related computational techniques and processing power by that time. More recently developments have come from Olshausen and Field (1996), Bell and Sejnowski (1997), and Hyvärinen and Hoyer (2000), who have used modern information theory optimization algorithms to sparsely analyze natural images. Their analyses, which strived for different interpretations of redundancy reduction, obtained decompositions which were very similar to the receptive fields our visual system uses. They have thereby made a very convincing argument that sparse and factorial coding are functions that are closely linked to perception, and its development. The other important point that came from their work was the fact that they employed natural image scenes to arrive to their results. Their systems were not trained on clean or synthetic data. They were then in a position to say that there is a strong relation between our visual system and its surrounding environment. As this field is now slowly opening up, further computational advances are gradually made that allow for more complex processing and better expression of these ideas on perception to develop.

The common thread through all this work, is that perception is an information processing system which could be analyzed with the appropriate theories. What makes this system coherent and possible to analyze, is the fact that it deals with a structured input.

Thus comes the second important observation highlighted in this work: we live in a highly structured world. If our perceptual system is indeed performing an information processing task, it would have to adapt to the statistical regularities of its environment. Our environment is biasing and shaping our perception, thus if we wish to examine perception itself we can start by examining the environment. In the case of computational perception, this means that it is worthwhile to implement perceptual systems that learn from their environment rather than try to emulate a set of observed perceptual behaviors.

1.2.2 Auditory Perception

Most of the auditory perception research is epitomized by the work of Bregman (1990). His book explains the workings of auditory scene analysis through a lot of careful experiments. An astounding mass of the computational auditory scene analysis research has been influenced by this work and applies direct translation of his observations onto computer programs. Such implementations aspire to construct a fairly complete listening system which could perform like a human listener, analyzing audio, making judgements on the existence objects and deal with short-term memory. The most notable work in auditory perception includes the systems build by Vercoe and Cumming (1988), Duda et al. (1990), Mellinger (1991), Cooke (1991), Brown (1992) and Ellis (1992, 1996). Most of these systems included a decomposition stage inspired by our hearing mechanisms, and then moved on to selectively group components of this decomposition to extract individual objects. Although these implementations were fairly successful in solving the problems they attacked, they did tend to give up in cases outside their expertise. That left a lot of researchers in this field hoping for more general models of computational audition.

It was soon quite evident that the even though Bregman's book is a very significant work in the field of experimental psychology, it does not reveal the foundations of audition. The fact that it is a collection of experiments and heuristic interpretations of isolated effects has been neglected, and many a researcher has used them as ground rules for audition. That use of experimental psychology observations has hindered the implementation of computationally robust systems, since it introduced fuzzy concepts such as 'similar', 'parallel', in an inherently deterministic and strict platform. The translation of these verbal descriptions to computer implementations is a daunting task, which often creates systems with pathological problems by design. This has created a research bias in the auditory research community, which has been lagging compared to visual research mainly due to lack of formal definitions and robust formulations. This is an issue that is now starting to draw attention, and we try to address in this thesis.

Aside from all the aforementioned research, there has been a significant amount of additional work on audio processing and analysis which can be related to human auditory processes. However, most of it is specific and does not fit a general framework of artificial perception, being instead mostly applications of pattern recognitions and signal processing. We will therefore not cover this work and the interested reader is referred to Kahrs and Brandenburg (1998) and Rhoads (1996) as a starting point.

1 . 3 Statistics for Perception

In this section we will present some of statistical concepts, especially as they relate to this dissertation and perceptual research.

1 . 3 . 1 Decorrelation and Statistical Independence

Especially after the observations by Barlow the concept of decorrelating or making sensory stimuli statistically independent became a central theme in computational perception. We will now deal with their definitions and present the relevant general purpose algorithms.

Decorrelation, also known as orthogonality or linear independence, is a well known and widely used principle in statistics. Two random variables x_1 and x_2 are said to be decorrelated when:

$$\mathrm{cov}(x_1, x_2) = E\{x_1 x_2\} - E\{x_1\}E\{x_2\} = 0 \tag{1}$$

where $E\{x\}$ is the expected value of x. The measure $\mathrm{cov}(x_1,x_2)$ is known as the covariance of x_1 and x_2. It is symmetric since:

$$\mathrm{cov}(x_1, x_2) = E\{x_1 x_2\} - E\{x_1\}E\{x_2\} = E\{x_2 x_1\} - E\{x_2\}E\{x_1\} = \mathrm{cov}(x_2, x_1) \tag{2}$$

and if $x_1 = x_2$ it equals the variance of x_1 (and x_2):

$$\mathrm{cov}(x_1, x_1) = E\{x_1^2\} - E\{x_1\}^2 = \mathrm{var}(x_1) = \mathrm{var}(x_2) \tag{3}$$

If x_1 and x_2 are correlated then $\mathrm{cov}(x_1, x_2) \neq 0$. If $\mathrm{cov}(x_1, x_2) > 0$, then x_1 increases as x_2 increases, whereas if $\mathrm{cov}(x_1, x_2) < 0$, x_1 decreases.

In order to evaluate whether a set of random variables are correlated or not we construct the *covariance matrix*, a matrix containing information about the covariances between all possible variable pairs. It is defined as:

$$\mathbf{V}(x_1, \ldots, x_N) = \begin{bmatrix} \mathrm{cov}(x_1, x_1) & \cdots & \mathrm{cov}(x_1, x_N) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(x_N, x_1) & \cdots & \mathrm{cov}(x_N, x_N) \end{bmatrix} \tag{4}$$

The diagonal elements cov(x_i, x_i), are equal to the variances of x_i, which are usually greater than zero. The off-diagonal elements, are the covariances which will be zero if our set of variables x_i are decorrelated. The covariance matrix is symmetric since we have cov(x_1, x_2) = cov(x_2, x_1), and it is also positive semidefinite.

In general decorrelation does not imply statistical independence. Statistical independence between a set of variables x_i exists if:

$$P(x_1, \ldots, x_N) = \prod_i P(x_i) \tag{5}$$

where $P(x)$ is the probability density function of x. We can visually inspect independence by plotting the product of the estimated marginal densities of x_i and comparing it to the estimated joint density of x_i. Figure 1 displays the real and reconstructed joint densities of a set of two dependent and two independent random variables.

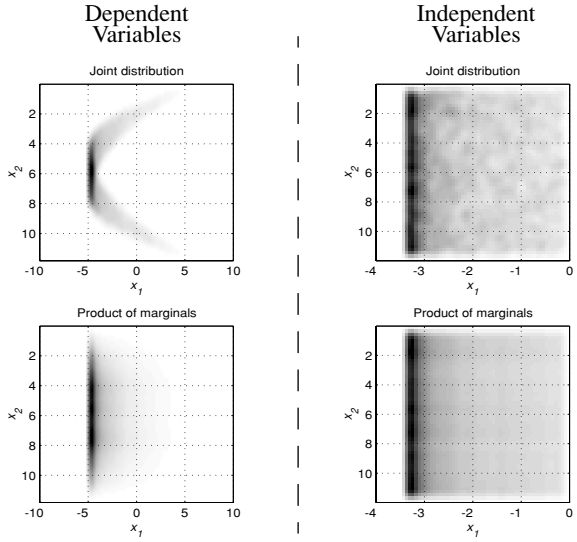

Figure 1

The two left figures display the joint probability density for two dependent variables, as measured from the data (top plot), and as calculated from the product of the marginals (bottom plot). Likewise the figures on the right, display the equivalent distributions of an independent set of variables. Note how the joint distributions of the dependent variables are different, an indication of dependence. The same effect does not take place for the independent variables.

This particular definition of independence is somewhat irksome and not very intuitive. An alternative condition more akin to decorrelation is defined for two variables x_1 and x_2 as:

$$E\{f(x_1)g(x_2)\} - E\{f(x_1)\}E\{g(x_2)\} = 0 \tag{6}$$

where $f(\cdot)$ and $g(\cdot)$ are measurable functions. This is the same condition as for decorrelation, with the added complication of the two functions. As is clear from this definition, independence implies decorrelation (the case where $f(x) = x$, $g(x) = x$), but the inverse does not hold. Decorrelation, by definition, denotes independence only up to second-order statistics. In the special case where the measured variables are drawn from the Gaussian distribution, decorrelation and independence are equivalent since the Gaussian distribution doesn't contain any information in orders higher than second.

Measurement of the amount of statistical independence between a set of variables can be done using many different approaches. The most straightforward one comes from the definition of statistical independence as presented in Equation (5) and the application of the Kullback-Leibler divergence (also known as the KL divergence or the KL distance), a measure of difference between two probability densities. The KL distance between two distributions $P(x)$ and $P(y)$ is defined as:

$$D(x \parallel y) = \int P(x)\log\left(\frac{P(x)}{P(y)}\right)dx \tag{7}$$

It assumes the value of zero if the two distributions are the same, and is a positive value otherwise. Using this and Equation (5) we can express the distance of the joint density of a vector $\mathbf{x}$ and the product of the marginal of its variates x_i. If these two sets of variables are independent their distributions will be the same and their the KL distance will be zero. This reasoning results in the formula:

$$D(\mathbf{x}) = \int P(\mathbf{x})\log\left(\frac{P(\mathbf{x})}{\prod P(x_i)}\right)d\mathbf{x} \tag{8}$$

This value is also known as the mutual information of $\mathbf{x}$ which is notated as $I(\mathbf{x})$. It is a measure of the common information between the variates of $\mathbf{x}$ and as, noted is positive and zero iff the variates of $\mathbf{x}$ are independent.

For practical reasons, the estimation of $P(\mathbf{x})$ is a hard and unreliable process. This means that for computational implementations the independence measures in Equation (5) and Equation (8) are not good candidates for optimization. To address this issue researchers have used alternative methods, one of the most popular being cumulant expansions. Such expansions involve a series approximation of the density functions by polynomials that include a set of easily estimated measures known as cumulants. These approximations can potentially extend to many terms prompting to a costly estimation. However through experience it has been found that using up to fourth order approximation the results can be satisfactory. Computationally a cumulant can be estimated by the k-statistics which are expressions of expectation values. They are defined as:

$$Cum\{x_1, x_2, \ldots, x_N\} = \sum (-1)^{p-1}(p-1)!E\left\{\prod_{i \in s_1} x_i\right\} \cdot E\left\{\prod_{i \in s_2} x_i\right\} \ldots E\left\{\prod_{i \in s_p} x_i\right\} \quad (9)$$

where the summation extends over all possible partitions $(s_1, s_2, ..., s_p)$ of all integers up to N. For example for $N = 2$ we have:

$$Cum\{x_1, x_2\} = E\{x_1 x_2\} - E\{x_1\}E\{x_2\} \quad (10)$$

since the only partitions of {1,2} are {1,2} and {{1} {2}}. From this definition it is easy to see that cumulants are symmetric functions with respect to their arguments. This particular case of $N = 2$, should be reminiscent of Equation (1), the expression of covariance. Second order cumulants are the covariances.

Even though the cumulants originate as density function approximations, they do necessarily not have to be used for this purpose. It is easy to show that when statistical independence exists between any two arguments of a cumulant, then the cumulant value equals zero (Comon 1994). Consider for example all of the second order cumulants. They are all expressed in the covariance matrix. The off-diagonal elements will be zero since independence implied decorrelation, and the diagonal will not. The off-diagonal elements, will be cumulants of two independent arguments, whereas the diagonal elements will not. Based on that observation and on that second and fourth order statistics are sufficient to approximately express independence, we can define the concept of the 'quadricovariance tensor' (Cardoso 1990, 1995a). This is a fourth-order tensor (can be loosely thought of as a four dimensional matrix), for which the element with the index $\{i,j,k,l\}$ is equal to $\text{Cum}\{x_i,x_j,x_k,x_l\}$. If we construct this tensor for two independent variables x_1 and x_2 we will only have nonzero values at indexes of the form $\{i,i,i,i\}$ (the diagonal of the tensor). In other words it will be diagonalized. This is the equivalent of the covariance matrix for fourth-order statistics. Similar extensions can be conceived for even higher orders, however the need for such hasn't arised yet since they do not seem necessary and their computation requires significantly more operations.

Other ways to evaluate independence have also been devised and tailored to different applications, however we feel that the ones presented are the most important for our purposes and we will not delve into others. In the following sections we will see how we can take advantage of these measures to impose decorrelation or independence on a set of variables.

1 . 3 . 2 Principal Components Analysis

Principal Component Analysis (PCA), is also known as the Karhunen-Loève transform (or KLT) or data whitening. It is a method to attain decorrelation between a set of variables by applying a linear transformation. Recall that decorrelation implies that the covariance matrix of our data will be diagonal. If we assume a zero mean set of variables x_i ordered as:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \tag{11}$$

then the covariance matrix will be defined as:

$$\mathbf{V} = E\{\mathbf{x} \cdot \mathbf{x}^T\} \tag{12}$$

We therefore need to apply a linear transformation on $\mathbf{x}$ that will result in a diagonal matrix $\mathbf{V}$. This is usually achieved by obtaining the eigenvectors of $\mathbf{V}$. If the matrix $\mathbf{U}$ contains the eigenvectors $\mathbf{u}_i$ of $\mathbf{V}$:

$$\mathbf{U} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{u}_1 & \cdots & \mathbf{u}_N \\ \downarrow & & \downarrow \end{bmatrix} \tag{13}$$

then by the properties of the eigenvectors

$$\mathbf{U}^{-1} \cdot \mathbf{V} \cdot \mathbf{U} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix} = \mathbf{\Lambda} \tag{14}$$

where λ_i are the eigenvalues of $\mathbf{V}$. Since $\mathbf{V}$ is positive semidefinite (all covariance matrices are), then the eigenvectors are mutually orthonormal[†], which, for real matrices, means that $\mathbf{U}^{-1} = \mathbf{U}^T$. Combining that with Equation (12) and Equation (14) we have:

$$\Lambda = \mathbf{U}^T \cdot E\{\mathbf{x} \cdot \mathbf{x}^T\} \cdot \mathbf{U} = E\{(\mathbf{U}^T \cdot \mathbf{x}) \cdot (\mathbf{U}^T \cdot \mathbf{x})^T\} \tag{15}$$

This means that by applying $\mathbf{U}^T$ on $\mathbf{x}$, we obtain a covariance matrix that is diagonal. So we have effectively decorrelated $\mathbf{x}$. We can additionally multiply $\mathbf{x}$ by $\sqrt{\mathbf{\Lambda}^{-1}}$ (where the square root operation applies element-wise), so that the resulting covariance matrix is

†. In general this property is not always true since $\mathbf{U}$ is not a unique eigenvector set; if multiplied by a constant it will still satisfy all conditions to be an eigenvector set. All standard numerical implementations though return the case where Equation (14) holds. For all other possible eigenvector sets the right hand side is multiplied by a constant. In either case the covariance matrix is diagonalized.

the unit matrix. If this extra step of normalization is applied then **x** is said to be whitened.

In practical terms the matrix **U** contains a set of vectors that point towards the directions of maximal variance of our data and **Λ** contains the variances in these directions. Applying $\mathbf{U}^T$ (= $\mathbf{U}^{-1}$) on **x**, rotates its joint distribution so that the directions with maximal variance are orthogonal to each other and coincide with the x_i axes. Additionally applying $\sqrt{\mathbf{\Lambda}^{-1}}$ will scale these variances to unity.

To illustrate this consider the bivariate joint distribution in Figure 2, generated by:

$$\mathbf{x} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \tag{16}$$

where n_i are two uncorrelated uniformly distributed random variables with zero mean and unit variance.

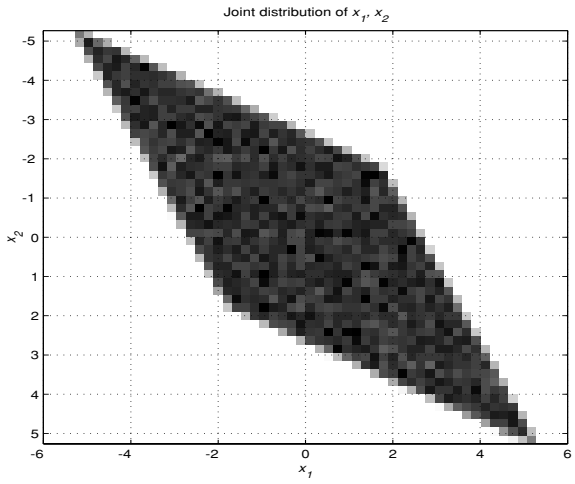

Figure 2 **The estimated bivariate joint probability density, of two correlated uniformly distributed random variables.**

Obviously since one element of **x** contains a portion of the other one, they will be correlated. The covariance matrix of **x** will be:

$$\mathbf{V} = E\{\mathbf{x} \cdot \mathbf{x}^T\} = E\left\{ \begin{bmatrix} n_1 + 2n_2 \\ 2n_1 + n_2 \end{bmatrix} \cdot \begin{bmatrix} n_1 + 2n_2 \\ 2n_1 + n_2 \end{bmatrix}^T \right\} =$$

$$= \begin{bmatrix} E\{n_1^2\} + 4E\{n_2^2\} + 4E\{n_1 n_2\} & 2E\{n_1^2\} + 2E\{n_2^2\} + 5E\{n_1 n_2\} \\ 2E\{n_1^2\} + 2E\{n_2^2\} + 5E\{n_1 n_2\} & 4E\{n_1^2\} + E\{n_2^2\} + 4E\{n_1 n_2\} \end{bmatrix} \quad (17)$$

The above is that due to the linearity of the $E\{\cdot\}$ operator we have $E\{x+y\} = E\{x\}+E\{y\}$ and $E\{ax\} = aE\{x\}$. We've also declared that $E\{x^2\} = 1$ (unit variance) and since n_1 and n_2 are decorrelated we also have that $E\{n_1 n_2\} = 0$. Combining all the above we have:

$$\mathbf{V} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \quad (18)$$

The eigenanalysis of $\mathbf{V}$ will be,

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix} \quad (19)$$

The set of vectors defined as the columns of $\sqrt{\mathbf{\Lambda}} \cdot \mathbf{U}$ will be pointing into the maximal variance directions (Figure 3).

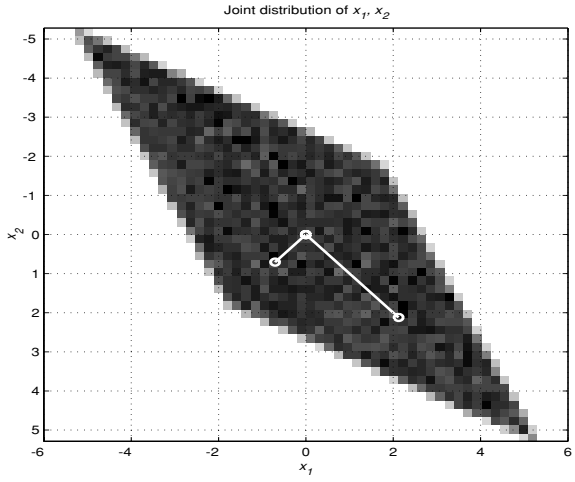

Figure 3 **The estimated bivariate probability density of x_1 and x_2 with the vectors of $\sqrt{\mathbf{\Lambda}} \cdot \mathbf{U}$ superimposed. As we can see there vectors coincide with the directions of maximal variance and have an analogous length.**

by applying $\mathbf{W} = \sqrt{\mathbf{\Lambda}^{-1}} \cdot \mathbf{U}^T$ on $\mathbf{x}$ we obtain a new random variable $\mathbf{y}$:

$$\mathbf{y} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ \frac{1}{3} & \frac{1}{3} \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = \frac{1}{\sqrt{2}} \cdot \begin{bmatrix} n_1 - n_2 \\ n_1 + n_2 \end{bmatrix} \tag{20}$$

The covariance matrix $\mathbf{V}_y$ of this linear transformation of $\mathbf{x}$ will be:

$$\mathbf{V}_y = E\{\mathbf{y} \cdot \mathbf{y}^T\} =$$

$$= \frac{1}{2} \begin{bmatrix} E\{n_1^2\} + E\{n_1^2\} - E\{n_1 n_2\} & E\{n_1^2\} - E\{n_2^2\} \\ E\{n_1^2\} - E\{n_2^2\} & E\{n_1^2\} + E\{n_1^2\} + E\{n_1 n_2\} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{21}$$

Which is the unit matrix, hence the variates of $\mathbf{y}$ are decorrelated and normalized to unit variance. The joint distribution of $\mathbf{y}$ is shown in Figure 4.

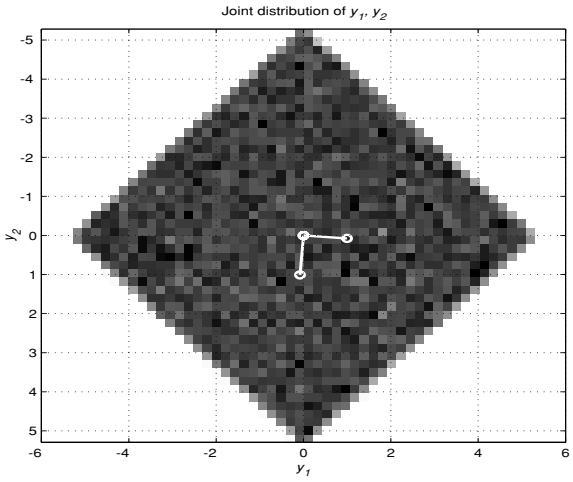

Figure 4 **The estimated probability density of the variables in Figure 2 after decorrelation. The covariance matrix of this linear transformation of x_1 and x_2 results in a diagonal matrix. Also note that the rotation which we achieve is the one that places the two maximal variance directions (denoted by the plotted vectors) on the axes of x_1 and x_2.**

Alternative algorithms for the estimation of the proper PCA rotation have been developed. The most notable set is that of adaptive (also known as online) algorithms, algorithms that estimate the rotation on a sample per sample basis, without using a

covariance matrix (Hertz et al. 1991). Such approaches are very useful in the case where the amount of data prohibits the calculation of the covariances, of when we work with real-time systems with little memory.

An alternative approach to PCA which produces similar results is to directly solve $\mathbf{V} = E\{\mathbf{x} \cdot \mathbf{x}^T\}$ to find the transformation we seek. This solution results in the transformation $\mathbf{W} = 2E\{\mathbf{x} \cdot \mathbf{x}^T\}^{-1/2}$. This is however a less common approach since it is computationally complicated (in particular the square root of a matrix is not an easy or always accurate computation), and it doesn't have the convenient semantic properties that PCA has. It has been used successfully though and has properties of its own, providing a symmetric transform (as opposed to the orthogonal transform of PCA), and a zero phase set of bases (Bell and Sejnowski 1996). This transform is referred to as Gaussian Components Analysis (GCA). In addition to PCA and GCA there are additional solutions to this diagonalization problem subject to varying constrains on the resulting transform, however they are not as common or useful and rarely appear in the literature.

Historically PCA has been a heavily used tool in statistics. It was used either as a mechanism to discover the 'interesting' directions of a multivariate distribution, or as a method to for data compression. Finding interesting directions has been one of the staples of pattern recognition where PCA is often used to discover significant features and axes of importance. The perceptual implications of PCA, and related methods, on sensory data were also examined by Linsker (1988), who paved the way to more the more sophisticated methods used today. In terms of compression and dimensionality reduction, by performing PCA on a set of data and discarding the variates with the minimal variance contribution we can reduce the amount of data and improve transmission and storage requirements, with limited degradation effect on the original data (in a least mean error sense). This approach has found a lot of applications on coding theory and signal processing and forms the basis of many industry standards on compression.

1.3.3 Independent Components Analysis

As seen from the above example, PCA performs a valuable task, it doesn't however return an even more desirable result. A result whose value uncovers the nature of $\mathbf{x}$ as it relates to a set of independent variables. Such an undertaking would imply an algorithm that not only decorrelates, but also makes the variates of $\mathbf{x}$ statistically independent. In the general linear model where $\mathbf{x}$ is defined as:

$$\mathbf{x} = \mathbf{A} \cdot \mathbf{n} \tag{22}$$

that process translates into finding the structure of $\mathbf{A}$ so as to recover the underlying data $\mathbf{n}$, by applying $\mathbf{W} = \mathbf{A}^{-1}$ on $\mathbf{x}$. The data $\mathbf{n}$ is in many cases a highly desirable discovery since it expresses $\mathbf{x}$ in the barest possible way. This problem has been studied rigorously and has provided a family of algorithms known as Independent Component Analysis (ICA).

In general most approaches to this problem implement the solution

$$\hat{\mathbf{n}} = \mathbf{W} \cdot \mathbf{x} \tag{23}$$

where the variates of $\hat{\mathbf{n}}$ were strived to be as independent as possible by fine tuning of $\mathbf{W}$. Since independence is defined in relation to either hard to estimate density functions (Equation (5)), or to a very large (if not infinite) set of constraints (Equation (6)), a process similar to PCA is hard to devise. These problems prompted researchers to use alternative ways to define independence, so as to be able to implement them.

One of the earliest examples of ICA at work was by Herrault and Jutten (1991), who devised a neural network that performed decorrelation of non-linear transformations of variables. In effect they imposed the constraint

$$E\{f(\hat{n}_1)g(\hat{n}_2)\} - E\{f(\hat{n}_1)\}E\{g(\hat{n}_2)\} = 0 \tag{24}$$

for a single set of the functions $f(\cdot)$ and $g(\cdot)$. Their results were highly dependent on a careful selection of these functions in relation to the statistical characteristics of the variables to decorrelate. They did however produce good results and effectively implemented a first practical, albeit limited, implementation of ICA. Later on, this approach was revisited by researchers working on non-linear PCA (Oja 1995). This is a form of PCA in which adaptive PCA algorithms were modified so as to decorrelate non-linear transformations of variables in hopes of achieving independence. This approach presented a unifying theory between PCA and ICA and provided a common platform for their analysis.

The first formal ICA definition came from Comon (1989), who coined the problem in terms of statistical independence and provided a solution based on mutual information minimization. As formulated by Comon, ICA attempts to minimize the mutual information of the output of the transformation in Equation (23). This definition contained the hard to estimate density functions, so Comon approximated mutual information using joint-cumulant approximations. In the process of doing so he proposed an algorithm and provided a wealth of information pertaining to this process. Cardoso (1993) also provided robust algorithms which were based on cumulants. He introduced the concept the ‘quadricovariance tensor’. By providing a diagonalization method for this tensor he performed an operation similar to PCA, but also for fourth order statistics, hence doing more than just decorrelation. Although technically this does not yield statistical independence (since there needs to independence for all orders, not just 2 and 4), for practical reasons it has proved to be satisfactory.

The seminal work on online ICA was presented by Bell and Sejnowski (1995), who were successful in formulating an online algorithm which provided relatively robust performance and efficiency. Their approach was termed information maximization (infomax), since it relied in maximization of joint entropy. Their network operated on a slightly modified version of Equation (23):

$$\mathbf{y} = f(\mathbf{u}) = f(\mathbf{W} \cdot \mathbf{x}) \tag{25}$$

where $f(\cdot)$ is an appropriately chosen function applied on all elements of its input (usually the logistic or the hyperbolic tangent). Maximization of entropy occurs on $\mathbf{y}$ by adaptation of $\mathbf{W}$. By doing so it effectively minimizes mutual information between the variates of $\mathbf{y}$ and provides the desired $\mathbf{W}$. The update rule assumed the very simple form:

$$\Delta\mathbf{W} \propto \mathbf{W}^{-T} + g(\mathbf{u}) \cdot \mathbf{x}^{T} \tag{26}$$

where the $^{-T}$ operator denotes inversion and transposition, and $g(\cdot)$ was applied to all elements of $\mathbf{u}$. The selection of $g(\cdot)$ is closely related to $f(\cdot)$ and is defined as:

$$g(u_i) = \frac{\partial}{\partial u_i} \log \left| y'_i \right| \tag{27}$$

Their results were obtained by observing that:

$$P(\mathbf{y}) = \frac{P(\mathbf{x})}{|J|} \tag{28}$$

and

$$H(\mathbf{y}) = E\{\log|J|\} - E\{\log P(\mathbf{x})\} \tag{29}$$

where $H(\cdot)$ is the entropy operator, $P(\cdot)$ the probability density function and J the jacobian of the transformation $f(\mathbf{W}\cdot\mathbf{x})$. From the equation above we can see that the effect of $\mathbf{W}$ on joint entropy is dependent solely on J. Hence maximization of entropy can be carried out by maximizing J. By carrying out the proper differentiations to ensure maximization by gradient descent, we obtain the rule in Equation (26).

