Identification of Stimulus Cues in Tone-in-Noise Detection with Sparse Logistic Regression

vorgelegt von
Dipl. Phys. (ETH) Vinzenz H. Schönfelder
aus Dresden

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss
Vorsitzender: Prof. Dr. Marc Alexa
Berichter: Prof. Felix Wichmann, DPhil
Berichter: Prof. Dr. Klaus Obermayer
Berichter: Prof. Dr. Bernhard Ronacher

Tag der wissenschaftlichen Aussprache: 15. Mai 2013

Berlin, 2013
D 83

Abstract

The central aim of psychophysics is to understand the functional relationship between the physical and the psychological world. Striving for that goal, modern research focuses on quantitatively measuring and explaining observer behavior in specific psychophysical paradigms. In this context, a principal question arises: Which particular stimulus features govern individual decisions in a behavioral task? As regards this problem, the classical psychophysical paradigm of narrow-band Tone-in-Noise (TiN) detection has been under investigation for more than 70 years. This particular experiment stands at the heart of a central notion in auditory perception: the "critical band". Yet no conclusive answer has been given as to which auditory features listeners employ in this task.

The present study describes how a modern statistical analysis procedure can be used to tackle this problem when modeling psychophysical data. The proposed technique combines the concept of relative linear combination weights with an L_1-regularized logistic regression—a procedure developed in machine learning. This method enforces "sparse" solutions, a computational approximation to the postulate that a good model should contain the minimal set of predictors necessary to explain the data. This property is essential when extracting the critical perceptual features from observer models after they were fit to behavioral data.

Using data generated from a simulated TiN detection paradigm, the method is shown to precisely identify observer cues from a large set of covarying, interdependent stimulus descriptors—a setting where standard correlation and regression methods fail. Furthermore, the detailed decision rules of the simulated observers were reconstructed, allowing predictions of responses on the basis of individual stimuli.

The practical part of this study aimed at using the sparse analysis procedure to investigate the perceptual mechanisms underlying the detection performance of human observers in a TiN detection paradigm. Therefore, a large trial-by-trial data set was collected with multiple listeners. Relative perceptual weights were then estimated for a diverse set of auditory features encompassing sound energy, fine structure and envelope. By expanding the common linear observer model to allow for behavioral predictors, sequential dependencies in observer responses were also taken into account. These dependencies generally impair detection performance and even arise when study participants are made aware of the purely random stimulus sequence. The fitted models captured the behavior of all listeners on a single-trial level. The estimated perceptual weights were stable across signal levels. They suggest that all observers depend on stimulus energy, and "critical band"-like detectors in the fine structure domain while a subset of the listeners exhibited an additional dependence on stimulus envelope. In addition to stimulus characteristics, earlier responses appeared to substantially influence the current decision of some observers.

In conclusion, by approaching a classical problem in auditory psychophysics with an advanced statistical analysis procedure, an already large pool of empirical knowledge was expanded in several important aspects. In that process, the power and efficiency of the proposed method was demonstrated. Based on very general concepts, it is flexible enough to be applicable in a wide variety of studies that investigate perceptual mechanisms.

Zusammenfassung

Das zentrale Ziel der Psychophysik ist das Verständnis der funktionellen Zusammenhänge zwischen der physischen und psychischen Welt. Aktuelle Forschung strebt nach diesem Ziel, indem sie sich darauf konzentriert menschliches Verhalten in bestimmten psychophysischen Paradigmen quantitativ zu messen und zu erklären. In diesem Umfeld ergibt sich eine prinzipielle Frage: Welche Reizeigenschaften bestimmen individuelle Entscheidungen in Verhaltensexperimenten? In Bezug auf dieses Problem wurde das klassische psychophysische Paradigma der Entdeckung von durch Rauschen maskierten Tonsignalen ("tone-in-noise detection", TiN) in den vergangenen 70 Jahren gründlich untersucht. Dieses Experiment bildet das Fundament für einen zentralen Begriff der auditorischen Wahrnehmung: das "kritische Band". Gleichwohl wurde bislang keine abschließende Antwort auf die Frage gefunden, welche auditorischen Reizeigenschaften die Hörer in dieser Aufgabe verwenden.

Die vorliegende Studie beschreibt, wie eine moderne statistische Analysemethode angewendet werden kann, um dieses Problem im Rahmen der Modellierung psychophysischer Daten zu lösen. Diese Prozedur vereint das Konzept der relativen linearen Kombinationsgewichte ("relative linear combination weights") mit einer L_1-regularisierten logistischen Regression–einer Methode die im Bereich des maschinellen Lernens entwickelt wurde. Die Methode erzwingt "spärliche" ("sparse") Lösungen, eine algorithmische Annäherung an das Postulat, dass ein gutes Modell nur diejenigen Prediktoren enthalten solle, die nötig sind um die beobachteten Daten zu erklären. Diese Eigenschaft ist entscheidend, um kritische Reizeigenschaften aus Beobachtermodellen zu extrahieren, nachdem diese an Verhaltensdaten angepasst wurden.

Mit Hilfe von Daten aus simulierten TiN-Experimenten wird gezeigt, dass die Methode die für die Beobachter entscheidenden Reizeigenschaften identifizieren kann–sogar dann, wenn eine Vielzahl von kovariierende, untereinander abhängigen Eigenschaften zur Wahl steht. In dieser Situation versagen bislang verwendete Methoden der Korrelations- oder Regressionsanalyse. Des Weiteren konnte die hier vorgestellte Analysemethode die detaillierten Entscheidungsregeln der simulierten Beobachter rekonstruieren und so Antworten auf der Basis individueller Stimuli vorhersagen.

Im praktischen Teil der vorliegenden Studie wird die "spärliche" Analysemethode angewendet, um die Wahrnehmungsmechanismen zu untersuchen, die der Entdeckungsleistung von Probanden in einem TiN-Entdeckungs-Paradigma zu Grunde liegen. Zu diesem Zweck wurden mit mehreren Beobachtern umfangreiche Daten mit Verhaltensinformationen auf der Ebene einzelner Stimulus-Antwort-Paare gesammelt. Anschließend wurden relative perzeptuelle Gewichte für einen Satz von vielfältigen auditorischen Reizeigenschaften geschätzt, bestehend aus Schallenergie, Ton-Feinstruktur und -Umhüllender. Das klassische lineare Beobachtermodell wurde außerdem durch behaviorale Prediktoren erweitert, um auch sequentielle Abhängigkeiten im Antwortverhalten der Beobachter zu berücksichtigen. Diese Abhängigkeiten verschlechtern generell die Entdeckungsleistung und treten auch dann auf, wenn die Probanden auf die Zufälligkeit der Reizreihenfolge hingewiesen werden. Die an die Daten angepassten Beobachtermodelle erklärten das Verhalten aller Probanden auf der Ebene von einzelnen Hörversuchen. Die geschätzten perzeptuellen Gewichte blieben über verschiedene Signalstärken hin konstant. Sie legen nahe, dass das Verhalten aller Beobachter von der Schallenergie abhing, sowie von Feinstruktur-Detektoren die den "kritischen Filtern" ähneln. Ein Teil der Beobachter zeigte eine zusätzliche Abhängigkeit von der Ton-Umhüllenden. Zusätzlich zu diesen Reizeigenschaften hatten bei einigen Probanden auch vorherige Entscheidungen einen wesentlichen Einfluss auf die aktuelle Antwort.

In dieser Arbeit wurde ein klassisches Problem in der auditorischen Psychophysik mit einer fortgeschrittenen statistischen Analysemethode untersucht. Dadurch konnte ein bereits großer empirischer Wissensschatz um einige wichtige Aspekte erweitert werden. Gleichzeitig wurde die Leistungsfähigkeit und Effizienz der vorgeschlagenen Methode demonstriert. Da sie auf sehr allgemeinen Konzepten beruht, ist sie zugleich so flexibel, dass sie in verschiedensten Studien über Wahrnehmungsmechanismen angewendet werden kann.

Acknowledgements

First of all, I would like to thank my doctoral supervisor Felix Wichmann for the opportunity to work on this stimulating research project, for his friendly advice, support, encouragement and calm patience during the past years. I also thank Bruce Henning for the original inspiration to this project, Daniel Tollin for advice regarding experimental equipment, as well as Virginia Richards and Yi Shen for fruitful discussions and helpful remarks that they offered during my stay in Irvine, California.

I also express my gratitude to the Bernstein Center Berlin, its associated graduate school (the PhD program coordinator Vanessa Casagrande in particular), Bernhard Ronacher of the Humboldt University of Berlin as well as Klaus Obermayer and Henning Sprekeler of the Technical University of Berlin. Thanks to their material and non-material support I was able to prepare with little time pressure the many aspects of this dissertation.

In addition, I thank Bernhard Ronacher for providing a large sound-proof booth for preparatory hearing tests, Ingo Fründ for inspiration and guidance concerning the analysis of sequential dependencies as well as Matthias Bethge and Jakob Macke for their advice on regularization techniques. In addition, I thank Bruce Berg and Kai Görgen for the inspiring ideas that they provided during different presentations of this research. Furthermore, I thank the anonymous reviewers of our submitted manuscripts for their insightful comments which also helped to improve the present thesis.

Finally and most importantly, I thank my family—whose head count doubled while I worked on this dissertation—for the accompaniment, advice and encouragement. In particular, the sharp and patient eye of Sigrun Beige tremendously helped to polish the work at hand during its final preparation.

This research was funded, in part, by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Programs Berlin (FKZ: 01GQ0414) and Tübingen (FKZ: 01GQ1002) and the Graduiertenkolleg 1589/1 (German Research Foundation DFG).

Declaration of Originality

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Berlin, May 2013

Pluralitas non est ponenda sine necessitate.

Frustra fit per plura, quod potest fieri per pauciora.

William of Ockham [c.1280-c.1349]

Ne pourroit on pas conjecturer que le Bruit n'est point d'une autre nature que le Son; qu'il n'est lui même que la somme d'une multitude confuse de Sons divers qui se font entendre à la fois & contrarient, en quelque sorte, mutuellement leurs ondulations?

Jean-Jacques Rousseau [1712-1778], Dictionaire de Musique (1768)

In telephone engineering circles it has been considered a real achievement to have developed a design which permits 30,000 wire terminals to be within the reach of a single operator. The area of such a switchboard panel is about 10,000 square centimeters. Nature has accomplished a similar thing in the hearing mechanizm occupying an area of only 1/10 of a square centimeter.

Harvey Fletcher [1884-1981], Auditory patterns (1940)

Publications & Presentations

Papers (peer-reviewed)

- Schönfelder, V. H., & Wichmann, F. A. (2012). *Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data.* The Journal of the Acoustical Society of America, 131(5), 3953-3969. doi:10.1121/1.3701832
- Schönfelder, V. H., & Wichmann, F. A. (2013). *Identification of Stimulus Cues in Narrow-Band Tone-in-Noise Detection using Sparse Observer Models.* The Journal of the Acoustical Society of America. 134 (1), (in press, publication on July 13th)

Talks

- Schönfelder, V. H., & Wichmann, F. A. (2008). *Machine Learning and Auditory Psychophysics: Unveiling Tone-in-Noise detection.* Berlin Brain Days, Berlin, Germany
- Schönfelder, V. H., & Wichmann, F. A. (2010). *Machine Learning in Auditory Psychophysics: System Identification with Sparse Pattern Classifiers.* KogWis 2010 - 10th Biannual Conference of the German Society for Cognitive Science, Potsdam, Germany
- Schönfelder, V. H., & Wichmann, F. A. (2011). *Peering into the Black Box – Using sparse feature selection to identify critical stimulus properties in audition.* Berlin Brain Days, Berlin, Germany

Posters

- Schönfelder, V. H., & Wichmann, F. A. (2008). *Machine Learning and Psychophysics: Unveiling Tone-in-Noise detection.* PhD Symposium at BCCN Conference Munich, Germany
- Schönfelder, V. H., & Wichmann, F. A. (2009). *Machine Learning in Auditory Psychophysics: System Identification beyond Regression Analysis.* ITD Processing 2009 meeting, Frauenwörth/Chiemsee, Germany
- Schönfelder, V. H., & Wichmann, F. A. (2009). *Machine Learning in Auditory Psychophysics: System Identification beyond Regression Analysis.* Berlin Brain Days, Berlin, Germany
- Schönfelder, V. H., & Wichmann, F. A. (2011). *Extracting Auditory Cues in Tone-in-Noise detection with a Sparse Feature Selection Algorithm.* Frontiers in Computational Neuroscience, Conference Abstract. BC11 - Computational Neuroscience & Neurotechnology Bernstein Conference & Neurex Annual Meeting 2011, Heidelberg, Germany

Contents

List of Figures

Chapter 1

Introduction

According to Gustav Fechner, the principal founder of the psychophysical research program, the central aim of this scientific discipline is to understand the functional relationship between the physical and psychological world (Fechner, 1860, chap. II). In experimental practice, this goal translates to explaining the connection between the stimulus and the percept. However, the percept—as a subjective entity—can not be objectively evaluated. Instead, modern psychophysical research measures behavior in specific psychophysical paradigms, such as perceptual detection or discrimination tasks, and attempts to describe the underlying mechanisms on a functional level. As natural scientists, psychophysicists work towards explaining perceptual processing not merely on a qualitative but also on a detailed quantitative level. In that process, a principal question arises: To what extent and in which way does observer behavior depend on specific stimulus features? The present doctoral dissertation describes a technique that offers answers to that question based on the analysis of psychophysical data. Concurrently, it allows the construction of detailed quantitative behavioral models. Although it is applied here in the context of audition, it is universally applicable in many kinds of perceptual studies. In the following, I discuss the general conceptual background related to the quantitative description of observer behavior and introduce the basic ideas behind the proposed data analysis procedure. The final paragraphs of this introduction provide a short summary of the following chapters.

Ever since Galileo Galilei began to capture natural phenomena in mathematical formulas, science has more and more refined the following general approach to provide quantitative explanations of scientific observations: constructing a simplified mathematical model that captures the essential causal components of the underlying mechanisms. Such a model is considered to provide a general account of the phenomenon and can usually be used to make independent predictions.

In a psychophysical setting the chemical, biological and neural processes that determine an observer's behavior might be manifold and extremely complex. Nevertheless, comparatively simple mathematical models provided an appropriate description of the observed behavior—at least in well-controlled experimental settings. Ideally, such models should describe human behavior not only in the broad terms of psychometric functions, which merely summarize the observer's behavior over a large number of trials, but on a fine-grained trial-by-trial level, the ultimate goal of "molecular psychophysics" (Green, 1964). In my thesis I work toward a precise quantitative description of the listener's decision mechanism in a basic hearing task. The models that are established for that purpose provide detailed insights into the stimulus processing that underlies individual behavior.

Generally, two techniques were employed for establishing models of psychophysical processes. On the one hand, beginning with Fechner, experimenters relied on their

intuition while establishing a model that may explain their observations (Fechner, 1860). Depending on the outcome of further experiments, details of these models were then adapted and refined. Fletcher and later researchers followed this classical approach while establishing the concept of the *auditory critical band* (Fletcher, 1938, 1940; Hartmann, 1998). Starting from the observation that the masking of a pure tone with a surrounding band of noise depended on the bandwidth of the noise only up to a certain critical value, they tried to explain this phenomenon with a model of an auditory band-pass filter. They further refined this model by allowing for non-rectangular shapes and asymmetries and determined the exact shape of the auditory filter in subsequent experiments (e.g., using notched noise).

While preparing his doctoral dissertation in the mid-1960s, Ahumada (1967) used an inverse approach following the advice of E. C. Carterette and M. P. Friedman: He tape-recorded a large number of stimuli as they were presented to the observers and then explored this data for correlates between stimulus properties and observer responses. This procedure, later termed "correlation analysis" and published in Ahumada and Lovell (1971), proved to be seminal for model construction in general and influential in a wide array of applications from auditory psychophysics to visual neurophysiology (Ahumada, 1996; Neri and Levi, 2006).

In contrast to the classical approach detailed above, this second method is essentially based on statistical data analysis that dispenses with a concrete a-priori hypothesis of how the investigated system behaves. It can be understood as a *system identification* or *reverse engineering* approach: The system under investigation is conceptualized as a "black box". Instead of attempting without avail to take it apart, the researcher probes its properties by recording the reactions to external stimulation. Essentially, he constructs a model of the system by posing questions and analyzing the corresponding answers.

This method is particularly suitable to the analysis of psychophysical observations. Because human sensory systems are such complex mechanisms, there is little hope that—by taking it apart—we can explain or even predict what happens during a perceptual experiment. Thus, one has to accept that it will remain a "black box" for the foreseeable future. On the other hand, the power of the researcher's intuition clearly has limits when it comes to explaining intricate phenomena whose structures do not allow a simple "intuitive" explanation. These factors—together with the newly available computational power—certainly represent some of the reasons why the indirect, data-driven approach has become such a wide-spread tool in the past decade, in particular in visual psychophysics (Murray, 2011; Kienzle *et al.*, 2009; Macke and Wichmann, 2010).

The "correlation analysis" method that is commonly used in auditory psychophysics for analyzing linear dependencies between stimulus and observed behavior is based on correlation coefficients. Strictly, such an analysis only applies when the stimulus components under investigation are statistically independent (Ahumada, 1996; Richards and Zhu, 1994). As a consequence, it can only be used in experiments where the investigated stimulus properties fulfil this condition, e.g., when the stimuli consist of randomly distributed multi-tone complexes (Richards and Tang, 2006) or follow random level variations over time (Pedersen and Ellermeier, 2008). Otherwise, some of the predictors may receive a significant weight even though observers completely ignore the corresponding stimulus features.

Often however, a researcher may be interested in investigating stimulus properties that are not under direct control and exhibit *interdependencies* that can not be avoided. This situation is particularly prominent in research on audition, where the stimulus is represented as a highly constrained 1D-time series often with strict conditions on spectral properties.[1] In these circumstances, observer models that are based on correlation coefficients may provide a misleading account of the mechanisms underlying the listener's behavior. In particular, a "correlation analysis" may suggest that the observer depends on a particular auditory feature, even though he exclusively relies on a very different

[1] The problem also arises in the field of vision research: As soon as images with "natural structures" are used, interdependencies are hard to avoid.

property that is merely correlated.

In theory, when applying a more sophisticated *multiple logistic regression*, a method that is more and more widely used for analyzing auditory psychophysical data, these interdependencies should be accounted for (Alexander and Lutfi, 2004; Dye *et al.*, 2005; Richards and Tang, 2006; Pedersen and Ellermeier, 2008). However, high levels of noise such as usually present in psychophysical data limit the ability of this method to correct for strong interdependencies.

The present dissertation attempts to overcome this obstacle by imposing an additional constraint during data fitting. In the following, I propose that *logistic regression* be combined with L_1*-regularization*, a method that was developed in machine learning and recently gained more and more theoretical and practical interest. Although extended with modern components, the entire analysis procedure that I suggest is firmly based on proven concepts of *"linear weight estimates"* that are built on the assumption that the decision mechanism depends on a weighted sum of a fixed set of predictors that represent different stimulus features (Ahumada and Lovell, 1971; Gilkey and Robinson, 1986; Berg, 1989; Richards and Zhu, 1994). In this respect, this study directly follows the tradition of earlier methods for estimating the relative importance of stimulus features which have long been used for successfully modeling observer behavior in auditory experiments.

Models comprising a large set of predictors and corresponding parameters are prone to *"over-fitting"*, a situation where the model parameters are to a large degree determined by the noise in a particular data set, instead of the universal structure of the generating process (Bishop, 2006, chap. 1.1). In machine learning, a powerful method termed *"regularization"* is commonly applied to prevent an over-fitting situation. In this study, I propose to employ an L_1*-regularizer* during data fitting which has a critical advantage compared to other regularization techniques—the resulting models obtain a property called *"sparseness"*. With such a constraint, weights are suppressed when the associated predictors are not critical for explaining the data. This is in perfect correspondence with the conditions and objectives of the present study: Often, even for the simplest auditory tasks, the number of features that is potentially available is much larger than the actual features an observer uses. Using a sparseness constraint, I aim at identifying these observer-specific *"cues"*. While all potential features are taken into account as predictors during fitting, the final models should be as simple as possible retaining only those components that are absolutely necessary to explain the data. These components are then considered the critical *"cues"* of the observers, as they are both necessary and sufficient to predict the behavioral decision.

The general procedure is as follows: After collecting individual behavioral data, a linear observer model is fit to stimuli and corresponding listener's responses using an L_1-regularized logistic regression. During that process, the strength of regularization is adjusted in order to strike a balance between model simplicity and quality of fit. When the model—as a result of the fitting procedure—correctly predicts *previously unseen data*, observer and fitted model are considered functionally equivalent, not just in regard to the psychometric function but on the basis of *individual trials*. The established models allow on the one hand to predict a listener's response to a particular stimulus on single-trial basis. On the other hand, by interpreting the fitted model parameters, they enable the extraction of the critical auditory features that determine individual behavior, notably even from interdependent predictors and under noisy conditions. Additionally, with a simple extension of the stimulus-dependent linear observer model, *sequential dependencies* in response behavior are taken into account. Even though undeniably present, they are often neglected during psychophysical data analysis—and certainly up till now in all types of "correlational analyses".

In the following, the content of each of the following sections of the dissertation is summarized. Chapter 2 provides information on the theoretical background of the investigation at hand, including the concepts of linear observer models, perceptual weight

estimation and cue identification as well as an introduction to machine learning procedures such as regularization and cross-validation followed by a discussion of "molecular" psychophysics and sequential dependencies. In chapter 3, the auditory task—Tone-in-Noise detection—is explained that served both as the initial motivation for the present study and as a vehicle for developing and testing the proposed data analysis method. In a historical overview, the results of previous scientific investigations of this task are discussed. Chapter 4 details the psychophysical task employed in my experiments including the properties of the stimuli used. In chapter 5, artificially generated "psychophysical data" from the auditory detection task is used to demonstrate the efficiency and reliability of the proposed analysis procedure. Chapter 6 presents the experimental methods and discusses the empirical results of psychophysical experiments with human listeners that were performed for the present study. As a conclusion, chapter 7 provides a broad discussion of the overall results of this dissertation as well as a general summary and outlook, how this method could be further developed and applied in different settings. Several Appendices A–F offer further details on the experimental setup and procedure as well as a number of mathematical derivations.

Chapters 5 and 6 represent expanded versions of two manuscripts submitted to the *Journal of the Acoustical Society of America*. The first of the two manuscripts has been published as Schönfelder and Wichmann (2012) while the other is currently under review. Several of the concerned sections provide both a broader and deeper account of the research described in the respective chapter as compared to the original manuscripts. The reader may notice that there exists partial overlap of the content in some respects. This is due to the fact that the text is based on two independent and self-contained papers that present different aspects of the same topic. I believe that these overlaps serve as reminders and improve comprehensibility of the present work.

Chapter 2

Theoretical Background: Modeling Observer Behavior

In this thesis a method is presented that addresses a central question in experimental psychology: What are the stimulus cues that observers rely on in making perceptual decisions? The present chapter firsts presents a number of classically employed techniques for cue identification based on linear observer models. Next, the field of machine learning, including the phenomenon of over-fitting and the concepts of regularization and sparseness is introduced. Finally, the concept of molecular and quasi-molecular psychophysics are discussed along with the phenomenon of sequential dependencies in response behavior.

2.1 Linear Weighting Models

Historically, the analysis of trial-by-trial dependencies between particular stimulus properties and measured behavior has proven to be a powerful method for identifying the stimulus features that govern observer behavior (Sherwin *et al.*, 1956; Ahumada and Lovell, 1971; Ahumada, 1996; Abbey and Eckstein, 2002; Murray, 2011). Aiming for a simple and robust, yet flexible foundation for a quantitative analysis, a majority of studies assumed that observer decisions depend on a linear combination of a set of stimulus descriptors on each trial (Ahumada, 1967; Berg, 1989; Lutfi, 1995; Ahumada, 1996; Alexander and Lutfi, 2004; Richards and Tang, 2006; Abbey and Eckstein, 2006; Pedersen and Ellermeier, 2008; Macke and Wichmann, 2010). A quantitative model of the decision mechanism can then be expressed as a weighted sum of predictors p_i (a set of values characterizing the stimulus) followed by a static nonlinearity S:

"The simplest transducers are linear, so let's consider them first." (Ringach and Shapley, 2004)

$$P = S\left[\sum_i w_i p_i + b\right] \tag{2.1}$$

with the model weights w_i, a bias term b and the model output P. Depending on whether S is a sign- or a sigmoid-function, P represents binary responses or response probabilities. For example, in a YES/NO-procedure, P may represent the response—"signal" or "no-signal"—or the probability for a "signal"-response. The model output is fully determined by the characteristics of the current stimulus, the model weights and the bias variable. Those predictors for which $w_i = 0$ are entirely ignored, effectively reducing model complexity. The model represents the basis for both the evaluation of stimulus cues and the prediction of responses.

In general, the predictors p_i may be linear or nonlinear transformations of certain stimulus features. Thus, the model is linear only with respect to the predictors, but not

necessarily with respect to simple stimulus descriptors—a predictor may represent any nonlinear transformation of the physical stimulus. Therefore, the overall decision function mapping the raw stimulus to the response may be highly nonlinear, depending on the employed stimulus preprocessing. While nonlinear decision models may ultimately be required to describe observers in full detail, linear models represent an important first step and the basis for more complex models.[1]

[1] One of the currently most powerful model fitting algorithm, the so-called support vector machine (SVM), is "only" such a linear combination of nonlinearly transformed stimulus features (Noble, 2006).

2.2 Weight Estimation with Regression and Correlation Analyses

Assuming a linear observer model, the problem of finding the mapping between stimulus input and observer responses simplifies into finding an estimate of the relative weights that best explains the empirical data. Ahumada (1967) and Ahumada and Lovell (1971) were the first to estimate relative weights in an auditory detection experiment using multiple linear regression (Ahumada (2002) offers a short personal history). In a wide-band Tone-in-Noise detection task, they estimated spectral weights by regressing sound energy in different frequency components with observer ratings. In a similar vein, Berg (1989) relied on pairwise regressions of response probabilities against the values of individual stimulus components to investigate "relative temporal weights" in a multiple tone task.

"*A multiple regression analysis found for each observer the linear combination of the energies in narrow bands around the tone frequency that best predicts his total ratings.*" (Ahumada and Lovell, 1971)

Richards and Zhu (1994) and Lutfi (1995) demonstrated that—given independent and normally distributed predictors—relative weights can be estimated more easily from correlation coefficients between predictor values and binary observer responses. Similarly, Ahumada (1996) determined visual "classification images" by correlating the local amplitude of noisy image presentations with observer decisions. In an analogous approach, the so-called "white noise analysis" (or "triggered correlation" in De Boer and Kuyper (1968), "reverse correlation" in Ringach and Shapley (2004)) has been extensively used in physiology to determine the visual response characteristics of single neurons, the "receptive field". These correlational procedures, usually termed "correlation analysis", "reverse correlation" or "classification image procedure", are based on the point-biserial correlation that is determined pairwise between the values of individual stimulus components and the observer's responses. Therefore, they do not allow for interactions between predictors. Correlation-based procedures can often be simplified to computing the difference of response-conditioned averages (Lutfi, 1995; Ahumada, 2002).

"*[T]he algorithm employed to estimate the weights is computationally trivial—the correlation between the observers' responses and the stimulus magnitude.*" (Richards and Zhu, 1994)

"*Reverse correlation is a technique for studying how sensory neurons add up signals from different locations in their receptive fields, and also how they sum up stimuli that they receive at different times, to generate a response.*" (Ringach and Shapley, 2004)

Estimates of relative combination weights using pairwise regression (Berg, 1989) or correlation (Richards and Zhu, 1994) are based on the assumption that the stimulus components are independently distributed. The predictors in the model usually correspond to those stimulus features that were independently manipulated during the experimental procedure. However, many features in auditory tasks are not independent, e.g., sound spectrum and envelope structure of sound stimuli are interrelated. In this case, pairwise weight estimation techniques may result in misleading conclusions about the weight distribution.

"*[I]t is important that the experimenter is able to manipulate the magnitude of the independent random variables assumed to be combined to form the decision variables.*" (Richards and Zhu, 1994)

In correlation analysis, covarying predictors represent a principal problem that may result in unstable or misleadingly large weights, misguiding experimenters when rating the relative importance of different cues. For example, let us assume an observer relies only on a predictor A. There is another predictor B, that is correlated with A, but that the observer does not have access to. Across many trials then, both predictors A and B correlate with the responses and a pairwise correlation analysis indeed assigns a significant weight to both.

For a perfectly deterministic observer, however, only predictor A explains the response on each and every trial. For a probabilistic observer, A is a better predictor of trial-by-trial responses than B. This critical difference between true and ostensible predictors can be taken advantage of to apply a perceptual weighting analysis even under

conditions where the stimulus features of interest can not be independently manipulated and statistical dependencies between stimulus features can not be entirely avoided, e.g., in speech perception.

