Learning Distance Functions to Gain Insights into Functional Representations of Complex Objects in the Brain

Thesis submitted for the degree of

"Doctor of Philosophy"

by

Roi Kliper

Submitted to the Senate of the Hebrew University of Jerusalem

March / 2013

This work was carried out under the supervision of:
Prof. Daphna Weinshall and Prof. Israel Nelken

To my wife.

Acknowledgements

This dissertation summarizes a wonderful period I spent at the Hebrew University. I was lucky to study in a great program and work with many gifted people who enriched me in numerous and often unexpected ways.

First and foremost, I would like to thank my advisors Prof. Daphna Weinshall and Prof. Israel Nelken. Daphna's curiosity for understanding fundamental principles, and seeing their implications to a wide array of scientific fields as well as her ability to simultaneously work in several research areas has been a continuous source of inspiration. Daphna has given me the time and encouragement to explore new questions, and to come up with with new ideas eventually helping me to separate the important from the redundant and bring my work to conclusion.

Eli, my second advisor, has been very supportive in our joint quest to apply machine learning techniques to neuroscience. He has helped me to see things through the eyes of an experimentalist, and to understand where novel methods could be most helpful. Eli's strong ability to identify and isolate interesting problems along with his strong analytic skills and a great intuition about the structure of solutions to problems have often proved themselves, although it often took me a while before I could formalize what she envisioned.

My research has taken a while, both my advisors' hands free style allowed to devote my efforts to a wide array of subjects and questions. It has also allowed me to make many mistakes along the way, and learn a whole lot from those mistakes. I must acknowledge and thank both Daphna and Eli for being patient and believing that good ideas take time to flourish. I hope this was the case here.

I would also like to thank the people at the ICNC which has been my academic home at the Hebrew University. The ICNC is an amazing institution and a great place to learn, visit, and spend time in. More then everyone, it was Ruthi Suchi who has given the ICNC community a true feeling of a family, and a visit to the her office was always as worm as a mother's hug.

I would like to acknowledge the contribution of my advising committee: Prof. Naftali Tishby, Prof. Shaul Hochstein, Dr. Gal Chechik who together have helped directing me to the right path of reseacrh. Specifically I would like to thank Prof. Tali Tishby for taking an interest in my research and proposing his point of view on science and life.

Two fellow student collaborators have had a major impact on my work for which I am in debt. Jonathan Rubin who has shared an office with me for the past few years, whose razor sharp intellect and knowledge and impressive analytic capabilities that are always displayed with humility and good spirit, made him always fun to be around and to learn from. Yael Bitterman who has given me endless support along the way, a fountain of knowledge and clear understanding that always comes with a smile and comfort.

I have also had the good fortune to collaborate with Israel Cohen and Michal Irani, on a project that evolved into one of the chapters of this dissertation. A visit to Carl von Ossietzky-University of Oldenburg resulted in a collaboration with Jörn Annemüler and Hendrik Kayser with whom I have also became close friends. Some other friends and colleagues I would like to thank for sharing my academic path are Shaul Druckman, Dana Bar-Niv, Uri Shalit, Yonatan Vaizmann, Shirley Portuguese, Thomas Serre and Tomer Hertz - I was privileged to have you guys around.

To my dear parents, Zmira and Menashe Kliper who has infused me with their intellectual curiosity from an early age and were always there for me when I needed them. Thank you for your love and support. I owe so

much of what I am to you. I love you both with all my heart, and wish all our family nothing but happiness in the years to come.

To my two little children Alma and Ariel, who always gave me a reason to smile and be thankful, who have suffered days of absence and an occasional miss of a night story read and did so with continuous cheerfulness and joy. Finally, to my best friend and beautiful and beloved wife, Yael, with whom I intend to spend the rest of my time on this earth, thank you for your endless patience and support through this long journey. Thanks you for carrying and caring for our family especially during this last stretch. Thanks you for allowing me to do what I enjoy without setting limits and thank you for waking up next to me every morning.

Abstract

This dissertation presents research in the field of computational neuroscience, machine learning and signal processing. We develop and deploy several algorithms that share both the inspiration of the brain as a computing machine and the attempt to understand, replicate and surpass the human brain and its performance. To achieve its impressive abilities the brain has to overcome many challenges, some of these challenges are also presented when researching the nature of brain representation. Key challenges are small sample size, lack of actual supervision, as well as very high dimensional space in both input and output.

It is clear that any empirical measurement of complex systems is bound to be partial. It is partial in more than one sense. First, the experimental setting is inherently limited in time, that is there is only so much that can be achieved in a given experimental session. Second, it is partial in the sense that only a subset of its units may be measured at a given time, resulting in the need to able to extrapolate and generalize form session to session and from unit to unit. Furthermore, it is always the case that only a very small subset of the stimuli space may be presented to the examined system: thus the small subset should be chosen wisely if one is to establish any reasonable conclusion. We are thus faced with the following fundamental question: what can one say about a system, given such a set of partial observations? An immediate extension of this question is how to choose the observations which would be most beneficial for studying a system. These two questions, and their various extensions are the focus of this dissertation.

These conceptual difficulties are not unique to brain research. They are manifested in the study of other complex systems such as biological networks, social graphs, and huge document databases such as the

world wide web. Machine learning is an important field of research which studies fundamental theoretic and algorithmic aspects of extracting rules from empirical measurements of such systems. We make frequent use of machine learning concepts and algorithms throughout this dissertation. One particular family of algorithms which is at the focus of this dissertation is the family of Distance function learning algorithms.

Distance functions are extensively used in various application domains and also serve as an important building block in many types of algorithms. We develop and employ distance function learning algorithm in the neuroscientific experimental setting to functionally characterize single cells with complex (non-linear) functional behavior. Our proposed approach allows the experimenter to automatically incorporate domain and expert specific knowledge from previous and concurrent experimental settings and to better understand the nature of representation of the outside stimulus space. Generalizing this attempt to the characterization of other subsystems in the brain as well as other communication systems is a straight forward task as is also explored in this dissertation.

An alternative and some times a complementary step to distance function learning is the step of feature extraction. A wise feature extraction may render any other learning task trivial. A wise feature extraction, may involve complex processing steps of the original input space, and should be done with care. Attention should be given, so that no processing artifact would be introduced into the data thus creating the problem of data contamination. When developing a feature extraction method in the neurosciences, a common demand is that the extracted features would be biologically feasible and that they would for the least part not contradict existing processing models. Otherwise they carry very little explanatory power with regards to the system they attempt to explain. A second consideration one may want to incorporate into a feature extraction scheme is that the chosen features would translate well into perceptual properties and be in reasonable congruence with observable behavior.

The main contribution of this dissertation are in the development of a computational framework that allows for the functional characterization of a communicating unit. Though at the focus of our research, this unit need not be a single neuron, it could be some subsystem in the brain or even a general communication channel in which case the learning of a representation would sometimes be called decoding. One of the main contribution of this dissertation is a framework to exploit available prior knowledge about both input representation, output representation, the type of noise in the channel and the nature of the channel itself for various inference tasks. This approach is shown to have numerous applications in domains extending far beyond neural coding.

Contents

1

Introduction

"As long as our brain is a mystery, the universe, the reflection of the structure of the brain will also be a mystery."

— Santiago Ramón y Cajal

1.1 Computation and Representation in the Brain

Comprising of a network of $\sim 10^{11}$ neurons and connected by as many as $\sim 10^{14}$ synapses the human brain is the most complex and sophisticated system we know of. It is also, perhaps, the most mysterious one: it is widely accepted that the enigma of the brain is the most challenging intellectual endeavor of the 21'st century, "The Century of the Brain".

The brain is the system whose simultaneous activity results in all human functions, from locomotion to emotion, from complex reasoning to lust and addiction. It is the prime source of our remarkable achievements, creative capabilities, emotional world, and belief systems. It is responsible for coordinating and controlling the body's activities, for representing, processing and interpreting information from the environment, for initiating actions, and reacting to the surroundings, and for our amazing ability to perform complex tasks rapidly and accurately. It is also the centre of memory, thought and is where consciousness emerges. But far and foremost, the brain is a learning machine. From the moment one is born - some say even sooner - one begins experiencing and learning from the world around us.

Understanding how the brain works carries many promises of replicating and improving on its amazing abilities, but also with regards to identifying and fixing that, that has gone wrong. The brain is at the core of many devastating diseases

such as Alzheimer's, Parkinson's, schizophrenia and autism. One must understand the basic principles that govern the functioning of the brain if one is to recognize a malfunction and develop systematic methods to repair it when needed.

Understanding the brain involves several extremely difficult paradigmatic and experimental challenges collectively gathered under the discipline of neuroscience. Neuroscience is the scientific study of the nervous system. The scope of neuroscience is broad and includes different approaches used to study the molecular, cellular, developmental, structural, functional, evolutionary, computational, and medical aspects of the nervous system. The techniques used by neuroscientists are also broad and diverse and are expanding enormously, from molecular and cellular studies of individual nerve cells to imaging of sensory and motor tasks in the brain, from electrophysiological recordings to psychophysiolocal experimenting.

Attempts to understand how the brain works, and what specific subsystems of it are responsible for, are as old as neuroscience itself. Studies of the brain became more sophisticated after the invention of the microscope and the development of a staining procedure by Camillo Golgi during the late 1890s. His technique was used by Santiago Ramón y Cajal and led to the formation of the *neuron doctrine* [1] , the hypothesis that the basic functional unit of the brain is the neuron. although a lot has been gained since(e.g. [2]) , this hypothesis remains as relevant today as it was a century ago. It is therefore not a surprise that the task of *functional characterization of a single neuron* stands as a key challenge in neuroscience - it is also at the focus of this dissertation.

1.2 Computational Neurosciece

Recent theoretical advances in neuroscience have been aided by studies in machine learning, signal processing and neural networks and by the general development of computational neuroscience as an independent field of study. Computational neuroscience focuses on characterizing and modeling the brain function, or the function of its subsystems, in terms of information processing properties. It implies that we ought to be able to exploit the conceptual and technical resources of computational research to help find explanations of how neural structures achieve their

effects, what functions are executed by neural structures, and be able to describe the nature of representations that are manifested by states and activation patterns of the nervous system. It is an interdisciplinary science that links the diverse fields of neuroscience, cognitive science, biology and psychology with electrical engineering, computer science, mathematics and physics. One particular point of interaction is that of *Machine learning*. Another focal point of interaction is that of *signal processing*. This dissertation explores several junctions of neuroscience, signal processing and machine learning.

1.2.1 Machine Learning and Computational Neuroscience

Machine learning is the discipline of study that deals with the construction and study of systems that can learn from data and experience. The core of machine learning deals with methods of representation, abstraction and generalization. Representation of data instances and of functions evaluated on these instances is a key part of all machine learning systems and a basic step that, done properly, simplifies any machine learning problem. Abstraction and generalization refer to the desired properties that the system will be parsimonious that is that it would do more than 'memorize' the instances it has seen, and be able to extract some concrete or abstract rule that will results in good performance on unseen data instances; the conditions under which this can be guaranteed are a key goal and hence a key object of study in the subfield of computational learning theory. There is a wide variety of machine learning tasks, algorithms and successful applications. Optical character recognition, speech recognition and collaborative filtering are classic engineering examples of machine learning. Very few of those algorithms have matured enough to compete with the abilities of the human brain. In many senses *Machine learning* is an attempt to imitate the way our brain learns form experience: it is the science of getting computers to act without being explicitly programmed. More specifically, *Machine learning* allows one to study patterns of activation, forms of representation, rules of regression and classification which may be thought of as abstractions of complex decision processes.

The neurosciences present some of the steepest challenges to machine learning, among those are diverse settings, inherent noise in both process and measurement,

as well as problems of diversity of data, scarcity of data and reproducibility. The wide diversity of problems and challenges posed by neuroscience, has spurred the introduction, development and application of many different approaches as well as concrete algorithms, some were applied with impressive success in handling specific aspects and challenges that arise in this setting.

While challenges in the field diversify in both problem settings and approaches, certain commonalities can be identified. A common concern in the neuroscientific experimental setting is the problem of small sample size, particularly when compared to the possibly very high-dimensional input structure - this is since each data point is usually gathered at a high cost in both time and money as well as due to constraints posed by physiological considerations. To avoid poor generalization, inference must therefore make considerable use of experts and domain knowledge from physics, biology, psychology, neurophysiology and anatomy, in order to regularize the acquired solution. A related concern is the inherent lack of "true" feedback, that limits the use of classic supervised algorithms.

Another concern is that the signal of interest typically occupy a relatively small subspace of the input representation space, the rest being made up of processes occurring in parallel that may be of much larger magnitude. In finding generalizable solutions, one usually has to contend with a high degree of variability, both between individuals and across time, leading to problems of covariate shift and non-stationarity.

In all cases, even the high-dimensional raw input is a vastly simplified reflection of the underlying processes and structures. This representation is typically chosen accommodate the relevant task to the level of idiosyncrasy in the representation choice. New ways of measuring what is being represented, and new ways of transforming the data, are therefore still waiting to be found, leading to feature representations that are more relevant, less noisy, or more transferrable between experimental sessions, subjects and experimental tasks.

In computational neuroscience and in particular in this dissertation, machine learning serves in a double role on the one hand it is a source of robust techniques used for the modeling and understanding of the various impressive operations the brain performs, and on the other hand it delves on the brain's impressive abili-

ties across its different faculties as a major source of inspiration for new exciting technological developments.

1.2.2 Signal Processing and Data Analysis in Neuroscience

Signal processing is concerned with the modeling, detection, identification and utilization of patterns and structures in a signal process. Applications of signal processing methods include audio hi-fi, digital TV and radio, cellular mobile phones, voice recognition, vision, radar, sonar, geophysical exploration, medical electronics, and in general any system that is concerned with the communication, representation or processing of information. One such system, and an extremely efficient one is the human brain. In a fact the basic abstraction of any subsystem of the brain is often that of a communication channel communicating information to and from our peripheral motor and sensory organs. This abstraction also holds when considering a single cell, or a subnetwork in the brain.

Signal processing theory plays an increasingly central role in the way we understand the flow of information in the brain, its processing, encoding, and decoding as well as in the development of modern telecommunication and information processing systems. It has a wide range of applications in multimedia technology, audio-visual signal processing, system identification and characterization, cellular mobile communication, adaptive network management, radar systems, pattern analysis, medical signal processing, financial data forecasting, decision making systems, etc.

Challenges in signal processing are diverse and numerous. They include among others signal detection, signal coding and decoding, compression, information extraction, channel equalization, noise reduction, signal reconstruction, signal enhancement, spatial processing and source localization, source separation as well as higher level processing such as inference, pattern classification/recognition. Most of these challenges are handled by the human brain in an almost seamless manner.

1.2.2.1 Signals and the Processing Stages

A signal can be defined as the variation of a quantity by which information is conveyed regarding the state, the characteristics, the composition, the trajectory,

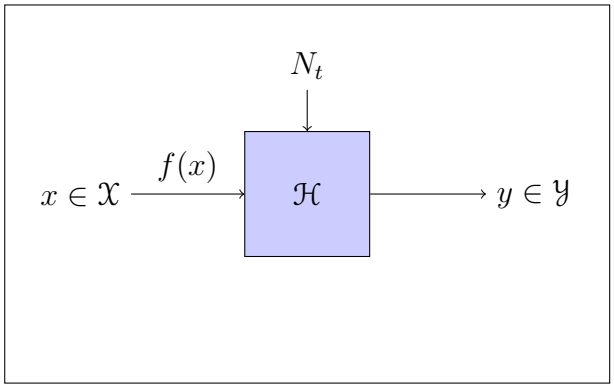

Figure 1.1: A channel abstraction: Data in the input space $\mathcal{X}$ is represented in some feature space $f(x)$ and passes through a possibly noisy communication channel $\mathcal{H}$ to reach an output representation space $\mathcal{Y}$.

the course of action or the intention of the signal source. *A signal is a means to convey information.* However, a received signal may convey more than what was originally intended to be transmitted. It may convey information about the nature of the channel, the environment from which it was encoded (e.g. location) or any other piece of information. The information conveyed in a signal may be used by humans or machines for communication, forecasting, decision-making, control, exploration etc. Figure 1.1 illustrates an information space ($\mathcal{X}$) followed by a possibly noisy system for signaling the information ($f(x)$): a communication channel for propagation of the signal ($\mathcal{H}$) from the transmitter to the receiver. This communication channel abstraction is a key abstraction of which we explore different parts through this work.

In the general case, there is a mapping operation H that maps the information $x \in \mathcal{X}$ or selected features of it $f(x)$ to the signal $Y = \mathcal{H}(f(x))$ that carries the information, this mapping function may be denoted as $\mathcal{H}[\cdot]$ and expressed as

$$y(t) = \mathcal{H}[f(x)(t)] \tag{1.1}$$

This channel may be contaminated with noise resulting in the expression in Eq. 1.2.

$$y(t) = \mathcal{H}[f(x)(t), N_t] \tag{1.2}$$

For example, in human speech communication, the voice-generating mechanism pro-

vides a means for the talker to map each word into a distinct acoustic speech signal that can propagate to the listener. To communicate a word w, the talker generates an acoustic signal realization of the word; this acoustic signal $x(t)$ may be contaminated by ambient noise and/or distorted by a communication channel, or impaired by the speaking abnormalities of the talker, and received as the noisy and distorted signal $y(t)$. In addition to conveying the spoken word, the acoustic speech signal has the capacity to convey information about the identity of the speaker as well as on the speaking characteristic, accent, emotional state of the talker, the location of the speaker, the type of channel the signal has gone through etc. The listener may be able to extract these information pieces by processing the signal $y(t)$, ultimately using these pieces of information to classify the signal according to the relevant task.

1.2.2.2 Signal Classification

Signal classification is used in signal detection, pattern recognition and decision-making systems. The aim is to design a minimum-error system for labeling a signal with one of a number of likely classes of signal. To design a classifier; a set of models are trained for the classes of signals that are of interest in the application. The simplest form that the models can assume is a bank, or code book, of examples, each representing the prototype for one class of signals. A more complete model for each class of signals takes the form of a probability distribution function. Many approaches for training systems to behave intelligently are based on components that learn automatically from previous experience. Among the various existing algorithms, one of the most recognized family of algorithms is termed Support Vector Machines (SVMs). Various flavors of SVMs are used across this dissertation both as an ultimate test of a representation scheme or as a learning infrastructure. SVMs have an attractive interpretation as a (multi-layer) neural network, thus when developing an SVM solution one can almost immediately link it to a possible physiological realization and is thus an attractive tool for studying an inherently representing system like the brain.

1.3 Learning Invariant Spaces In Experimental Neuroscience Setting

Transformations of representations are an immanent part of our cognitive processes. Objects and sensations with distinct physical properties are analyzed, synthesized and transformed through the brain's different pathways, taking detailed forms of membrane potentials and action potentials that eventually compose larger mental representations. The question of what exactly is being coded and represented in different brain regions has attracted great interest and some remarkable work, but while the picture may seem clear in some domains, most remain only partially understood.

A case of special interest is the functional characterization of sensory neurons. One of its main goals is to characterize dimensions in stimulus space to which the neurons are highly sensitive (causing large gradients in the neural responses), or alternatively dimensions in stimulus space to which the neuronal response is invariant (defining iso-response manifolds). This challenge is especially pronounced when trying to learn the representation of complex visual/auditory objects in higher brain areas. In higher areas, cells, presumably, no longer represent simple features and tend to represent more complex, non-trivial semantic classes.

Let $\mathcal{S}$ and $\mathcal{R}$ be the spaces of possible stimuli, and of possible responses accordingly. The dominant approach to the functional characterization of sensory neurons attempts to learn the input-output relations $f : \mathcal{S} \to \mathcal{R}$ of a neuron. The "goodness" of such a function is usually measured by its ability to "explain" the previously seen data and by its ability to predict the response of the neuron to novel stimuli. Prediction is often estimated from a set of known stimuli-response pairs of a given neuron, by modeling some linear or quasilinear filter over the stimuli space. One typical method of building such predictors uses linear models and their second order variants, in order to approximate the function that is assumed to generate the response [3, 4, 5]. However, these models may fail when responses are highly non-linear or when the smoothness of the response as dependent on the stimulus space is lost [4, 6] - a process which is hypothesized to occur, when moving from earlier processing stations to higher ones e.g. as one moves along the visual pathways from V4 to

the ITC and further to the PFC [7]. In addition, filtering models tend to be limited in their ability to use information from one neuron, for functional characterization of another - that is, to generalize from cell to cell.

The family of algorithms termed *Distance function learning algorithms* provides a robust computational tool for the study of relations between objects that reside in a given space. Taking a distance function learning approach means avoiding learning the input-output relation f and focusing on the relation within a given space. Motivated by the view that different neurons impose different partitions of stimulus space which are not necessarily simply related to the simple feature structure of the stimuli [8], we attempt to learn the specific "geometric" structure induced by the neuron on the stimulus space by learning a *cell specific distance function*. In other words, we try to learn the non-linear partition of stimulus space as induced by the neuronal responses.

Specifically, we characterize a neuron by learning a pairwise distance function over the stimulus space that is consistent with the similarities between the responses to different stimuli. A distance function is a function defined over pairs of data-points $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, which assigns a real (and possibly bounded) valued number to any pair of points from the input space $\{s_i, s_j\} \in \mathcal{S}$. The assigned number measures the distance between pairs of points, which reflects the similarities between them. Intuitively, a good distance function would capture the desired structure and assign small distance values to pairs of stimuli that elicit a similar neuronal response, and large values to pairs of stimuli that elicit different neuronal responses. Knowledge of the structure is in itself valuable, as it can be used to understand the kind of information coding that a single cell performs on the stimulus space. Interestingly, it could also be used for prediction, e.g. by using some variant of k-Nearest Neighbors (k-NN) with the learnt distance function, or by applying avariant of SVM in case the distance function is also a kernel.

The approach we have explored and implemented, offers several advantages: first, it allows us to aggregate information from a number of neurons and reach a good hypothesis even when the number of known stimuli responses per neuron is small (and even zero), which is a typical concern in the domain of neuronal characterization. Second, unlike most functional characterizations that are limited to linear or weakly non-linear models, distance learning can approximate functions that

are highly non-linear and implicitly embed complex feature spaces. Metaphorically speaking, learning a cell specific distance function allows the investigator to bypass the question of "how the cell speaks", or "how many spikes are fired to a given stimuli?", and attempt, instead, to touch upon the more fundamental question of what exactly the cell is saying, or "what partition does the cell induce on stimulus space?".

In the following we start by describing the distance function learning paradigm and its different flavors. We continue by developing the idea of invariant spaces and stressing its importance in the neuroscientific experimental setting. We then explain why the use of distance function learning for functional characterization of single cells is the right approach for this problem and draw other closely related use case. We then shift our focus to the area of feature extraction and draw some parallels between feature extraction and distance function learning.

1.4 Distance function learning

While *distance function learning* is a somewhat new area of research, the concept of a *distance function* is well known and is widely used in various applications and for various computational tasks. Recent years have seen a lot of interest in distance function learning algorithms (e.g. [9, 10, 11] and also [12] and references therein for a comprehensive review. Distance function learning algorithms aim at automatically incorporating domain specific knowledge and/or side information into the applied distance function. In the general setting of such an algorithm, side information (typically in the form of equivalence constraints) is used to learn a pairwise function that adequately captures the structure of the space and the relations between the data-points.

Unlike the classical learning scenario, where one attempts to learn some function $f : \mathcal{S} \rightarrow \mathcal{R}$ that approximates the relation between a given input and its resulting output using a training sample $\mathcal{S} \times \mathcal{R} = \{(s_1, r_1), (s_2, r_2), \ldots (s_N, r_N)\}$, distance functions provide information about the similarity of pairs of points - essentially capturing relations within the input data-points themselves $D : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ rather than an input-output relation. In some cases capturing the relations between data-points

can provide information which cannot be easily extracted from directly estimating input-output relations. Such cases occur when the input-output function is highly complex, multi staged, or just difficult to estimate. In other cases, the structure of the space is what one is interested in - rather than the transformation.

In the case of single cell characterization, both incentives for using distance function learning appeal: on the one hand estimating the actual input-output relation is hard and poses many physiological and technical limitations. On the other hand, we argue, information about the partition of space may in many cases be much more interesting than predicting the actual response. Ideally, learning an adequate distance function may render the task of response prediction redundant.

1.4.1 Distance function learning from equivalence constraints

Recent years have seen a growing interest in various learning algorithms which make use of side-information in the form of equivalence constraints, either positive ('a is similar to b') or negative ('a is dissimilar from b') or both e.g. [10, 13, 14, 15, 16, 17]. Since a distance function is a function defined over pairs of points, equivalence constraints are a natural form of supervision in this context, as they are binary labels in the product space $\mathcal{S} \times \mathcal{S}$, the space of all pairs of points. To use equivalence constraints one labels pairs of points which are negatively constrained by -1, and pairs of points which are positively constrained by 1, making an adaptation of classification algorithm a natural approach. In contrast to explicit labels that are usually provided by a human instructor, in some scenarios, equivalence constraints may be extracted with minimal effort or even automatically. Two examples of such scenarios are *(1)* Temporal or spatial continuity (or discontinuity), such as the one observed in multiple surveillance cameras where consecutive frames may be positively constrained and simultaneous frames from different cameras may be negatively constrained. *(2)* Generalized Relevance Feedback: e.g. when multiple subject may label subsets of data but possibly not with the same labels (thus allowing distribution of a laborious task).

It is, therefore, not surprising that most of the work done on learning distance functions centered on algorithms that learn distance functions using equivalence constraints. On a different aspect, a major focus in the field of distance function learning has been given to the learning of Mahalanobis metrics which have a nice interpretation and are in essence a linear transformation of the input space. Recently several authors have suggested algorithms that explore wider spaces of non-linear distance functions [18, 19, 20]. In Chapter 2, we follow this path and formalize the problem of single cell characterization as a problem of distance function learning from equivalence constraints.

As typically done in learning algorithms, a distance function learning algorithm had to balance between a fitting term and a regularization term. The fitting made sure that the a model was trained such that positively constrained points were modeled as closer, and negatively constrained points were modeled as far from each other. A successfully trained model would have immediately resulted in improved classification results, which is, typically, also the ultimate test for a distance function learning algorithm.

In a common distance function learning setting, positive and negative constraints were typically derived from a dataset where actual class labels existed. In a sense, the supervision given to the algorithm was a degenerate form of the available information. Most algorithms developed in the field e.g. [9, 10, 11, 14, 21] among others, mentioned very elaborate settings where distance function learning from equivalence constraints would be the right approach to take and where equivalence constraints were, indeed, the only form of available supervision, but then came back to exemplify their strength on classification datasets, typically going back to the famous UCI repository [22].

This state of affairs was dictated, at least in part, by the availability of data, and the type of problems which are studied by machine learning researchers. Scenarios (and datasets) where, actual learning of the distance function is the relevant task are hard to come by, while those of classification are abundant. Thus distance function learning was apprehended as a tool to improve classification rather than a goal in itself. Our work and in particular chapters 2-4, challenges this approach in putting the focus on distances themselves as well as in using the developed learning algorithms in settings where semi-supervision is the only supervision.

1.4.2 Distance function learning as a regression problem

Linear regression is an approach to modeling the relationship between a scalar dependent variable y and one or more explanatory variables denoted $\vec{x}$. When performing regression one typically learns a function $f : \mathcal{X} \to \mathcal{Y}$. This path of learning may seem orthogonal to the path taken by researchers of distance function learning algorithms however, in Chapter 3 we show that a distance function learning problem may be thought of as a regression problem in the space of pairs of points. Furthermore, we show that a natural representation of a pair of points when learning a Mahalanobis distance function is the outer-product of the points (or their difference vector). This apprehension of the problem brings into focus a variety of algorithms designed for regression which is maybe the oldest machine learning problem.

Regression problems and their solution depend on the existence of sample points $\{x, y\}$. In the distance function regression setting, it may be unclear what is the dependent variable and how do we obtain its samples. We maintain that in some settings, and specifically (but not only) in the neuro-scientific experimental setting the dependent variable (the regressed distance) can be readily defined.

Numerous extensions and variant of linear regression have been developed including various formulations of nonlinear regression [23, 24, 25] and see [26] and reference therein. Generally speaking the various regression flavors differ in the way the loss function is defined and in the way the hypotheses space is regularized. In [24] a large margin approach to regression was offered as an extension to support vector machines (SVMs) and was termed support vector regression (SVR). Similarly to the model produced by SVMs for classification that depends directly on a subset of the training data (Support Vectors), the model produced by SVR depends only on a subset of the training data. This is thanks to a loss function formulation that produces a sparse solution and ignores the loss of any training data that is within an ϵ-insensitive margin. In Chapter 3 we adopt the regression formulation of SVR to regress the *Gram matrix* given as supervision and learn a cell specific distance function.

1.5 The Question of Invariant Spaces in Neurophysiology

The common view of sensory systems considers neurons as feature detectors arranged in an anatomical and functional hierarchy. The best example of such an understanding is evident in the way current understanding models the visual system, but similar views exist for the auditory pathway as well as other processing streams.

Visual processing in cortex is classically modeled as a hierarchy of increasingly sophisticated representations, naturally extending the model of simple to complex cells due to Hubel and Wiesel [27]. In this view of the visual system, information from simple feature detectors in the retina converges at the level of the primary visual cortex (V1) to represent elaborate features of direction and orientation. Simple feature representations are then being gradually neglected in favor of distributed complex object representations at the level of the inferotemporal cortex (ITC) [28, 29].

This hierarchical view continues to dominate models characterizing both the *ventral pathway*, associated with object recognition and form representation, and the *dorsal pathway* associated with determining objects' position in space. Several models have been proposed to account for the impressive human ability to locate, identify, recognize and track objects in a visual scenery e.g. [30, 31, 32]. However, little quantitative modeling has been done to explore the biological feasibility of this class of models to explain aspects of higher-level visual processing such as object recognition. A recent study [33] showed that it was possible to reliably extract position and scale invariant object category information from a small population of neurons (~ 200) in ITC. A second study pointed out a possible functional shift that might be taking place between the closely connected ITC and prefrontal cortex (PFC) regions [7, 34]. Others [35] point out the manifestation of non-linear representation already in the V4 regions and offered a model that may account for it. Understanding the actual manifestation of this shift and characterizing the role of single cells in such schemes as well as gathering information to support such models of representation has proved to be a major challenge in visual neuroscience.

In addition, physiological evidence accumulated over the past decade remains controversial, particularly because models of object recognition in the cortex have been mostly applied to tasks involving the recognition of isolated objects presented on blank backgrounds. Ultimately models of the visual system have to prove themselves in real world object recognition tasks, such as face detection in cluttered scenes, a standard computer vision benchmark task. In Chapter 2 we use a distance function learning algorithm in an effort to characterize the role of single cells in the ITC and PFC and contribute to the understanding of the role these cells play in the visual processing stream.

A similar view to that of the visual system, exist in the auditory domain. While certain basic features, such as frequency sensitivity, are shared by auditory neurons of all stations, several studies suggest that dramatic changes in stimulus representation occur as information flows from Inferior colliculus (IC) through Medial Geniculate Body (MGB) to Auditory Cortex (A1)[36]. Perhaps the first difference between IC neurons and neurons in higher stations appears at the basic characterization of an auditory neuron - its Frequency Response Area (FRA). The FRA maps the neurons response properties in a two dimensional frequency level space. Two fundamental characteristics of neurons in IC, MGB and A1 are their best frequency (BF; the frequency at which they have their lowest threshold) and their bandwidth. IC neurons typically exhibit narrow FRAs, while FRAs characteristically become wider as one advances to the thalamus and on to the cortex. Another difference is the temporal resolution of the neurons: IC neurons typically follow repetitive stimuli up to about 100 Hz while typical A1 neurons follow this type of stimuli up to about 10 Hz and MGB neurons are intermediate. Such characterization suggest that neurons in A1 (and MGB) are less sensitive to low level features of auditory stimuli than neurons in the preceding station, the IC. In Chapter 3 we take adistance regression approach and try to characterize and decipher the different role neurons in the IC and A1 play.

In both neural settings, the role of a single cell in such complex tasks is a subject of great debate. It places a very difficult challenge on the experimental setting: that of revealing the invariant response space of the cell. While researchers have acknowledged the need to account for properties of invariance and specificity, the

prediction requirement of a single cell model is still, to a great degree, phrased in terms of response predictions.

The accepted approach to functional single cell characterization has focused on characterizing the Spatio-Temporal (Spectro-Temporal) Receptive Fields (STRFs) of a neuron: computing the average stimulus that elicits spikes, yielding the so-called ST receptive field (STRF) which is (assuming linearity) a function that maps time-varying visual (auditory) inputs to neural responses [3, 6, 37, 38]. Classically, STRFs have been used to implement linear and sometimes second-order models of neural responses [5]; ultimately, evaluating the success of such a model by measuring the model's response prediction against the real responses. This approach has frequently been used to study both the visual system and the auditory system and has been proven quite efficient for the functional characterization of early stages of both systems; however, using these models to characterize neurons in higher stages has not been proven as efficient.

Unfortunately for filtering models, the stimulus average may be a coarse descriptor of the complex features that excite neurons. For example, the average of all stimuli that excite a visual, face sensitive, neuron would be an elongated lump, rather than a face [39], losing on the way distinctive spatio-temporal features, e.g. high-order sensitivities which could allow the neurons to identify a class of complex objects even when the spatio-temporal structure in that class varies. Similarly STRF models in the auditory domain, fail to account for effects of context and the coding of abstract features in the stimuli [8, 36, 40].