Their approach was subsequently improved by Amari et al (1996), for more robust performance. They performed adaptation using a more proper form of gradient for the problem (known as the natural gradient) that allowed for better convergence characteristics, such as performance independent of the mixing conditions, and faster adaptation speed. The resulting rule was the original rule in Equation (26) right multiplied by $\mathbf{W}^{T}\mathbf{W}$:

$$\Delta\mathbf{W} \propto (\mathbf{I} + g(\mathbf{u}) \cdot \mathbf{u}^{T})\mathbf{W} \tag{30}$$

Similar results were independently obtained by Cardoso (1995b) who coined this form of gradient the relative gradient, and by McKay (1996) who recast ICA in the maximum likelihood context and employed Newton's method to obtain similar results. It is interesting to note in retrospect that this approach is effectively performing decorrelation between $g(\mathbf{y})$ and $\mathbf{x}$ since $\mathbf{\Delta W}$ becomes $\mathbf{0}$ when the product $g(\mathbf{y})\cdot\mathbf{x}$ equals $\mathbf{I}$. When that is the case we have the implication that $g(\mathbf{y})$ and $\mathbf{x}$ are decorrelated. Referring back to Equation (6) we see that this is a condition for likely statistical independence. Bell and Sejnowski (1995) note that a suitable choice for the function $f(\cdot)$ is the cumulative density function of the sources. Although there has been work on the optimization of $f(\cdot)$ in addition to the $\mathbf{W}$ matrix (Nadal and Parga 1994), this is not always necessary. It has been observed that if $f(\cdot)$ is a cdf of a super-Gaussian distribution (a peakier distribution than the Gaussian, exhibiting a positive kurtosis), then it is sufficient to perform ICA on any other super-Gaussian distribution, and like-wise for sub-Gaussian. Although this effect has not been explained so far, it is a fortunate one since application of ICA only requires knowledge of super- or sub- Gaussianity of the sources. For audio, due to its mostly super-Gaussian character, employing $f(\cdot)$ as the logistic or the hyperbolic tangent is sufficient for most cases.

The latest major algorithm to emerge is FastICA (Hyvärinen 1999), a fixed-point algorithm that exhibits fast convergence by employing an algebraic algorithm based on fixed-point iterations. This approach yields very fast results and ample flexibility, and has also risen to the forefront of ICA algorithms.

In order to build a better intuition on ICA we refer back to the numerical example we used in the previous section. Applying any of the ICA algorithms to the data in Figure 2, we obtain an different rotation. A resulting transformation was:

$$\mathbf{W} = ICA(\mathbf{x}) = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \tag{31}$$

which when applied on the input data obtains the original independent distributions we used to create it.

$$\mathbf{W}\cdot\mathbf{x} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \cdot \left(\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \right) = \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \tag{32}$$

This result[†] can be interpreted in a number of ways. One is that we have linearly transformed the data so as to make them statistically independent. This is equivalent to the notion of factorial coding, in which we desire an transformation that results in maximally independent output. We can also claim that we have discovered the underlying

model and data that generated our observations. A result of significance when we are concerned with the generative process of our data. The resulting $\mathbf{W}$ matrix is also worth examining since it provides a sparsifying transformation. In many cases knowledge of this transform is more important than the independent data.

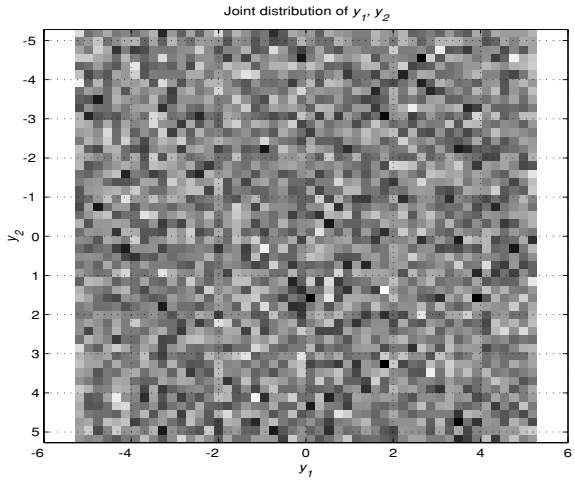

Figure 5 **The estimated probability density of the data in Figure 2 after application of ICA. We can see that we have managed to obtain an independent set of two uniform distributions.**

1 . 3 . 4 Applications of ICA

As should be evident, ICA performs a very important task in terms of statistics. It is a relatively new technique which is still being explored. So far its applications have been limited in scope, it does however contain a lot more potential (and part of our work is to present additional uses of ICA).

The most common application of ICA, up to the point of being virtually synonymous to it, has been blind source separation. Blind source separation is defined as the process where a set of linearly superimposed sources that are randomly mixed are recovered by only observing only their mixtures and having no knowledge about the mixing conditions. It is formulated by using a set of sources s_i and mixing them using what is referred to as the mixing matrix $\mathbf{A}$. This produces the mixtures x_i:

†. It should be noted that ICA algorithms are invariant to scaling and permutations. Another valid solution to our data would be any matrix of the form $\mathbf{L} \cdot \mathbf{P} \cdot \mathbf{W}$, where $\mathbf{L}$ is a diagonal scaling matrix, $\mathbf{P}$ a permutation matrix and $\mathbf{W}$ the result we have obtained above. This is not generally a problem though since we are often interested in just obtaining an independent set of outputs, and scaling and ordering are not as important.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix} \tag{33}$$

By using this formulation we can apply an ICA algorithm on **x** to obtain the unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ and use it for recovering **s** (Bell and Sejnowski 1995). This approach was initially used for auditory mixtures attempting to solve the scenario of *N* sources captured by *N* microphones. It was also applied on electroencephalographic (EEG) and magnetoencephalographic (MEG) data as well as to other electrophysiological data, in which case multisensor signals of neurological activity are recorder and a need for isolation of specific sources is key for subsequent analysis (Makeig et al 1996). Additional applications of this particular technique took place in other domains such as image, financial analyses, they are fields that have still to mature.

Blind source separation using ICA was also applied for signal deconvolution. This was done by observation that convolution is equivalent to mixing of time delayed versions of the same signal:

$$y(t) = \sum_{k=0}^{L} f(k)x(t-k) \Rightarrow \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} f(L) & 0 & \cdots & 0 \\ \vdots & & 0 & \vdots \\ f(1) & & & 0 \\ 0 & f(1) & \cdots & f(L) \end{bmatrix} \cdot \begin{bmatrix} x(1) \\ \vdots \\ x(N) \end{bmatrix} \tag{34}$$

Using this interpretation of filtering we can represent the filter as a mixing matrix, and recover its inverse with application of ICA on the *y* vector (Bell and Sejnowski 1995). By subsequent application of the inverse filter on **y** we can recover an estimate of the original signal **x**.

Due to the nature of sound mixing in the real world, the simple blind separation model was expanded to include delay propagations and source filtering, an approach that combined the two problems above. Although it is still a generally unsolved problem, under reasonable conditions, solutions are possible (Torkkola 1996, Smaragdis 1997, Lee 1997, Ikeda 1999). Effords in this field are exaustivelly covered by Torkkola (1999).

An additional application for which ICA was used was that of feature extraction. Although the ICA applications so far were concerned with obtaining a particular linear transformation of the input, the transformation itself is also worth examining. In the above examples it represents the inverse mixing and/or filtering conditions from the ones present. If we however supply arbitrary data as an input, then the transformation will be one that ensures a maximally sparse coding. In particular the obtained matrix will contain a set of sparsifying filters in it. There filters can be associated with detectors for extracting important features. This approach was applied for natural image coding and the results were of great perceptual importance since the discovered filters proved to be close to neurologically measured filters in our visual cortex (Olshausen and

Field 1996, Bell and Sejnowski 1997, Hyvärinen 2000). To further strengthen the value of this research, the approach agreed with the information redundancy theories on perception and vision which makes a plausible argument on environmental perception development.

Apart from separation and feature extraction, additional applications of ICA have been scarce so far. There has been some work on substituting PCA with ICA for coding purposes, and research that has produced denoising algorithms (Hyvärinen et al 2000).

1.4 Putting It All Together

The concepts that we have presented in this chapter are the backbone of this thesis. So far we have made a connection between perception and information processing, and described some popular algorithms that achieve data transformations which can reveal useful information about data and be paralleled with low level perceptual processes. In the succeeding sections, we will use these principles, and the inspiration from the work linking information theory and perception, to construct various perceptual-like functions in the auditory domain. These functions will be shaped by their environment and abide by a common and simple set of evolutionary rules. Our focus will be to make a machine behave similarly to how we do, but under these restrictions. The focus will be on constructing a system that learns to perform its function by data observation, and in contrast to past approaches for computational audition we will avoid any form of external influence or guidance.

Chapter 2. Auditory Preprocessing and Basis Selection

2 . 1 Introduction

In this chapter we will consider the preprocessing step of auditory systems. We will present an introduction to preprocessing and basis selection as it applies in both the perceptual and the computational domains. A recently introduced method for discovering bases for perceptual data will be applied in the auditory domain, and will be used to argue a connection between the statistics of natural sounds and the development of our hearing interface.

2 . 2 Computation Background

Basis selection is an important part of the design of analysis systems. Basis decomposition provides a translation of raw data to a collection of known functions. Having transformed our data to this known set of functions, which usually have some desirable properties, we can manipulate the original data with greater ease.

2 . 2 . 1 Fixed Bases

The use of basis functions came to light with the work of Jean Baptiste Joseph Fourier, early in the 19th century. His work on the expansion of functions into trigonometric series sparked the beginning of the field of functional analysis. Functional analysis deals with the translation of a function to a sequence of simpler functions that converge to it. Studies of linear transformations by Jacobi, Kronecker, Cayley and others, created an algebra dealing with operators that perform such translations and thus helped define basis decomposition for discrete sequences as we use it today.

The idea behind basis decomposition is fairly simple. Assume we have an unknown set of data like the one depicted in Figure 1.

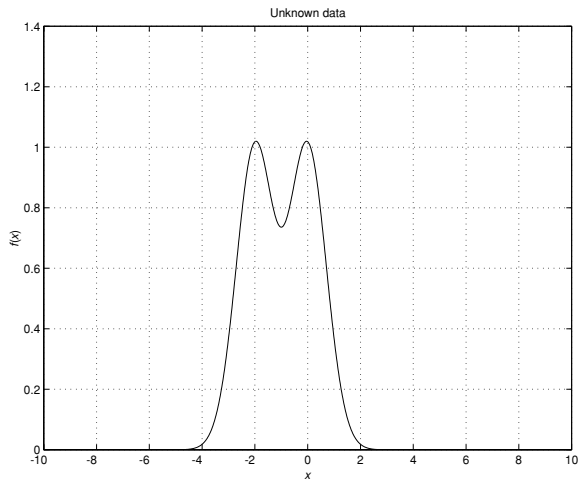

Figure 1 **An unknown function.**

Some calculations and mathematical manipulations on this curve might be hard because it exhibits an unknown structure. We have no list of properties to take advantage of, and no knowledge of its statistics. However, if we were to decompose it into known and simpler functions, we could instantly manipulate it with greater ease. These simpler functions will be the basis functions we will extract from that data. For the specific data above, a possible set of basis functions is the one shown in Figure 2. The superposition of these bases reconstructs the original data exactly.

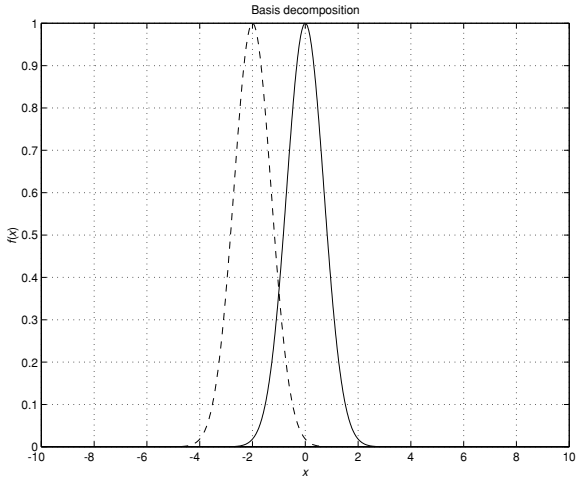

Figure 2 **Two basis functions of the data in Figure 1**

With our newfound knowledge that the initial function can be decomposed to the sum of the two functions e^{-x^2} and $e^{-(x+2)^2}$ we are now free to perform various analysis procedures taking advantage of their properties. Simplified integration/differentiation can be performed using linearity and the properties of simple exponentials, local minima can be discovered by examining the intersection of the two bases, etc. Although this was a simple example, in more complicated data sets basis decomposition is crucial for most analysis procedures.

It should be obvious here that there is never just one set of basis functions for any data. The bases used in the preceding example were exponential and were chosen as such because the original data set was created using them. When we deal with unknown data this is not the case, and the selection of a set of basis functions is determined by the nature of our problem, and the techniques we wish to apply to it. Even then, some basis functions have become more popular than others and are more frequently used.

One of the most popular set of basis functions are the ones originally introduced by Fourier. He was able to prove that any integrable function $f(x)$ can be decomposed to a sum of trigonometric functions scaled in amplitude by the scalars $\hat{f}(\xi)$:

$$\hat{f}(\xi) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-ix\xi} f(x)dx \tag{1}$$

In the discrete case, we use the Discrete Fourier Transform (DFT), which is defined as:

$$\hat{\mathbf{f}} = \mathbf{F}_N \cdot \mathbf{f} \tag{2}$$

where $\hat{\mathbf{f}} = \left[\hat{f}(1) \cdots \hat{f}(N)\right]^T$, $\mathbf{f} = \left[f(1) \cdots f(N)\right]^T$ and $\mathbf{F}_N^{(k,n)} = e^{-\frac{2\pi i}{N}kn}$. The latter matrix is called the Fourier matrix or DFT Matrix.

His work was found to be invaluable for the analysis of time-series, and through the work of Norbert Wiener (1949) and Andrey Nikolayevich Kolmogorov it has become a staple of communications and signal processing sciences. Figure 3 illustrates a simple function and its corresponding discrete Fourier transform. As is evident the Fourier transform gives us a clear interpretation of the signal it terms of frequency ordered sinusoids.

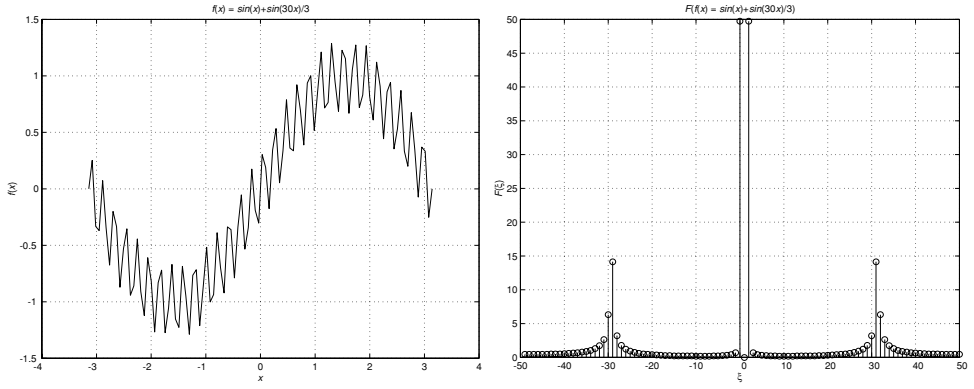

Figure 3 **The left figure displays the function $\sin(x) + \frac{1}{3}\sin(30x)$. The magnitude of its Fourier analysis is shown in the right figure. The stems signify the amplitude of each trigonometric basis, as we expect the bases with the strongest amplitude are the ones with frequencies mapping to 1 and 30 (note that the output of the magnitude of the Fourier transform is symmetric about 0, hence the four peaks).**

With the advancement of telecommunications and signal processing theory, a number of variants of the Fourier transform were introduced, most notably the Cosine transform, the Sine transform, the Haar transform and others. Of these, the Cosine transform (Ahmed 1972) was found to be very popular, since it was purely real (unlike the Fourier transform which contains imaginary parts) and because it had a straightforward interpretation, as well as some beautiful properties. In discrete time it also has a matrix formulation same as the Fourier transform (Equation (2)), in which the transformation matrix is:

$$\mathbf{C}_N^{(i,j)} = \begin{cases} \frac{1}{\sqrt{2}}\cos\left(\left(j+\frac{1}{2}\right)\frac{i\pi}{N}\right) \text{for } i = 0 \\ \cos\left(\left(j+\frac{1}{2}\right)\frac{i\pi}{N}\right) \text{for } i > 0 \end{cases} \quad (3)$$

As we will see later the Discrete Cosine Transform (DCT) has a special property in decomposing some time series such as sounds.

Although the aforementioned transforms have proved to be very useful, they lack the ability to track the evolution of bases across time. Consider as an example the signal in Figure 4; it exhibits a simple amplitude modulation of a single Fourier base, but the Fourier transform doesn't reveal that in any intuitive way. It returns an interpretation that contains many neighboring sinusoids, whose sum creates a beating effect resulting in the original amplitude modulation. However in terms of bases we would rather have an interpretation that a single base is amplitude modulated through time. In other words, we wish to introduce the time dimension in the Fourier transform.

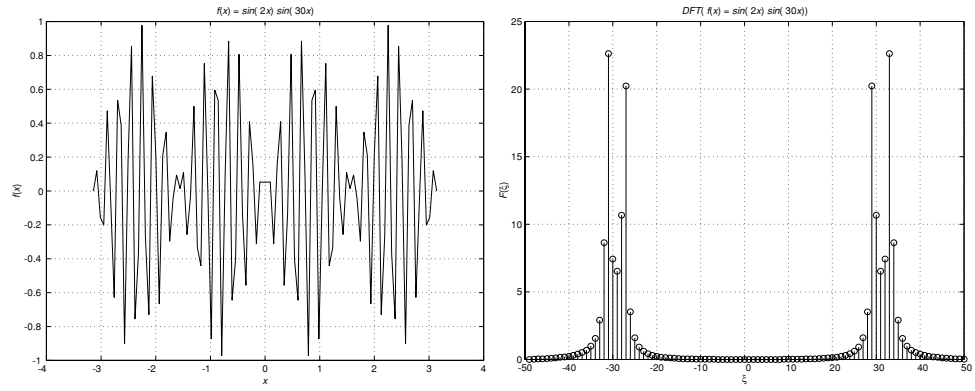

Figure 4 **A constant-frequency time-modulated signal has a Fourier transform that obscures its true structure.**

To overcome the time insensitivity illustrated above, the Short Time Fourier Transform (STFT) was introduced by Gabor in 1946. The STFT performs Fourier transforms on a moving time window across our data (Figure 5).

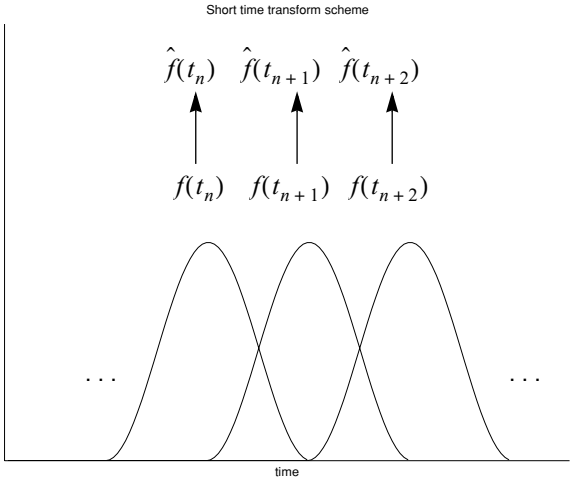

Figure 5 **Short Time Fourier Transform process. A set of windows are positioned on the time axis, selecting time-localized parts of the function to be analyzed. They are then individually Fourier-transformed and result in time-localized Fourier analyses.**

Thus for each time instance t we have a corresponding Fourier transform $\hat{f}(\xi)$ decomposing just the data from $x(t)$ to $x(t+T)$. For practical reasons that data can be optionally scaled by a window $w(t)$ before the Fourier transform, so as to remove some analysis artifacts caused by the, otherwise abrupt, boundaries. In the continuous time it is formulated as:

$$\hat{f}(\tau, \xi) = 2\pi^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-it\xi} f(t) w(t-\tau) dt \tag{4}$$

and in the discrete case:

$$\hat{\mathbf{f}}_\tau = \mathbf{F}_N \cdot (\mathbf{f}_\tau \otimes \mathbf{w}) \tag{5}$$

where $\hat{\mathbf{f}}_\tau = \left[\hat{f}(\tau, 1) \cdots \hat{f}(\tau, N)\right]^T$, $\mathbf{f}_\tau = \left[f(\tau+1) \cdots f(\tau+N)\right]^T$, $\mathbf{F}_N^{(k,n)} = e^{-\frac{2\pi i}{N}kn}$, $\mathbf{w}$ is the scaling window and the $\otimes$ operator denotes the Hadamard product (element-wise multiplication).

Using this system we only perform basis estimation for a given time neighborhood, thus improving the time localization of our decompositions. The rate by which the variable τ advances is dependent on how dense we want our decompositions to be in time. How-

ever it seldom exceeds the time span of the transformation, and is typically an integer fraction of the transform size. This type of transform is of course not only restricted to Fourier analysis. We could substitute the DFT matrix in Equation (5), with other transform matrices such as the DCT matrix. Revisiting the example in Figure 4, we can visually display the STFT of that time series (Figure 6). Unlike the plain Fourier analysis, the STFT forms a peak corresponding to the modulated frequency in our signal, which is modulated in time. This interpretation is more intuitive and easier to understand than the one given by Fourier.

This kind of joint time-frequency analysis, is also referred to as a spectrogram or a sonogram (Figure 7), and it has been an invaluable tool for harmonic analysis of sounds (Risset 1965). Just as the Fourier analysis is a great tool for making manipulations on the harmonic structure of a sound, the STFT has extended this facility on the time axis. Using the STFT it is easy to perform time-variable filtering and complicated convolution operations. A derivative of the STFT is the phase vocoder which has been a most valuable tool for the manipulation of speech (Flanagan and Golden 1966, Dolson 1986).

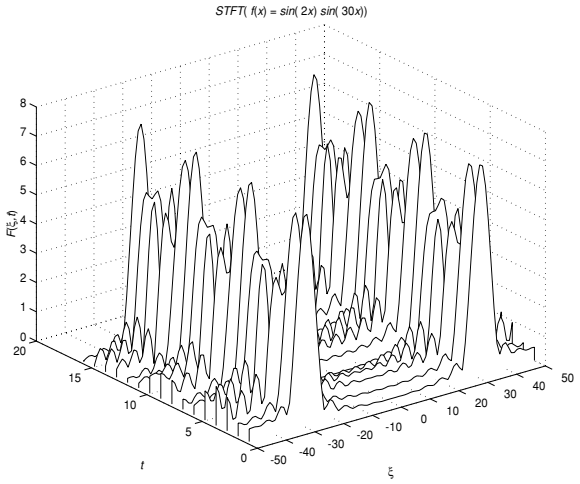

Figure 6

STFT analysis of the signal in Figure 4. Note how, unlike the Fourier transform, we get basis estimates in different times, thereby representing the nature of the signal in a more comprehensible way.

It should also be noted that the STFT is intimately linked with filterbanks. Filterbanks are sets of filters that operate concurrently on a signal to distribute its energy on a frequency axis. Most commonly they are used for spectral analysis by employing bandpass filters in series, a function equivalent to what the STFT performs. In fact the STFT is a filterbank (Allen and Rabiner 1977, Moore 1990) and it can be analyzed and treated as such.

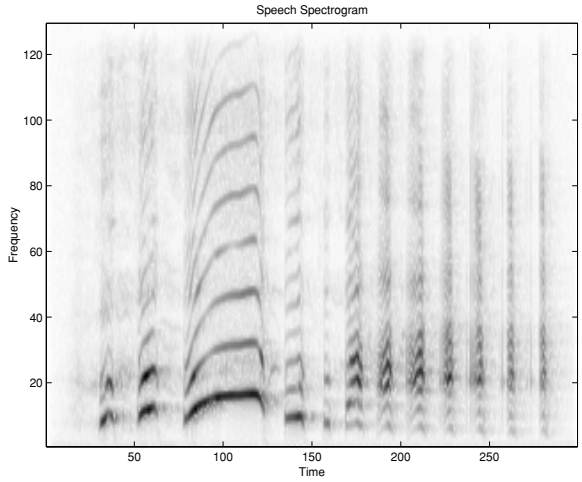

Figure 7

An STFT analysis of speech. Each column of this graph is an individual Fourier transform at a corresponding time point. Darkness denoted amplitude. Using this type of analysis we can visualize features which an ordinary Fourier transform would not display. In this example it is clear to see the formant structure of speech.

The move to composite time-frequency transforms opened the way for additional exploration that resulted in the field of time-frequency analysis. Heisenberg's uncertainty principle applies to this type of analysis, and it translates the momentum/position trade-off, to a time/frequency trade-off. It states that a higher frequency tracking resolution will lessen the time tracking resolution and vice versa. In particular, for a basis function $f(t)$, where $\|f\| = 1$ (f is unit norm), the time spread $\sigma = \|t \cdot f(t)\|$ and the frequency spread $\hat{\sigma} = \left\|\hat{f}(t)\right\|$, are related by:

$$\sigma\hat{\sigma} \geq \frac{1}{2} \tag{6}$$

Whenever equality exists the basis offers optimal resolution for both frequency and time (this occurs when $f(t) = e^{-t^2}$). However, in both the Fourier transform and by inheritance the STFT, as we examine bases of higher frequencies, $\hat{\sigma}$ remains the same whereas σ grows. That means that we are not performing an optimal detection for the higher frequencies. This is easy to understand intuitively, by observing that for a fixed time spread the lowest frequency will only occupy one period, hence we will only need to estimate its amplitude with one number. Higher frequencies, however, will contain many more periods whose mean energy will be what the DFT will return. We are obviously discarding information about the temporal changes within this frame, but if we were to make the time window smaller to compensate we would not be able to track the lower frequencies. This observation was the starting point of multiresolution analysis

theory, which culminated with the development of the constant-Q and wavelet transforms that address this problem. These transforms are still basis decompositions, albeit with more elaborate structure so as to avoid suboptimal resolution problems.

Of course harmonic decompositions presented in this section are not the only type of basis analyses, they are however the most popular and are indeed some of the most useful in the analysis of auditory signals. Other non-harmonic popular bases are radial basis functions, mixture models, sums of sigmoids (Haykin 1994). However, they have not yet been put to any significant use in the audio processing research literature, and they lack the intuitive appeal of the harmonic transforms.

2.2.2 Data-dependent Bases

Although basis decomposition to a set of known and well-understood bases is something to pursue, it is not necessarily the only way to perform analysis. Sometimes it is not the form of the bases that is important, but rather their mutual properties and their relation to the original data. A lot of these transforms derive the bases from the data given some property constraints. Usually these transformations are statistical in nature and are closely related to statistical analysis.