Generally, *multiple* regression procedures, such as used by Ahumada and Lovell (1971), are able to discern critical and ostensible cues by taking interactions between different predictors into account.[2] In contrast to methods based on the pairwise analysis of predictors and responses (Richards and Zhu, 1994; Berg, 1989), the model weights are *jointly* adjusted to optimize the model fit. The study by Ahumada and Lovell (1971) mentioned above represents one of the first occasions where multiple regression was applied to analyze auditory psychophysical data.

[2] Lutfi (1995) suggests to account for statistical dependencies between predictors in correlation analysis by iteratively computing partial correlation coefficients. However, depending on the number of predictors and the correlational structure, this process quickly becomes impractical. It provides no advantage over multiple regression techniques.

In another early attempt at modeling single-trial behavior in an auditory task, Gilkey and Robinson (1986) used multiple logistic regression to estimate the relative weights attributed to auditory features. In contrast to linear procedures, logistic regression provides estimates for response *probabilities*, which is particularly appropriate in psychological studies where observer behavior is generally better described in a probabilistic fashion. Logistic regression in general has a long history rooted in population growth statistics (Cramer, 2003; Verhulst, 1838), the description of binary sequences (Cox, 1958) and was later incorporated in the theory of Generalized Linear Models (Nelder and Wedderburn, 1972; Dobson and Barnett, 2008). More recently, a number of auditory studies relied on multiple logistic regression to derive relative spectral weights (Alexander and Lutfi, 2004; Dye *et al.*, 2005; Richards and Tang, 2006) or temporal weights (Pedersen and Ellermeier, 2008). However, weight estimates from multiple regression procedures become less reliable with increasing covariance and, in particular with large numbers of predictors, results become sensitive to noise. In addition, when predictors are linear combinations of each other, the solution of a multiple regression becomes ill-defined.

"*We believe that the generalized linear models here developed could form a useful basis for courses in statistics. They give a consistent way of linking together the systematic elements in a model with the random elements.*" (Nelder and Wedderburn, 1972)

Recent methods from machine learning allow a dissociation of critical cues even in the presence of strong covariances, including linear dependencies, and noise. By adding constraints on the observer models during fitting that impose the assumption of "parsimony", they prevent over-fitting and avoid instabilities of earlier weight estimation procedures as explained in detail in the following section.

2.3 Machine Learning, Over-Fitting and Regularization

During recent years, powerful statistical analysis tools have been developed to form the field of machine learning:

"*Building on thirty years of analysis of learning processes, in the 1990s the synthesis of novel learning machines controlling generalization ability began.*" (Vapnik, 2000, p. 8)

Since then, techniques such as support vector machines (SVM) and regularized multiple logistic regression have been successfully applied in a number of scientific fields (Noble, 2006; Schölkopf and Smola, 2001). Generally, machine learning algorithms can be characterized as methods for identifying complex multidimensional input-output mappings from diverse kinds of data, similar in principle to fitting the linear observer model in Eq. 2.1 to stimulus-response data:

"*SVM analysis can be applied to a wide variety of biological data. As we have seen, the SVM boasts a strong theoretical underpinning, coupled with remarkable empirical results across a growing spectrum of applications.*" (Noble, 2006)

"*The result of running the machine learning algorithm can be expressed as a function $y(x)$ which takes [...] x as input and that generates an output vector y, encoded in the same way as the target vectors.*" (Bishop, 2006, chap. 1)

In general, these methods aim at identifying hidden structure in complex data sets. Typically, machine learning algorithms are simultaneously trained on sample data and the corresponding *correct* "classes", the so-called "ground truth", e.g., pictures of hand-written digits ("data") and their true identity ("class"). In behavioral data analyses, a different approach is taken. The goal is to obtain an algorithm that implements the observer's

"*We used Support Vector Machines, a well-known classification algorithm in the field of machine learning, to classify the calls of the different bats.*" (Yovel *et al.*, 2009)

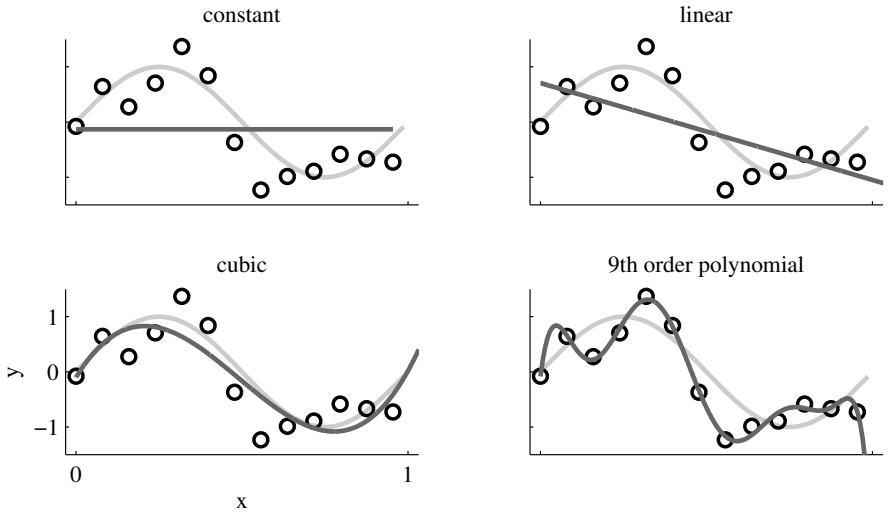


Figure 2.1: Illustration of over-fitting. Plots of polynomials having various orders M, shown as dark curves, fitted to the black data points that were probabilistically generated from the light gray curve. For constant and linear fits, the model under-fits the data, for the 9th order polynomial, the model over-fits and is mostly determined through noise in the data. The cubic fit best follows the original sine function from which the data was generated. This plot was inspired by Fig. 1.4 in Bishop (2006).

decision mechanism instead of one that responds "correctly". Thus, the "machine" is trained to predict the observer's answer, even though it might oftentimes be wrong. In the context of behavioral studies of perception, this technique has been employed for system identification in visual psychophysics in humans (Wichmann *et al.*, 2005; Kienzle *et al.*, 2009; Macke and Wichmann, 2010) and auditory processing in bats (Yovel *et al.*, 2008, 2009).

In machine learning, a major concern when adjusting complex models to limited amounts of data is over-fitting. This term describes a situation where the parameters of an overly complex model are to a large degree determined by the noise in a particular data set, instead of the universal structure of the generating process. Such a model reproduces the current data, seemingly being "correct". But as a consequence of over-fitting, it is in reality a poor descriptor of the process underlying the data, and the model is unable to predict future data originating from the same process. This directly conflicts with a central aim when fitting a model: to be able to accurately predict new data.[3] A plot from the book illustrating the phenomenon is shown as Fig. 2.1. In addition, when the model fit depends on noise fluctuations, the estimated parameters become fundamentally unstable. In particular for models with a large number of parameters, it is important to rule out the possibility of over-fitting. Therefore, the quality of fit of a model should not be tested on the same data that was used for fitting—as it is usually done in the behavioral sciences.

[3] An extensive discussion of the phenomenon and consequences of over-fitting can be found in the introductory chapter 1.1 ("Example: Polynomial Curve Fitting") of Bishop's *Pattern Recognition*" (Bishop, 2006).

As a common practice in machine learning, a model is first "trained" to a subset ("training set") of the data, i.e., model parameters are fit to match the data:

"The precise form of the function y(x) is determined during the training phase, also known as the learning phase, on the basis of the training data." (Bishop, 2006, chap. 1)

This step is equivalent to model fitting in classical correlation or regression techniques. To confirm that the model indeed captures the general structure underlying the data instead of over-fitting, it is necessary to measure how accurately the model predicts *new*

data. Therefore, in a second step, the quality of the fit is evaluated, or "tested", on the remaining chunk of data ("test set") that was not touched during training:

"Once the model is trained it can then determine the identity of new [input samples], which are said to comprise a test set. The ability to categorize correctly new examples that differ from those used for training is known as generalization." (Bishop, 2006, chap. 1)

During both processes, the data on which the model parameters are estimated (training set) must be treated strictly independently from the data on which the agreement between model and data is determined (test set), including any kind of data transformations.

While splitting the data helps to identify whether over-fitting has occurred, a second method called "regularization" is required to prevent over-fitting. The initial motivation for introducing regularization was to transform ill-conditioned optimization procedures into well-posed problems (Tikhonov, 1963; Vapnik, 2000, p. 9).

"Regularization theory was one of the first signs of the existence of intelligent inference." (Vapnik, 2000, p. 9)

"In the second half of the century a number of very important real-life problems were found to be ill-posed. In the middle of the 1960s it was discovered that if [...] one minimized [a] so-called regularized functional [...], then one obtains a sequence of solutions that converges to the desired one. [...] The influence of the philosophy created by the theory of solving ill-posed problems is very deep. Both the regularization philosophy and the regularization technique became widely disseminated in many areas of science, including statistics." (Vapnik, 2000, p. 9-10)

Regularization prevents over-fitting by introducing a penalty on overly complex models. It is adjusted by training the model using different regularization strengths resulting in models of varying complexity. The optimal strength is chosen as to maximize model agreement during *testing*. In this way, a model is identified that is complex enough to explain the data and simple enough to not over-fit (Duda *et al.*, 2001, chap. 6.11). The intrinsic complexity of the data, generally unknown a-priori, determines the necessary complexity of the model. While optimizing regularization as described above, the complexity is adjusted to be appropriate for the data set.[4]

[4] For an extensive discussion of the theory and practice of model selection, refer to the "Model selection"-special issue of the *Journal of Mathematical Psychology*, e.g., Forster (2000) or Zucchini (2000).

Commonly, a regularized fitting procedure is based on a two-part error function that is minimized by adapting the model parameters (i.e., the weights **w**):

$$E^* = E(\mathbf{P}, \mathbf{p}, \mathbf{w}) + \lambda \sum_j |w_j|^n \tag{2.2}$$

The first term expresses how much the model diverges from the data (consisting of the stimulus-dependent predictors **p** and observer responses **P**). The second "regularization" term constrains (or "shrinks") the values of the model weights w_j, reducing the complexity of the model. The regularization parameter λ controls the trade-off between quality of fit and compliance to the regularization constraint. The norm-parameter n usually takes the values 1 or 2, and controls in which way model complexity is reduced. Up until recently, machine learning focused almost entirely on L_2-regularization ($n = 2$), due to its ease of implementation—optimizing a square-law function is much easier than a non-differentiable absolute-value function.

"Despite the additional computational challenge posed by L_1-regularized logistic regression, compared to L_2-regularized logistic regression, interest in its use has been growing." (Koh *et al.*, 2007)

An L_2-regularization ($n = 2$, termed "ridge regression") generally results in weights that are more uniformly distributed minimizing their squared norm. Such a model generally exhibits a smooth input-output function. In contrast, L_1-regularization ($n = 1$, termed "lasso") promotes zero-valued weights, effectively reducing the number of predictors and resulting in "sparse" models (Tibshirani, 1996; Koh *et al.*, 2007).

"We propose a new technique, called the 'lasso' [...]. It shrinks some coefficients and sets others to zero, and hence tries to retain the good features of both subset selection and ridge regression." (Tibshirani, 1996)

"Sparseness" is a meaningful and rational constraint in the context of modeling. As already stated by William of Ockham (c.1280-c.1349): *"Pluralitas non est ponenda sine necessitate"* (Plurality is never to be proposed without necessity) and *"Frustra fit per plura, quod potest fieri per pauciora"* (It is in vain to do with more what can be done

with fewer) (Thorburn, 1918). These statements are known as the postulate of "Ockham's Razor" or the "law of parsimony" (Forster, 2000). Whether in machine learning, physics or psychology, it is now generally assumed that a model that uses fewer parameters to explain a phenomenon is to be preferred against more complex ones. In the present case, this corresponds to a model that relies on fewer stimulus descriptors to predict observer decisions.

2.3.1 L_1-Regularization, Sparseness and Feature Selection

The central purpose of fitting a model to behavioral data was to identify stimulus features that determine an observer's response. Striving after model "sparseness", the present study relies on L_1-regularization as a method for feature selection. Essentially, it suppresses those weights that do not contribute to improving model predictions while the remaining weights and associated predictors are identified as critical for explaining the data (Donoho, 2004; Hastie *et al.*, 2009, chap. 4.4.4).

"The objective of this paper is to show how many kinds of geophysical data fitting can be made to be robust. In particular, all the calculations we now do in solving overdetermined linear simultaneous equations [...] can be made robust, instead, by minimizing summed absolute values of errors. [...] With robust modeling methods [...], [t]he earthquake may be properly located even if it knocks down some of the telephone lines." (Claerbout and Muir, 1973)

In general, finding the smallest subset of features that optimally predicts observer decisions amounts to a combinatorial problem that grows exponentially with the number of candidate features and thus quickly becomes intractable. Fortunately, sparse regularization offers a reliable and efficient approximation to solve this problem. The general idea of using L_1-norm minimization for feature selection and robust data analysis is quite old, e.g., it has already been applied in seismics in the 1970s by Claerbout and Muir (1973) (for a short history see Tropp (2006) and Koh *et al.* (2007)).

In the context of regularization, the L_1-norm was later valued for providing results that were sparser than those from L_2-based ridge regression and that could be used for subset selection (Tibshirani, 1996). Meanwhile, Donoho (2004) has mathematically proven that the *unique sparsest solution* can indeed be found via convex optimization of the L_1-regularized problem, a finding that also holds in the presence of noise (Donoho *et al.*, 2006). *Convexity* of an optimization problem implies that the surface of the objective function, which measures the agreement between data and model depending on the model parameters, has a unique extremum, which can be found through gradient descent or related methods. Non-convex optimization problems always run the risk that the optimization algorithm gets stuck in non-optimal local extrema.

In fact, a regularizer with L_1-norm represents the closest convex approximation to an L_0-regularizer. An L_0-regularizer simply counts the number of nonzero weights, or equivalently the number of predictors used to model the data. However, the corresponding optimization task amounts to a non-convex problem, for which no efficient optimization algorithms exist and which is as difficult as testing all combinations (Donoho, 2004).

"In general, solving (P_0) [the sparsest possible representation] requires combinatorial optimization and is impractical. The L_1-norm is in some sense the convex relaxation of the L_0-norm." (Donoho, 2004)

By contrast, L_1-regularization can even be applied in challenging settings, e.g., with large numbers of predictors, where other heuristic attempts to enforce sparsity perform poorly (Donoho, 2004). In addition, L_1-regularization is highly insensitive to correlations between predictors, e.g., correlation coefficients up to 0.9 did not effect the weight estimates in a two-predictor case (Tibshirani, 1996). L_1-regularization still poses a computational challenge, as the objective function, containing an "absolute value" function, is non-differentiable. Consequently, gradient-descent methods commonly used in model fitting can not be applied. In recent years, however, efficient algorithms for L_1-regularized optimization have been developed (Koh *et al.*, 2007; Park and Hastie, 2007; Hastie *et al.*, 2009, chap. 18.4).

"Sparse approximation problems arise throughout electrical engineering, statistics, and applied mathematics. One of the most common applications is to compress audio, images, and video. Sparsity criteria also arise in linear regression, deconvolution, signal modeling, preconditioning, machine learning, denoising, and regularization." (Tropp, 2006)

L_1-algorithms have received increasing attention in a wide array of applications (Tropp, 2006), in particular in machine learning in the context of support vector machines (Matthew *et al.*, 2005) and logistic regression (Ryali *et al.*, 2010). The L_1-regularizer is particularly well suited to the analysis of behavioral data, where a large number of potential features might be studied while often only few are critical to the observer. An L_1-regularized procedure is able to identify this small subset because it favors the model

with the "simplest explanation" of the data, i.e., the one that relies on the smallest set of predictors as the explanatory variables.

Additionally, L_1-regularization provides a critical advantage regarding the central purpose of this study as it offers a natural solution to a problem that was discussed in section 2.2: the problem of correlated predictors that may arise during the analysis of psychophysical data, in particular in the context of auditory perception. An L_1-regularized regression allows a dissociation of critical auditory cues among a set of correlated predictors, because it identifies exactly those that are both necessary and sufficient to predict the data. As discussed earlier, this would have been much more difficult or even impossible with standard non-regularized regression procedures. As discussed in the following section, these critical predictors are then considered to correspond to the observer cues. During prediction, only these cues are used while ignoring all non-critical features.

2.3.2 Extracting Observer Cues from Sparse Models

This section presents the principal ideas behind using sparse linear-weighting models to identify a listener's perceptual processing from his responses in a psychophysical experiment.

Using a regression procedure, an observer model is fit to auditory stimuli and corresponding observer responses. If the fitting is successful, the model should then be able to capture the perceptual decisions of the listener on a trial-by-trial level. In terms of behavior, it can thus be considered at least as functionally equivalent. The proposed cue extraction procedure goes one step further by making the central assumption that the mechanisms the model employs to predict the observer's decisions are similar to the actual perceptual processes of the listener. Essentially, it assumes that the equivalence between model and observer may not only be functional, but that there may be a stronger mechanistic equivalence, too. The decision mechanism of the model is entirely defined by the model weights (and the bias parameter). The weight values determine to which extent and—depending on the sign—in which direction a certain predictor influences the decision. Thus, by analyzing the model weights, one can infer whether and to what extent a particular stimulus property, which is represented by a predictor, determines the listener's behavior.

Using a sparse fitting technique, the present study identifies the unique necessary and sufficient predictors that explain the behavioral data. To my knowledge, this is the first time that a sparse L_1-regularization technique is applied to identify perceptual cues from psychophysical data. Of course, as any other cue extraction technique based on linear-weighting models, the proposed method can only identify stimulus cues that are made available as predictors. The extracted cues must always be interpreted in the light of the present model predictors. However, weak predictive power of a fitted model is a strong indicator that essential observer cues are not represented by the predictors. Generally, there is no unique correct answer to the question which stimulus cues observers use. Often, there are multiple ways to project particular stimulus properties on model predictors. As demonstrated in chapter 6, the weight distribution itself may provide hints as to how to adjust the stimulus preprocessing to obtain more meaningful and simpler representations of observer cues.

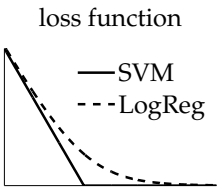
2.3.3 Discussion of Alternative System Identification Techniques

Restricting the analysis to the set of linear models may appear as a strong limitation. Powerful nonlinear techniques for data analysis and model fitting have been developed in recent years (Schölkopf and Smola, 2001), methods based on multi-layered neural networks have been available even earlier (Rumelhart *et al.*, 1986; le Cun, 1988). However, there are serious difficulties when applying these techniques in the current setting, due

to which I conclude that a linear logistic regression provides an excellent choice for the present kind of analysis.

First of all, techniques based on fitting multi-layered neural networks have the main disadvantage that they can not guarantee that the optimal model given the data and model constraints can be found (Rumelhart *et al.*, 1986; Vapnik, 2000) . Depending on the starting conditions of the fitting procedure, the optimization may get stuck in local extrema that do not correspond to the globally optimal model estimate. Further, there is a large number of a priori unrestricted model parameters which have to be optimized, such as the number of hidden layers, the number of neurons in each layer and the structure of connections within and across layers. Given the high amount of noise in observer data and the resulting uncertainty when estimating the model likelihood, it may be very difficult or even impossible to distinguish the optimal model structure in such a high dimensional space. By contrast, a sparse linear logistic regression represents a convex optimization problem that always converges to the global optimum while only one hyper-parameter—controlling the regularization strength—has to be optimized (Candes and Tao, 2005). In addition, it is very hard to analyze the decision mechanism underlying a trained multi-layer neural network.[5]

[5] One way is to "project-out" a polynomial-function of very high degree, that one has to "prune" afterwards (Gary Green, pers. communication).



While the support vector machine (SVM) relies on a hinge-loss function, the loss curve for logistic regression (LogReg) is smooth. Both have the same asymptotes.

Generally, linear support vector machines are the binary analogue to a regularized logistic regression with only a small difference regarding the loss function, which determines how strongly incorrectly classified data samples are weighted (Schölkopf and Smola, 2001). In practice, this difference is minor—the weights estimated with a linear SVM or a logistic regression hardly differ. Both SVMs and logistic regression offer the same advantages regarding model fitting (convexity and few hyper-parameters). However, SVMs have one significant disadvantage regarding model predictions: Generally, they only provide a binary prediction instead of a graded probability estimate. They are thus not able to accurately capture the probabilistic nature of behavioral data.

Nonlinear support vector machines, based on polynomial or radial-type basis functions have proven to be a very successful extension to linear SVMs and are able to capture hidden structure in a wide range of data types (Noble, 2006). However, the focus in such studies commonly lies on successful prediction while the mechanisms underlying the decision process remain hidden. In fact, the optimization of the model weights takes place in a so-called "dual space" or "kernel-induced feature space", while the transformation of the raw data from the primal space is only defined implicitly and not directly accessible (Schölkopf and Smola, 2001). As a result, the precise decision mechanism of a fitted nonlinear SVM can in general only be obtained through sophisticated indirect methods, e.g., Wiener and Volterra-series for polynomial SVMs (Franz and Schölkopf, 2006). More easily tractable approximate solutions have been suggested (e.g., see Kwok and Tsang (2004)) and some authors successfully applied that approach—mostly thanks to an unexpectedly simple data fitting result (Kienzle *et al.*, 2009). Still, the interpretability of such an analysis is highly constricted by the amount of noise in the data and the complexity of the resulting solution.

In conclusion, for the present purposes, a sparse logistic regression provides the best trade-off in terms of its capabilities and limitations: Models can be fit very reliably to data even in the presence of a strong noise component, as should be expected for data from naïve observers in psychophysical experiments. Second, once the model is fit, it provides probability estimates for the predicted responses that can be directly compared with the response probability of the observers. And finally, and most importantly, the mechanism by which the model makes predictions—which is used to make inferences on the perceptual processing of the observers—can be directly obtained from the model weights. If necessary, nonlinear aspects of the perceptual processing may be introduced by using nonlinear stimulus preprocessing steps before the data are fit with the linear model. Such a preprocessing also allows the incorporation of domain knowledge collected from anatomy, physiology or psychophysics. Only if such an approach—the combination of

nonlinear preprocessing and a linear decision stage—should fail, would it be necessary to turn to nonlinear model-fitting techniques.

2.4 Cue Identification through Stimulus Manipulation

In principle it is possible to control some stimulus properties separately from others through careful signal manipulations, the realm of classical psychophysics. For the canonical example of Tone-in-Noise detection presented in chapter 3, studies relied on equal-energy stimuli (Richards *et al.*, 1991), envelope modulated pure tones (Richards and Nekrich, 1993), roving-level sounds (Kidd *et al.*, 1989), or chimeric stimuli (Davidson *et al.*, 2009b). The underlying idea is the following: If behavior changes as a function of a particular feature, all other stimulus properties being uninformative, this feature must be critical for the observer decision and vice versa.

"One approach that has been successfully used to evaluate the role of envelope and fine-structure cues [...] involves the use of chimeras. Chimeras are stimuli formed by combining the envelope from one stimulus with the fine structure from another." (Davidson *et al.*, 2009b)

Often however, controlling a single stimulus property without affecting others is difficult in practice. In addition, the necessary manipulations may introduce differences from non-manipulated stimuli, e.g., subtle artifacts, which in turn may unintentionally influence observer judgements. Furthermore, this approach may even become practically impossible when individual features are strongly interdependent or with large sets of candidate features. For these reasons, the present study strictly relies on non-manipulated stimuli that correspond to the original perceptual task as presented in chapter 3.

2.5 Molecular Psychophysics and Sequential Dependencies

Almost half a century ago, Green (1964) introduced the concept "molecular psychophysics":

"[T]he subject matter of molecular psychophysics must be the individual response of the individual subject."

For David Green, one of the goals of psychophysics is to be able to explain and predict observer behavior on a single-trial level. He mentions two main virtues of such an analysis:

"First, a trial-by-trial analysis of the observer's response may provide a critical and efficient test of a hypothesis that is impossible on the molar level. [...] Second [...], molecular analysis provides a rigorous and demanding test of ideas about the manner in which nonstimulus variables influence the judgment process." (Green, 1964)

He noted that, unfortunately, the observer's responses most likely not only depend on the stimulus, as the "external factor", but also on "internal factors" that are not accounted for by models that only consider stimulus predictors. In order to estimate the relative influence of external and internal determinants of the decision, he proposes to quantify observer "consistency" in multi-pass experiments. In such a configuration, the exact same stimulus is presented multiple times. For example, observer consistency could be quantified by presenting a set of 100 stimuli twice and measuring the proportion of stimuli to which the listener responded identically in both runs.

Among those models that explain and predict the listener's decision based on the relationship of stimuli and corresponding responses, there can not be a better estimator of the observer's decision than the observer himself in an earlier trial. Necessarily then, the consistency determines how precise the behavior of a listener can be predicted by any such observer model. In Appendix G I derive the corresponding upper limit for model-observer agreement depending on observer consistency in a two-pass experiment. A more complex derivation of the same limit can be found in Neri and Levi (2006).

Given that internal fluctuations do not allow a perfect "molecular" prediction of responses on every single trial, Green proposed that the behavior of the observer can still be quantified precisely in a "quasi-molecular" manner by estimating the *probability* that a listener would respond in a certain way to a particular stimulus.

Most "internal factors" determining behavior, such as attention or fatigue, is hard to assess or even control. Often their effect is simply subsumed under the term "internal noise", being *"synonymous with the presence of effects that are not fully understood"* (Green, 1964).

Even so, there is a unique component that is likely to influence a listener's decision and that an experimenter has direct access to: *the sequence of past responses*. In his extensive discussion of observer consistency, Green (1964) pays particular attention to this point and argues that *"[t]here is little question that the subject is influenced by the sequence of past responses"*. Anyone who has ever participated in a psychophysical experiment will agree: After repeatedly pressing only one button in a binary choice experiment, the urge increases to press the other button at some point—presumably because such a monotonous stimulus sequence appears increasingly unlikely. Listeners may also have a tendency to try and discover patterns in the stimulus sequence which also results in sequential dependencies in their responses.

At the time when Green wrote his early manuscript, the importance of the response factors had almost exclusively been estimated in conditions with little or no sensory stimulus and it had yet to be investigated how pronounced this effect was with supra-threshold stimulation. The strength of sequential dependencies in concert with other uncontrolled internal factors critically determines how well an observer can be predicted by a model that is based on the stimulus alone.

As Green (1964) writes,

"If response mechanisms could be understood in detail, then the experimenter [...] could, in theory, predict the trial-by-trial behavior of the subject."

As long as the influence of other internal factors remains unknown, this ambitious aim will never be fully achieved. Nevertheless, being able to quantitatively describe the mechanisms underlying the response-sequence effect at least improves the quality of single-trial predictions.

Since Green's early publication, other studies investigated the effect of sequential response dependencies and its existence was proven early on (Howarth and Bulmer, 1956; Green *et al.*, 1977). Later, sequential dependencies were studied in the framework of signal detection theory with a focus on their influence on criterion setting (Treisman and Williams, 1984; Lages and Treisman, 1998) or estimates of the psychometric function (Fründ *et al.*, 2011a). Even more recently, Busse *et al.* (2011) found particularly strong sequential dependencies in an animal study with mice, where in extreme cases the immediately preceding trial determined the current response as strong as the current stimulus. An analysis of a number of data sets from human psychophysical studies collected in both visual and auditory tasks concluded that a majority of observers exhibited these effects albeit to a weaker extent (Fründ *et al.*, 2011b, 2012).

As a consequence, this study attempts to directly model the effect of sequential dependencies in response behavior. For that purpose, the linear observer model in Eq. 2.1 which traditionally only allows for the influence of the current stimulus on observer behavior, is supplemented with information on the listener's behavior. In concrete terms, the decisions in several immediately preceding trials are included as individual binary predictors which are associated to a corresponding relative "behavioral" weight. This approach is closely related to the method used in an animal study by Busse *et al.* (2011), though these authors considered only the most recent trial. To my knowledge, the proposed analysis method is the first attempt of this kind in the context of human psychophysics, and almost certainly the first attempt to jointly estimate causal perceptual

cues together with behavioral determinants based on stimuli and response history.

Chapter 3

The Auditory Task: Tone-in-Noise Detection

The central intention of the present study was to develop a quantitative description of the perceptual processing in Tone-in-Noise (TiN) detection—an as yet not fully resolved problem rooted in classical auditory psychophysics. The following section discusses the historical developments surrounding this paradigm and its connection to the concept of the critical band. Later sections shed some light on the evolution of Signal Detection Theory as well as an experimental technique called "Frozen-Noise" both of which were tightly linked to the TiN paradigm. Subsequently, I discuss studies of the perceptual cues and quantitative models for TiN detection which provide the immediate foundation for the present study. In the final section, I present the details of the experimental procedure used while collecting TiN detection data in a hearing laboratory.

3.1 The History of Tone-in-Noise Detection

The auditory system is a highly intricate, sensitive and hard-to-access mechanism, whose function should best be studied in an intact system—as we know now, it includes active feedback components that are essential to its capacities (Gold, 1948; Kemp, 1978). Generally, it is efficient to study such complex neural systems in parallel both on a *functional/algorithmic* level as well as from a *physiological/implementational* perspective. Psychophysical hearing experiments offer a powerful approach to elucidate the overall function of the auditory processing mechanism on the algorithmic level, which constitutes the general focus of the present thesis.