These and other findings seem to suggest that highly complex representations of the environment cannot be accounted for by the mere use of simple linear or near-linear models, nor can its success be measured purely by its ability to predict neuronal response, and that a new approach to the functional characterization of these cells may be in order. A first step in this direction was taken by [41], introducing the learning of distance functions into this domain and showing some preliminary and promising results. In the following two chapters, we pursue this research direction further, trying to identify and characterize complex cells in the visual system and auditory system. We propose a repertoire of algorithms and demonstrate their applicability to the problem at hand. In Chapter 2 we continue on the mission taken by Weiner et al. [41] and propose an improved learning algorithm that learns

a single cell's nonlinear distance function from equivalence constraints. In Chapter 3 we pose the problem of distance function learning as a regression problem. We develop and use a distance function learning algorithm that learns a single cell's invariant spaces directly from the responses. Unlike the traditional approach to distance learning, that rely on equivalence constraints for supervision, this algorithm uses the actual distances between responses for supervision. Our results show a dramatic improvement over previously published results and may help to make this approach an accepted one in experimental neuroscience setting.

1.6 Data Representations, Feature Selection and Feature Weighting

Many classification and clustering algorithms operate over a dataset S in which each datapoint is represented by some feature vector where each dimension within the vector is assumed to represent some measurement or a specific feature (e.g. color, size) of the input instance. Finding a good representation of the input data is known as the data representation problem and is a fundamental step in all machine learning approaches. This problem has been extensively studied over the last several decades, and numerous representation schemes have been suggested for various input domains. For example, if one wants to represent an acoustic object, one can represent the waveform using a vector of its amplitude measurement across time[1], use some dimensionality reduction method such as Principal Component Analysis (PCA) over the original waveform vectors, represent the acoustic object using its power spectral density (psd) function (thus neglecting time), or represent the object using a more elaborate set set of coefficients (e.g. wavelet coefficients, or the output of some filterbank). These are a few of the numerous vector representations which have been suggested and explored for images. It should be clear that for each different input domain, different representations can be defined, with varying degrees of complexity and incorporation of domain knowledge.

[1]Note that even such a crude representation would already results in a dramatic simplification of the original space due to sampling and quantization

The reason the representation problem has received enormous attention is that by improving the representation of the data, one can significantly enhance the performance of the clustering (or classification) algorithm over which it is applied. In fact, finding the optimal (or ideal) representation can make the employment of fancy learning algorithms redundant altogether.

Suppose for example that our task is to classify a set of input images into two categories - 'cats' and 'dogs'. If we had a good representation of these images - i.e. one that would represent cat images in a very different way than dog images , the classification task would become very easy, and a very simple classifier such as a linear separator would easily provide perfect classification performance. However, if our representation mapped 'cat' images and 'dog' images in a similar, possibly overlapping way, the classification task would become much harder, and in order to obtain good performance we would probably need to use more sophisticated classifiers to solve the problem. Note, however, that a good representation for the 'dog'-'cat' challenge would not necessarily be any good for other challenges over the same data - e.g. distinguishing a male pet form a female one - this task would probably require a completely different set of features.

In general the success of many learning algorithms is often strongly dependent on various assumptions which they make about the data representation - i.e. about the feature space in which the input data objects are represented: Classes are assumed to be convex, or at least continuous, and at least some of the features are expected to be relevant for predicting the class label of the input instances. However, in most cases, the data representations used in various application domains are far from ideal, and some of these assumptions do not hold. Put differently, in most cases, the data representation is rather 'weak' and is usually based on some standard off-the-shelf domain specific features.

One way to improve such 'weak' representations is to apply some pre-processing transformation $\mathcal{F}$ to the input data-points. The transformation attempts to map the data-points into a feature space in which the data are 'better' represented. In many cases, this pre-processing transformation can also be used to reduce the dimensionality of the input space, by selecting 'relevant' dimensions, or by eliminating 'non-relevant' dimensions. This problem is also known as the feature selection problem, for which various algorithms have been suggested (see [42] for a review of

traditional methods and [43] for a more recent summary). Another approach to the problem involves the design of a carefully chosen set of features (sometimes hand-crafted) - A set that would allow a learning algorithm a convenient way to measure a relevant dimension.

The representation problem is closely related to the distance learning problem. Since, in many cases learning a distance function can be thought of as a transformation of the input-data, furthermore, many learning algorithms are distance based (i.e. they only require as input the distances between data-points), selecting a 'better' distance function will improve the performance of these algorithms. In other words, there is an analogy here - finding a better distance function for a distance based algorithm is just like finding a better representation for a feature-based algorithm.

In some cases, the connections between data representation and distance functions can be made more explicit. Consider for example distance-based clustering algorithms such as linkage algorithms or a codebook based retrieval system. It is well known that the performance of these algorithms depends to a great extent on the quality of the distance function used. Therefore, one way to improve the performance of these algorithms is to improve the quality of the distance function used to compute their input distance matrix. However, note that if we had an ideal representation of our data (i.e. one that would map every data-point into a set of well separated feature vectors, each of which would represent a single class within our data), any simple (and non-trivial) distance function we chose would provide perfect clustering results. Another more formal example of the connection between distance functions and data representation can be seen when considering the relation of the *Mahalanobis distance metric* to linear transformations. Since a Mahalanobis distance matrix A is a symmetric Positive Semi Definite (PSD) matrix, it can be decomposed using singular value decomposition (SVD) as follows:

$$A = U\Sigma U^T = (U\Sigma^{-\frac{1}{2}}2)(\Sigma^{-\frac{1}{2}}U^T) = B^T B \tag{1.3}$$

where U is an orthonormal matrix $\left(UU^T = I\right)$ where each column is an eigen-vector of the matrix A, and Σ is a diagonal matrix which holds the singular values of the matrix A (that are equal to the squared eigen-values of the matrix), and $B = A^{\frac{1}{2}}$ It

can therefore clearly be seen that using the Mahalanobis distance metric defined by A is equivalent to applying a linear transformation of the data using the matrix B and then measuring the Euclidean distance between the transformed datapoints[1]:

$$\sqrt{z^T A z} = \sqrt{z^T \left(B^T B\right) z} = \sqrt{\left(z^T B^T\right)\left(Bz\right)} = \sqrt{\left(Bz\right)^T \left(Bz\right)} \tag{1.4}$$

Therefore, finding an optimal Mahalanobis metric A is equivalent to finding an optimal linear transformation B and then using the Euclidean distance metric over the transformed space. When the Mahalanobis metric considered is of low rank, it is equivalent to a linear projection of the data. Linear projections have been widely used in various application domains. One well known supervised algorithm for learning a linear projection of the data is Linear Discriminant Analysis (LDA, also known as FLD), which was originally suggested by Fisher [44]. Another well known linear transformation which is an unsupervised one is the Mahalnobis distance proposed by non other than Mahalanobis himself [45].

However, in the case of a general distance function D there is no principled way to find the (possibly nonlinear) transformation which can be used to represent the data in some feature space where the Euclidean distance between the data-points would provide the same pairwise distances. Furthermore, as a distance function is not necessarily a metric (that is some or all of the metric properties may be violated), it is safe to say, that in some cases such a transformation may not exist. In many cases defining the 'right' feature space for a task is an effort worth spending.

One possible approach to this problem, which has been explored in the literature, is based on embedding. The classical problem of embedding is to take a set of datapoints for which pairwise distances are provided, and to embed them into some low-dimensional Euclidean space, in which the pairwise Euclidean distances between the datapoints would be minimally distorted. Several algorithms have been suggested for this problem, including Local Linear Embedding (LLE) [46], Isomap [47].

Distance learning algorithms are a principled way for going in the opposite direction - they are provided with a set of data-points that lie in some vectorial feature

[1] z should be defined here as the difference vector $z = x - y$

space, and with some additional side-information on the distances (or more commonly relations) between some pairs of data-points, and attempt to learn a distance function in which the distances between pairs of data-points will reflect the side information provided. Learning distance functions can therefore be seen as an alternative to finding a good representation of the data, using some standard representation of the data and some additional side-information. Therefore, distance learning can also be seen as an alternative approach to feature selection, or to finding a strong data representation.

The problem of learning distance functions is also closely related to the problem of feature weighting. In this setting, each data object is represented using a set of features (or a feature vector), and the objective is to learn a set of weights over these input features. The distances between a pair of objects are defined as the weighted sum of their corresponding feature vectors. In most cases, feature weighting methods are used in the context of K-Nearest-Neighbor (KNN) classifiers (see [48] for a review). Feature weighting can be seen as a generalization of feature selection. More importantly, it can be seen as a special case of distance function learning, in which a diagonal Mahalanobis metric is learnt. Therefore recent works in this area are sometimes described as "feature selection" algorithms, and sometimes described as âĂIJdistance-learningâĂİ algorithms.

1.7 Summary and Outline

With over a century of neuroscience research, the brain remains as enigmatic as ever. While understanding as to some of the brain's impressive mechanisms is accumulating rapidly, basic questions, as well as basic subsystems remain in the dark. Paradigmatic difficulties as well as technical ones are making it difficult to isolate a functional unit from its surroundings and study its computational properties; it is even more difficult to establish a reasonable interpretation as to its behavior.

Any subsystem can be abstracted as a simple communication channel with a defined (noisy) input space possibly non-linear operator and an output space. Many of the attempt to functionally characterize a subsystem insisted on estimating the unknown operator. Distance function learning algorithms offer an opportunity to study

the behavior of a communication channel (e.g. a neuron) in terms of its changing space representation. Taking this approach means accepting a 'black box' approach to the characterized subsystem, and focusing instead on the implied semantic of the changing representation between the input and the output.

This approach simplifies the problem in several ways and is based on combining several ingredients. First, we use natural and complex stimuli, reflecting our belief that interesting properties of high level processing (presumably taking place in the auditory cortex) can be revealed in the responses to such stimuli. Such properties however cannot be discovered using standard linear methods. Secondly, electrophysiological recordings from a series of auditory processing stations allows us to compare the representations of these complex stimuli, and the way they change along the processing hierarchy, thus reflecting the computational processes that the system applies. Our goal is to identify design principles that underlie the changes in these representations. Thirdly, we use advanced machine learning technique to quantify how sensory neurons interact with the environment to represent stimuli, we develop machine learning technique to study what the cells represent. Our belief is that the combination of these components can reveal novel evidence about the principles that underly neural coding.

The paradigmatic approach to single cell characterization that was introduced in Chapter 1 is implemented and experimented with and in Chapter 2 and in Chapter 3. Two algorithms for this task are discussed suggesting the use of two different types of supervision for the learning of a single cell's functional characterization. In Chapter 2 we show how one can characterize single cells in the visual cortex by learning a cell specific distance function from equivalence constraints We then develop in Chapter 3 a distance regression algorithm that is used to characterize complex cells in the auditory pathway. In Chapter 4 we present a novel algorithm for signal reconstruction in the presence of unreliable channel properties. This setting is shown to be a good candidate for the deployment of our distance regression algorithm, which in this context is used to better the retrieval of code-words in a code-book based speech enhancement algorithm. Observing the amazing abilities of the brain, one cannot but ask how do such abilities come about? Understanding, how the brain accomplishes a task, or what can go wrong in an attempt to achieve a task is a recurring theme across this disertation. While not a direct application of distance function

learning, in Chapter 5 we explore the use of physiologically inspired features, as well as a simulation of the filtering of the human ear for the task of monaural and binaural localization. Finally in Chapter 6 we employ advanced signal processing and machine learning techniques in an attempt to translate a long standing notion in the psychiatric community into an automatic diagnostic tool. Specifically we try to quantify perceptual concepts of speech prosody into signal processing measures which in turn highlight changes related to an underlying mental state.

2

Characterizing Visual Neurons Through Distance Learning

"If the human brain was so simple that we could understand it, we would be so simple that we couldn't."
— Emerson Pugh

2.1 Learning neural representation with equivalence constraints

In this chapter we follow the path charted by Weiner et al [51], and attempt to characterize the invariant spaces of a single cell by learning a cell specific distance function. Specifically, we employ the Kernel-Boost algorithm [19] to study a cell specific distance function of neurons in the visual pathway. Kernel-Boost is a robust variant of the Dist-Boost algorithm [18] employed in [51] to study the non-linear wrapping of acoustic stimulus space in neurons in the auditory pathway. The Kernel-Boost algorithm uses a boosting scheme to learn a non-linear distance function from equivalence constraints and was shown to be extremely efficient for learning from small samples - a common consideration in the neuroscientific experimental setting. The resulting distance function features the desired property that the learnt distance function is also a kernel which makes a family of kernel based algorithms available for use along with the learned kernel.

2.1.1 Problem formulation

Our approach is based on the idea of learning a distance function over the stimuli space, using side-information extracted from the response space. The initial

data consists of stimulus-response paired representations. To generate the side-information, we use a similarity measure over pairs of points in the response space. These are used in turn to generate equivalence constraints on pairs of stimuli: two stimuli are related by a positive equivalence constraint if their paired responses are highly similar; they are related by a negative equivalence constraint if their paired responses are highly dissimilar.

Next, this side information is used to train an algorithm that learns a distance function between pairs of stimuli points, thus capturing implicitly the structure of the stimuli space as induced by the cell. In our framework, the cell becomes a teacher, specifying similarities between stimuli using its own language of action potentials. These similarities are then used to learn a cell-specific distance function over the space of all possible stimuli. This learned distance function should reveal what exactly is represented by the changes in the response of the specific cell.

An outline of the computational task is as follows:

Input: Given a set of stimuli-response pairs $\{s_i, r_i\}_{i=1}^{N}$

1. Represent the responses and stimuli in their own 'natural' feature space, along with a predefined similarity measure in the responses space.
2. Use the responses to extract equivalence constraints on stimuli, as described above.
3. Learn a distance function over the stimuli space $D(s_i, s_j) \rightarrow \mathbb{R}$ using these constraints.
4. Use the generated distance function to understand and predict the nature of stimuli space.

In the remainder of this section we present the details of our suggested scheme and how it is used for the characterization of visual neurons.

2.1.2 Neural computational setting implementation

The data was collected by Freedman et al [7, 34] and consisted of stimuli-response pairs of data recorded in the ITC and PFC of macaque monkeys. The stimuli used

consisted of a continuous set of cat and dog stimuli constructed from six prototypes with a three-dimensional morphing system. The stimuli were generated by morphing different amounts of the prototypes (see Figure 2.1).

This allowed to continuously vary the stimulus shape, and precisely define a category boundary. The category of a stimulus was defined by whichever category contributed more (50%) to a given morph. The behavioral paradigm required monkeys to release a lever if two stimuli (separated by a 1 sec delay) were from the same category (a category match), see [7, 34] for details. The continuos nature of the stimuli in combination with this behavioral task presumably broke smoothness of response over stimuli space making the task difficult for simple models.

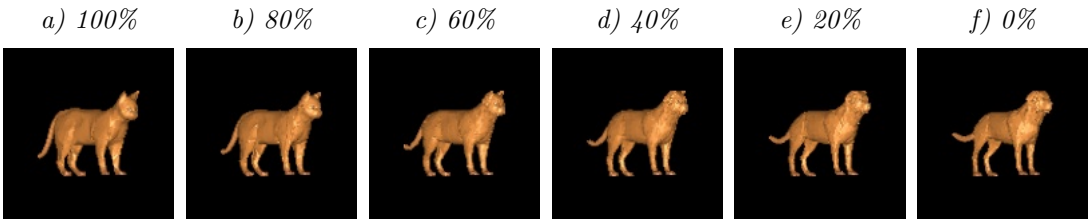

Figure 2.1: A morph line between 'Cat I' (left) and 'Dog I' (right). Three Prototypes of dogs and three prototypes of cats were morphed in six levels of inter-species. 100% (Figure 2.1a)„ 80% (Figure 2.1b) ,60% (Figure 2.1c), 40% (Figure 2.1d), 20% (Figure 2.1e), 0% (Figure 2.1f) and 4 levels of within species morphing 100% 60% 40% 0% yielding 42 Cat-Dog morphs and 12 within species morphs.

2.1.2.1 Data representation

As input for our learning algorithm we used the first 20 principal components of a gray scale representation of the images. The neuronal response for each stimulus was represented as a vector containing in each of its entries the spike rate of one out of multiple stimulus presentations (trials), where the number of trials per stimulus varied in the range $7 - 13$.

2.1.3 Obtaining equivalence constraints from neural data

For the sake of simplicity, the distance between responses was measured using the absolute value of *Cohen's d* statistic. Cohen's d statistic may be seen as a distance

measure defined over vectors of measures of different length. It is defined as the difference between two means divided by the pooled standard deviation. This is very similar to taking the F-statistic of a 1-way ANOVA between vectors of responses over multiple trials. By choosing Cohen's d length as our response distance measure, we implicitly assume that responses of the cells to different stimuli are normally distributed across trial repetitions to same stimuli and are generally of equal variance (homoscedasticity assumption), but may be characterized by different means. Calculating Cohen's d for each two pairs of responses, we end up with a distance matrix over all pairs of responses.

We then used the complete linkage algorithm to cluster the data into 8 clusters. We found that choosing this number of clusters does not over constrain the problem not positively nor negatively. All of the points in each cluster were marked as similar to one another, thus providing *positive equivalence constraints*. Negative constraints were determined to exist between points in the 4 furthest clusters. This left most of the pairs neither positively nor negatively constrained.

2.1.4 Training a distance function

We took the scheme described above and implemented it using the Kernel-Boost distance learning algorithm described in [19] and illustrated in Figure 2.2 [1]. Kernel-Boost is a variant of the Dist-Boost algorithm [52] which was used in [41] and showed promising results. Kernel-Boost is a semi-supervised distance learning algorithm that learns distance functions using unlabeled data-points and equivalence constraints. While the Dist-Boost algorithm has been shown to enhance clustering and retrieval performance, it was never used in the context of classification mainly due to the fact that the learnt distance function is not a kernel (and is not necessarily metric). Therefore it cannot be used by the large variety of kernel based classifiers that have shown to be highly successful in fully labeled classification scenarios. Kernel-Boost alleviates this problem by modifying the weak learner of Dist-Boost to produce a 'weak' kernel function. The 'weak' kernel has an intuitive probabilistic interpretation - the similarity between two points is defined by the probability that

[1] In our comparative study Kernel-Boost performed extremely well, especially when given only a small amount of data.

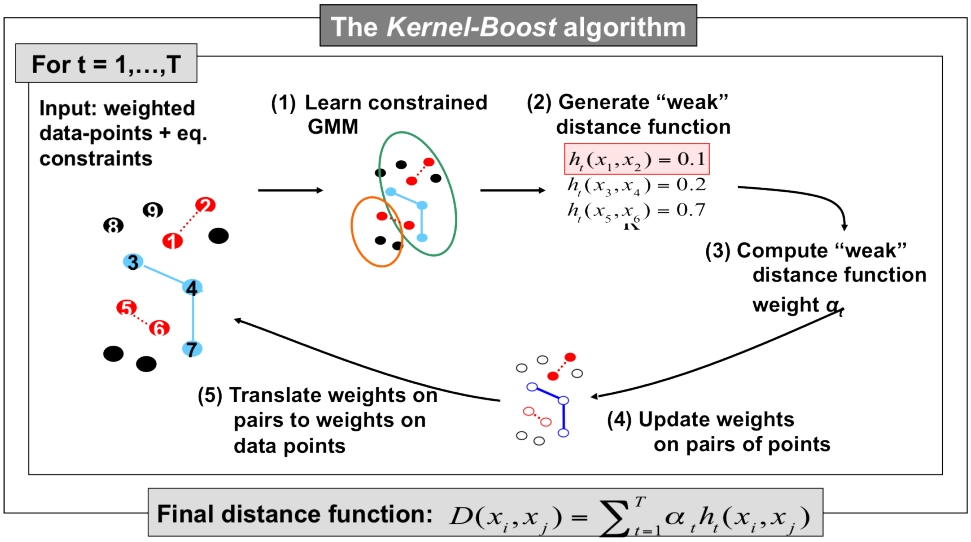

Figure 2.2: The Kernel-Boost algorithm: - A schematic view of the Kernel-Boost algorithm: *1*) Get an input data points and equivalence constraints. *2*) Train a constrained GMM model. *3*) Estimate a weak kernel from the probabilistic model trained using the cGMM algorithm. *4*) Adjust weights *5*) reiterate.

they both belong to the same Gaussian component within the constrained Gaussian Mixture Model (cGMM) learned by the weak learner.

An additional important advantage of Kernel-Boost over Dist-Boost is that it is not restricted to model each class at each round using a single Gaussian model, therefore removing the assumption that classes are convex. This restriction is dealt with by using an adaptive label dissolve mechanism, which splits the labeled points from each class into several local subsets. An important inherited feature of Kernel Boost is that it is semi-supervised, and can naturally accommodate unlabeled data in the learning process.

Specifically, for each neuron, a subset of all pairs of stimuli was selected such that the responses of the two stimuli in a pair were either very similar or very dissimilar as described above. The distance function was trained in a leave one out manner using a cross validation scheme to fit these constraints. That is for each left out stimuli a 5-fold cross validation scheme was conducted to properly select the algorithm parameters and then a model was trained on the entirety of the training data and tested against the test stimuli. We tested whether the resulting distance functions generalized to predict the distances between the responses of a test stimulus and all

trained stimuli.

Evaluation

We used a number of ways to evaluate the quality of the learned distance function. First, we evaluated the learned distance function by calculating the rank correlation (Spearman) between the learned distances and the actual distances as measured over the cell responses using *Cohen's d*. High correlation suggests that the learning algorithm managed to capture the inter-point relations in the stimulus-space. We compared our results to the rank correlation results obtained when measuring the distances between the gray scale representation of the images and correlating them to the actual results. In a second evaluation step, the distances were used to cluster the stimulus data, which was then compared with the clustering induced by the cell responses; the agreement between the two clusterings was measured using the Rand index. Finally, the distance function was used along with a k-NN classifier to generate predictions for novel samples in a cross validation scheme.

2.2 Results and evaluation

We narrowed our analysis to neurons which displayed some stimulus selectivity (not necessarily category selectivity). Such selectivity was established by performing 1-way ANOVA with each of the 54 samples as grouping factors at $p < 0.01$. This analysis resulted in 162 ITC neurons and 61 PFC neurons. However, while this procedure filters out non discriminative cells it still keeps cells that are selective to a very small number of stimuli (1-3 stimuli). Since this analysis is a demonstration of a new technique we limit it to the ten most selective cells in each setting [1]. We start the analysis by measuring the success of the learner to fit and generalize the distances as defined by the responses (Section 2.2.1). We then continue to show how this knowledge can be used for response prediction (Section 2.2.2) and stimulus classification (Section 6.2.5).

[1] All reported results pertain to this subset of the data.

2.2.1 Fitting Power and Generalization

As a first step we examined the ability of our algorithm to fit the distances as induced by the cell responses. We evaluated this by measuring the mean Spearman rank correlation between the distances computed by our distance learning algorithm and those induced by the cell as measured by the 'actual' distances over the responses: we use our distance function to calculate the distances between the stimuli and check wether close pairs of stimuli (as defined by a cell) are indeed close as measured by our cell specific distance function. The results as seen in Figure 2.3 (left most bar) show dramatic improvement after learning. This suggests that some significant learning of stimuli space was indeed performed by the algorithm.

After establishing the improvement in fitting power across both data sets, we turned to evaluate the generalization properties of the algorithm. We tested three scenarios: Leave-One-Out (LOO) where at each simulation one stimulus was left out of the training set and used for testing, Leave-Five-Out (LFO) and Leave-Ten-Out (LTO) where a random sample of five and ten (in accordance) stimuli were left out of the training set for later test (20 repetitions). For each stimulus that was tested in one of this manners, we measured its distance to all other stimuli using the learned distance function. We then computed the rank order correlation coefficient between the learned distances in the stimulus domain, and the similarity of the corresponding response vectors. This procedure yielded a single correlation coefficient for each of the simulations using the stimuli which were left out. To measure performance, we took the average of the correlation coefficient over all runs for each cell.

The results seen in Figure 2.3 show persistent improvement in test correlation in all scenarios, as compared to baseline the correlations measured over the naive (gray scale) distances. These results highlight the fact that the algorithm can generalize well, and some aspects of topology of the stimulus space have indeed been captured. Test correlation scores are generally reduced as the size of the sample used for training is reduced, but the reduction is minor. Finally, individual cells also displayed very strong correlation between training performance and test performance.

Interestingly, we have observed that some of the stimuli's distances were more predictable than others' when left out of the training set. Presumably good results

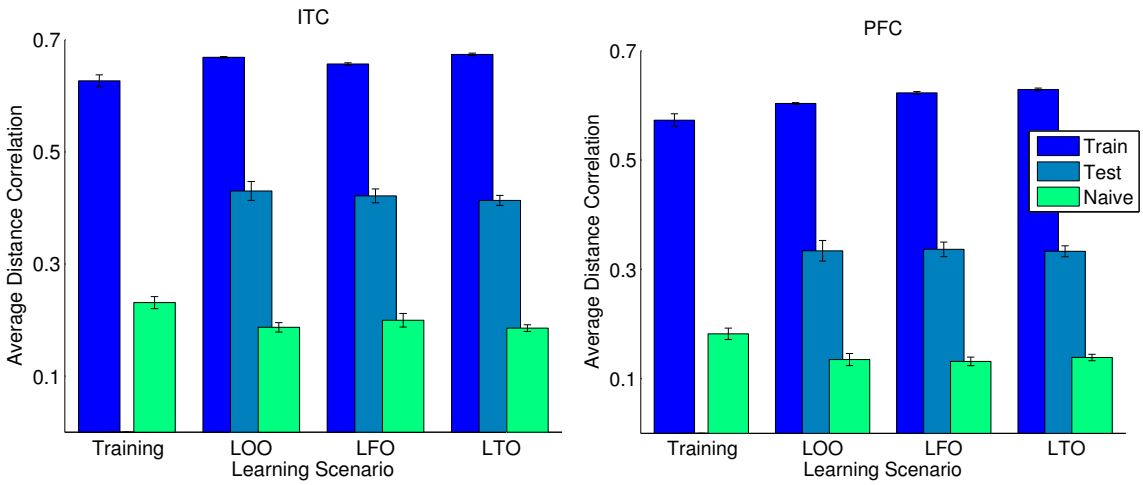

Figure 2.3: Distance correlation score. 1) Trained vs. Test vs. Naive (distances between images) performance: The algorithm achieved a dramatic increase in average rank correlation, suggesting that some aspect of the structure of stimulus space was indeed captured. Correlation between train and test scores were in the order of 0.9 (not shown).

will ultimately depend on the quality of sampling of stimulus space, a task to which this algorithm can be helpful by pointing out interesting areas in the stimulus space.

2.2.2 Generating response predictions

Having proved the generalization properties of the learning algorithm, we attempt to use the newly learnt distance function to generate a response prediction to novel stimuli. We do this by matching to each novel stimulus its 1^{st}-Nearest Neighbor (1-NN) from the training sample using our learnt distance function. We then predict that the cell would respond to the novel stimulus in a similar way to its response to this matched 1-NN stimulus. We note that since the stimulus space is rather restricted in our experimental setting, responses tend to be relatively smooth locally, in the sense that similar (gray level) stimuli tend to give rise to similar responses. Thus, even the naive 1-NN predictor, based on the inter-stimuli distance in the original grey-level representation, is expected to perform well in most cases. We compared our prediction scheme to both the linear regression predictor and a naive prediction based on the 1-NN as computed using the gray levels of the stimuli. It is

worth noting that the regression method is optimized for response prediction while our method is optimized for distance learning rather than response prediction and is, thus, expected to do well on this task. A trend that favors our method over both competitors can be observed in Figure 2.4. The success in prediction, and our advantage over the alternative predictors, is highly dependent on the size of the training sample - it increases with sample size, which is consistent with the increase in test distance correlations demonstrated earlier. Specifically, we do better or equal to other predictors in 64%, 66% and 61% for ITC cells, and 68%, 65% and 61% for PFC cells, with $53, 49, 44$ training samples respectively.

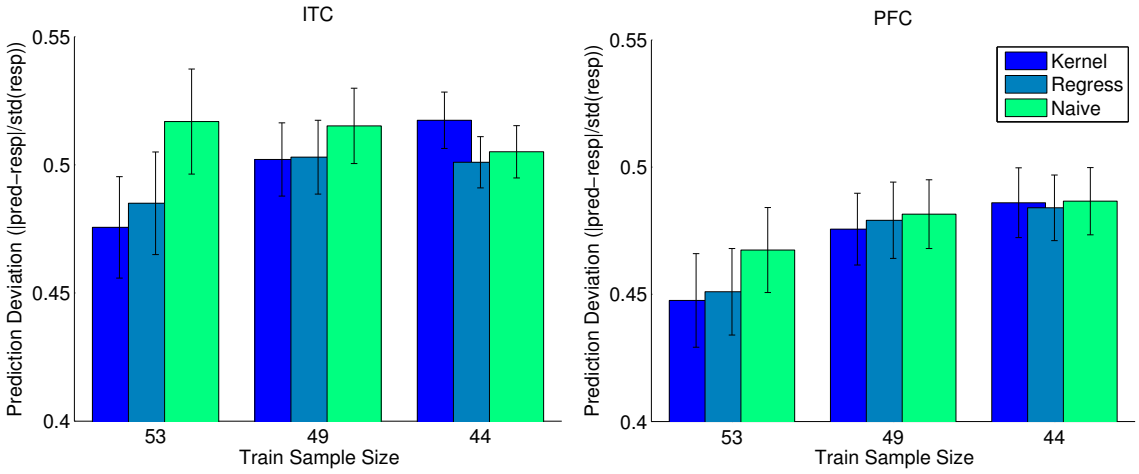

Figure 2.4: Prediction advantage: shown above is a measure for the success of response prediction. The quality of prediction success is measured by the normalized absolute distance to actual response: $\frac{|prediction-response|}{std(response)}$. In other words we measure how many standard deviation our prediction is from the actual response. Performance is compared against naive KNN predictor (Naive) and against a linear regressor (regress). This advantage of our algorithm becomes evident as sample size increases.

2.2.3 Peeking into the semantics: Inferring iso-response manifolds

A single cell can divide the world it sees to a few iso-response sets. This division is reflected by distinct response patterns to each of these sets. The application

and performance of the learning algorithm in this setting is limited, we believe, by the richness of representation of the response space. We believe that richer representations of the response space would result in better learning of the stimulus space. In our setting, for the sake of simplicity and due to the rather simple nature of the response patterns (rate code), we assume that a cell induces a binary partition on the stimulus space and thus defines two iso-response sets.

The learnt distance function can be thought of as somewhat analogous to a transformation of the feature space. We now ask - in the transformed space, is the division to iso-response sets more evident? does this transformation make the input space easier to analyze and understand? To answer this question, we visualize the data using 2-D embeddings. We use classic MDS [53] to compare three embeddings: one based on the true distances in response space (presumably the goal embedding), a second based on the Euclidean distance in the original grey-level features space, and a third based on the learnt distances. Specifically, we show two examples of cells in a leave-ten-out setting in Figure 2.5. Clearly the embedding based on the learnt distances is more similar to the goal embedding than the original one. Moreover, these response-induced partitions are evident in stimulus space after distance transformation, but hardly evident in the original feature space. This result was found for other cells as well. The point to appreciate here is the fact that while creating the cell embedding (left) is illustrative in figuring out the semantics of the cell, it is dependent completely on the actual recording experiment. Our distance based embedding, however is not and can be used on any number of (possibly unused) data-points.

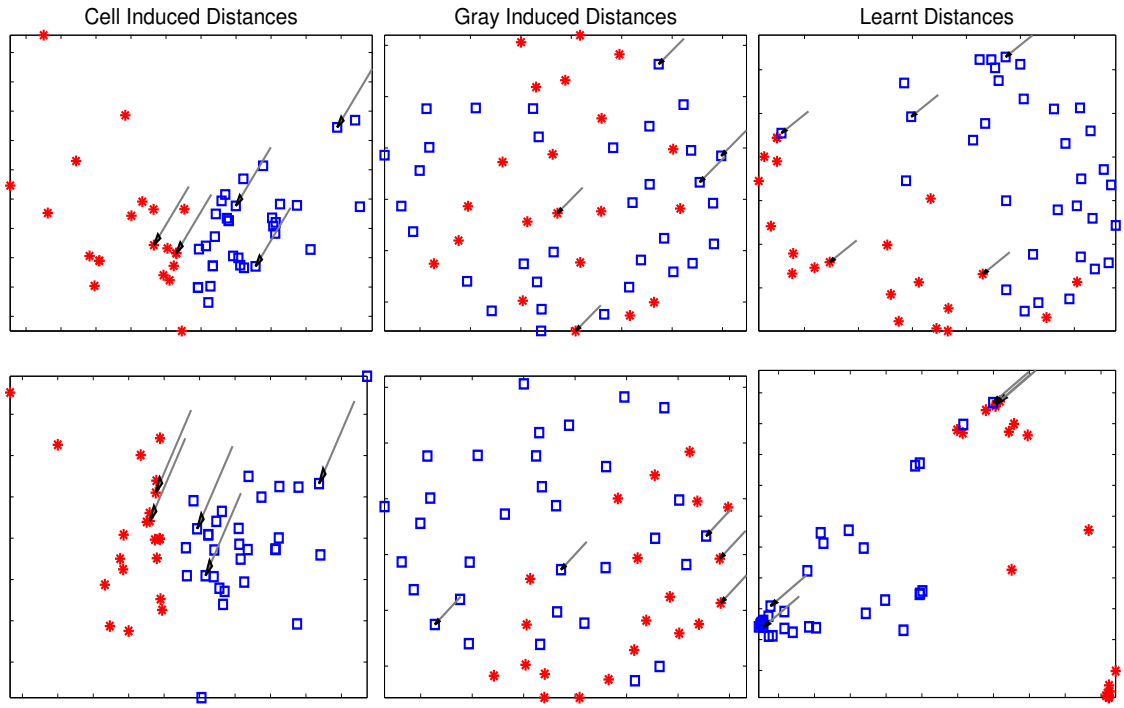

Figure 2.5: Visualizing classification effect: Two cells were selected to display the effect of learning on the feature space topology, which is visualized via the embedding of points in a 2-D space. For each cell we show embeddings based on response distances (left column), original gray levels values (middle column), and learnt distance function (right column). In each case we used classic MDS to embed the data points in 2-D. Points were partitioned into two iso-response sets according to the response they elicit: red circle indicate points that elicited high response while blue square points that elicited low response. The goal of the learner was to be able to replicate the partition induced by the cell (left column). Arrows mark points that were left out during the training phase.