Linear algebra provides multiple examples of such basis decompositions that are defined by their properties rather than their form. One particularly useful example is the Principal Components Analysis (PCA, also known as the Karhunen-Loève transform, or KLT). In PCA the goal is to deduce from the data a set of basis functions which are orthogonal to each other and ordered in rank of variance contribution. They additionally have to apply a transformation to the data that results into decorrelated outputs. The significant difference of this approach from the transforms in the previous section is the fact that the basis functions will be derived by observation of the data.

Generally, when dealing with time series an appropriate way to apply any linear algebra decomposition is as follows. For a zero-mean discrete time series $f(t)$ we construct the vector $\mathbf{f}_\tau = \left[f(\tau+1) \cdots f(\tau+N)\right]^T$ where N is an integer equal to the time span of the bases we wish to discover. We can then construct a matrix containing the entire series:

$$\mathbf{F} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{f}_1 & \cdots & \mathbf{f}_M \\ \downarrow & & \downarrow \end{bmatrix} \tag{7}$$

Where M is the index of the last possible frame, and the temporal spacing of the columns can take place at an arbitrary rate (in some cases there is no need that the **f** vectors are even spaced in strict time order, nor that all possible frames exist). Having this representation, our goal is to find a set of basis functions that will be able to reconstruct

matrix **F**. We can proceed by analyzing this matrix according to our favorite linear algebra decomposition method†. Since PCA exhibits a lot of interesting features for sound analysis, we will proceed with its formulation in this framework.

As we have covered before, in order to perform PCA we need to obtain the covariance matrix of our data. As described above our data is now the matrix **F** (Equation (7)). The covariance we need is defined as:

$$\mathbf{C} = \mathbf{F} \cdot \mathbf{F}^T \tag{8}$$

We then proceed by factorizing **C** as:

$$\mathbf{C} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^T \tag{9}$$

where **Q** is an orthogonal and Λ a diagonal matrix. In **Q** we obtain an orthogonal set of bases $\mathbf{q}_i$:

$$\mathbf{Q} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{q}_1 & \cdots & \mathbf{q}_N \\ \downarrow & & \downarrow \end{bmatrix} \tag{10}$$

These estimated basis functions comply with the PCA orthogonality requirement since $\mathbf{q}_i \cdot \mathbf{q}_j^T = 0, \forall i \neq j$, and they are fine-tuned to our particular data set. The factorization in Equation (9) can be numerically performed by either the Singular Value Decomposition (SVD) or by symmetric matrix eigenvector analysis (Golub and Van Loan 1983).

As mentioned previously, PCA has a special feature making it a good analysis tool for sounds. It is very interesting to note its relationship with the Discrete Cosine Transform. We can draw an analog by observing that PCA and DCT both feature orthogonal bases. So it not unimaginable that some sets of data, when analyzed by PCA, will result in the DCT bases. It turns out that if we were to use the aforementioned method of deriving PCA bases from a set of sounds which exhibit some temporal coherency, we actually get very similar bases with the DCT (Ahmed 1972). The bases in Figure 8 were derived from a recording of speech, note how they are approximately sinusoids ordered in frequency, just like the DCT bases. N was 36.

†. Keen observers will notice that this is a similar scheme to the STFT transform in the discrete domain. Had our linear decomposition been multiplication with the DFT matrix, we would be implementing Equation (5) with no windowing.

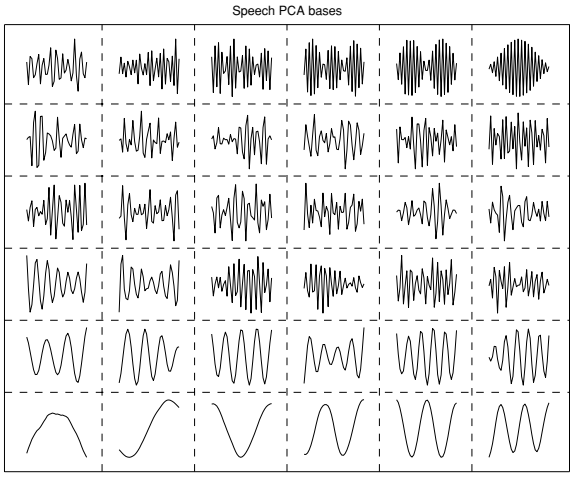

Figure 8 **The set of basis functions derived using PCA on sounds exhibiting harmonicity. The requested basis size was 36 samples. Note that these functions are similar to DCT bases, in the sense that they are frequency localized and sinusoidal in nature.**

This interesting result has been studied in the field of data coding, where PCA is often used. It can be shown that the DCT, is indeed asymptotically equivalent to such PCA analysis of time coherent time series. Rao and Yip (1990) show that the DCT bases can be derived from a Karhunen-Loève transform of a first order Marcov process (a process with a covariance matrix with elements $\mathbf{C}^{(i,j)} = \rho^{|i-j|}$ for any $0 < \rho < 1$). Strang (1999), proves the same for the case where the covariance matrix is the second difference matrix, and Sánchez et al (1995), repeat the same task for all finite order Markov processes. Interestingly enough the same results apply to the DFT (Grenander and Szegö 1958, Gray 1972), although it is less optimal than the DCT.

We did however mention that PCA should be highly dependent on the type of input data, and there are cases where the above constraints do not hold. To illustrate the effect of the analyzed time series data on the estimated bases, we consider two more examples of input sounds. First we assume that $f(t)$ is white noise. That implies that there are no time regularities as in natural sounds, henceforth we would not be able to estimate as coherent a set of bases like before. In fact the individual bases turn out to be white noise (Figure 9).

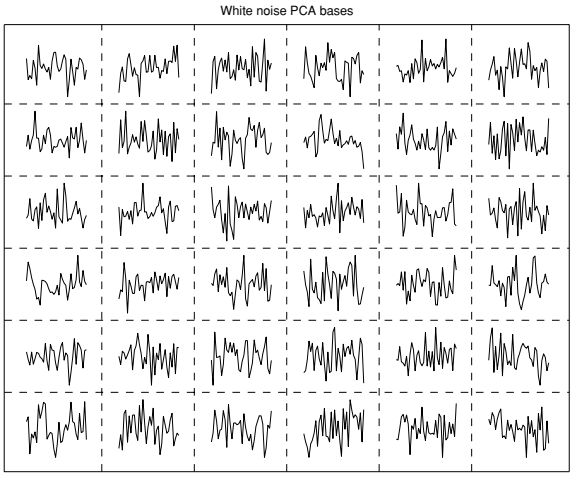

Figure 9 **PCA-derived bases from a white noise signal. The bases themselves are white noise since there was no temporal regularity to be discovered.**

We repeat the same experiment for a signal that was generated by a sparse and random placement of the signal {1,-1} on the time axis. This time, temporal regularities are present (all 1's are followed by a –1), but they are not enough to deduce the harmonic series from them. Our analysis produces the bases in Figure 10.

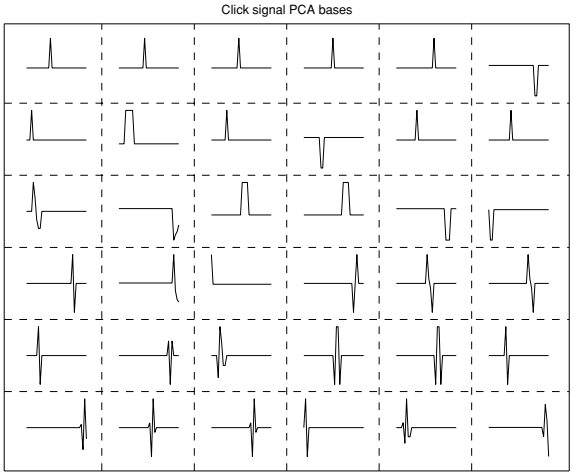

Figure 10 **PCA-derived bases for sparse time-delayed instances of {1,–1}. Note that the bases are influenced by the structure of the input signal in the form of maintaining a high-frequency character.**

Other basis decomposition operations in the linear algebra domain include the Singular Value Decomposition (SVD) and eigenvector decomposition (which are both closely related to PCA), the LU, QR, QZ decompositions, as well as a multitude of more specialized ones. All of these result in bases that have unique features. Depending on our goals, we can pick the most relevant transformation to help us.

As in the case of the Fourier transform we can apply these bases in a sliding window fashion so as to emulate the STFT process. This allows us to track the change of bases through time. Unlike with harmonic transforms there is no time-frequency trade-off, since these bases do not have a frequency by design, neither is there a need for a window in STFT type analysis, since the bases don't have any specific features that result in artifacts.

2 . 2 . 3 Auditory Perception and Basis Decompositions

In the studies of auditory perception, basis decompositions are highly prominent. This fact is caused by the excessive study of the preprocessing parts of the auditory system, in particular the mechanics of an organ known as the cochlea.

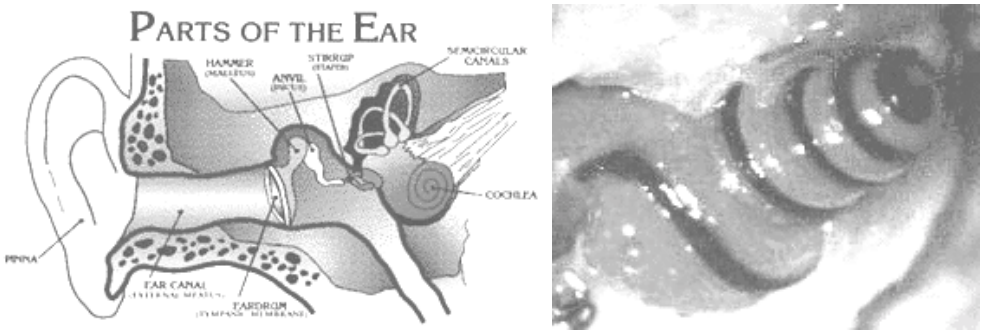

Figure 11 **A schematic of the physiology of the ear (left) and a picture of a cochlea (right).**

The cochlea is the most dominant organ in the physiology of the mammalian ear. As its name implies, it is a snail-like organ and it resides at the inner ear (Figure 11). Uncoiled it measures about 35mm in length. It is mechanically connected to vibrating parts of the ear and it is responsible for the transduction of physical energy to neural impulses. Inside it lies the basilar membrane floating in the cochlear fluids. The basilar membrane extends throughout the length of the cochlea, starting out as being narrow at the beginning and gradually becoming three to four times wider at the other end. As vibrations caused by sounds excite the basilar membrane, it tends to resonate with the higher frequencies near its base at the beginning of the cochlea, while progressively lower frequencies create displacements towards its apex (Figure 12).

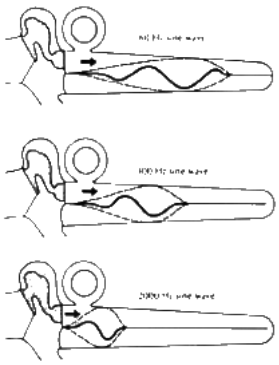

Figure 12 **Resonating behavior of the basilar membrane under different frequencies.**

In effect it distributes the energy of different frequency bands throughout its length. These displacements of the basilar membrane are detected by a series of hair cells inside the cochlea, that upon stimulation release chemical transmitters through a connection with the nervous system and cause neural pulses in proportion to the detected activity.

As the cochlear function suggests, it performs a decomposition very similar to a harmonic analysis. This observation was noted by many researchers who have worked on computational audition models and spawned an entire culture of research dealing with front-end design for audition.

The visual appeal of harmonic analyses and the further justification that our hearing system includes one, have been catalysts for their adoption in audio analysis systems. The STFT has been, and still is, a dominant model for a front-end. It was used early on by Moorer (1975) and Parsons (1976) to create sound separation systems. It is easy to manipulate, efficient and well-understood (albeit limited). A later model, closer to the cochlear function, as well as a better estimator of time-frequency analysis, are constant-Q transforms (harmonic transforms in which the frequency spread versus the time spread are constant throughout the bases). They were used by Petersen (1980) and Brown (1991), as front-ends for audio analysis systems, reporting a better analysis performance as compared to STFT. McAuley and Quartieri (1986) introduced a sinusoidal analysis technique, in which tracks of sinusoidal partials are formed by event formation over peaks in a time-frequency analysis (Figure 13). Serra (1986), extended the sinusoidal model by employing a shaped noise generator to model noisy sound segments, which the McAulay-Quatiery system failed to represent well since their finite number of sinusoids were unable to cover the frequency spread of noise. Ellis and Vercoe (1992), also made use of the McAuley-Quartieri sinusoidal analysis technique in conjunction with a constant-Q transform to mimic the behavior of the auditory system and provide what was coined as a ‘perceptual representation’. Grossman, Kronland-Martinet and Morlet have published a lot of work on audio analysis incorporating wavelets as the front-end (Grossman et al 1990). Using the unique multiresolution properties of wavelets they were able to perform very accurate analyses with superior results as compared

to the STFT. Additional transform methods such as the cepstrum (Oppenheim and Schafer 1989) and LPC (Makhoul 1975) have been employed, their uses however are specifically application based and we will deal with them.

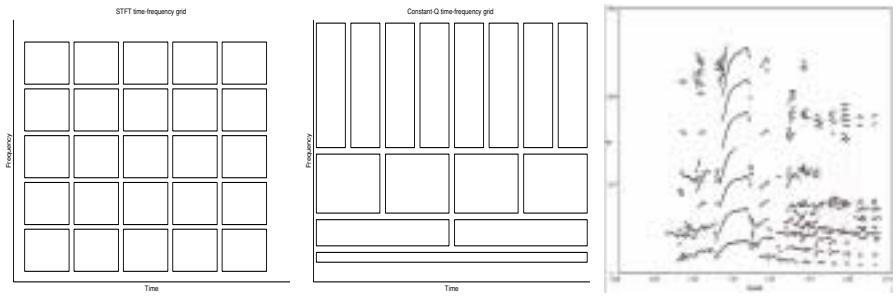

Figure 13 **Illustration of various auditory front-ends. The first two figures illustrate the time-frequency tiling of bases in the STFT (left), and a constant-Q transform or wavelet (center). In the STFT the tiles are the same throughout frequencies; as a result the upper tiles capture more periods of the corresponding harmonic, resulting in an averaged estimate. This problem is solved in a constant-Q tiling since the time spread is reduced for higher frequencies. The right figure illustrates sinusoidal analysis, from which sinusoidal tracks are formed by examination of an underlying time-frequency transform (the original sound is displayed in Figure 7).**

Although the aforementioned decompositions are inspired by the function of the cochlea, they were by no means meant to be biologically accurate. The accurate reconstruction of the cochlear function has been extensively studied and has become a field of study on its own. (Flanagan 1960). Today, by convention, the dominant model employes a gammatone filterbank (Johannesma 1972) to approximate the function of the cochlea. A gammatone filterbank is composed of basis functions which are sinusoidal tones modulated by gamma distributions, that is:

$$g_{f,\alpha,\beta} = \sin(f \cdot t) P_\gamma(t, \alpha, \beta) \tag{11}$$

where f, α and β are arbitrary values and P_γ is defined as:

$$P_\gamma(x, \alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}, x \geq 0 \tag{12}$$

Examples of some instances of gammatones are displayed in Figure 14. The various reasons attributed to its relative success are somewhat ad-hoc (at least as far auditory perception is concerned), and primarily include the ‘pseudo-resonant’ transfer function of gammatones, the simple description of the filterbank, and the efficiency of implementations in either digital or analog form (Lyon 1996).

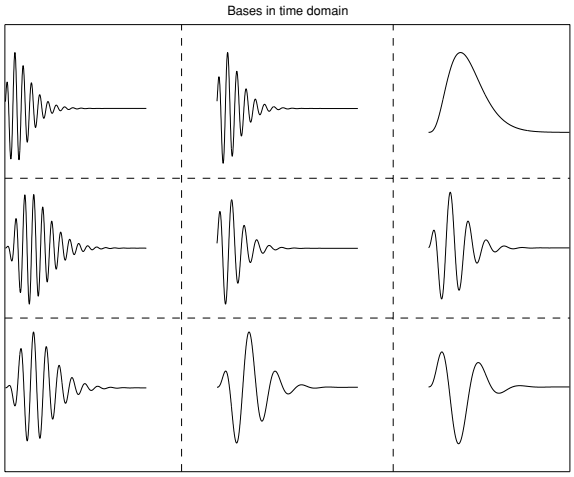

Figure 14 **Gammatone examples for various values of *f*, α and β.**

Various ruminations of the basic model exist, but mostly dealing with implementation issues and efficiency matters.

Researchers in computational audition, drawing inspiration from cochlear modelling research, have employed even more complex front-ends, most notably the correlogram and its derivatives (Slaney et al 1994). Although it is arguable as to whether this is just a preprocessing process or not, many people have used it as a front end for listening functions. The use of correlogram usually consists of a cochlear-like filterbank extended by an extra dimension, which represents the lag time of autocorrelations applied on the energies of every frequency channel. Building on that model, the weft was introduced by Ellis and Rosenthal (1995), as an element for decompositions of primarily harmonic sounds. The weft is extracted from the correlogram data and is inherently linked to common modulation between frequency bands.

2 . 3 Environmental Statistics

Barlow (1989) speculated that the processing of sensory signals from neural mechanisms performs factorial coding. Factorial coding performs a decomposition which results in statistically independent from features, thereby storing maximal information in a non-redundant manner. The fact that our environment contains strong and consistent statistical regularities which our sensory system has to parse, makes it conceivable that our neural mechanisms are fine-tuned to them in order to perform such a coding.

Atick and Redlich (1990), following Barlow's redundancy reduction theory, were able to derive transfer functions for visual encoding which were excellent approximations to experimental data derived from visual cortices of monkeys. They had effectively made the point that the visual system is indeed tuned to the regularities of visual scenes, and it

can be evolved by the use of some information based transformation. The idea that the receptive fields of neurons in our visual system, are tuned to natural image statistics had been around since Barlow (1987), but this was one of the first practical implementations to confirm it. Olshausen and Field (1996), and Bell and Sejnowski (1997), made a similar argument by experimentally deriving filters equivalent to the ones found in the visual cortex, by performing optimization striving for filter sparseness and optimal encoding. This work was followed by Hyvärinen (2000), who extended the framework to deal with shift invariant bases. Further work on this subject is been pursued, and by now it is believed as very probable that the receptive fields of the V1 neurons, one of the early stages of the visual processing system, are indeed tuned to natural images and their statistics.

This work has shed new light on the development of visual front ends through their environment. It presents plausible argument which is grounded on a reasonable theory of perception. However, no matter how compelling these results have been they have not been applied in the auditory domain. The following section performs the auditory analogue to these visual experiments in the auditory domain and results in an equally interesting outcome.

2.4 Measuring Bases from Real Sounds

Although the idea of deriving bases from ensembles of real sounds[†] was touched upon (Bell 1996), it was never examined in depth neither compared to physiological data from our auditory system. We will address this area in this section.

We begin by employing a method similar to that we used for PCA analysis of sounds as introduced in a preceding section. We construct a matrix which features segments $\mathbf{f}_\tau = \left[f(\tau+1) \cdots f(\tau+N)\right]^T$ from the data $f(t)$ to be analyzed:

$$\mathbf{F} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{f}_1 & \cdots & \mathbf{f}_M \\ \downarrow & & \downarrow \end{bmatrix} \tag{13}$$

According to our theory, our goal would be to estimate a set of bases that result in maximally sparse outputs. That is that instead of looking for second order decorrelating transforms as we do in PCA, we look for a maximally statistically independent output. As it was shown in the previous chapter Independent Component Analysis (ICA), provides us with the means to perform such a decomposition.

†. Real sounds is a loose definition describing mostly naturally created sounds. These are sounds that encapsulate some statistical features that specialized synthetic sounds might not have.

We start by whitening (decorrelating) the data using PCA, by which we incidentally obtain the approximate DCT bases $\mathbf{C}_N$. The whitening matrix $\mathbf{W}_P$ is:

$$\mathbf{W}_P = PCA(\mathbf{F}) \approx \mathbf{C}_N \tag{14}$$

and the whitened data, $\mathbf{F}_p$ are obtained from

$$\mathbf{F}_P = \mathbf{W}_P \cdot \mathbf{F} \tag{15}$$

We then employ ICA on the whitened data $\mathbf{F}_P$ from which we obtain $\mathbf{W}_I$

$$\mathbf{W}_I = ICA(\mathbf{F}_P) \tag{16}$$

Our derived bases, will be contained in the product of the two analysis matrices,

$$\mathbf{W} = \mathbf{W}_I \cdot \mathbf{W}_P = \begin{bmatrix} \leftarrow \mathbf{w}_1 \rightarrow \\ \vdots \\ \leftarrow \mathbf{w}_N \rightarrow \end{bmatrix} \tag{17}$$

where the vectors $\mathbf{w}_i$ will be our desired bases.

We first try this method on a set of speech sounds at a sampling rate of 8kHz. The input was segmented into 36 sample blocks (9 ms) and ordered into a 36 by 6000 matrix as in Equation (13). The PCA bases were obtained first, using the Singular Value Decomposition†, and they were used to whiten the data. An ICA step was then applied to the whitened data. We employed an Amari update rule, using the tanh(·) for $\mathbf{f}(\cdot)$, for 800 sample batches and a learning rate of 10^{-4} for 500 epochs. The initial form of the ICA bases before training was Gaussian noise. The experiment was repeated multiple times, always resulting in quantitatively similar results. The derived bases of both steps are displayed in Figure 15.

†. Singular Value Decomposition (SVD) decomposes a matrix $\mathbf{A}$ to the product $\mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$. where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{\Sigma}$ is a diagonal. In the case where $\mathbf{A}$ is symmetric, $\mathbf{V}^T = \mathbf{U}$ and also contains the eigenvectors of $\mathbf{A}$.

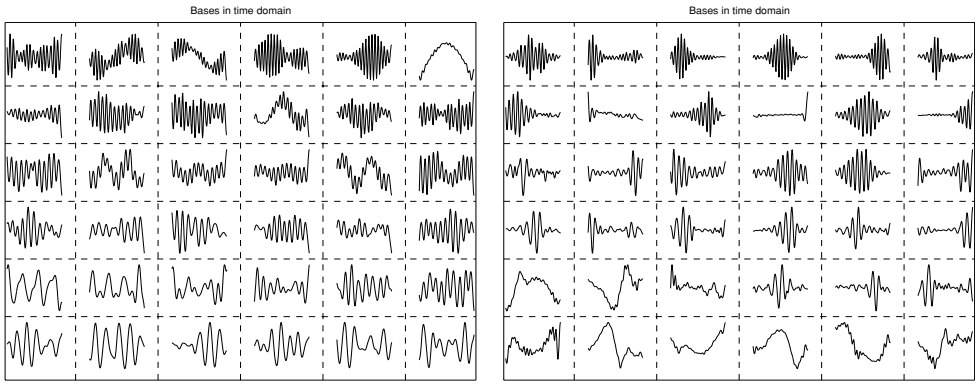

Figure 15 **Derived bases from speech segments. On the left figure are the bases derived from PCA and on the left the bases derived from the subsequent ICA processing.**

As we expected from previous experiments, the PCA bases are approximate sinusoids with variable frequency. The subsequent ICA step though, extends the expressiveness of the bases by introducing an additional variation in the time axis, very much like the STFT extends the DFT. Low frequencies do not have a significant time variance since they do occupy a longer time frame, but, as we examine higher frequencies, we notice an increasing time localization. This is a structure similar to the wavelet and constant-Q transforms, where higher frequency bases are designed to be more localized in time so as to produce an optimal time/frequency analysis according to the Heisenberg constraint (Equation (6)).

Differences between the PCA and ICA bases are more clearly illustrated in Figure 16, where we can see the time and frequency distributions of each set. The PCA bases are clearly clustered in the center of the time axis, extending significantly towards the sides, whereas the ICA bases have a more uniform spread through time, with shorter time spans. The frequency distributions are fairly uniform as expected. The ICA bases have more localized frequency distributions caused by the fact that the bases adapt to the dominant frequencies of the input. In the case of PCA the orthogonality constraint forced the bases to extend evenly throughout the frequency axis. Also displayed are the joint distributions of time and frequency which help us visualize the joint localization features of both analyses. From these joint-distributions we can see once more how PCA has a wide time spread which attempts to cover all the length of the analysis, whereas ICA forms time localized bumps uniformly through time.

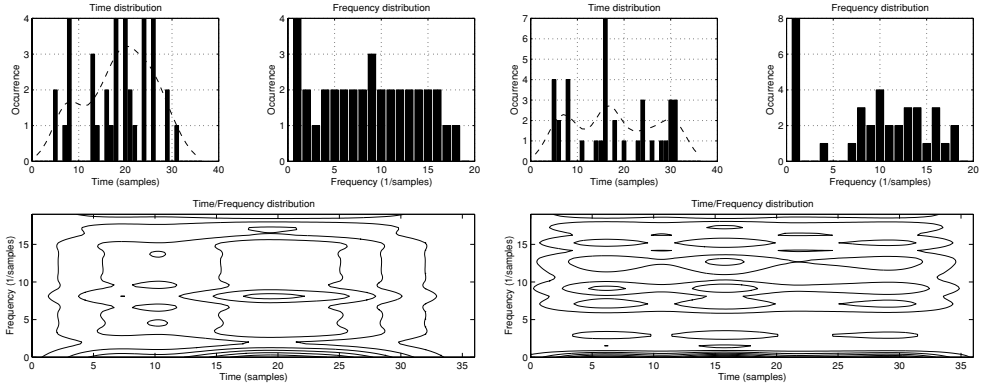

Figure 16 **Time and frequency distributions of the PCA bases (left), and the ICA bases (right). Note how the PCA bases are mostly wider and centered in time. The ICA bases have a more uniform time distribution, forming localized time-frequency bumps.**

The difference in the results is attributed to the fact that the more stringent constraint of statistical independence that ICA imposes (Figure 17), provides for a more compact set of bases that optimally decompose the input sound. Given the nature of speech, which includes rapid frequency changes and busy temporal activity, the optimal set of bases to capture this structure are time localized sinusoids (constant amplitude sinusoids are not expressive enough to efficiently capture such features). To illustrate this we can consider the analysis of a sinusoid with constantly decreasing frequency and a silent part in the middle of its progression. To model this with the PCA/DCT set of bases, we would have to use almost all of them so that with proper cancellations we would create the sliding effect and the discontinuities from the silence. This would imply a coding which would be fairly redundant since it activates many bases at once. The ICA bases, can handle this more elegantly (very much like the STFT), by using the advantage of time localization, to analyze this scene with much less simultaneous base activation.

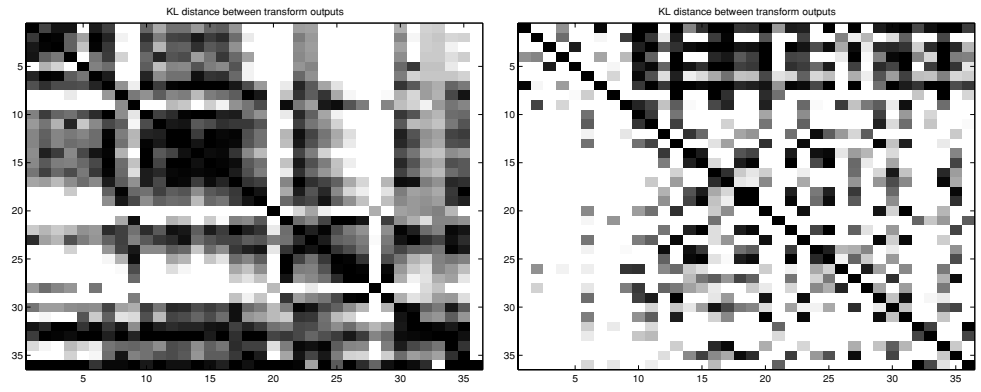

Figure 17 **Kullback-Leibler distances between the outputs of the PCA and ICA transforms. The left figure depicts the PCA transform, the right figure the ICA transform. Lighter color implies lesser statistical dependence. Note how the ICA outputs are in general more independent as compared to the PCA outputs.**

Another interesting thing to note about the ICA bases is the fact that, as bases, they are significantly more sparse than their PCA counterparts. One way to note this difference is to examine their probability distributions (Figure 18). From the kurtosis (a measure of peaky-ness of the distribution and a rough approximation of sparsity) of their distributions we can validate this.