When studying a complex system, such as the human auditory system, it is convenient to start with its most fundamental functions and properties. In *Steven's Handbook of Experimental Psychology* (Pashler and Yantis, 2002) Laurel Carney writes

"A primary function of the auditory system is to encode the frequency spectrum of a sound, that is, the level of energy at each frequency within a stimulus. The physical dimensions of frequency and sound level convey most of the important information in sounds."

Harvey Fletcher, as the Director of Research at the Bell Telephone Laboratories, was the first to rigorously study auditory frequency analysis with psychophysical methods. Swets *et al.* (1962) summarize his seminal studies on the hearing process:

"For the better part of a century, attempts to specify the process of auditory frequency analysis were based almost exclusively on anatomical and physiological evidence. Then,

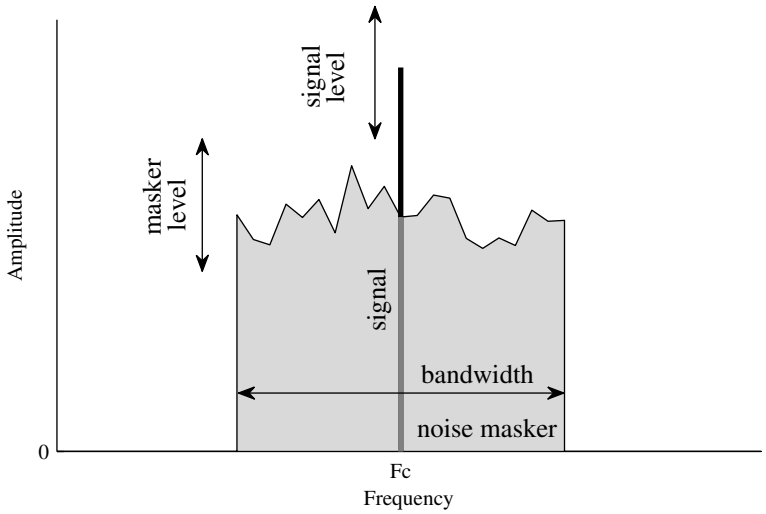


Figure 3.1: Illustration of a standard Tone-in-Noise detection stimulus with a band-limited noise masker (gray) centered on a signal tone (bold black) at frequency Fc. In each trial of a YES/NO detection task, the signal may be present or not. The signal-to-noise ratio (SNR) may be adjusted by modifying the masker or signal level.

in 1940, Fletcher presented psychophysical data that gave a new form to the problem. He reported an experiment showing that only noise components in a narrow region about a pure tone are effective in masking the tone. This region he termed the 'critical band'."

"*The simple psychophysical task of detecting a pure tone in the presence of a random noise has been of fundamental importance in the development of auditory theory. Essentially all modern models of auditory processing include an initial array of narrowband filters, and for detection of a tone in noise, it is generally assumed that a filter tuned near the frequency of the tone is used.*" (Carney *et al.*, 2002)

The Tone-in-Noise (TiN) detection experiment was the central paradigm that gave rise to this discovery. During such an experiment, a listener is asked to detect a pure tone masked by background noise, a sample stimulus as illustrated in Fig. 3.1. The masker either consists of wide-band Gaussian noise, band-limited noise centered on the signal or notched noise with an "empty region" surrounding the tone. As described by Swets *et al.* (1962), it was first introduced by Fletcher (1938, 1940) who used it to determine the characteristics of a hypothetical auditory filter. With his idea of the "critical band", Fletcher established one of the most important concepts in the theory of auditory perception. The concept of the auditory filter is useful for explaining a range of psychophysical phenomena, including TiN detection, loudness perception, nonlinear distortion, frequency discrimination, binaural interactions as well as the perception of pitch, individual harmonics and relative phase (Hartmann, 1998).

"*The concept of auditory filtering, or critical band, is the single most dominant concept in auditory theory. Mainly, it is defined by masking experiments, but it is also said that the critical band is that frequency bandwidth or frequency separation where perceptual properties change suddenly.*" (Hartmann, 1998, Sec. "The Ubiquitous Critical Band")

Relying on the TiN detection paradigm, several early studies attempted to characterize the properties of the auditory filter hypothesized by Fletcher. Critical bandwidth was usually measured by varying the bandwidth of the noise (Fletcher, 1940; Schafer *et al.*, 1950; Swets *et al.*, 1962) or by comparing the performance of the observers with theoretical detection systems (Hawkins and Stevens, 1950; Jeffress, 1964; Green and Swets, 1966). Fletcher concluded from his data (see Fig. 3.2)—or at least assumed for simplicity—that the auditory filter was rectangular in shape. Later on, more precise measurements found that the transition between the rising and constant slopes of the threshold curve was in fact not as sharp but gradual instead, suggesting a corresponding filter that was smooth in shape. The shape of the auditory filters was not only of interest as regards the frequency discrimination, but also the absolute hearing capability. The joint curve of the combined critical band filters is assumed to generate the hearing threshold curve (Schafer *et al.*, 1950).

Unfortunately, depending on the particular assumptions and experimental settings, the width of the critical filter varied over almost one order of magnitude (Ahumada,

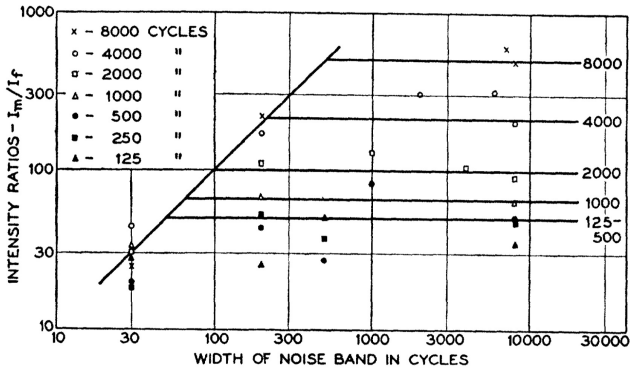


FIG. 17. Ratio of the intensity of the mask tone to the intensity per cycle of the noise plotted against the width of the noise band in cycles. The critical band width in cycles is numerically equal to the ratio of the intensity of the tone masked to the intensity per cycle of the noise producing masking and always corresponds to $\frac{1}{2}$ mm of length on the basilar membrane.

Figure 3.2: Tone-in-Noise detection thresholds (y-axis) for varying noise bandwidths (x-axis) and signal frequencies (symbols) as measured by Harvey Fletcher. Reprinted figure with permission from H. Fletcher (1940), "Auditory patterns", Reviews of Modern Physics 12, 47–65. Copyright 2013 by the American Physical Society.

1967). Very early it was recognized that there may not be a single bandwidth, but that the properties of the auditory system are adapting depending on the task (French and Steinberg, 1947; Swets *et al.*, 1962; Jeffress, 1964). As Swets *et al.* (1962) state:

"[I]t seems unlikely that all of these experiments are measuring the critical band, a fixed property of the auditory system that exists independent of experiments. It seems more reasonable to suppose that the parameters of the auditory system are not fixed, specifically, that they may vary from one sensory task to another under intelligent control."

In conclusion, even though over the course of the 20th century the critical band had a unifying effect in auditory theory, it has been more critically scrutinized in recent decades. As Hartmann (1998) writes:

"Historically, the critical band has been a unifying element in the explanation of many psychoacoustical effects, particularly their dependences on bandwidths or frequency separations. [...] According to contemporary view, most psychoacoustical effects are actually rather complicated, and although auditory filtering is an important part of any explanatory model, it is rarely enough by itself."

In any case, whether auditory researchers attempt to verify or disprove the idea of the "critical band", they all agree that it is a meaningful and valuable concept to base their studies on. There is no doubt that Fletcher's original TiN detection experiment and his interpretation in terms of an auditory filter lead to significant progress in the field of auditory theory and still today—after more than 70 years—inspire research projects like the present study.

3.2 TiN Detection and the Origins of Signal Detection Theory

"*Den Punct, wo die Merklichkeit eines Reizes oder eines Reizunterschiedes beginnt und schwindet, wollen wir kurz die Schwelle nennen.*" (Fechner, 1860, chap. X, "Die Tatsache der Schwelle")

Until the middle of the twentieth century, most psychophysicists believed that observers could best be understood as behaving according to Fechner's "threshold model", i.e., whenever a perceptual signal was larger than a certain threshold, it could be correctly detected; whenever it was below, observers could merely guess (Fechner, 1860). As a direct consequence of this assumption, the theory had difficulties of explaining meaningful ordering that was observed in data comprising so-called "subthreshold" events[1] (Green, 1960). Several attempts were made to "fix" the threshold model. The suggested changes often bore similarities with signal detection theory (SDT), a new concept that was introduced into psychophysics in the 1950s and '60s (Swets, 1961). Eventually detection theory was accepted as the appropriate general framework for explaining psychophysical performance data (Krantz, 1969). This early history of the development of signal detection theory is closely linked to the classical auditory paradigm of Tone-in-Noise detection, as detailed in the following paragraphs.

[1] Fechner was well aware of this fact: "*An die Unmerklichkeit kleiner Unterschiede knüpft sich von selbst eine feine und nicht unwichtige Frage für das Massverfahren der Empfindlichkeit nach der Methode der richtigen und falschen Fälle. [...] Trotzdem, dass ein Unterschied für sich unmerklich ist, wird er doch bei einer hinreichenden Anzahl Verglei-chen ein Uebergewicht richtiger Fälle zu Gunsten des schwereren Gewichtes, allgemein des grösseren Reizes finden lassen.*" (Fechner, 1860, Sec. X.3 "Allgemeinere Betrachtungen bezüglich der Schwelle.")

"*The mathematical model of signal detections is applicable to the problem of visual detection. [...] The experimental data supports the logical connection between the forced-choice and yes-no techniques developed by the theory.*" (Tanner and Swets, 1953)

The theory of signal detection is generally grounded in the statistical theory of hypothesis testing. Without reference to any kind of psychophysical paradigm, Peterson and Birdsall (1953) and Peterson *et al.* (1954) provided a very general quantitative account of the problem of detecting a signal in stationary, band-limited, white Gaussian noise. In terms of possible applications, their manuscripts mainly deal with the problem of detecting a target in a radar signal. Even though the authors do not explicitly elaborate on applications in psychophysics, this early work represents a central foundation of the "theory of signal detectability" (later termed "signal detection theory", SDT). Tanner and Swets (1953, 1954) immediately drew a connection to psychophysics. Based on the manuscript by Peterson and Birdsall (1953) (and "a series of seminars conducted by Dr. H. R. Blackwell"), they developed a "New Theory of Visual Detection" which—in their opinion—explains the psychophysical data better than earlier theories.

"*It is found that, except for being approximately 13 db less sensitive, subjects behave very much like the ideal detector—that is, in accordance with the mathematical predictions—when the signals are pure tones.*" (Marill, 1956)

Subsequently, for his doctoral thesis entitled "Detection Theory and Psychophysics", Marill (1956) critically studies and improves upon the work by Tanner and Swets (1953), and—also based on Peterson and Birdsall (1953)—concludes that "*the two-category forced-choice technique is found to be particularly advantageous on theoretical grounds.*" He then for the first time applies SDT to an auditory psychophysical paradigm, namely the detection of pure tones masked by broadband Gaussian noise. Using the newly developed theory, he attempts to "*predict the probability that the subject will perceive the signal*". He observes that his listeners are much less sensitive than a theoretically derived "ideal detector" and infers the width of the hypothetical critical band from the measured discrepancy (see section 3.1).

In 1960, Green (1960) reviewed the current developments in detection theory and its application to signal detection in wide-band masking noise while he notes that there seems to be "*some confusion both about the theory and its applications*". He concludes that the experimental data "*seriously conflict with [...] the threshold model and give some measure of support to the decision-theory analysis.*" In the same manuscript, Green also elaborates on the mathematical concept of the ideal observer and derives the shape of the psychometric function.

"*[This paper] is beholden to TSD [Theory of Signal Detectability] for many basic concepts and hopes to pay part of the debt by casting a little more light into one or two dark corners.*" (Jeffress, 1964)

A few years later, Jeffress (1964) "*re-invented detection theory*" (his own words! (Jeffress, 1967)) and theoretically analyzed the problem of detecting a tonal signal in a background of *narrow-band* Gaussian noise using "receiver operating characteristic" (ROC) curves. This same paradigm of "narrow band Tone-in-Noise detection" then becomes one of the central paradigms in the now classical textbook on SDT by Green and Swets (1966). The authors theoretically derive the detection performance of an ideal observer depending on the characteristics of the background noise and the signal to be detected while the

listener is assumed to employ sound energy as the decision variable.

In conclusion, the Tone-in-Noise detection paradigm was one of the earliest test cases for detection theory. Its successful application in this context certainly helped the young theory to be embraced in the psychophysical community. As soon as detection theory was accepted as a fundamental framework for the analysis of psychophysical data, subsequent Tone-in-Noise detection studies naturally relied on the corresponding concepts to analyze and explain the data.

3.3 Frozen Noise in TiN Detection

Fletcher (1938) used a thermal noise generator to generate the masking noise and was able to find dependencies between the statistics of the stimuli and detection performance. However, because in Fletcher's and most of the subsequent studies (Hawkins and Stevens, 1950; Schafer *et al.*, 1950; Marill, 1956; Swets *et al.*, 1962; Jeffress, 1964) the sounds were never recorded, the authors principally could not investigate relations between the characteristics of particular sound samples and the corresponding observer responses.

"*The signal and noise outputs were added and recorded on [...] a dual channel tape recorder. In this manner one obtains about 5 or 6 minutes of exactly reproducible stimulus material.*" (Sherwin *et al.*, 1956)

The "frozen noise" technique, i.e., the ability to record and replay the auditory noise waveforms exactly as they were presented to the listeners, was a critical technical development for the progress in the study of TiN detection. A study by Sherwin *et al.* (1956) represents one of the first accounts where recorded sound samples are used in the investigation of the TiN paradigm. The recordings were used to measure a correlation between an energy detector and a listener responses.

"*While it is impossible to guarantee that the two presentations of the tape were exactly alike, every precaution was taken to approach that goal. The largest source of stimulus variability probably arose because of variation in the speed of the tape through the heads and slight variation in contact between the playback head and the magnetic tape.*" (Green, 1964)

Even though the frozen noise technique is perfectly suited to investigate the effect of particular characteristics of the stimulus on the observer's responses, even a decade later only a few more studies had relied on this method. Watson (1964) performed a rating experiment and correlated the rating judgements with "voltage peaks" within a narrow frequency band centered on the stimulus. Green (1964) used noise waveforms replayed from magnetic tape to measure observer consistency during multiple presentations of identical stimuli. And finally, Pfafflin and Mathews (1966) used reproducible noises to find dependencies between observer behavior and the output of a digital filter. They also made an observation they only could have made with "frozen noise": "*Special characteristics of certain noises appear to affect the subject's response when these noises appear in either signal or non-signal trials.*"

"*Twelve noise samples were used as stimuli in these experiments. They were generated by a random number-generating program on an IBM-7094 computer that produced 240 independent numbers for each noise stimulus. The stimuli were stored in numerical form in the memory of a Packard Bell 250 computer.*" (Pfafflin and Mathews, 1966)

During that time, the computer revolution was about to greatly simplify the controlled (re-)presentation of stimulus samples and subsequent data analysis. In fact, already Pfafflin and Mathews (1966) did not record their noise samples on tape, but instead generated and stored them digitally on a computer from where they were directly delivered to headphones. Even more advanced digital equipment allowed Ahumada and Lovell (1971) to not only repeatedly present a small number of computer-generated narrow-band TiN bursts, but also to correlate the recorded rating responses with the energy present in each of several frequency bands surrounding the signal using multiple regression.

Meanwhile, computers and sound cards have became a standard tools in auditory psychophysics replacing analog "thermal noise generators" and tapes. As a consequence, employing the "frozen noise" technique no longer implies a difference in terms of the way the experiments are set up and executed. It only means that the computer-generated sound samples are stored and used during later analysis.

Today, an ordinary lab computer can store and analyze hundreds of thousands of high-definition sound samples. Together with the corresponding listener responses, these data can be comprehensively investigated in any conceivable aspect, whether that are complex stimulus features or observer reaction times. The present study makes use of this impressive progress in terms of equipment, in particular the advances in storage and computing power.

3.4 Perceptual Cues in TiN Detection

A central problem concerning the paradigm of TiN detection has been the question as to which perceptual cues observers employ in order to detect the masked tone signal. The present section discusses a number of suggestions that have been made regarding this problem.

The underlying assumption when estimating critical bands with a TiN detection experiment is that observers "measure" some property of the auditory stimulus that passes the critical band filter surrounding the signal tone. In early studies, this property was usually assumed to simply be the overall energy of the sound (sum over the squared amplitudes) (Fletcher, 1940; Hawkins and Stevens, 1950) or signal amplitude (Jeffress, 1964). Fletcher (1940) assumed that the signal could be detected as soon as the intensity of the tone was equal to the average noise intensity. Sherwin *et al.* (1956) performed the first experiment to actually measure a correlation between the output of an energy detector and a listener's response, though the observer turned out to perform better than the detector. Using reproducible noises, Pfafflin and Mathews (1966) correlated the observer's responses with a digital filter of fixed width. They concluded that responses could be explained in part, but not entirely, by the presence of sound energy near the signal tone. Using concepts from the newly developed "theory of signal detectability" (see section 3.2), Jeffress (1964) analyzed TiN detection data and suggested that observers use an "envelope detector", i.e., the maximum amplitude height is detected and used as the critical cue. He discusses in detail whether a detector based on voltage (amplitude/envelope) or voltage-squared (power) better describes the detection data and concludes *"that amplitude [envelope height], rather than power [energy], is the basis for detection."*

"However, the observer's false alarm rate is about an order of magnitude lower than that calculated for the fluctuating threshold detector so it is clear that the model is deficient in some important respect." (Sherwin *et al.*, 1956)

Nevertheless, in their discussion of TiN detection in the context of the psychophysical theory of signal detection, Green and Swets (1966) suggested that observers rely on differences in energy between noise and signal stimuli to detect the tone:

"[T]hey propose that this detection process may be represented conceptually by a band pass filter, a nonlinear transformation of the output of the filter (a square-law device in their example), an integrator and finally a detection mechanism." (Kidd *et al.*, 1989)

However, in order to quantitatively match the data with their theoretical predictions Green and Swets (1966) had to assume a very high level of internal noise. To get a more direct estimate of auditory filter width, Ahumada (1967) correlated the energy passed by filters *of varying bandwidths* with observer responses. The author found that a simple energy-detection model could not explain the data, in particular for those trials where the signal was absent. He concludes that this *"sort of result"* would be expected *"if the observer is looking at a set of measures, only one of which represents essentially the energy [...]"* and that *"this kind of behavior would be exhibited by a filter-bank model [i.e.] the observer is monitoring a set of narrow filters tuned to different frequencies [...]."*

"The obvious next research step is to see whether such a model can actually give an output that correlates better with the observer's responses than does the single-filter energy-detection model." (Ahumada, 1967)

Thus, over the three decades following Fletcher's original experiment, ample evidence had been collected that the energy or amplitude of the sound passing through a single auditory filter were not sufficient to explain the observed data. Over the following years, even more observations were made that were incompatible with the simple "critical-band energy model". For example, the effect of "Comodulation Masking Release" demonstrated that information in frequency bands far away from the critical filter is able to *positively* effect signal detection thresholds (Hall *et al.*, 1984). A simple way to test whether a certain stimulus features serves as the basis for detection is to observe the effect of making this potential cue more or less reliable. Based on this idea, several studies using roving level paradigms (Kidd *et al.*, 1989; Berg *et al.*, 1992) or equal energy stimuli (Richards *et al.*, 1991) have shown that the presence or absence of the energy cue does not necessarily have a substantial effect on detection performance.

In a roving level experiment, the overall sound level of the stimuli is randomly

modulated from stimulus to stimulus, so that the energy cue necessarily becomes less informative. Kidd *et al.* (1989) "roved" the sound level of the stimuli in the two intervals of a two-interval forced choice task by up to 32 dB. Both at masker bandwidths above and below the width of the critical filter the masking threshold was very similar in the roving and non-roving (constant level) condition, or at least the difference between the two conditions was lower than predicted from an energy-detector model. The opposite effect of making the energy cue more informative can be achieved by adjusting the energy of every *noise masker* in order to perfectly equalize their sound energy across the two intervals of a trial. In this way, because the random variability of stimulus energy due to fluctuations in the noise token was eliminated, detecting the signal based on stimulus energy should become easier. Richards *et al.* (1991) applied this technique and found that "*[n]either increasing nor decreasing the variance of the noise-alone and tone-plus-noise energy difference distributions altered the detectability of a tone added to noise.*" concluding that *"the changes in energy that are concomitant with the addition of the tone are not the sole cue for the detection of the tone."*

"[T]he traditional critical-band energy-detector model could not account for the results, which are attributed to discrimination based on spectral shape or on wave shape." (Kidd *et al.*, 1989)

Obviously, the traditional model based on "energy detection" had become insufficient. Auditory detection models that incorporated energy from multiple critical-band detectors not centered on the tone were considered a natural alternative in conditions with noise bandwidths close to or wider than critical bands. Gilkey and Robinson (1986) found that *"[l]inear combinations of the outputs of several detectors that differ in center frequency or integration window provide even better fits to the data [than a single-tuned filter] [...] suggesting that a subject's decision is based on a comparison of information in different spectral or temporal portions of the stimulus."* This observation was later confirmed by Davidson *et al.* (2006).

"A model based on a weighted combination of energy in multiple critical bands performed best, predicting up to 90% of the variance in the reproducible-noise data." (Davidson *et al.*, 2006)

In addition, a number of stimulus features were proposed that relate to sound fine structure (i.e., phase modulation) as well as envelope. Mathematically, any narrow band signal can be described in terms of a quickly varying instantaneous phase $\phi(t)$ and a slowly varying envelope $A(t)$ (Hartmann, 1998):

$$S(t) = A(t) * sin\phi(t) \tag{3.1}$$

Both can be derived as the angle and magnitude of the complex "analytic signal"

$$X(t) = S(t) + iH[S(t)] \tag{3.2}$$

with the Hilbert H transform of the original signal $S(t)$

$$H[S(t)] = \frac{1}{\pi} \int_{-\infty}^{+\infty} dt' \frac{S(t')}{t - t'} \tag{3.3}$$

While the envelope determines the time-varying amplitude of the signal, the instantaneous phase characterizes the fine structure, e.g., position of zero crossings or variations in frequency. In practice, the envelope can also be approximately extracted by a half-wave rectifier and a low pass filter adapted to the frequency range of the sound.

A number of studies focused on the relative contribution of energy, fine structure and envelope features (Richards *et al.*, 1991; Richards and Nekrich, 1993; Davidson *et al.*, 2009b). Richards (1992) analyzed an extensive collection of "engineering type" fine-structure and envelope descriptors, such as envelope mean, kurtosis, crest factor (peak divided by average), average slope or the distance of zero-crossings and variance in instantaneous frequency. She concluded that either fine structure or envelope characteristics could explain detection performance.

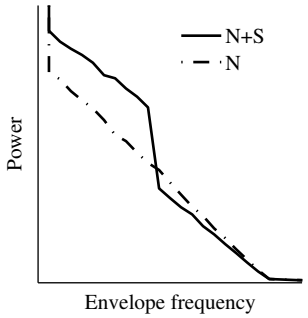
"The simulations indicated that changes in both the fine structure and envelope were sufficient for the detection of a tone added to noise when reliable energy-based cues were not present." (Richards, 1992)

Dropping the assumption, that TiN detection is dominated by a single stimulus characteristic, Richards and Nekrich (1993) investigated whether observers may rely on a combination of multiple auditory features and found that a substantial fraction of observers appeared to be simultaneously relying on level-dependent (i.e., related to sound

energy) as well as level-invariant cues. Davidson *et al.* (2009a) corroborate the idea of cue-combination by observing that *"dependence upon both envelope and fine structure [...] is required to predict the detection results"*.

Following a suggestion by Green *et al.* (1992), the envelope (or "modulation") power spectrum has been scrutinized regarding its involvement in narrow-band TiN detection:

"Basically, we assume that the auditory system can compute the envelope of a narrow-band stimulus, and that changes in the shape of the envelope power spectrum are used to detect changes in the power spectrum of the original waveform."



The average power spectrum of the envelope of a narrow band masker (N) with and without a pure tone signal (S).

Adding a pure tone to a narrow-band noise masker characteristically modifies the typical triangular shape of the envelope spectrum (see graph on the side). Based on this concept of envelope extraction, Dau *et al.* (1996a) designed a complex "quantitative model of the 'effective' signal processing in the auditory system". In this model, the envelope was approximately computed through half-wave rectification and low-pass filtering. The temporal activity pattern of the signal processed in such a way was then compared to a stored template representing the signal to be detected. The model was able to predict thresholds for a range of auditory detection conditions, and Dau *et al.* (1996b) concluded *"that the present model is a successful approach to describing the detection process in the human auditory system."* Later, the model was expanded with a modulation filterbank (Dau *et al.*, 1997; Verhey *et al.*, 1999).

"Discharge cadence may be as viable as tonotopic activity in providing a mechanism for detecting a tone in noise, but it may also contribute to the discrimination of complex periodic sounds." (Berg, 2004)

Berg (2004) suggested that a physiologically more plausible way to extract envelope information was to observe the output of a leaky integrator, i.e., the cadence of neural discharge would serve as the decision variable. By comparing human detection thresholds with computer simulations, he demonstrates that—at least in principle—such *"a temporal code can account for critical bands in Tone-in-Noise detection"*.

However, one can not immediately conclude that models based on the envelope spectrum completely explain auditory processing, even though they may be successful at predicting *psychophysical thresholds* in TiN detection tasks. An observer model that is supposed to capture the causal details of the observer's decision mechanism should aim at predicting behavior not merely in terms of *thresholds* averaged over a set of stimuli, but on a "molecular" level—in the terminology of Green (1964):

"[T]he subject matter of molecular psychophysics must be the individual response of the individual subject. [... A] trial-by-trial analysis of the observer's response may provide a critical and efficient test of a hypothesis that is impossible on the molar level—impossible either because there is no difference in prediction on that level or because the amount of data required to select one hypothesis over another would be excessive."

"Comparisons of the dependencies of each model on envelope and fine-structure cues to those in the data suggested that dependence upon both envelope and fine structure, as well as an interaction between them, is required to predict the detection results." (Davidson *et al.*, 2009a)

As regards trial-by-trial predictions, current models for TiN detection are less successful. After testing the performance of a number of these models —including single and multiple critical-band as well as fine-structure and envelope-based detectors—in predicting TiN detection data, Davidson *et al.* (2009a) concluded that even though these models successfully predicted listeners' thresholds, they *"cannot explain diotic detection patterns for reproducible noise maskers"*, i.e., they were not able to explain the data on a single trial level.[2] Notwithstanding the poor overall performance, the study concluded that the *"data were best predicted by an energy-based multiple-detector"* while *"complicated temporal models [including an envelope-template model] [...] were weakly correlated with subjects' responses"*.

[2] Davidson *et al.* (2009a) did not test models whose decision mechanism is based on an envelope filter bank such as proposed by Dau *et al.* (1997). Such models may have been able to explain the data on a single-trial level, but so far they have not been tested in this regard.

In conclusion, despite decades of research in TiN detection, no firm answer has been given to the question which cues observers employ under narrow-band conditions. A number of models have been established to predict response patterns, but none of these can reliably predict the listeners' behavior on the basis of individual stimuli (Davidson *et al.*, 2009a)—possibly because the choice of candidate features was too limited and did not include the actual cues.

Chapter 4

Principal Research Methods

This chapter presents details on how the TiN stimuli were generated and preprocessed both for the computer simulations and behavioral experiments described in chapters 5 and 6. Furthermore, some implementational details concerning the psychophysical experiments are discussed.

4.1 Stimuli

Tone frequency, noise masker bandwidth and length of the Tone-in-Noise stimuli were chosen according to standard values in the literature (Ahumada and Lovell, 1971; Richards and Nekrich, 1993; Evilsizer *et al.*, 2002). The masker stimulus consisted of a 100 Hz wide band of noise centered at 500 Hz. In signal trials, a pure tone at 500 Hz was added to the masker. At this frequency range, human observers should be able to access fine structure cues which may not be available at substantially higher frequencies (in the range of kHz) due to the limited firing rate of auditory nerve fibers (Rose *et al.*, 1967). The width of the auditory filters at 500 Hz center frequency is expected to be slightly narrower than the noise bandwidth of 100 Hz, although estimates vary widely (Ahumada, 1967).