3

Characterizing Auditory Neurons Through Distance Regression

"It is essential to understand our brains in some detail if we are to assess correctly our place in this vast and complicated universe we see all around us."
— Francis Crick, What Mad Pursuit"

In this chapter we introduce a new method for characterizing complex cells by learning cell specific distance function over pairs of data-points. Contrary to previous attempts (including our own) that have used supervision in the form of equivalence constraints to learn a cell specific distance function, we use the distances themselves as a supervision and pose the problem as that of a regression problem over pairs of points. Specifically, we take a *spectral regularization* approach and pose the problem as a regression problem over the outer product of pairs of points. This formulation which has a natural justification, turns out to be a large margin problem. We, thus, use Support Vector Regression (SVR) to solve this computationally hard regression problem and show that a simple and reasonable assumption about the nature of the kernel results in a significant simplification of the solution. We experiment with our newly proposed method in an attempt to functionally characterize cells in the Inferior Colliculus (IC) and Primary auditory cortex (A1) and show very promising results.

3.1 Regressing Distances

In the following, we start by a short description of the problem, we then formulate and solve a translation invariant formulation of the distance regression optimization problem. We discuss in details the linear case and offer a kernelized version. We

then formulate a second more general formulation of the distance regression problem and discuss its solution.

3.1.1 Posing the problem

In the machine learning community the Mahalanobis distance between points $x, y \in \mathbb{R}^n$ is defined as:[1]

$$d(x,y) = (x-y)^T A(x-y) \tag{3.1}$$

Where A is an $n \times n$ *Positive Semi Definite* (PSD) matrix. We note that

$$\begin{aligned}
(x-y)^T A\,(x-y) &= tr\left((x-y)^T A\,(x-y)\right) \\
&= tr\left(A \cdot (x-y)\,(x-y)^T\right) \\
&= \sum_{i,j} A_{i,j}\left((x-y)\,(x-y)^T\right)_{i,j} \\
&= \left\langle A, (x-y)\,(x-y)^T\right\rangle_{\mathbb{R}^{n^2}}
\end{aligned} \tag{3.2}$$

Where $\langle\cdot,\cdot\rangle_{\mathbb{R}^{n^2}}$ represents the standard inner product in $\mathbb{R}^{n^2}$. From Eq. 3.2 we see that Mahalanobis distance can be viewed as a linear function in the space of outer products $xy^T \in \mathbb{R}^{n^2}$. Thus, the problem of learning a Mahalanobis distance function from a training set $\{x_{i_1}, x_{i2}, \hat{d}(x_{i_1}, x_{i_2})\}_{i=1}^M$ is naturally posed as a linear regression problem, where the instances are matrices of the form $(x_{i_1} - x_{i_2})(x_{i_1} - x_{i_2})^T$, the target variables are $\hat{d}(x_{i_1}, x_{i_2})$, and the function learned is the Mahalanobis distance matrix A.

The form of the matrix A learned via this process will depend on the regression method used, but typically it will be a linear combination of the training instances $(x_{i_1} - x_{i_2})(x_{i_1} - x_{i_2})^T$. Such a solution is symmetric, but not necessarily PSD. However, if all the coefficients of the linear combination are positive, the result will be PSD. This point is crucial if one wishes to interpret the learned distance as a linear transformation over the entirety of the space. In practice we note that the learning scenario pushes the matrix towards the PSD cone, and in the worst case we have a matrix that behaves "nicely" in the locality of the training set.

[1] We discard here of the square root, which is a monotonic transformation.

3.1.2 Support Vector Regression

The regression scheme we will look into is that of support vector regression (SVR), described in [54]. Support vector machines are easily extended to handle regression problems while at the same time preserving the property of sparseness. In simple linear regression, we minimize a regularized error function given by:

$$\frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2 + \frac{\lambda}{2}\|w\|_2^2 \tag{3.3}$$

To obtain sparse solutions, the quadratic error function is replaced by an ϵ-insensitive error function [55] which gives zero error if the absolute difference between the prediction $y(x)$ and the target t is less than ϵ where $\epsilon \geq 0$. A simple example of an ϵ-insensitive error function, having linear cost associated with errors outside the insensitive region, is given by Eq. 3.4 and illustrated in Figure 3.1.

$$E_\epsilon(t) = \begin{cases} 0, & \text{if } |t| \leq \epsilon \\ |t| - \epsilon, & \text{otherwise} \end{cases} \tag{3.4}$$

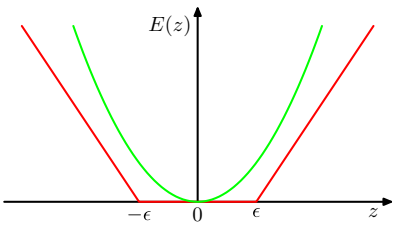

Figure 3.1: A linear ϵ-insensitive error function - The error increases linearly with the distance beyond the insensitive region (red). Square loss (green) given as reference.

SVR tries to find a continuous function such that the maximum number of data points lie within the epsilon-wide insensitivity tube (see Figure 3.2). Predictions falling within epsilon distance of the true target value are not interpreted as errors. In SVR one minimizes the so-called ϵ-insensitive loss, where ϵ is a user defined parameter. The epsilon factor is a regularization setting for SVR which balances the margin of error with model robustness to achieve the best generalization to new data. When ϵ is set to 0, SVR degenerates to an L_1 regression problem. Typically, epsilon is data dependent and is determined through a cross validation scheme over the given sample data.

Specifically, we minimize the following regularized error function

$$\mathcal{L}(A) = \frac{1}{2}\|A\|_2^2 + C\sum_{i=1}^{M} E_\epsilon\left(d_A\left(x_{i_1}, x_{i_2}\right), d_i\right) \tag{3.5}$$

Where the regularization term is the Frobenious norm of A (also known as *Hilbert Schmidt norm*) - $\|A\|_F^2 = \sum_{i,j} a_{i,j}^2$ which can also be stated as $tr\left(AA^T\right)$. It is equivalent to the SVM/SVR regularization term $\|w\|^2$, which, in matrix context, can be viewed as the sum of squares of A's singular values - $\sum \sigma_i^2$; and where the fitting term is the sum of the individual loss incurred by a sample $d_A(x, y)$ as given by Eq. 3.1. By convention the (inverse) regularization parameter denoted C appears next to the error term. Let $z_i = x_{i_1} - x_{i_2}$ and $d_i = \hat{d}(x_{i_1}, x_{i_2})$, we restate the regularized error function given in Eq. 3.5 as follows:

$$\mathcal{L}(A) = \frac{1}{2}tr\left(AA^T\right) + C\sum_{i=1}^{M} E_\epsilon\left({z_i}^T A z_i, d_i\right)\Big) \tag{3.6}$$

As typically done, we re-express the optimization problem by introducing slack variables. For each pair of data point x_{i_1}, x_{i_2} we now need two slack variables $\xi_i \geq 0$ and $\hat{\xi}_i \geq 0$ each corresponding to a slack at one side of the insensitivity tube. Introducing the slack variables allows points to lie outside the tube provided the slack variables are nonzero. The error function for the SVR formulation can then be rewritten as

$$\frac{1}{2}tr\left(AA^T\right) + C\sum_{i=1}^{M}\left(\xi_i + \hat{\xi}_i\right) \tag{3.7}$$

which must be minimized subject to inequality constraints over the slack variables $\xi_i \geq 0$ and $\hat{\xi}_i \geq 0$,

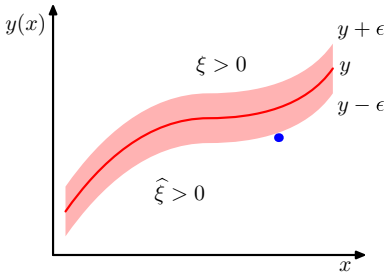

Figure 3.2: ϵ insensitive tube: - The formulation using an ϵ-insensitive loss creates an insensitivity tube where point are not penalized for deviating from the regression line.

as well as the following constraints

$$d_i \leq d_A\left(x_{i_1}, x_{i_2}\right) + \epsilon + \xi_i \tag{3.8}$$

$$d_i \geq d_A\left(x_{i_1}, x_{i_2}\right) - \epsilon - \hat{\xi}_i \tag{3.9}$$

The minimization is achieved by introducing *Lagrange multipliers* $\mu_i, \hat{\mu}_i, \nu_i, \hat{\nu}_i$ resulting with an optimization Langrangian as follows:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2} tr\left(AA^T\right) + C\sum_{i=1}^{M}\left(\xi_i + \hat{\xi}_i\right) - \sum_{i=1}^{M}\left(\mu_i\xi_i + \hat{\mu}_i\hat{\xi}_i\right) \\ &\quad - \sum_{i=1}^{M}\nu_i\left(\epsilon + \xi_i + d_A\left(x_{i_1}, x_{i_2}\right) - d_i\right) \\ &\quad - \sum_{i=1}^{M}\hat{\nu}_i\left(\epsilon + \hat{\xi}_i - d_A\left(x_{i_1}, x_{i_2}\right) + d_i\right)\end{aligned} \tag{3.10}$$

We now substitute for $d_A\left(x_{i_1}, x_{i_2}\right)$ using Eq. 3.1 and then set the derivatives of the Langrangian with respect to the variables A, ξ and $\hat{\xi}$ to zero.

$$\frac{\partial L}{\partial A} = 0 \Rightarrow A = \sum_{i=1}^{M}\left(\nu_i - \hat{\nu}_i\right) z_i z_i^T \tag{3.11}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \nu_i + \mu_i = C \tag{3.12}$$

$$\frac{\partial L}{\partial \hat{\xi}_i} = 0 \Rightarrow \hat{\nu}_i + \hat{\mu}_i = C \tag{3.13}$$

Plugging the results back to the Langrangian we get the dual problem which involves maximization over the dual variables ν, $\hat{\nu} \in \mathbb{R}^M$:

$$\begin{aligned}\mathcal{L}\left(\nu, \hat{\nu}\right) &= -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\nu_i - \hat{\nu}_i\right)\left(\nu_j - \hat{\nu}_j\right)\left\langle\left(z_i z_i^T, z_j z_j^T\right)\right\rangle_{\mathbb{R}^{n^2}} \\ &\quad -\epsilon\sum_{i=1}^{M}\left(\nu_i + \hat{\nu}_i\right) + \sum_{i=1}^{M}\left(\nu_i - \hat{\nu}_i\right) d_i\end{aligned} \tag{3.14}$$

This Lagrangian should be maximized under the constraints:

$$0 \leq \nu_j, \hat{\nu}_j \leq C \qquad \forall j = 1 \dots M \tag{3.15}$$

It should be noted that typically the vectors $\tilde{\nu}$, $\hat{\nu}$ are sparse, and specifically in the case of SVR, the Karush-Kuhn-Tucker (KKT) conditions force a complementarity condition which states that either $\nu_j \neq 0$ or $\hat{\nu}_j \neq 0$, but not both meaning that a point may be wrong on one side of the ϵ-tube but not on both sides. Finally, the inner product term $\left\langle z_i z_i^T, z_j z_j^T) \right\rangle_{\mathbb{R}^{n^2}}$ can be replaced by a kernel of choice, thus allowing implicit mapping to high dimensional feature space.

Before we do that however, note that naively calculating this inner product is $O(n^2)$. Instead, we can use the trace identities and derive

$$\left\langle \left(z_i z_i^T, z_j z_j^T \right) \right\rangle_{\mathbb{R}^{n^2}} = tr\left(z_i z_i^T z_j z_j^T \right) = \left(z_i^T z_j \right)^2 \tag{3.16}$$

and calculate this term in $O(n)$.

3.1.2.1 Introducing Kernels

Introducing kernels, we replace $\left\langle (z_i z_i^T, z_j z_j^T) \right\rangle_{\mathbb{R}^{n^2}}$ with $\hat{k}(z_i z_i^T, z_j z_j^T)$ for some kernel function $\hat{k}(\cdot, \cdot)$. We could use any kernel function defined on pairs of vectors in $\mathbb{R}^{n^2}$, but doing this naively would result again in $O(n^2)$ complexity. Recall that kernel functions rely on an implicit (possibly infinite) feature mapping $\Phi : \mathbb{R}^l \to \mathbb{R}^\Omega$. Also note, that all our examples in this case are symmetric outer products of the sort zz^T. Thus, a natural feature mapping in our case is: [1]

$$\hat{\Phi} : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{\Omega^2} \tag{3.17}$$

$$\hat{\Phi}(zz^T) = \Phi(z)\Phi(z)^T \tag{3.18}$$

Now, given a kernel k derived from an implicit mapping Φ, we can create a new

[1] This mapping is (almost) well defined, because any symmetric rank one matrix A can be decomposed into an outer product of two identical vectors $A = vv^T$. The "almost" is due to the fact that we may replace v with $-v$, and this may matter for inhomogenous polynomial kernels

kernel $\hat{k}$ for the space of outer products based on the implicit mapping $\hat{\Phi}$:

$$\begin{aligned}
&\hat{k}\left(z_1 z_1^T, z_2 z_2^T\right) \\
&= \left\langle \hat{\Phi}(z_1 z_1^T), \hat{\Phi}(z_2 z_2^T) \right\rangle_{\mathbb{R}^{\Omega^2}} \\
&= tr\left(\hat{\Phi}(z_1 z_1^T) \cdot \hat{\Phi}(z_2 z_2^T)\right) \\
&= tr\left(\Phi(z_1)\Phi(z_1)^T \Phi(z_2)\Phi(z_2)^T\right) \\
&= \left(\Phi(z_1)^T \Phi(z_2)\right) \cdot tr\left(\Phi(z_1)\Phi(z_2)^T\right) \\
&= \left(\Phi(z_1)^T \Phi(z_2)\right)^2 \\
&= k\left(z_1, z_2\right)^2
\end{aligned} \tag{3.19}$$

So given any kernel defined for vectors in $\mathbb{R}^n$, we can use the square of this kernel for the outer product problem, with no cost in computation complexity. Restating the problem: given a training set consisting of pairs of n-dimensional vectors $\{x_{i_1}, x_{i_2}\}_{i=1}^M$, and their distances $\{\hat{d}(x_{i_1}, x_{i_2})\}_{i=1}^M$. Denote $z_i = (x_{i_1} - x_{i_2})$, $\hat{d}_i = \hat{d}(x_{i_1}, x_{i_2})$. Choosing any kernel function $k(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we solve the following optimization problem:

$$\begin{aligned}
&\max_{\nu,\hat{\nu}} L\left(\nu, \hat{\nu}\right) = \\
&\quad -\tfrac{1}{2}\textstyle\sum_{i=1}^M \sum_{j=1}^M (\nu_i - \hat{\nu}_i)(\nu_j - \hat{\nu}_j) k(z_i, z_j)^2 \\
&\qquad -\epsilon \textstyle\sum_{i=1}^M (\nu_i + \hat{\nu}_i) + \sum_{i=1}^M (\nu_i - \hat{\nu}_i) d_i \\
&\quad s.t. \qquad 0 \le \nu_j, \hat{\nu}_j \le C \qquad \forall j = 1 \ldots M
\end{aligned} \tag{3.20}$$

This problem is quadratic in the number of support vectors and linear in the sample dimension n (assuming the kernel function $k(\cdot,\cdot)$ is linear in n), which is exactly the computational complexity of ordinary kernel support vector regression in $\mathbb{R}^n$.

Prediction for the distance of a test data pair x_1^{test}, x_2^{test} is given by:

$$d(x_1^{test}, x_2^{test}) = \sum_{i=1}^{M} (\nu_i - \hat{\nu}_i)\, k\left(x_1^{test} - x_2^{test}, z_i\right)^2 \tag{3.21}$$

3.1.3 General Pairwise Relations Learning

In the way the problem of learning distances was stated above, only the differences between vectors were taken into account, thereby necessarily making any model learned translation invariant in the input space. However a different version can be stated as follows:

Suppose we wish to learn a *similarity* measure parametrized by a matrix A, $s(x, y) = x^T A y$. If A is symmetric positive definite, then this measure obeys the axioms of an inner product; however, we will deal with general A. Our training set is now of the form $\{x_{i_1}, x_{i2}, \hat{s}(x_{i_1}, x_{i_2})\}_{i=1}^{M}$, where $\hat{s}(x_{i_1}, x_{i_2})$ is understood as a measure of similarity, and will be henceforth denoted s_i. Again, learning the matrix A can be viewed as a linear regression problem in the space of outer products xy^T. Note that in this version, the training instances are generally not symmetric matrices (They may even not be square matrices). However, when computing the "inner product of outer products" we can still use the same algebraic manipulation as in *Eq.* 3.2 and obtain

$$\begin{aligned} \left\langle x_{i_1} x_{i_2}^T, x_{j_1} x_{j_2}^T \right\rangle_{\mathbb{R}^{n^2}} = & \\ & tr(x_{i_1} x_{i_2}^T (x_{j_1} x_{j_2}^T)^T) = \\ & x_{i_1}^T x_{j_1} \cdot tr\left(x_{i_2} x_{j_2}^T\right) = \\ & \left(x_{i_1}^T x_{j_1}\right) \cdot \left(x_{i_2}^T x_{j_2}\right) \end{aligned} \tag{3.22}$$

Thus, following the same analysis as above and using the same kind of feature mapping [1] $\hat{\Phi}\left(xy^T\right) = \Phi(x)\,\Phi(y)^T$ and will obtain the following Lagrangian:

[1] Note that now the feature mapping is not well defined - for any scalar α, we have $xy^T = (\alpha x)\left(\frac{1}{\alpha} y^T\right)$. This may be addressed by a preprocessing step of normalizing all input data point, however this correction applies to the linear case only.

$$L(\nu, \hat{\nu}) = \\ -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(\nu_i - \hat{\nu}_i)(\nu_j - \hat{\nu}_j)k(x_{i_1}, x_{j_1})k(x_{i_2}, x_{j_2}) \\ -\epsilon\sum_{i=1}^{M}(\nu_i + \hat{\nu}_i) + \sum_{i=1}^{M}(\nu_i - \hat{\nu}_i)d_i \tag{3.23}$$

where again $k(\cdot,\cdot)$ is any kernel defined on pairs of vectors in $\mathbb{R}^n$. The predicted similarity of a new pair x_1^{test}, x_2^{test} is then:

$$d(x_1^{test}, x_2^{test}) = \sum_{i=1}^{M}(\nu_i - \hat{\nu}_i)k(x_1^{test}, x_{i_1})k(x_2^{test}, x_{i_2}) \tag{3.24}$$

3.1.4 Implementation Remarks

The square of a kernel function is itself a kernel. If our original kernel was (non-) homogeneous polynomial of degree p, the new kernel (in the case of distance functions) will also be (non-) homogeneous, of degree $2p$. Thus any even polynomial kernel can be used for $\hat{k}(\cdot,\cdot)$. It is readily seen that using an odd degree polynomial kernel would lead to an anti-symmetric distance function: $d(x,y) = -d(y,x)$. If our original kernel is RBF with width γ, $k(a,b) = e^{-\gamma\|(a-b)\|^2}$, then the new kernel will be simply half the width $\hat{k}(a,b) = e^{-2\gamma\|(a-b)\|^2}$.

3.2 Experiments: Learning neural representations from paired distance measures

3.2.1 Data Representation

We used neural recordings made by Bar-Yosef et al [8] of single cell responses to variations of bird chirps taken from the Cornell Laboratory of Ornithology[1]. In this

[1]Library of Natural Sounds, Cornell Laboratory of Ornithology, Ithaca, NY

setting, described in details elsewhere ([8, 40, 41], $8-10$ variations of 4 different bird chirp stimuli, each of length $60-100$ ms were presented to an anesthetized animal.

3.2.1.1 Acoustic Stimuli

The stimuli consisted of main tonal component with frequency and amplitude modulation and of background noise consisting of echos and unrelated components. Each of these stimuli was further modified by separating the main tonal component from the noise, and by further separating the noise into echos and background. In addition, a stylized artificial version that lacked the amplitude modulation of the natural sound was used. In total, 8 versions of each stimulus were used in the A1 experiments 2 other versions were used in the IC experiment. Sample stimuli are shown in Figure 3.3

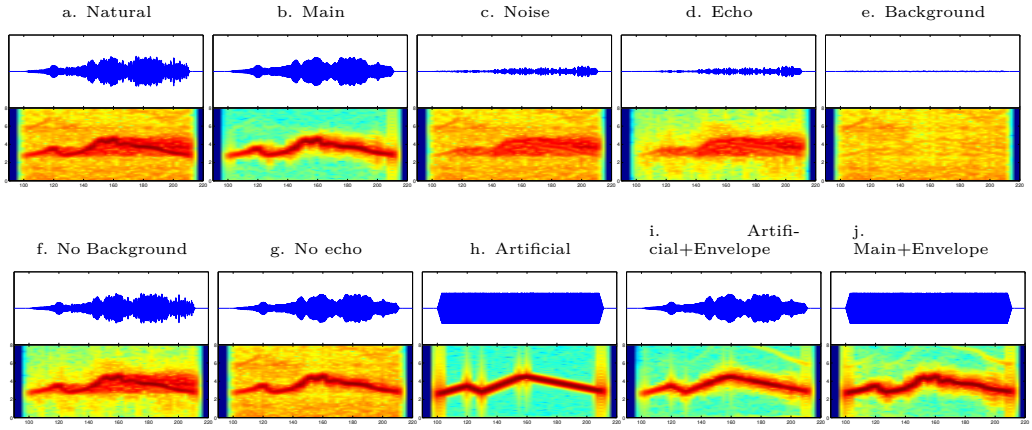

Figure 3.3: The bird song stimuli and their modifications. Each version is represented both as an oscillogram and as a spectrogram. The frequency range for all spectrograms is $0-8$ kHz. All spectrograms share the same color scale (covering a range of 60 dB), and all the oscillograms share the same scale. The time scale is identical for all versions of the same stimulus. The versions and their relationships are: Natural (Main+Echo+Background) (Figure 3.3a), Main (Figure 3.3b), Noise (Echo+Background) (Figure 3.3c), Echo (Figure 3.3d), Background(Figure 3.3e), Main+Echo (Figure 3.3f), Main+Background (Figure 3.3g) Artificial (Figure 3.3h) An Artifical version with the natural envelope (Figure 3.3i) and finally,the main component with an artificial envelope (Figure 3.3j). The last twi versions were not presented to cells in the auditory cortex.

As input to our learning algorithm we used the first 60 ms of each stimulus to extract vectorial representations. Each stimulus was represented using the first 15

real cepstral coefficients.[1] This kind of representation was previously used in [41] and elsewhere and is a well accepted representation. Alternative representations have been proposed by others (e.g. [56] and have proven beneficiary in terms of phonemes recognition.

3.2.1.2 Neural Responses

Neuronal response were represented by creating Peri-Stimulus Time Histograms (PSTHs) using 20 repetitions recorded for each stimuli. Response duration was 100 ms. An example of such responses in the form of spike rasters along with the PSTH are given in Figure 3.4 and show complex but repeating response patterns to different stimuli classes.

3.3 Results & Evaluation

3.3.1 Evaluation

In order to evaluate the quality of the learned distance function, we measured the correlation between *(1)* the distance computed by our distance regression algorithm, and *(2)* the distances as measured by the χ^2 distance over the responses.

We conducted our experiments in a leave one out (LOO) manner, in each experiment one stimulus was left out of the training sample. The rest of the stimuli-response pairs (39 in for IC cells, 31 for A1 cells) were then used to learn a cell specific distance function. For each test stimulus that was left out, we measured the distances to all other stimuli using the learnt distance function. To evaluate the quality of the learnt distances (and distance function), we then computed the rank order (Spearmen) correlation coefficient between these learnt distances in the stimuli domain and the χ^2 distances of the appropriate responses. A typical result for one left out stimulus in IC (IC-26 stimulus 1) and one left out stimulus in A1 (A1-38 stimulus 1) is presented in Figure 3.5 in the form of a scatter plot. This

[1] The real cepstrum of a signal x was calculated by taking the natural logarithm of magnitude of the Fourier transform of x and then computing the inverse Fourier transform of the resulting sequence.

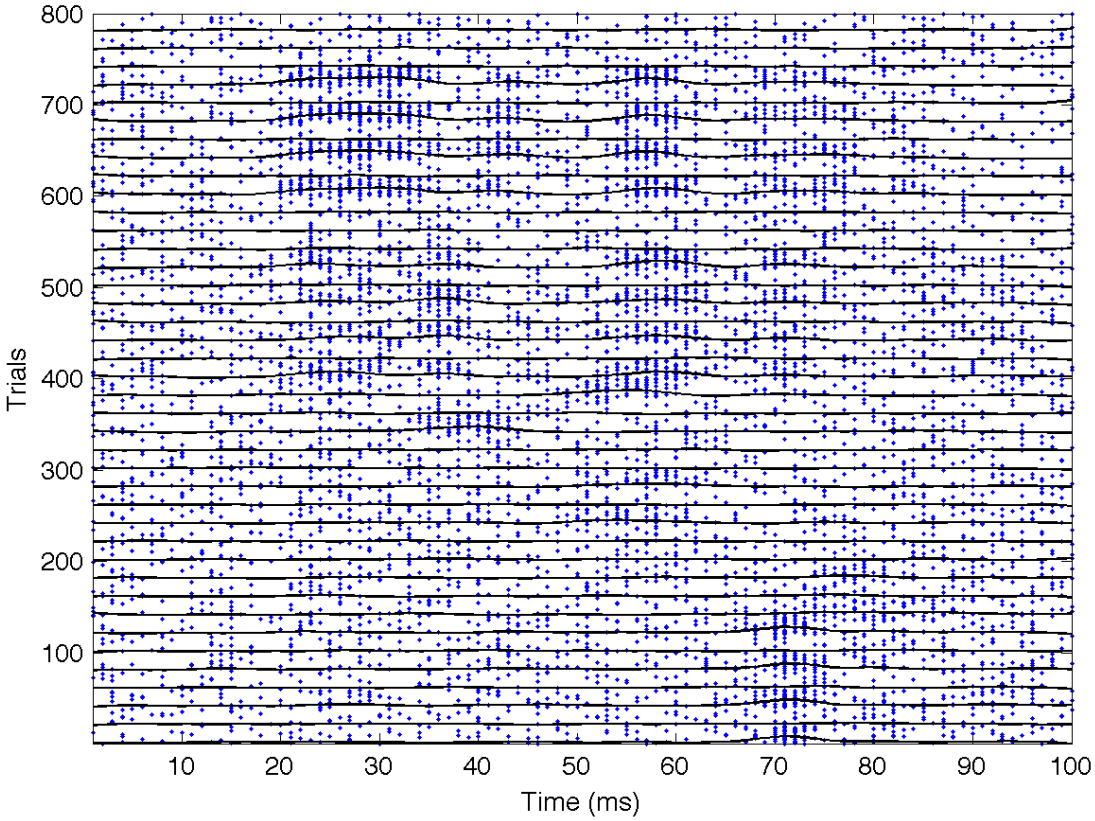

Figure 3.4: IC - Cell 26 spike Raster plot: - 100 ms (1 ms resolution) of the spike responses of an Inferior Colliculus (IC) neuron in response to 40 different stimuli, each stimuli was presented 20 times yielding a total of 40×20 trials

procedure produced a single correlation coefficient for each left out stimulus. In our evaluations we calculate the average correlation coefficient across all LOO scenarios tested on each cell.

3.3.2 Parameter Selection

The algorithm proposed has several degrees of freedom that need to be tuned to provide robust performance, those include the type of kernel used (and the proper kernel parameter and the parameters C, ϵ and σ. The proper way to adjust these parameters is through cross-validation (CV). Roughly speaking, when number of parameters is smaller than training examples CV guarantees, in theory, that only

a minor difference will exist between CV performance and testing performance and generally provides good results in practice.

3.3.3 Fitting Power of cell specific distance functions

We begin our analysis with an evaluation of the fitting power of our method, by training A1 and IC neurons with the entire stimuli set (40 and 32 respectively). For each cell, rank order correlation coefficient values between learnt distances in the stimulus domain and the χ^2 distances between the appropriate responses were calculated. The fitting power of a method is a measure of the fraction of the information that the model can explain on previously seen examples. In the present context, this is a measure of how well our method captures the relevant structure of the auditory stimulus space as induced by a specific cell, while trained with all the stimulus set. The average Spearman rank correlation coefficient between the learnt distances and the given distances as averaged over all cells was $(0.9973, 0.078)$ for IC cells and $(0.9681, 0.0243)$ for A1 cells.

These very high results suggest the hypotheses space is wide enough to capture the required structure; however it may also suggest an over-fitting problem. To circumvent this we used a cross validation approach to learn a regularized model such that generalization properties would be maximized. High correlations between mean train results and mean test results ($r(28) = 0.7099$ $p << 0.01$ for IC cells and $r(22) = 0.6849$, $p << 0.01$ for A1 cells) suggest that this was indeed the case and the learned model and hypothesis space were adequate for the data modeling (see Figure 3.6)

3.3.4 Generalization

In order to evaluate the generalization potential of our approach, we used a leave one out cross-validation paradigm. In each run, we removed a single stimulus from the dataset of N stimuli (40 for IC and 32 for A1), trained our algorithm on the remaining $N - 1$ stimuli using a 5-fold cross validation to determine the C, ϵ and σ parameters , and then tested its performance on the datapoint that was left out. The train results in this paradigm is the mean correlation for output distances

between the N-1 stimuli (that is $(N-1) \times (N-2)/2$ pairs) and the distances as defined between the responses. Test results are again mean correlations between the distances of LO stimulus and the train stimuli and the distances between the corresponding responses.

As can be seen, results for test performance are varied, even when cells from the same sub-system (A1 or IC) are considered. Average rank order correlations for IC neurons are at $\sim 0.64, 0.094$ and reaches an average rank-order correlations (over LOO runs) as high as ~ 0.83 (cell 26) and as low as ~ 0.39 (cell 20). The results for A1 neurons are more varied and are in the range of ~ 0.79 (cell 38) and ~ 0.23 (cell 51) with an average performance of $(0.435, 0.12)$. This larger variability in test performance may be attributed at least in part to the training sample size effect (32 for A1 vs. 40 for IC). These results compare favorably with the results obtained for the same data using the DistBoost algorithm [18] and an STRF model.

3.3.5 Comparison to Other methods

We compared our results to the results obtained using the DistBoost algorithm on the same data set as reported in [51] and to the results obtained using a filter estimation approach (STRF). In order to compare our approach to the STRF approach we used the same procedure as described in [51] and performed the following procedure using STRFPAK-5.3 package by [57, 58]. For each cell and each LOO scenario an optimal filter was computed using the $N-1$ training stimuli. A response was subsequently calculated for the remaining left out stimuli. Rank Order Spearman correlation coefficients were computed between the predicted responses' distances and true responses distances (as defined by the χ^2 distance). For 28 IC cell in the dataset mean rank-order correlation coefficient was $(0.39, 0.09)$. For A1 neurons dataset mean rank-order correlation coefficient was $(0.22, 0.067)$. Comparison results are shown in Table 3.1.

3.3.6 Stimuli Selection

After computing the test performance per cell we measured the predictability of each stimulus by averaging the LOO test results obtained for each stimulus across

	Distance Regression	Dist Boost	STRF
IC	(**0.64,0.094)**	(0.34, 0.034)	(0.39, 0.09)
A1	(**0.435,0,12)**	(0.24, 0.0019)	(0.22,0.067)

Table 3.1: Performance Comparison: Distances were calculated according to the three methods. Average and STD of rank order correlation coefficient between the estimated distances and the 'true' distances are shown in the table. A clear advantage to our method is indicated in bold.

all cell separately in IC and A1. A previous study [51] suggested that wide-band stimuli were better predicted than narrow-band stimuli. In our experiments this result could not be replicated neither in the IC nor in the A1. However, in the following we have trained the model using a small subset of the stimuli. In each setting either the subset of narrow-band was selected to train and test on, or the wide-band stimulus set. Results shown in Figure 3.9 for IC and Figure 3.10 for A1, show that training and testing on wide-band stimuli yields significantly better results. While there are cells that 'prefer' narrow-band stimuli most cells train and generalize better when given a set of wide-band results.