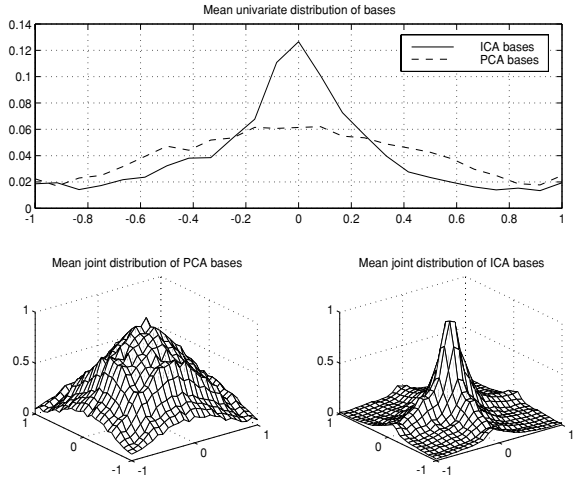

Figure 18 **Comparison of mean univariate and pairwise joint distribution of the PCA and ICA bases. The top figure displays the univariate distributions, PCA depicted with the dashed line, and ICA with the solid.**

As we can see both joint and univariate distributions of ICA are considerably more sparse than their PCA counterparts. This peakyness of the ICA distributions is attributed to the high occurrence of near-zero values in the bases. This is an interesting fact, since it implies that not only the transformation that ICA produces is statistically optimal, but also that the transform itself, is actually quite compact and simple.

This experiment was repeated with a bigger base size ($N = 64$), which again extended about 9ms. The same convergence parameters were used only this time the PCA step was omitted, so that the ICA bases had to converge from the initial Gaussian noise state. Even without the preprocessing, the results through multiple runs have been consistent and still produced time-frequency localized bases. A representative outcome is shown in Figure 19.

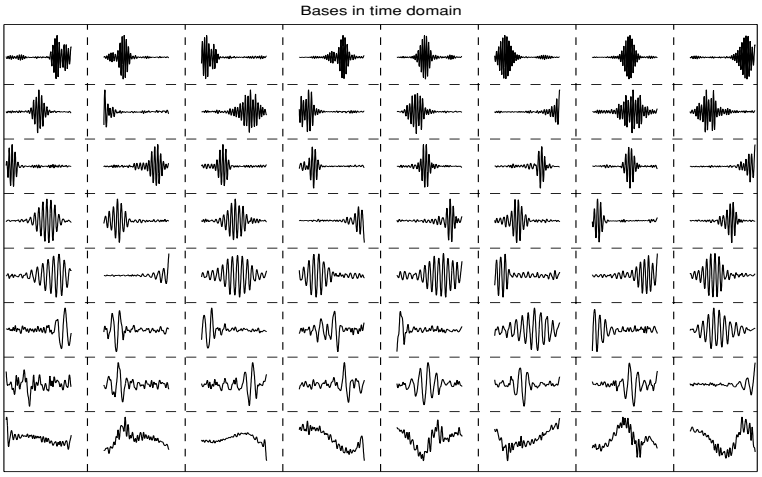

Figure 19 **ICA bases for $N = 64$**

Similar experiments with varying base sizes have yielded similar results. Different types of sources had an effect on the strength of the time-localization of the ICA bases. Musical signals did tend to have worse localization due to their constant frequency characteristic, and sources of natural scenes (rain, rustling leaves, etc.) did tend to have better time localization due to the density of their events. In general, however, time localization was always present and much more pronounced from its PCA counterpart.

2.5 Conclusions

At this point we can make the connection to the psychoacoustical research. As mentioned before, gammatone filterbanks have been measured to be accurate models for the frequency response of the cochlea. According to an ecological evolution point of view such a design would have been an adaptation to the statistics of natural sounds. As our experiment proves, we can obtain very similar bases by using a sensory coding scheme which conforms with leading theories on perception. To further strengthen our point, the same principle has been used to derive the corresponding visual bases, with equivalent results. The significance of this experiment is to validate factorial coding theories of perception as they apply to audition. This was a first application to these theories on a domain other than the visual. We can therefore make a strong argument on a common manner that our neural mechanisms treat sensory signals and open the way for experimentation in the additional domains (Olfaction is a sense that is also heavily modelled with basis decompositions (Hopfield 1991)).

Other than the psychophysical conclusions we can make from this experiment, we can also use this knowledge for designing sensory preprocessors for alternative domains. It is conceivable that this method can be used to derive optimal filters for parsing text streams, electromagnetic sensor fields, messages routed between network nodes, data

traffic inside a computer, etc. Employing this technique, we have a tool to construct perceptual front-ends for arbitrary data stream domains, as long as their exhibit some consistent statistical coherence.

Chapter 3. Perceptual Grouping

3 . 1 Introduction

Perceptual grouping holds a rather prominent spot in the world of auditory research. Due to the attention it has received in the psychoacoustics literature and its speculated importance on discriminating between sounds, it came to be a key operations in implementations of many auditory scene analysis systems. In this chapter we will present the principles of auditory grouping and reformulate them so as to fit our information theoretic viewpoint. The resulting formulation proves to be more compact and computationally efficient than previous approaches. It provides a good interface to the Barlowian theories and exhibits various desirable properties, such as invariance to data representation, bypassing of parameter estimation and good scaling.

3 . 2 Perceptual Grouping

3 . 2 . 1 The Gestalt School

In 1912 Max Wertheimer and his two colleagues Wolfgang Köhler and Kurt Koffka published the paper that founded the gestalt school of psychology. Although the early work of the gestalt school was concerned with visual perception, the field quickly grew to encompass perception in general, and in time extended to other branches of psychology such as learning, thinking, social psychology, personality etc.

The goal of the gestalt school was to encompass the qualities of form and meaning that were traditionally ignored in psychology. Its very name, gestalt, comes from the German word that translates to “placed”, or “put together”. This expressed one of the central issues explored, the perceptual integration of parts to form a whole. An early example of the gestalt work was an investigation of the perception of movement from series of rapidly interchanged stills. In time similar phenomena, where a percept was formed from smaller and seemingly irrelevant observations were also studied. The organizational principles which were observed, were formulated as the law of Prägnanz, which states that the formation of a whole from its parts is dependent on a good configuration of these that has such properties as simplicity, stability, regularity, symmetry, continuity, unity and etc.

Famous examples of gestalt perception are some optical illusions, where prägnanz principles form a percept that isn’t there (Figure 1).

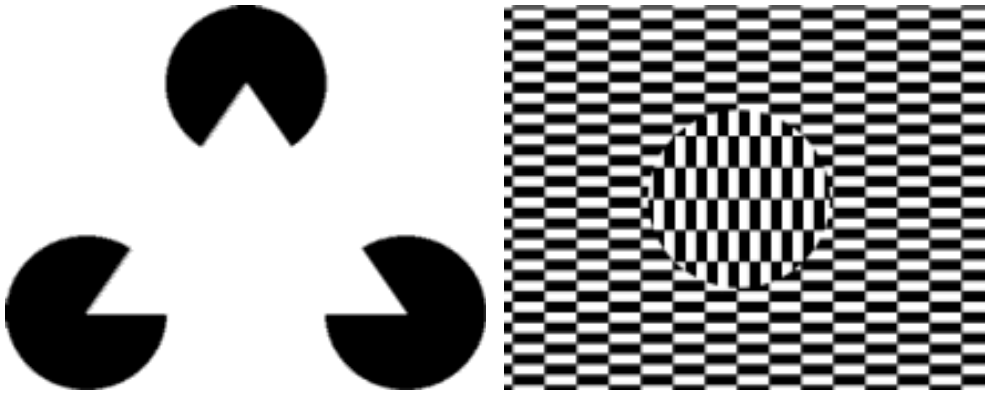

Figure 1 **Gestalt at work. The left figure creates a percept of a white triangle covering three black circles. Although there is not an explicit triangle drawn, we infer its existence from the configuration of the black circles. Likewise the ouchi illusion, on the right, makes us form a percept of a circle hovering over a plane, even though the drawing just reorients some of rectangles.**

One of the Gestalt school concepts with major impact in the perceptual computing world was the notion of perceptual grouping. Perceptual grouping was defined as the process of finding out which components of an analyzed scene should be grouped

together in such a way so as to form an individual object. This notion that individual objects are formed by a proper combination of simpler components had a profound impact the field of machine listening.

3.2.2 Auditory Grouping

The effect of gestalt psychology in auditory perception was epitomized by Bregman (1992). His work collectively presented a set of experiments that highlighted different cases of auditory grouping. It was speculated that these rules were crucial to the perception of individual auditory objects. This work inspired many computational implementations which worked by integrating component decompositions with the appropriate grouping systems. That way it was possible to decompose a scene to basic elements that when grouped correctly would reconstruct individual sounds. Such work was pursued by many researchers, most notably by Parsons (1976), Weintraub (1986), Cooke (1991), Mellinger (1991), Ellis and Vercoe (1992), Brown (1992) and Ellis (1994). In some form or another these systems performed a decomposition of the input sounds to auditory atoms, that were subsequently grouped together with various algorithms. The preferred component representation was that of a sinusoid, or in some cases some more abstract frequency localized objects. The most important cues utilized for grouping these elements were the following.

a. Common Frequency/Amplitude Modulation

These are two of the major and most important cues for auditory grouping. They are also referred to as common fate cues. They are basically cues that originate due to common frequency and/or amplitude modulation between two auditory components. Such component groups that exhibit these commonalities are generally perceived as one object, making it hard for the human listener to focus on the individual components. An example is pointed out by the arrows in Figure 2.

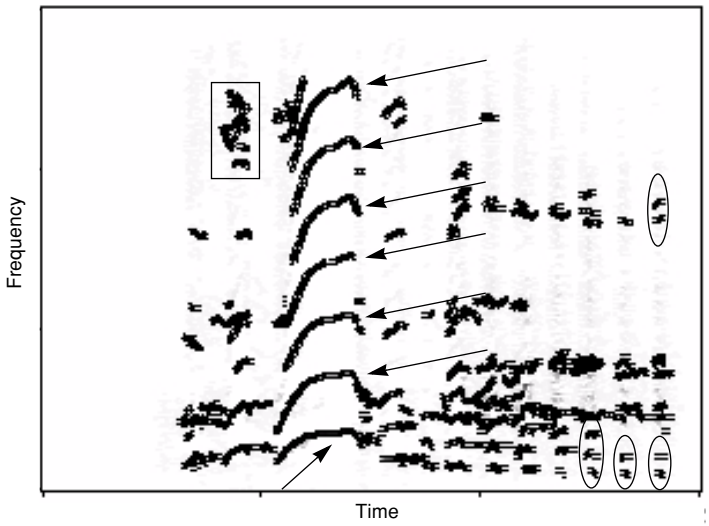

Figure 2

Grouping examples. The sound described in this figure contains a set of sinusoidal components derived from a speech segment. The arrows point out to components that should be grouped due to harmonicity and common frequency modulation. The circles point out subgroups with common onsets and offsets. The squared cluster points out a group due to time and harmonic proximity.

Our behavior under these cues is attributed to the fact that physical sound producing systems, under manipulation that would change their frequency or amplitude characteristics, will imprint that modulation in all of their auditory components. For example, changing pitch as we speak results in a common shift on all modal oscillations of our vocal tract. Likewise, whenever we change the level of our voice, we again amplify or suppress all frequency components in a common manner. Given our auditory development in a world where this auditory consistency exists, it is only natural to create a perceptual mapping of such common behavior to the indication that the modulated components originate from the same source, hence they constitute an individual object. This mapping is known as the common frequency and common amplitude grouping cue.

In order to computationally take advantage of these cues, it was deemed necessary to estimate frequency and amplitude. Estimation of common amplitude modulation is a simple matter to measure. In the case of frequency modulation though we have issues of resolution and estimation artifacts which can bias the estimate and conceal the true value we wish to obtain. This is a well recognized problem which has been addressed by a continuous revision of front ends to more accurate estimation methods. However the fact that the parameter estimation has to be done remains, as does the possibility of errors.

b. Common Onsets

Common onset is another key cue in computational implementations. Although it is considered a cue of its own, it can be seen as a special case of common amplitude modulation, where the modulating function is a rectangle window. This cue can be accredited to the fact that most physical systems when excited produce complex sounds whose components have the same event timing. This regularity in our auditory environment has shaped an interpretation of common onsets between components, as an indication that these components belong to the same sound. Although not as strong a cue, sometimes the common offset is also taken into account. It is not however as reliable a cue as the onset. Figure 2 also presents some cases of common onset groups (the groups in circles). Estimation of these features is easy to make and is generally not an issue.

c. Harmonic Relationship

This is a cue that is unique to audio, and rather intriguing. Although the preceding cues can be adapted to other time based senses (e.g. vision), this is a cue that is only so strong in audio, and it is probably one of the most important ones.

Components that are harmonics of a common fundamental frequency tend to fuse together. In fact most pitched sounds we perceive are summations of fairly harmonic modal oscillations of physical systems. We are usually unaware of this structure, due to strong fusing. Due to the high occurrence of this regularity, our ears have adapted to detecting harmonic series as one object and we have very little ability to detect individual harmonic components. Estimation of harmonicity is a complex operation, which is often achieved using frequency estimates. As also mentioned before, such estimates cannot always be reliable and sometimes derail the grouping process.

Another issue in harmonicity which is related to grouping is that of harmonic proximity. It has been observed that harmonic pairs of components that are closer in frequency are more likely to fuse. For example a fundamental and its first harmonic will fuse much stronger than the fundamental with its seventh harmonic. When given harmonic components with frequency gaps between them, the frequency distance between the components becomes a factor on whether they will be grouped or not. This can be again attributed to the environment since the occurrence of harmonic gaps is very rare. Most pitched sounds will never miss more than a few consecutive harmonics. Estimation of this cue is once more dependent on frequency detection.

d. Higher Level Cues

Various other cues have been put to use, many of them basic and fundamental in nature. An example is spatial cues, which arise from timing commonalities between sounds as they are captured across our two ears. We should also consider the case where our components are more complex sounds rather than sinusoids. In this case common modulation assumes new meanings that are dependent on the

characteristics of the components. Such grouping criteria can include common spectral shaping, or some commonly modulated statistical property.

e. Streaming Cues

There is an additional class of grouping cues which we will not deal with in this chapter. These are cues that relate to the formation of streams. Although these cues have been used extensively, they are not as low-level as the ones presented so far. The reason why they are regarded as important, in both the psychological and computational literature, has to do with the interpretation of an individual sound, and sometimes with the chosen representation†. The goal of many research projects is to extract entire auditory 'sentences' of a sound, not the individual instances. For example, in a scene with a piano melody, the effort is usually to extract the entire piano performance as one sound, not the individual piano notes. Such streaming cues have to do with event structure and are sometimes a cognitive process rather than a perceptual one. Examples of these are closure, continuity, pattern repetition, etc. This is an issue which we are not interested in at the moment and we'll cover in a later chapter. In this chapter we will be dealing with grouping criteria that deal mostly with grouping of simultaneous components. We will however refer the reader to the discussion at the end of the chapter on possible integrations of these cues in this chapter's framework.

Most computational implementations had to deal with these cues. Parsons (1976) dealt with Fourier components and performed grouping based on harmonicity, Weintraub (1986) employed a cochlear filterbank and a similar grouping approach. Mellinger (1992), bootstrapped his model for musical scenes, used mostly common onsets and common frequency modulation. Brown and Cooke (1994), used a filterbank, and most of the aforementioned grouping cues. Using similar cues Ellis (1992, 1994), employed a constant-Q transform, and later (1996*),* an alternative representation called the weft.

However most of these implementation had to deal with some problems of semantic nature. A major problem in computationally implementing all of the above rules is their complex description. Although these principles are well understood and regarded as simple in a heuristic sense, they are hard to translate to mathematics and input to a computer implementation. For example consider the case of frequency modulation, although it is easy to describe semantically and even visually detect it from a time-frequency decomposition, it is a very complicated cue to detect. It involves, precise frequency tracking, and a smart correlation algorithm. In addition to this, considering that many time-frequency analyses produce a lot of artifacts in analysis it is easy to see that this problem becomes very hard to deal with.

To further muddle matters, all said in the auditory grouping is quite irrelevant to other perceptual domains. Cues such as harmonicity and common frequency modulations are

†. Upon examination of Figure 2, we see a lot of small seemingly noisy components in temporal and frequency proximity. These are mainly artifacts of the analysis, quite unrelated to the true components of the original sound (laughing speech). Yet they have to be grouped together for a proper resynthesis. Certain implementations were obliged to use proximity grouping criteria as lower level grouping. This was however something which was mandated by the analysis procedures used and not as much by perceptual studies.

not easily mappable to the visual domain. For the researcher of perception in general, this is a very perplexing matter, since it forms an obvious barrier in finding unifying principles of perception. In the auditory domain the selection of the cues is highly dependent by the sinusoidal component representation (harmonicity and frequency modulation, two of the strongest cues, are sometimes hard to define for other component types). This particular component selection, is an obvious influence by the observed cochlear filtering and the visual appeal that such representations have (sadly a lot of researchers strive to find a decomposition in which the auditory grouping problem translates to an image grouping problem). Although this is not something that necessarily effects the performance of systems, it is an upsetting issue when in comes to integrating such systems in a more general perceptual theory.

Recent work on blind source separation did come close to the concept of grouping since it is based on matching across multisensor recordings, however it was never conceptually connected with perceptual grouping. It was however free of the aforementioned problems and gives a good example to imitate. Our approach will build on this observation.

3.3 Grouping as redundancy reduction

3.3.1 The Theory

As we do throughout this thesis, we will adopt the Barlowian approach to perception (1959). We will try to formulate the auditory grouping process in terms of minimizing statistical dependencies of the inputs and coming up with a compact and sparse code. Although the following is an obvious approach, it was ignored by many computational auditory researchers. Most grouping principles, even more evidently for audio, are interpretations of statistical dependencies. It is easy to see that common modulations between two signals will introduce some form of statistical dependency between them. This stems from the fact that components that are generated by the same physical system are by design dependent, and we have adapted to using the various consistent relations between them to perceive them as one sound. By using this statistical layer of abstraction in a computational implementation, we can bypass psychological literature interpretations and deal with the realities embedded in the data. This forms a much more compact description of grouping rules and creates a straightforward computational basis. It also provides an plausible interface to the role of the environment in the development of our auditory behavior.

Throughout this chapter we will assume that successful grouping is a result of minimum mutual information partitioning (a form of sparse coding). We will show that perceptual fusion is stronger whenever two components exhibit high mutual information. By partitioning a set of these components so that each group has maximum mutual information, we will be reducing the overall mutual information of the groups, and making a sparse partition.

3.3.2 The Method

In the following sections we will iterate the same procedure for different grouping scenaria. We will pick a simple two or three object auditory scene which we will parameterize over a variable n. For a specific value of that variable we would get a scene that would exhibit perceived auditory fusion between the components. If our suspicions are correct, this value of n will also give a maximum at their mutual information. In effect, mutual information will be an indication of strength of fusing.

One way to easily evaluate mutual information is numerically. Information theoretic measurements are not easy to make, and they are often very approximate, so at first we will only use these results as evidence, and we will try to explain how they arise. In a later section these results are considered in the design of a grouping mechanism, which does not rely on discrete approximations. The numerical operations we will perform are based on the discretization of the mutual information definition. Recall the definition of mutual information:

$$I(x_1, \ldots x_N) = \sum_{i=1}^{N} H(x_i) - H(x_1, \ldots x_N) \tag{1}$$

where the $H(x_1,...,x_N)$ is the joint entropy of all the x_i and $H(x_i)$ the marginal entropy of each x_i. We will estimate these entropies from their definitions:

$$H(x_1, \ldots x_N) = -\sum_{\forall x_1, \ldots x_N} P(x_1, \ldots x_N) \log P(x_1, \ldots x_N) \tag{2}$$

and

$$H(x_i) = -\sum_{\forall x_i} P(x_i) \log P(x_i) \tag{3}$$

where $P(\cdot)$ is the probability density function of a variable. We will estimate $P(\cdot)$ from a histogram $\hat{P}(\cdot)$, which is normalized so that $\sum \hat{P}(\cdot) = 1$. That makes our computation:

$$\hat{I}(x, y) = \sum_{\forall x, y} \hat{P}(x, y) \log \hat{P}(x, y) - \sum_{\forall x} \hat{P}(x) \log \hat{P}(x) - \sum_{\forall y} \hat{P}(y) \log \hat{P}(y) \tag{4}$$

Although for probability density estimation, it is a popular practice to use more sophisticated methods than histograms, in this particular case it not a recommended approach.

The objects we will be dealing with are sometimes sinusoidal and they feature peculiar density functions which have sharp peaks and abrupt ends at their extrema. Using a kernel estimation method such as a Partzen window, or Gaussian mixtures, would be altering the density functions dramatically and will remove some very important features. We employed a straightforward algorithm for histogram estimation and took special care in using histograms which were not plagued by gaps and gross inaccuracies.

3 . 4 Harmonicity - Frequency Proximity

Harmonicity is one of the major gestalt principles in auditory grouping. Auditory grouping theory states that two sinusoids, where one of which has a frequency which is an integer multiple of the other's (harmonic relation), are grouped together.

Harmonic relations exhibit more mutual information than non-harmonic relations. We will first attempt to present an intuitive justification for this before we proceed to the numerical tests. Mutual information has a direct relationship with the amount of possible value sets between our variables. Consider the case where our variables are the identical series $x = \{1, -1, 1, -1, ...\}$ and $y = \{1, -1, 1, -1, ...\}$. In this case we only witness two kinds of sets, $\{x,y\} = \{1,1\}$ and $\{x,y\} = \{-1,-1\}$, which will create a joint probability density with only two points at $\{1,1\}$ and $\{-1,-1\}$, both having values of 0.5. This pointy and uniform in values probability density implies a low joint entropy, which assuming a constant set of marginal entropies, translates to a high mutual information (refer to Equation (1)). In this example we do indeed have a lot of mutual information since each of the two sequences is fully described from the other. In the other end we can have the sequences $x = \{1,1,1,1,...\}$, $y = \{1, 2, 3, 4, ...\}$, in which we have an infinite amount of value pairs, hence zero mutual information. That is easy to intuitively see by noting that knowing one of the sequences it is impossible to deduce any information about the other. Back in the harmonicity case, if we have two sinusoids of frequencies 1 and 2 the possible pairs will be:

$$\{\cos(t),\ \cos(f \cdot t)\} \tag{5}$$

Assuming infinitely long sinusoids and given the periodicity of the signals, we only need to examine the pairs in the interval of one *joint period*, i.e. the time span between which the two sinusoids assume the same phase twice. In this case the joint period is from 0 to 2π, which is the interval in which both sinusoids assume the phase pair $\{0,0\}$ twice. The number of pairs of these functions in that interval is roughly all the points in between[†]. Assuming frequencies of 1 and 1.5, the interval which we will examine will be from 0 to 6π. The number of pairs will be three times as many, which will result in a reduction of pair repetition by three, hence a reduced mutual information. Similar reasoning can show that the more irrational the frequency ratio becomes, the less mutual information we will have since the number of possible pairs tends to be larger (for a

†. We will have a repetition of the pair {0,0}, at $t = 0$ and $t = \pi$, so not all samples should be counted, but for now we assume the effect of these solutions to be negligent. Later we will see the effect of this on harmonic proximity.

more rigorous derivation see Smaragdis 2001). We will now use some numerical results to validate these theories.

The experiment below attempts to detect grouping based on harmonicity by measuring mutual information numerically. We construct a parameterized scene from a set of two sinusoids as:

$$s(n) = \begin{cases} \cos(f \cdot t) \\ \cos(n \cdot f \cdot t) \end{cases} \tag{6}$$

where n was swept from 1 to 6, and f was randomly chosen to be 1321. As discussed above, we would expect the mutual information to be peaking when n is an integer, the case when the two sinusoids would be harmonic and have a stronger tendency to group. The results are shown in Figure 3.

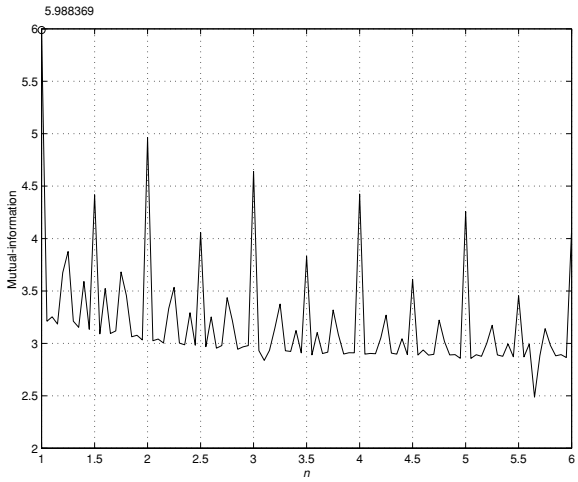

Figure 3 **Mutual information measurements for harmonicity experiment**

As is evident, our prediction about the relationship of mutual information and harmonicity was right. We obviously have the strongest fusion at the case where the two sinusoids have the same frequency, then as the second sinusoid assumes the role of harmonics we get local peaks. However we will also make some additional important observations. First, we appear to have local peaks at points where n = 1.5, 2.5, 3.5, 4.5 .... At these points we have an interesting effect. The two sinusoids act as a first and second harmonic to a non-existing fundamental at $\frac{f}{2}$. This is also a case of harmonicity, albeit with a missing fundamental. We can observe additional diminishing local peaks at progressively more irrational frequency ratios. These peaks correspond to a progressively more distant fundamental at $f \cdot \text{frac}(n)$, where $\text{frac}(n)$ is the fractional part of n.

Although the resolution in Figure 3 is not high enough to display this effect consistently, there will be values that will have zero mutual information when the frequency ratios are fully irrational.

We also note that when $n = 2$ we get a stronger predicted grouping than when $n = 3$, something which we can (subjectively) verify by ear. This leads us to yet another observation to be made; harmonic proximity can also be tracked using mutual information. Note that the more distant the second sinusoid becomes in frequency, the less the mutual information peaks. This is an effect caused by the increasing number of common pairs in Equation (5). The bigger the joint period is, the more repetitions of {0,0} will be present. This peak of the joint probability at {0,0} causes the joint entropy to increase, hence lowering the mutual information (once again see Smaragdis 2001, for a more rigorous treatment).

In order to further verify the proximity point numerically we can insert another sinusoid in Equation (6) and transform it to:

$$s(n) = \begin{cases} \cos(f \cdot t) \\ \cos(4 \cdot f \cdot t) \\ \cos(n \cdot f \cdot t) \end{cases} \tag{7}$$

We would expect this to make the peaks near 4 more prominent, this the second sinusoid assumes this frequency. It should also introduce a maximum at the third harmonic ($n = 4$), since that value along with $n = 1$ produces the most harmonic set. The results are shown in Figure 4.