Instead of preparing Gaussian noise and afterwards applying band-pass filters, the noise was generated as the sum of individual pure tone components of the desired frequencies with Rayleigh-distributed amplitudes and uniformly distributed random phases (Hartmann, 1998, chap. 23).[1] For signal stimuli, a signal tone was added with random phase. The strength of the signal relative to the masker, or signal-to-noise ratio (SNR), for TiN stimuli is usually measured as the unit-less ratio E/N_0, signal energy E (Power · Time) divided by noise spectrum power N_0 (Power / Frequency). The noise level was fixed while the signal level was varied to achieve a wide range of SNRs covering the entire range of psychometric performance (–6 to 18 dB for the simulations, and 7 to 15 dB for the psychophysical experiments). On each trial of the single-interval task, a noise or signal-plus-noise stimulus could appear with equal likelihood.

All stimuli were 200 ms in length. Stimuli were windowed with comparatively long 50 ms cosine-squared ramps, in order to minimize the energy present in frequency components outside the narrow noise-band (for details see Appendix D). Nevertheless, the remaining steady-state portion of 100 ms had a length similar to earlier TiN detection studies (Swets *et al.*, 1962; Ahumada and Lovell, 1971; Richards and Nekrich, 1993). The stimulus magnitude during digital processing was adjusted to minimize the effects of sound clipping, which is discussed in detail in Appendix C.

[1] The norm $\sqrt{a^2 + b^2}$ of two normally distributed quantities a and b is Rayleigh-distributed. A random noise signal can be represented as a sum of independent and normally distributed sine- and cosine-signals of different frequencies. The amplitude of each frequency component is computed as the norm of the two orthogonal signals and consequently has a Rayleigh distribution (Hartmann, 1998, App. I).

4.2 General Aspects of Data Preprocessing

Generally, data preprocessing is regarded as a critical step when applying machine learning algorithms to mine large sets of data. In particular, preprocessing is often necessary in order to reduce the dimensionality of the input, as calculations may be practically unfeasible due to time or memory constraints. The raw time series data of the auditory TiN stimuli consist of many more dimensions than the actual number of degrees of freedom of the signal, which is governed by the number of noise components. Ideally, the preprocessing should scale down the fitting problem to the actual dimensionality. Reducing the dimensionality not only makes the problem more manageable in terms of computation time and memory requirements, but also reduces the risk of over-fitting.

In addition, the preprocessing needs to be adapted both to the space of mappings the model provides and, in the context of a psychophysical study, the perceptual mechanisms the observers are expected to rely on. The mathematical function that maps stimulus input to behavioral output usually covers a limited space of mappings. For example, a linear model—such as used in the present study—can not reproduce a nonlinear mapping from the input data to the output, unless the raw data are nonlinearly transformed. In addition, dimensions that are very unlikely to be relevant for data fitting, e.g., sound properties that are not accessible to the observer for physiological reasons such as phase information of high frequency tones, should be eliminated.

The preprocessing of stimuli may be considered the most decisive aspect of data analysis when attempting to explain and predict behavioral data from stimulus characteristics. Not only does it critically effect the chances of successfully fitting a model, but also the ability of meaningfully interpreting the fit—since the preprocessed stimulus properties form the basis for explaining the decision mechanism of the fitted model. The interpretation of the resulting decision cues is substantially simplified when a straightforward, comprehensible and reproducible preprocessing technique is employed.

4.3 Stimulus Processing

Stimuli were preprocessed to extract three potentially relevant auditory features and corresponding predictors as depicted in Fig. 4.1:

- **Sound energy.** This feature is extracted by integrating (or summing, for a discretized signal) over the squared instantaneous amplitudes of the sound waveform $S(t)$:

  $$E = \int_{t=0}^{T} S(t)^2 dt \tag{4.1}$$

  This transformation represents a simple quadratic mapping.

- **Power spectrum of the sound or fine structure.** During the simulations (chapter 5), the spectrum of the waveform was extracted through a discrete Fourier transformation. In principle, the Fourier transform is a linear mapping. However, the extraction of the amplitude of each component and the squaring of these is a nonlinear process. While analyzing the experimental data (chapter 6), the spectrum of the fine structure (extracted through a Hilbert transform) was employed, not the raw waveform. In this way, a complete separation between spectral features and energy was achieved, because the fine structure—by definition—is independent of sound energy (see Eq. 3.1).

- **Power spectrum of the envelope.** The extraction of the sound envelope via the Hilbert transform represents a highly nonlinear mapping (see Eqs. 3.1–3.3). Subsequently, the envelope spectrum is extracted via Fourier power transform, another nonlinear processing step.

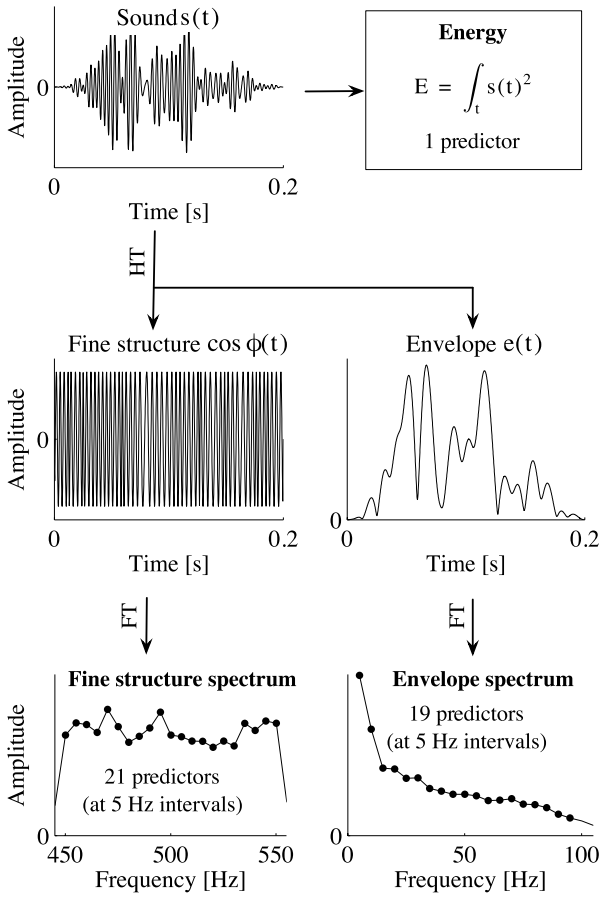


Figure 4.1: Stimulus preprocessing. Schema of the preprocessing steps that transform the raw stimulus time series (top left) into three sets of stimulus descriptors (marked in bold) through Hilbert (HT) and Fourier (FT) transforms resulting in altogether 41 stimulus predictors. These predictors were used for analyzing the experimental data. For the simulations, the sound waveform was directly Fourier transformed (without first computing the fine structure) to extract the sound power spectrum.

These preprocessing steps were chosen both for their clear mathematical interpretability as well as physiological plausibility—all of them should be available to a human listener in similar form and have been proposed in earlier studies of the problem (see section 3.4). The total sound energy, which has been proposed as a cue in the majority of early TiN studies, is approximately represented by overall neural activation in the auditory nerve (Sachs, 1974). The spectral properties of the sound, though not as fine-grained as through a Fourier transformation, should be partially accessible through the tuning properties of multiple auditory nerve fibers centered close to the stimulus frequencies. Finally, the envelope spectrum could be extracted by nerve cells in the inferior colliculus which are sensitive to specific amplitude modulations of pure tones (e.g., see Snyder *et al.* (1995) for data from cats). I also paid attention to choose features that are independent of assumptions about the characteristics of hypothetical auditory critical bands. Otherwise, one might end up with a dangerous case of circular reasoning. These bandwidths are usually derived in TiN detection experiments under the assumption that observers rely solely on energy. This assumption, however, is one of the main open questions I am about to investigate.

Both the waveform/fine structure power spectrum and the envelope (or "modulation") power spectrum were extracted using standard functions for Fourier and Hilbert transformations of the computing software (*fft*, *hilbert* in Matlab, The Mathworks, Inc., 2010). Each of the spectra was truncated to isolate those components that contained stimulus energy. The sound power spectrum was entirely defined with 21 components ranging from 450-550 Hz with 5 Hz distance (the maximum spectral resolution available from a stimulus of 200 ms length), all other frequencies do not contain any energy. Similarly for the envelope power spectrum: 21 components ranging from 0-100 Hz with 5 Hz distance entirely defined the stimulus in this domain. The 0 Hz component was excluded being equivalent to sound energy, and the 100 Hz component as it contained almost no energy. In total, the vector of predictors describing the stimulus then consisted of 41 entries: energy (one entry), waveform/fine structure power spectrum (21 entries) and envelope power spectrum (19 entries). During model fitting, all predictors were presented collectively independent of observer or condition.

In order to uncouple the scale of the model weights from mean and variance of the associated predictors, the predictors were standardized, i.e., they were individually rescaled to zero-mean and unit-variance across samples. More precisely, standardization is performed across the training data set alone. The test set is then scaled and shifted by the same amount. In this way, preprocessing of the training and test stimuli is identical, but strictly independent of the characteristics of the, a priori unknown, test set. Standardization is a common processing step in machine learning to prevent input data with abnormal original values to bias the model weights and to improve numerical tractability (Hastie *et al.*, 2009, chap. 11.5.3). As a result, the weights of a linear model directly correspond to the extent to which a predictor controls the model output based on the stimulus data: Predictors with small associated weights have little influence, while those with large weights govern the responses.

4.4 Correlations between Features

Interdependencies between model predictors generally pose a challenge during model fitting and should be taken into account when estimating relative weights. A correlation matrix for the current set of predictors is plotted in Fig. 4.2. Only the components defining the sound spectrum do not exhibit any correlations with each other, since they are independently sampled during stimulus generation. By contrast, most components in the envelope spectrum display correlations among each other and with components in the sound spectrum. In addition, sound energy is correlated to the sound power

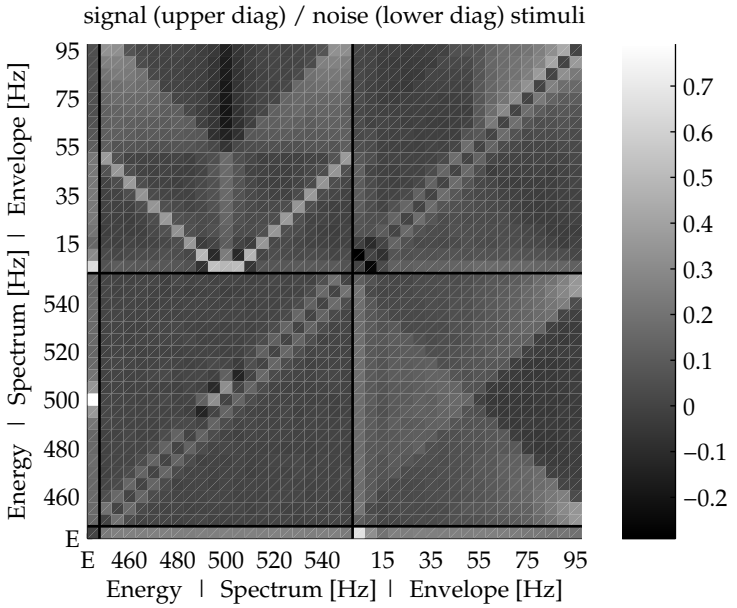


Figure 4.2: Correlation matrix for stimulus features. Pairwise correlations between individual predictors for signal (E/N_0 = 12dB, upper left diagonal) and noise (lower right diagonal) stimuli, values range linearly from –0.3 (black) to +0.7 (white).

spectrum.

For signal stimuli, particularly robust correlations appear between the total sound energy and the amplitude of the signal frequency, and between frequency pairs in the sound spectrum and corresponding beat frequencies in the envelope domain, as combinations of components in the sound spectrum add up to single components in the envelope spectrum. This reflects the fact that the combination of pure tones results in a beating sound which generates a single frequency component in the envelope domain. Overall, pairwise correlations range from −0.3 to 0.7, with a majority clustering between −0.1 - 0.3 for noise stimuli. For signal stimuli, the values depend on signal level. As a typical example, for signal stimuli at 12 dB SNR they range between −0.05 - 0.2.

The fact that the correlation is lower for signal than for noise stimuli may be counterintuitive. Essentially, the underlying reason is that the overall variance of some of the predictors is much higher for signal than for noise stimuli. Therefore, even though the overall covariance between features moderately increases for signal stimuli, the correlation still decreases as it corresponds to the covariance *divided by* the individual variances.

4.5 Details of the Behavioral Experiments

In the following, I briefly discuss some of the choices that were made during the implementation of the human psychophysical experiments which are further described in chapter 6.

The first and probably most important decision concerned the question whether to realize the TiN detection task as a two-interval forced-choice (2IFC) or a one-interval YES/NO-paradigm. A 2IFC-implementation has several theoretical (e.g., observer are expected to use a fixed decision criterion (Marill, 1956; Green and Swets, 1966)) and practical advantages (according to Blackwell (1952) it generates the “best” threshold

estimates, at least for highly trained observers). Based on the concept of the ideal observer from signal detection theory, observers in a 2IFC task have long been assumed to follow the ideal "difference strategy": The single decision variable is computed as the difference between a perceptual measure that characterizes each of the two stimuli (Green and Swets, 1966). Thorough empirical tests of this assumption have shown, however, that an average naïve observer does not appear to employ such a strategy (Jäkel and Wichmann, 2006; Yeshurun *et al.*, 2008), unlike highly trained ones as in the study by Blackwell (1952). Instead of weighing both stimuli equally strongly, as in the difference strategy, observers may, for example, attempt to make a decision already after the first interval (which may appear more salient) or only based on the second interval (which is closer to the response window). Or they may weigh each interval according to the respective perceptual evidence. From the response data alone, one can hardly judge which decision rule they followed.

"We find little evidence supporting the claims that 2-IFC is unbiased [...] and we also reject the two claims associated with the Difference Model [...]." (Yeshurun *et al.*, 2008)

To summarize, in a 2IFC paradigm there exists a whole range of possible decision strategies. All of them need to be considered when attempting to capture the listener's decisions, which generally requires a substantially more complex behavioral model compared to the simple linear mechanism that was assumed here (Eq. 2.1). In consequence, I decided to employ a YES/NO-paradigm, where listeners are presented with a single stimulus before giving a response and which is quite common for this kind of task (Isabelle and Colburn, 1991; Evilsizer *et al.*, 2002; Davidson *et al.*, 2006, 2009b). Essentially then, there is only one simple strategy a listener can follow—pressing a button depending on the stimulus currently presented—which corresponds exactly to the way the observer model behaves.

The second critical decision is related to experimental feedback. The feedback that observers are presented during a psychophysical experiment generally has a critical influence on their behavior (Blackwell, 1952; Richards, 2002). Providing feedback is an important tool at least for two purposes: First, with feedback observers can learn a task implicitly, i.e., without direct instructions. In this way, it may be easier for listeners to find an efficient strategy while remaining completely naïve as to the purpose of the experiment (Pedersen and Ellermeier, 2008). Second, during prolonged monotonous tasks, feedback may serve as a motivational factor and increase attention. Unfortunately, providing feedback also has a significant downside: In particular at low signal levels, when even an optimal observer is only slightly better than chance, immediate feedback may prevent observers from developing a consistent decision strategy. As they regularly receive negative feedback, they continuously try and adjust their behavior to improve their "score".

In the psychophysical experiments presented in chapter 6, a hybrid strategy was employed as a compromise between the benefits and downsides of feedback. During an initial training phase spanning the first few sessions, immediate feedback was presented during all trials of an experiment. Later on, when listeners had grasped the general idea of the task and stabilized in performance, immediate trial-by-trial feedback was shown only during the beginning of every block. In this way, after taking a short break between blocks, listeners were able to readjust to the task, which was particularly important when the signal-to-noise ratio had changed. In addition, as a motivational incentive at the end of each block, observers were informed about their current percent-correct and percent-valid scores (averaged over the preceding block). During the subsequent data analysis, all trials with immediate visual feedback were discarded, as were the initial sessions before the observers stabilized in psychophysical performance.

Another important choice concerned the range of signal-to-noise ratios (SNRs) that observers were presented with. I chose to perform the experiment at four different SNR conditions, to be able to discover potential differences in observer decision strategy. For theoretical and practical reasons, only intermediate ranges of observer performance are useful for the analysis of observer strategy. At very high SNRs all observers reach the top

plateau of the psychometric function and thus give identical (namely correct) responses. Consequently, no information about observer strategy can be extracted—the decision already being known beforehand. At very low SNRs, observers mostly guess and merely reach chance performance. In principle this range provides the most information about observer strategy because individual differences in response strategy are most prominent. In psychophysical practice, however, low SNRs are to be avoided, as human observers are not able to learn a useful strategy and are quickly frustrated by the lack of success. Thus, during the initial training sessions, the SNR for all observers was adjusted in order for them to reach about 60/70/80/90%-correct scores in each of four different SNR levels.

Chapter 5

Computer Simulations

The theoretical analysis presented in this chapter aims at quantitatively assessing how specific and reliable the proposed cue-identification procedure works. Therefore, simulated behavioral data were generated from "artificial observers" for which the true perceptual cues and decision rules can be independently verified. The observers perform the Tone-in-Noise detection (TiN) task presented in chapter 3 according to known preprogrammed decision strategies. Representing a classic auditory task, TiN detection, was considered as a specific test case to evaluate the capabilities of the proposed cue identification method. In addition, the present simulations serve as a preparation for the analysis of "real" experimental data which is presented in chapter 6.

In the following, I demonstrate that by employing a sparse regression procedure perceptual cues can be identified even among linearly dependent stimulus features. At the same time, observer models were constructed that reliably predicted response behavior. In particular after introducing observer noise, estimates of the model parameters were robust while the required amount of data remained in a range that can be collected during moderately expensive psychophysical experiments. The advantages of the proposed method are corroborated by directly comparing the results with two earlier (non-regularized) techniques for identifying perceptual cues—logistic regression and reverse correlation—that were discussed in chapter 2.

5.1 Methods

In this study, simulated behavioral data from a Tone-in-Noise detection task was analyzed using different techniques for fitting a linear observer model. The procedure followed the subsequent steps (numbers referring to Fig. 5.1):

1. Tone-in-Noise sound samples were generated (section 4.1) and preprocessed to extract a set of stimulus features (section 4.2).

2. Observer responses were computed according to one of three strategies in a YES/NO-paradigm: detection of Energy, Spectrum or Envelope cues (section 5.1.1).

3. The fitting procedure was applied to sound stimuli and corresponding observer responses for a subset of the data, the training set (section 5.1.2). For regularized procedures, the ideal regularization parameter λ was determined during a previous parameter optimization (section 5.1.4).

4. As a measure for the quality of the fitted model, the agreement between observer and model was estimated for the remaining data, the test set (section 5.1.3).

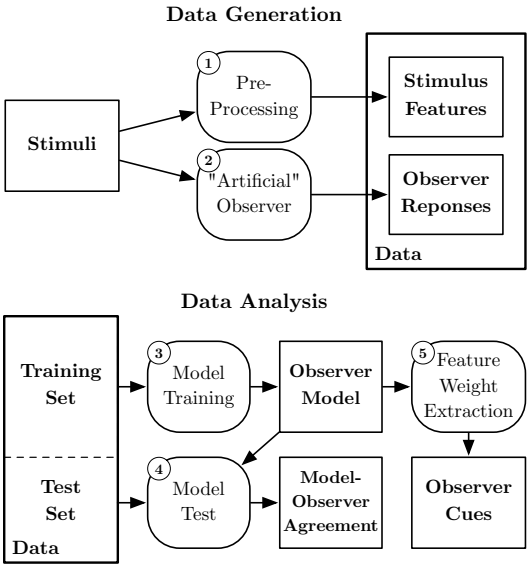


Figure 5.1: Diagram of the simulation sequence, the circled numbers refer to the description in section 5.1.

5. Predictor weights were extracted from the model to estimate the cues underlying the observer responses. The estimated observer cues were then compared to the true decision strategy (section 5.1.5).

All simulations were performed on a desktop computer running a custom-written script under standard scientific computing software.

5.1.1 Response Generation with Artificial Observers

In order to test the utility and efficiency of the proposed method, three different observers were simulated relying on the following strategies:

- **The Energy Detector** forms the decision variable from total sound energy. This observer exploits the fact that signal stimuli have a higher energy on average (Green and Swets, 1966).

- **The Spectral Shape Detector** (top left in Fig. 5.2) uses three adjacent Gammatone filters (Patterson *et al.*, 1991), with parameters chosen according to standard multiple detector models (Gilkey and Robinson, 1986): one centered on the signal (500 HZ) and two sideband filters (450 and 550 Hz). Center and sideband filter outputs are subtracted to generate a 1D-decision variable. This observer can be considered a differential analyzer of the power spectrum. It ignores overall changes in sound energy (or loudness), and instead relies on relative changes in power between different spectral bands. Exploiting the peak in the sound spectrum resulting from the added signal tone, it is in fact barely correlated with the energy observer for noise stimuli.

- **The Envelope Shape (or Modulation) Detector** (bottom left in Fig. 5.2) relies on the power at the output of a low-frequency (25 Hz) and a high-frequency (75 Hz) 2nd-order bandpass-filter (50 Hz bandwidth) operating on the Hilbert envelope (Appendix E offers details on the filtering algorithm). This detection mechanism was inspired by envelope-frequency models (Dau *et al.*, 1996a) that are based on

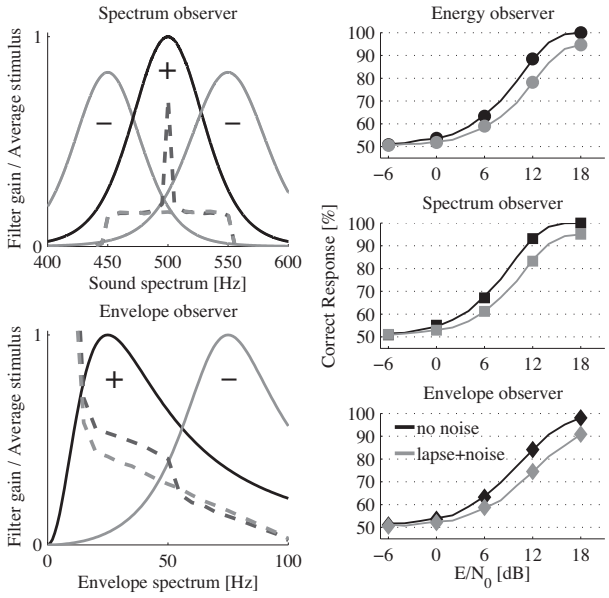


Figure 5.2: "Artificial" observers for Tone-in-Noise detection. **Left:** Observer strategy and underlying signal filters in the frequency (top) and envelope (bottom) domain. The average spectra in each domain are shown as dashed lines for signal (black) and noise (gray) stimuli. **Right:** Psychometric functions for the three simulated observers (top to bottom). Each plot displays results for deterministic observers (black, no lapse or noise) and probabilistic observers at high noise levels (gray, 10% lapses, 200% decision noise). Because of the long windowing, the average spectra in the envelope domain (bottom left) deviate slightly from standard envelope spectra of Tone-in-Noise stimuli (Green *et al.*, 1992).

a decision rule first proposed by Green *et al.* (1992). The two envelope filters are centered at the positions where the spectra of noise and signal stimuli diverge most. Their outputs are subtracted resulting in a 1D-decision variable. This observer mainly depends on an increase of low frequency envelope components resulting from an interaction (also called "beats") of signal and noise components.

During a trial, the simulated observers are presented with a single sound stimulus and respond in a YES/NO-fashion. An ideal observer applies a probability ratio criterion. Instead, a threshold criterion was applied, which generates exactly the same responses for most but the rare extreme values of the decision variable—assuming normal distributions with similar variance for the decision variable (Green and Swets, 1966). The decision threshold was set to the median value of the decision variable of all presented stimuli in individual conditions so that observers were unbiased.[1] With a value of the decision variable above this threshold, observers responded "Yes" (signal present), otherwise "No" (signal not present). These decision mechanisms give rise to psychometric functions as shown in the right column of Fig. 5.2. Best performance is achieved by the spectral shape detector, while the envelope shape spectrum detector shows the worst performance.

[1] More precisely, for the observers to be truly unbiased, the median was calculated from an equal amount of signal and noise stimuli.

To test the robustness of the procedure under more realistic conditions, simulations were also performed with observers exhibiting two kinds of probabilistic behavior: lapsing and decision noise. The lapse rate determines the percentage of (randomly distributed) trials, on which observers give arbitrary responses independent of the stimulus. To simulate decision noise a Gaussian random number was added to the decision variable before the threshold-based decision. Following the concepts of signal detection theory (Green and Swets, 1966), the amount of noise was measured relative to the standard deviation of the noise distribution, i.e., the distribution of the decision variable for noise stimuli. A decision noise value of 100% means that the width of the noise distribution has doubled. The same amount of noise was added to the signal+noise distribution. Probabilistic observers were simulated with decision noise of 100% ("medium noise level") and 200% ("high noise level"). The respective lapse rates were set to 5% and 10% broadly covering experimentally observed values of around 5% (Wichmann and Hill, 2001).

In order to test whether complex cue-mixing strategies can also be identified, conditions were included where observers switched between different decision mechanisms across trials. For example, such an observer would rely on the spectral cue in 25% of the trials and on the envelope cue in the remaining 75%, corresponding to a mixing proportion of 1:3. Altogether, a set of nine mixed-cue strategies was implemented each of which combined two of the three proposed observer cues (Energy/Spectrum, Energy/Envelope or Spectrum/Envelope) to a proportion of 3:1, 1:1 or 1:3.

In conclusion, the simulated observers relied on energy and spectral filters in the frequency as well as envelope domain. A spectral filter can be described as a weighted sum of the squared Fourier amplitudes of the sound. Thus, a linear model provided with the sound energy, as well as the squared Fourier components of the sound and the envelope is capable of reproducing the three observer strategies. The corresponding preprocessing steps were performed accordingly as detailed in section 4.2.

5.1.2 Model Fitting

In contrast to earlier studies (Gilkey and Robinson, 1986; Richards and Tang, 2006; Pedersen and Ellermeier, 2008), a *regularized* version of multiple logistic regression was applied in order to profit from the sparse regularization procedure described in section 2.3. The toolbox *LibLinear* (Fan *et al.*, 2008) served as an implementation for regularized logistic regression, which offers both sparse (L_1-) and non-sparse (L_2-) regularization. Model fitting as well as model predictions are both based on binary response data by

default.

Two classical procedures for fitting linear models were included, reverse correlation and (non-regularized) multiple logistic regression, as a baseline comparison to evaluate the advantage of the proposed method to cope with probabilistic data and dependencies between predictors. The logistic regression was implemented with *glmfit* in Matlab (The Mathworks, Inc., 2010). Reverse correlation analysis was approached similar to "classification images": First, the average of all predictors in each response class was computed. To estimate the correlation coefficients between predictors and responses, the difference between the response-averaged predictors was then calculated. The coefficients are directly proportional to the relative weights.[2]

In all cases, observer responses are predicted following a linear-binary rule (see Eq. 2.1): The predictors are linearly weighted according to the extracted model parameters and then combined to generate a "signal" or "no-signal" response according to an overall positive or negative sum, respectively.[3] As described in section 2.3, model "training" (fitting the model to the data) *with all procedures* was performed on a training subset of the data, while model "testing" (estimating agreement between data and model) was done on the remaining chunk of data.

[2] For binary response data from an equal number of signal and noise trials, reverse correlation (point-biserial correlation between stimuli and responses) and the classification image procedure (stimulus averages conditioned on responses) are mathematically identical (Ahumada, 2002).

[3] We ignored the bias term since it does not influence extracted model weights and, for unbiased observers and zero-mean predictors, has little effect on the resulting predictions.

5.1.3 Quantification of Model Agreement

After the linear observer model was trained to the data, it was verified whether it reliably captured observer behavior. Therefore, the agreement between observer and model was quantified in terms of the proportion of trials (in a 400-sample test set) where the model accurately predicted the observer's response. Only when predicted responses agree well with the observer, the model and observer are expected to be functionally equivalent and model weights to have a meaningful interpretation. Failure in training the model may have several reasons: an insufficient amount of data, too weak or too strong regularization or an inappropriate stimulus representation which does not allow a prediction of the responses with a linear decision rule. These obstacles were avoided by first optimizing the parameters controlling the amount of data and regularization (see the following section), and choosing predictors which allow a linear modeling of responses.

For noisy and consequently unreliable observers, the precision of predicting single trial responses is necessarily limited. Observer reliability was quantified in terms of (self-) consistency: the proportion of trials that observers give the same response on repeated presentations of identical stimuli (Green, 1964). For observer consistency measured in a "double-pass" experiment, i.e., in presenting two passes of the same stimulus set, Neri and Levi (2006) provided an upper and lower bound for the single-trial prediction accuracy achieved by the best-possible model. The agreement of the observer models was compared against this theoretical limit by directly estimating self-consistency of the "artificial observers" in each experimental condition by computing responses twice for the same set of stimuli (1000 samples).