3.4 Conclusions and Future Work

We have introduced an efficient computational method for the characterization of sensory systems with a special focus on the characterization of single cells in the auditory processing stations. This computational method has the potential to reveal principles that govern information processing and frameworks of representation in a reacting system

The algorithm shows promising results in the fundamental task of single cell characterization. It achieves better trained models with better generalization properties than both existing distance learning algorithms which have been applied to the problem, as well as other models used for single cell characterization. Our distance regression algorithm has managed to capture key features in the way auditory stimuli are represented in both the IC and the A1. The results also suggest that there is a basic difference between the neurons in the IC and neurons in A1 the former being 'simpler' in terms of represented geometry.

This algorithm can easily and efficiently be applied to other problem in the field of neuroscience as well as in other fields. Possible applications are *a*) Characterization of the difference between cells as defined by their different response patterns. *b*) Learning spike metrics through distance regression. *c*) Large system characterization using multi-unit responses.Additionally, while we have applied the method for single cell characterization from electrophysiological recordings, the suggested framework, as well as the specific algorithm can be easily applied to characterize other, more elaborate systems, and with other types of response patterns such as EEG, EMG and fMRI traces to name a few.

An online version of the algorithm is a natural extension that can be of dramatic service in the experimental setting: It can be used to provide real time guidance to the experimenter leading the experiment to 'interesting' stimulus regions. An online version can be easily implemented in the linear case (when learning a Mahalanobis metric), but extra care should be given in the case of the kernelized version. Extrapolation from multiple cells is another extension that should be experimented with - the implementation of which - bring forth different variants of weighted SVR.

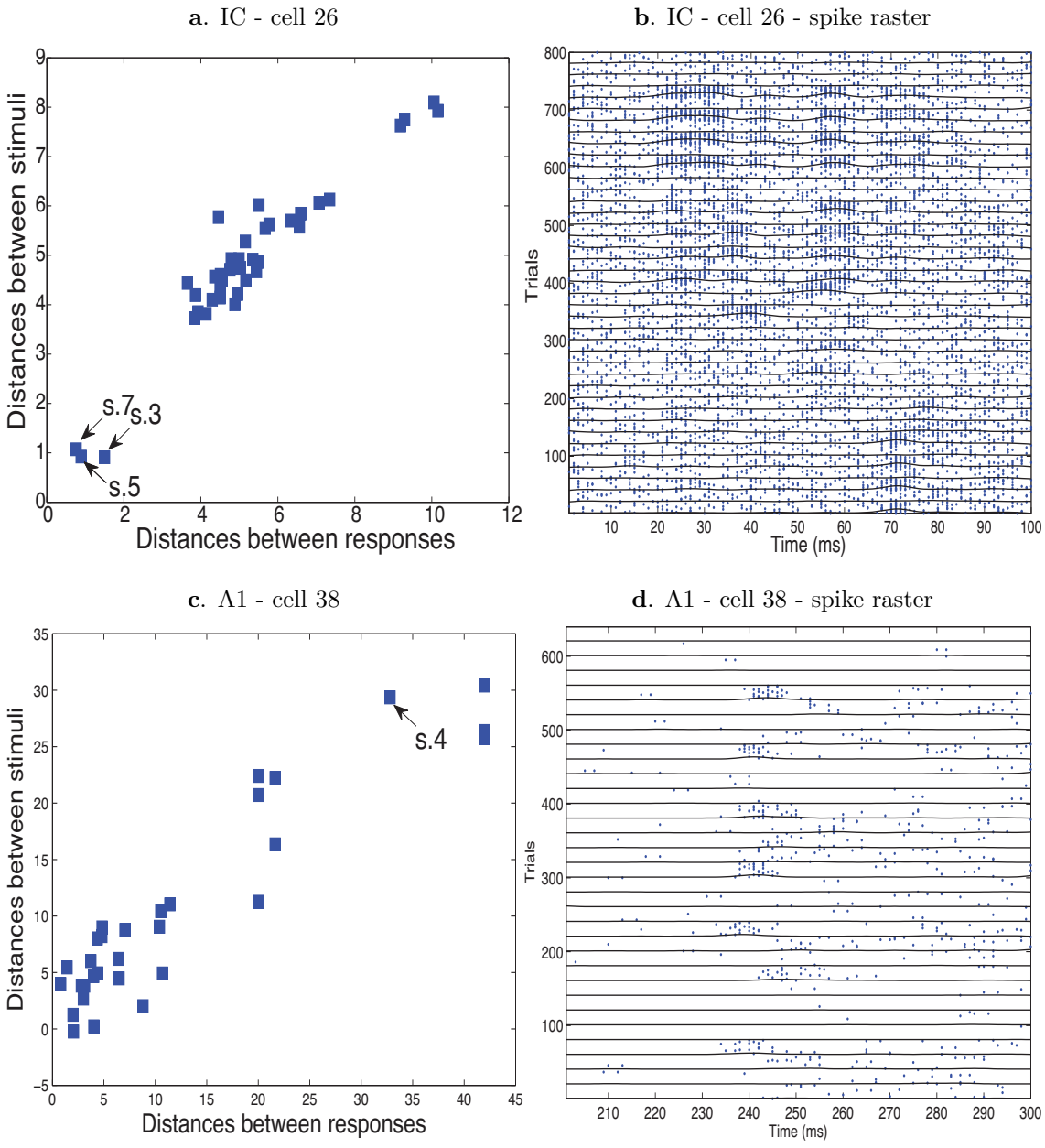

Figure 3.5: Distance regression results scatter: Figure 3.5a: Results for a single LOO run of Cell 26 in the IC. Figure 3.5c: Results for a single LOO run of Cell 38 in A1. The x axis of the figure is the χ^2 distances between the LOO stimulus' and the 39 (32) remaining neural responses. The Y axis is the distances induced by our DR algorithm between the LOO stimulus and the remaining 39 (32) train stimuli. The correlation is measured between these two vectors. In both cases stimulus number 1 s_1 was left out. Spike rasters are given as reference: bottom stimulus is stimulus 1 (s_1) (that was left out) second form below is s_2 etc. See Figure 3.5b for cell 26's raster plot and note that s_3, s_5 and s_7 indeed share a common response pattern with the LOO stimulus - A property very well captured by the algorithm.

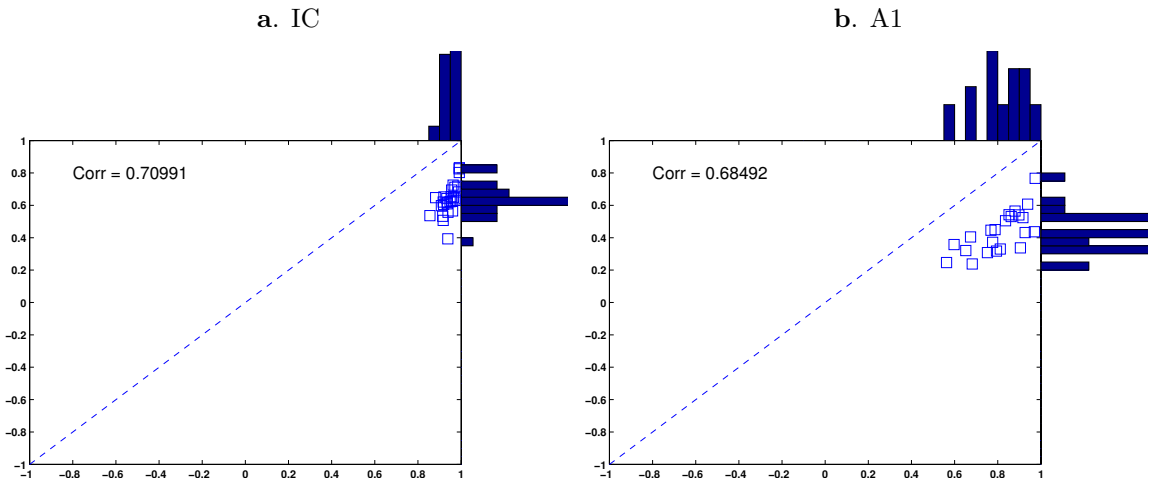

Figure 3.6: Train vs. Test correlations: Each point marks the average correlation of a single cell. The distribution of train and test correlations is displayed as histograms on the top (Train) and right (Test) IC is shown in Figure 3.6a. A1is shown in (Figure 3.6b)

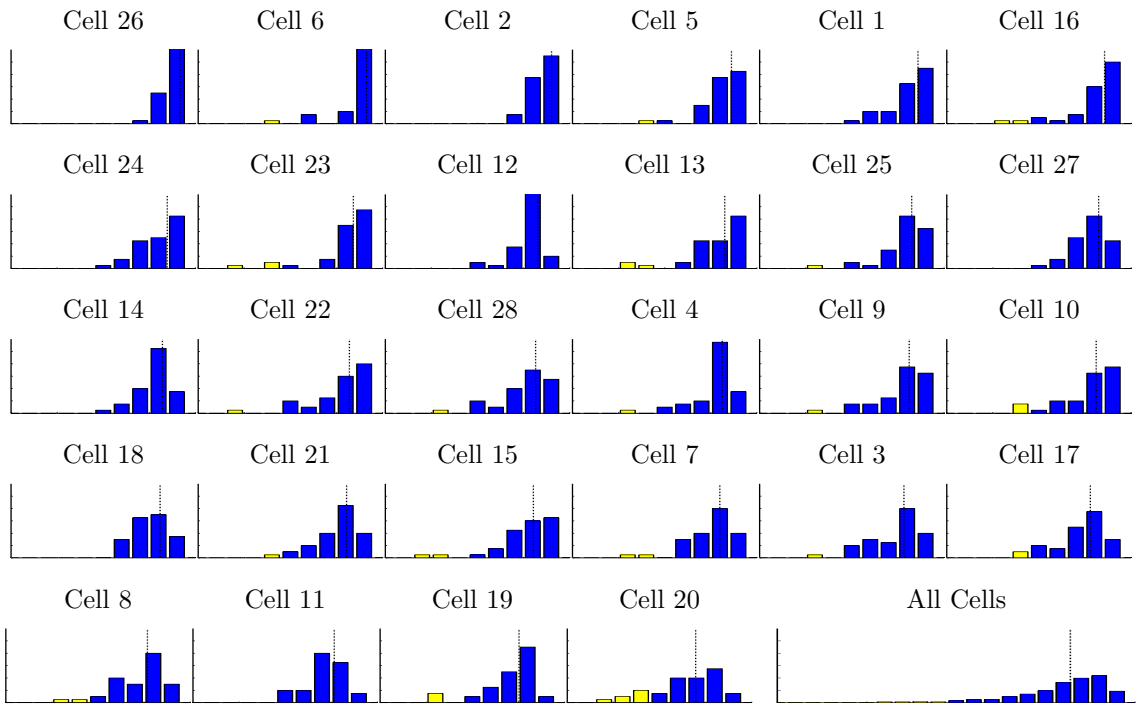

Figure 3.7: Inferior Colliculus (IC): single cells test results: Histograms of cell specific test-rank-order correlations for the 28 IC cells in the dataset. The given correlations are the rank-order correlations between the predicted distances and the distances between the recorded responses, measured on a single stimulus which was left out during the training stage. For each cell 40 LOO experiments were performed. Results displayed are histograms over these 40 experiments. For Visualization purposes, cells are ordered (columns) by their average test correlation per stimulus in descending order. Negative correlations are in yellow, positive in blue. Average over all LOO experiments is given by the dotted line at 0.64

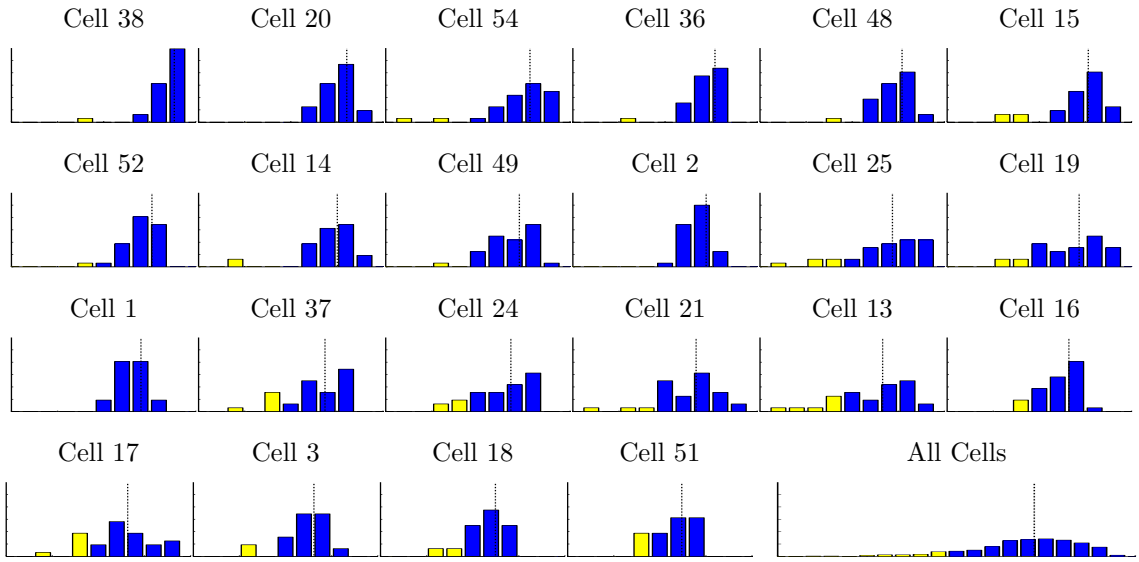

Figure 3.8: Primary Auditory Cortex (A1): Single cells test results: single cells test results: Histograms of cell specific test-rank-order correlations for the 22 A1 cells in the dataset. The given correlations are the rank-order correlations between the predicted distances and the distances between the recorded responses, measured on a single stimulus which was left out during the training stage. For each cell 32 LOO experiments were performed. Results displayed are histograms over these 32 experiments. For Visualization purposes, cells are ordered (columns) by their average test correlation per stimulus in descending order. Negative correlations are in yellow, positive in blue. Average over all LOO experiments is given by the the dotted line at 0.435

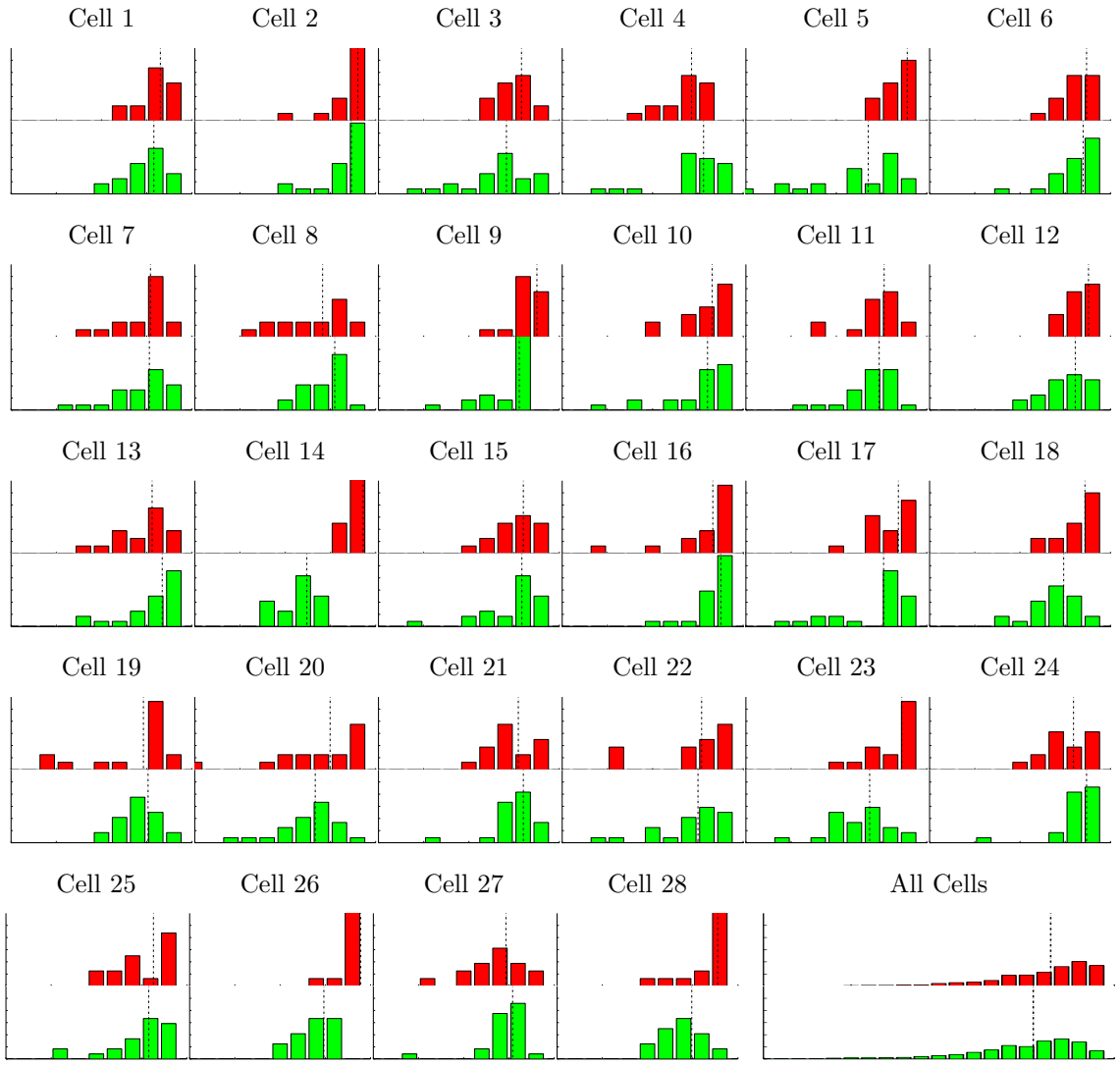

Figure 3.9: IC: Performance over a selection of the stimuli: Data was trained on a selection of the data, upper panel in each subfigure (red) indicates a model trained on wide-band stimuli and tested on wide band stimuli, lower subfigure (green) indicates a model trained on narrow band stimuli and tested on narrow band stimuli.

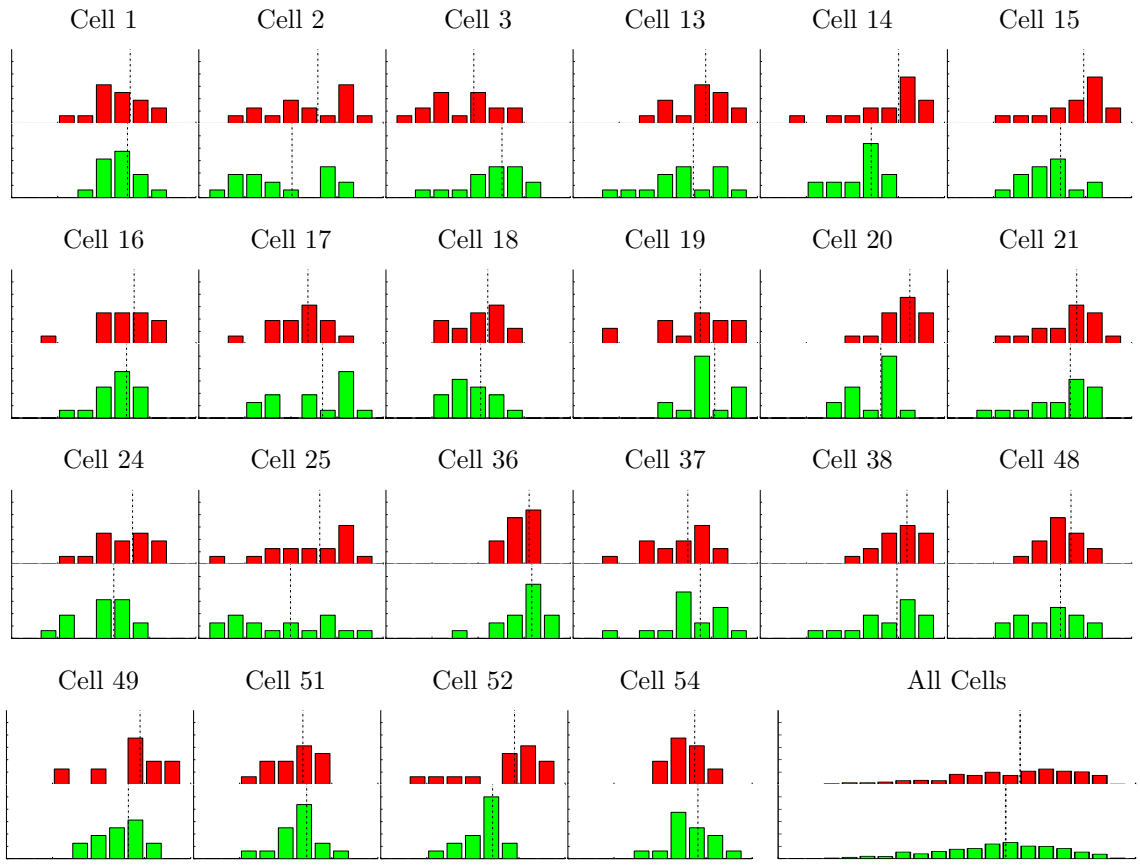

Figure 3.10: A1: Performance over a selection of the stimuli: Data was trained on a selection of the data, upper panel in each subfigure (red) indicates a model trained on wide-band stimuli and tested on wide band stimuli. lower subfigure (green) indicates a model trained on narrow band stimuli and tested on narrow band stimuli.

4

Speech Reconstruction using Local-Non Local Self Similarity

"A brain is only capable of what it could conceive, and it couldn't conceive what it hasn't experienced"

— Graham Greene, Brighton Rock.

The terms *signal enhancement* and *signal cleaning* refer respectively to the improvement in the quality or intelligibility of a signal and the reversal of degradations that have corrupted it. In practice however, the two terms are used interchangeably. In this chapter we present a single channel signal enhancement technique that uses a variant of *non-local means* method over the spectral domain representation to estimate and enhance the spectral amplitude of a noisy speech signal. Adding on previously reported techniques, the proposed technique makes use of two pieces of information that may be available when one is challenged with a distorted signal: the first one is previous knowledge of the nature of the signal and possibly samples of the same signal that are of better quality, the second one is an understanding or an assumption about the degradation process the signal has gone through.

In speech processing and communication, it is often the case that signals appear at different qualities at different times. Changing quality may be the result of different encoding schemes, unstable channel properties, changing environmental noise, or any number of other reasons. Such cases are evident in our daily use of radio transmissions, cellular technology or Voice over IP (VoIP). Single channel speech enhancement is particularly difficult when the noise is non-stationary (such as traffic noise). In many such cases the receiving end has no control - or little control - over the encoding scheme or over the quality of the channel and is left at the later end of the communication channel with a distorted signal.

Speech enhancement algorithms usually involve a transformation of the degraded signal into its spectral domain representation. A common practice in many such

cases, focuses on altering the amplitude of the signal (e.g. spectral subtraction [59, 60], Wiener filtering [61, 62], signal subspace approach (SSA) [63], Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimator (MMSE-STSA)[64], Time-Frequency masking [65] and model based enhancement [66]). These methods are general purpose and rely on general signal properties such as signal to noise ratio (SNR) estimation, or harmonic component identification to selectively enhance or suppress certain time frequency frames and are successful in enhancing certain degradation scenarios. While impressive improvement has been achieved in some of the cases (for some types of noise, and for some specific settings) the full potential of signal enhancement has yet to exhaust itself.

Signal degradation may appear in many different forms and settings, in some of the cases, some often overlooked at, pieces of information are available at the scene: firstly, one may have some prior idea about the nature of encoding or distortion the signal has gone through and that has supposedly caused the degradation in quality. Alternatively, one may be able to infer, estimate or learn the nature of degradation e.g. the level of quantization or clipping. Notably, knowing the type of degradation does not guarantee the ability to reconstruct the original signal as most degradations are not invertible - that is they come with an inherent loss. Secondly, one may have samples of the same or similar source which are of better quality - these were presumably received and collected earlier - when better communication conditions were available.

In this chapter, we assume that a degraded signal with known degradation characteristics has been received and try to correct it. We also assume that prior samples of better quality are available at the receiving end of the channel. As a running example, we treat the case of speech signal that has been quantized at low bit rate. We attempt to reconstruct the given distorted signal by using a dictionary built earlier from previously seen, high quality, signal samples.

The remainder of the chapter is structured as follows: in Section 4.1 we briefly review the challenge of signal degradation. We focus the discussion on quantization noise and discuss solutions proposed to handle different aspects that arise in the context. While the scope of this study is restricted to handling quantization noise (see Section 4.1.2.1) which is a challenging task in itself, we maintain the claim that the same approach can be used for other types of noise and in more elaborate

scenarios. In Section 4.2, we then introduce the building blocks of our proposed approach and continue to present a detailed algorithmic scheme. We continue in Section 4.3.2 with a small detour and attempt to reconstruct a high quality signal from samples of the same source and quality. By doing so we establish the notion and the limitations of self-similarity in speech signals, our use of the non-local means method as well as our parameters choice. We then explore the performance of this method in the absence of same speaker samples (but with no degradation) and show that to a great degree performance is maintained. This is followed by the results of some informal experiments (Section 4.3) in same speaker scenario. We conclude this chapter with a discussion of the results, possible extensions of the working framework and suggestions for future work.

4.1 Related Work

4.1.1 Nonlocal means

The nonlocal means (*NL-means*) algorithm recently proposed by [67] for image de-noising has proved highly effective for removing additive noise in images while to a large extent maintaing image details. The algorithm performs de-noising by averaging each pixel with other pixels that have similar characteristics in the image. The NL-means method takes advantage of two general assumptions. The first assumption is that additive noise in a signal is uncorrelated and has zero-mean. The second assumption is that every ‘small‘ window in a signal which encloses many samples has a number of perceptually similar windows in the same signal. This property is termed *self-similarity*.

NL-means approach was applied to speech reconstruction, speech de-noising, and speech enhancement (e.g [68]); however, its application has been limited bymed both assumptions to settings where reconstruction a signal from itself was possible. In [68] self-similarity property was assumed to be the result of a local harmonic structure rather than to the more ’global’ property of statistical regularities of speech. This assumption directed the authors to take a very local NL-means approach and the application was thus very limited in nature. The authors of [68] also mention that the

assumption of zero mean noise is not met in their setting and propose a separation of real and imaginary parts to correct for that and average out the noise.

We employ the notion of NL-means to reconstruct degraded speech signal. Similarly to other speech enhancement techniques [60, 61, 63, 64, 65, 66] we attempt to enhance the short term spectral magnitude of the speech. We do so by reconstructing it from other time slices with similar spectral shapes. In that respect we assume that given a "large enough" codebook or a "small enough" word the *self similarity* assumption is fulfilled. The other assumption, of zero mean noise, is approximately valid since we reconstruct the noisy patch as a linear combination of clean high quality patches and thus, to the extent that no bias was built into the codebook, the second assumption is valid as well.

4.1.2 Signal degradation and reconstruction

When handling speech signal one can encounter various types of phenomena causing the signal of interest to degrade in both quality and intelligibility. Possible degradations include but are not limited to: additive acoustic noise, acoustic reverberations, convolutive channel effects and non-linear distortion such as the distortion that arises from clipping or quantization.

Different settings may cause different types of degradation, pose diverse challenges and call for different approaches to speech and signal enhancement and reconstruction. In the general sense, enhancing or reconstructing a signal demands the utilization of some degree of freedom (e.g. an additional microphone, a statistical model of the signal or/and of the noise process etc). Knowing the type of noise that was introduced to the signal can help but does not guarantee success, as degradation may be irreversible.

In the following we discuss reconstruction from a single microphone mainly quantization distortion which is a degradation one commonly encounters when handling digital signals especially when communication constraints are involved. Quantization noise is usually handled at a preprocessing station using the technique of *dithering* (Section 4.1.3). Other techniques are specialized for other types of noise and are available also at the post processing station: a popular one is that of *spectral subtraction* which is discussed in brief in Section 4.1.3.2.

4.1.2.1 Quantization and Quantization Distortion

Quantization, in digital signal processing, is the process of mapping a large set of input values to a smaller set. Such a process may be achieved, for example by rounding values to some unit of precision. The simplest form of coding analog signals, Pulse-code modulation (PCM) is a widely used quantization method used to digitally represent sampled analog signals. (Telephone applications frequently use 8-bit quantization.) Quantization also forms the core of essentially all lossy compression algorithms which in turn may degrade signal quality significantly. The objective in the design of a coder is to achieve high fidelity with as few bits per sample as possible, at an affordable implementation cost. Crude quantization may result from hard bandwidth constraints and may introduce substantial quantization noise into the signal. Because quantization is a many-to-few mapping, it is an inherently non-linear and irreversible process and thus, without additional knowledge or assumptions a signal can only be approximated to some degree from its quantized version. We note that while we treat the case of crude linear PCM quantization, in most compression schemes the quantization in not of the PCM levels directly but rather of a different representation of the signal. Having said that, we maintain our claim that our reconstruction scheme is invariant to what was quantized as long as the type of transformation/compression is known or can be efficiently estimated.

In analog-to-digital conversion, the difference between the actual analog value and the quantized digital value is called *quantization error* or *quantization distortion*. This error is the result of naive compression either due to rounding or truncation. Generally speaking as the step size increases the quantization effect becomes severe. The error signal is sometimes considered as an additional random signal called quantization noise because of its stochastic behavior. Perceptually, quantization noise significantly degrades both pleasantness and intelligibility of speech and poses a significant challenge both to human listeners and to automatic speech recognition (ASR) systems in particular when the systems are trained on clean speech. Poor adaptivity has become one of the major barriers for the widespread deployment of ASR systems and their derivatives. As seen in Figure 4.1, signal quantization results in a rough signal temporally which in turn manifests itself in general spread over

time and frequency axes as well as pronounced high frequency component in the spectral representation.

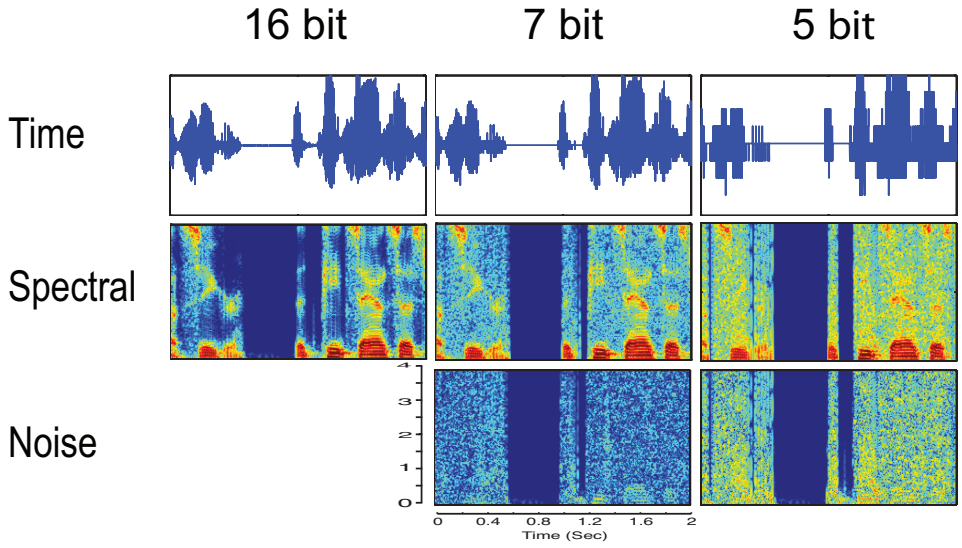

Figure 4.1: Quantized Signals - Three extracts of the same 2 seconds in its original form (left), 7 bit quantization (center), and 5 bit quantization (right). Signal is given in time domain (upper row), spectral domain (middle row), the difference in spectral amplitude between original and noisy signal is given for reference (bottom row).

4.1.3 Speech Denoising

4.1.3.1 Dithering

Dither (Dithering) is an intentionally applied form of noise used to randomize quantization error, allowing modeling of the noise as additive uncorrelated noise and preventing large-scale patterns such as ”banding” in images and transient high frequency components in sound [69]. Dither is routinely used in processing of both digital audio and digital video data, and is often one of the last stages of audio production to compact disc. This step is applied to the signal pre quantization to prevent in advance the adversarial effects of quantization. While it is indeed successful in limiting these effects it requires access to the pre-quantized data - an access sometimes wanting. A different approach to de-noising - one that admits post hoc speech enhancement - is that of spectral subtraction.

4.1.3.2 Spectral subtraction

One of the most popular methods of reducing the effect of background (additive) noise is *Spectral Subtraction*. The spectral subtraction approach incorporates a family of algorithms that essentially attempts to estimate the noise component in a spectral slice and reduce it form the noisy slice. In power spectral subtraction, the power spectrum of noise is estimated during speech inactive periods and subtracted from the power spectrum of the current frame resulting in the power spectrum of the speech. A crucial assumption in power spectral subtraction is that the expected phase difference between the signal and noise is zero. Generally, spectral subtraction is suitable for stationary or very slow varying noises (so that the statistics of noise could be updated during speech inactive periods). In addition, when performing spectral subtraction two reasonable assumptions (which are not met when quantization noise is at stake) are made: *a*) Noise and speech magnitude spectrum values are independent of each other. *b*) The phase of noise and speech are independent of each other and of their magnitude.