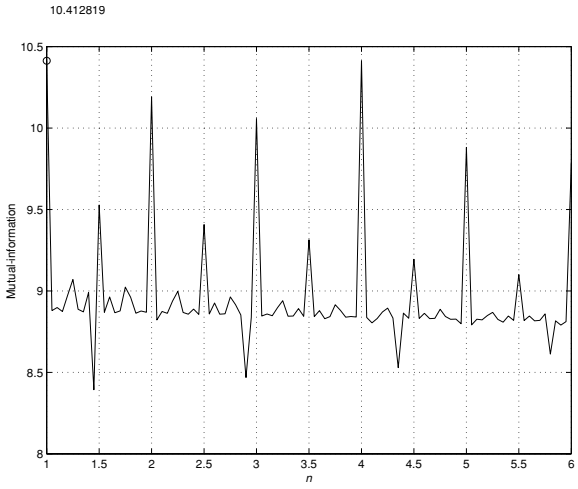

Figure 4 **Mutual information measurements for proximity experiment**

As can be seen, at $n = 4$ we have a maximal peak (sharing the honor with the peak at $n = 1$), and the neighboring peaks at integer values of n are relatively higher than before. This demonstrates the ability to track harmonic proximity.

3.5 Common Fate

3.5.1 Frequency and Amplitude Modulation

Common fate in the auditory domain is usually linked to two parameters, frequency and amplitude, and it is usually described as their common modulation. If a set of sinusoids feature common modulation in either amplitude or frequency (or both), then they are most likely perceived as one sound. Mutual information in this case is created from the statistical dependencies that a common frequency or amplitude track will generate.

According to our theory, we would expect the mutual information measure to peak when we have co-modulated sinusoids. The following two short experiments prove to behave as the gestalt rules predict for common modulation cases.

In the first case the auditory scene $s(n)$ is a parameterized set of two frequency modulated sinusoids. One will be modulated by an arbitrary function a and the other by $a + n \cdot e$, where e will be a noise signal having the standard normal distribution. That is:

$$s(n) = \begin{pmatrix} \cos(f \cdot a \cdot t) \\ \cos(f\sqrt{2} \cdot (a + e) \cdot t) \end{pmatrix} \tag{8}$$

where f was chosen to be 1321. The square root of two is used as an irrational frequency ratio between the sinusoids so as to remove any possible harmonicity which might skew the results. Intuitively what this equation does is maintain a modulation on the two sinusoids which is common only when the $n = 0$. The scene at this value, according to psychoacoustics, will increase the chance of perceptual grouping. If we were to measure the mutual information of $s(n)$, with respect to n, we would expect a peak at that point.

Similarly we set up the same experiment with amplitude modulation rather than frequency modulation, where:

$$s(n) = \begin{pmatrix} a\cos(f \cdot t) \\ (a + e)\cos(f\sqrt{2} \cdot t) \end{pmatrix} \tag{9}$$

Our measurements shown in Figure 5, validate our hypotheses.

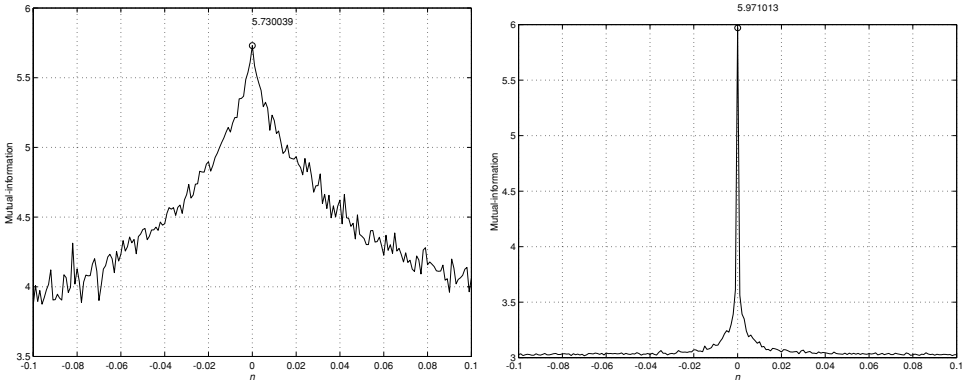

Figure 5 **Averaged mutual information measurements for common amplitude modulation (left), and common frequency modulation (right). The results are the average of 12 trials. For better visualization purposes *n* ranged from -0.1 to 0.1. At the limit however, the plots are symmetric around 0 and the negative values are redundant.**

We can see how the system is much more sensitive when it comes to frequency modulation, as compared to amplitude modulation. This applies in human perception too, where we are more sensitive to frequency discrepancies as opposed to amplitude ones. Similar results arise for sound sources other than sinusoids, with the only difference being the scale of the n parameter and the relative size of the mutual information peak. Simple waveforms such as triangle, sawtooth and square waves display virtually identical results. Using real sounds, such as speech and environmental noises, the plots were similar, but extended throughout a larger scale of the parameter n, so as to average out the modulation already existent in these sounds.

We can refer to the intuitive foundation we formed in the previous section to analyze frequency modulation The two sinusoids will have a fixed number of pairs without the modulation applied. By frequency modulating them with the same function, we will approximately maintain this number of pairs since the harmonic relationship of these sounds will still be the same, although at a local level. When we modulate by different functions, we offset this number to a smaller value due to the instantaneous relative inharmonicities we introduce (although it is possible to envision cases where this doesn't hold, they are not as probable)†.

3.5.2 Common onsets/offsets

Common onset and offset of partials is another example of common fate and also a key clue in auditory grouping. If two sounds coincide in time, then they exhibit a correlation which results in additional mutual information. Just like in the previous section, this is maximized when the on and off boundaries are exactly the same for both sounds.

†. It should be noted that grouping of sinusoids with a constant frequency slopes is a special case of common frequency modulation. When the slopes of the sinusoids are parallel, we perceive them as a group, and this is obviously a case of common frequency modulation where the function α in Equation (8) is a line, and the noise injection e is altering its slope.

The following scene is set up:

$$s(n) = \begin{cases} f(t) \\ f(\sqrt{2} \cdot t + n) \end{cases} \tag{10}$$

where:

$$f(t) = \begin{cases} t_1 \le t \le t_2,\ f(t) = \cos(t) \\ \text{otherwise}, f(t) = 0 \end{cases} \tag{11}$$

The function f is essentially a time-bounded sinusoid from t_1 to t_2. By varying the value of n we produce two sinusoids that are misaligned in time for $x \neq 0$, and perfectly aligned otherwise. We would expect to see a peak in the mutual information of $s(n)$ at the point where the two sinusoids align. The results are shown in Figure 6. Similar results are obtained when we change only the onset or the offset of the second sinusoid rather than both. The same results can be obtained by using real sounds instead of sinusoids. The mutual information peak is formed by the fact that the more common silence the two sounds will have the more prominent the pair {0,0} will be, forcing the joint entropy to minimize, hence the mutual information to maximize.

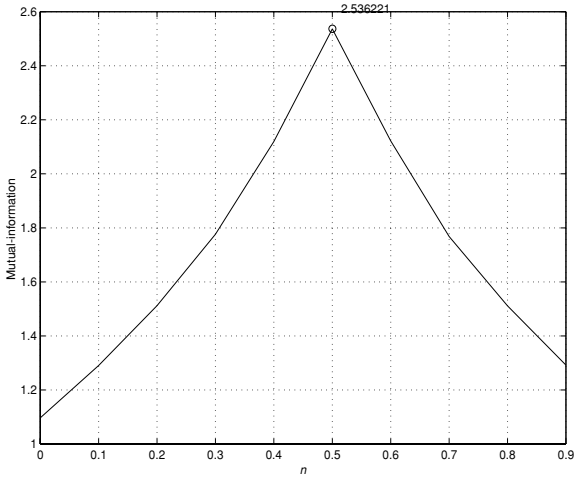

Figure 6 **Mutual information measurements for common time onset/offset**

It should be noted that the common onset/offset can also be seen as a special case of common amplitude modulation, where the modulating functions are instances of time localized rectangle windows or gate functions.

3 . 5 . 3 Time Proximity

Time proximity is also a factor in grouping, especially in cases of large concentrations of short-timed components. We set the following experiment to deal with it. Our scene is defined as:

$$s(n) = \begin{cases} g_1(t, n)\cos(f_1 \cdot t) \\ g_2(t, n)\cos(f_2 \cdot t) \\ g_3(t, n)\cos(f_3 \cdot t) \end{cases} \tag{12}$$

where f_i were random frequencies and g_i defined as:

$$g_i(t, n) = \begin{cases} 1, \frac{C_{1,i}}{i} \le n \le C_{2,i} \\ 0, \quad \text{otherwise} \end{cases} \tag{13}$$

where $C_{1,i}$ and $C_{2,i}$ are time constants, which are different for each component. What this scene effectively performs is, place three time localized sinusoids in temporal distances whose proximity is dependent on n. As n tends towards 0, the three sinusoids will increasingly overlap until they coincide in time (for $n = 0$). For larger values of n, the three sinusoids will not overlap as much if at all. The results are shown in Figure 7.

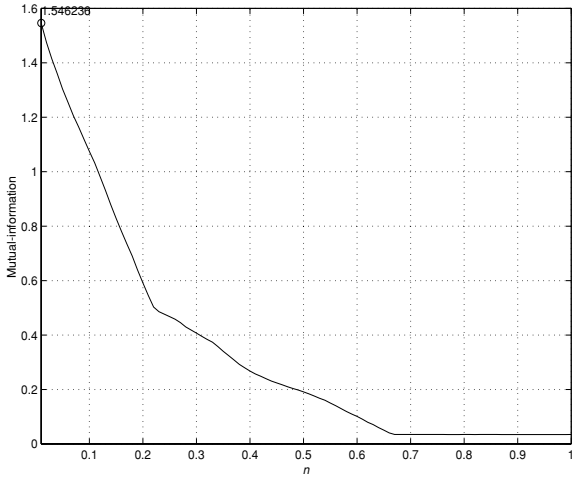

Figure 7 **Time proximity experiment. As the sinusoids move closer in time they increase their mutual information.**

Similar results were obtained when instead if sinusoids, real sounds were used.

The behavior of this experiment is explained by the same observation used for the common onset. The gradual peaking in the mutual information is caused by the increasing cases of coinciding silence across the three sounds (the triplet {0,0,0}). In other words the mutual information is in the silence and not the contents of the sinusoids. This interpretation fits in very well with the higher level status of this cue, since that form of grouping is quite ignorant of the contents of the components and more involved with their time positioning.

3.6 Prägnanz and Higher-order principles

Due to the fact that gestalt principles were formulated to be general and applicable to all perceptual functions, there are a lot of special statistical dependencies that are not described. As an effort to include all of these unknown and vague factors, the prägnanz principle was used. The prägnanz principle states that "of several geometrically possible organizations that one will actually occur which possesses the best, simplest and most stable shape." (Koffka 1935). It is arguable that descriptions such as simple, best and stable, tend to denote strong statistical dependencies. A square for example which would fall under the best/simple/stable description exhibits a lot of structure and redundancy, when compared to the complicated and unstable random polygon. This is of course a direct statement of this thesis' argument. Simplicity, stability, predictability and order are features of high mutual information systems. Instead of using them by name and having to resort to their estimation, we can indirectly solve the problem by measuring their side effects from statistics.

3.7 Putting it all together

The main observation that we derived from the experiments above, is that perceptual grouping can be judged by measuring the redundancy of a scene. We've noticed that whenever two elements should be grouped, they exhibit common information which in turn generates a lot or redundancy. By measuring the amount or redundancy we can make a decision on grouping. In the case of many elements and multiple groups, based on our experiments we can theorize that by manipulating the scene so that we reduce redundancy we would infer the proper groups. One way to do so and maintain an interface to the other chapters of this thesis, is to employ ICA. ICA is designed to find a linear combination of its inputs that will minimize the redundancy between its outputs. In effect the linear transformation we obtain from it will be some kind of an adjacency matrix which will attempt to group strongly related inputs. An example of how we could use this algorithm to perform auditory grouping is presented in the following experiment.

We assume that we have a set of sinusoids which comprise a simple auditory scene. We would like to partition them in such a way so that we identify the different groups in the scene. Ten sinusoids were ordered as rows on a 10 row matrix. Out of these rows, the

sinusoids at rows 1, 8 and 9 were related, as were the sinusoids at rows 2, 6 and 7 and at rows 3, 4, 5 and 10. In effect we had three groups, being the rows {1,8,9}, {2,6,7} and {3,4,5,10}. The relations between the sinusoids in each group, were harmonic ratios and common frequency and amplitude modulation. Across different groups the sinusoids had no such common information. Figure 8 displays the spectrograms of each of the three groups, as well as the entire scene.

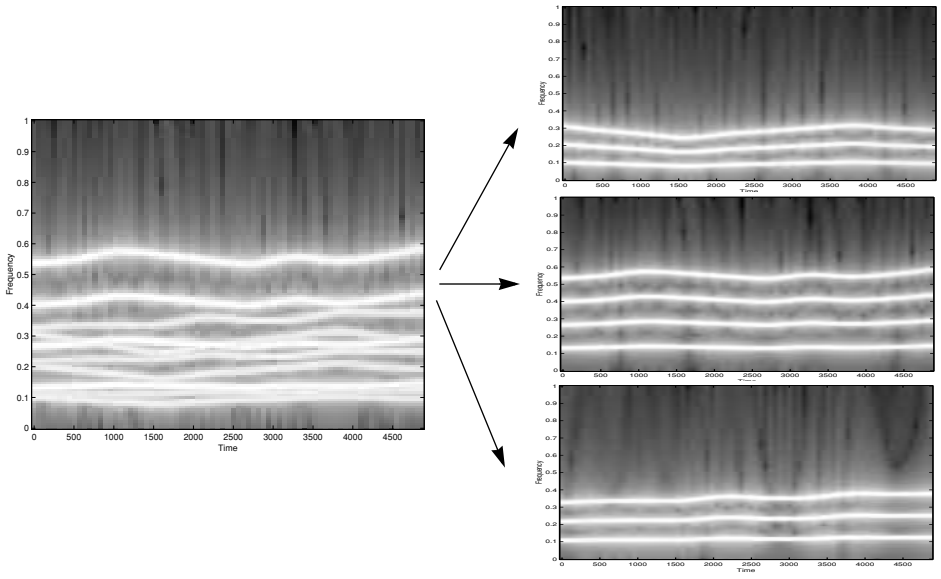

Figure 8 **A spectrogram of all the sinusoids combined (left) and the three groups individually (right). Each group exhibited some gestalt grouping criterion between its members.**

The matrix that contained the ten sinusoids was submitted, without any partitioning hints, to an ICA algorithm. As we described before, we would expect the ICA algorithm to return something like an adjacency matrix which would point relationships between elements, based on their mutual information. Its values with larger magnitude, will indicate a relationship between the two elements they link, whereas smaller in magnitude values will indicate the opposite (this is a function similar to the covariance matrix, it is however derived from information measures as measured by ICA rather than just second order statistics that we use for covariance). The algorithm we used was the same as in the previous chapter, Amari's rule with a hyperbolic tangent nonlinearity. We trained until convergence. The magnitudes of the matrix that we obtained are shown in Figure 9.

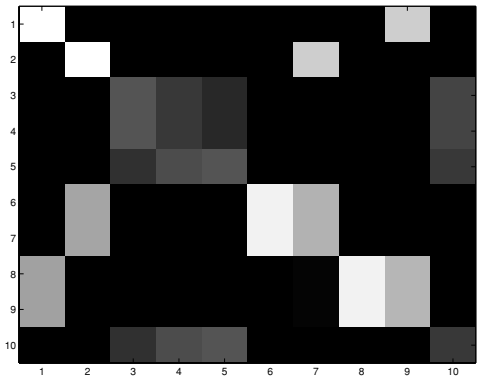

Figure 9 **Magnitudes of the weight matrix returned from ICA analysis of auditory components. This can be perceived as a higher order covariance matrix, where values greater in magnitude (lighter color) will be denoting a significant statistical dependence between the elements they link.**

As it is clearly evident, the values that have a some magnitude reveal the correlations of the scene. To make our results more readable we can also force this weight matrix to be symmetric (thereby making all measured correlations be bidirectional) and rounding the values to either 0 or not 0 (Figure 10). For the case of the first group, which contains the sinusoids 1, 8, and 9, we can discover its structure as reflected by our result, by observing that the values that correspond to these elements, {8,1}, {9,1}, {1,9}, {1,8}, {8,9} and {9,8} are non zero. Likewise we can infer the rest of the groups.

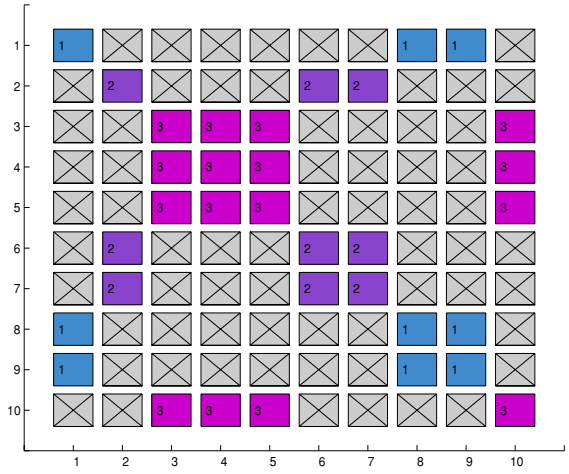

Figure 10 **Resolved weights. By simply rounding the elements of the ICA output (Figure 9), to 0 and not 0, and forcing it to be symmetric, we can reveal the exact groups in the given scene. The answer is given in the form of an adjacency matrix. In this case it easier to see the three groups which are formed.**

This approach yields results even in the presence of approximately common modulation between the sinusoids of the same group. In general about 5% difference between modulation functions which are supposedly common, would be the tolerance limit for successful grouping. Identical results arise by using different types of elementary waveforms instead of sinusoids, such as triangles, sawteeth, etc. (refer to appendix A, for a multimodal example that uses non-uniform elements across the visual and auditory domains).

What is unique to this approach from past work is that, it is exhibits more generalization potential by abolishing the notion of a parameter. In terms of representation, there is the problem that any parameter based approaches will be irrelevant to other forms of objects. By operating on pure data, without parameter definitions such as frequency and amplitude, we can abstract the algorithm and apply it on arbitrary representations. In terms of estimation, there were no pitch track, amplitude or on/offset estimates. Considerable research has been expended on just the parameter tracking of the objects to group, which complicates matters. Not only is it sometimes hard to perform such estimations, but there is the increased risk that the estimation algorithm peculiarities will skew the results. Instead of focusing on the parameters, we bypass them by examining their statistical side-effects.

An additional point to make is that there was no indication of how many groups to extract from the scene. The number of groups was inferred from the data and their cross-statistics. It is encoded as the rank of the adjacency matrix that we deduce from the analysis.

Finally, perhaps the most practical feature of this approach, is that the required computational load is comparatively small. ICA algorithms execute operations on the scale of $O\{N^3\}$, spending most of their time on matrix multiplications equal in order to the number of components. Alternative approaches based on parameter extraction can utilize some kind of exhaustive pairing algorithm that would examine common traits between all possible groups. If such a complete search was to be chosen, the complexity could rise up to $O\{B_N\}$ where B_N is the bell number (the number of ways a set of N elements can be partitioned to non-empty subsets). This number grows with an extremely high rate†, making grouping of more than a handful components a daunting task. Even disregarding the parameter estimation steps, which can be also expensive, for cases where we are presented with more than 20 sinusoids the problem becomes almost intractable. Luckily, most implementations do not perform a complete search, but they still have to deal with more complex operations than the ICA model.

3.8 Conclusions

In this section we presented an information theoretic treatment of the auditory grouping problem. It was shown that many auditory grouping cues are just specific cases of statistical dependencies and that they are easy to detect by employing standard statistical

†. The first few values are $B_1 = 1$, $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, $B_6 = 203$, $B_7 = 877$, $B_8 = 4140$, $B_9 = 21147$, $B_{10} = 115975$, ...

measurements. This approach is characterized by many advantages, as compared to previously used heuristic methods. The most important advantage is a perceptual justification. Most approaches have been grounded on experimental psychoacoustic literature, but there was never an attempt to explain these findings as they relate to our environment, the development of our listening capabilities, and perception in general. By adhering to a Barlowian perspective, we were able to address these issues and come up with a more formal definition of grouping, which was easily translated to a computational algorithm. The implementation successfully performed in a similar manner as our auditory grouping process.

This approach also provided us with a solution to trade-off and fusing decision problems. Many implementations that are hard wired to fuse components exhibiting some fusing cues, are unable to deal with the case where two cues compete. This is because fusing decisions tend to be binary (components either fuse, or not), which sometimes removes the flexibility of judging more complex situations or even sharing components between two different sounds.

We have also purposely avoided the definition of a component. Although the examples we presented were done on a sinusoidal component framework, the resulting rule for grouping is quite independent of the component properties. In fact the only reason we used this representation is to construct a familiar interface to psychoacoustics. This is not a necessary representation though, in fact it is easy to use whatever seems appropriate instead. By using the abstraction of statistics we have made no assumption about the nature of a component, something that we will employ in the next chapter. The observation that many gestalt behaviors are stemming from statistics of natural sound producing systems might also be an indicator that it is possible to use this technique in different domains where consistent statistics exist. The visual domain is an obvious example since it exhibits strong structure.

One issue that we have not addressed in this section is that of time related grouping and streaming effects. In fact the common onset and proximity measures that we presented are unable to perform robust grouping for very long sequences and are only valid for short time scales. Unfortunately the algorithms required to deal with time related ICA processing are not yet developed and we are reluctant to go ahead in this direction. However, entropy for time stamped random series is defined as the entropy rate and the relevant mathematical theories exist. But until a computational framework is build we will not be in a position to validate such experiments. In the following chapter we will deal with streaming effects though in a different way that does not require generalization to time dependent information measures.

Chapter 4. Auditory Scene Analysis

4.1 Introduction

In the previous chapters we focused on low level perceptual elements. The operations that we performed were taking place in short timescales and they were modelling subconscious perceptual processes. In this chapter we will describe similar techniques that deal with larger time spans and result in seemingly more complex tasks. We will explore the applications of ICA on extracting features from auditory scenes and relate this to the perception of separate sounds. Our approach will draw inspiration from two very different fields with opposing viewpoints. One is the field of multichannel statistical methods, and the other the field of psychoacoustic-based monaural scene analysis.

4.2 Source Separation and Scene Analysis

Source separation is the holy grail of audio processing. Its goal is to extract auditory objects from a scene. It is an elusive process that has attracted a lot of attention and has commanded a lot of research. Due to this there has been development on many different approaches towards solving it. The two dominant methods that relate to this thesis, are the statistical and the psychoacoustic approaches.

4.2.1 Psychoacoustic Separation

Psychoacoustic separation methods have been investigated ever since the definition of the cocktail party problem. This problem was defined by Cherry (1953) and raised the issue on how it is possible that a human at a cocktail party, as subjected in many auditory sources, can understand and extract from the scene only the source that draws the listener's attention. Through time this problem has been transformed to finding a method that can computationally isolate auditory sources of a scene from a monophonic recording. This work was termed as Computational Auditory Scene Analysis (CASA) and drew a wide array of approaches.

In their most basic form most of the CASA approaches are attempting to extract objects by some form of perceptual grouping in the frequency domain. The usual path has been to utilize a time-frequency transformation and then extracting individual elements based on a set psychoacoustic grouping procedures. This approach flourished in the early 90's where any conceivable transform and set of grouping rules were used to this extend. Notable work in this vein has been by Weintraub (1986), Cooke (1991), Mellinger (1991), Ellis and Vercoe (1992), Brown (1992) and Ellis (1994, 1996), and has been partly described in the preceding chapter.

These various approaches introduced different methods that work best within their domain of sources, but they were not universal extractors. Most were bootstrapped for either music or vocal data, exploiting knowledge and procedures specific to these domains. In most of these cases separation extracted distorted sources, partly because of the loss of information due to mixing. In light of the realization that exact reconstruction is not always possible, recently the opinion of many researchers has changed to that accurate reconstruction should not be the goal of CASA. Auditory perception is more concerned with the detection and outlining of sound objects, rather than the exact reconstruction (which is not always feasible). This is a step closer to the initial cocktail party problem and presumably to human listening, which brought CASA back against a more reasonable challenge.

In a hostile viewpoint, most of CASA is still rather primitive and short-sighted. This is partly because of the fuzzy nature of its goals and unconditional faith to psychoacoustical literature. The most important problem however is that most of CASA approaches still don't have a definition of what a source is. Most often a source is heuristically defined as something that fulfills the requirements of being a separate object according to perceptual grouping. Given that perceptual grouping itself is also poorly defined, it is no surprise that there is a high incidence of extremely complex systems that produce extremely basic results. There is also a significant disregard towards general perception

which results in highly specific systems that cannot shed any light on the inner workings of the human mind. Finally the strong belief into psychological literature, prohibits the use of formal mathematical definitions. These problems are acknowledged by many in this field and are not personal musings. The desire to overcome these obstacles has been noted, and is an active subject of debate.

4 . 2 . 2 Multichannel Blind Source Separation

In contrast to CASA approaches, multichannel separation is a clean, well defined and understood operation. It deals with the case where we have a number of sound sources as recorded by a set of sensors. Each of these sensor recordings will contain a superposition of the sources, thereby concealing their individual form. By modelling the recording process and using knowledge of the setup, or some measurable statistics, we can cancel out interfering sources to extract the desired ones. This operation is easy to formulate in the context of linear algebra. Assuming a set of sources $\mathbf{s}(t) = \left[s_1(t) \cdots s_N(t)\right]^T$ we can model a multisensor recording process as:

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t) \tag{1}$$

where $\mathbf{x}(t) = \left[x_1(t) \cdots x_M(t)\right]^T$ are the recorded series and $\mathbf{A}$ is a real M by N matrix, known as the 'mixing matrix'. The structure and contents of the mixing matrix are dependent on the sensor and source numbers and physical locations (Figure 1).

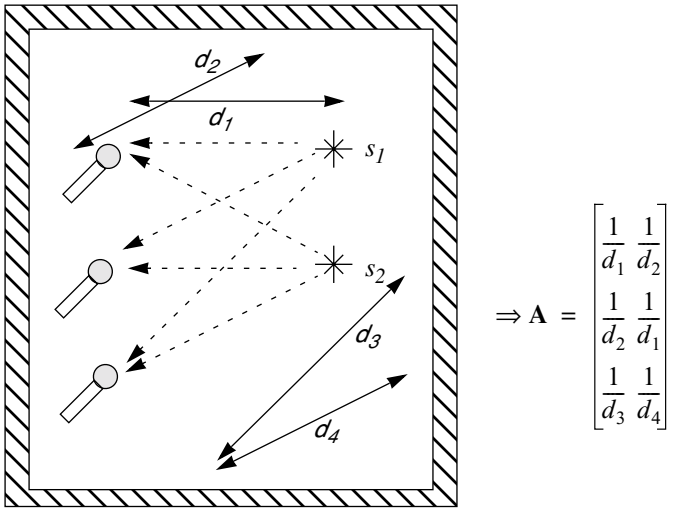

Figure 1 **Example of a multichannel setup. The stars s_1 and s_2 are the sources and the dotted lines their paths to the microphones. The solid lines indicate their corresponding distances. Assuming a simple inverse distance attenuation rule for the propagation of sound, we can deduce the mixing matrix for this scene**

If $M \geq N$ and $\mathbf{A}$ is full rank it is possible to recover the exact original sources by:

$$\hat{\mathbf{s}}(t) = \mathbf{A}^{+} \cdot \mathbf{x}(t) \tag{2}$$

where $\mathbf{A}^{+}$ is the inverse (or pseudoinverse if $M \neq N$) of $\mathbf{A}$. In practice the invertibility constraints that $M \geq N$ and full rank of $\mathbf{A}$, mean that no two or more sensors or sources can be at the same position, and that we have at least as many sensors as sources (If these constraints do not apply, it is only possible to try to reinforce the sources as much as the situation allows, but in general it not possible to fully recover them. We will not consider these cases here). This scenario is referred to as the instantaneous mixing scenario, since it does not model propagation delays and other filtering issues, thereby assuming an instantaneous transmittance of sound

More realistic mixing systems were subsequently formulated, modeling the mixing process as:

$$\mathbf{x}(t) = \underline{\mathbf{A}} \cdot \mathbf{s}(t) \tag{3}$$

where $\underline{\mathbf{A}}$ is an FIR matrix. That means that instead of scalar elements the matrix contains FIR filters and matrix multiplication of this object with a vector is defined as:

$$\underline{\mathbf{A}} \cdot \mathbf{x}(t) = \begin{bmatrix} a_{(1,1)} & \cdots & a_{(1,N)} \\ \vdots & \ddots & \vdots \\ a_{(N,1)} & \cdots & a_{(N,N)} \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix} = \begin{bmatrix} \sum_i a_{(1,i)} * x_i(t) \\ \vdots \\ \sum_i a_{(N,i)} * x_i(t) \end{bmatrix} \tag{4}$$

where the * operator denotes convolution and $a_{(i,j)}$ are FIR filters. This model accounts for the effect of reflections, propagation delay and reverberation in a recording environment.