5.1.4 Parameter Optimization

Before running the full set of simulations, an estimate of the amount of data necessary for reliable and stable solutions was determined. For a limited set of conditions, *training* data sets with different numbers of trials were analyzed within a range that was both appropriate for the algorithm and computationally feasible (25–6000 samples in logarithmic steps). An independently generated *test* data set of fixed size (400 samples) was used to estimate agreement between the trained model and the data, as shown in Fig. 5.3. Around 1000 to 2000 training data samples allowed the fitted model to reliably achieve more than 95% test agreement for deterministic observers. To be confident to

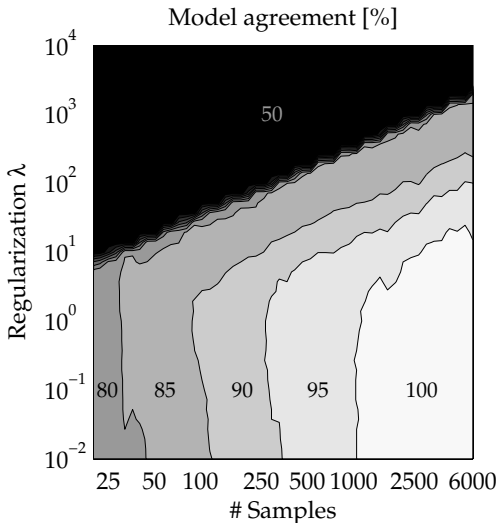


Figure 5.3: Grid search for optimizing simulation parameters. Agreement of model predictions with observer responses is indicated as the gray level, depending on the number of training data samples (x-axis) and regularization parameter λ (y-axis). Values range linearly from 50% (black) to 100% (white) and are indicated as numbers on the surface. Data are shown for an L_1-logistic regression fit to data obtained with a deterministic Envelope observer at 8 dB signal level.

operate in a safe regime for probabilistic observers too, 4000 training samples were used in subsequent analyses while retaining the number of 400 test samples.

When fitting the observer model with regularized procedures, the value of the regularization parameter λ had to be optimized to concurrently maximize prediction performance and sparseness of the weights. Therefore, in preliminary tests, the outcome of the fitting algorithm was analyzed with different values of λ. To choose the best λ, an automated procedure was followed: Observer-model agreement was plotted against the regularization parameter λ. This usually resulted in a high plateau for small λ corresponding to a moderate sparseness constraints. Beyond a critical value for λ, agreement fell off abruptly where regularization was too strong pushing too many weights to zero. A plot showing the relationship between number of data points, regularization and the resulting model agreement is shown in Fig. 5.3. As the ideal λ, a value was chosen that was both on the plateau, where the algorithm provided good predictions, but also close to the fall-off, favoring sparse weights. This is illustrated with an example in Fig. 5.6. This procedure was repeated for each condition (signal-to-noise ratio/SNR, observer strategy and observer noise/lapse) to determine the corresponding optimal λ that was used during the subsequent final model fit.

5.1.5 Analysis of Model Weights

The output of the observer model is computed from a weighted linear sum of model predictors. The weights w_i obtained through model fitting are represented as a vector $\mathbf{w}$. This weight vector can be used to generate predictions from new data and to analyze the underlying decision mechanism: The weight the model applies to each predictor is a direct measure of the influence of the associated stimulus feature on the model output.

To determine the main stimulus cue that governs model predictions, the relative weights were computed that are associated with the three sets of predictors: energy, spectrum and envelope. These *set weights* W_s indicate on which of those three predictor

sets a particular observer strategy is based. The full weight vector had as many weights as there were predictors (41 entries, see section 4.2). By combining the weights belonging to each set into a single number, this vector was projected from 41 down to three dimensions, where each new dimension represented the weight on the three sets of predictors respectively. Mathematically, this corresponds to separately computing the vector norm (or RMS value) of the weights associated with each set. As a final step, these (positive) set weights were scaled to sum to one in order to represent the *relative* importance of each predictor set:

$$W_s = \frac{\sqrt{\sum_{i \in s} w_i^2}}{\sum W_s} \tag{5.1}$$

with s designating each of the three predictor sets. For example, the set weight for the spectrum is defined as the vector norm of the weights associated with the sound spectrum predictors, divided by the sum of all set weights. Predictor sets that receive a significant nonzero weight are considered to be critical for explaining behavior. The metric of these set weights actually corresponds to the proportion of trials an observer relies on a particular cue, as confirmed in abstract numerical simulations. If the investigated predictors are statistically independent and if sufficient data are given, these proportions are precisely reproduced.

To investigate model strategies in more detail, the complete ensemble of model parameters was examined. The weights associated with the sound spectrum and envelope spectrum predictors can be interpreted as spectral filters and directly compared to the filters of the Spectrum and Envelope observers. However, as derived in Appendix F, they first need to be rescaled according to the equation $f_i = w_i/\sigma_i$, with the reconstructed observer filters f_i, the model weights w_i and the standard deviation of the original predictors σ_i. Finally, an appropriate comparison of filters also needs to take into account the weight on the energy feature, which is equivalent to a constant shift of all spectrum weights. Thus, to provide a fair model comparison, the extracted power spectrum filter was adjusted to reflect the combination of both the energy and the spectrum weights.

5.2 Results

Results reported here refer to three principal model fitting procedures: L_1-regularized logistic regression, "standard" (non-regularized) logistic regression and reverse correlation. In addition, some results for an L_2-regularized logistic regression are provided. Simulations were performed for all combinations of signal-to-noise ratios (SNR), observers (including deterministic and probabilistic decision strategies) as well as fitting procedures. Each condition was repeated six times with identical parameters. The confidence intervals in the present study were estimated using bootstrap sampling: The simulation procedure was repeated several times with different sets of randomly generated stimuli. The error bars in Figs. 5.4–5.9 display the mean and standard deviation of the results across repetitions.

5.2.1 Model Agreement

For deterministic ("no noise") observers, both logistic regression models mimic observer behavior across the entire range of tested SNRs. As shown on the left of Fig. 5.4, a model-observer agreement around 98% was achieved across all SNRs. Thus, the model predictions were identical to the observer's response in almost all trials. For model estimates from reverse correlation the predictive quality depends on the tested SNR, dropping to 85% agreement for the lowest levels. In conclusion, for deterministic observers, both logistic regression algorithms outperform reverse correlation at low SNRs. For noisy observers, percent agreement drops to 60–70% for all training procedures and

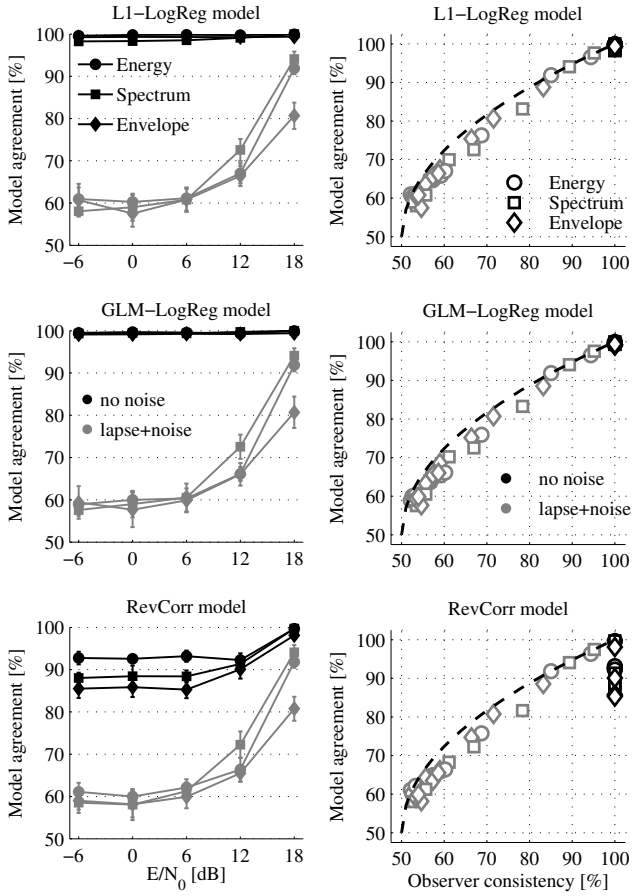


Figure 5.4: Model-observer agreement of L_1-logistic regression (top), non-regularized (GLM) logistic regression (middle) and reverse correlation (bottom) models with different observers (different symbols; gray level designates deterministic and probabilistic decision rules). **Left:** Agreement as a function of SNR (data for probabilistic observers at high noise levels). **Right:** Agreement as a function of observer self-consistency. Data for probabilistic observers include both medium and high noise levels. The optimal agreement given at a certain level of observer consistency is represented as a dashed line (following Neri and Levi (2006)).

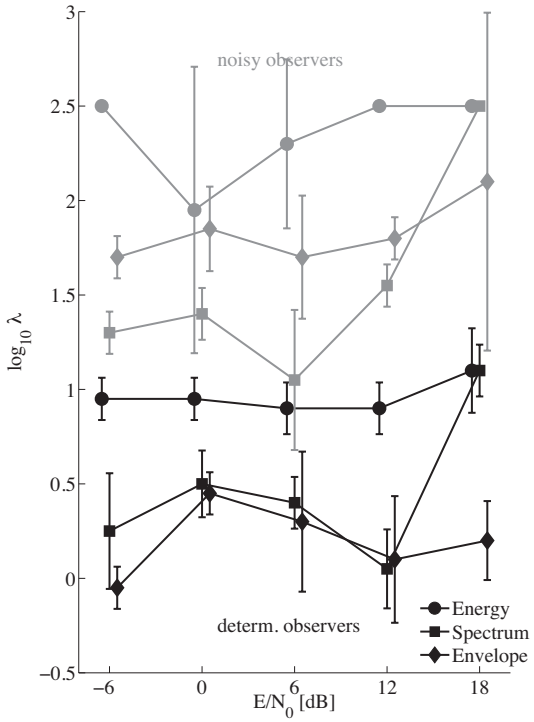


Figure 5.5: Variations of the regularization parameter λ depending on observer (symbol, see legend), noise condition (black—deterministic, gray—noisy), and signal level (x-axis).

is highly dependent on the SNR—because decision noise is more effective in interfering with the observer's strategy at low SNRs.

Nevertheless, the models perform almost ideally within the theoretical limits dictated by observer reliability, as demonstrated on the right of Fig. 5.4. Across the entire range of tested SNRs and observer noise levels as well as corresponding observer self-consistencies (ranging from 50%–100%), the prediction of single trials with the proposed methods attains close-to-optimal values—it lay less than 5% below the theoretical limit determined by observer reliability when predicting probabilistic observer responses.

5.2.2 Regularization Parameter λ

The absolute values of the parameter λ have no clear interpretation, but differences across conditions provide valuable insights indeed. Figure 5.5 shows both the average and the variation of λ across conditions for the three observers. Between different signal levels, there are no systematic changes. For the energy observer, the regularization parameter was larger than for the other observers across all conditions. Generally, stronger L_1-regularization results in sparser weights. The energy observer is the "sparsest" observer, it can be modeled with only a single predictor. Therefore, a model fit to the corresponding data profits most from increased regularization.

For probabilistic observers, the regularization parameter was much larger (2 orders of magnitude) than for deterministic observers. When fitting a model to noisy data, there is an increased risk of "over-fitting", i.e., fitting irrelevant noise structure in the training data set. Regularization is applied to prevent over-fitting. Accordingly: The noisier the data, the stronger the regularization that has to be applied to prevent over-fitting and achieve good test data predictions.

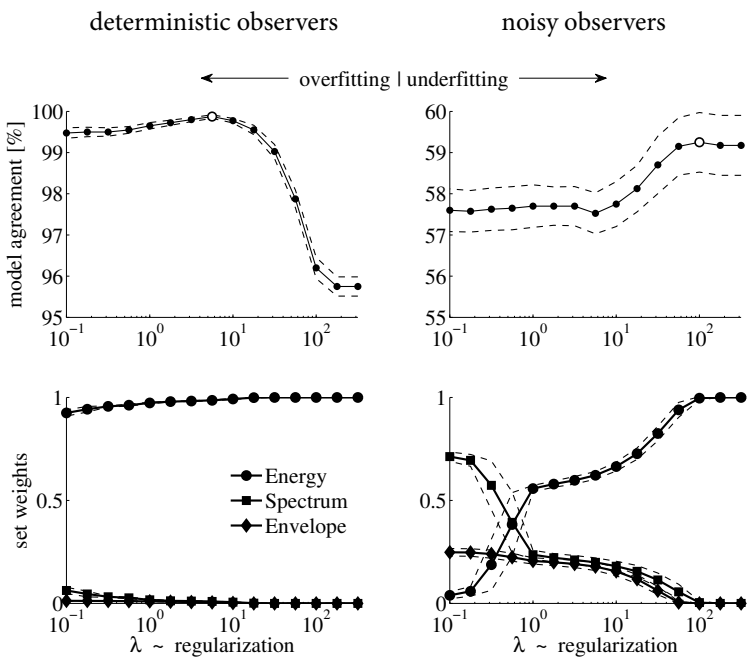


Figure 5.6: Regularization in practice. Data sample from a single fitting procedure, for an energy observer at SNR=-6 dB under deterministic (left) and noisy (right) conditions. **Top:** Model-observer agreement on a test set depending on the strength of regularization. The open circles indicate the position of the maximum model agreement. **Bottom:** Relative weights on the feature sets (symbols, see legend) at different regularization levels. In particular for the noisy data set, strong regularization increases model agreement. At the same time, the stimulus energy is being identified as the critical cue.

While for most conditions, λ varies very little across repetitions, in rare cases there are some outliers that increase the variance of λ. These outliers are the result of flat regularization/model agreement-profiles around the maximum agreement, i.e., model agreement changes very little even with strong variations in λ. However, even though λ varies strongly across conditions, the weights (just as the model agreement) change very little and the resulting models are stable nonetheless. Figure 5.6 contains a sample plot (including a deterministic and noisy energy observer at -6 dB) to illustrate this point.

5.2.3 Set Weights

The set weights represent a measure for the relative importance that the model attributes to the three sets of predictors associated with energy, sound spectrum, and envelope spectrum. In the following, they are directly compared with the true observer cues.

As shown on the left of Fig. 5.7, the L_1-logistic regression procedure almost perfectly extracts the correct cues for all observers: When training with data from the Energy observer, a set weight close to unity is attributed to energy, while spectral and envelope predictors are ignored. Similarly, for the Spectrum and Envelope observers, the set weights are almost entirely placed on the sound or envelope predictors, respectively. Weights on predictors corresponding to non-critical features are efficiently suppressed, as expected given the sparseness property (section 2.3.1). Results are stable across a wide range of SNRs, except for the very highest observer performance levels.

As expected given the interdependence of predictors (section 4.2), the results for reverse correlation on the right of Fig. 5.7 show little variation among different observers

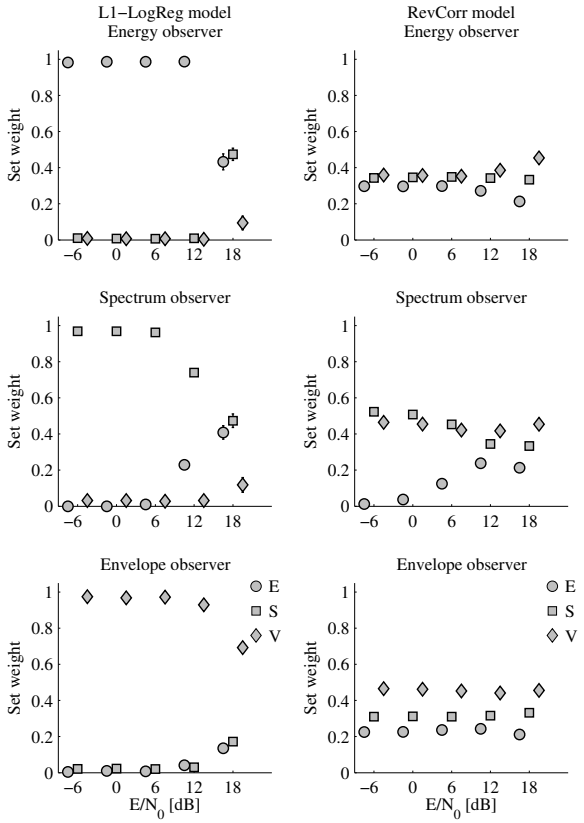


Figure 5.7: Set weights for L_1-regularized logistic regression (**left**) and reverse correlation models (**right**), for deterministic Energy, Spectrum and Envelope observers (top to bottom). The markers represent relative set weights on the energy (E), spectrum (S) and envelope (V) predictor sets (different symbols, see legend) across SNRs.

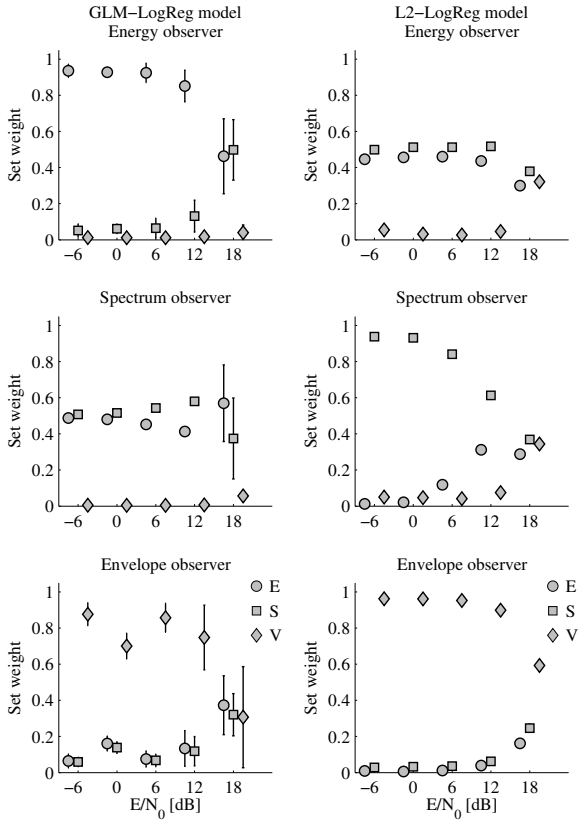


Figure 5.8: Same as Fig. 5.7, results for non-regularized (left) and L_2-regularized (right) logistic regression fits.

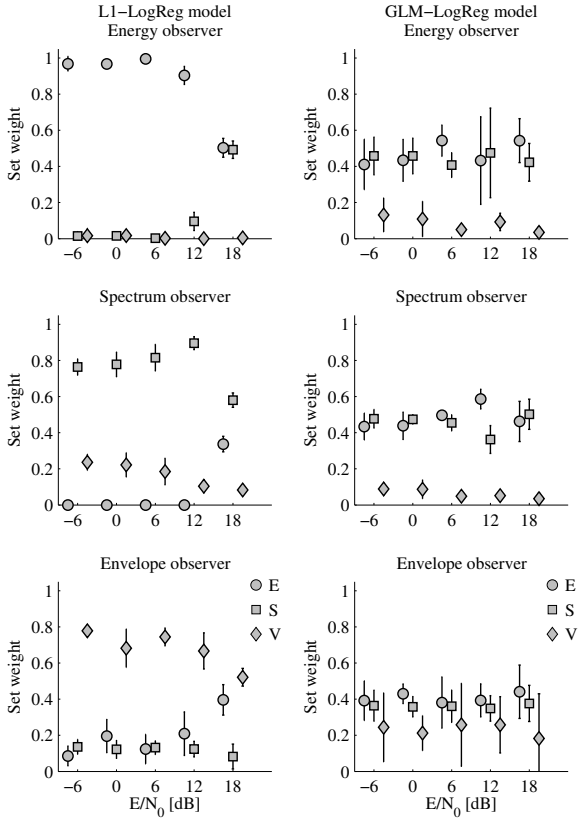


Figure 5.9: Same as Fig. 5.7, results for probabilistic observers at high noise levels.

in terms of set weights: For all observers, spectrum and envelope predictors receive a significant proportion of set weights around 0.3-0.5, usually with a smaller part attributed to energy. When dropping the regularization constraint, multiple logistic regression is still able to reconstruct differences among observer cues, but non-critical features are less efficiently suppressed, in particular for the Spectrum observer, as seen on the left of Fig. 5.8. In addition, extracted weights become less stable across repetitions resulting in larger error bars.

As discussed in section 2.3, different regularization paradigms correspond to diverging constraints on the model which in turn determine the resulting weights. While an L_1-regularization favors a solution with few nonzero weights, an L_2-regularization generally results in weights distributed across many predictors. Indeed, L_2-regularized logistic regression (Fig. 5.8, right) reconstructs the Energy observer with most of the weights attributed to the sound spectrum in contrast to the L_1-logistic regression which almost exclusively relies on the single energy predictor (compare top panels in Figs. 5.7 and 5.8).

In terms of predictive accuracy and stability, there is little reason to prefer either the L_1- or L_2-regularized algorithm. However, this study was based with one central assumption: Given multiple equally performing models, the one that uses the least number of predictors is preferred. Consequently, the sparse L_1-norm regularization should be strongly favored.

Next, the stability of the procedures was tested with data from probabilistic observers. As expected, prediction performance declines (Fig. 5.4) and extracted set weights generally become more noisy as observers become less reliable (Fig. 5.9). However, the

extraction of critical cues with L_1-logistic regression remains robust. The Energy observer is precisely identified (a set weight on the energy predictor reaching unity), while for the Spectrum and Envelope observer the critical cues consistently achieve a set weight of 0.7-0.9 on the correct cue.

In the same conditions, the non-regularized logistic regression algorithm generates strongly fluctuating weights which do not reflect the critical observer cues. For all observers, a weight of only 0.5 or less is attributed to the critical predictor set. Results for reverse correlation are stable and mostly unaffected by observer noise, but exhibit the same shortcomings as for the deterministic case. In conclusion, L_1-regularized logistic regression outperforms both standard logistic regression and reverse correlation in terms of robustness and reliability of observer reconstruction under conditions with high noise and lapse levels.

5.2.4 Model Filters

As a next step, the extracted spectral and envelope filters were compared with the strategies underlying the Spectrum and Envelope observers. Again, the efficiency of the logistic regression models in comparison to the reverse correlation procedure was evident: Both L_1- and standard logistic regression algorithms precisely reconstructed the filter shapes employed by the Spectrum and Envelope observers, except for the highest SNR (Fig. 5.10). The weights from the reverse correlation fit strongly depended on signal level (Fig. 5.11). For low levels of SNR, they were most similar to the underlying observer filters. For higher SNRs, however, filters degenerated both in the sound spectrum and envelope domain and diverged from the actual observer strategy.

For independent predictors, reverse correlation is expected to extract the true weights. Because predictors characterizing the sound spectrum are mostly independent, the corresponding filters are well reconstructed. Only for high SNRs, spectrum predictors that refer to the signal frequency and its sidebands become correlated due to spectral splatter. Consequently, the extracted spectrum filters show distortions in this region. In comparison, envelope filters are less well reconstructed overall even for low SNRs, reflecting the fact that envelope predictors are generally correlated for TiN stimuli (see section 4.2). Taken together with the results for the set weights, the extracted cues barely represent the underlying observer strategy, even though reverse correlation reliably predicted observer behavior (Fig. 5.4).

Even for highly probabilistic observers (10% lapse, 200% decision noise), both non-regularized and L_1-regularized logistic regression robustly reconstructed the shape of the spectral filters underlying the decision mechanism as demonstrated in Fig. 5.12. Variability is increasing, as was expected for more noisy data. Only for the Spectrum observer at very high SNRs the extracted filters degenerate. At these conditions, however, none of the fitting procedures reliably extracts information about the observer strategy, as already observed in Fig. 5.9. Results for the reverse correlation procedure are almost completely unchanged for noisy data, and still depend on SNR.

5.2.5 Comparison across Conditions

In order to evaluate the results across all SNRs, observers and noise conditions, collapsed data in terms of estimation error and variability for both set weights and extracted filters are displayed in Fig. 5.13. The set weight error corresponds to the fraction of weights erroneously attributed to predictors corresponding to non-critical features. For example, when a set weight of 0.8 is attributed to energy for the Energy observer, the error amounts to $1 - 0.8 = 0.2$. Set weight variability is defined as the average standard error of the estimates in each condition across repetition (error bars in Fig. 5.7, averaged across the three set weights for each SNR). In almost all situations, L_1-logistic regression

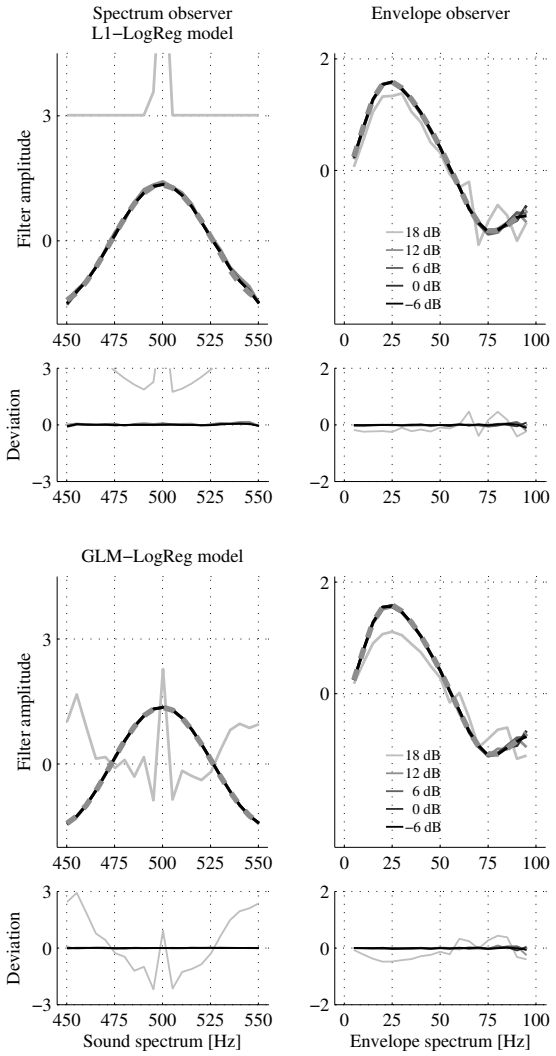


Figure 5.10: Observer filters as extracted by L_1-logistic regression (**top**) and standard logistic regression (**bottom**). The extracted filters for the Spectrum observer (**left**) and Envelope observer (**right**) are displayed for different SNRs (gray level, see legend). The dashed gray line represents the true underlying observer filter. The narrow graph below each plot displays the deviation of the extracted filters from the true filter shape.

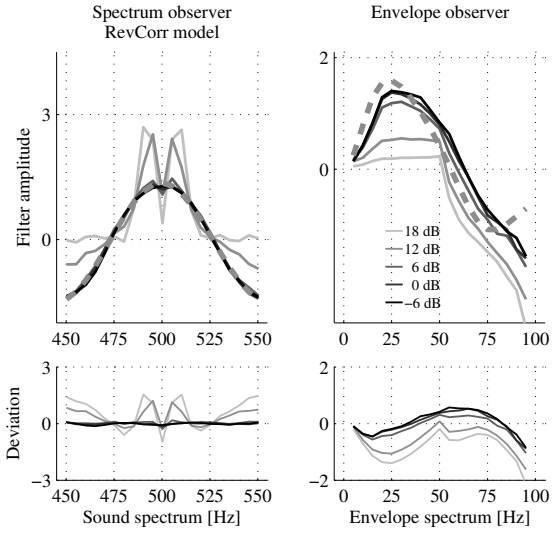


Figure 5.11: Same as Fig. 5.10 for reverse correlation.

results exhibit little error and variability. Only at high psychometric performance levels, observer identification fails. In contrast, standard logistic regression leads to increased error and variability in the majority of conditions. Reverse correlation results are stable, but consistently display large set weight errors (around 50% or more).

As an overview over the quality of the extracted filters, filter error and variability were computed. Error was defined as the RMS error of the filter amplitudes compared between the true observer strategy and the estimated model filter. For an indifferent model filter (constant at zero), the error was scaled to unity. Variability designates the average standard error of the filter amplitudes across repetitions of the simulations.

In terms of extracting the true observer filter shapes, non-regularized logistic regression has a small advantage over the L_1-regularized method (average filter RMS error for standard logistic regression $\overline{E_f}$ = 0.20, for L_1-logistic regression $\overline{E_f}$ = 0.23).[4] Thus, the introduction of the sparseness constraint generally leads to slightly distorted filters. This effect is small, however, compared to the overall range of the filter errors ($E_f \approx 0\ldots1$). While reverse correlation is generally able to extract the shape of the spectral filters for the Spectrum observer ($\overline{E_f}$ = 0.23), it fails with the envelope filter ($\overline{E_f}$ = 0.41), as already seen in Fig. 5.11. In terms of filter variability, the L_1-logistic regression offers a small improvement (average filter variability $\overline{V_f}$ = 0.25) against standard logistic regression ($\overline{V_f}$ = 0.28), i.e., results are slightly more stable against random fluctuations in the data. In terms of filter variability, reverse correlation seems to operate even better than the other procedures ($\overline{V_f}$ = 0.13). In practice, however, this does not provide a genuine advantage, because filter error was substantially larger, in particular for the Envelope observer.

[4] These and the following values were determined across all data points in Fig. 5.13, except for those where psychometric performance exceeded 95% to exclude situations with strongly degenerate model fits.