As a result of the fluctuations of noise spectrum (whether power or magnitude) around its mean (expected) value, there is always some difference between the actual noise and its mean value. Hence some of the noise remains in the spectrum in the case that the value of noise is greater than its mean and some of the speech spectrum also is removed in the case that our estimate of noise is greater than the actual value of noise. The latter produces negative values in spectrum. These negative values are prevented or set to a floor (sometimes zero) using different techniques. The overall effect puts a noise in the output signal known as *residual*. The narrow band relatively long-lived portion of residual noise is sometimes referred to as *musical noise*

4.2 Speech Reconstruction

Speech analysis is often conducted in the spectral domain on a frame by frame basis. The Short Term Fourier Transform (STFT) is a very efficient and simple tool for audio signal processing it allows for separate processing of magnitude and phase, and under reasonable assumptions is also invertible. For each frame the spectrum is

obtained through the fast Fourier transformation (FFT) resulting in a 2-D display where the horizontal axis is the time (frame) index and the vertical axis represents the central frequency (cf). Each element in this 2-D representation represents the magnitude of a time-frequency component. Similarly to other methods [60, 61] we focus our efforts on enhancing the amplitude of the degraded signal and leave the phase untouched. Phase estimation has previously been thought of as having less importance in speech enhancement [70] however recent work has shown phase estimation may play a more significant role in pushing the envelope of speech enhancement applications [71, 72]. While we believe both that phase estimation may help improve our results and that a variant of our method can be efficient for the task of phase estimation as well, we leave phase estimation for future work.

In the proposed framework we produce a 2-D spectrogram and treat this representation of the magnitude as an image. Phase of the degraded signal is left for later use when inverting the enhanced magnitude back to the time domain using the inverse Short Term Fourier Transform (iSTFT). To allow a proper adaptation of the NL-Means algorithm the equivalents of *a*) a pixel, *b*) its environment (A patch) and *c*) An adequate distance function need to be properly defined.

In this study we take a single STFT frame to be the equivalent of a pixel. A k frames time slice is taken as the patch. The size of the patch determines the significance we assign to the context of the frame we wish to reconstruct, but at the same time it also reflects our confidence in the robustness of our codebook: a "thin" codebook (having less entries) will not allow large patches as enlarging the patches pushes each frame (and its context) towards idiosyncrasy. Once patches are defined, a proper distance function between patches needs to be defined. The choice of a distance function is crucial for the success of any NL-Means application; however, a second crucial component of this algorithm involves efficient search in the codebook. Regretfully, these demands may be competing. As our data structure we chose the well known *KD-tree* which forces the use of a linear distance function. We, thus used the Euclidean distance function to determine similarity of patches. However, pre-emphasis was given on more local points in the patch using pseudo-Gaussian weighting over the time axis. In addition, to specifically handle the effect of quantization, we reweighed along the frequency axis. We are left with a linear distance function that highly weighs higher frequency components which tend to be

ignored due to low energy. While the framework itself is not restricted to linear distance functions the choice of a KD-tree as an efficient database restricts this choice. Other, more flexible, options are possible and are discussed in Section 4.5

The outline of the proposed algorithm is as follows: *1*) Given a degraded signal (that has gone through a known or estimated type of degradation). *2*) Take a high quality signal of similar nature, degrade it in similar manner and keep the degraded signal aligned to its origin to construct a codebook of key-value pairs (where the degraded part serves as the key and the original "good quality" part is the value. *3*) For each point in the degraded signal, find the k-"closest" keys in the codebook. *4*) Extract the corresponding k values and combine them together (with weights relative to their similarity to the query patch). *5*) Combine the values together and get an enhanced signal.

Figure 4.2 illustrates the preprocessing steps taken to construct a code book and allow for high quality reconstruction. In (A) high quality signal is purposefully degraded to create a signal of similar quality to the one we wish to enhance. The two signals are then transformed to spectral domain (B) from the spectrogram we create a code book of key-value patches (C). We use the spectral representation of the high quality signal to build a codebook. A "word" in the code book comprises of concatenated spectral time frame centered a a given time with approximately Gaussian weights around the center time frame (see D).

To reconstruct a degraded signal we again transform to spectral domain and work patch by patch [1]. Figure 4.3 illustrates the steps taken to reconstruct a queried patch. Given a degraded patch we find the $k-$Nearest keys in the dictionary. We then take the k corresponding values and combine them together according to their distance form the queried patch. After querying the entirety of the degraded signal we take the overlapping patches, combine them with the degraded signal's phase and use the iSTFT to get back to the time domain with an enhanced signal.

[1]This need not be done sequentially and can be used in a distributed setting e.g. with an implementation of the Map reduce paradigm.

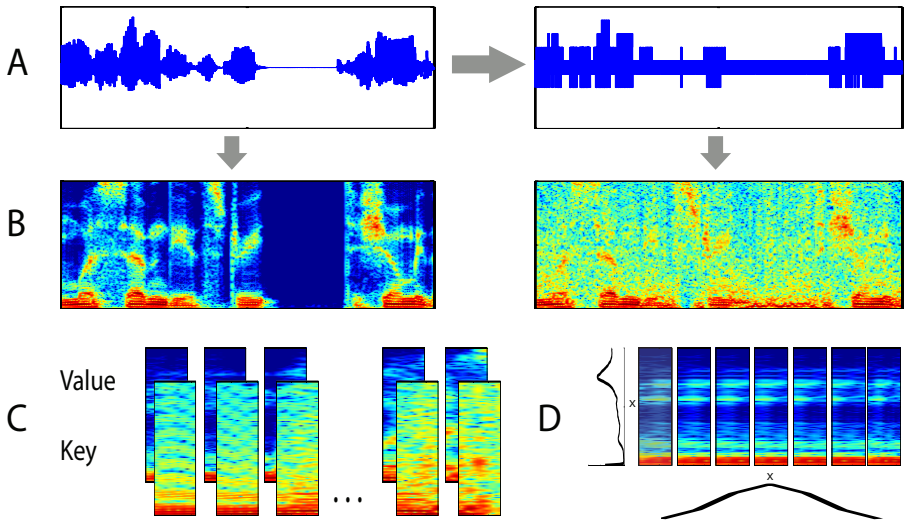

Figure 4.2: Preprocessing: - high quality signal (A) is transformed into low quality one (B) time domain signal is transformed into spectral domain representation and used to extract segments and build a key-value dictionary (C). D illustrates the construction of a patch: a k-neighborhood of a segment is re-weighted in both time and frequency axis to emphasize the desired segment (time emphasis) and also highlight high frequencies (frequency weighting)

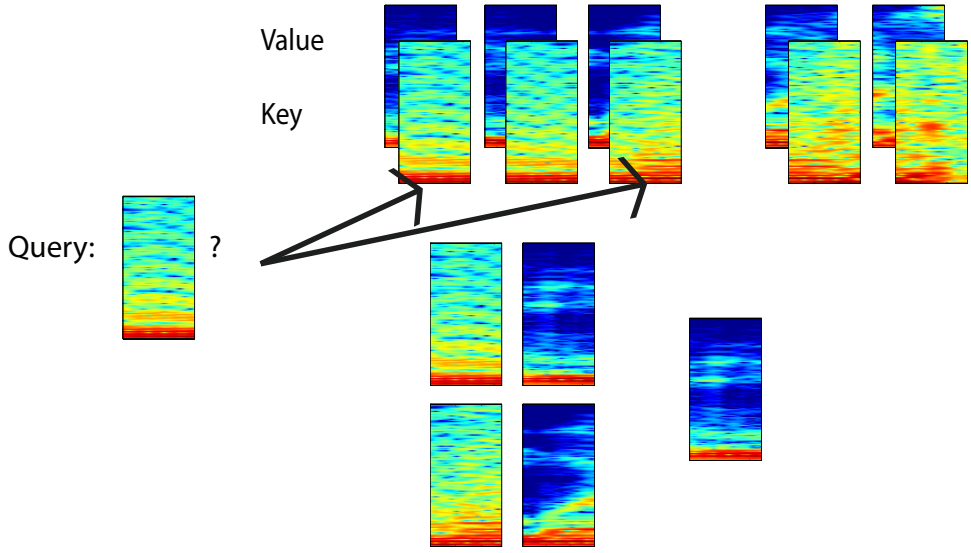

Figure 4.3: Querying: - A patch of degraded signal is queried to the codebook. K matching key-value pairs are taken and reweighed according to their distance and linearly combined to form a new enhanced patch.

4.3 Experiments and Results

In the following we present first the evaluation measure used to evaluate the reconstruction framework. We then present an attempt to reconstruct a speech signal

from other signals of high quality, thus establishing the existence of the self-similarity property in the current setting. Finally, we demonstrate the reconstruction of distorted signals that have been corrupted by coarse quantization. All speech samples were extracts taken from a variety of audio books (see details below). Samples of degraded speech and their enhanced counterparts as well as other examples are available on Roi Kliper's Home Page.

4.3.1 Evaluation

The aims of speech enhancement vary according to the application and may include: *(i)* Improvements in the intelligibility of speech to human listeners. *(ii)* Improvement in the quality of speech that make it more acceptable to human listeners. *(iii)* Modifications to the speech that lead to improved performance of automatic speech or speaker recognition systems. *(iv)* Modifications to the speech so that it may be encoded more effectively for storage or transmission. Most of the literature on the subject has concentrated on the last three points. The distinction is important because it has been found that these demands may at times be competing. For example, some approaches that are effective in improving the signal to noise ratio, may actually damage intelligibility. Despite its apparent similarity to that of improving intelligibility, improving the performance of speech recognition systems is significantly different because such systems ignore many of the cues used by humans and, in particular, normally use a coarse spectral representation of the signal from which most voicing information has been removed.

For the evaluation of the reconstructed signal, we employed Perceptual Evaluation of Speech Quality *PESQ* as implemented in [73]. PESQ is a worldwide applied industry standard for objective voice quality testing used by phone manufacturers, network equipment vendors and telecom operators. PESQ was particularly developed to model subjective tests commonly used in telecommunications to assess the voice quality by human beings. Consequently, PESQ employs true voice samples as test signals; that is it is a "Full Reference" (FR) algorithm and analyzes the speech signal sample-by-sample after a temporal alignment of corresponding excerpts of reference and test signal. PESQ results range from -0.5 to 4.5 and principally

model Mean Opinion Scores (MOS) that cover a scale from 1 (bad) to 5 (excellent). A mapping function to MOS-LQO is outlined under P.862.1 [74] and added as reference in Figure 4.5B.

4.3.2 Validity of Self Similarity in Audio Representation

Real speech data is made of many variable pattens that repeat. This property is termed in other contexts *self-similarity* and was used effectively for example in the context of speech motif pattern matching [75]. Given such a property one can construct a low dimensional representation of spoken language and use such a model for various application. However, this property is highly dependent on both inter-speaker and intra-speaker variability and is generally constrained by intrinsic speech variability. Self similarity in our context is a crucial property of the algorithm as each spectral slice in the reconstructed signal is essentially reconstructed from a linear combination of (other) existing segments.

To see that self similarity is indeed a valid assumption in our setting we reconstructed the 11'th minute of *Bob Dylan's Chronicles: Volume One* an audio book read by Sean Penn [76] from patches of the same high quality (16kHz, 16 bit) taken from one, three, five and ten minutes of 4 audio books including the original one. The first segment was taken from *The Old man and sea* read by Donald Sutherland [77], *Stardust* read by Burt Reynolds [78] and lastly from *Fifty shades of Grey* read Becca Battoe [79]. This experiment isolates the effect of the noisy phase (since we take the correct phase), and puts the focus on our ability to find similar spectral amplitude patches in other speech signals that is on the self-similarity property.

A seen in Table 4.1 the signal was nicely reconstructed from itself, but with some degradation caused by the inability to find direct matches in the dictionary, degradation is decreased as dictionary size increases but reaches some level of saturation. Reconstruction form the voice of other readers reaches a level of coherence but does not match the levels of reconstruction from self extracted segments with a minimum of ~ 0.6 PESQ-MOS difference. Interestingly reconstruction ability is stable even when trying to reconstruct Sean Penn's voice from that of Becca Bettoe. This is interesting because the codebook built from Battoe's voice lacks segments containing Sean Penn's pitch ($85Hz$), the result contains as it seams somewhat of a phantom

Self-Similarity					
		Codebook Size			
		1 min	3 min	5 min	10 min
Source	Self	3.25	3.42	3.51	3.65
	The Old Man and The Sea	2.8	2.89	2.9	2.94
	Stardust	2.6	2.64	2.66	2.66
	50 Shades of Grey	2.68	2.83	2.87	2.89
	Diversified	2.89	2.99	3.03	3.08

Table 4.1: Self-similarity in high quality Speech: PESQ results of reconstruction of one minute of Bob Dylan's Chronicles from different sources.

pitch. Diversifying the sources of information, so that it simultaneously includes voices of different speakers improves performance as the dictionary is enriched.

4.3.3 Reconstructing Distorted Speech

The first one, three, five and ten minutes of Bob Dylan's Chronicles: Volume One an audio book read by Sean Penn [76] were used to construct a codebook. 16kHz mpeg files were scaled and normalized to reside in the $[-1, 1]$ range and have zero mean. It was then quantized using *round-to-nearest* method, which produces an output that is symmetric about zero at an appropriate linear quantization rate ($1-8$ bits). Both the original (high quality) and the quantized (low quality) signals were represented in the spectral domain through the use of STFT (32ms *Hanning* window with 4ms overlap resulting in a densely sampled spectrogram of size 257×15000 per minute. For the purpose of dictionary construction (to be used in the context of segment comparisons) we used the spectrogram of the quantized signal restricting our use to the bands between 75Hz and 4kHz. This was done both from memory and space consideration but was also justified by the fact that this spectral range carries most of the relevant information for speech intelligibility. The truncated spectrogram was scaled to balance energy components at different frequency bands. A single patch in the dictionary was created by concatenating 7 consecutive time frames in the truncated spectrogram resulting in a "patch" of $56ms$ symmetric around the focus of reconstruction and a super vector of length 875 (7×125). Frequency weighting

was done by taking the 0.1 power of each time-frequency component thus giving more weight to low energy (high frequency) components.

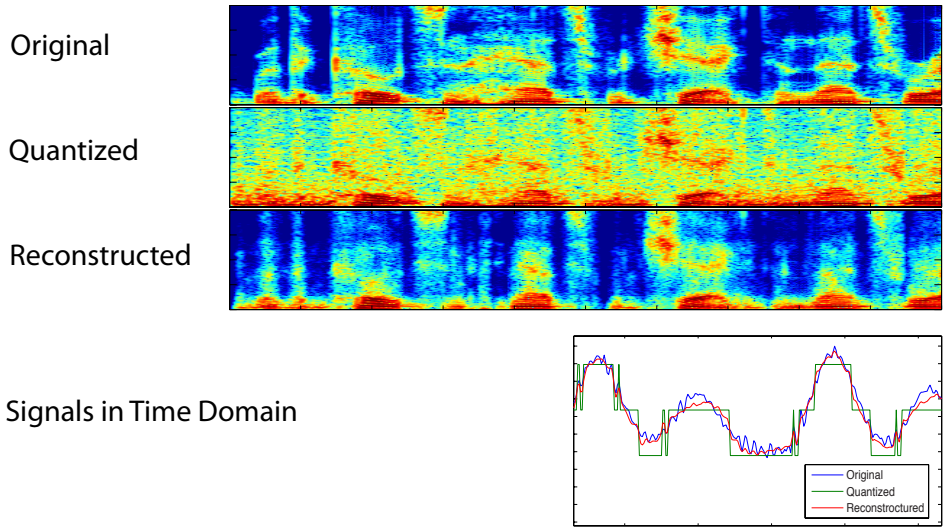

Figure 4.4: The effect of reconstruction: - Three signals: Original signal (upper panel), 4bit quantization signal (middle panel), reconstructed signal (bottom panel).

The overall evaluation of the results for the enhanced noisy input are presented in Figure 4.5 where the curve of the noisy (quantized) signal is given as reference. It can be seen that our technique consistently enhances the distorted signal with improvement of up to 0.86 PESQ MOS. Significant enhancement can already be achieved when using only one minute of codebook memory. Not very surprisingly, the algorithm continues to improve as we increase the size of the codebook from one minute to three and then to five and ten minutes. Improvement is dramatic at low bit rate but is consistent across all bit rates. The reader is invited to listen to some samples of degraded speech and their enhanced counterparts on Roi Kliper's Home Page where one can find in addition other examples.

We then used the 11'th minute of the audio book to evaluate our ability to reconstruct. We used a similar process to create super vectors of the distorted signals. Each super vector in the distorted signal was queried to the codebook: a query that resulted in its 3-nearest neighbors along with Euclidian distance scores for each neighbor. Each of the three neighbors was indexed to its high quality counterpart

and those high quality patches were reweighed according to their distances. The resulting patch was used as an estimate for the magnitude of the signal. The patches of the degraded signal were then concatenated (with overlap) and the signal was inverted to the time domain with the phase taken form the distorted version. An illustration of the effect in both time domain and spectral domain is given in Figure 4.4 and shows that to a great degree reconstruction is able to recapture the fine structure of the signal.

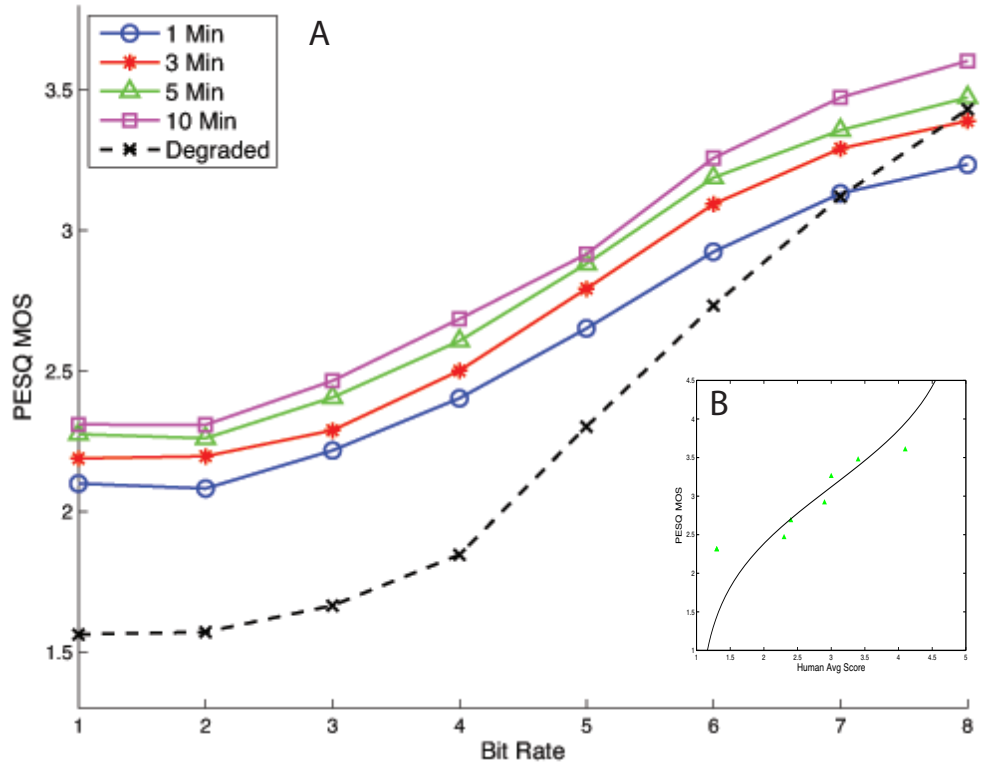

Figure 4.5: A. PESQ Results: - PESQ results of the reconstructed signal as evaluated in the different bit depth quantization. **B. Listening experiments:** MOS (n=10) results are given vs. the corresponding PESQ results. (PESQ to MOS curve (black line) is given as reference.

4.4 Learning the Retrieval Distance Function

A key component of our reconstruction scheme involves the retrieval of the relevant patches from the dictionary. As described earlier a naive choice (that works well)

would be to chose the spectral slice in the dictionary that is closest in the Euclidean sense to the queried slice. Up till now, (up to a scaling of higher frequencies) this has been our approach.

For our approach to work for a general (known) channel a general way to define the distance measure according to which one retrieves a patch from the dictionary would be to learn a distance function automatically. Given that the distances between the high quality patches are defined in some reasonable way (e.g. according to a weighted Euclidean distance measure), the adoption of our distance regression learning algorithm described in Chapter 3 would allow us to learn an adequate distance function and improve our retrieval scheme. That said, the algorithm, at its current form is limited by our choice of data-structure (KD-Tree) to handle Euclidean distances.

This limitations can be handled by changing the data structure to handle a more general distance function, but for the moment it directs us to limit the learning to the family of Mahalanobis distances, which can be interpreted as a linear transformation of the data and can thus be used in a framework that allows only Euclidean distances. Preliminary experiments with a 4 and 5 bit quantization signals show an additional improvement of ~ 0.2 PESQ-MOS in both scenarios highlighting the potential of this approach.

4.5 Discussion

In this study we have shown how one can reconstruct a distorted signal using previously acquired good quality samples of the same or similar signal. This situation, while to some extent idiosyncratic, is abundant in modern communication where stability of the channel and available bandwidth is constantly changing. The methodology we have proposed was demonstrated over an artificially generated dataset that has gone through coarse quantization and was shown to be robust especially for very low SNR; however this methodology can be served as general purpose reconstruction methodology across a wide range of settings and a wide range of signals. The specifics of the application to specific problem need to be determined according to the specifics of the setting and the assumptions one makes, and may require some

trial and error exploration to determine the appropriate parameters. Specifically, the assumptions one makes with regards to the nature of the noise and the type of degradation determine to a great degree the choice of an appropriate feature representation: the way one handles the preparation of the dictionary and the parameters used to create the spectral segments.

Complementary to that, one may employ machine learning techniques to estimate the signal degradation and optimize the representation accordingly. One promising path for improvement is that of learning a distance function. The described setting offers a natural opportunity for the employment of such algorithms, and specifically to the employment of distance regression algorithms. As we are interested in finding a "good" pairwise relationship on the patches of the degraded signal (To allow retrieval of appropriate a values) we can use the high quality patches to calculate the target distances. Finding a good distance function may render the search for a good representation obsolete and vice versa.

One challenge the algorithm poses for real time application is that of space and time management. We have shown that significant enhancement can already be achieved when using only one minute of "memory". We have also shown, not very surprisingly, that the algorithm continues to improve as we increase the size of the dictionary from one minute to three and then to five and ten minutes. However, increasing the size of the dictionary comes at a cost of both building the code-book (which is a non recurrent effort) but more importantly the price of search time. In this study we have used the KD-Tree as a data-structure; while generally a very efficient data structure, using it for real-time application is not an option due to prohibitive long search times. In future work we suggest using the *Locally sensitive hashing* (LSH) technique, that while has its flaws in terms of construction effort and optimality, offers significant improvement ($O(1)$) in retrieval time and thus allows for real time application.

5

A Model for Monaural and Binaural Sound Localization

"Brains are survival engines, not truth detectors."

— Peter Watts, Blindsight.

In this chapter we offer a physiologically derived model for sound localization. Using a simulation of natural environment, we show how by intercepting a signal at its end point (the eardrum), while maintaing model of its expected statistical properties, one can possibly reach interesting conclusions about the nature of the source, the medium in which it traversed and the locus of emission. This discussion highlights the information interchange that is played out in the battle between signal and noise where noise and signal compete for primacy and what is once called signal can simultaneously be called noise when the point of interest changes: that is, the "noise" introduced to the emitted signal along the way between the source and the target carries valuable pieces of information that may be crucial for efficient behavior in the environment.

Voice (or vocalization) is the sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box. It is generated by airflow from the lungs as the vocal folds are brought close together. When air is pushed past the vocal folds with sufficient pressure, the vocal folds vibrate and speech signals are created and are used for communication. From a signal processing point of view speech signals constitutes a unique family of signals that carries distinct statistical properties which make it different from other signals. This restrictiveness of statistical properties is an important feature of speech as it allows a listener to build a, possibly implicit, statistical model of the spoken language and interpret what is being heard even when the original signal has been corrupted or degraded. In addition a proper statistical model of the signal can be served to

deduce a hypothesis about the nature of distortion, and particularly the location of the source which in turn can be used to reinforce other human behaviors such as attention mechanisms and situation awareness among others.

The ability to perform speech localization is an important survival skill and an essential prerequisite for daily human communication. In addition, accurate speech localization is a fundamental building block for advanced speech processing that handles problems such as stream segregation, source separation, source enhancement and de-noising, ultimately improving speech intelligibility. The ability of humans and other animals to extract various types of information particularly the ability to pinpoint the location of sound sources from a received auditory signal is remarkably good. While humans with normal hearing have a surprisingly good ability to determine sound origins, individuals who use hearing aids often have only a very limited ability to localize sounds. Researching into the way humans localize sounds is therefore a worthy challenge.

In the following we present Amplitude Modulation Spectra (AMS) patterns (see Section 5.2.3) as efficient representation of speech signals to perform monaural localization. While the efficacy of this representation for speech recognition has already been demonstrated [80], we show that this representation can simultaneously capture location information included in a non-specific speech signal due to direction-dependent filtering by the human head and pinna. We show that the existence of this information allows monaural localization. We press this point as an argument for favoring directional microphones over omnidirectional ones. Furthermore, we propose and demonstrate the idea and framework in which the two ears are treated as two parallel processors, each processing monaural information and reaching a hypothesis about the location of a given source. In this framework, integration between the two ears is achieved, at the level of the decision rather than at the level of analysis, making the very restrictive demand of intensity and time of current binaural models redundant.

The rest of the chapter is organized as follows: in Section 5.1 we review different models of human sound localization as well as some physiological and psychoacoustic finding related to sound localization, localization cues, their representation and integration. In Section 5.2 we describe our model for monaural sound localization and binaural integration, as well as describe the experimental setting and the data

used. Section 5.3 continues with a description of the experiments and results. We conclude this chapter in Section 5.4 with a discussion of the findings and future work.

5.1 Models of Human Sound Localization

A sound wave generated by an external source is diffracted by the head and external ear (as well as other objects in the environment). The resulting changes in the temporal and intensity characteristics of the acoustical stimuli provide cues about the locus of the sound relative to the head. These localization cues have traditionally been divided into *Binaural cues* and *Monaural cues* and various studies have explored their relative efficacy in determining sound localization [81, 82]. Recent efforts have focused on understanding the integration of these different types of information, these efforts have spread from physiological research [83, 84, 85] through psychoacoustic research [82, 86] to application derived research [87, 88].

5.1.1 Binaural localization

Binaural hearing is a well established source of information for the localization of sound sources in space, it builds upon two distinctive properties of incoming sounds: interaural time differences (ITDs) and interaural level differences (ILDs). These differences arise from the fact that the two ears are separated by both space and an acoustically opaque mass (the head). Models of the processing of these cues build upon carefully constructed coincidence detectors, and are supported by both physiological and psychoacoustic findings.

While such models for sound localization are both elegant and robust, binaural cues are limited in their effective frequency band [82] and present a major challenge to the integrative capabilities of the nervous system requiring very high accuracy in highly structured labeled lines. Furthermore, to the extent that the head and ears are symmetrical, interaural differences should provide no cue to the vertical location of a sound source on the median plane nor would they contribute to the resolution of the front to back confusion. While acknowledging these limitations,

artificial system designers have favored models inspired by the binaural models for sound localization. Some challenges to the primacy of binaural localization cues have indeed been made (e.g. [89]); however, the role of monaural cues in sound localization is, to a great extent, only brought into focus in situations where binaural differences are nonexistent.

5.1.2 Monaural localization

For a single ear, the changes in temporal and intensity characteristics of the signal are generally described by head related impulse responses (HRIR), a generalization of which are Monaural Room Impulse Reposes (MRIR). This parsimonious description captures all the directional influence a signal may have received on its way to the ear drum. For example, to some abstraction the influence of the pinna is to produce multiple paths to the ear canal, among them a direct path and a reflection from the structure of the pinna. The addition of a direct path with a delayed path of the same signal produces a comb-filtered spectrum containing a characteristic structure of peaks and notches. The pinna thus acts as a direction-dependent filter which strongly affects the HRIR at high frequencies. Head related transfer function (HRTFs) are the spectral representation achieved through the Fourier transform of the HRIR. Figure 5.1 shows the different HRTFs as an illustration of the different spectral effects caused by different locations.

Several attempts to explain and exploit monaural cues for sound localization have been made. Generally speaking, these studies rely on the differential way in which the HRIR affects the spectrum of the input signal and require some interaction and/or comparison between different spectral bands of the signal. In this respect, these studies compel to the fact that the signal appearing at the eardrum has no reference point but itself. A second requirement for monaural localization is some statistical model of the source signal or equivalently some restriction to its statistical properties; directional sensitivity in itself cannot be exploited for localization of a general signal. If no assumptions on the nature of the signal are made, such a system is underdetermined and may result in ambiguous localization as each sound signal can be manipulated such that it is perceived as coming from all other locations. Assuming that the source to be localized is the statistically restricted set of speech

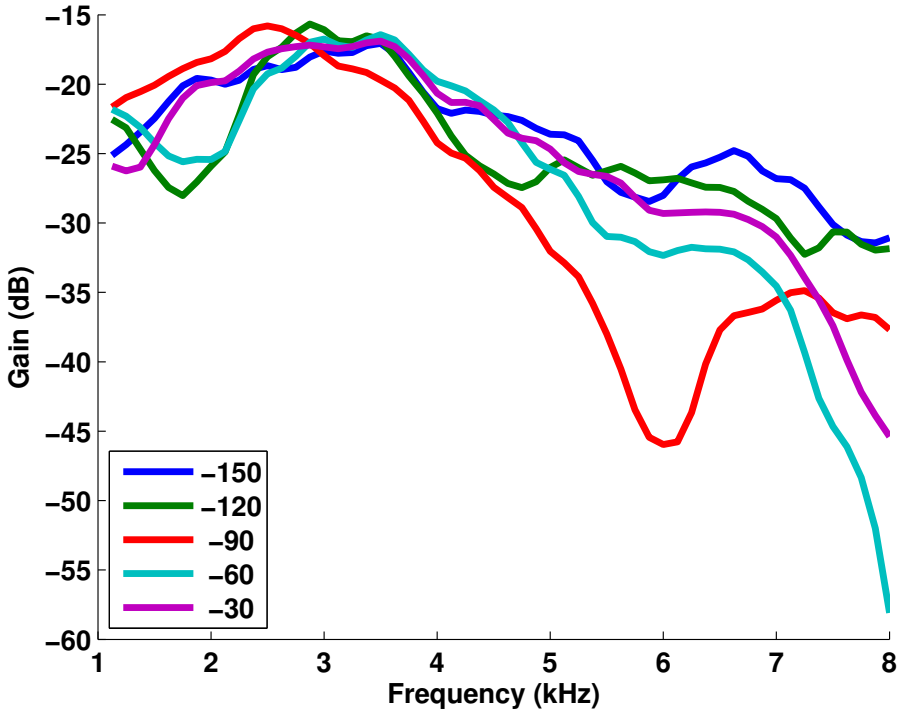

Figure 5.1: Head Related Transfer Functions (HRTFs) - The head related transfer function as captured by the left ear over the frontal horizontal plain, as recorded in an anechoic room. (For more details see Section 5.2.1.2)

signals is a step towards resolving this ambiguity. In monaural localization of sound, and more specifically in monaural localization of speech one should bear in mind that high intelligibility and accurate localization often represent competing requirements where intelligibility requires minimum distortion while localization requires direction dependent distortions.

5.1.3 Physiological and psychoacoustic findings

Psychoacoustic experiments have demonstrated monaural localization of a sound source on both the horizontal and the vertical planes [82]. In a recent paper Shub et al [90] have demonstrated the ability of human subjects to monaurally discriminate and classify different directions. Our experiments follow their paradigm and show that by employing a simple machine learning approach these results can be reproduced and surpassed (see Section 5.3.2).

Chase & Young [91] explored how different acoustic localization cues are coded in the inferior Colliculus (IC). The rationale was that the IC integrates binaural cues (ITD, ILD) and monaural cues (spectral information). Their results suggest that different cues converge to different degrees in different neurons: ITD and ILD are coded most strongly by spike rate while the spectral envelope of the signal is coded by the temporal pattern of the spikes. Localization in the vertical plane is often thought to be a purely monaural ability but recent psychophysical studies [92] have shown that both ears are used to determine the elevation, with the relative contribution of each ear varying with horizontal location. Our binaural experiment successfully implements this idea for localization in the horizontal plane (see Section 5.3.2).

5.2 Experimental setting

5.2.1 Data

Experiments were carried out using a well known speech database (see Section 5.2.1.1) and a database of HRIRs (see Section 5.2.1.2), these together provide a rich and reproducible setting with complete control over the experimental conditions.

5.2.1.1 Speech

Speech data was taken from the TIMIT Speech corpus [93] which provides recordings of 630 different speakers each reading ten phonetically rich sentences recorded at a sampling rate of 16 kHz. The segmentation into train and test data was adopted from the corpus. Both of the sets were further divided into direction specific subsets sampled randomly from the dataset and concatenated. No intersection between any of the subsets was allowed, resulting in 155 s train and 57 s test data for each of 37 directions.