In both of the above cases the objective of a source separation algorithm is to estimate an inverse of the mixing matrix ($\mathbf{A}$ or $\underline{\mathbf{A}}$). In general this matrix is not available to us, and it is deduced from observation of the recordings, and/or knowledge of the sources and the geometry of the setup. Modern techniques, known as blind processing techniques, have no knowledge of the environment (hence the term blind), and only use the statistics of the recordings to make their estimation. The only assumptions made are that different sources are not dependent to each other, so by trying to find an unmixing matrix that produces maximally independent outputs it is possible to obtain the original sources. Although there are many ways to attack this problem when armed with this assumption, the most relevant to our work is the application of ICA methods. It is quite

obvious that the outputs of Equation (2) can be forced to be independent using an ICA algorithm. This has been exploited by many researchers (see Torkkola 1999 for exhaustive reference) and in the case of instantaneous mixing is regarded as a solved problem. Equation (3), commands more complex implementations of ICA, it has been however successfully solved for a reasonable number of sources and filter sizes (Smaragdis 1997, Lee et al 1997).

Overall the ICA approaches to source separation yielded impressive results, superior to their CASA counterparts. However, the use of multisensor methods for source separation has not been accepted by the CASA community since it deviates on a key point for perceptual studies. Auditory perception is based at most on a two channel input (our two ears), and the constraint that current ICA algorithms impose is that we have at least as many inputs as sources. Although there is this obstacle in bridging ICA and CASA, ICA can boast a relation to redundancy reduction and the relevant perceptual theories which by far supersede the CASA philosophies in both depth and breadth.

Scene analysis, as has been pursued in the auditory community, has been intimately linked to source separation. In the rest of this chapter we will not abide by this convention. We feel that it is important to realize that understanding a scene and extracting useful information from it, is a distinct task from source separation (it is however driven by the same principles). This is a realization that has come to mature only recently (Ellis 1997, Scheirer 2000), and source separation is nowadays seen as a probably impossible task, and not crucial to perceptual processing. We will adopt this point of view and continue with this chapter presenting methods that lie in the fine line between detection and separation. Our primary goal will be awareness of the structure of an auditory scene. Issues of separation will be examined as side effects. Our approach will start from a methodology introduced by Casey and Westner (2000), which combines the formal rigor of the multichannel approach, with the realistic constraints of the perceptual approach.

4.3 Decompositions of Magnitude Transforms

4.3.1 Object Detection

In this section we will examine the problem of detecting individual sounds from an auditory scene. Our assumption, as in the previous chapters, is that by performing redundancy reduction we can yield interesting results. We will examine redundancy reduction as it applies to the time-frequency energy of signals. We will show that by extracting the energy of components of scenes we can deduce a lot of information about the events that take place. This is an approach that will try to use the philosophy of CASA approaches, but backed by the more rigorous theories relating to multichannel approaches.

Having a sound scene $s(t)$ we can analyze its energy content with respect to a set of bases using the magnitude of the transformation that these bases dictate. For example, the magnitude of a STFT transform will reveal the energy of all the sinusoidal compo-

nents in the scene as they evolve through time. In terms of linear algebra that is equivalent to:

$$\mathbf{F} = \left\| \mathbf{A} \cdot \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{s}_1 & \cdots & \mathbf{s}_M \\ \downarrow & & \downarrow \end{bmatrix} \right\| \tag{5}$$

where $\mathbf{A}$ is a N by N matrix expressing the desired transformation, the $\| \ \|$ operator performs element-wise magnitude extraction, and $\mathbf{s}_i = \begin{bmatrix} s(i+1) & \cdots & s(i+N) \end{bmatrix}^T$.

The resulting matrix $\mathbf{F}$ can be interpreted in two directions. If we observe it by traversing its columns one after the other, we can think of them as instantaneous magnitude transforms of the scene, localized in successive times. Observing its rows, we can think each one as a time series denoting the amount of involvement of a basis. We will refer to the columns of this matrix as spectra and the rows as energy tracks.

Since our thesis is that redundancy reduction of data leads to perceptual-like outputs, we will apply the same techniques we've used in the previous chapters to this representation. In order to build some intuition we will start with a very simple two object scene and its magnitude STFT transform (so that the $\mathbf{A}$ matrix in Equation (5) will be the Fourier matrix). The scene is one that displays interesting structure in both the time and the frequency axis. The two components of this scene are two amplitude modulated sinusoids:

$$s(t) = \cos(1321t)\mathrm{g}(6t) + \cos(3211t)\mathrm{g}(9.7t) \tag{6}$$

where $\mathrm{g}(\cdot)$ is a periodic gate function defined as:

$$\mathrm{g}(t) = \begin{cases} 1, \text{ if } \sin(t) \geq 0 \\ 0, \text{ if } \sin(t) < 0 \end{cases} \tag{7}$$

The odd nature of the constants defined in Equation (6) was such so as to avoid any gestalt grouping problems. The two gated sinusoids are intended to be separate and independent sources.

The magnitude of the STFT of this scene is depicted in Figure 2. To obtain it we used an FFT size of 128 and a hop size of 32 samples. This data will constitute our $\mathbf{F}$ matrix in Equation (5). Due to the symmetry properties of the FFT we only to retained the first 65 samples from each transform (64 frequencies and the constant offset).

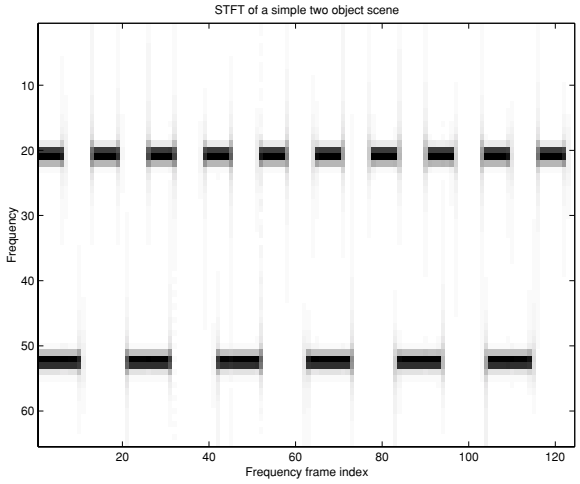

Figure 2 **The magnitude STFT of a simple two object scene. Note that the two objects feature different frequency and time characteristics**

By observation of the data we can extract basis functions for the matrix **F**. We can do so in either of the two dimensions. We'll first consider basis functions that describe the spectra (the columns of **F**). They would be 65 samples long and they should be able to reconstruct every column of **F** (the vertical slices in Figure 2). We will try to estimate these basis functions using techniques for redundancy reduction presented in earlier chapters.

We will start by using the PCA transform on **F**. PCA will provide a matrix $\mathbf{W}_p$ so that the operation:

$$\mathbf{C}_p = \mathbf{W}_p \cdot \mathbf{F} \tag{8}$$

will result in a matrix $\mathbf{C}_p$ whose rows will be decorrelated. The rows of the $\mathbf{W}_p$ matrix will be a set of bases comprising the transformation that results in $\mathbf{C}_p$. If we wish to resynthesize the matrix **F** using $\mathbf{W}_p$ and $\mathbf{C}_p$ we only need to solve Equation (8) with respect to **F**, from which we get:

$$\mathbf{C}_p = \mathbf{W}_p \cdot \mathbf{F} \Rightarrow \mathbf{W}_p^{-1} \cdot \mathbf{C}_p = \mathbf{W}_p^{-1} \cdot \mathbf{W}_p \cdot \mathbf{F} \Rightarrow \mathbf{F} = \mathbf{W}_p^{-1} \cdot \mathbf{C}_p \tag{9}$$

The columns of matrix $\mathbf{W}_p^{-1}$ will contain a set of bases that are required to resynthesize $\mathbf{F}$. In terms of transforms, $\mathbf{W}_p^{-1}$ will be the inverse transform† of $\mathbf{W}_p$.

In our case we only require a small number of cases and we do not care as much for an accurate reconstruction. In order to perform this dimensionality reduction on our transform $\mathbf{W}_p$ we only keep the first few rows, which correspond to the most significant‡ bases. In order to then recover the inverse transform we need to take the pseudoinverse of $\mathbf{W}_p$:

$$\mathbf{W}_p^{+} = (\mathbf{W}_p^T \cdot \mathbf{W}_p)^{-1}\mathbf{W}_p^T \tag{10}$$

The columns of $\mathbf{W}_p^{+}$ will be the bases of the inverse transform. They will effectively be the elements that when multiplied with $\mathbf{C}_p$ will reconstruct the scene. We will think of them as the building blocks of our scene. We will now examine these bases and their resulting transformation.

In the case at hand we will obtain a set of resynthesis bases which will be 65 samples long and can be used to recreate all of the spectral instances of Figure 2 (the columns of matrix $\mathbf{F}$). Since we only have two simple objects we will only keep two bases. The results we obtained are shown in Figure 3.

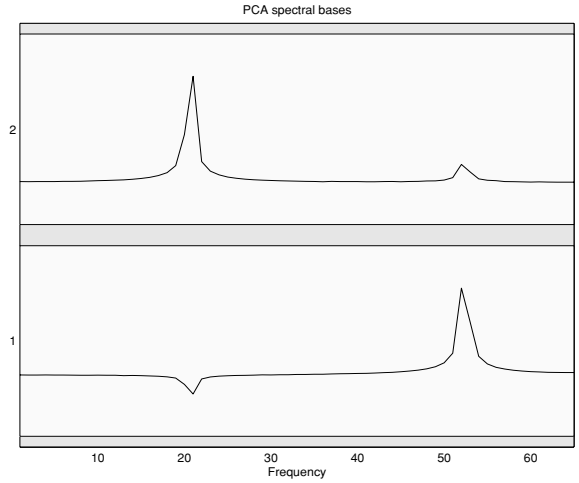

Figure 3 **Results of the PCA analysis on the data in Figure 2. This figure displays the estimated spectral basis functions, which almost reveal the two sinusoidal spectra.**

†. In the special case of PCA, the inverse transform matrix is just the transpose of the forward transform matrix (since $\mathbf{W}$ is an orthogonal matrix), and most references are as such. However, we generalize and use $\mathbf{W}^{-1}$ to accommodate a more abstract framework, which will help later on.

‡. In the case of PCA, 'significant base' implies a base that has a significant contribution in terms of variance contribution.

We note that these two basis functions roughly correspond to the two spectra that construct each of the two sinusoids. In order to increase the quality of the results, we will continue the analysis by a subsequent application of ICA on the resulting matrix $\mathbf{C}_p$. By doing so we will make the operation:

$$\mathbf{C}_I = \mathbf{W}_I \cdot \mathbf{C}_P = \mathbf{W}_I \cdot \mathbf{W}_P \cdot \mathbf{F} = \mathbf{W} \cdot \mathbf{F} \tag{11}$$

Where $\mathbf{C}_I$ is the output of the ICA transform $\mathbf{W}_I$, and $\mathbf{W}$ the overall transform that we have applied on $\mathbf{F}$ (the transform that includes both the PCA basis reduction and ICA steps). The inverse (or pseudo inverse if we reduce the number of bases) of $\mathbf{W}$ will contain the new ICA resynthesis bases is its columns. Applying this to our scene we obtain the bases in Figure 4.

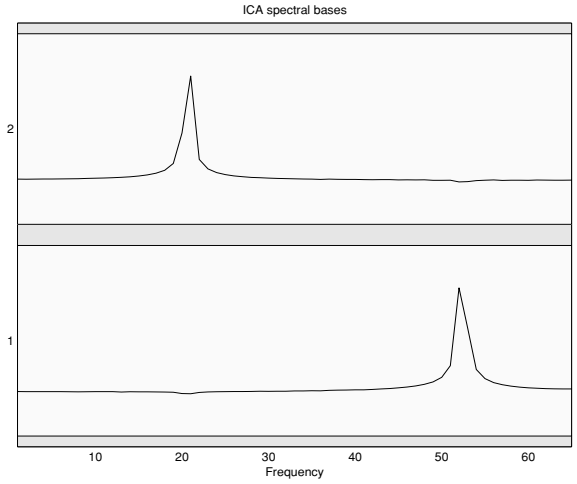

Figure 4 **Spectral bases of the scene in Figure 2 as extracted using ICA. Each of the two bases corresponds to one of the sinusoids in the scene, effectively discovering its composition.**

We now notice that the extracted bases are much more clean, and have a interesting relation to the spectral structure of the scene having highlighted the existence of the two different frequency peaks, each corresponding to one object. In effect we have isolated the frequency contribution of each object. It is also interesting to examine the resulting transformation $\mathbf{C}_I$ (Figure 5).

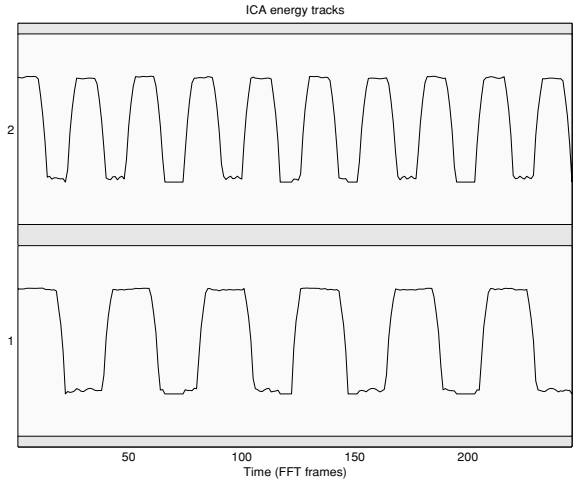

Figure 5 **The resulting transformation of the data in Figure 2. Note how the independent components are in fact the amplitude envelopes of the two objects.**

The rows of $\mathbf{C}_I$ contain the transformed energy tracks of $\mathbf{F}$ so that they are now maximally independent. We notice that this transformation results into two energy tracks describing the amplitude of the two objects in our scene.

The operations that were performed on $\mathbf{F}$ were that of reducing unnecessary data by the PCA dimensionality reduction, and then forcing the remaining elements to be statistically independent. As is easy to see from Figure 2 we had two kinds of significant energy tracks (the horizontal slices of $\mathbf{F}$), the tracks that corresponded to the amplitudes of the two objects. These were the two tracks that were retained by the PCA transform, they were not however returned in a clearly separated form (and in general they will never be), they were returned as two mixtures. The ICA step was responsible for the fine tuning required for separating these two components into two independent tracks. In effect this last operation is equivalent to the work in multichannel separation, only we are performing it in a subsampled transformed domain. We have produced the extra channels by the magnitude STFT transform and that allowed us to run multichannel ICA on an originally monophonic sound. The motivation to do this operation stems from the fact that the individual objects in a scene are independent. That means that they feature a set of energy tracks that would also be independent. By attempting to decorrelate these energy tracks and then seeing which spectra correspond to them, we hope to reveal some of the structure in a scene. This is the same reasoning applied in the multichannel ICA work, only on a different domain.

We will now examine the same operation as applied on the other dimension of the data. We will extract time bases by striving for independence of the spectral bases. To do so we only need to repeat the experiment above by setting the input to our analysis being $\mathbf{F}^T$ rather than $\mathbf{F}$. Just like before we obtain a set of time bases and spectral components, which highlight the two objects in the scene. The results after the ICA step are shown in Figure 6.

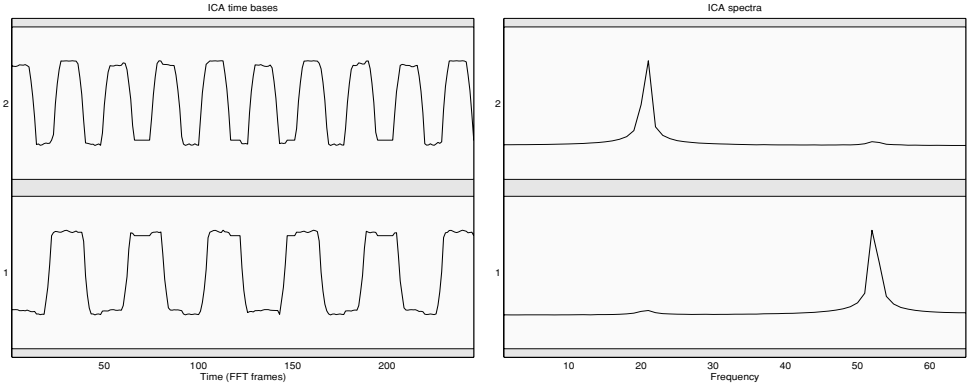

Figure 6

PCA analysis of the data in Figure 2. These results, although similar to Figure 4 and Figure 5, were obtained by applying the same procedure, but this time on the other dimension. This operation has the effect of making the transformed spectra independent, as opposed to the previous approach that made the energy tracks independent

It is interesting to note that we obtain qualitatively the same results as we have had using $\mathbf{F}$ instead of $\mathbf{F}^T$. This fact highlights an interesting point; independence between objects in a scene exists in both their frequency and time axes. Interestingly enough by looking at one domain we can extract information about the other. In the previous case we obtained our results by forcing the energy tracks to be independent and then seeing which spectral bases corresponded to them. In this case we performed the opposite, trying to get the spectral components to be independent and then extracting their corresponding energy tracks. Since clues of the objects' identities existed in both the time and frequency axis, it became possible to detect both by forcing a constraints on only one of them.

Obviously the scene we used so far was quite a simple example where there was, on average, little overlap in the time domain and none in the frequency domain and good discrimination between the two objects. We'll now use a more complex scene. It will be a three element scene comprised by the sum of the components displayed in Figure 7. These three components are drum instruments, a snare drum, a bass drum and a hi-hat. This is an example that displays a more complex mix of spectral and temporal variation. The solution is non-trivial since for some time instances we have overlap of multiple components and between all instruments we have spectral overlaps.

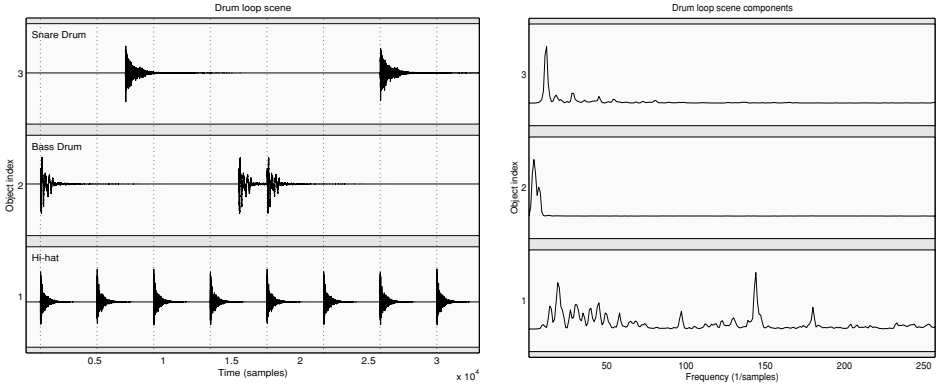

Figure 7 **The left figure displays the individual components of an auditory scene in the time domain. The right figure displays their corresponding power spectra. The scene itself is composed by the summations of these components**

By noting the time and frequency characteristics of the instruments we can make some observations that will help us in deciphering our results later. The bass drum contains a lot of energy at the lower side of the spectrum and its only isolated instance is its second one. The first and third instances occur simultaneously with the hi-hat. The snare drum has a resonant character in the low-middle frequencies and is somewhat wideband. Of its two instances only the first is isolated, the other coincides with a hi-hat. The hi-hat has a wideband tone with some high resonant frequencies. In addition to its temporal overlaps we mentioned above there is also significant spectral overlap with the snare drum.

We analyze this scene using the magnitude STFT again and obtain the matrix $\mathbf{F}$ shown in Figure 8. We used an FFT size of 512 and a hop size of 64 samples. As we had done before we truncated the frequency axis to keep only one half of the symmetric spectrum plus the constant offset.

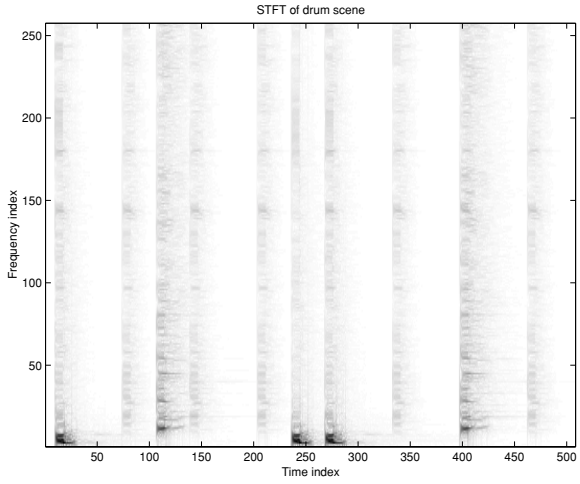

Figure 8 **The STFT of the drum loop scene. The individual components are easily identified by eye from their temporal and spectral characteristics.**

We apply our previous procedure using PCA dimensionality reduction on $\mathbf{F}$, down to four bases, and subsequently follow this step with application of ICA. This operation makes the temporal features maximally independent and provides the results in Figure 9.

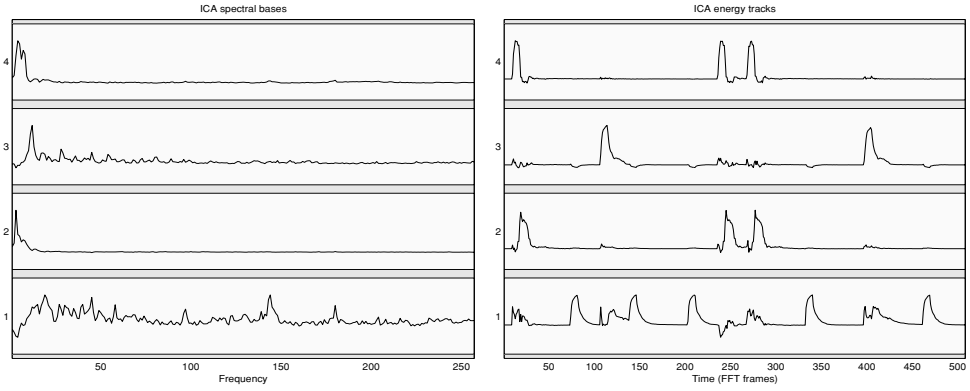

Figure 9 **Analysis of the data in Figure 8. The left figure displays the estimated resynthesis spectral bases, and the right figure their energy through time. In contrast to the original objects we have a fair decomposition of the scene to its original comprising elements**

Observation of the results unveils the structure of the scene. Easier to decipher is the resulting transform $\mathbf{C}_p$ (right figure in Figure 9). We can easily see that the second and fourth energy tracks correspond to the bass drum. The fourth appears to deal with the snap in the attack portion of the bass drum, whereas the second one deals with its low

frequency resonance. By examining their spectral counterparts, we see that they are both exhibiting a low frequency content. The one corresponding to the attack portion has a slightly higher frequency content, whereas the other has most of the low end energy. Likewise we see that the snare drum energy track is well captured in the third component. As we might expect the corresponding spectral base has the same resonances as the snare drum. Finally the first component is the estimated energy of the hi-hat. Due to the fact that it temporally coincides with the bass and the snare drum for some instances, the time track is not perfect. The spectral base though is very accurate and encapsulates all of its characteristics.

In order to deal with the temporal problems in our estimation, we repeat the process using the other dimension of the spectrum by analysis of $\mathbf{F}^T$. This will perform statistical separation of the spectra of the objects, rather than the time energies, which should alleviate the time tracking problems we have. The results are very similar, the spectra indeed correspond to the sounds we had in the scene, and the temporal functions are more complete now. We do note however that the energy track corresponding to the hi-hat does contain some extra events. Due to the spectral similarity of the snare drum and the hi-hat (they both feature a wideband spectrum), their separation was not perfect and we notice that the hi-hat track also includes the snare drum events.

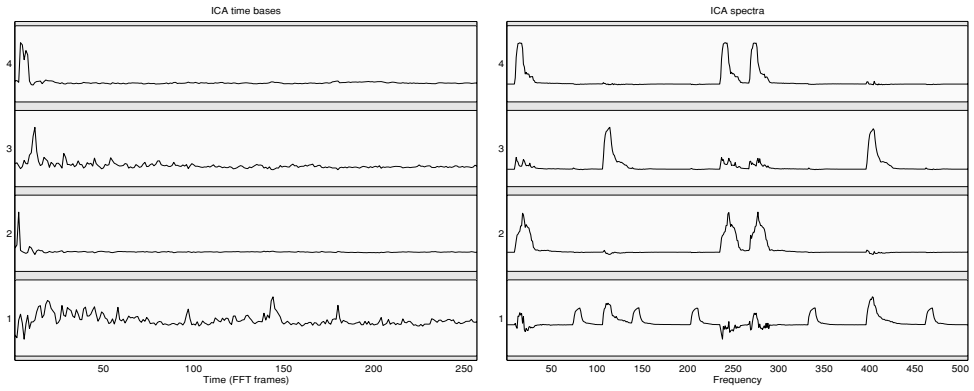

Figure 10 **ICA analysis of the data in Figure 9. The two figures are the corresponding ICA outputs to Figure 9. In contrast we see that the temporal estimation of the hi-hat was more accurate, although there is confusion with the snare drum.**

In general, the issues that arise with these kinds of overlaps are not easy to solve. The process of monophonically mixing a set of sounds results in information loss that we cannot reconstruct without additional knowledge. For this reason it is hard to extract exactly what occurs at all points. In many cases this requires the use of a knowledge system to interpolate missing data. In our case the only form of knowledge is the knowledge discerned from the presented scene. Considering that the drum scene encapsulates all of the auditory experiences of our system, the results are not bad at all.

As hinted by the initial definition of this process, it is not imperative to use the magnitude STFT transform for this process. The magnitude STFT was used due to its desir-

able phase-invariance properties and since it is a fair approximation to the bases that we have discovered in the second chapter which does not complicate the implementation. It is quite possible to substitute this transform with a DCT transform, or even PCA or ICA derived bases. In fact, the transform can even be the unit matrix (or for that matter any linear transform). The performance of this algorithm however is strongly biased depending on the transform type. The main goal of the transform step is to obtain a sparse decomposition of the data which can consequently assist the PCA and ICA steps to have better convergence. By taking the magnitude of the transform we create an invariance to the sign of the data, and in the case of the STFT an invariance to the phase of the data. Perceptually these invariances exist in the human auditory system, and they help provide a good sparse coding of what we hear. Computationally these invariances venture in the domain of nonlinear ICA since they deal with a non-linear and non-invertible transformation, and complicate matters considerably. It is outside the scope of this thesis to delve into this territory (it is also a volatile and relatively unexplored subject), but we will however bring up the subject once more in the last chapter.