5.2.6 Observers with Mixed Cues

Additional analyses were carried out for data generated from strategies that mix different cues. The goal was to investigate whether individual cues were still identified and whether the set weights provide a quantitative measure of the mixing proportions. The data were analyzed in exactly the same way as for the single-cue observers. In the majority of conditions, the variations in decision strategy are well reflected in the set weights, as displayed in Fig. 5.14. In particular for the observers with Energy/Envelope and Spectrum/Envelope mixing strategies, the weights agree well with the true mixing

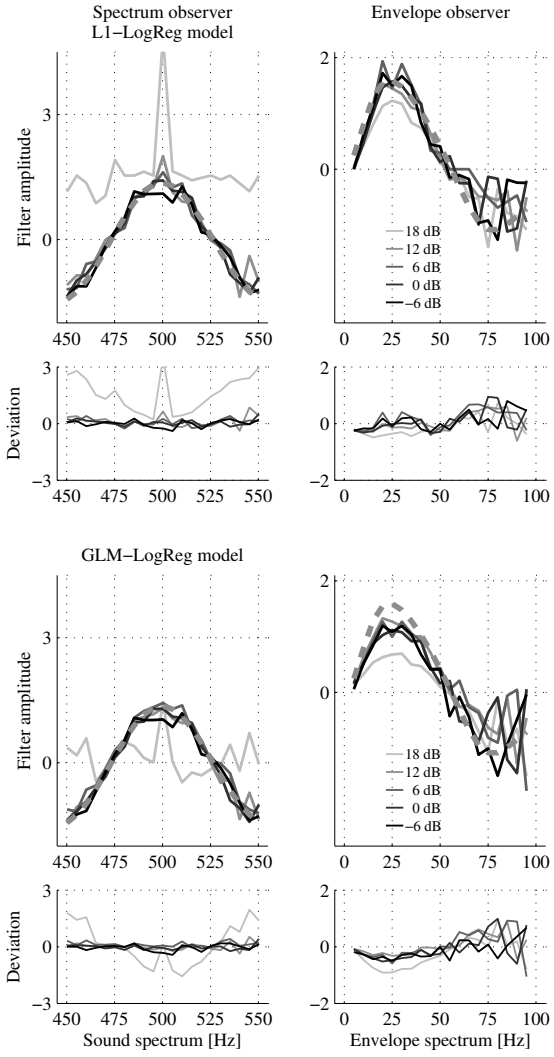


Figure 5.12: Same as Fig. 5.10 for probabilistic observers at high noise levels.

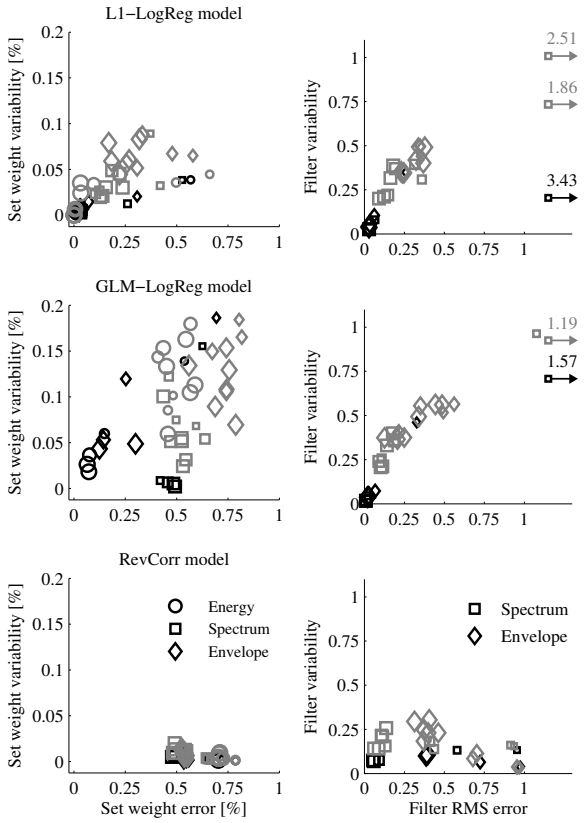


Figure 5.13: Comparison of modeling procedures across conditions. Different symbols correspond to observers (see legend), gray level denotes deterministic (black) and probabilistic observers (gray, for both medium and high noise levels), marker size represents observer performance with the smallest symbols indicating performance close to 100%-correct. **Left:** Errors and variability of set weights across conditions (compare with Figs. 5.7, 5.8 and 5.9). **Right:** RMS error and variability of extracted filters (compare with Figs. 5.10, 5.11 and 5.12). Several outliers are marked with arrows, the values given on top indicate their true position on the x-axis.

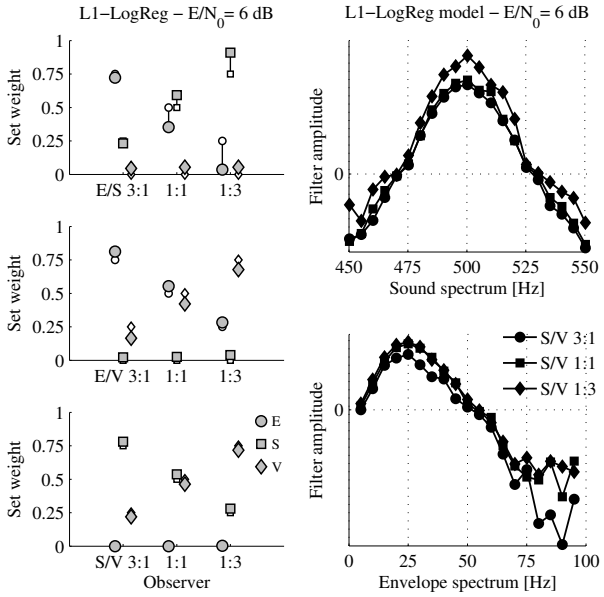


Figure 5.14: Modeling results for observers that mix different decision strategies across trials. **Left:** Set weights for observers that mix Energy/Spectrum (E/G, top), Energy/Envelope (E/V, middle) and Spectrum/Envelope (G/V, bottom) strategies at proportions as indicated on the x-axis. The connected white symbols correspond to the correct solution, the true mixing proportions. **Right:** Spectrum and envelope filters extracted for observers mixing Spectrum and Envelope strategies at different proportions (see legend).

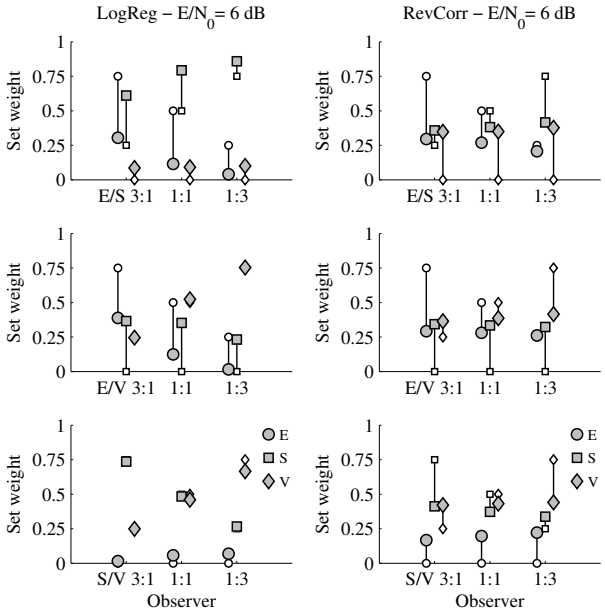


Figure 5.15: Set weights for mixed observers (top to bottom: Energy-Spectrum, Energy-Envelope, Spectrum-E nvelope) for classical model fitting procedures (**Left:** standard logistic regression, **Right:** reverse correlation). Different symbols indicate the relative weight on individual predictor sets (energy, spectrum, envelope). The connected open symbols represent the true proportion.

proportions. The underlying filters of the mixed strategies are accurately reconstructed corresponding to the intermixed cues for Spectrum/Envelope observers. Filter estimates for mixed-strategy observers are not as reliable as in the single-cue case, since, effectively, fewer trials are available to infer the individual cues. For comparison, Fig. 5.15 shows the results for mixed observers with classical model fitting procedures. Non-regularized logistic regression failed in conditions that include the energy observer cue with set weights deviating up to 0.5 from the true mixing proportion, while reverse correlation failed under all conditions, designating significant weights around 0.25-0.35 to predictors that were not relevant as observer cues.

5.3 Discussion

Taken together, the present results demonstrate that the L_1-regularized multiple logistic regression procedure reliably reproduced the predefined decision strategies of different "artificial observers" performing a TiN detection task, in contrast to a non-regularized multiple logistic regression and a reverse correlation procedure.

The primary advantage of using an L_1-regularizer is the ability to explicitly distinguish the critical observer cues even under challenging conditions with covarying model predictors and noisy observers. Results for L_1-logistic regression are close to the ideal solution and stable across a wide range of SNRs and corresponding psychometric performance levels as shown in Figs. 5.7 and 5.10. For deterministic decision strategies, observer identification and reconstruction was very good, i.e., critical cues and employed filters are confidently reproduced by the model, as long as psychometric performance does not exceed 95%. In contrast to non-regularized procedures, correlated but non-critical features are efficiently suppressed (Fig. 5.7).

The application of regularization prevents over-fitting and gives rise to a robust estimation of model parameters. This is particularly obvious when comparing the data with results from non-regularized logistic regression which exhibit strongly fluctuating weights in particular with noisy observer data (Fig. 5.9). These fluctuations appear because predictors associated with the energy and sound spectrum are linearly interdependent. The model is "ill-conditioned" because the observers strategies are not uniquely defined in terms of the considered predictors.[5] Even for the Envelope observer, completely ignoring both energy and spectrum, the model placed considerable weights on these predictors—essentially, energy and spectrum weights canceled each other out and were then effectively ignored in the decision mechanism. The large variations across repetitions are driven by the combination of small random fluctuations in the data and the fundamental instability of the under-constrained solution. This particular problem was one of the reasons, why regularization was originally introduced (see section 2.3): By adding a constraint, an ill-posed optimization problem becomes well-defined and robustly solvable.

[5] An increase in weight on energy is equivalent to a constant increase across all spectrum weights.

On the other hand, using reverse correlation for model fitting provides stable, but wrong weight estimates, that do not allow a valid reconstruction of the underlying observer cues in most conditions (Fig. 5.7). This result is expected, since the assumption that predictors are independently distributed is violated—an assumption that is central when applying reverse correlation for directly estimating relative weights.

In terms of trial-by-trial predictions, both regularized and non-regularized logistic regression procedures display a clear advantage over reverse correlation, anticipating the observer's decision in almost 100% of the trials, independent of signal level (Fig. 5.4). This advantage vanished for probabilistic observers. While the reverse correlation method incorporated non-critical features into the observer model, this only marginally impaired the predictive power of the resulting model. Nevertheless, only the L_1-regularized logistic regression procedure was able to consistently uncover the underlying decision strategies with their corresponding observer cues. Non-regularized regression attributed a large

proportion of the weights on predictors associated with non-critical features, and overall results showed stronger fluctuations (Fig. 5.13) due to over-fitting.

Apart from identifying the critical features, the L_1-logistic regression also reproduced the precise weighting applied to the associated predictors in terms of spectral filters (Figs. 5.10 and 5.12). The sparseness constraint of the L_1-regularization pushes individual weights towards zero. Potentially, this property may have distorted the extracted filter shapes compared to a non-regularized logistic regression. However, this effect was barely noticeable (see section 5.2.5), demonstrating the flexibility of L_1-regularization to even recover mildly sparse solutions (Donoho, 2004). A reverse correlation procedure only reconstructed filter shapes for non-correlated stimulus features, but failed for those that covaried (Fig. 5.11).

Overall, the L_1-logistic regression was able to accurately extract individual observer decision rules from stimulus-response data alone, both qualitatively in terms of the employed cues and quantitatively in terms of the relative weights. The other two methods compared here, non-regularized logistic regression and reverse correlation failed in different aspects of this task. Traditional reverse correlation could perform better in terms of predictions if only trained on a subset of predictors corresponding to the current observer cue. However, in practice one typically does not know in advance which predictors are relevant. Instead of running through multiple combinations of predictors, the proposed method recovers the relevant ones in a single step, without the experimenter knowing which predictors to preselect. In addition, this study explored the impact of correlated predictors. A priori, reverse correlation was expected to produce misleading results because of the violation of the fundamental assumption of independent predictors. Still, it was not clear in advance *how strong* the effect of this violation on model estimates would be. The main reason for including reverse correlation was to show how much better a regularized regression method can do with both deterministic and noisy data under statistically challenging conditions explicitly violating the independence assumption.

5.3.1 Limitations – Correlations, Signal Level and Features

As any method, the application of L_1-logistic regression has limits. First of all, principally no method can tell apart two perfectly correlated stimulus features, one of which determines behavior. Generally, multiple logistic regression takes into account correlations between input variables, in contrast to reverse correlation. Accordingly, both L_1- and non-regularized logistic regression procedures deliver good fits overall for deterministic observers. When the decision mechanism is noisy, however, the results indicate that only by introducing the regularization constraint correlated but non-critical features are suppressed. Even though this modern statistical approach is less effected by correlations and data fluctuations than earlier weight estimation methods, there is a limit in how much interdependent predictors of non-critical features can be suppressed. This limit depends on the amount of data as well as the structure and strength of the correlations. When strongly correlated predictors are included in the analysis, the results need to be interpreted with appropriate care.

Furthermore, since sparse regularization is equivalent to the prior assumption that the observers employ only few features, the amount of regularization must be well balanced. Otherwise, by applying too much regularization, features that have lower priority to the observer, but are still consistently used, may remain undetected. The simulations demonstrate that even when observers rely on features corresponding to a large proportion of predictors (e.g., the mixed observers described in section 5.2.6 were modeled with 40 out of 41 predictors) the sparseness constraint still allows for all of these cues to be detected.

As mentioned in sections 5.2.3, 5.2.4 and 5.2.5, recovering the observer strategy was increasingly more difficult at high signal levels. When an observer correctly classifies

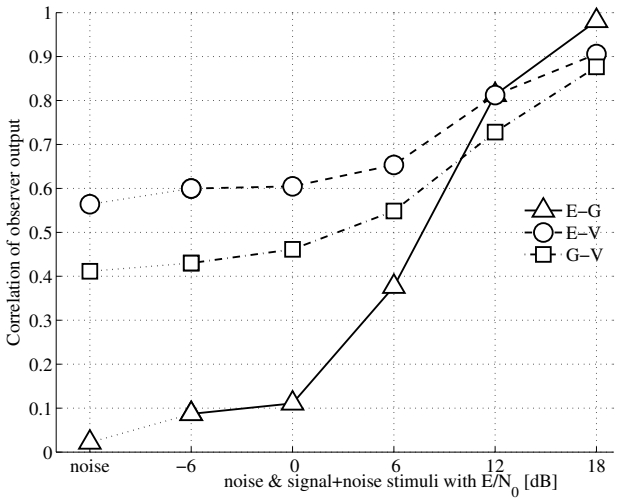


Figure 5.16: Correlations of the decision variables of the simulated observers (E–energy observer, G–spectral/gammatone Filter observer, V–envelope observer, symbols indicate different combinations, see legend) for noise-only stimuli (leftmost position on the x-axis) and mixed noise & signal+noise stimuli at different SNR (indicated on the x-axis). The output of the energy and spectral observer are hardly correlated for noise-only stimuli.

the stimulus in every single trial, there is no way to identify which cue she was relying on. The particular stimuli for which an observer gives correct (hits+rejects) or wrong (misses+false alarms) responses can be understood as a signature of his decision strategy. Even at identical observer performance, responses in individual trials may still differ between observers. With increasing signal levels the psychometric performance increases as well, and observers must become more similar and in consequence harder to discriminate. Figure 5.16 shows the relationship between signal-to-noise ratio and the correlation between the responses of different observers. In the present simulations, the strategy of deterministic observers could be reliably discriminated at a psychometric level of up to 95%-correct—even though at this level observers give a correct (and thus identical) response on 3800 out of the 4000 trials. In the extreme case of 100%-correct, all observers respond identically and are principally indistinguishable.

On the other extreme end, the lower the signal level of the stimuli, the larger the variance in individual behavior, and the easier the strategy can be reconstructed, at least for deterministic observers. For probabilistic observers, the lower the signal level, the more their responses are governed by decision noise. Thus, more data are required to discover regularities in the relationship between stimuli and responses.

The choice of stimulus features and associated predictors is considered the most critical part of the approach determining to a large extent the success of the method. Obviously, if the critical cues are not among the features assessed during the analysis, they cannot be identified. Clearly, the machine learning-based system identification technique still relies on the user's knowledge, skill and intuition at this crucial step. In this respect, the approach reveals the same limitations as classical correlation or regression analyses, but it is not worse either. However, it allows to simultaneously investigate large sets of predictors, thereby massively increasing the search space of stimulus features. Instead of just a handful, several dozens of presumed observer cues and their combinations can be tested in a single step. In addition, they need not be under direct control of the experimenter and may be statistically interdependent.

If the cue that the observer relies on is different from the features analyzed in the

investigation, two situations may arise: First, the fitted model may not be able to predict observer responses, giving a strong hint that the critical cue is not among the features. Second, the model may provide reliable predictions, because a subset of the candidate features is related to the observer cue. In this case, a mixture of cues is identified as underlying the observer's decision. A thorough analysis of this mixture may offer hints which other features ought to be investigated to establish an even simpler observer model. Whatever the case, it is only the analysis that needs to be repeated, not the entire experiment.

5.3.2 Application of the Method to Experimental Data

Although this study demonstrated the potential of the proposed cue extraction method with Tone-in-Noise detection in a YES/NO-paradigm, the proposed cue identification approach is not limited to this particular experiment. It can equally well be applied to a variety of different paradigms, e.g., 2IFC-experiments, and perceptual discrimination tasks. The only essential change compared to the procedure presented here would be a different choice of sound features and associated predictors according to the task at hand.

The data quantity used for training in the simulations (4000 trials per condition) can be collected during moderately expensive psychophysical experiments. As the results are solid, not only for deterministic observers, but also for unpredictable behaviors such as decision noise and lapse, it is expected to be directly applicable to real observer data. The method is based on the common assumption that response behavior is stationary. Thus, care must be taken that this is the case with the investigated observers. Periods of perceptual learning ought to be excluded from the analysis.

In the present simulations, data for different levels of SNR were separated with stable results across a wide range of signal levels. Consequently, if observers are expected to employ a static decision mechanism independent of level, data from different SNRs, e.g., from an adaptive procedure, can be pooled for the analysis. Conversely, if data from different SNRs is analyzed separately, SNR-dependent changes in strategy can be uncovered, e.g., observers may be switching between different cues depending on how informative they are and how much attention they require. Even if observers switch between several stimulus cues during the experiment, the proposed method is able to quantitively recover the underlying mixing ratios (see section 5.2.6).

The results demonstrate that L_1-regularized logistic regression has a major advantage over earlier methods: As long as the critical cue is present, including many non-critical stimulus features in the analysis is not a problem, even if the associated predictors are not statistically independent from the critical one. This observation is the major advantage over using correlational techniques, where correlated predictors usually result in misleading weight distributions, as well as over non-regularized logistic regression, where these dependencies and the presence of noise may lead to ill-conditioned solutions and instabilities in the estimated model variables.

When the regularization parameter λ is properly chosen (see description in section 5.1.4), the regularized procedure produces a model that is "as sparse as possible" without compromising the fit to the data. There is little risk of obtaining a model that fails to capture the data because it is overly sparse. Even when an L_1-regularized logistic regression is not able to extract a sparse solution (because of noise in the data, or because a sparse solution simply does not exist), it approaches a standard logistic regression both in terms of quality of fit and reliability. In that sense, the regularized logistic regression has no disadvantage with respect to the non-regularized method.

In these simulations, generating more "experimental" data was cheap. For the sake of simplicity, the optimal regularization parameter λ was therefore optimized beforehand with additionally generated data sets (see section 5.1.4). For real behavioral experiments,

where the amount of data is usually limited, the application of cross-validation is recommended instead. This method is a means to generate independent data subsets that can be used for proper model training and testing (as described in section 2.3) and is widely used in machine learning (Bishop, 2006, chap. 1.3); (Hastie *et al.*, 2009, chap. 7.10). It offers a more elaborate but equally reliable way to optimize regularization.

As discussed in section 5.2.1 and shown in Fig. 5.4, the relationship between model agreement and observer consistency provides an informative measure to assess the quality of the established observer model. For this reason, experimenters are encouraged to measure consistency of their observers, e.g., by using double-pass data for a subset of the stimuli.

Even though this study focused on binary response data, the current method can be generalized in a straight-forward way to experiments that measure response probabilities to individual stimuli, e.g., frozen noise experiments: Model training proceeds in the same way, with the only difference that individual stimuli are presented multiple times with a distribution of responses corresponding to the measured probabilities. For example, if an observer had a 80% chance of answering "Yes" to a particular stimulus, this stimulus is added to the data altogether five times: four times with a "Yes" and once with a "No" response. The resulting model can then be used to probabilistically predict responses on a stimulus basis. The predicted and measured response probabilities could then be compared as a measure for the goodness of fit.

5.4 Conclusions

This simulation study compared different methods for estimating relative combination weights from behavioral data under the assumption of a linear decision process. In simulations of a standard auditory psychophysics paradigm, it was demonstrated that an L_1-regularized multiple logistic regression procedure reliably reproduced the detailed decision strategies of deterministic as well as probabilistic observers in a Tone-in-Noise detection task. The resulting "sparse" observer models also allowed reliable trial-by-trial predictions of observer behavior. In addition, they followed the law of parsimony, relying on the smallest subset of stimulus features that was both necessary and sufficient to explain observer responses.

Importantly, near-optimal solutions were obtained under conditions where classical procedures hit fundamental hurdles. In addition to random fluctuations in the data, the features investigated in this study exhibited significant interdependencies, a condition that regularly appears in auditory experiments. Consequently, both correlation analysis and non-regularized multiple logistic regression procedures failed to extract the unique decision mechanisms underlying the simulated behavioral data. As an advantageous alternative, the regularization-based technique provides robust estimates of relative weights from psychophysical data in conditions that were not accessible due to constraints and limitations of classical techniques.

In conclusion, L_1-regularized logistic regression represents an efficient and reliable method for extracting stimulus cues that are critical for observer behavior. The proposed procedure holds substantial benefits over classical weight estimation techniques, and at the same time, its scope is general enough such that it can be applied in a wide range of auditory discrimination tasks. It opens up new opportunities by giving experimenters more freedom in choosing their set of potential cues, independent of which stimulus properties they directly control.

Chapter 6

Psychophysical Experiments

In a psychophysical experiment, which was thoroughly prepared and executed, large amounts of data were collected from several listeners performing the classical narrow-band TiN detection experiment presented and discussed in chapter 3. These data were then analyzed using the sparsely regularized regression procedure presented in chapter 2 that had the main purpose of extracting the perceptual cues and decision strategy of the observers. It is demonstrated that listeners follow highly idiosyncratic decision strategies and employ an individual mixture of auditory features including sound energy, spectrum and envelope. A linear decision model combined with nonlinear stimulus preprocessing appears to be largely sufficient to predict observer behavior on a "molecular" trial-by-trail level. In addition, responses of half of the observers were found to depend on decisions in preceding trials which generally calls for a consideration of such behavioral factors in psychophysical experiments.

6.1 Methods

6.1.1 Subjects, Stimuli and Setup

Six listeners, 5 male and 1 female with a mean age of 25.8±2.8 years (mean ± sd) performed the experiment after giving written informed consent. Subjects were naïve with respect to the purpose of the experiment. They were seated in a quiet and darkened laboratory room in front of an LCD screen providing instructions as well as visual feedback. In order to determine whether remaining environmental sounds were sufficiently shielded through the headphone inserts, hearing thresholds were measured both in our laboratory and a double-walled IAC sound-insulated chamber. Thresholds fulfilled standard audiometric criteria for normal hearing and differed on average by no more than ±2 dB, except for one observer (see Appendix B for details on the listener hearing tests).[1]

[1] Subject S6 reported a transient monaural Tinnitus-like percept during one of the hearing tests, which significantly increased the hearing threshold. He did not report such a sensation during the main experiment and the concerned narrow band of frequencies (3–4 kHz) was 2–3 octaves away from the experimental stimuli.

The masker stimulus consisted of a 100 Hz wide band of noise centered at 500 Hz and presented at an average sound pressure level of 70 dB. In signal trials, a pure tone at 500 Hz was added (see section 4.1 [p. 25] for details). The signal tone was presented at different sound levels resulting in four conditions of signal-to-noise ratio (SNR or E/N_0): 9, 11, 13 and 15 dB for observers S3 and S6; for the remaining four listeners the signal was presented at 7, 9, 11 and 13 dB. These settings were chosen so that observers roughly spanned the same range of psychophysical performance (the choice is discussed in more detail in section 4.5). The signal tone was presented simultaneously with the noise mask. All sounds were 200 ms in length with 50 ms cosine-squared on-/off-ramps to minimize spectral spread. Thus, the steady state section was 100 ms in length.

Stimuli were presented binaurally with *Etymotic ER-2* (Etymotic Research, Inc., USA)

in-ear headphones providing −30 dB external noise exclusion according to the manufacturer. The analog signal to the headphones was generated with an external 24-bit *RME Fireface 400* (Audio AG, Germany) sound card and amplified with a fan-less *Naim Headline* (Naim Audio Ltd., UK) headphone amplifier. Headphones were calibrated with a *G.R.A.S. RA0045* (G.R.A.S. Sound&Vibration A/S, Denmark) headphone coupler-microphone, as well as a *B&K Falcon 2669* microphone preamplifier and *B&K Nexus 2690* conditioning amplifier (Brüel & Kjær A/S, Denmark). Digital signal generation and output, as well as response registration and feedback display were controlled from an *Apple Mac Pro* desktop computer running Matlab with the *Psychtoolbox-3* extension (Kleiner *et al.*, 2007). Further details concerning the setup and calibration procedure are presented in Appendix A.

6.1.2 Experimental Procedure

Subjects had to perform a single interval identification task ("YES/NO-task") where the signal was present ("signal trials") or absent ("no-signal trials") with 50% probability. In each trial, the presentation of the sound (200 ms) was followed by a response window adapted to the typical response speed of the subject (800 ms for S1 and S2, 700 ms for S5 or 600 ms for all other subjects measured from sound offset). Observers indicated the perceived presence ("Yes") or absence ("No") of the signal by pressing one of two buttons on a dedicated USB-device that recorded high-precision response times with an internal clock (*Response Time Box*, developed by Xiangrui Li).[2] Trials where observers did not respond or gave responses too early (before sound offset) or too late (after the response window) were discarded as invalid. The beginning of the next trial was determined by a rest interval lasting 1000 ms (from the end of the response window to the onset of the next sound).

[2] Department of Psychology, University of Southern California, xiangrui.li@usc.edu, http://lobes.usc.edu/RTbox (last accessed on 09/12/12)

Stimuli were presented in blocks with a fixed signal-to-noise ratio E/N_0 to allow listeners to form a consistent response strategy in each signal level condition. The number of trials per block (50 trials for observers S1, S2; 60 trials for S3, S4, S5, S6) was adjusted so that each listener could comfortably generate the required amount of data. The first 10 trials of each block included immediate visual feedback indicating the correct response in order for subjects to (re-)adjust to the task. However, only the remaining trials in a block—those without feedback—were used in the subsequent analysis. At the end of each block, subjects were visually presented with their average performance (proportion of correct and valid responses in the preceding block). After six blocks, i.e., 300–360 trials (including 240–300 trials without feedback), the signal-to-noise ratio was lowered to the next level. During each session 1200–1440 trials were performed with all four E/N_0-levels in decreasing order. In this way, observers always started with the easiest condition and slowly progressed towards more difficult levels.

Our data analysis assumes that observers employ a consistent decision strategy. Although feedback was required for learning the task, it may confuse listeners in particular at low signal levels. Under these conditions they would regularly receive a negative trial-by-trial feedback prompting them to try and adjust their strategy in order to avoid mistakes. This behavior would violate the initial assumption. The feedback strategy is discussed in more detail in section 4.5.

A session lasted about one hour, including self-paced breaks between blocks. In all, each subject performed between 25 to 33 sessions. The initial 7 to 12 sessions were used for subject training and to find the appropriate E/N_0-levels as well as secondary experimental parameters (number of trials per block, length of the response interval). Afterwards, the E/N_0-levels and other parameters were fixed for the remaining 18 to 22 sessions so that each subject produced between 4'983 and 5'696 valid trials at each E/N_0-level. Only the latter data with fixed parameters was used for the investigations presented in this study. Across conditions, listeners completed 7,726–14,335 valid trials for training and 20,024–22,752 for analysis. In total, 217,621 valid trials were collected, 125,560 of which were analyzed and form the basis for the conclusions drawn below.

Across observers, roughly 1.5% of trials recorded during this period of data collection were registered as invalid. While observers S4 (5.6%) and S5 (1.9%) produced a considerable proportion of invalid trials, presumably because they tended to not respond when in doubt, the other four observers achieved a value of only 0.2–0.7%.