5.2.1.2 Head-related impulse responses

A database of head-related impulse responses [94] was employed to introduce directional characteristics to the speech data. The database provides, among others, HRIRs (or MRIRs, respectively) measured on an head and torso simulator equipped with in-ear microphones under anechoic conditions. The azimuthal resolution is 5° covering the full azimuthal plane. MRIRs from an elevation angle of 0° (no elevation) and a radius of 3 m were taken. The initial delay contained in the MRIRs was removed and the remaining impulse response was cut to a length of 10 ms.

The audio data was convolved with the MRIR corresponding to a specific direction, resulting in monaural sound signals containing the input to either the left or the right ear. All data was scaled to unit variance to compensate for overall level differences between signals according to different directions. For experiments carried out in the presence of diffuse noise, pink random noise was generated artificially. The noise signal was convolved with each direction's MRIR from the full circle except for the one the target speech source was impinging from, thus approximating an isotropic noise field. After convolution, the noise was added to the directional speech source with the desired SNR. The coordinate system for the azimuth angles is relative to the center of the horizontal plane and is 0° in front of the head $-90°$ and 90° in front of the left and right ear.

5.2.2 Classification

Linear support vector machines (SVM [95]) were employed to conduct training of discriminative models using AMS features. The task consists of either binary classification, where a model was trained to discriminate between the directions of two speech sources, or multi-class classification, where a set of models was trained to estimate the absolute direction of an impinging speech source in a *1 vs. all* approach given a set of more than two directions.

5.2.3 AMS features and their extraction

Following the structure of the temporal envelope has shown to be crucial in human and machine recognition of speech. As a consequence, low modulation frequencies

were employed for several tasks in speech processing and acoustic scene analysis. Such features deliver a robust and generalized representation of speech signals and are known to be largely invariant to speaker and channel variations such as pitch and spectral distortions in the input signal. They showed to be a robust representation of speech even in challenging conditions, e.g. in speech detection experiments [96]. The features employed here are based on AMS [80] and are calculated as follows (cf. Figure 5.2):

The amplitude modulation spectrogram analyzes sound signals with respect to their modulation content by decomposing them into time, frequency and modulation-frequency components. It is computed by first extracting the spectral envelopes of the acoustic signal via a short-term Fourier Transformation (STFT) with a 32 ms Hamming window with a shift of 2 ms followed by computation of the log-energy. To extract modulation energy in each spectral band, another STFT is applied employing a 100 ms Hamming window with a shift of 50 ms.

Finally, the log-energy is computed again and the DC-component of the modulation spectrum, containing the acoustic frequency spectrum of the input signal, and the first acoustic frequency component of the resulting AMS pattern are removed to disregard spectral properties of the signal. Modulation frequencies above 100 Hz are also removed. The resulting AMS spectra for each time frame span 256 frequency bands from $32 - 8000$ Hz and modulation frequency bands covering the range from

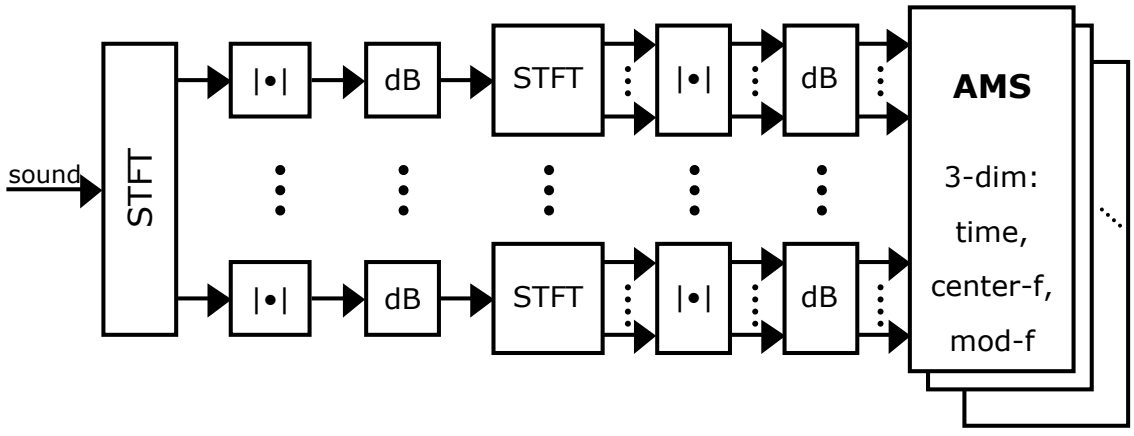

Figure 5.2: AMS calculation flow: Extraction of amplitude modulation features generates a 3-dimensional representation of the input. For each 100 ms signal segment a 256×9 pattern is extracted, overlap of neighboring patterns is 50 ms.

20 Hz to 100 Hz with a resolution of 10 Hz. The employed range patterns lie above modulation frequencies used in speech recognition and detection (typically ≤ 16 Hz). In Figure 5.3 one can observe the nature of this representation as well as the evident notch that appears when speech arives form -90^o around frequency 6 kHz.

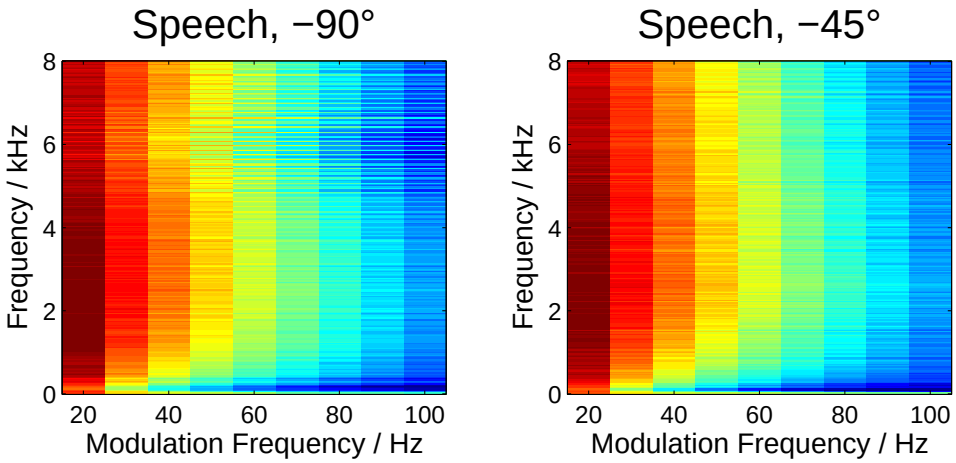

Figure 5.3: AMS Patterns - Two AMS Patterns are displayed as averaged on 800 ms period.

5.3 Experiments and results

In the following we first present experiments of a discrimination task which is a popular task in psychoacoustics used to evaluate minimum audible difference. We then show results of multi-direction classification experiments under different noise conditions. Finally, results of a straightforward *winner takes all* formalism of binaural integration are presented.

5.3.1 Minimum audible angle difference

In this experiment a set of models for binary classification was trained to discriminate between two sound sources s_1 and s_2 impinging from different directions. s_1 (target) was located at positions from 0° to −180° in steps of −60° while the angular position of s_2 (reference) is varied with a resolution of 5° on the same half circle. The results are shown in Figure 5.4.

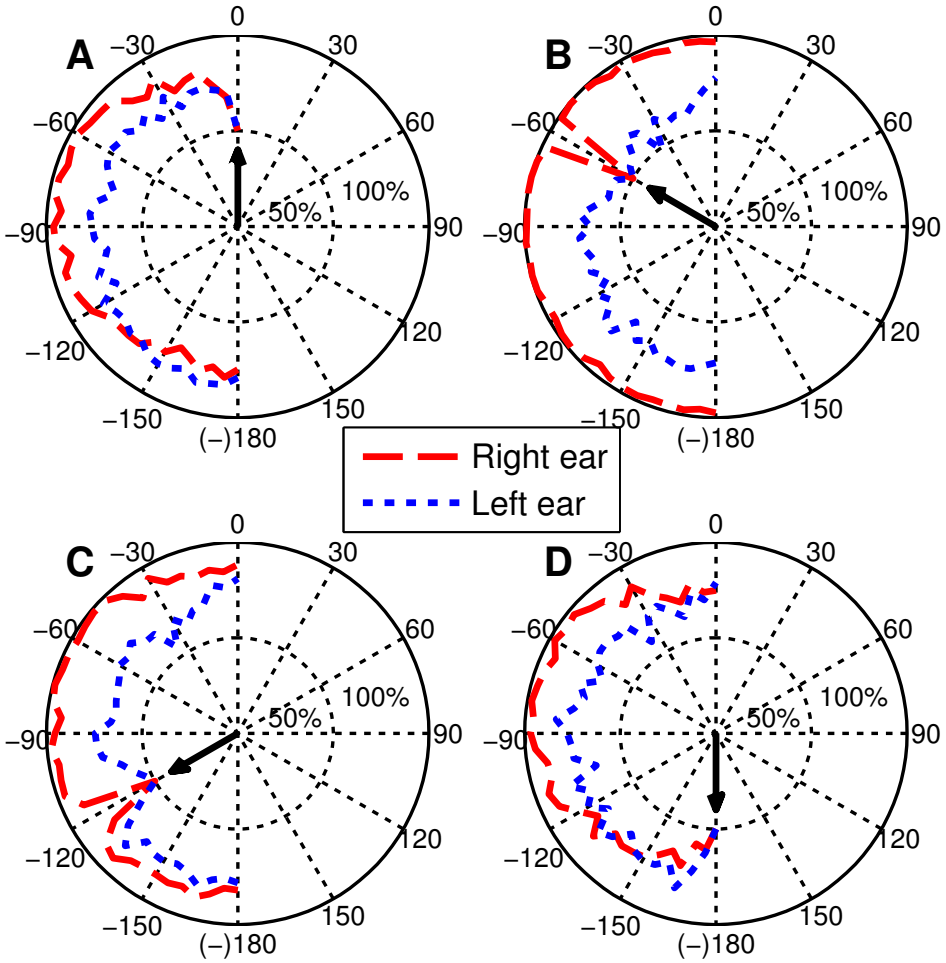

Figure 5.4: Binary classification: - The accuracies gained by the right (contralateral) ear are shown in dashed red and by the left (ipsilateral) ear are shown in dotted blue. Four target signals where classified against a reference signal drawn one at a time from the left hemisphere at a 5° resolution. Performance in terms of testing accuracy is shown on the radial axis. Position of reference signals are shown on the circular axis. Target signal is pointed by a black arrow.

The superiority of the contralateral monaural cues over the ipsilateral can be read from the consistent improved accuracy of the right ear on the left hemisphere (red line above blue line in e.g. Figure 5.4 B). Contralateral cues achieve a nearly perfect performance discriminating the target from references around it at a resolution of 5°. This makes sense as heavier distortion has been introduced in the contralateral case thus allowing better directional sensitivity. We note that, as we normalized the

energy at the eardrum, our results cannot be explained away using level differences (which will appear in real scenarios). One can also note (1) The appearance of moderate front to back confusion in the cases where s_1 is located exactly in front or behind the listener (see Figure 5.4 A & D). (2) Increased performance in the frontal half space (compare Figure 5.4 B & C). Similar discrimination characteristics can be found in psychoacoustic monaural localization experiments carried out in [90]; however, although their experiment was done with stimuli ten times longer, our results surpass theirs.

5.3.2 Localization of sound sources in the presence of noise and the integration of both ears

In the multi-class classification experiment we first evaluated the influence of averaging the features on classification performance. For that, a sliding average over time was applied to the AMS patterns before models were trained. Averaging was conducted with a sliding window of length ranging from 0 s (no averaging) to 5 s in steps of 0.25 s. Training and testing was conducted without noise (infinite SNR). Furthermore robustness against diffuse pink noise was investigated: noise was introduced as described in Section 5.2.1 with an SNR varying from 10 dB to 60 dB in steps of 10 dB.

Figure 5.5 displays the results of this experiment for 12 directions distributed equally spaced over the complete horizontal plane from $-180°$ to $150°$. The accuracies gained by the left and the right ear independently in addition to the performance achieved by the integration of information from both ears are shown. For each ear, one model was trained and tested on one condition given by the SNR and length of averaging. Different noise conditions appear in Figure 5.5 A-G and show a clear dependency on noise level ranging from guessing probability (8.3%) at SNR 10 dB to 56.5% and 56.2% for the single ears and 76.5% for the combination of two ears where no noise is present (inf). Average accuracy dependent on the averaging length is read against the x-axis in each figure and shows a global maximum at around 1 s suggesting that longer averaging windows are erasing localization information. Integration of the information from both ears was done in a *winner takes all* manner

where the more confident ear was taken as the reliable one. Confidence was assumed to be positively correlated to the margin from the model's separating hyperplane. Integration of the two ears achieves up to 20% absolute boost in accuracy.

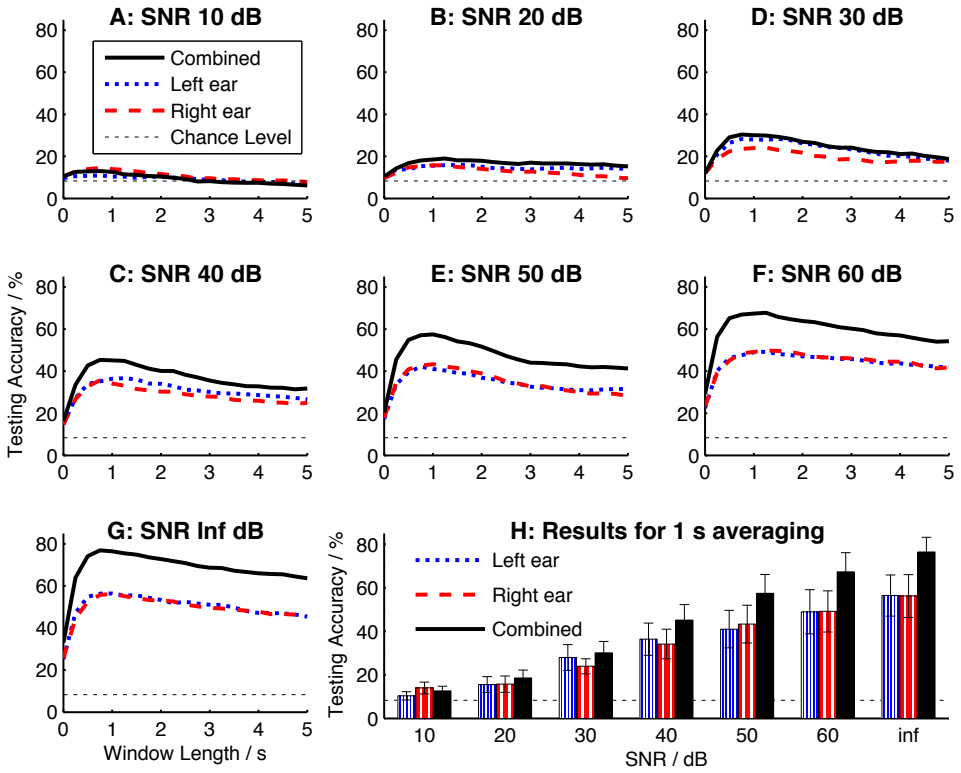

Figure 5.5: Multi-class classification at different SNRs: - A-G: Performance of the localization on 12 directions for different SNRs and averaging times of the features. Blue and red lines denote the accuracy achieved by the single ears, black the performance of the integrated approach. H: Average and standard error of accuracy for each SNR.

Besides the overall average testing accuracy the confusion between different directions is a significant figure of merit. The confusion matrices obtained for an averaging length of 1 s in the clean condition from the left and the right ear and from the integration of both ears are shown in Figure 5.6. Clear preference for the contralateral ear can be read in Figure 5.6 A & B. The ipsilateral ear regains significant classification capabilities only when the angular distance is larger than 60° (allowing directional cues to come into effect). Integration of the information from the two ears solves most of the confusion, leaving a slightly increased confusion around the center supposably due to noise in the confidence of the classifiers.

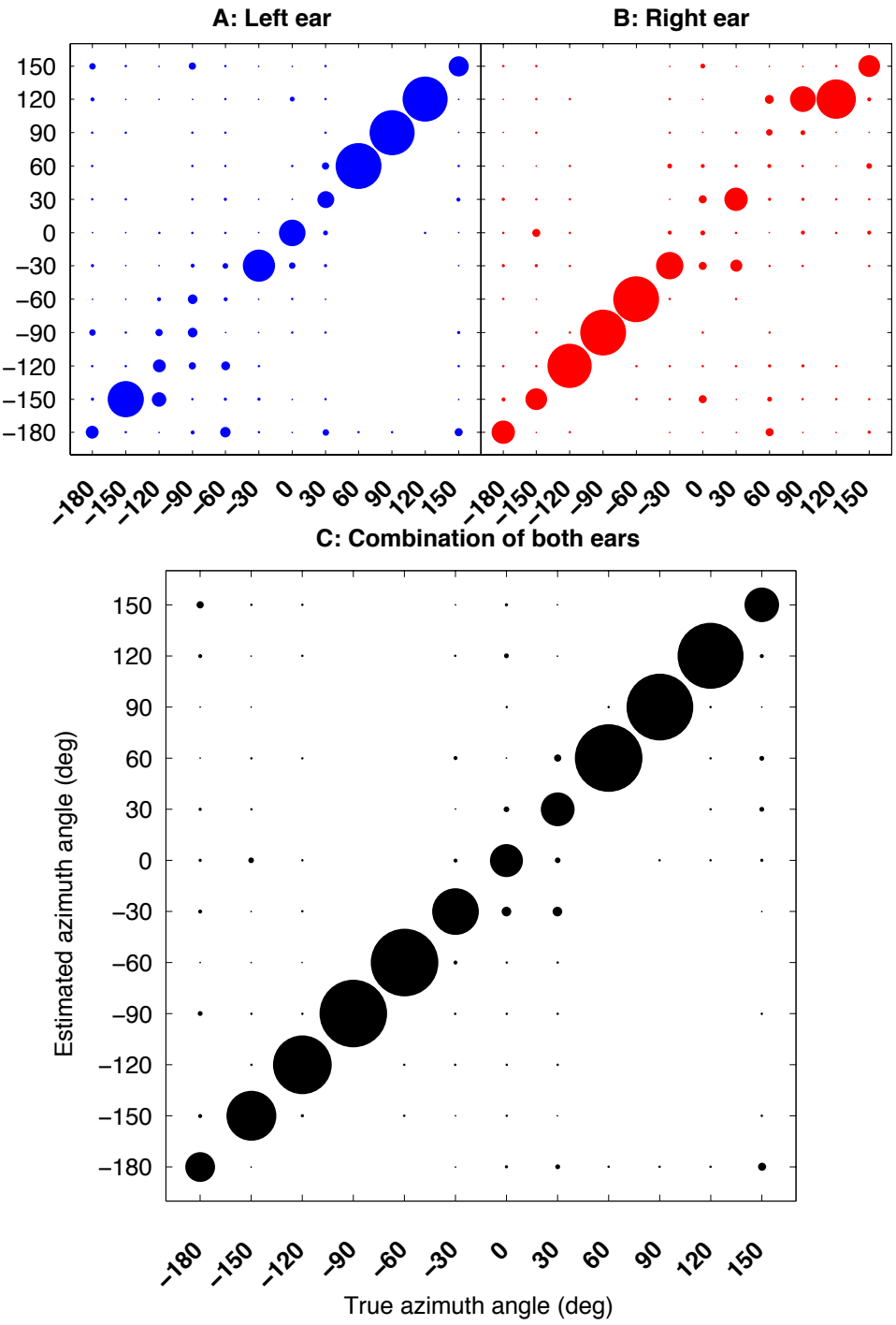

Figure 5.6: Confusion matrix for multi-direction classification: - Confusion matrices for single ears (upper row) and after integration of both ears for the clean condition and averaging window length of 1 s. Coincidence of target and reference is given by the radius of the circle.

5.4 Discussion and Conclusions

We have demonstrated that amplitude modulation spectra patterns in modulation frequencies above those commonly used for other speech applications are efficient in monaural speech localization. Our results suggest that under clean and moderate noise conditions, accurate speech localization can be achieved using the information obtained by a single ear, without distorting intelligibility related information. The experiments showed a maximal performance for an integration time of around 1 s, which corresponds to the choice of parameters in monaural localization experiments conducted by Shub et al. Keeping in mind that the analysis of the stimuli was done using a rather technical approach, further investigation incorporating auditory models as a preprocessing stage is a natural step. While monaural localization was proven to be feasible, the integration of information processed parallel in the two ears is highly beneficial as each ear performs better on the contralateral hemisphere. Relating to physiological research we hypothesize that the IC may host such an integration mechanism.

This success in speech source localization highlights the potential use of the described method in the context of other speech processing applications such as source separation, signal to noise ratio estimation, and noise reduction. Drawing to the design of artificial systems, the implementation of such an approach in ear-worn hearing assistive systems is an interesting application as it does not necessarily demand for a technically costly cross-linking in the case of a bilateral means.

6

Prosodic Analysis of Speech and the Underlying Mental State

"The best listeners listen between the lines."

— Nina Malkin, Swoon

Speech is a measurable behavior that can be used as a biomarker for various mental states including schizophrenia and depression. In this chapter, we show that simple temporal domain features, extracted from conversational speech, may highlight alterations in acoustic characteristics that are manifested in changes in speech prosody - these changes may, in turn, indicate an underlying mental condition. We have developed automatic computational tools for the monitoring of pathological mental states - including characterization, detection, and classification. We show that some features strongly correlate with perceptual diagnostic evaluation scales of both schizophrenia and depression, suggesting the contribution of such acoustic speech properties to the perception of an apparent mental condition. We further show that one can use these temporal domain features to automatically and correctly classify up to 87.5% and up to 69.75% of the speakers in a two-way and in a three-way classification tasks respectively.

6.1 Speech as a Biomarker of an Underlying Mental State

Psychiatry is a medical discipline in search of objective and clinically applicable assessment and monitoring tools. The acoustic characteristics of speech are a measurable behavior, hence can be used in the assessment and monitoring of disorders

such as schizophrenia and depression. This observation has not gone unnoticed in the psychiatric community and previous attempts to quantify this acoustic effect is the psychiatric setting have been made. However, these attempts have been limited, in part by technical and technological limitations. Recent technological advancement has made the recording, storage and analysis of speech an available option for both researchers and practitioners.

The use of acoustic characteristics of speech in the description of pathological voice qualities has been studied in various contexts and with a variety of goals including mental health evaluation [97, 98]. Studies have correlated acoustic features with perceptual qualities [99] and to a lesser extent with physiologic conditions at the glottis [100, 101]. These studies looked at the use of syntactic structures, richness of vocabulary, time to respond and many other qualities [97, 99, 102, 103, 104, 105].

Speech prosody is the component of speech that refers to the way words are spoken. It includes the rhythm, stress, and intonation of speech. Prosody may reflect various features of the speaker or the utterance: the emotional state of the speaker; the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or choice of vocabulary. Changes in the acoustic characteristics of speech prosody in the course of mental disorders, notably depression and schizophrenia, are a well documented phenomenon [97, 99, 104, 105], and the evaluation of aspects of speech constitutes, today, a standard part of the mental status examination.

The acoustic changes in schizophrenia patients' speech are a well recognized clinical phenomenon [106, 107]. In fact, references to the acoustic properties of schizophrenia subjects' speech have been made as early as 1913: Kraepelin, in his seminal work, described dementia preacox subjetcs as having "indifferent tone" and "distorted turns of speech" [108]. These speech changes are currently conceptualized as a component of the negative symptoms [109]. The most accepted scale for negative symptoms is the Scale for the Assessment of Negative Symptoms (SANS) [110]. Negative symptoms are divided into five domains including blunted affect, alogia, asociality, anhedonia, and avolition [111], where speech acoustic changes are especially reflected in two different domains - blunt affect (diminished expression of emotion) and alogia (poverty of speech).

Similarly, speech in depression patients has distinct acoustic patterns, reflected in its own clinical rating scales such as Montgomery and Absberg Depression Rating Scale (MADRS) [112] and the Hamilton Depression Scale (Ham-D) [113]. The recognition of these changes can also be dated back at least to the time of Kraepelin [108]. Today, speech in depression patients is generally described as having smaller pitch variability, more pauses while speaking, and slower speech production [114, 115, 116, 117]. Distinguishing negative symptoms from depressive symptoms in schizophrenia subjects is an important clinical challenge which existing acoustic research was unable to meet.

Speech prosody is currently measured by subjective rating scales requiring highly trained staff. The results of these ratings are summarized by ordinal variables which may be insensitive to subtle changes over time and are not ideal for statistical analysis. Also, the measured variables tend to be skewed and restricted in range in comparison to computerized acoustic measures, which may lead to rating bias in the global clinical impression. [99, 105, 118]

Several attempts at using speech cues for the automatic quantification of specific mental effects have been made in the past [99, 104, 119, 120, 121]. These attempts have made an effort to first correlate specific acoustic measures to their perceptual (clinical) counterparts, and second to quantify and asses different aspects of subjects' speech using different acoustic measures [105, 122]. The advantage of automatic quantification of effects apparent in different mental states has been highlighted as early as 1938 [123] and is very well outlined in [124].

Acoustic analysis can be performed using either fully computational methods, where no human involvement is required in the analysis process (e.g. [99]), or partly computational signal decoding where human observers are involved in the processing, sometime in an iterative manner with a computerized system (e.g. [115, 125]). The fully computational decoding methods have the advantage of objectivity and processing speed allowing real time evaluation, but they also pose major challenges including, but not limited to, voice activity detection, signal quality assessment, speaker separation and identification, and more. Partially computational settings handle some of these issues by referring to the help of human observers, and consequently are also able to offer other advantages such as referring to the content of the word.

In [121] the researchers propose lexical analysis as a measure of mental deficits; but while some success has been shown, it has been claimed and shown that significant aspects of speech are missed when focusing on the lexical level [126]. In a previous study [104] the researchers have extracted several acoustic features from both structured speech and semi structured interview of depression subjects. The study shows that depressed patients tend to be less fluent, exhibiting what the authors term reduced prosody in both emphasis and inflection when participating in a semi-structured interview. Results of the automated speaking tasks also show a difference between depressed patient and the control group, but are incomparable to the semi-structured task since idiosyncratic features have been measured. A later study [105] showes high correlations between speech fluency measures and basic prosody measures (mainly inflection) and clinical ratings of negative symptoms of schizophrenia. In these and other studies results are highly task specific, a fact that makes any conclusion with regards to speech generation mechanisms and / or pathogenetic mechanisms very hard to establish. This has not prevent researchers from hypothesizing about the existence of such source [127, 128]. Lastly, in a recent study [99] the authors have focused their efforts on the characterization of negative symptoms of schizophrenia and identified *Inflection* and *speech rate* as discriminative features between schizophrenia subjects and control. The authors highlight the general advantages of computerized measures over traditional clinical-based measures from the psychometric point of view. The authors also outline a clear path for the incorporation of computerized measures into the global clinical impression.

Speech samples can be obtained from either structured tasks, such as passage reading [129], counting [104, 130] or short monologues [128] or by non structured tasks such as interviews or dialogues [127, 131]. A structured task may enhance the assessment accuracy and replicability of the test but is less clinically applicable than an analysis based on conversational speech. Analyzing a conversation allows one to asses the mental state of the conversing subjects ”en passant” without recruiting the patient to perform special tasks or tests. Hence, while much more challenging, analysis of speech during the clinical interview was chosen in hope to facilitate the integration of this tool in the clinical practice.

Developing automatic computational tools required two corresponding research efforts. First, study the physical signal properties specific to schizophrenia and de-

pression. Second, develop an automatic, real-time, reliable and objective assessment of these properties. Referring to the first task, we chose to focus our efforts on simple temporal domain features. These features, while not easy to extract or measure, provide a meaningful interpretation and may be referenced in the context of existing clinical evaluation scales. Referring to the second task we adopted a discriminative approach and trained a Support Vector Machine (SVM) classifier over the data. These family of discriminative models have proven its robustness in a variety of tasks and offers in addition to classification a probabilistic interpretation.

The remainder of the chapter is organized as follows: in Section 6.2 we describe in detail the subject population, the experimental setting, and the signal processing methods we used as well as the statistical analysis we have conducted to establish the significance of the results. This is followed in Section 6.3 by a description of the results. We conclude in Section 6.4 with a discussion of the results, their significance and a few words regarding future research.

6.2 Materials and Methods:

6.2.1 Subjects

62 subjects participated in the study, giving written consent approved by McLean Hospital Institutional review board. The study subjects (described in Table 6.1) comprised of three groups, including schizophrenia patients (n=22, 13 male, 9 female), patients with clinical depression (n=20, 9 male, 11 female), and healthy participants (n=20, 10 male, 10 female). The subjects were matched according to age ($mean = 39.98, std = 11.37, p = 0.8489$) years of education ($mean = 14.8, std = 2.3, p = 0.063$) , and gender ($\chi^2$ test of Independence, $q = 0.86$, $dof = 3$, $p = 0.65$). Subjects from all races were included Few of the subjects were recorded in two separate occasions; however, for the purpose of this study, unless stated otherwise, the two recordings were concatenated and treated as a single recording which eventually comprised a single vector of measures.

Table 6.1: Subjects' description: Mean and standard deviation

Subjects									
	Schizophrenia			Depression			Healthy		
	N	Age (Y)	Education (Y)	N	Age (Y)	Education (Y)	N	Age (Y)	Education (Y)
Male	13	(36.46,9.62)	(14.15,1.67)	9	(39.22,12.26)	(14.22,2.39)	10	(37.3,11.13)	(14.9,1.96)
Female	9	(43.67,8.47)	(13.67,2.65)	11	(42.82,12.29)	(15.55,2.11)	10	(41.5,14.53)	(16.3,2.58)
All	22	(39.41 ,9.66)	(13.95,2.08)	20	(41.2,12.09)	14.95,2.28	20	(39.4,12.78)	(15.6,2.35)

The schizophrenia subjects were previously hospitalized in a closed psychiatric unit and participated in a prior study which included Structured Clinical Interview for DSM-IV (SCID-IV) [132] to verify the diagnosis. They were later discharged from the hospital and were considered clinically stable at least 6 months before being interviewed for the current study. The depression subjects were recruited during hospitalization in a closed psychiatric unit. The healthy subjects were recruited by advertisement.

All subjects were interviewed and recorded by an experienced psychiatrist. Inclusion criteria included ages between 18 and 65 and English as first language. Exclusion criteria included significant medical or neurological conditions other then those of interest (e.g. traumatic brain injury), mental retardation, and hearing or speech impairments.

6.2.2 Clinically rated symptom measures

The subjects completed a clinical interview which included Semi structured Clinical Interview for DSM-IV (SCID IV) [132] , Positive and negative Syndrome Scale for Schizophrenia (PANSS) [133], Scale for the Assessment of Negative Symptoms (SANS) [110], and the Montgomery and Absberg Depression Rating Scale (MADRS) [112] as well as the Hamilton Depression Scale (Ham-D) [113].

6.2.3 Acoustic Recordings

The recordings were made by a headset without sound isolation or calibration. To prepare the recordings for acoustic analysis, the audio tapes were digitized at a 44.1 kHz sampling rate. Acoustic analysis was conducted using MATLAB [134] (details are given in 6.2.4). Average length of a clinical interview was: schizophrenia: 57m

13s, depression: 30m 31s, healthy: 48m 46s. Silence was automatically removed at the beginning and end of the recording and the remaining data was normalized to have 0 mean and 1 variance thus avoiding effects caused by constellation of headset. To enable efficient handling of the data each interview was divided into 2 minutes segments, which were subsequently analyzed independently. All results from a single person's 2 minutes segments were later used together for classification.

6.2.4 Speech Prosody and Feature Extraction

Two minutes segments of interview were used to measure nine diagnostic features. This research focuses on a very small and simple set of features extracted in the temporal domain. This choice was motivated by the fact that temporal domain features are, in general, easier to relate to perceptual properties of speech and thus provide better infrastructure for further use in the psychiatric community. For example, the physical property termed (lack of) pitch variability is perceived at least in part as (lack of) inflections, and the physical property of power variability is perceived at least in part as emphasis or lack of emphasis [135]. Thus, though spectral features have indeed been proven beneficial in various speech analysis tasks e.g. [136, 137], the examination of the relevance of these features was left for future research.

Alterations of the speech signal in abnormal conditions can occur at different time-scale levels, including the *macro-scale* level (above 1 sec) which refers to variables such as speaking rate, the *meso-scale* (25 ms to 1 sec), in which variables like pitch and its statistics are measured, and finally, the *micro-scale* (10 ms or less) level in which cycle to cycle measures are taken and is considered as contributing to the naturalness of the speech sound. While *macro* and *meso* scales are influenced by voluntary aspects of speech, the *micro-scale* is involuntary in nature and, thus, can better serve as a reliable biomarker. All scales contribute to the prosodic structure of speech signal and using them as an ensemble may provide insight into the possible role of prosody in the characterization of pathological mental states. The focus on prosodic features follows reports showing the relevance of meso-scale and micro-scale level to the tasks of mental evaluation [97, 99, 101] and emotion detection [138, 139].

An adaptation of an energy based Voice Activity Detection (VAD) algorithm [140] was used to segment the signal. This algorithm determines the presence or absence of speech by comparing the observed signal statistics in the current frame with the estimated noise statistics according to an adaptive decision rule. An initial (global) decision is modified by a hangover scheme to minimize mis-detections at weak speech tails. Furthermore, we have adapted the algorithm so as to avoid speech detection of the interviewer voice, by integrating an automatic optimization phase that determines the proper thresholds and leaves out speech parts that are not of the person of interest (see Figure 6.1). Removing the interviewer's speech from analysis proved to be a tricky task as automatic voice detection tends to ignore amplitude and focus on fine structure. Our approach to the problem was very conservative: any segment that was suspected as being uttered by the interviewer was discarded form the analysis. We have sampled and listened to several segments and found this procedure to be fairly accurate.