Similar results to the ones presented can be extracted from a variety of scenes, and the results are dependent mostly on the density and overlaps of the scene objects. The issue of invariance comes up again in the case of complex sources. So far we have dealt with auditory objects that have a fairly static spectral character. Once we attempt to extract sources with a more complicated structure, it is quite impossible to obtain results by forcing the spectra to be independent due to the absence of consistent spectral characters. For such complex scenes it is preferable to force independence of the energy tracks instead. This implies analysis of $\mathbf{F}$ rather than $\mathbf{F}^T$ and we will deal with that in later sections.

4.3.2 Applications in Music Processing

The estimation of independent spectral bases opens avenues of exploration for music processing. A traditionally hard operation on audio processing is music transcription. It involves the translation of a raw audio stream to musical semantics (most systems attempt only the extraction of musical notes, a daunting task by itself).

Fitting our framework, we present a brief example of how we could use the method developed in the previous section for note detection. We used the first two bars of Bach's first invention in C major BWV 772 (Gould 1966). This being a solo piano piece, it offers a highly structured scene. There is a strong dependency to frequency templates which are the present piano notes. This is a structure we can easily extract by the aforementioned analysis method. We repeatedly performed the analysis using windows lasting one bar (total two windows to cover the entire scene). The first window of analysis contained a segment with one note sounding at any time, whereas the second one contained a polyphonic passage. We performed ICA on $\mathbf{F}^T$, thereby striving for independence of frequency components, since we know that the structure we care for lies there. We kept five components from the first window analysis and twelve from the second. The resulting frequency bases from the first analysis are displayed in Figure 11.

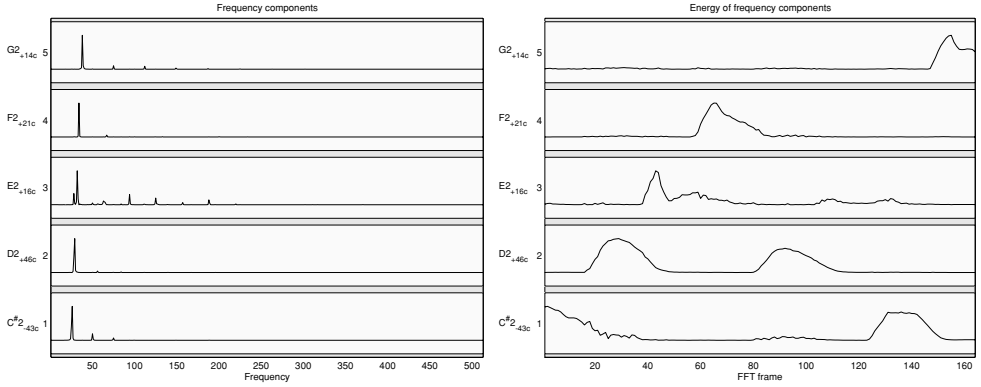

Figure 11 **Frequency components (left) and their corresponding energy (right), from analysis of the first bar of the Bach's first invention. Note how the frequency bases correspond to note spectra and the energies to their respective time locations.**

As expected the frequency bases were the spectra of the notes in the scene, and their energy tracks provide their corresponding time location. Cross referencing this data with the score of the piece proves that we have correctly identified the position of every note†. The second bar is a more challenging case, since it exhibits polyphony. To compensate for the added harmonic activity we extracted twelve bases and used a longer FFT length. The results were similar and are displayed in Figure 12. The first six bases correspond to the left hand notes, and we have only one undetected note and an octave ambiguity for the last G. The top six bases capture the trill between C and B, as stretched out and jagged energy tracks.

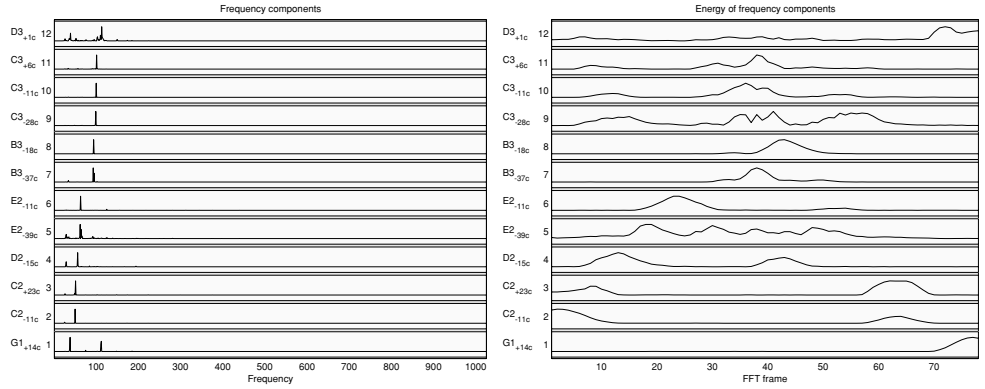

Figure 12 **Frequency components (left) and their corresponding energy (right), from analysis of the second bar of Bach's first invention.**

†. An historically interesting point to make, is one that is reflected in our results. There is a reason why the third basis does not have a smooth energy track, and seems to accent the start and end of every note instance and is rather low and noisy during the middle. This is an effect due to the fact that the piano used in this recording was damaged during a transfer, and was recorded having a middle register that suffered from a mechanical tic that resulted in an audible hiccup. In the analyzed passage, the E note, represented by the third basis, is one of the affected notes, and the energy track has reflected that!

From an evolutionary standpoint this is a very interesting results, since it highlights the connection between statistically strong features and musical building blocks. The interesting point to make is that the notion of notes was implied by examination of the sound scene and it was not specifically pointed out. We have thus, created an analysis from which our system can deduce musical information from listening, rather than being explicitly told what to look for.

4.3.3 Auditory Object Extraction

Since we do have a way to detect and isolate objects in a transformed domain it is conceivable that we can perform an inverse operation to extract individual objects. Using the magnitude transform to obtain our data, we face the ambiguity of phase as we attempt to perform the inverse transform. Having kept only the magnitude data, we have no information about the phase. One, admittedly poor, way of doing this is to use the phase values from the initial scene and modulate them with the newfound amplitudes. Although this is not a precise and clean way to reconstruct the data is provides reasonable results (we should keep in mind that exact reconstruction of each auditory object is not always possible, and this pursuit is merely to satisfy our curiosity of how the components sound).

To explain this more formally, having obtained a set of bases $\mathbf{W}$ and their coefficients $\mathbf{C}$ from the original input $\mathbf{F}$, we can reconstruct $\mathbf{F}$ by:

$$\mathbf{F} = \mathbf{W}^{-1} \cdot \mathbf{C} \tag{12}$$

As we mentioned we do not need to keep all of the bases in $\mathbf{W}$. If $\mathbf{W}_r$ is the matrix that contains the reduced set of bases, we can reconstruct the data-reduced $\mathbf{F}$ matrix (denoted as $\hat{\mathbf{F}}$) by first doing:

$$\mathbf{C}_r = \mathbf{W}_r \cdot \mathbf{F} \tag{13}$$

to obtain the coefficients for the reduced bases, and then:

$$\hat{\mathbf{F}} = \mathbf{W}_r^{+} \cdot \mathbf{C}_r \tag{14}$$

to get the thinned-out $\hat{\mathbf{F}}$. For the sake of illustration had we wanted to obtain a reconstruction using only one component, we could use:

$$\hat{\mathbf{F}}_i = \mathbf{W}_{(i)}^{+} \cdot \mathbf{C}_{(i)} \tag{15}$$

where the subscript on the right hand matrices, selects the *i*th column of $\mathbf{W}^+$ and *i*th row of **C**. This will give us the magnitude spectrum of only the *i*th component. In order to convert this to an invertible spectrum we can use the phase of the original input **F** and amplitude modulate it by our new $\hat{\mathbf{F}}_i$ as:

$$\mathbf{S}_i = \hat{\mathbf{F}}_i(\cos(\angle\mathbf{F}) + i\sin(\angle\mathbf{F})) \tag{16}$$

$\mathbf{S}_i$ will contain the approximate time-frequency spectrum of one component, and we can use it as the input to an inverse STFT to obtain the component in the time domain. Using this reconstruction we can extract audible representations of the sinusoids, drums and piano notes presented in the previous examples.

Following is an example of a section from the pop song “Da da da” (Trio 1981). The particular section contains instances of the word “da”, a synthetic clave sound, a bass line, some guitar strumming and drums.

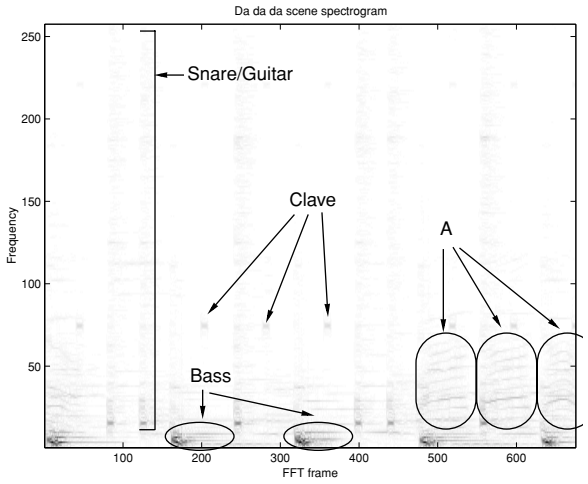

Figure 13

Looking at the spectrogram we can locate the individual sounds. The utterances of the word da can be distinguished by the formant structure of speech. The clave is the high frequency ‘beep’ which we see halfway through the spectrum. The snare drum and the guitar strums coincide in time and are the large vertical columns. The bass and the bass drum are located at the bottom of the plot. For legibility purposes not all instances of each sound are marked.

Applying on this scene the aforementioned analysis and resynthesis method we can effectively, detect and pseudo-extract the instruments that are present. In this case we analyzed **F** (looking for independence of energy tracks) and used the STFT with an FFT size of 64 and a hop size of 8. We also applied a hanning window to improve clarity of reconstruction and avoid frame to frame clicking transitions. We used PCA to reduce

the number of energy components down to nine. Out of these nine components, at least six corresponded to the various instruments in the scene (Figure 14).

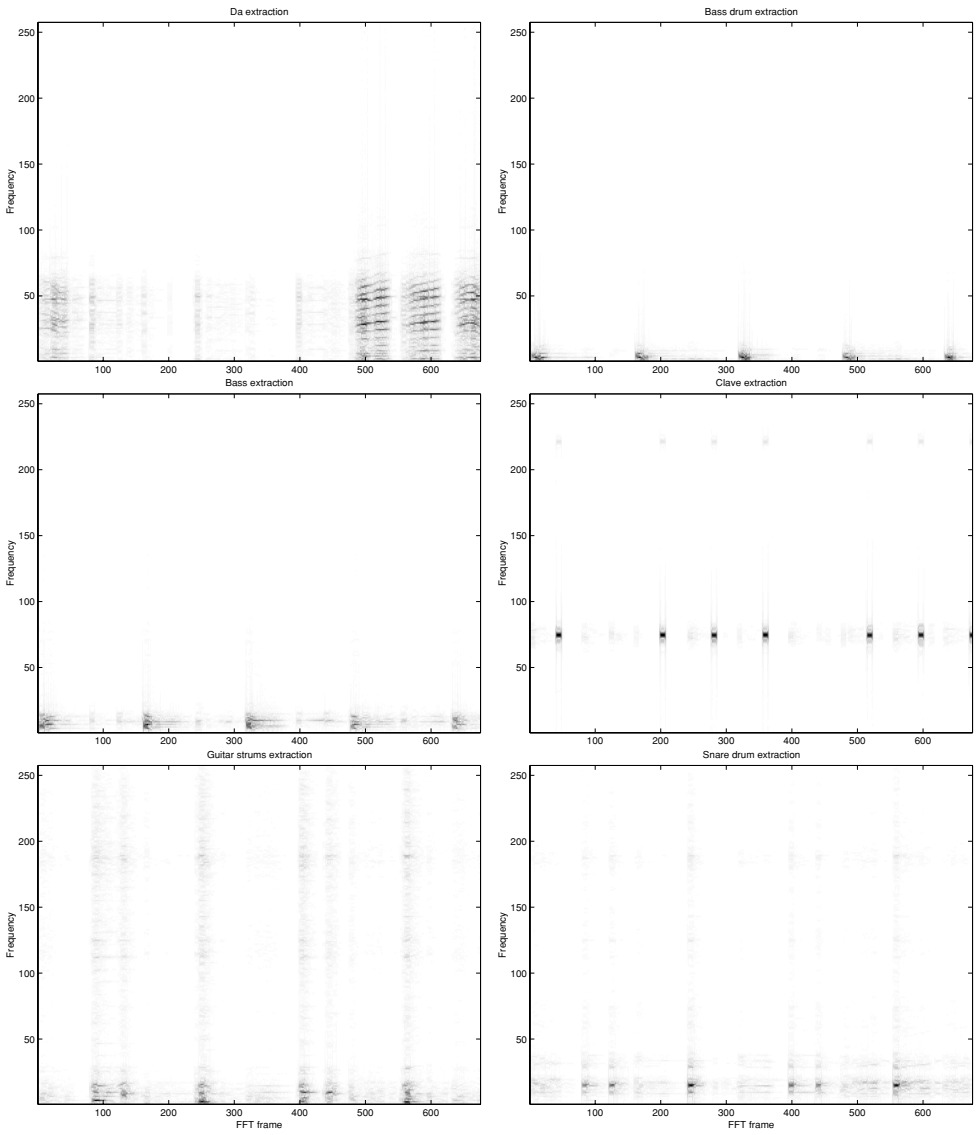

Figure 14

Clockwise from top left: extraction of the 'da' utterances (note the formant structure), the bass drum (low and short bursts), the clave (high pitched short 'beeps'), the snare drum (wideband bursts, low-mid resonance), guitar strumming (somewhat contaminated by the snare drum, but exhibits a more harmonic structure than the snare drum track), and the bass part (with a clear harmonic structure and low notes).

The results have been quite satisfying. The algorithm successfully tracked the major components of the scene and provided a reasonable reconstruction. Extraction was not

comparable to multichannel techniques though. For example the guitar and snare drum, whose instances were synchronized, posed a hard, if not impossible, separation problem. The missing information due to their mixing cannot be reconstructed and the resulting components are not clearly separated (however, we can distinguish them quite easily as the guitar and snare drum tracks).

4 . 3 . 4 What is a Source? What is an Auditory Object?

An issue that arises at this point is the definition of a source, an auditory percept, or an auditory object. It might have been noticeable that in the preceding section we used the term "auditory object", rather than "source", to describe what it was we extracted from the scene. Preceding researchers have consistently defined their extracts as sources, a term we did not use. In this section we will attempt to make a distinction between a source and an object, as they relate to the extraction process.

The concept of a source is usually defined in terms of human perception. It is easy to make for most sounds; a passing car, a person speaking, or music playing from a speaker can be a source. These are percepts that provide some sense of continuity which makes us perceive them as sources. Even though the car can shift gears, the speaker stop speaking for a while, or the music be composed from a changing setting of instruments, we still perceive them as one auditory source, regardless of the drastic changes they undergo. This is because we have a prior knowledge of how these sources are produced and we are able to make higher level judgements on how to piece such varied sounds together to construct a source. For different reasons, both the CASA and the multichannel approaches perform extraction of sources. In the CASA case extracting sources is a hardwired function and a primary objective; in the case of multichannel systems the inputs are spatially constrained in such a way so that they imply a source structure.

Our approach, although heavily indebted to CASA and multichannel work, extracts a different kind of auditory entity. We will use the term auditory object. An auditory object is a simpler element which is easily defined as an independent set of data in the scene. The particular method we provide to extract this independent set works by defining spectral or temporal templates. In the case of spectral templates, an object is extracted as a frequency response which, in the analyzed scene, is amplitude modulated in a statistically independent manner. Likewise, extracting temporal templates results in statistically independent amplitude tracks, being harmonically modulated. It should be quite clear that this type of analysis is inadequate in capturing a source as defined in the previous literature. So it is important at this point to make a distinction between a statistically independent object in the scene and a source. In a strict mathematical sense a person speaking in between long pauses, although a source by itself, provides many auditory objects which can be either the syllables, the words, the sentences, or some other coherent mass of sound (a fact determined by our method and length of observation). Using our particular method, we extract objects that have some common frequency or time modulation. We cannot expect to extract speech, or music as one object because it is not. We can only pick out coherent pieces that compose them. If we bring ourselves to the 'mindset', of our algorithm it is easy to understand this. The submitted scene is the only stimulus the algorithm receives and all knowledge we use is extracted

from it. Had the scene had only spoked word, the algorithm not having the knowledge of what speech is, it would be quite natural to say that each different coherent section of the speech signal is a different object. We can not expect at such a low level to extract complex representations. We perceive this as one source, but this is because we also analyze the semantic content, and use the timbral continuation to deduce that it originates from a single speaker, thereby concluding it is one source. Our algorithm, does not have access to all this information and makes a simple statistical judgement. Now if the spoken word was such that it offered some kind of repetition to imply continuity, then the connection would be surely made (such is an example of the "da da da" extraction, where the repetition of the utterance da makes the algorithm extract all of the sung part as one object).

The following example highlights the difference between a source and an object. It is a segment from the jazz song "Blue moon" as sung by Billie Holiday (Holiday 1945). The section we analyzed contains Holiday singing "you saw me standing alone" over a bass line, a piano part and some faint drums (Figure 15). Our goal was to extract just the vocal part. Since this part is hard to represent with only one component, we tried to extract more components and selectively combine them to reconstruct all of the sung part.

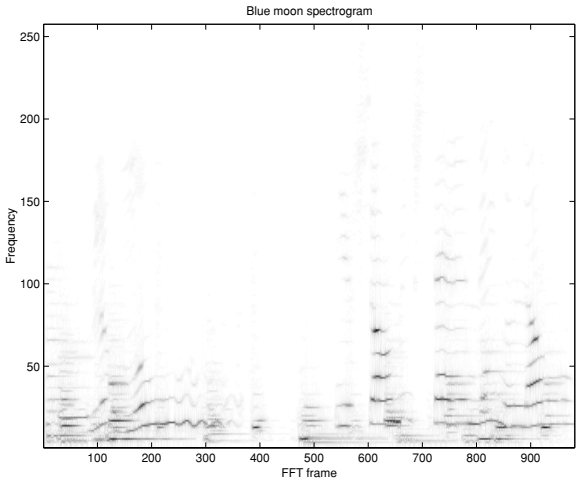

Figure 15

Extract from 'Blue Moon'. The vocal parts, easily seen by their formant structure, cannot be extracted by a single frequency template as we have before. To overcome this problem we need to combine a set of components to reconstruct the entire sung line.

We extracted nine components, and some of them were corresponding to speech. Individual components have adapted to represent dominant and locally static parts of the singing line (the auditory objects). These were mostly the vowel parts which are sung most prominently (Figure 16) and display statistical coherence amongst them.

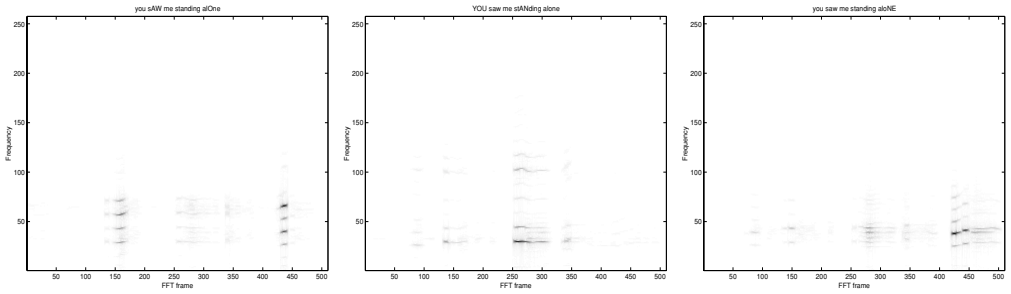

Figure 16 **Extracted vocal components. The three components highlight different sections of the sung part. Complete reconstruction of the vocal part requires the combination of all these components.**

Once the vocal components are extracted we can reconstruct the entire vocal part by just combining them. Likewise, when we analyze complex scenes we need multiple components to extract complex sources such as speech, lines of a solo instrument etc.

Another way to deal with the issue of needing many components to represent one source is the use of analysis frames. As we have done in the previous example analyzing the Bach invention, we can subdivide the scene into multiple segments and analyze each one individually. Each frame, due to limited content, will require fewer components to represent all the sources, and by proper recombination of components we can reconstruct a source over the course of many analysis frames. Unfortunately, finding the successor of each component in the consecutive frames is a very hard problem and this method of analysis is rather hard in all but trivial cases. To remedy this, online algorithms can be used to provide a running estimate of the components on a sample by sample basis. Although this approach eliminates a lot of the work required to match components over time, it provides poorer estimates due to its short time frame of operation. In any case though, the results we have so far are promising and perform fairly well under reasonable conditions, the only major drawback being that this process requires human guidance. Most work in this area has been preliminary though, and we are confident that in time these problems can be addressed in a more mature manner.

4.4 Conclusions

In this chapter we presented a methodology for decomposing auditory scene to a set of independent features. The approach was a hybrid of the multichannel statistical approaches, and the monaural psychoacoustical models. This was inspired by the fact that although the psychoacoustic approaches were dealing with a perceptual problem (one input, many outputs), they were not using the elegant formulations of the multichannel approaches (which in turn were not dealing with a perceptually realistic problem). The formulations used for the multichannel approaches, aside from rigor and depth, had a strong link with perceptual theories and provided a, conceptually, very strong platform for scene analysis. Once combined with the scenario of the psychoa-

coustic approaches they resulted in, what we feel, is a much more plausible process for scene analysis.

This particular approach is however a starting point rather than a solution. It has many unclear points, and a fair amount of manual fine-tuning to obtain results. Its relative simplicity and conformance to perceptual theories make it a compelling approach which can potentially lead to more elegant auditory scene analysis. Hopefully it will mature in time and provide greater autonomy and robustness.

Chapter 5. In Closing

5 . 1 Thesis Overview

The driving impetus behind this thesis was the lack of representation of modern perceptual theories on auditory perception. As we have described in previous chapters, the majority of work on auditory perception has been rather limited in scope, and highly dependent on psychoacoustics. The approach we have followed, was that of abstract perception, with no particular ties to auditory representations. We also avoided complex and heuristic verbal descriptions stemming from psychology, but instead formulated various stages of low level auditory perception to fit a compact redundancy reduction principle. In order to avoid using our own biases and knowledge when it comes to listening, we examined perception from an evolutionary perspective, and pondered on what it means to evolve a sense of perception from raw observation of data.

In the first section of this thesis we presented the formation of auditory preprocessing filters as they relate to their stimuli. We showed that speech signals are optimally decomposed, in terms of redundancy reduction, by time-frequency localized sinusoids, very much like the filters used to model human auditory systems. We then proceeded to show that the perceptual grouping rules that apply to audition are also products of the same process. Auditory atoms that belong together can be grouped by attempting redundancy reduction. We speculated that the reason why we tend to fuse the patterns we fuse, is a direct consequence of their statistics. The various cases where fusing happens are cases where the auditory atoms are dependent, something that implies a common origin. Given our development in an environment where these dependencies always accompany the auditory atoms of sound producing mechanisms, it makes sense that our system interprets these dependencies as a hint of unity. Finally we formulated the most complex process of lower level listening, scene analysis, by employing the same principle once more. We presented results in which we have managed to detect individual sounds and extract them. We did so by building a small kernel of knowledge that was deducted by the scene we examined. By exploiting repetition of spectral and temporal characteristics and by trying to find a maximally independent set of these, we were able to distinguish the different objects that made up a scene.

Although as presented these three formulations seem rather irrelevant to each other, they do have more than just redundancy reduction in common. The sinusoid representation that was deducted in out pursuits for optimal basis selection was used later on in the form of the STFT front end for scene analysis. Having strong indication that an STFT type transform is an approximation to the optimal in terms of redundancy reduction we used that instead of attempt to derive custom filters using the little data we had. Admittedly a constant-Q type transform would be a better interpretation of our results for preprocessing, but in order to accommodate clarity and simplicity the STFT was used instead. This by no means precludes the use of a more advanced transform type, perhaps one that has better characteristics (Wigner-Ville, or wavelets), or one that is entirely derived from the data itself (this however requires large scenes for better results). The way we reduced and combined the energy tracks for scene analysis was intimately linked to the grouping section. Although we had used sinusoids with variable frequencies for grouping, we essentially repeated the same procedure for the constant frequency narrowbands of auditory scenes. Due to the data invariant formulation of grouping we used the same computational method, albeit in a different way to accommodate the different form of outputs we required.

Taking a step back and observing our work in a global manner, allows us to see a point we wish to stress (and have been doing so throughout). Perception is a process that has evolved to interpret the statistics that exist in our environment. Things that are common and repeating we tend to ignore (for example the existence of partials in harmonic relations), and things that do not make statistical sense confuse us (auditory illusions, noise). Unfortunately with the advent of sound synthesis technology psychoacoustics has been focusing on simple scenarios involving highly improbable and unnatural sounds, such as the scenes we used in the third chapter. This allows observation of interesting effects, but also skews our interpretation of what perception does. For example, the grouping rules that we have observed are merely side-effects of our experience on parsing scenes. It would be quite improbable that our listening mechanisms are using such simple rules to parse complicated inputs. It is much more likely that the

expertise we have acquired for detecting sources and recovering statistics, is influencing this behavior. This is not however a widely understood point and, in our opinion, is a cause for misplaced research efforts. Through our work we hope to stress the importance of the structure of natural sounds when it comes to modelling and understanding auditory perception.

5.2 Future Work

As it should be obvious by now, we are staunch believers in the relation between redundancy reduction, perception and environmental statistics. This is a philosophy that has infiltrated other fields of perception but not so much audition. We provided some indication on how it relates to lower level auditory perception, but there is no reason to stop there. Redundancy reduction can be used at many levels of listening processes, and could be a unifying tool for the creation of compact and elegant systems. Preliminary results were successful in application of these principles to the discovery of themes in contrapuntal music, pitch detection based on examination of common information with pitch templates, and identification of simple waveforms. More adventurous topics that delve into the realm of aesthetics include the relation between the notion of mutual information maximization with consonance. Just like harmonically related sinusoids display are maxima of mutual information, musical intervals and chords can yield similar characteristics. Remarkably enough, we note that the intervals and chords that are known to be consonant are in fact maxima of mutual information. To further entice our interest we discovered that the more consonant an interval or a chord is believed to be, the more mutual information it possessed. This can prompt theorizing about the aesthetics of listening based on informational measures. Although out of the scope of this thesis, this is a project that has been pursued in other domains and looks very promising in the case of audition.

In terms of computational backup we have the fortunate assistance of excellent research on redundancy reduction and related principles. Some of the most interesting directions, as far as this thesis is concerned, are those of multidimensional (Cardoso 1998), nonlinear (Hyvärinen 2000) and invariant ICA. They all deal with problems that are very hard and can find immediate application in our work. Multidimensional ICA deals with the issues that deal with groups of components, and not necessarily individual components. In both the grouping and the auditory scene analysis chapters, we had to deal with many components, some of which belonged together. Fortunately in the first case we were able to deduct the number of groups, but in the latter we always had to make a decision on how many components to use, and which of them were describing the same source. Multidimensional ICA deals with the issue of grouping the proper sets of components that 'belong together'. It is largely unsolved as a problem although some solutions are slowly emerging (Hyvärinen 2000). Nonlinear ICA deals with extracting nonlinear mixtures of components. We had a taste of nonlinear ICA in the scene analysis section where we used the magnitude of the spectral data. By doing so we were able to perform redundancy reduction in a phase invariant manner. Nonlinear ICA is a very general definition and introduces a lot of complications, it is however promising when it comes to the design of invariant models. Such properties are highly

desired in perception, and these approaches are bound to play a more prominent role in perception when they are better understood.