6.1.3 Observer Consistency Estimate

A subset of the stimuli was used to estimate observer consistency, i.e., the fraction of identical responses in repeated presentations of identical stimuli. Since I had not intentionally collected two-pass data with identical stimuli, I used stimulus samples that exhibited a strong similarity according to the correlation in their time series. I chose a threshold for the correlation of 0.95, so that at least 100 pairs per observer and signal level were classified as sufficiently similar to be regarded as two-pass pairs. The number of similar pairs for the different conditions varied widely, usually ranging from a few hundred for low SNR to more than a thousand for high signal levels. The estimation error for these values was computed from the number of pairs assuming a binomial distribution. For four observers, identical stimuli multiple times were unintentionally presented.[3] For these observers, the consistency computed from truly identical double-pair stimuli was compared with the consistency for similar stimuli, as assessed by the 0.95-correlation criterion, and found no significant difference in estimated consistency (1.3 ± 4.2 percent points)—a post-hoc corroboration of the (arbitrarily chosen) correlation coefficient criterion.

[3] Quite likely, the random seed was reset and the same random numbers were generated after the computing environment was restarted.

6.1.4 Linear Observer Model

The critical features, or "cues", that govern observer decisions were identified by analyzing trial-by-trial dependencies between stimulus characteristics and individual responses. Each stimulus feature is represented by a set of one or multiple predictors. I followed the common assumption that observer behavior depends on a linear combination of these predictors in each trial.

Consequently, the observer model was expressed as a weighted sum of predictors p_i (a set of values characterizing the stimulus) followed by a static nonlinearity, the logistic function S:

$$P = S\left[\sum_i w_i p_i + b\right] \tag{6.1}$$

with the model weights w_i and a bias term b (identical to Eq. 2.1). The function S gives rise to the model output P which corresponds to an estimate of the probability of a Yes-response given a particular stimulus. The weight parameters w_i are obtained by fitting the model to data from individual observers using a sparse logistic regression procedure (see section 6.1.6). The weights represent the basis for the prediction of trial-by-trial responses as well as for evaluating the relative importance of particular stimulus features—a feature is considered to be a perceptual cue if the fitted model attributes a substantial weight to the corresponding predictors.

6.1.5 Stimulus Processing

Stimuli were preprocessed to extract three potentially relevant auditory features and corresponding predictors as described in detail in section 4.3 (p. 26):

- the sound energy
- the fine structure power spectrum
- the envelope power spectrum

In addition to these stimulus characteristics, the analysis included the sequence of earlier responses as a behavioral predictor in order to account for sequential dependencies in response behavior: For each trial the responses in the five preceding trials within the same block were collected, coded as ±1 for "Yes"/"No", and 0 where no previous response

was available. The complete information for each trial, including stimulus features as well as response history, was then combined into a vector with 46 entries. The response in the current trial was collected as a separate variable.

6.1.6 Data Analysis

The data analysis was then performed as follows: For each subject and each signal level, the collected data were entered into an L_1-regularized logistic regression to estimate the optimal model parameters, with the vector of predictors as regressor and the observer's responses as output. This method is similar to a standard logistic regression with the additional property of enforcing a sparse weight distribution, i.e., model weights that are not explicitly necessary to predict the listeners' responses are suppressed (Schönfelder and Wichmann, 2012). For model fitting the toolbox *LibLinear* (Fan *et al.*, 2008) was used which provides an efficient implementation of L_1-regularized logistic regression. The entire data analysis was based on custom-written scripts for standard scientific computing software running on a desktop computer.

6.1.7 Three Measures of Predictive Power

In machine learning, model training and testing are generally performed on independent subsets of the data in order to rule out over-fitting (Bishop, 2006). As depicted in Fig. 6.1, for each observer and signal level a model was fit to a "training set" of data—a large random subset of the data (⅝ of trials). As a second step, the predictive accuracy of the model was measured on the smaller "test set" that consisted of the remaining chunk of data (⅛ of trials) for the same observer and E/N_0-level.

Three measures were used to estimate the predictive accuracy of the model: model-observer agreement, model likelihood and model deviance. The first two measures are strongly related—the agreement only relies on the predicted binary responses and provides an easily interpretable measure. The likelihood, as a standard measure in machine learning contexts, takes into account the response probabilities and, in consequence, offers more fine-grained results. It is useful for comparing models but hard to interpret on an absolute scale. A model with a large likelihood necessarily exhibits a strong agreement, but not vice versa.

In the following, r_i represents the empirically measured response of the listener, taking the values 0/1 for "No"/"Yes". The probability P_i corresponds to the probability of a Yes-response estimated by the logistic model given a particular stimulus. This probability P_i was converted into a binary response R_i ("Yes"/+1 whenever $P_i > 0.5$ and "No"/0 otherwise).

The model-observer agreement A is based on binary response predictions and corresponds to the percentage of trials where the model agreed with the observer:

$$A = \frac{1}{N} \sum_{i=1}^{N} I(R_i, r_i) \tag{6.2}$$

where $I(a, b) = 1$ if $a = b$ and 0 otherwise. The likelihood is computed as:

$$L = \prod_{i=1}^{N} P_i^{r_i} (1 - P_i)^{1-r_i} \tag{6.3}$$

Essentially, model agreement weighs all predictions equally, whereas with the likelihood a correct (or wrong) outcome that was predicted with higher confidence is rewarded (or punished) stronger.

Because likelihood is hard to interpret in absolute terms, I developed a third measure termed model deviance that is based on probabilities and provides an absolute quantification by comparing the *empirical* with the *predicted* probability for a Yes-response. Ideally, I would have directly compared this on the basis of individual stimuli. However, since every stimulus was only presented once, observer response probability could not be estimated on a stimulus-basis. Instead, empirical estimates of response probability were obtained by pooling stimuli. More precisely, all stimuli were first sorted according to the associated response-probability of a Yes-response *as predicted by the fitted observer model*. Stimuli with similar predicted response probabilities were then pooled into bins with a fixed number of items. Next, for the stimuli in each bin the average *empirical* probability was determined as the proportion of trials with recorded Yes-responses. Finally, the predicted probability averaged over each bin was compared with the empirical probability (for an example data set the result of this procedure is shown in Fig. 6.4).

To summarize the relationship between predicted and empirical response probability for the stimuli in each bin, the deviance was computed for each observer and signal level, i.e., the log-likelihood ratio between a hypothetical optimal model and the current model. For binomial data, deviance is asymptotically χ^2_N-distributed, with N designating the number of bins. When deviance is divided by N, its mean value represents an absolute measure independent of bin size. Asymptotically, this measure attains a theoretical lower bound of unity when a model completely describes all aspects of the data. In practice a model with average deviance values near unity (below 1.5) can be considered a "very good" fit, at least in psychophysics (Goris *et al.*, 2008; Wichmann, 1999; Wichmann and Hill, 2001; Collett, 2003).

Theoretically, the best estimate of model deviance would be attained in the limit of small bin sizes approaching the ideal level of individual stimuli. However, the smaller the number of trials collected in a bin, the larger the estimation error for the empirical response probability. On the other hand, larger bins increasingly pool stimuli that have large differences in terms of associated response probability, running counter the original idea of grouping "similar" stimuli.

Because there is no a-priori "correct" bin size, the analysis was performed with varying bin sizes ranging from 10 to 250. For bin sizes larger than 50 trials/bin, model deviance started to increase as shown for a single observer in Fig. 6.6 (left). Accordingly, to obtain the single estimates for model deviance per observer and signal level shown in Fig. 6.6 (right), the results were averaged over bin sizes ranging from 10 to 50 trials/bin.

6.1.8 Optimized Regularization with Cross-Validation

The parameter that controls the L_1-regularization, i.e., the amount of sparseness, needs to be optimized during model fitting. Therefore, a cross-validation (CV) procedure was performed *within the training set* as shown in Fig. 6.1 (for an extensive discussion of cross-validation procedures refer to Browne (2000)). For cross-validation, the training data (see section 6.1.7) were further subdivided into ten folds. The model was then trained on nine CV-folds and tested on the remaining one. This was repeated with all permutations of training and test folds resulting in an estimate of the likelihood for the training data set at a fixed regularization parameter (in that process the independent test data set remains untouched). The procedure was reiterated with different values of the regularization parameter. Increasing the parameter typically did not effect the estimated likelihood up until a critical value, where the likelihood would drop steeply with increasing regularization due to under-fitting.[4] In an automated procedure, the largest value of the parameter right before the drop-off was chosen as the optimal trade-off for a regularization strength that results in maximally sparse models that are still good predictors. Thereafter, the model was fit to the *entire* training set using the optimal regularization parameter, before the definite quality of fit was evaluated on the—as yet

[4] Usually, likelihood also drops with weak regularization due to over-fitting. However, because of the large amount of data compared to the number of parameters the effects of over-fitting at low regularization strength were small.

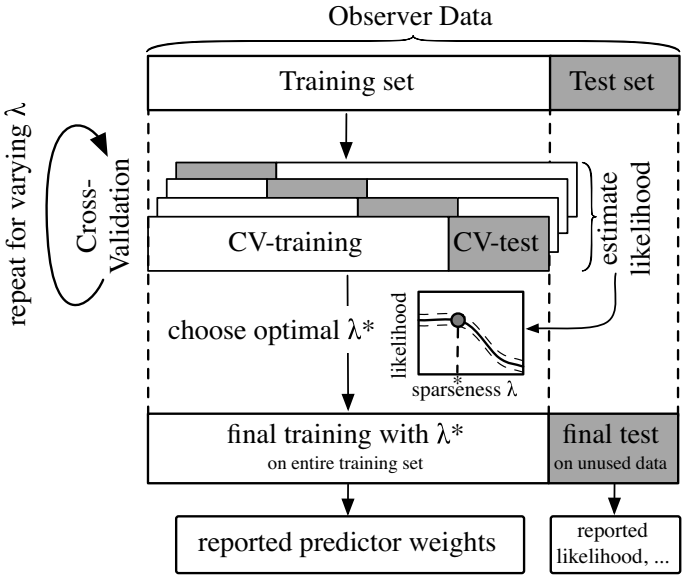


Figure 6.1: Schema depicting model fitting including parameter optimization through cross-validation (CV) and subsequent training and testing. **Top:** training/test-set splitting. **Middle:** optimization of the regularization-parameter through cross-validation. **Bottom:** final training and testing which result in the data reported for model parameters (predictor weights) and measures of predictive power (likelihood, agreement and deviance, see section 6.1.7).

untouched—independent test set.

After fitting, the predictor weights were extracted from the observer models. They were used for two purposes. First, the relative weighting of each of the four sets of predictors was determined (three "stimulus feature" sets, one "behavioral" set with earlier responses), i.e., the relative importance of each set for predicting observer decisions.[5] These set weights indicate whether and to what extent a particular stimulus or behavioral feature was relevant for predicting listeners' responses. Second, the weights on individual feature components in the fine structure and envelope domain were interpreted as spectral filters to uncover the detailed perceptual processing that determined the response. As explained in Appendix F, the predictor weights need to be rescaled before they can be interpreted as spectral filters. In this second step, the weights on previous responses were also analyzed to see how earlier decisions influenced the current trial. The obtained filters and weights provide insights into the specific quantitative processing that governed the behavior of individual observers.

[5] The full set of weights (associated to the 46 predictors) was projected onto the four dimensions of the weight sets by separately applying a vector norm to the weights w_i in each set s: $V_s^2 = \sum_{i \in s} w_i^2$. Each of the four obtained vector norms was then divided by the sum of all norms, resulting in relative set weights summing to one: $W_s = V_s / \sum V_s$.

The confidence intervals in the present study were estimated by repeating the analysis procedure multiple times with different subsets of stimuli used for model training and testing. If these subsets had been independent, the standard error of the estimate across repetitions could be directly interpreted as a confidence interval. However, the data subsets were overlapping and not independent: The data were split into six subsets and in each repetition one of these subsets was removed before data fitting. Generally, this results in an underestimation of the confidence interval of the model parameters. In order to account for this effect, the estimated standard error across repetitions had to be multiplied with a correcting factor. This factor was estimated numerically by computing the variance of a Gaussian random variable from both independent and overlapping data sets. The ratio between these variances was found to depend on the overlap between subsets and was not related to the overall size of the data sample. For the amount of

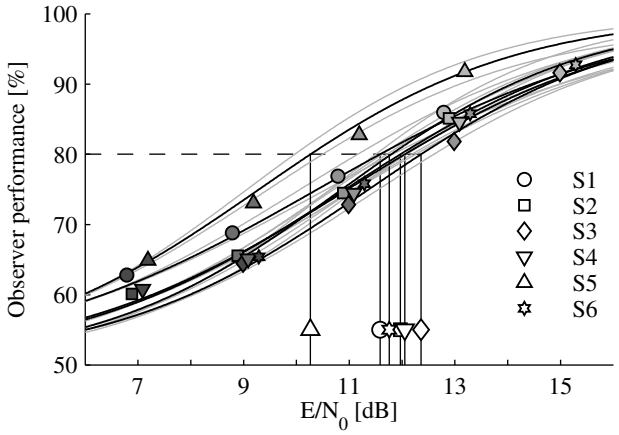


Figure 6.2: Observer psychometric performance. Markers represent raw data, lines represent fits of the psychometric function for all observers. Symbols indicate different observers, fill color corresponds to the index of the signal level for each observer (all observers were tested at four levels, but in different ranges of E/N_0: 7–13 dB or 9–15 dB). The vertical lines marked with open symbols represent 80%-correct thresholds for individual listeners.

overlap used during the main data analysis, the ratio was estimated to be 8.34 ± 0.01. Accordingly, the standard error of the model weights determined across repetitions was multiplied with the square root of this value. The values obtained for predictive power (agreement, likelihood and deviance) were estimated on the non-overlapping test sets. For this reason, their confidence intervals had not to be corrected. Unless indicated otherwise, the reported values for the estimated perceptual weights and the measures of predictive power represent the mean and standard error of the modeling results across repetitions.

6.2 Main Results

6.2.1 Analysis of Psychophysical Measures

In terms of raw psychometric performance, most listeners were hard to distinguish. The curve fits of the individual psychometric functions shown in Fig. 6.2 were computed with *psignifit* by Wichmann and Hill (2001). Average 80%-correct thresholds amounted to 12 dB (ranging between 11.6–12.4 dB), except for observer S5 who achieved a significantly lower threshold (10.3 dB). While most observers were unbiased with a proportion of Yes-responses within ±5% around 50%, listener S2 was more conservative with an average bias of 36%. Response times (measured from the offset of the stimulus) typically ranged between 200–400 ms.

With respect to general psychophysical performance measures, percent correct and response bias, the model generally reproduced the listeners' behavior with one consistent deviation as shown in Fig. 6.3 (left): Model responses resulted in a significantly increased percent correct score as compared to the observers (on average +9 ± 2 percent points). This effect can most likely be explained by the fact that the model implements the observer decision mechanism in a strictly deterministic fashion without any internal noise component (see Eq. 6.1). As regards response bias, no significant difference between model and observer was found (−1.8 ± 2.6 percent points).

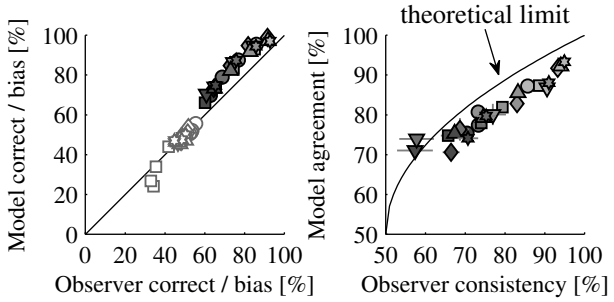


Figure 6.3: Comparing raw performance and agreement of observer and model. **Left:** Comparison of observer and model in terms of task performance ("percent correct", filled symbols, gray level represents signal level as in Fig. 6.2) and bias (percent Yes-responses, gray open markers). **Right:** Comparison of model-observer agreement against observer consistency (see section 6.1.3), observer symbols as in Fig. 6.2. The black line represents the theoretical upper limit for a hypothetical optimal model as derived by Neri and Levi (2006).

6.2.2 Model Predictive Power

Two measures were used to assess predictive power of the individual observer models on an absolute scale. First, observer consistency and model-observer-agreement were compared with an upper bound for a hypothetical optimal model as given by Neri and Levi (2006) and depicted in Fig. 6.3 (right). By relying on observer consistency, this bound takes into account the probabilistic nature of observer behavior when estimating the best possible prediction rates. Across the observed consistencies, the empirical agreement is close to the theoretical limit and follows this upper bound with an average distance of 5.2 ± 2.9 percent points, i.e., the model fit was close to ceiling.[6]

[6] This value was estimated from a model fit that excluded the predictor representing previous responses. The derivation by Neri and Levi (2006) follows the assumption that the model relies on the current stimulus only while the inclusion of "internal" determinants such as previous responses is not accounted for. Consequently, the model agreement shown in Fig. 6.3 was also estimated from models that did not rely on the "internal" predictors. Model agreement improved minimally when these predictors were included. The respective plots can hardly be distinguished while the reported average distance to the bound marginally decreases to 4.9 ± 3.0 percent points.

Second, the predicted and empirical response probabilities were compared for groups of individual stimuli as described in section 6.1.7. Overall, the model estimates showed a close match with the empirically observed behavior across the entire range of response probabilities (see Figs. 6.4 (black dots) and 6.5). As a summary measure for each observer and signal level, model deviance was computed. This analysis results in excellent values between 0.8–1.4 as shown on the right of Fig. 6.6. Across all observers and signal levels, model deviance achieved an average value of 1.15 ± 0.10.

As an independent post-hoc confirmation that response probability—as predicted by the model—orders stimuli in a behaviorally meaningful way, I analyzed the relationship between the estimated response probability and the measured response times of the listeners, which were not used during model fitting. According to Pièron's law, response times should become shorter with increasing sensory evidence (Pièron, 1914). Indeed, when the model assigned strong sensory evidence to a stimulus (Yes-response probability close to 0 or 1), observers responded faster compared to stimuli for which a high uncertainty was predicted (response probability close to 0.5) as shown in Fig. 6.4 (gray graph) with an example data set.

To quantify the effect, a quadratic function was fit that captured the dependence of observer response time on the predicted response probability for *individual* (not binned) stimuli. The function was centered at 0.5-probability, while offset and curvature were determined by minimizing the RMS-error. Sample data and the associated fit are presented on the left of Fig. 6.7, the same procedure was carried out for each observer and signal level separately. The quadratic fits provided an estimate of the average difference in response time ΔRT for stimuli for which the model predicted a large uncertainty (response probability close to 0.5) and with which it associated a strong confidence

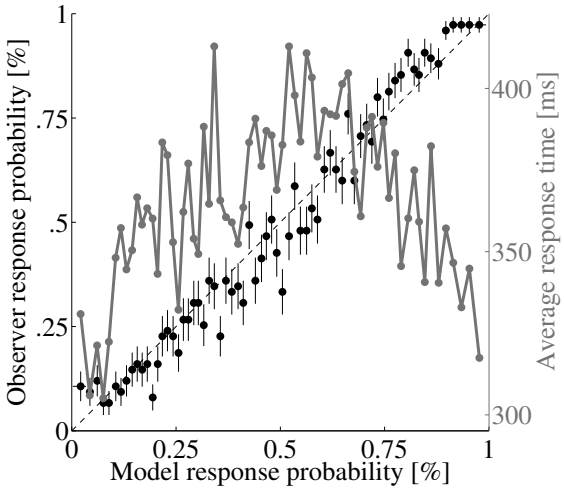


Figure 6.4: Comparing response probability between model and observer for individual stimuli (sample data for listener S4 at E/N_0 = 9 dB). Stimuli are binned according to the average response probability predicted by the model (x-axis) and compared with estimates of the empirical response probability (left y-axis). The average response time for stimuli in each bin is plotted in gray (right y-axis).

(response probability at 0 or 1). Although response time data were, as always, noisy overall (see left graph in Fig. 6.7), the parameters of the quadratic regression could be estimated with little error thanks to the large number of data points per observer. Mean response time differences across observers and signal conditions are plotted in Fig. 6.7 (right). They were significantly larger than zero and varied from 25 up to 140 ms in strict accordance with the behavioral prediction. Overall, the values increased with stronger signal levels, presumably because "easy" stimuli—resulting in fast responses—were *particularly* easy in these conditions as compared to "difficult" stimuli which appeared under all conditions.

In conclusion, the analysis of model-observer agreement and model deviance as well as the comparison of estimated response probability and measured response times confirm that the fitted models capture the perceptual decision behavior of the listeners on the level of individual stimuli for all individual observers and experimental conditions.

6.2.3 Relative Importance of Predictor Sets

After having confirmed that the trained models accurately capture the listeners' trial-by-trial behavior, the distribution of model weights for the individual observers was analyzed. The resulting relative predictor set weights for each listener and signal level are shown in Fig. 6.8. They are highly idiosyncratic in terms of the distribution of set weights, i.e., the relative importance of each of the four sets of predictors—a general summary is provided in Table 6.1. Only for two observers, S1 and S5, energy was the dominant auditory cue with set weights larger than 50%. For the other four, fine structure went head to head with energy at around 30–40% weight assigned to each of them.

Half of the listeners, S1, S2 and S4, also relied on envelope cues (with a set weight of 20%–30% when averaged across signal levels), while for the remaining subjects this cue was negligible (12% or below) compared to the other two stimulus predictor sets (30%–50%). The influence of previous responses showed a similarly diverse picture: While three observers, S3, S4 and S6, showed a clear interaction of responses with earlier trials (10% or larger averaged across signal level), serial dependencies almost vanished for the others (around 5% or below).

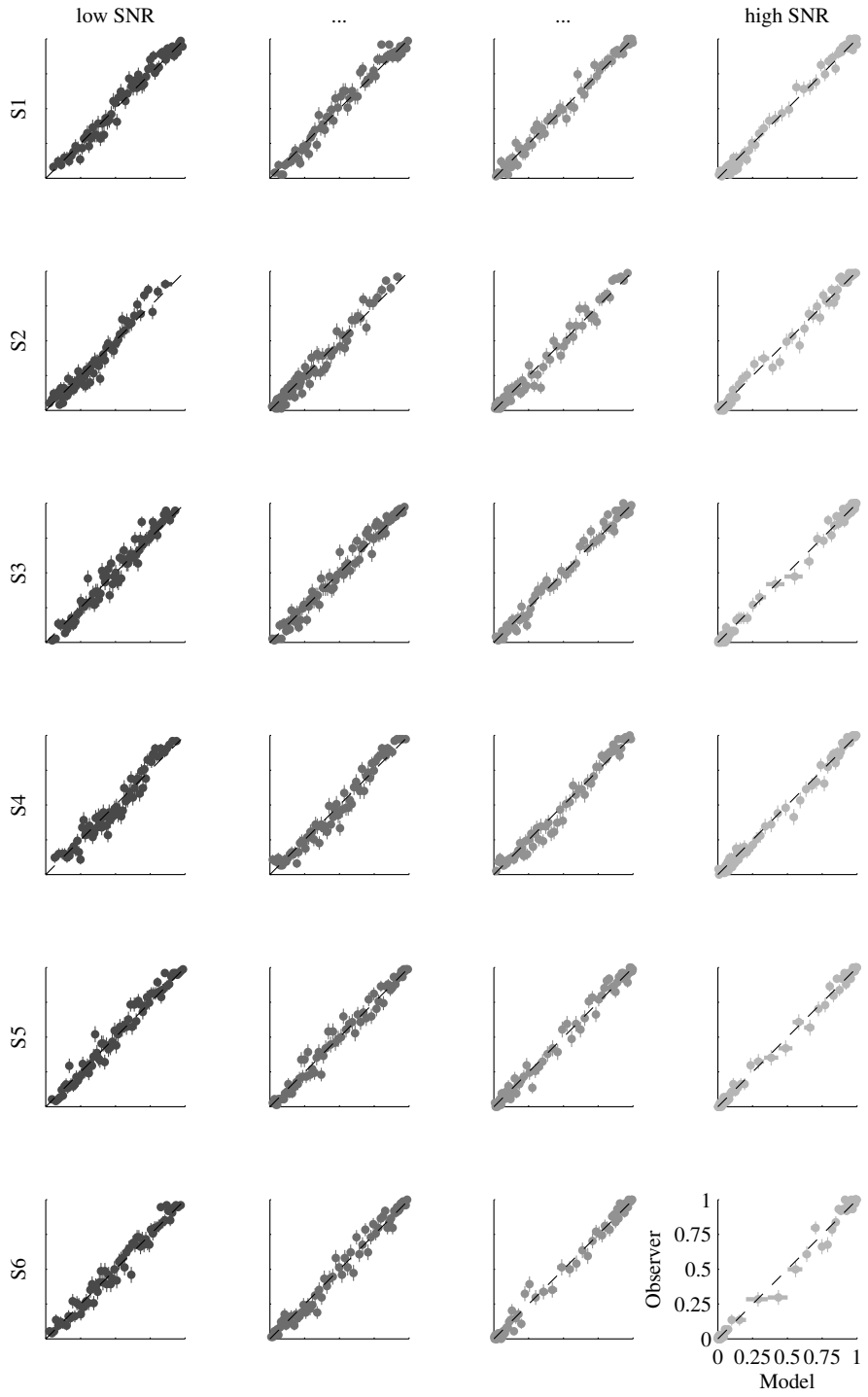


Figure 6.5: Direct comparison of predicted (x-axis) and empirical (y-axis) response probability estimated from binned stimuli (75 trials per bin). Data are shown for each observer (top to bottom) and signal level (increasing from left to right). The deviance value reported in the manuscript is estimated based on these comparisons (after averaging the results from multiple bin sizes). Most data points are close to the diagonal which represents the ideal result. Accordingly, the deviance computed from all of these plots is close to 1.

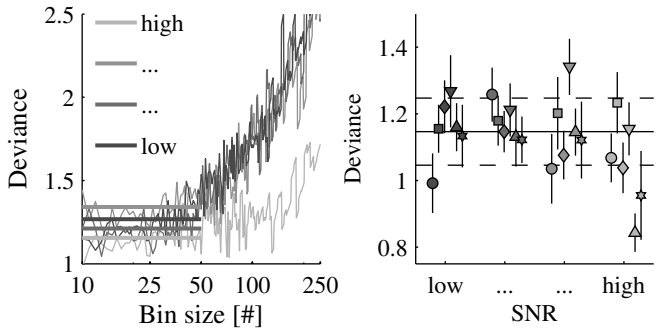


Figure 6.6: Model deviance. **Left:** The computed deviance between empirical and predicted response probability for binned stimuli is shown for one example data set (observer S4 at multiple signal levels represented by gray level). The estimated deviance varied depending on the bin size (x-axis) used for stimulus pooling. Deviance averages and standard errors shown on the right were estimated from bin sizes 10–50 as indicated by the straight lines. **Right:** Average deviance estimates across all observers (symbols as in Fig. 6.2) and signal levels range from 0.8 to 1.4. In addition, average deviance across all observers and signal levels is represented by a straight line, the corresponding standard error by dashed lines.

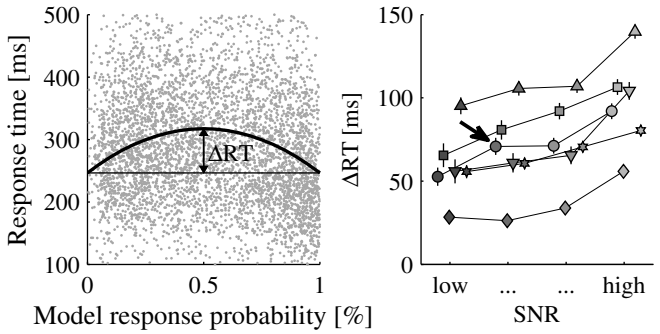


Figure 6.7: Response times to individual stimuli depending on predicted response probability. **Left:** Sample data (≈5000 trials from observer S1 at E/N_0 = 9 dB) showing measured response times to individual stimuli against the response probability as predicted by the observer model. The curve represents a quadratic regression fit to the data points, the indicated distance ΔRT corresponds to the estimated difference in response time for stimuli with 0.5 and 0/1 predicted response probability. **Right:** Estimated difference in response time ΔRT (between stimuli with Yes-response probabilities of 0.5 and 1/0) for all observers across signal levels. The data for the sample shown on the left is marked with an arrow.