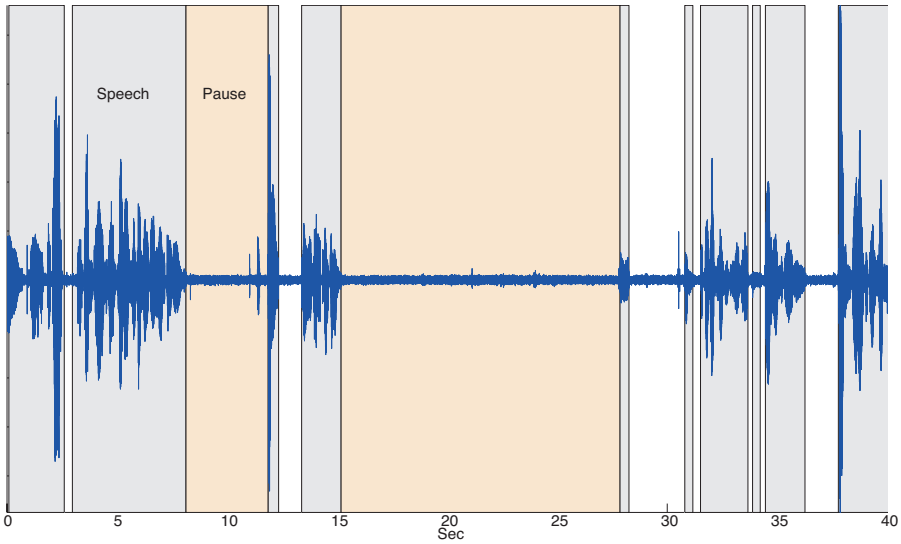

Figure 6.1: An example of speech signal segmentation: - 40 seconds of speech were segmented into *1*) Interviewee Utterances (Gray). *2*) Interviewer Interventions (White). and *3*) Gaps (Peach). Segments that contained interviewer interventions were later discarded of analysis.

From this segmentation the following utterance and gap statistics were extracted: *1*) Mean utterance duration. *2*) Mean gap duration. *3*) Speaking ratio. (See below for a more detailed explanation.) These measures were termed by others fluency

features [104, 127] and are considered key features for various detection tasks. They have been shown elsewhere to be highly influenced by one's mental state; however, they tend to be very task dependent [116] and also, in a dialogue setting, may reflect speaking properties of the interviewer rather than the interviewee [141].

To extract the meso-scale level features we have used the YIN pitch detection algorithm [142]. The algorithm was used to detect pitched segments (among uttered segments) and to further segment the uttered segments into Voiced-Unvoiced segments. This was followed by the extraction of the following power and pitch related statistics: *4*) Pitch range. *5*) Standard deviation of pitch. *6*) Standard deviation of power. (See below for a more detailed explanation.) Pitched segments were then used to refer to the finer resolution of 'cycle-to-cycle' variation and extract micro-scale statistics. For each identified iso-pitched segment (the free speech equivalent of sustained vowels) the following measures were obtained: *7*) Mean wave form correlation (MWC). *8*) Mean jitter. *9*) Mean shimmer. (See below for a more detailed explanation.)

Macro scale measures: Mean utterance duration, Mean gap duration, Mean spoken ratio

An utterance was considered any segment identified as speech that exceeded 0.5 seconds. Segments of shorter length are rare and were generally observed in preliminary experiments to be misclassification of speech (laugh, cough etc...); they were thus discarded from the analysis. A "gap" was considered any segment of recording with no subjects' speech. The attempt was to study the amount of gaps that occur in a natural stream of a subject's speech, thus neutralizing as much as possible the role of the interviewer. Such a measure may possibly reflect the coherence form of thought of the subject. Gaps that exceeded 3 seconds were ignored as again they were rare and were generally associated with an intervention of the interviewer in the stream of speech. *Spoken ratio* was calculated as the ratio between the total volume of speech occupied by the patient, that is the sum of the length of all utterances divided by the total conversation length. It is believed to be reflective of the subject's initiative and willingness to engage in conversation; however it also may

reflect, to some degree, the conversational behavior of the interviewer rather than the interviewee [141].

Meso scale measures: Pitch Range, Pitch standard deviation, Power standard deviation

Pitch range was calculated as the difference between maximum estimated pitch and minimum estimated pitch normalized by the mean pitch over the entire 2 minutes segments. No significant between group differences were observed for mean pitch, which was (Males: 112.3Hz, 18.99Hz; Females: 176.39Hz, 18.53Hz) ($F = 0.18, df = 2, 56, p = 0.83$) nor for interaction with gender ($F = 0.75, df = 2, 56, p = 0.47$). Between gender differences were strong as expected ($F = 182.9, df = 1, 56, p << 0.01$). Such an estimate gives a pitch range feature as percentage of the mean pitch. 2.5% of pitch estimation points were discarded on each side of pitch estimation to increase stability of this measure and to avoid outliers.

Standard deviation (STD) of pitch was calculated for each utterance and was then averaged for each speaker. STD of pitch was again normalized by the mean pitch, but this time normalization was done for each utterance. *STD of pitch* is, thus, given as relative to the current utterance and then averaged over all utterances. STD of pitch gives the dynamics of an uttered sentence and is expected to depend and correlate to some degree with the length of an utterance.

An attempt to measure *emphasis*, a key component of blunt effect, was not made in [99] due to what the authors termed 'variability in recording conditions'. Accurate measurement of power and its related statistics is indeed a challenging task in real world applications, as it is highly effected by small changes in equipment and subject configuration. As mentioned before, we normalized all spoken segments to have approximately the same mean power. This was done by dividing the signal by the power estimate based on the first 5 seconds, thus, controlling some of the between segments variability in that measure.

Mean active power and its variance were measured in decibels (dB) in reference to a calibration level and were 64 dB, with a standard deviation of 8.6 dB. A trend was observed for a group difference in mean power ($F = 2.99, df = 2, 56, p = 0.06$). No main effect was observed for gender ($F = 0.02, df = 2, 56, p = 0.87$)

nor for interaction of group with gender ($F = 1.12, df = 2, 56, p = 0.33$). The mean dynamic range measured from voiced segments within the 120 second excerpts was approximately 29 dB. Standard deviation of power within an utterance was measured, as normalized by the mean power in the entire segment in order to avoid effects caused by noise in the location of the microphone. This gives a notion of the 'local' emphasis used by the speaker in a given utterance. Results are given in units of standard deviation (z-score). An example of the processing done on each utterance is given in Figure 6.2.

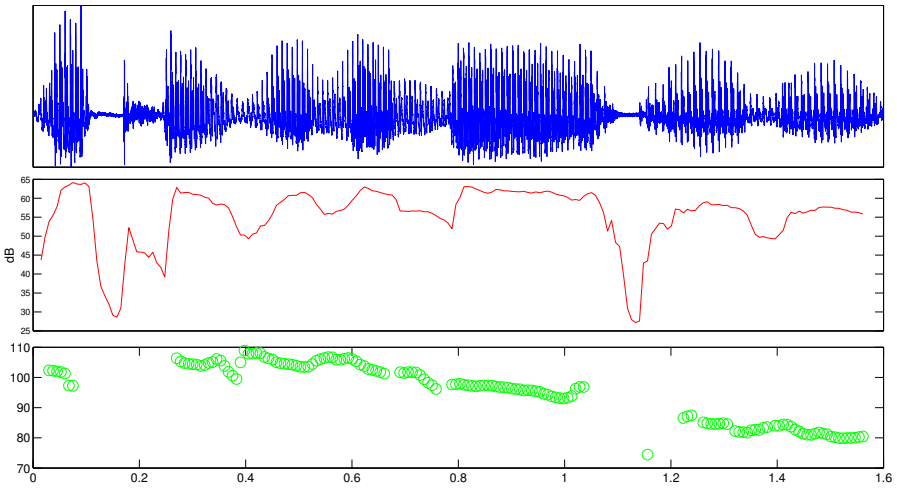

Figure 6.2: An example of a 1.6 **seconds utterance** - Transcription: "I came home one day and am...". Speech signal is given in upper panel, power trace is given in the middle panel and pitch estimation is given in bottom panel. This estimation served to extract meso-scale prosodic features e.g. standard deviation of power, pitch range, pitch standard deviation

Micro scale measures: Mean waveform correlation, Mean jitter, Mean shimmer

The mean of all correlation coefficients evaluated for every pair of consecutive periods was used as the acoustic measure termed *Mean Waveform Correlation (MWC)*. It indicates the overall similarity between the cycles of the time signal. When applied to pitched segments it measures the level at which the speaker sustained its constant pitch. The evaluation of *MWC* is relatively robust compared to jitter and shimmer calculation (see below). Its upper limit of 1 is reached for signals with identical

periodic shapes. MWC Decreases with increasing differences in length or shape between consecutive periods. i.e. with short term variations of the time signal.

Following [97] jitter and shimmer were calculated as the period perturbation quotient *PPQ* and the energy perturbation quotient *EPQ* respectively where the locality parameter was chosen to be 5. Perturbation Quotient (PQ) measures the local deviation from stationarity of a given measure and is defined in Eq. 6.1. It measures the ratio of deviation of a given measure in a local neighborhood defined by the locality parameter K. Put in simple terms, the jitter measured the stability of the period in a 5-local cycles environment and the shimmer measured the stability of the energy in a given 5-local cycles environment.

$$PQ = \frac{100\%}{N-K} \sum_{\nu=\frac{K-1}{2}}^{N-\frac{K-1}{2}-1} \left| \frac{u(v) - \frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} u\left(v+k\right)}{\frac{1}{K}\sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} u\left(v+k\right)} \right| \tag{6.1}$$

We chose to use the energy shimmer since it is expected to be considerably less susceptible to noise than the amplitude shimmer often used in acoustic analysis. An example of 5 cycles is shown in Figure 6.3: an iso-pitched segment is given in Figure 6.3a, 5 cycles are shown in Figure 6.3b, and the same five cycles aligned are shown in Figure 6.3c. One can observe the relative punctuation of the pitch period as well as the existence of an evident energy noise (shimmer).

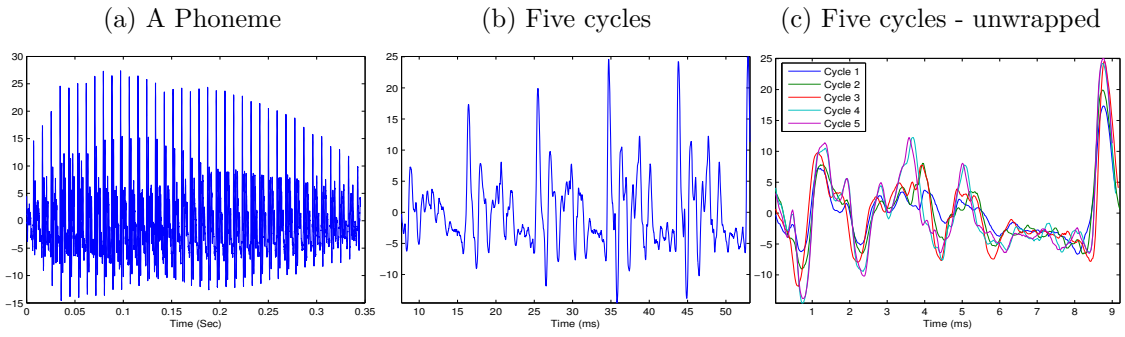

Figure 6.3: **(a)** A Phoneme, **(b)** first five cycles in an iso-pitched segment, **(c)** same first five cycles aligned.

6.2.5 Classification

We used the extracted acoustic features and a basic linear classifier [143] to classify the different conditions: Healthy (HL), Schizophrenia (SZ) and Depression (DP) in a two-way classification task. For each classification scenario (e.g. HL vs. SZ), one subject was left out for testing and the rest were used to train a classifier. In a single classification setting all 2 minutes segments of a given speaker were left out and later used for testing. All the left out segments were used with a majority vote decision rule so are in fact a averaged decision over all 2 minutes segments.

6.2.6 Statistical Analysis

To check the statistical significance of the results over each of the individual features, when the distribution was roughly normaly distributed we used 1-way ANOVA, otherwise we have used the nonparametric version of the 1-way ANOVA, called the Kruskal-Wallis test (which many call a "test of whether or not the medians of > 2 groups are the same"). It is actually a more general test, in that it is comparing distributions rather than medians. Checking the statistical significance of the effects brings up the problem of multiplicity (multiple comparisons). We consider the problem of testing simultaneously 9 null hypotheses where within each are nested 3 pair comparisons. We therefore used a sequential Bonferroni type procedure, which is very conservative and assures the statistical soundness of the results. Specifically, the results were first tested for significance of main effect using a Bonferroni correction, and later multiple comparison was done using a second Bonferroni correction[1].

6.3 Results

In the following section we describe first an analysis of the individual features that were extracted as indicated above. We start by describing the distributions of the

[1]Note that with only three treatment groups, it's overly conservative to adjust the alpha levels with a Bonferroni method as with only 3 treatment groups, there is little risk in an increasing Type I error rate.

different features according to the different groups. The distributions of the different features are described one by one and are compared to previous reports, when such exist, in the literature. When of interest, we also report features not used in the analysis as a measure of appropriateness of the calculation.

6.3.1 Isolated Features - between group analysis

Figure 6.4 contains the mean and standard error of isolated features extracted from semi structured interview. Statistically significant deviations between any two groups are indicated with a horizontal bar. The analysis was performed, while taking into consideration the issue of multiple comparisons as explained in Section 6.2.6 and is thus very conservative in nature. In the following we provide detailed discussion of each of the single features.

Spoken ratio

In [99] the authors have measured alogia as number of words per second. In some sense we follow their approach. However, without human interference we are unable to reliably establish word boundaries and we are thus reduced to measuring total spoken volume ratio. As seen in Figure 6.4a, the ratio of spoken volume for healthy subjects (47.19%,1.98%) is larger than that of Schizophrenia subjects (37.16%,2.4%) and Depressed subjects (29.52%, 2.4%). The difference between the groups is indeed significant ($\chi^2 = 21.63; df = 2, 59; p < 0.001$) and possibly reflects the subject's initiative and willingness to engage in conversation; however it may also reflect, to some degree, the conversational behavior of the interviewer rather than the interviewee.

Utterance Duration

Healthy subjects appear to speak in longer utterances (∼1.35s, 0.07s) as compared to Schizophrenia subjects (1.26s, 0.05s) and depressive subjects (1.03s,0.05s). Significant ($\chi^2 = 16.06; df = 2, 59; p < 0.001$) differences were observed between the depressed group and both the group of healthy subjects and schizophrenia subjects. Only a trend was observed between normal controls and schizophrenia subjects. In

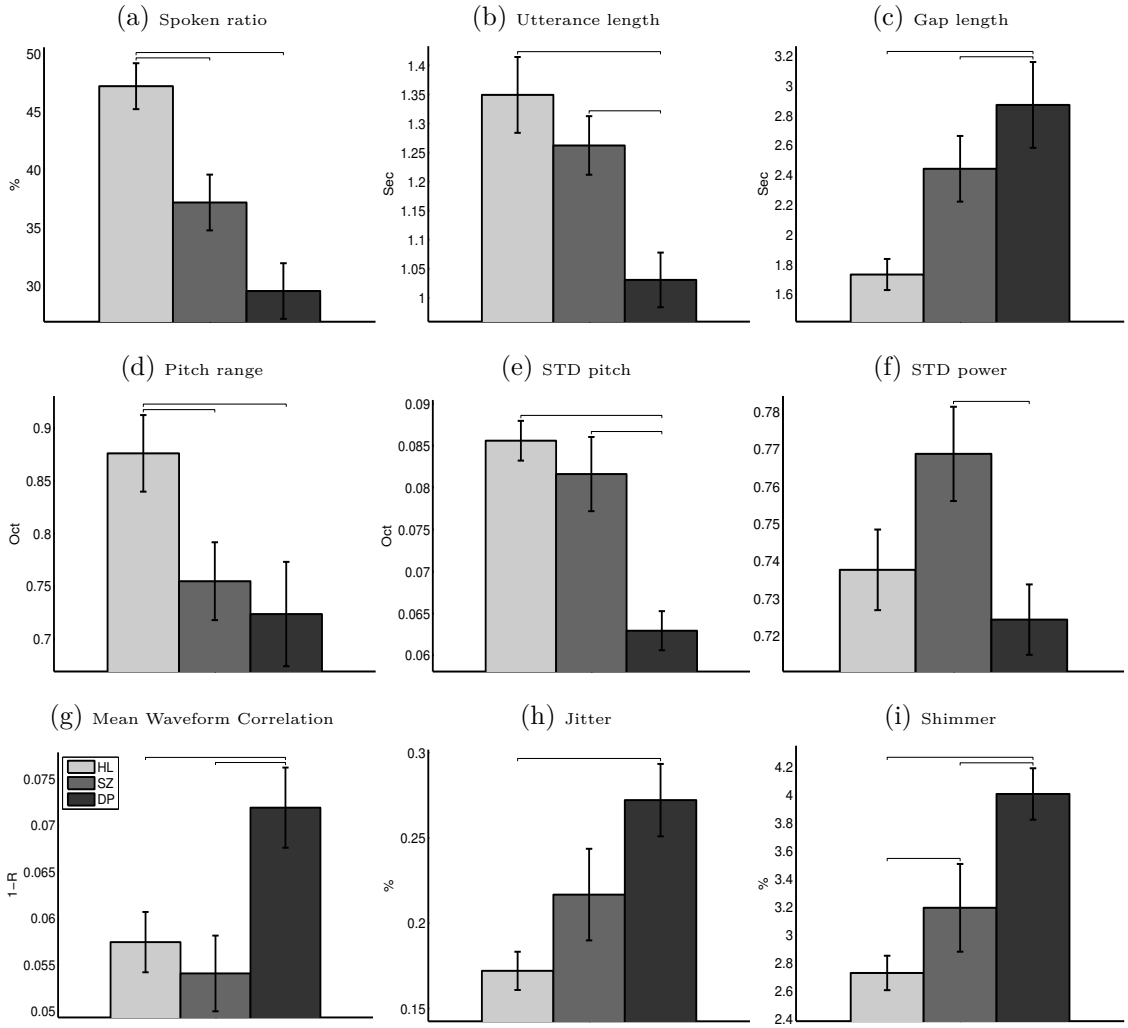

Figure 6.4: Individual features: 9 features are analyzed showing mean and standard error (STE) bar for each feature in each group of patients. Significant differences are indicated with a horizontal bar. Significance was established using a Kruskal-Wallis procedure and a Bonferonni correction for multiple comparisons.

depressed recordings an average utterance length that exceeds 3 seconds (as averaged over a two minutes segment) never occurred, which is reflective of the reported difficulty of engaging in conversation. These results tend to agree with previous reports [127].

Gap Duration

Normal subjects tend to pause less and for less time (1.73s, 0.1s) as compared to schiziprenia subjects (2.44s,0.22s) and depressive subjects (2.87s,0.29s) ($\chi^2 = 14.63$; $df = 2, 59$; $p < 0.001$). Also, the pauses of normals are of more regular pattern as reflected by the small error bar. These results are in congruence with previous reports in the literature that have shown that both length and frequency of within-clause pauses tend to be positively correlated with Alogia [118] and Ham-D [144] measures. Contrary to this study, the researchers in [118] have used a human observer along with a computerized system (VOXCOM). In addition the authors of [118] analyzed the syntactic word choice of the subjects and suggests that alogic patients have difficulty in word finding and in thought formulation, that is manifested in a general increase in the duration of pauses.

Pitch Range

Healthy subjects evidently displayed larger pitch range of ($\sim$0.88,$\sim$0.04) (mean, ste) of an octave, which is in agrrement with reported literature [105, 119, 135]. Pitch range is significantly reduced for schizophrenia (0.75, $\sim$0.04) with an even lower pitch range of (0.67,$\sim$0.05) for depressed patients ($\chi^2 = 7.49$; $df = 2, 59$; $p < 0.05$). While between group differences are significant, the differences between schizophrenia subjects and depressed ones show only a trend; this may suggest that global pitch range might not be the adequate measure for the perceptual quality of inflection and a possibly a more local concept might be the adequate one.

Standard Deviation of Pitch

Our measure of standard deviation of pitch within an utterance represents a temporally local perception of inflection. Normals display a wider dynamics of pitch (0.0992,0.0024) within an utterance as compared to schizophrenia subjects (0.0924, 0.0044) and depressed subjects are significantly different (0.0629,0.0023) ($\chi^2 = 21.28$; $df = 2, 59$; $p < 0.001$). This difference may also be contributed by the fact that normals have longer utterance, as suggested also by the large correlations between utterance length and standard deviation of pitch (see Table 6.2).

Schizophrenia patients display diverse behavior covering the entire behavioral range, as suggested by the large error bar. This may also be attributed to the lack of distinction between flat patients and "non flat" patients, which was shown previously [99] to be of significance when measuring inflection. In [119] the researchers showed that flat effect patient have reduced inflection as compared to non-flat patient. A later study [135] showed that blunt effect patients have reduced inflection when compared to healthy controls. Lastly, in [99] the authors found a significant difference in computer measure of inflection between flat effect subjects and control subjects, but only a mild trend between non flat subjects and control. Note, that in [99] the authors took inflection as the standard deviation of fundamental frequency in all spoken segments (not taking into account neither the level of aperiodicity in the segment, nor the mean fundamental frequency) an approach which may have effected the values reported[1].

Power standard deviation

Differences in power standard deviation were evident between Schizophrenia subjects ($\sim$0.77, $\sim$0.01) and Healthy subjects (0.74,$\sim$0.01) on the one hand, and Depressed subjects (0.7244,0.01) on the other hand ($\chi^2 = 6; df = 2,59; p < 0.05$). Only differences between schizophrenia and depression remained significant after correction for multiple comparisons. While the reduced variability in standard deviation of power for depression subjects is as reported in the literature, the increased standard deviation of power in schizophrenia subjects contradicts in part previous reports.

Mean Waveform Correlation

The mean of all correlation coefficients evaluated for every pair of consecutive periods is used as the acoustic measure termed *Mean Waveform Correlation (MWC)*. It indicates the overall similarity between the structures of the time signal. When applied to pitched segments it measures the level at which the speaker sustained a constant pitch. The evaluation of *MWC* is relatively robust compared to jitter and shimmer. Its upper limit of 1 is reached for signals with identical periodic shapes.

[1] Indeed when controlling for gender the reported effect reduced to avery mild trend.

MWC Decreases with increasing differences in length or shape between consecutive periods. i.e. with short term variations of the time signal [97]. The results show a significant deviation of the depressed subjects ($\chi^2 = 10.42; df = 2, 59; p < 0.01$). This criterion indicated a reduced ability to maintain a constant pitch, which may indicate a problem in the generation mechanism. The result also agrees with the reported results of both jitter and shimmer.

Mean Jitter

Following [97] jitter was calculated as the period perturbation quotient *PPQ* with a locality parameter of 5. (That is, jitter was measured on 5 cycles at a time with a lag of 1 cycle). Generally speaking, we allowed only pitched segments with aperiodicity factor (*ap0*) of up to 0.1. This criterion left clearly separated phonemes for analysis as validated both by eye and by ear. Calculation of *ap0* was done as suggested in [142] and is based mainly on the shape of the waveform (rather than magnitude); we therefore expect very low values of jitter. Indeed, this strong thresholding on the *ap0* gave very small values of jitter for all conditions. This restriction in dynamic range of the jitter measure still allowed for significant differences between healthy subjects ($0.1722, \sim 0.01$)) and both schizophrenia subjects ($\sim$0.22, 0.01) and depressed subjects ($\sim$0.27, 0.0212). The way jitter was calculated puts a focus on the physiological ability to maintain a constant period and suggests a deficiency in this ability in both schizophrenia and depression subjects.

Mean Shimmer

Following [97], shimmer was calculated as the energy perturbation quotient *EPQ* with a locality parameter of 5. Since the shimmer measure is based on the energy sequence it is expected to be considerably less susceptible to noise than the amplitude shimmer often used in acoustic analysis. Healthy subjects displayed the lowest shimmer ($2.73\%, 0.12\%$) whereas depressed subjects displayed an elevated shimmer ($4\%, 0.18\%$) with an intermediate level($3.22\%, 0.12\%$) for schizophrenia subjects. Again these results may suggest some problem in spontaneous control of the glottal production mechanism. All post-hoc between group comparisons were significant ($\chi^2 = 26.41; df = 2, 59; p << 0.01$).

6.3.2 Reliability of measures

In psychometrics, reliability is used to describe the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consistent conditions. For example, measurements of people's height and weight are often extremely reliable. Some researches have claimed that speech measurements have arguable validity and poor reliability, pointing in part to the large within subject variability [139]. This issue is indeed crucial in all applications of speech analysis, and poses a major problem for the development of reliable identification and classification tools. Ideally, for the purpose of classification, we would have liked to employ measures which are time and task independent on the one hand, and (mental) condition-dependent on the other hand. Such a tool would ultimately be invariant to the recorded task and to the specific time a sample has been taken.

In a previous study [116] we observed that both macro-scale and meso-scale measures seem to incorporate a larger variability component due to the specific task (between task-variability: (S_b) as compared to the within speaking-task variability (S_w). This task dependency is significantly reduced for micro-scale measures where one sees the ratio stabilize and generally around 1. This observation does not disqualify the larger scale measures from being useful in a classification task (as the task is known in advance); however, it highlights micro-scale features as candidates to be used by a robust general-purpose classifier.

6.3.3 Correlation

In order to compensate for excessive skew in the clinical measures we followed [99] and employed non-parametric statistics (Spearman's ρ correlation coefficient). Rank order correlation (Spearman) were computed between the acoustic and clinical based symptoms of the subjects. Pairwise correlations were computed for all combinations of acoustic features and clinical measures. This data is presented in Table 6.2.

Next we correlated the acoustic measures with the diagnostic rating as seen in Table 6.3. In this table the correlation scores were only calculated within the relevant group, that is, Schizophrenia clinical ratings were correlated with acoustic

Feature Correlations

	Acoustic features								
	1	2	3	4	5	6	7	8	9
1. Spoken Ratio	1	0.77^*	-0.9^*	0.29^*	0.52^*	0.21^*	-0.32^*	-0.62^*	-0.65^*
2. Utterance Duration	–	1	-0.49^*	0.32^*	0.68^*	0.47^*	-0.59^*	-0.43^*	-0.55^*
3. Gap Duration	–	–	1	-0.22^*	-0.32^*	-0.05	0.12	0.59^*	0.56^*
4. Pitch Range	–	–	–	1	0.56^*	-0.1	-0.24^*	0.09	-0.22^*
5. STD Pitch	–	–	–	–	1	0.33^*	-0.53^*	-0.33^*	-0.51^*
6. STD Power	–	–	–	–	–	1	-0.46^*	-0.06	-0.2^*
7. $1 - MWC$	–	–	–	–	–	–	1	0.09	0.34^*
8. Jitter	–	–	–	–	–	–	–	1	0.72^*
9. Shimmer	–	–	–	–	–	–	–	–	1

Table 6.2: Within feature correlations: * indicates $p << 0.01$

measures of subjects diagnosed with schizophrenia and depression clinical ratings were correlated with acoustic measures of depressed subjects only.

	Schizophrenia Scales				Depression Scales	
	PANSS	SANS (total)	SANS (affect)	SANS (alogia)	MADRAS	HAM-D
1. Spoken Ratio	-0.11	-0.5^{**}	-0.58^{**}	-0.64^{**}	-0.11^*	0.11
2. Utterance Duration	-0.19	-0.44^{**}	-0.55^{**}	-0.49^{**}	-0.27^*	0.05
3. Gap Duration	0.09	0.45^{**}	0.45^{**}	0.54^{**}	-0.01^*	-0.14
4. Pitch Range	-0.16	0.04	0.13	0	-0.35^*	-0.33^*
5. STD Pitch	0	-0.07	-0.17	-0.17	-0.18	-0.15
6. STD Power	-0.14	-0.27	-0.39	-0.31^*	-0.01	0.1
7. $1 - MWC$	0.03	0.24	0.4	0.32^*	0.19	0.19
8. Jitter	0.34	0.38^*	0.21	0.45^*	-0.1	-0.3
9. Shimmer	0.41^*	0.4^*	0.18	0.31^*	-0.01	0.2

Table 6.3: Acoustic Features- Psychiatric Scales Correlations * indicates $p < 0.05$, ** indicates $p << 0.01$

Some findings form Table 6.3 bear mention here. Spoken ratio was defined to agree with the description of Alogia as a reduction in quantity of speech; it is not surprising that it is highly correlated with the SANS-alogia clinical rating $(0.64, p << 0.01)$. Contrary to reported results in [99] high correlation between STD of pitch and spoken ration were observed.

6.3.4 Classification

Linear support vector machines (SVM [95]) were employed to conduct training of discriminative models using the extracted measures. The task consists of either binary classification, where a model was trained to discriminate between two distinct mental states, or multi-class classification, where a set of models was trained to identify the mental state of a specific speaker in a *1 vs. all* approach given a balanced set of training speech data. Table 6.4. contains the average success rate over all leave-one-out test realizations. Pair wise classification rates are shown on the off-diagonal entries. Multi-class classification results are shown on bottom right entry. As hinted by the single feature analysis, pairwise classification is easier for depression vs healthy subjects. Multi-class classification success rate are at 69.77 (baseline success is at 33.3%).

		Class			
		Hl	Sz	Dp	All
	Hl	x	76.19%	87.5%	x
	Sz	x	x	71.43%	x
	Dp	x	x	x	x
Class	All	x	x	x	69.77%

Table 6.4: Classification results: results of two-way classification and three-way classification.

6.4 Discussion

Speech acoustics is a measurable behavior that might be utilized as a biomarker in the clinical setting. The change in the acoustics of speech is not the only aspect of speech that changes in the course of various disorders [135] but these changes are a well documented phenomenon in both schizophrenia and depression. In both disorders speech acoustics can change over time. Schizophrenia speech might change over the course of years and might correlate with the course of other negative symptoms [99, 115]. There is little data of speech in the first psychotic episode and the podrom period. Depression symptoms' time course is wax and wane, hence speech

acoustics are expected to present accordingly. The depression vegetative symptoms usually fade early in the course of a successful treatment [145], speech behavior might be a good indicator for the response to treatment. The application of this tool might Include screening and treatment assessment. Negative symptoms might appear before the positive we might be able to predict the course of the disorder. The information acquired in this study will help design future computerized acoustic measure studies. In a later stage this method might be applied on first degree relatives of schizophrenia patients to better assess the risk of future psychotic illness. The method might also be beneficial in other mental disorders such as autism. Further research is required to replicate the results and determine the possible use of this biomarker in the clinical setting.

Significance and Future Studies

This study potentially provides an important tool in *Brain & Behavior Research Foundation's* (Formerly NARSAD) mission to find better treatments and ultimately prevent severe mental illnesses. The study contributes to the identification of a possible biomarker for schizophrenia and major depressive disorder. The development of reliable, objective, low-priced, and readily applicable assessment tool would enhance the accuracy of the clinical evaluation for diagnosis and severity. In the future the potential use of acoustic measures in primary prevention (mapping populations at risk like relatives of schizophrenia patients), secondary and tertiary prevention (assessment of the severity and course of the disease and treatment response) can be explored. Suitable technological apparatus including speech recognition software could allow this tool to be applied for screening or monitoring mental health status from remote on the phone by a computer [146].

In this study, we explored some of the difficulties related to the rating of the negative symptoms of schizophrenia, as well as those of major depression. We contrasted and compared the use of the clinical rating scale SANS and MADRAS with an acoustic method that offers the possibility for a more objective assessment of these clinical phenomena.

The process for completing a clinical rating scale (e.g. SANS) is achieved within a psychophysical paradigm, the physical being the stimulus presented to the inter-

viewer by the patient's speech paralinguistics, the psychological being the rater's clinical impressions. The model behind a clinical rating scale assumes that flat affect and alogia are measurable signs and not symptoms that require attention to the content of the patient's report. The interview, a primary psychiatric assessment tool, provides the opportunity for the patient to report subjective distress while the clinician observes the patient's speech, demeanor and dyadic interactions. In addition, the model assumes that paralinguistic aspects of talking provide a sufficient basis for the ratings. To enhance the reliability of the SANS ratings, SANS rating is usually based on the consensus of two calibrated raters. However, "Reliability is a necessary but not sufficient condition for validity" [147] and this study suggests that SANS ratings, while reliable, are not measuring what they claim.

For example, we found that the relation between the SANS rating for *Reduced Vocal Inflection* and the acoustic measures of *Pitch Range* or *Pitch STD* produced an insignificant correlation that was virtually zero. Yet, the acoustic measure has face validity. It appears that the SANS is completed on a 'topdown' basis. With the use of objective measures it is possible to demonstrate that there is an indirect path from patient behavior to global impression. This impression, then, leads to the item rating. [148], in psychometric studies of the SANS, reported results consistent with this view. Her report noted a Cronbach's alpha of 0.86 for the five global ratings. The five scales behaved like members of a single scale. Such high coherence suggests that SANS ratings derive from an undifferentiated global impression, which would constitute a 'top-down' rating.