As far as audio is concerned, this work will hopefully inspire some long overdue studies on the statistics of real-world sounds. Unfortunately most computational audio research deals with sounds as being either simple stochastic processes, or even worse, a random variable (most often presumed Gaussian). It is our hope that we have provided some proof that sounds are not just a Gaussian variable, and require the appropriate treatment for any meaningful statistical operation. We also hope that we have shown that when considering better tailored statistics for audio, it is quite simple to extract meaningful information from sounds. Many of the contrasts between PCA and ICA in this thesis were meant to be a demonstration of what happens when a better model is used. What we did not examine though, and is still an open question, is the statistics of sound as a process. A lot of our work was deriving sound structures by statistical analysis of audio. It was however constrained in terms of time scales exposing only specific behavior. Our preprocessing method was only used for short time windows seldom exceeding 128 samples, and likewise the scene analysis work was presented using scenes of a fairly short length, yet enough to contain enough information. This ad-hoc selection of the analysis windows was an unfortunate case of external influence which we have tried to avoid otherwise. We can speculate that statistics across different window sizes are drastically changing, prompting the same analysis procedure to result in different outputs, intimately linked to the analysis scale. It should be noted, as an example, that the preprocessing and the scene analysis methods as presented differ only slightly in the incorporation of the frequency transform[†]. Their truly primary difference is the difference of scale, fractions of a second for preprocessing vs. seconds for scene analysis. Analysis of the possible transitions in the data and the response of these methods as it relates to analysis window length is an area still unexplored and open for experimentation. Pursuit of this could result in a better understanding of the nature of sound, leading to better tailored approaches of dealing with it.

In the philosophical side, considering the breadth of applications that such an approach has, we could stand out on a limb and propose that it might help formulate a set of goals for audition. For example noting the relations between consonance, grouping, identity of a source, and mutual information we could assume that the sensory mechanism is most satisfied when reaching such optima. We can theorize that the goal of our auditory system is to derive satisfaction by decomposing complex input into simpler structures. In such a framework, sounds like white noise or simple tones would be highly unsatisfactory since they cannot be decomposed any more and they provide little or no structure to exploit, whereas music, exhibiting a rich structure in many levels, would come as a interesting task for examination. We of course will not go as far so as to back-up such an ambitious claim, we do find it however satisfying that we can use the same measure we have exploited so far to express global goals of perception.

Hopefully, thinking along the lines of statistics and redundancy reduction can yield a complete model for audition (or perhaps perception). This is a very vibrant and rapidly

†. As described before, there is no real reason why to only use the STFT as a preprocessing step. Any other form of transform will yield some sort of reasonable results. If this transform is the identity, then we will be presenting batches of the original sound, exactly as we do for extracting the audio bases.

growing field which promises many interesting development in the near future. Unfortunately, the auditory community has been largely unaware of these approaches. We hope that through our work we have provided some proof and an encouraging interface so that people involved in auditory perception research will consider joining this trend.

Appendix A. Multimodal Examples

A . 1 Introduction

As vigorously preached in this thesis, the fact that we avoid to measure parameters and assume no particular data format, facilitates the generalization of our work to other domains. Particularly interesting is the case of multimodal perception. In this appendix we present some simple examples on how it is possible to apply our work in video and joint audio/video scenaria. In particular we examine the cases of grouping and scene analysis. Preprocessing similar to our work has been adequately covered in the visual domain by Olshausen and Field (1996), Bell and Sejnowski (1997), and Hyvärinen and Hoyer (2000).

A . 2 Multimodal Grouping Example

The grouping process that we described in chapter 3 was purely in the auditory domain and it featured simple and similar elements. In this section we present an example which performs grouping across multiple types of stimuli.

In order to illustrate multimodal grouping we construct a scene consisting of one sound and a visual stream. In particular the sound was a speech segment and the visual stream was composed of two regions of activity. One region was intensity modulated by the speech signal, whereas the other was modulated by a sinusoid of one period (Figure 1).

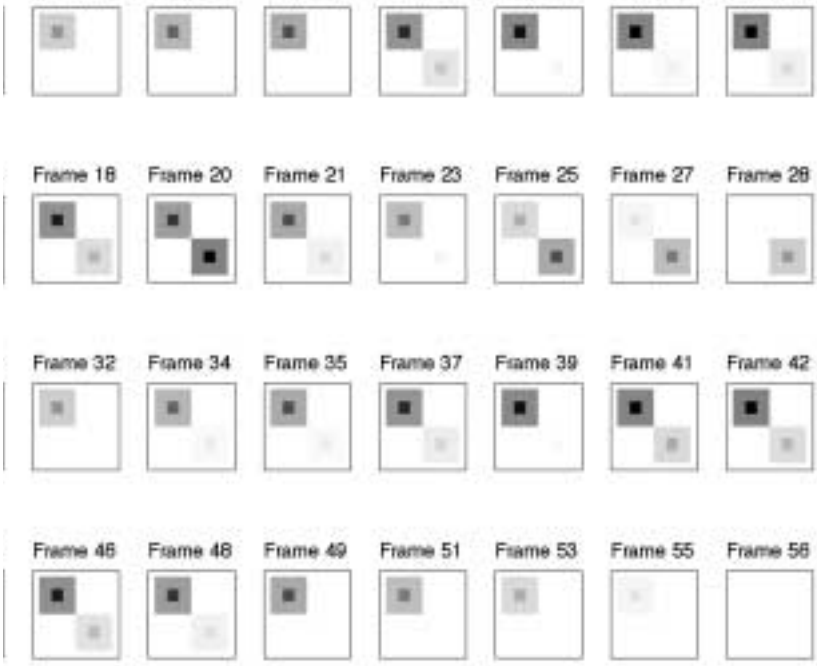

Figure 1 **The unraveled visual stream. The composition involved the visual configuration at the top left corner, whose intensity was modulated by a sinusoidal period, and the configuration at the lower right corner similarly modulated by a speech signal.**

We employed the same scheme we used for grouping in chapter 3. The speech signal was assumed to be one component. The other components were the time series that were formed from the intensity of each pixel through time. Since the movie was 8 by 8 pixels, that translated to a grouping problem of 65 objects (1 auditory + 64 pixels). We once again performed ICA on this input until convergence. The resulting matrix transformation was threshholded so that all values were either 0 or 1 depending on their magnitude, and is displayed in Figure 2.

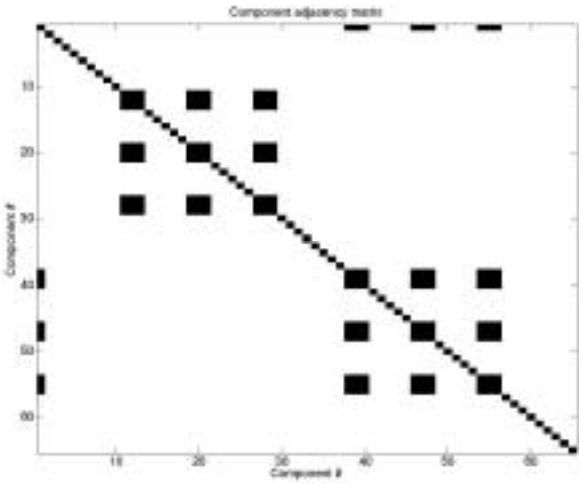

Figure 2 **The grouping matrix for the data in Figure 1. Note how we see the formation of two regions which correspond to the two visual configurations. The matrix is grouping the rasterized picture so that the two clusters we see correspond exactly to the rectangles in the visual stream. Note also the grouping of the first component which is the audio stream, which has been related to the lower right configuration. This means that the grouping operation has recognized the relation of the image and the sound.**

By examination of the results it is clear to see that the proper grouping has taken place. The two visual configurations have been clustered and the audio track appropriately associated to the lower right configuration.

The point of this experiment was to emphasize the versatility of our non-parametric approach. It was possible to use the same algorithm we used to group sinusoidal data to for a different representation, with equally good results. This opens makes an even stronger point about collapsing grouping rules to one principle since we are now able to deal with even more complicated rules which are not clearly defined.

A . 3 Multimodal Scene Analysis

In this section we present an example that employs our scene analysis method for joint audio/visual analysis. The data we used was a movie clip of two hand palms, opening and closing in synchrony with the soundtrack. The soundtrack was the same “Da, da, da” audio clip used in the scene analysis system described in chapter 4. In the movie the right palm was closing whenever the word “Da” occurred, whereas the other one was mostly opening and closing in synchrony with the snare drum (Figure 3).

Figure 3 **The three states of the video scene. One of the two palms closed or both open. The first state occurred in synchrony with the word "Da" in the audio soundtrack, whereas the second one in some synchrony with the snare drum.**

The movie data was ordered in a matrix where the row index indicated frame number, and the contents of that row were the respective frames rasterized. We performed PCA on the movie data to compress it down to three components (the "eigenmovies" of our data), and subsequently applied ICA so as to minimize their statistical dependencies. Upon processing we obtained the overall transformation matrix, $\mathbf{W}$ (the inverse?*** of the product of the PCA and ICA transformations, $\mathbf{W}_I \cdot \mathbf{W}_P$) and the rasterized image components $\mathbf{C}$. The original movie can be approximated by the multiplication $\mathbf{W} \cdot \mathbf{C}$. This analysis step is identical to the process described in chapter 4, only this time it is applied on visual data, rather than magnitude transforms. By visualizing the three components individually we can see that they correspond to meaningful independent regions of activity (Figure 4).

Figure 4 **The energy of the three visual components. The first one concentrates on the region of the left hand, the second on the right hand, and the third on the still elements on the scene. Linear combinations of these three elements can approximately reconstruct each frame of the input movie.**

By visualizing the columns of the $\mathbf{W}$ transformation matrix we can see the temporal activity of each of these components (Figure 5). By doing so we can validate that the three components indeed behave as we hope they should.

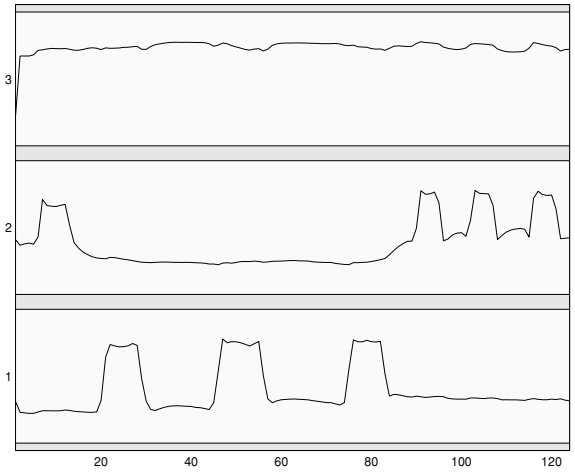

Figure 5 **The temporal weights of the three visual components. We see that the first component has three equidistant distinct instances (corresponding to the proper times in the movie), the second component has four instances which coincide with the "Da" utterances, and the third component has a steady value, which is to be expected since it is the still elements of the movie.**

Upon obtaining this decomposition of the visual stream, we can repeat the same process on the audio portion of the movie by which we derive similar results to the figure 14*** in chapter 4. For reference, the seventh component corresponded to the snare drum, whereas the eighth component was the "Da" utterances.

At this point we should point that the same analysis has occurred on the streams of both modalities. We want to stress this point since it makes a good starting point of dealing with multimodal perception. The statistics of the stimuli, regardless of the modality, point to the objects in the scenes. This is a fact of the way that nature works (independent systems, produce independent signals), and by using only this we can efficiently move cross-modally without having to change representations, of feature detectors.

Now that the scenes are decomposed to a set elementary components, we can use the grouping procedure we introduced in the third chapter to perform grouping between the temporal auditory and visual elements. We constructed a matrix which was composed of both the auditory and the visual temporal weights. In most cases the amount of data will not be sampled at the same rate (in our case the auditory matrix is a 9 components by 5584 spectral windows matrix, whereas the visual one is 3 components by 124 frames matrix). To resolve this problem we resort to interpolation of the smaller matrix. For our data the visual matrix was interpolated in the frame axis to as to have 5584 columns (frames) and the two matrices were stacked row-wise (with the auditory data on top) to form a 12 by 5584 matrix. This matrix contains the temporal characteristics of both auditory and visual stream components. Having that we perform ICA on it, as in chapter 3, and deduct a grouping matrix (Figure 6).

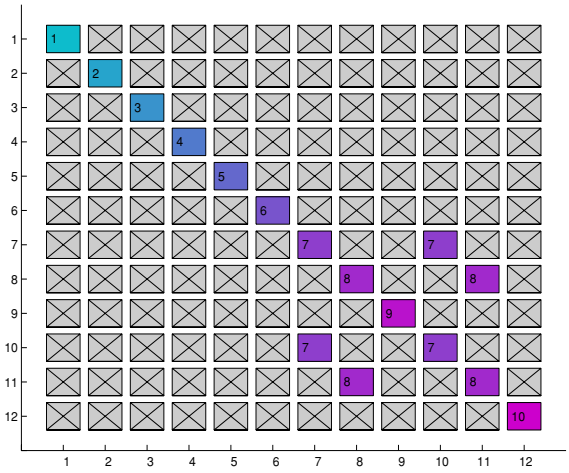

Figure 6 **The grouping matrix across the auditory and the visual data. Each element is obviously grouped with itself, hence the diagonal. The set of non diagonal elements, link the 7th to the 10th component and the 8th to the 11th. The auditory data are the first 9 rows, hence the 7th and 8th correspond to the snare drum and the "Da"'s respectively. The remaining components from 10th to 12th, correspond to the three visual components. Keeping this in mind we can see that the snare drum was grouped to the first visual component, and the "Da"'s to the second component. These visual components correspond to the two palms that were in synchrony with the respective sounds.**

Upon decoding our grouping matrix, we see that the snare drum component was grouped with the right hand component, and the "Da" component with the left hand. We have thus performed cross-modal scene analysis and grouping using one recurring computational method.

A . 4 Conclusions

As these experiments should help understand, the generality of our approach allows very easy formulations of the same operations we have applied on the auditory domain to, not only other domains, but also to multimodal settings. It is our hope that further refinement of our work and the imminent advances in statistics will help us generalize our work to formulate an abstract theory of computational perceptual development, devoid of data dependencies and our own subjective preconceptions.

Bibliography

Referenced Bibliography

The following bibliography is referenced from this thesis.

Ahmed N., T. Natarajan and K.R. Rao (1974) Discrete Cosine Transform in IEEE Transactions on Computers, January 1974. pp 90-93

Allen J. B. and L. R. Rabiner. (1977) A unified approach to short-time Fourier analysis, synthesis. Proc. IEEE, 65(11):1558-1564, November 1977.

Amari, S-I., A. Cichocki, and H.H. Yang. (1996) A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems **8***. MIT Press, Cambridge, MA.

Amari, S-I., S.C. Douglas, A. Cichocki, and H.H. Yang, (1997) Multichannel Blind Deconvolution and Equalization Using the Natural Gradient. In *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, Paris, France, pp. 101-104.

Atick, J.J. and A.N. Redlich. (1990) Towards a theory of early visual processing. *In Neural Computation **2***. pp. 308-320. MIT Press, Cambridge, MA.

Attneave, F. (1954) Informational aspects of visual perception. *Psychological Review **61***, pp. 183-193.

Barlow, H.B. (1959) Sensory mechanisms, the reduction of redundancy, and intelligence. In *National Physical Laboratory Symposium No. 10*, The Mechanization of Thought Processes.

Barlow, H.B. (1961) Possible principles underlying the transformation of sensory messages, In *Sensory Communication*, W. Rosenblith, ed., pp. 217-234. MIT Press, Cambridge, MA.

Barlow, H.B. (1989) Unsupervised learning. In *Neural Computation **1***, pp. 295-311. MIT Press, Cambridge, MA.

Bell, A.J. and T.J. Sejnowski. (1995) An information maximization approach to blind separation and blind deconvolution. In *Neural Computation **7***. pp. 1129-1159. MIT Press, Cambridge, MA.

Bell A.J. and Sejnowski T.J. (1996) Learning the higher-order structure of a natural sound, Network: Computation in Neural Systems, 7

Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. Vision Research, 37(23) 3327-3338

Bregman, A.S. (1990) Auditory Scene Analysis, MIT Press, Cambridge, MA.

Brown, G.J. (1992) Computational auditory scene analysis: A representational approach. Ph.D. dissertation, Uniersity of Sheffield, Computer Science Dept., Sept, 1992.

Brown, J. C. (1991). "Calculation of a constant Q spectral transform." Journal of the Acoustical Society of America 89(1): 425-434.

Cardoso, J-F. (1990) Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *Proceedings ICASSP*, pages 2655-2658.

Cardoso, J-F, and A. Souloumiac (1993) Blind beamforming for non Gaussian signals. In IEE Proceedings-F, 140(6):362-370.

Cardoso, J-F. (1995a) A tetradic decomposition of 4th-order tensors: application to the source separation problem. In M. Moonen and B. de Moor, editors, *Algorithms, architectures and applications*, volume III of SVD and signal processing, pp 375-382.

Cardoso, J-F. (1995b) The invariant approach to source separation In Proc. NOLTA, pages 55-60.

Casey, M. and Westner, W., (2000) Separation of Mixed Audio Sources by Independent Subspace Analysis, in *Proceedings of the International Computer Music Conference*, ICMC, Berlin.

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

Comon, P. (1989) Independent component analysis - a new concept? In *Signal Processing* ***36***, pp. 287-314.

Cooke, M.P. (1991) Modeling auditory processing and organization. Ph.D. thesis, University of Sheffield, Dept. of computer science.

Clarke, R.J. (1981) Relation between the Karhunen Loeve and Cosine Transforms, IEE Proceedings, vol. 128, pt. F, no. 6, pp.359-360.

Dolson, M. (1986). The phase vocoder: A tutorial. Computer Music Journal, 10(4):14-27.

Duda, R.O., R.F. Lyon, and M. Slaney. (1990) Correlograms and the separation of sounds. In *Proceedings Asilomar Conference on Signals, Systems and Computers 1990*.

Ellis, D.P.W. (1992) A perceptual representation of sound. Masters thesis, MIT EECS Department.

Ellis, D.P.W. (1994) A computer implementation of psychoacoustic rules. In 12th International Conference on Pattern Recognition, Jerusalem.

Ellis, D.P.W. (1996) Prediction driven computational auditory scene analysis. Ph.D. thesis. EECS Department.

Ellis, D.P.W., Vercoe, B.L. (1992). A perceptual representation of sound for auditory signal separation, Presented to the 123rd meeting of the Acoustical Society of America, Salt Lake City.

Field, D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells.Journal of the Optical Society of America A, 4 2379-2394.

Field, D.J. (1994) What is the goal of sensory coding? Neural Computation, 6 559-601.

Flanagan, J. L. and Golden, R. M. (1966). Phase vocoder. Bell System Technical Journal, 45:1493-1509.

Gabor D (1946) Theory of Communication, Journal of the IEE 93:429-441

Golub, G.H., and C.F. Van Loan (1983) Matrix Computations, North Oxford Academic, Oxford.

Gould, G. (1966) Invention No. 1 in C Major, BWV 772, Track 1 in Bach: Two- and Three-Part Inventions, released by Sony Classical 1993.

Gray, R.M. (1972) On the asymptotic eigenvalue distribution of Toeplitz matrices, IEEE transactions on information theory, vol. IT-18, pp. 725-730.

Grenander, U. and G. Szegö (1958) Toeplitz forms and their applications, University of Califormia Press, Berkeley, CA.

Grossman, A., Kronland-Martinet, R. and Morlet, J. (1990). Reading and understanding continuous wavelet transforms. In Combes, J. M., Grossman, A. & Tchamitchian, Ph. (Eds), Wavelet – Time-frequency methods and phase space (pp. 2–20). Berlin, Heidelberg, New York: Springer (2nd edition).

Haykin, S. (1994) Neural networks, a comprehensive foundation. McMillan, New York, NY.

Helmholtz, H.von (1925) Physiological Optics. Volume III. The Theory of the Perceptions of vision. Dover Publications, New York, 1962

Herault, J., C. Jutten. (1991) Blind separation of sources, part I, An adaptive algorithm based on neuromimetic architecture. *Signal Processing **24***, pp. 1-10.

Hertz, J., A. Krogh, and R.G. Palmer. (1991) Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, CA.

Holiday, B. (1945) Blue moon, Track 19 CD 1 in Complete-On Verve 1945-59, released by PGD/Verve.

Hopfield, J.J. (1991) Olfactory computation and object perception. *Proceedings of the National Academy of Sciences **88***, 6462-6466.

Hyvärinen, A. (1999) Fast and Robust Fixed-Point Algorithms for Independent Component Analysis IEEE Transactions on Neural Networks 10(3):626-634.

Hyvärinen, A and P. Pajunen, (1999) Nonlinear Independent Component Analysis: Existence and Uniqueness results. Neural Networks 12(3): 429--439.

Hyvärinen, A, P. Hoyer and E. Oja, (2000) Image Denoising by Sparse Code Shrinkage. To appear in S. Haykin and B. Kosko (eds), Intelligent Signal Processing, IEEE Press.

Hyvärinen, A and P.O. Hoyer, (2000) Emergence of phase and shift invariant features by decomposition of natural imagesinto independent feature subspaces. Neural Computation, 12(7):1705-1720.

Ikeda S. and N. Murata, (1999) A method of ICA in time-frequency domain, In Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation, pp.365-371.

Johannesma, P.I.M. (1972) The pre-response stimulus enseble of neurons in the cochlear nucleus. Proc. of the symposium of hearing theory, IPO, Eindhoven, Netherlands.

Kahrs M, and K. Brandenburg. (1998) Applications of digital signal processing to audio and acoustics. Kluwer Academic Publishers Group, Dordrecht, Netherlands.

Koffka, K. (1935) Principles of Gestalt Psychology , Lund Humphries, London.

Lambert, R. H., (1996) Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures. Ph.D. dissertation, University of Southern California, EE dept. May 1996.

Lee, T-W., A.J. Bell and R. Orglmeister. (1997) Blind Source Separation of Real World Signals, Proceedings of IEEE International Conference Neural Networks, June 97, Houston, pp 2129-2135.

Linsker, R. (1986a) From Basic Network Principles to Neural Architecture - emergence of spatial opponent cells. Proceedings of the National Academy of Sciences, USA, 83 7508-7512.

Linsker, R. (1986b) From Basic Network Principles to Neural Architecture - emergence of orientation selective cells. Proceedings of the National Academy of Sciences, USA,, 83 8390-8394.

Linsker, R. (1986c) From Basic Network Principles to Neural Architecture - emergence of orientation columns.Proceedings of the National Academy of Sciences, USA, 83 8779-8783.

Linsker, R. (1988) Self-Organization in a perceptual network. In *Computer **21*** (March), pp. 105-117.

Lyon, R. F. (1996) The All-Pole Gammatone Filter and Auditory Models, Structured transatlantic session: Computational models of signal processing in the auditory system, Forum Acusticum '96, Antwerp, Belgium, April 1-4, 1996.

Makeig S., Bell A.J., Jung T-P., and Sejnowski T.J., (1996) "Independent component analysis of electroencephalographic data." Advances in Neural Information Processing Systems 8, 145-151.

Makhoul, J. (1975). "Linear prediction: A tutorial review." Proc IEEE 63(4): 561-580.

McKay, D. (1996) Maximum Likelihood and Covariant Algorithms for Independent Component Analysis, *Draft paper available at:* `ftp://wol.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz`

McAulay, R. J. and T. F. Quatieri (1986). "Speech analysis/synthesis based on a sinusoidal representation." IEEE ASSP 34(4): 744-754.

Mellinger, D.K. (1991) Event formation and separation in musical sound. Ph.D. thesis, CCRMA, Stanford University.

Moorer, J. (1975) On the segmentation and analysis of continuous musical sound, Report STAN-M-3, Stanford University Department of Music.

Moore, R.F. (1990) Elements of computer music, Prentice Hall, NJ.

Nadal J-P. and N. Parga. (1994) Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. Network: Computation in Neural Systems Volume 5, Number 4 (November 1994), pp 565-581

Nakatami, T., H.G. Okuno, and T. Kawabata. (1994) Auditory stream segregation in auditory scene analysiswith a multi-agent system. In *AIII Conference Proceedings*, 1994.

Oja, E. (1995) The nonlinear PCA learning rule and signal separation - mathematical analysis. Helsinki University of Technology, Laboratory of Computer and Information Science, Report A26.

Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. Nature, 381 607-609.

Oppenheim A.V. and R.W. Schafer. (1989) Discrete-time signal processing, Prentice Hall, Englewood Cliffs NJ.

Parsons, T.W. (1976) Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America,* vol. ***60***, pp. 911-918.

Petersen, T.L., (1980). Acoustic Signal Processing in the Context of a Perceptual Model, Thesis in Computer Science, Univ. of Utah.

Platt, J. and F. Faggin. (1992) In *Advances in Neural Information Processing **4***, J. Moody, S. Hanson, R. Lippmann, eds., pp. 730-737, Morgan-Kaufmann.

Risset, J-C. (1965) Computer study of trumpet tones, in Journal of Acoustical Society of America 38:912 (abstract only)

Redlich, A.N. (1993) Redundancy reduction as a strategy for unsupervised learning. *Neural Computation **5***. pp. 289-304. MIT Press, Cambridge, MA.

Rao K.R., P. Yip, (1990) Discrete cosine transform, algorithms, advantages, applications. Academic Press Inc, San Diego CA.

Roads, C. (1996) The computer music tutorial. MIT press, Cambridge, MA.

Sánchez, V., P. García, A.M. Peinado, J.C Segura and A.J. Rubio (1995) Diagonalizing properties of the Discrete Cosine Transforms, IEEE transactions on Signal Processing Vol. 43 No. 11, pp. 2631-2641.

Serra, X. (1986) A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition, PhD dissertation, Stanford University.

Shannon, C. and W. Weaver (1963) "The Mathematical Theory of Communication" University of Illinois Press.

Slaney, M., D. Naar and R.F. Lyon (1994) Auditory model inversion for sound separation. In Proc. ICASSP94, Adelaide, Australia.

Smaragdis, P. (1997) Information Theoretic Approaches to Source Separation, Masters Thesis, MAS Department, Massachusetts Institute of Technology.

Smaragdis P. (2001) Gestalt and Entropy. Machine Listening Group technical report, *in preparation*.

Stockham, T.G., T.M. Cannon, and R.B. Ingerbretsen. (1975) Blind deconvolution through digital signal processing. In *Proc. IEEE*, vol. ***63***, pp 678-692.

Strangn G. (1999) The discrete cosine transform, SIAM Review 41, pp 135-147.

Trio, (1981) Da da da, I don't love you, you don't love me. Track 8 in album Trio, released by Mobile Suit.

Torkkola. (1996) Blind separation of convolved sources based on information maximization, in proceedings of Neural Networks for Signal Processing 96.

Torkkola, K. (1999) Blind separation for audio signals - Are we there yet?, Proceedings of ICA'99, Assois, France.

Watanabe, S. (1960) Information-theoretical aspects of Inductive and Deductive Inference. I.B.M. Journal of Research and Development, 4 208-231.

Weintraub, M. (1985) A theory and computational model of auditory monaural sound separation. Ph.D. dissertation, Stanford University, EE Dept.

Wiener N. (1949) Extrapolation, Interpolation, and Smoothing of Stationary Time Series, MIT press, MA

Vercoe, B. and D. Cumming. (1988) Connection machine tracking of polyphonic audio. In *Proceedings of International Computer Music Conference 1988*. pp. 211-218.