A number of factors may contribute to this 'topdown' rating process. The SANS requires discriminations which exceed the abilities of clinicians. For example, the mean utterance duration of our subjects ranged from $\sim$ 1 s to $\sim$ 1.35 s, with a standard deviation of about 0.25 s. To rate the SANS item, *Poverty of Speech*, the rater might be expected to assign a zero (normal) to latencies at or below 1.35 s. Latencies around 1 s, or less, would be assigned a five, the highest SANS score. Between the zero rating and the five rating would be increments each 0.05 s; a score of one for durations of 1.05 s, etc. A listener would have to have a very accurate time perception to be able to do that even when given multiple opportunities to average over. The demands of this rating exceed the sensory capacity of the common rater. Newman and Mather (1938) [123] reported an early systematic attempt

to operationalize important clinical signs of depression. They found that an experienced speech pathologist required many repetitions of the recordings of the patient's speech before the type and severity of vocal changes could be established with confidence. Yet, clinicians are expected to capture subtle differences 'on line' and while distracted by other tasks and ratings. A similar analysis could be applied to the item, *Reduced Vocal Inflection*. The rater is expected to judge the variance of F_0 across syllables, and to discriminate fractions of a semitone. The subtleties of the required discriminations exceed the abilities of most raters, especially if we consider that the rater is expected to process the several speech parameters simultaneously.

The ratings were obtained within the assessment core of a clinical research center, and reflect state of the art procedures. Raters are trained to reliability criteria, and are checked periodically to prevent drift. Another factor that might contribute to a 'top-down' process is that a number of the items, such as *Blocking* or *Poverty of Content of Speech*, or the distinction between *Spontaneous Movements* and *Expressive Gestures* are not well operationalized. To rate blocking, the patient should stop in the middle of a sentence and report that that they had forgotten what they had intended to say. The impression of blocking appears to result from global impressions of alogia. The lack of operational definitions invites a 'top-down' approach. It appears that the high rater reliability is at the expense of rater validity. There has not been a great deal of enthusiasm for investing in the infrastructure of clinical laboratory procedures in psychiatry; rather there is a preference for the more exciting, cutting-edge technology of imaging and genetics. However, progress in these exciting areas is limited by the imprecision of the assessment infrastructure.

Recently, negative symptoms have been added to DSM as an area contributing to diagnosis, heightening the importance of valid assessment. Our findings that the acoustic measures can separate schizophrenia for depression subjects, without reference to the content of the words provides converging evidence of the usefulness of these methods. There are other areas in DSM that might profit from a focus on validity measures. While others have used various combinations of descriptive features from both the spectral and the temporal domain e.g. [97, 99, 105], we focused on a relatively simple set of features extracted in the temporal domain. We have divided our set of features into three groups of features, according to the time scale required for their extraction. We have shown that while macro-scale

correlate with distinct components of the SANS rating scale meso-sclae features show poor correlations. Our results suggests that micro-scale measures may be of good potential as a diagnostic tool both in terms of reliability and in those of validity.

7

Discussion and Concluding Remarks

"A story has no beginning or end: arbitrarily one chooses that moment of experience from which to look back or from which to look ahead."

— Graham Greene, The End of the Affair

In this thesis we have developed, presented and experimented with several algorithms in machine learning, and signal processing, where the main application domain has been the neuroscience experimental setting to its different flavors.

Firstly, we have developed and employed algorithms for learning distance functions to characterize the invariant spaces of a reacting system. The introduction of an efficient computational method for the characterization of sensory systems with a special focus on the characterization of single cells in both the higher visual and higher auditory brain regions is a key contribution of this dissertation.

This computational method has the potential to reveal principles that govern information processing and frameworks of representation in the those regions. We have taken the concept of distance function learning introduced earlier to the field, and adapted the learning scheme so as to better take advantage of the available information gathered before or during the experimental setting. Thus achieving better trained models with better generalization properties. The developed algorithms show promising results in the fundamental task of single cell characterization and the learned models surpass both existing distance learning algorithms and other models which have been applied to the problem of single cell characterization.

While we have applied the method and learned a single cell's distance function from the electrophysiological recordings of its responses, the suggested framework, as well as the specific algorithms can be easily applied to characterize other, more

elaborate systems, with other types of response patterns such as EEG, EMG and fMRI traces to name a few. Additionally, the developed algorithms may also play a role in a wide variety of application domains including data clustering, image retrieval, data classification, signal enhancement, de-noising, collaborative filtering and interactions prediction as well as the analysis of other neuronal data recordings.

Our use of *KernelBoost* distance function algorithm has shown promising results in characterizing the invariant spaces of Visual neurons in the IT and PFC. The application and performance of the learning algorithm in this setting has been limited, we believe, by a poor representation of the response space. We believe that richer representations of the response space would result in better learning of the stimulus space.

Our distance regression algorithm has managed to capture key features in the way auditory stimuli are represented in both the IC and the A1. The results also suggest that there is a basic difference between the neurons in the IC and neurons in A1 the former being 'simpler' in terms of represented geometry. This algorithm can easily end efficiently be applied to other problem in the field of neuroscience as well as in other fields. Possible applications are *a*) Characterization of the difference between cells as defined by their different response patterns. *b*) Learning spike metrics through distance regression. *c*) Large system characterization using multi-unit responses.

An online version of the algorithm is a natural extension that can be of dramatic service in the experimental setting: It can be used to provide real time guidance to the experimenter leading the experiment to 'interesting' stimulus regions. An online version can be easily implemented in the linear case (when learning a Mahalanobis metric), but extra care should be given in the case of the kernelized version. Extrapolation from multiple cells is another extension that should be experimented with - the implementation of which - bring forth different variants of weighted SVR.

Secondly, we have developed a framework for signal enhancement of a signal that has gone through a degradation channel. This framework, similarly to the distance function learning approach, shifts the focus form the channel and does not attempt to learn an inverse operator. Instead, it bypasses the channel altogether, and attempts to reconstruct the signal using the pairing of examples of good and bad quality. A key component of this framework is a similarity based retrieval

mechanism of high quality spectral patches - a fact that makes the employment of our distance regression algorithm in this context, a natural choice. We learn the weights of over the spectral domain to better retrieve high quality patches.

Thirdly, we have demonstrated robust performance of signal localization using monaural cues, and suggested a *winner take all* model for binaural localization. Our model makes use of the information available at the human eardrum to try and localize speech signals on the horizontal plane. Reported monaural human performance is replicated by the model and the winner takes all approach brings the results close to human binaural performance without the use of complex delay lines and comparison mechanisms.

Finally we follow reports in the psychiatric community of changes in speech prosody in the course of different mental states. We have developed an automatic tool for real time analysis of clinical interviews. We have analyzed a set of clinical interview samples of speakers of three groups: schiziphrenia subjects, depression subjects and healthy subjects. We developed a robust feature extraction mechanism, that employs voice activity detection, a segmentation process, as well as pitch detection algorithms to extract a set of temporal domain features in an attempt to quantify these changes. While speech is a very variate phenomena both within a speaker and between speakers we mange to extract reliable features and show state of the are results in a classification task of the three mental states.

References

[1] Camillo Golgi. **The neuron doctrine: theory and facts**. *Nobel Lectures: Physiology or Medicine*, **1921**:189–217, 1901. 2

[2] Theodore H Bullock, Michael VL Bennett, Daniel Johnston, Robert Josephson, Eve Marder, and R Douglas Fields. **The neuron doctrine, redux**. *Science*, **310**(5749):791–793, 2005. 2

[3] F.E. Theunissen, K. Sen, and A.J. Doupe. **Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds**. *The Journal of Neuroscience*, **20**(6):2315–2331, 2000. 8, 16

[4] D.T. Blake and M.M. Merzenich. **Changes of AI receptive fields with sound density**. *Journal of neurophysiology*, **88**(6):3409–3420, 2002. 8

[5] N. C. Rust, O. Schwartz, J. A. Movshon, and E. P. Simoncelli. **Spatiotemporal elements of macaque v1 receptive fields.** *Neuron*, **46**(6):945–956, Jun 2005. 8, 16

[6] S.V. David and J. L. Gallant. **Predicting neuronal responses during natural vision.** *Network*, **16**(2-3):239–260, 2005. 8, 16

[7] D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. **A comparison of primate prefrontal and inferior temporal cortices during visual categorization.** *J Neurosci*, **23**(12):5235–5246, 2003. 9, 14, 26, 27

[8] O. Bar-Yosef, Y. Rotman, and I. Nelken. **Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context**. *J Neurosci*, **22**(19):8619–8632, Oct 2002. 9, 16, 45, 46

[9] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. **Distance metric learning with application to clustering with side-information**. *Advances in neural information processing systems*, pages 521–528, 2003. 10, 12

[10] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. **Learning distance functions using equivalence relations**. In *International Conference on Machine Learning (ICML).*, 2003. 10, 11, 12

[11] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I.S. Dhillon. **Information-theoretic metric learning**. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007. 10, 12

[12] T. Hertz. *Learning Distance Functions: Algorithms and Applications*. PhD thesis, Citeseer, 2006. 10

[13] Noam Shental, Aharon Bar-Hillel, Tomer Hertz, and Daphna Weinshall. **Computing Gaussian mixture models with EM using equivalence constraints**. *Advances in neural information processing systems*, **16**(8):465–472, 2004. 11

[14] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. **Distance metric learning for large margin nearest neighbor classification**. In *In NIPS*. Citeseer, 2006. 11, 12

[15] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained clustering: Advances in algorithms, theory, and applications*. Chapman & Hall/CRC, 2008. 11

[16] Hakan Cevikalp, Jakob Verbeek, Frederic Jurie, Alexander Klaser, et al. **Semi-supervised dimensionality reduction using pairwise equivalence constraints**. In *3rd International Conference on Computer Vision*

Theory and Applications (VISAPP'08), pages 489–496, 2008. 11

[17] Rubi Hammer, Tomer Hertz, Shaul Hochstein, and Daphna Weinshall. **Category learning from equivalence constraints**. *Cognitive processing*, **10**(3):211–232, 2009. 11

[18] T. Hertz, A. Bar-Hillel, and D. Weinshall. **Boosting margin based distance functions for clustering**. In *Proceedings of the twenty-first international conference on Machine learning*, page 50. ACM, 2004. 12, 25, 50

[19] T. Hertz, A. Bar-Hillel, and D. Weinshall. **Learning a kernel function for classification with small training samples.** In *International Conference on Machine Learning (ICML)*., 2006. 12, 25, 28

[20] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul. **A new kernelization framework for Mahalanobis distance learning algorithms**. *Neurocomputing*, **73**(10):1570–1579, 2010. 12

[21] A. Globerson and S. Roweis. **Metric learning by collapsing classes**. *Advances in neural information processing systems*, **18**:451, 2006. 12

[22] CL Blake and CJ Merz. **UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science**, 1998. 12

[23] Peter J Rousseeuw. **Least median of squares regression**. *Journal of the American statistical association*, **79**(388):871–880, 1984. 13

[24] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik. **Support vector regression machines**. *Advances in neural information processing systems*, pages 155–161, 1997. 13

[25] J.A.K. Suykens and J. Vandewalle. **Least squares support vector machine classifiers**. *Neural processing letters*, **9**(3):293–300, 1999. 13

[26] Douglas M Bates and Donald G Watts. *Nonlinear regression: iterative estimation and linear approximations*. Wiley Online Library, 2008. 13

[27] DH Hubel and T.N. Wiesel. **Receptive fields of single neurons in the cat's striate visual cortex**. *J. Phys*, **148**:574–591, 1959. 14

[28] Nikos K Logothetis and David L Sheinberg. **Visual object recognition**. *Annual review of neuroscience*, **19**(1):577–621, 1996. 14

[29] Keiji Tanaka. **Inferotemporal cortex and object vision**. *Annual review of neuroscience*, **19**(1):109–139, 1996. 14

[30] M. Riesenhuber and T. Poggio. **Hierarchical models of object recognition in cortex**. *nature neuroscience*, **2**:1019–1025, 1999. 14

[31] T. Serre, M. Kouh., C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. **A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex**. AI Memo 2005-036 / CBCL Memo 259, MIT, Cambridge, MA, 2005. 14

[32] T. Serre, A. Oliva, and T. Poggio. **A Feedforward theory of visual cortex accounts for human performance in rapid categorization**. *Proceedings of the National Academy of Science*, **104**(15):6424–6429, 2007. 14

[33] C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo. **Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex**. *Science*, **310**:863–866, November 2005. 14

[34] D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. **Categorical representation of visual stimuli in the primate prefrontal cortex.** *Science*, **291**(5502):312–316, Jan 2001. 14, 26, 27

[35] C. Cadieu, M. Kouh, A. Pasupathy, C. E. Connor, M. Riesenhuber, and T. Poggio. **A model of V4 shape selectivity and invariance.** *J Neurophysiol*, **98**(3):1733–1750, Sep 2007. 14

[36] G. Chechik and I. Nelken. **Auditory abstraction from spectro-temporal features to coding auditory entities**. *Proceedings of the National Academy of Sciences*, **109**(46):18968–18973, 2012. 15, 16

[37] A. Aertsen and PIM Johannesma. **The spectro-temporal receptive field**. *Biological cybernetics*, **42**(2):133–143, 1981. 16

[38] D.L. Ringach. **Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex**. *Journal of Neurophysiology*, **88**(1):455–463, 2002. 16

[39] S.V. David, W.E. Vinje, and J.L. Gallant. **Natural stimulus statistics alter the receptive field structure of v1 neurons**. *The Journal of Neuroscience*, **24**(31):6991–7006, 2004. 16

[40] O. Bar-Yosef and I. Nelken. **The effects of background noise on the neural responses to natural sounds in cat primary auditory cortex**. *Frontiers in computational neuroscience*, **1**, 2007. 16, 46

[41] I. Weiner, T. Hertz, I. Nelken, and D. Weinshall. **Analyzing Auditory Neurons by Learning Distance Functions**. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1481–1488. MIT Press, Cambridge, MA, 2006. 16, 28, 46, 47

[42] Pat Langley et al. *Selection of relevant features in machine learning*. Defense Technical Information Center, 1994. 18

[43] Isabelle Guyon and André Elisseeff. **An introduction to variable and feature selection**. *The Journal of Machine Learning Research*, **3**:1157–1182, 2003. 19

[44] Ronald A Fisher. **The use of multiple measurements in taxonomic problems**. *Annals of Human Genetics*, **7**(2):179–188, 1936. 20

[45] Prasanta Chandra Mahalanobis. **On the generalized distance in statistics**. In *Proceedings of the National Institute of Sciences of India*, **2**, pages 49–55. New Delhi, 1936. 20

[46] Sam T Roweis and Lawrence K Saul. **Nonlinear dimensionality reduction by locally linear embedding**. *Science*, **290**(5500):2323–2326, 2000. 20

[47] Joshua B Tenenbaum, Vin De Silva, and John C Langford. **A global geometric framework for nonlinear dimensionality reduction**. *Science*, **290**(5500):2319–2323, 2000. 20

[48] Dietrich Wettschereck, David W Aha, and Takao Mohri. **A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms**. *Artificial Intelligence Review*, **11**(1):273–314, 1997. 21

[49] Amir Globerson. *The Minimum Information Principle in Learning and Neural Data Analysis*. PhD thesis, Hebrew University of Jerusalem., 2005.

[50] Gal Chechik. *An information theoretic approach to the study of auditory coding*. PhD thesis, Hebrew University of Jerusalem, 2003.

[51] I Weiner. *Analyzing Auditory Neurons by Learning Distance Functions*. Master's thesis, Hebrew University of Jerusalem, 2005. 25, 50, 51

[52] T. Hertz, A. Bar-Hillel, and D. Weinshall. **Learning distance functions for image retrieval**. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004*, **2**, pages II–570–II–577 Vol.2, 2004. 28

[53] Warren S Torgerson. **Multidimensional scaling: I. Theory and method**. *Psychometrika*, **17**(4):401–419, 1952. 34

[54] C.M. Bishop and SpringerLink (Service en ligne). *Pattern recognition and machine learning*, **4**. springer New York, 2006. 39

[55] V. Vapnik. *The nature of statistical learning theory*. springer, 1999. 39

[56] T. Ezzat, J. Bouvrie, and T. Poggio. **AM-FM demodulation of spectrograms using localized 2D max-Gabor analysis**. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, **4**, pages IV–1061. IEEE, 2007. 47

[57] F.E. Theunissen, S.V. David, N.C. Singh, A. Hsu, W.E. Vinje, and J.L. Gallant. **Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli**. *Network: Computation in Neural Systems*, **12**(3):289–316, 2001. 50

[58] J. Zhang and P. Gill. **Strfpak documentation**. Technical report, Tech. rep., Theunissen and Gallant Laboratories, UC Berkeley, 2006. 50

[59] MR Weiss, E. Aschkenasy, and TW Parsons. **Processing speech signals to attenuate interference**. In *Proc. IEEE Symp. Speech Recognition*, pages 292–293, 1974. 60

[60] S. Boll. **Suppression of acoustic noise in speech using spectral subtraction**. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **27**(2):113–120, 1979. 60, 62, 66

[61] B.N. Levinson. **The Wiener RMS (root mean square) error criterion in filter design and prediction**. *Journal of Mathematical Physics*, **25**:261–278, 1947. 60, 62, 66

[62] J.S. Lim and A.V. Oppenheim. **Enhancement and bandwidth compression of noisy speech**. *Proceedings of the IEEE*, **67**(12):1586–1604, 1979. 60

[63] Y. Ephraim and H.L. Van Trees. **A signal subspace approach for speech enhancement**. *Speech and Audio Processing, IEEE Transactions on*, **3**(4):251–266, 1995. 60, 62

[64] Y. Ephraim and D. Malah. **Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator**. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **32**(6):1109–1121, 1984. 60, 62

[65] N. Virag. **Single channel speech enhancement based on masking properties of the human auditory system**. *Speech and Audio Processing, IEEE Transactions on*, **7**(2):126–137, 1999. 60, 62

[66] I. Cohen. **Relaxed statistical model for speech enhancement and a priori SNR estimation**. *Speech and Audio Processing, IEEE Transactions on*, **13**(5):870–881, 2005. 60, 62

[67] A. Buades, B. Coll, and J.M. Morel. **A non-local algorithm for image denoising**. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, **2**, pages 60–65. Ieee, 2005. 61

[68] H. Xu, Tan Z., P. Dalsgaard, and B. Lindberg. **Robust Speech Recognition by Nonlocal Means Denoising Processing**. *IEEE Signal processing Letters*, **15**:701–704, 2008. 61

[69] J. Lim and A. Oppenheim. **Reduction of quantization noise in PCM speech coding**. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **28**(1):107–110, 1980. 64

[70] D. Wang and J. Lim. **The unimportance of phase in speech enhancement**. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **30**(4):679–681, 1982. 66

[71] K. Paliwal, K. Wójcicki, and B. Shannon. **The importance of phase in speech enhancement**. *Speech communication*, **53**(4):465–494, 2011. 66

[72] M. Krawczyk and T. Gerkmann. **STFT Phase Improvement for Single Channel Speech Enhancement**. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pages 1–4. VDE, 2012. 66

[73] P.C. Loizou. *Speech enhancement: theory and practice*, **30**. CRC, 2007. 69

[74] I. Rec. **P. 862.1: Mapping function for transforming P. 862 raw result scores to MOS-LQO**. *International Telecommunication Union, Geneva*, 2003. 70

[75] A. Muscariello, G. Gravier, and F. Bimbot. **TOWARDS ROBUST WORD DISCOVERY BY SELF-SIMILARITY MATRIX COMPARISON**. *Acoustics, Speech and Signal Processing (ICASSP)*, 2011. 70

[76] Bob Dylan. **Chronicles Volume 1 Read by Sean Penn**, 2005. 70, 71

[77] E. Hemingway. **The Old Man and the Sea. Unabridged Audio Read by: Donald Sutherland**, 1952 (May 2006). 70

[78] R.B. Parker. **Stardust. Unabridged Audio Read by: Burt Reynolds**, 1991. 70

[79] E L James. **Fifty Shades of Grey. Unabridged Audiobook Read By: Becca Battoe**. Random House Audio, 2012. 70

[80] Birger Kollmeier and René Koch. **Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction**. *The Journal of the Acoustical Society of America*, **95**:1593, 1994. 78, 84

[81] A.D. Musicant and R.A. Butler. **Influence of monaural spectral cues on binaural localization**. *The Journal of the Acoustical Society of America*, **77**:202, 1985. 79

[82] J.C. Middlebrooks and D.M. Green. **Sound localization by human listeners**. *Annual Review of Psychology*, **42**(1):135–159, 1991. 79, 81

[83] D. McAlpine, D. Jiang, A.R. Palmer, et al. **A neural code for low-frequency sound localization in mammals**. *Nature neuroscience*, **4**(4):396–401, 2001. 79

[84] S.M. Chase and E.D. Young. **Limited segregation of different types of sound localization information among classes of units in the inferior colliculus**. *The Journal of neuroscience*, **25**(33):7575–7585, 2005. 79

[85] B. Grothe, M. Pecka, and D. McAlpine. **Mechanisms of sound localization in mammals**. *Physiological Reviews*, **90**(3):983–1012, 2010. 79

[86] M.M. Van Wanrooij and A.J. Van Opstal. **Contribution of head shadow and pinna cues to chronic monaural sound localization**. *Journal of Neuroscience*, **24**(17):4163, 2004. 79

[87] Jie Huang, Tadawute Supaongprapa, Ikutaka Terakura, Fuming Wang, Noboru Ohnishi, and Noboru Sugie. **A model-based sound localization system and its application to robot navigation**. *Robotics and Autonomous Systems*, **27**(4):199–209, 1999. 79

[88] Futoshi Asano, Masataka Goto, Katunobu Itou, and Hideki Asoh. **Real-time sound source localization and separation system and its application to automatic speech recognition.** In *INTERSPEECH*, pages 1013–1016, 2001. 79

[89] H.G. Fisher and S.J. Freedman. **The role of the pinna in auditory localization.** *Journal of Auditory research*, 1968. 80

[90] D.E. Shub, S.P. Carr, Y. Kong, and H.S. Colburn. **Discrimination and identification of azimuth using spectral shape**. *The Journal of the Acoustical Society of America*, **124**:3132, 2008. 81, 87

[91] S.M. Chase and E.D. Young. **Cues for sound localization are encoded in multiple aspects of spike trains in the inferior colliculus**. *Journal of neurophysiology*, **99**(4):1672, 2008. 82

[92] P.M. Hofman and A.J. Van Opstal. **Binaural weighting of pinna cues in human sound localization**. *Experimental brain research*, **148**(4):458–470, 2003. 82

[93] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren. *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. U.S. Dept. of Commerce NTIS, Gaithersburg, MD, 1990. 82

[94] H. Kayser, S.D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier. **Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural RoomImpulse Responses**. *EURASIP Journal on Advances in Signal Processing*, page 10, 2009. 83

[95] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. **LIBLINEAR: A Library for Large Linear Classification**. *Journal of Machine Learning Research*, **9**:1871–1874, 2008. 83, 111

[96] J.-H. Bach, B. Kollmeier, and J. Anemüller. **Modulation-based detection of speech in real background noise: Generalization to novel background classes**. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 41–44. IEEE, 2010. 84

[97] D. Michaelis, M. Fröhlich, and H.W. Strube. **Selection and combination of acoustic features for the description of pathologic voices**. *The Journal of the Acoustical Society of America*, **103**:1628, 1998. 92, 97, 102, 108, 114

[98] S.B. Davis. **Acoustic characteristics of normal and pathological voices**. *Speech and language: Advances in basic research and practice*, **1**:271–335, 1979. 92

[99] A.S. Cohen, M. Alpert, T.M. Nienow, T.J. Dinzeo, and N.M. Docherty. **Computerized measurement of negative symptoms in schizophrenia**. *Journal of psychiatric research*, **42**(10):827–836, 2008. 92, 93, 94, 97, 100, 104, 107, 109, 110, 111, 114

[100] Asli Ozdas, Richard G Shiavi, Stephen E Silverman, Marilyn K Silverman, and D Mitchell Wilkes. **Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk**. *Biomedical Engineering, IEEE Transactions on*, **51**(9):1530–1540, 2004. 92

[101] Elliot Moore, Mark A Clements, John W Peifer, and Lydia Weisser. **Critical analysis of the impact of glottal features in the classification of clinical depression in speech**. *Biomedical Engineering, IEEE Transactions on*, **55**(1):96–107, 2008. 92, 97

[102] M.D. Carpenter. *The Use of Syntactic Structure in the Speech Perception of Schizophrenics*. Microfilm, 1974. 92

[103] Doris Appel Graber. *Verbal behavior and politics*. Univ of Illinois Pr, 1976. 92

[104] M. Alpert, E.R. Pouget, and R.R. Silva. **Reflections of depression in acoustic measures of the patient's speech**. *Journal of affective disorders*, **66**(1):59–69, 2001. 92, 93, 94, 99

[105] M. Alpert, R.J. Shaw, E.R. Pouget, and K.O. Lim. **A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia**. *Journal of psychiatric research*, **36**(5):347–353, 2002. 92, 93, 94, 106, 114

[106] E. Kraepelin. **Translated by RM Barclay**. *Dementia praecox and paraphrenia. Edinburgh: E. and S. Livingstone*, 1919. 92

[107] N.C. Andreasen. **Symptoms, signs, and diagnosis of schizophrenia**. *The Lancet*, **346**(8973):477–481, 1995. 92

[108] E. Kraepelin. **Lectures on clinical psychiatry**. *The Journal of Nervous and Mental Disease*, **40**(6):423, 1913. 92, 93

[109] American Psychiatric Association and American Psychiatric Association. Task Force on DSM-IV. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. American Psychiatric Publishing, Inc., 2000. 92

[110] N.C. Andreasen. **Scale for the Assessment of Negative Symptoms (SANS).** *British Journal of Psychiatry*, 1989. 92, 96

[111] N.C. Andreasen. **Negative symptoms in schizophrenia: definition and reliability**. *Archives of General Psychiatry*, **39**(7):784, 1982. 92

[112] S.A. Montgomery and M. Asberg. **A new depression scale designed to be sensitive to change**. *The British Journal of Psychiatry*, **134**(4):382, 1979. 93, 96

[113] M. Hamilton. **A rating scale for depression**. *Journal of neurology, neurosurgery, and psychiatry*, **23**(1):56, 1960. 93, 96

[114] J.C. Mundt, P.J. Snyder, M.S. Cannizzaro, K. Chappie, and D.S. Geralts. **Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology**. *Journal of neurolinguistics*, **20**(1):50–64, 2007. 93

[115] HH Stassen, M. Albers, J. Puschel, C. Scharfetter, M. Tewesmeier, and B. Woggon. **Speaking behavior and voice**

sound characteristics associated with negative schizophrenia. *Journal of psychiatric research*, **29**(4):277–296, 1995. 93, 111

[116] R. Kliper, Y. Vaizman, D. Weinshall, and S. Portuguese. **Evidence for depression and schizophrenia in speech prosody**. *ExLing 2010*, page 85, 2010. 93, 99, 109

[117] H. Ellgring and K.R. Scherer. **Vocal indicators of mood change in depression**. *Journal of Nonverbal Behavior*, **20**(2):83–110, 1996. 93

[118] M. Alpert, A. Clark, and E.R. Pouget. **The syntactic role of pauses in the speech of schizophrenic patients with alogia.** *Journal of abnormal psychology*, **103**(4):750, 1994. 93, 106

[119] N.C. Andreasen, M. Alpert, and M.J. Martz. **Acoustic analysis: an objective measure of affective flattening**. *Archives of General Psychiatry*, **38**(3):281, 1981. 93, 106, 107

[120] M. Alpert, P. Homel, F. Merewether, J. Martz, and M. Lomask. **Voxcom: A system for analyzing natural speech in real time**. *Behavior Research Methods*, **18**(2):267–272, 1986. 93

[121] J.W. Pennebaker, M.R. Mehl, and K.G. Niederhoffer. **Psychological aspects of natural language use: Our words, our selves**. *Annual review of psychology*, **54**(1):547–577, 2003. 93, 94

[122] M. Alpert, E.R. Pouget, J. Welkowitz, and J. Cohen. **Mapping schizophrenic negative symptoms onto measures of the patient's speech: Set correlational analysis**. *Psychiatry research*, **48**(3):181–190, 1993. 93

[123] S. Newman and V.G. Mather. **Analysis of spoken language of patients with affective disorders**. *American Journal of Psychiatry*, **94**(4):913, 1938. 93, 113

[124] W.W. IsHak. *Outcome measurement in psychiatry: A critical review*. American Psychiatric Pub, 2002. 93

[125] J. Püschel, HH Stassen, G. Bomben, C. Scharfetter, and D. Hell. **Speaking behavior and speech sound characteristics in acute schizophrenia**. *Journal of psychiatric research*, **32**(2):89–97, 1998. 93

[126] A.M. Kring and J.A. Bachorowski. **Emotions and psychopathology**. *Cognition & Emotion*, **13**(5):575–599, 1999. 94

[127] M. Alpert, A. Kotsaftis, and E.R. Pouget. **Speech fluency and schizophrenic negative signs**. *Schizophrenia Bulletin*, **23**(2):171–177, 1997. 94, 99, 105

[128] M.S. Cannizzaro, H. Cohen, F. Rappard, and P.J. Snyder. **Bradyphrenia and bradykinesia both contribute to altered speech in schizophrenia: a quantitative acoustic study**. *Cognitive and behavioral neurology*, **18**(4):206, 2005. 94

[129] V. Rapcan, S. D'Arcy, S. Yeap, N. Afzal, J. Thakore, and R.B. Reilly. **Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia**. *Medical engineering & physics*, **32**(9):1074–1079, 2010. 94

[130] A Nilsonne. **Speech characteristics as indicators of depressive illness**. *Acta Psychiatrica Scandinavica*, **77**(3):253–263, 1988. 94

[131] I.A. Rubino, L. D'Agostino, L. Sarchiola, D. Romeo, A. Siracusano, and N.M.

Docherty. **Referential Failures and Affective Reactivity of Language in Schizophrenia and Unipolar Depression**. *Schizophrenia bulletin*, **37**(3):554, 2011. 94

[132] IV DSM. **American Psychiatric Association**. *Diagnostic and Statistical Manual of Mental Disorders*, 1994. 96

[133] S.R. Kay, A. Flszbein, and L.A. Opfer. **The positive and negative syndrome scale (PANSS) for schizophrenia.** *Schizophrenia bulletin*, **13**(2):261, 1987. 96

[134] MATLAB. *version 7.11.0 (R2010b)*. The MathWorks Inc., Natick, Massachusetts, 2010. 96

[135] M. Alpert, S.D. Rosenberg, E.R. Pouget, and R.J. Shaw. **Prosody and lexical accuracy in flat affect schizophrenia**. *Psychiatry Research*, **97**(2-3):107–118, 2000. 97, 106, 107, 111

[136] Eric Scheirer and Malcolm Slaney. **Construction and evaluation of a robust multifeature speech/music discriminator**. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, **2**, pages 1331–1334. IEEE, 1997. 97

[137] Michael J Carey, Eluned S Parris, and Harvey Lloyd-Thomas. **A comparison of features for speech, music discrimination**. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, **1**, pages 149–152. IEEE, 1999. 97

[138] R. Cowie and R.R. Cornelius. **Describing the emotional states that are expressed in speech**. *Speech Communication*, **40**(1-2):5–32, 2003. 97

[139] J. Rong, G. Li, and Y.P.P. Chen. **Acoustic feature selection for automatic emotion recognition from speech**. *Information processing & management*, **45**(3):315–328, 2009. 97, 109

[140] J. Sohn, N.S. Kim, and W. Sung. **A statistical model-based voice activity detection**. *Signal Processing Letters, IEEE*, **6**(1):1–3, 1999. 98

[141] Benjamin Pope and Aron W Siegman. **Interviewer warmth in relation to interviewee verbal behavior.** *Journal of Consulting and Clinical Psychology*, **32**(5p1):588, 1968. 99, 100

[142] A. De Cheveigné and H. Kawahara. **YIN, a fundamental frequency estimator for speech and music**. *The Journal of the Acoustical Society of America*, **111**:1917, 2002. 99, 108

[143] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. **LIBLINEAR: A library for large linear classification**. *The Journal of Machine Learning Research*, **9**:1871–1874, 2008. 103

[144] HH Stassen, S Kuny, and D Hell. **The speech analysis approach to determining onset of improvement under antidepressants**. *European neuropsychopharmacology*, **8**(4):303–310, 1998. 106

[145] Aaron T Beck and Brad A Alford. *Depression: Causes and treatment*. University of Pennsylvania Press, 2008. 112

[146] Michael S Cannizzaro, Nicole Reilly, James C Mundt, and Peter J Snyder. **Remote capture of human voice acoustical data by telephone: A methods study**. *Clinical linguistics & phonetics*, **19**(8):649–658, 2005. 112

[147] C Nunnally Jum and H Bernstein Ira. **Psychometric theory**, 1978. 113

[148] Nancy C Andreasen, Michael Flaum, Victor W Swayze, Gary Tyrrell, and Stephan Arndt. **Positive and negative symptoms in schizophrenia: A critical reappraisal**. *Archives of General Psychiatry*, **47**(7):615, 1990. 113