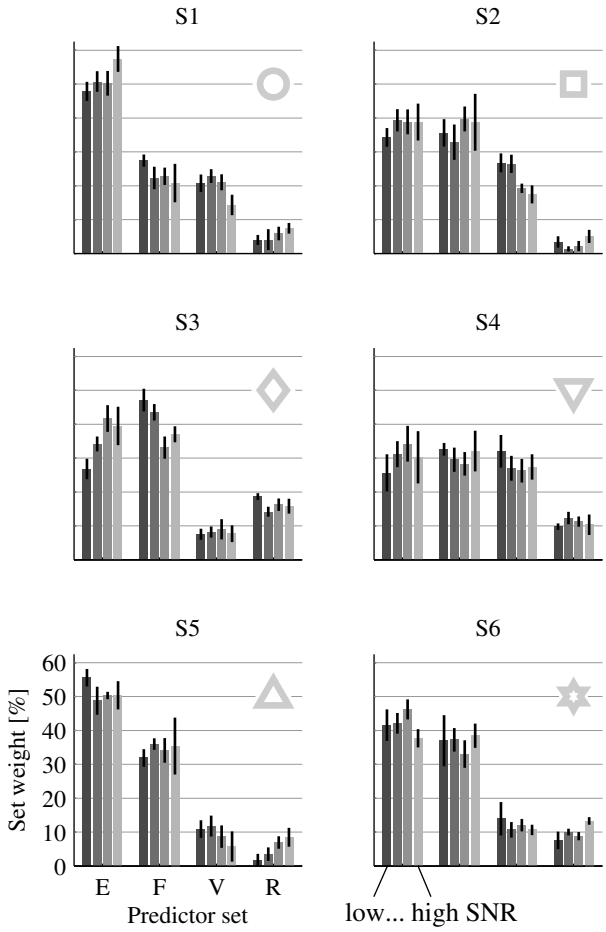


Figure 6.8: Set weights for all observers. Relative weights for the four predictor sets (E)nergy, (F)ine structure, en(V)elope and previous (R)esponses for all observers across signal levels (represented by gray level). The large gray symbols represent the observers as in Fig. 6.2.

Observer		**Energy dominant**	**Symmetr. Filters**	**Envelope**	**Sequential Depend.**
S1	○	yes (no*)	no	yes	no
S2	□	no	yes	yes	no
S3	◇	no	no	no	yes
S4	▽	no	no	yes	yes
S5	△	yes	yes	no	no
S6	✡	no	no	no (yes*)	yes

Table 6.1: Summary of observer decision strategies, * for more information see section 6.4 “Additional Analysis and Results”.

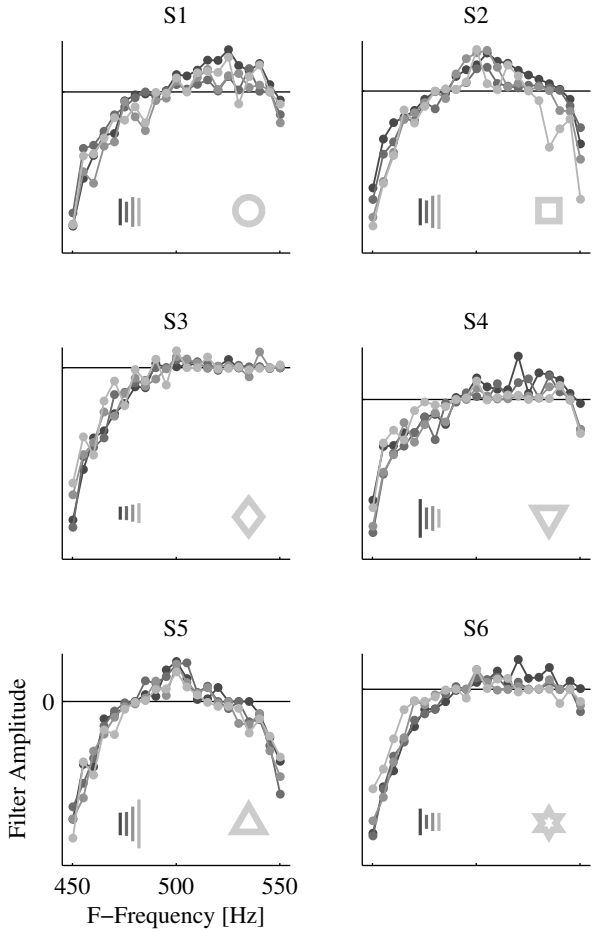


Figure 6.9: Fine structure filters. Filter amplitudes of the observer models in the fine structure domain across signal levels (gray level as in Fig. 6.8). The inset parallel lines (bottom left) depict standard errors of the estimated filters averaged over each signal level condition.

Perhaps surprisingly, the results were stable across all observers and signal levels—the decision strategy of the observers changed only marginally as a function of task difficulty (for individual observers standard error amounted to 2–3.5 percent points across signal levels when averaged over predictor sets). Thus, each observer was unique in the cues she used to perform the TiN-task even for this nominally "easy" low-level task of TiN detection, but the strategy itself did not depend on task difficulty.

6.2.4 Spectral and Behavioral Weights

Individual spectral filters in the fine structure and envelope domain were reconstructed from the individual model weights and are shown in Figs. 6.9 and 6.10. For four observers, S1, S3, S4 and S6, the filters in the fine structure domain were strongly asymmetric, according to visual inspection. The peak frequency was typically centered above the signal tone with a negative lobe below and a positive or neutral filter weight above. Data from observers S2 and S5 resulted in largely symmetrical filters centered on the signal frequency.

For the three observers for which the envelope achieved a strong set weight, S1, S2 and

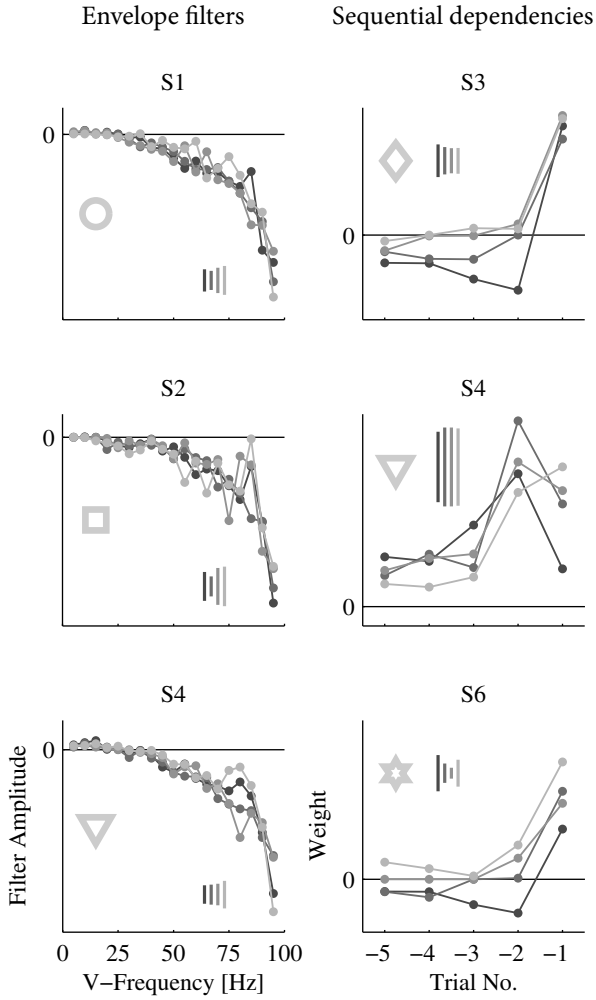


Figure 6.10: Envelope filters and weights on previous responses. **Left:** Filter amplitudes of the observer models in the envelope domain across signal levels for observers considered relying on an envelope cue. **Right:** Weights on individual preceding trials for observers considered depending on previous responses. Gray level indicates signal level as in Fig. 6.8, inset parallel lines depict standard errors as in Fig. 6.9.

S4, the corresponding filters in the envelope domain have a similar shape (see Fig. 6.10, left): At low frequencies, the filter's amplitude vanishes, but with rising frequency it exhibits an increasingly negative weight.

For the three listeners S3, S4 and S6 for which substantial serial dependencies in responses were observed, Fig. 6.10 (right) shows the precise interaction pattern in terms of the weights on earlier trials. For subjects S3 and S6, the response during the immediately preceding trial exerted a dominant positive influence on the current trial, i.e., they tended to press the same button in consecutive trials. In the lowest SNR-level, the most difficult condition, they also preferred to choose the opposite response to the second-to-last and earlier trials. In contrast, for observer S4, a dominant positive influence appeared for the second-to-last trial. For observer S3, previous responses had the strongest influence. As an extreme example, fast responses (i.e., trials with response times shorter than the median for each signal level) could be predicted from earlier decisions alone with 58–60%-accuracy—an almost 10% improvement from chance even when completely ignoring stimulus information.

6.2.5 Additional Confirmation of Consistency

A model may be a good predictor of the listener's behavior, even though it does not capture anything about the internal decision strategy of the observer. For high signal-to-noise ratios, where observers respond correctly in a large proportion of trials, a model might have been trained to merely discriminate between the true stimulus classes (signal and no-signal). In order to rule out this possibility, the analysis was repeated based on no-signal and signal trials separately.[7] In this case, since only one class of stimuli is used for training, there is no way the model can pick up ground-truth differences between no-signal and signal stimuli.

[7] Ahumada and Lovell (1971) performed the same kind of analysis, though with little success: *"As a further check on the validity of [the multiple regression procedure] for predicting response totals, regression analyses were done separately for the SN stimuli and the N stimuli. [...] These estimates [...] suggest that most observers are not adequately described by the model."*

Predictions from these data subsets were almost as accurate even though only half the number of trials entered each fit (model agreement dropped by 2.5 ± 2.2 and 2.3 ± 2.5 percent points on average when separately fitting noise and signal data, respectively). The relative weights on predictor sets were not significantly affected either with mean differences across all conditions and observers amounting to 4.5 ± 6.4 and 3.6 ± 5.6 percent points for noise and signal stimuli, respectively. These results validate and confirm the earlier model fits and corresponding conclusions on the distribution of relative weights.

To verify that the extracted set weights do indeed relate to the observer's use of the corresponding features, one of the predictor sets was separately removed during the analysis to see how model predictions were affected. As presented in Fig. 6.11, the improvement of model likelihood when adding a particular predictor (envelope or sequential dependencies) is clearly related to the corresponding set weight: The larger the set weight, the more the likelihood increased when that predictor is included compared to being excluded. This result also validates the earlier classification of observers into those that depended on, or ignored stimulus envelope and earlier responses.

We assumed that differences in the trained models correspond to variations in individual observer decision mechanisms and are not merely the consequence of noise in the data. The fact that the weight distribution varies much more strongly across observers than across SNR-conditions lends support to this claim. Stimulus generation and data analysis for each observer as well as each condition were performed completely independently. It is highly unlikely that noise in the data by chance resulted in inter-individual, but not intra-individual differences. On the contrary, similarities should rather be expected for the same SNR-condition across listeners (i.e., opposite as to what was observed), since the overall stimulus variability was determined by the signal level.

As a further confirmation that inter-individual differences in terms of perceptual cues resulted not merely from noise but reflect actual differences in observer behavior, decisions of one observer were predicted with models trained to another observer's

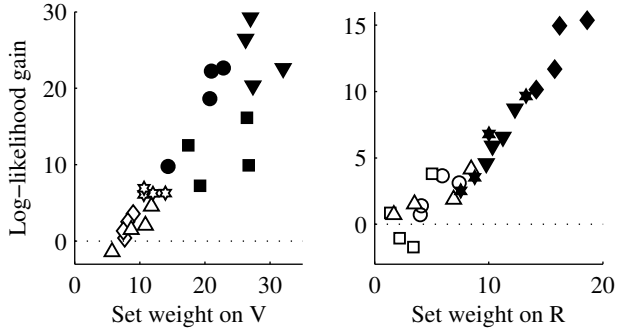


Figure 6.11: Likelihood gain and relative set weight. Improvement in model log-likelihood by inclusion of the envelope feature (V, left) or previous responses (R, right) during model fitting plotted against the corresponding set weights in the full model (data plotted for all SNR condition, symbols represent observers as in Fig. 6.2). Filled symbols designate observers that were considered relying on the envelope feature or previous responses, respectively. The axes are differently: Compared to previous responses, stimulus envelope was on average about twice as important to predicting observer responses, both in terms of set weight and log-likelihood gain.

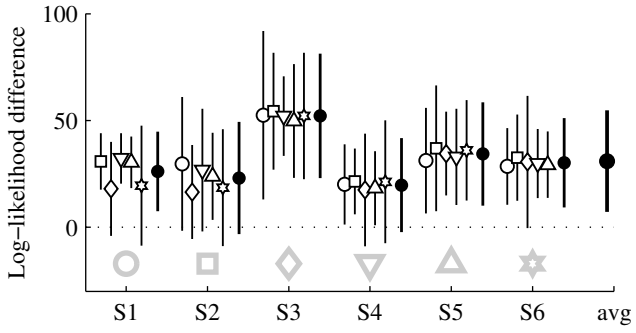


Figure 6.12: Prediction improvement in terms of log-likelihood when training and testing models with data from the same vs. a different observer. The big gray symbol and corresponding label on the x-axis represent the observer whose decisions were predicted. The black open symbols depict the listener with whose data a model was fit. The y-axis represents the improvement in log-likelihood when the observer's behavior was predicted from its own data instead. The black marker next to each group represents observer averages, the bold marker on the far right indicates the grand average over all observers.

data. If variations in the weights were mainly determined by noise and observers were behaving largely identically, then each listener should predict the behavior of the other observers as good as her own decisions. This analysis was confined to conditions with the lowest signal levels ($E/N_0 = 7$ and 9 dB), because this is where individual differences are most prominent. At higher signal levels, differences in decision strategy are harder to distinguish behaviorally because the majority of meaningful strategies leads to the correct and thus identical response. The results of the present analysis can nevertheless be transferred to high SNR conditions since the weights for all observers varied little across signal levels.

In addition, I accounted for the fact that individual differences in response bias were observed. This may lead to a deterioration in predictive power in cases where the strategy of two observers is perfectly identical in terms of the weights w_j, and merely differs in terms of the response bias (b in Eq. 6.1). In this analysis, observer behavior was therefore predicted with the model *weights* w_j from *different* observers while the model *bias* b was fixed to the value obtained from the *same* observer. As a consequence, degradations in predictive power represent a true measure for individual differences in terms of predictor weights or, equivalently, the perceptual decision rule.

The extent to which the predictive power suffers when a listener's behavior is explained with a model trained to a different observer is displayed in Fig. 6.12. Model predictions measured in terms of log-likelihood are deteriorated across all listeners although for two observers the difference did not reach significance. I conclude that variations in model parameters resulted from individual behavioral differences, not just from random noise fluctuations.

6.3 Preliminary Discussion

Based on linear models fit with a sparse regression procedure, the behavior of six observers in a TiN detection task was predicted with high accuracy. Using the empirically estimated observer consistency, I confirmed that agreement between observer and model was close to the theoretical upper bound. This analysis also illustrated how the observers' inconsistencies prevent a perfect prediction of responses on a trial-by-trial level. Nevertheless, the observer models were able to identify sets of stimuli to which observers responded with "Yes" or "No" with large confidence or where listeners appear to be very uncertain. Using response time data collected during the experiment, this claim was independently validated. Thus, the observer models are truly "molecular" in the sense of David Green. Having shown that little room is left for improving predictions with any kind of observer model (Fig. 6.3, right), I conclude that relying on a simple linear model (in combination with nonlinear stimulus features) for explaining observer behavior does not seem to represent a strong restriction.

Even though the listeners could hardly be discriminated according to their psychometric performance (Fig. 6.2), they were using very different decision strategies as quantified by the extracted weighting of predictors (Figs. 6.8, 6.9, 6.10) and confirmed by the drop in prediction performance when predicting one observer's behavior from the decisions of another (Fig. 6.12). These idiosyncrasies in cue distributions, which are detailed in Table 6.1, have already been observed in earlier studies (Richards *et al.*, 1991; Richards and Nekrich, 1993). In the current analysis, stimulus energy was partially predictive for the behavior of all observers, but with two out of six observers, only a minority relied predominantly on this cue. While corroborating that sound energy is an important factor in TiN detection, I also confirmed that it is not the only cue listeners rely on in agreement with earlier experiments employing level variation (Kidd *et al.*, 1989). Instead, observers seem to be relying on a combination of multiple characteristics of TiN stimuli comprising energy, fine structure as well as envelope.

In addition to the energy cue, behavior was best explained with asymmetrical spectral filters exhibiting strong negative weights on the lower half of the noise spectrum for a majority of four listeners. For the remaining two observers, the analysis predicted a symmetrical filter centered on the signal. The behavior of half of the observers also depended on the envelope characteristics of the stimuli, with increasingly negative filters for high modulation frequencies and little to no weight attributed to frequencies below the central 50 Hz. The interpretation of this result is discussed in more detail in the following section.

Responses for three out of six observers did not only depend on the current stimulus, but also on decisions in immediately preceding trials. The analysis of the associated predictor weights uncovered very diverse interaction patterns where the last or second to last response had the strongest influence on the current decision, but dependencies were observed for up to five preceding trials. Thus, it may generally be beneficial to take sequential dependencies into account when explaining behavioral data on a trial-by-trial level, though their influence appears to be less strong in humans than in behaving animals (Busse *et al.*, 2011) and shows pronounced individual differences.

Our "best" observer in terms of psychometric thresholds, S5, did not show any sequential dependencies (which always degrade performance) and instead relied dominantly on energy. In addition, his behavior is best predicted with a symmetrical spectral filter centered on the signal, i.e., closer to an "optimal" observer (Peterson and Birdsall, 1953). Interactions with envelope characteristics were small compared to other stimulus cues—thus, this observer may be regarded as resembling the archetypical energy observer as initially postulated by Green and Swets (1966).

To validate the consistency of the present results, I demonstrated that differences between observer models in terms of predictor weights are not merely a result of noise in the data. Otherwise, the prediction rate should not have declined after a model fit to data from one observer was used to predict the behavior of another. In consequence, the idiosyncrasies in terms of decision mechanisms must have been real, not just an effect of noise. Conversely, since I have shown that the procedure was principally able to discriminate different strategies—namely across observers—the stability of individual strategies across signal levels must have been real as well and could not have been merely an artifact of the analysis method (e.g., by relying on a too constrained model).

The smooth spectral filters in the frequency domain that were estimated for the listeners can be construed as gammatone-like auditory filters or combinations thereof. The interpretation for the weights in the envelope domain is however not as clear-cut. Observers do not seem to rely on the most predictive portions of the envelope spectrum as proposed by Green *et al.* (1992), otherwise a significant proportion of the weights should be concentrated near the central component of the envelope spectrum (above and below 50 Hz), and not at the high frequency components as observed.

The above-mentioned interpretation of *"increasingly negative filters for high envelope frequencies"* is probably not the most meaningful description of the observers' reliance on the stimulus envelope. A different representation of the envelope characteristics in the observer model, i.e., a different set of predictors characterizing the envelope may result in a clearer picture regarding the listeners decision strategy. Therefore, the following section extends the analysis to include an earlier proposed set of six envelope descriptors that listeners may employ in TiN detection tasks (Richards, 1992).

When expanding the set of predictors in this fashion, I rely on a critical advantage of the newly proposed sparse weight estimation procedure (Schönfelder and Wichmann, 2012): Even for data sets of moderate size, large sets of predictors can be investigated even though they may overlap in terms of the stimulus information they contain. The L_1-regularizer then identifies the smallest set of predictors that is both necessary and sufficient for explaining behavior, while suppressing the weights on the remaining predictors. Because the method is flexible in this particular regard, two additional descriptors

of the fine structure were included to see whether they would obtain a significant relative weight in addition or at the expense of the fine structure predictors employed so far.

6.4 Additional Analysis and Results

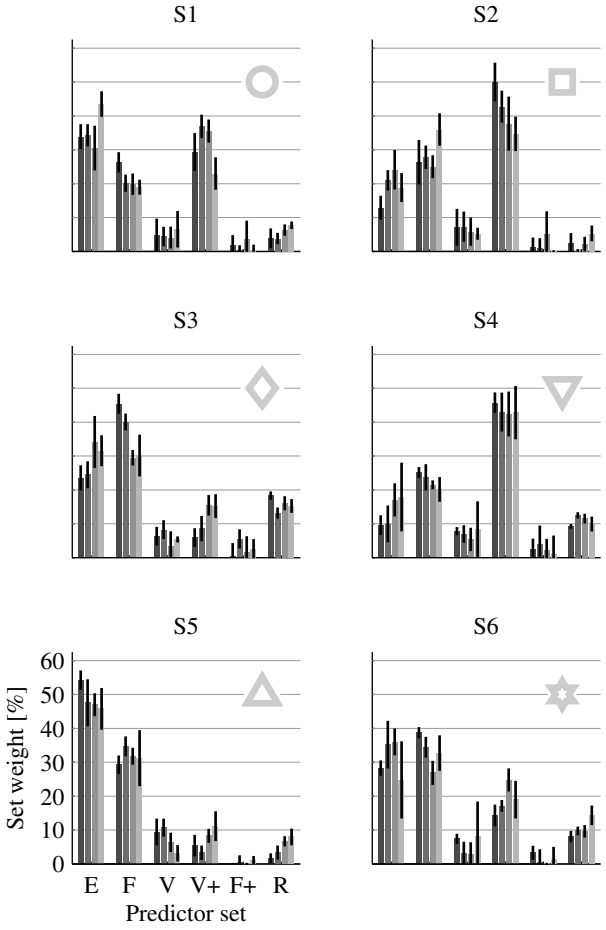


Figure 6.13: Set weights for all observers when fitting the data with the fully extended set of predictors (including alternative envelope [V+] and fine structure [F+] predictors). Otherwise same as Fig. 6.8. The set weights on the envelope power spectrum (V) as well as the alternative fine structure predictors (F+) nearly vanish, while the alternative envelope predictors (V+) gain a large proportion of the overall model weights for observers S1, S2, S4, and S6.

The additional analysis aimed at identifying a simpler and possibly more meaningful description of the observers reliance on the stimulus envelope by using a different set of predictors to represent envelope characteristics.

The model fitting procedure was repeated in the same manner as for the earlier analyses, relying on the previously proposed set of predictors (energy, fine structure spectrum, envelope spectrum, previous responses) and—*in addition*—a set of six "classical" envelope predictors that had been collected by Richards (1992): envelope variance, maximum divided by minimum, envelope crest factor [peak divided by average] and kurtosis as well as the overall number of extrema and average envelope slope. All of these measures

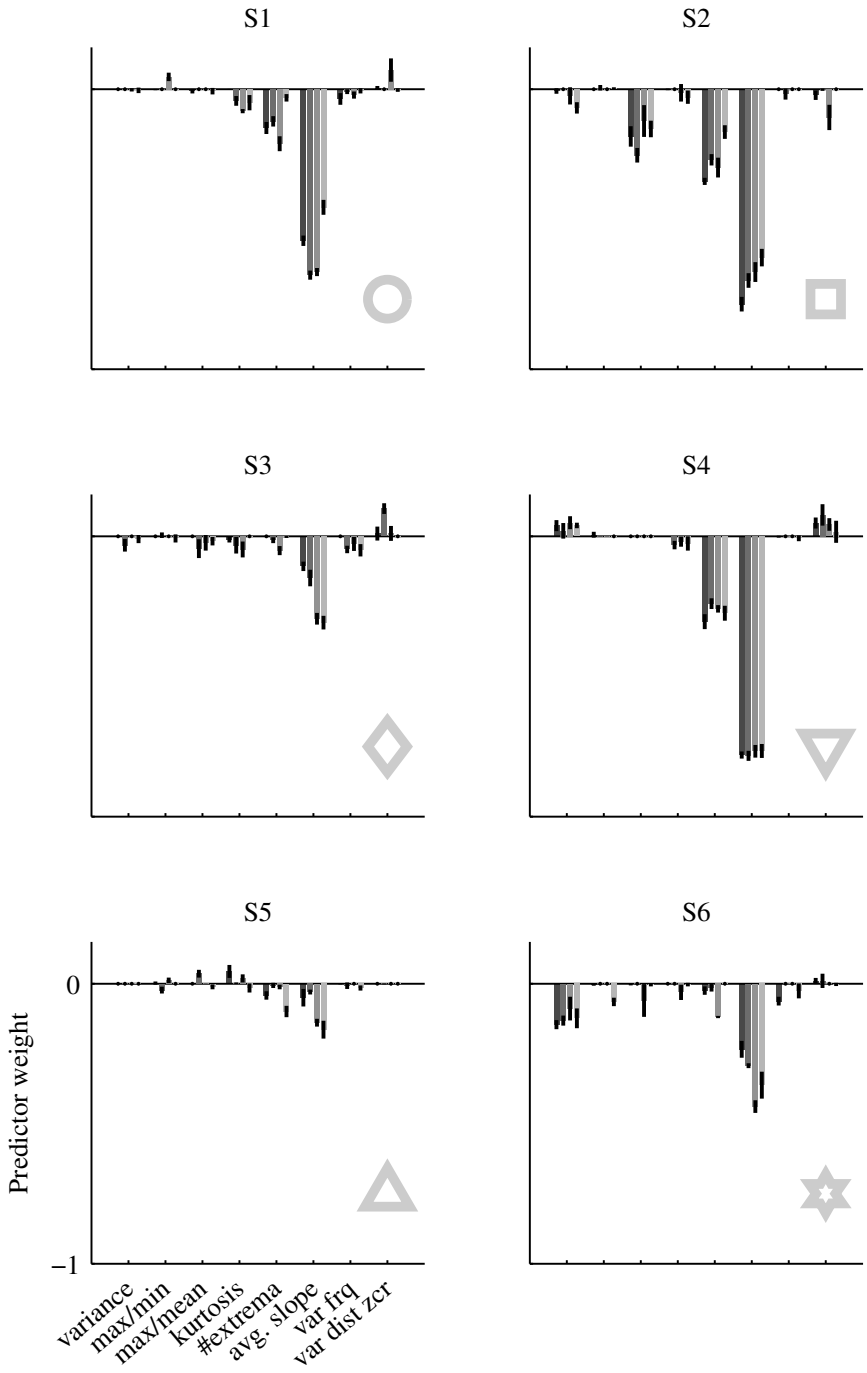


Figure 6.14: Model weights associated with individual predictors for alternative envelope and fine structure descriptors ([envelope] variance, maximum/minimum, maximum/mean, kurtosis, extrema count, average slope; [fine structure] variance in instantaneous frequency, variance in zero crossing distances). For most observers, almost all of the alternative weights are concentrated on envelope extrema count and average slope.

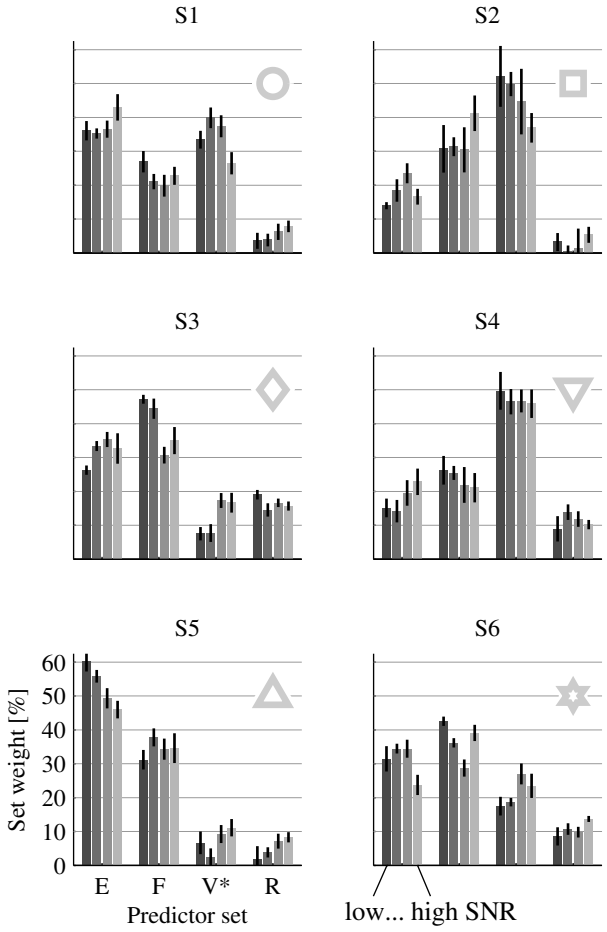


Figure 6.15: Set weights for all observers using alternative envelope predictors for model fitting. Relative weights for the three predictor sets (E)nergy, (F)ine structure, en(V*)elope ("alternative" predictors average slope and extrema count) and previous (R)esponses for all observers across signal levels (represented by gray levels).

were determined after normalizing the envelope to make sure that overall differences in sound level did not influence these variables. Richards (1992) had also suggested two fine structure predictors, which were included as well (variance in the temporal distance of zero-crossings, variance in instantaneous frequency), totalling a number of 54 model predictors.

As a result of using this supplemented set of stimulus predictors, a substantial shift of the weight distribution was observed for those listeners that were previously identified as relying on the envelope (S1, S2, S4): As displayed in Figs. 6.13 and 6.14, almost all of the weights on the envelope spectrum were transferred to the predictors associated with envelope slope and extrema count (the weight on the envelope spectrum set fell by -17.3 ± 1.3 percent points to $6.2 \pm 2.1\%$ averaged over all observers and conditions). For the remaining three observers, the weight on the envelope spectrum had already been comparatively small and did not change much ($-3.5 \pm 1.2\%$). As regards the fine structure, no such shift of weights occurred: The weight for the two predictors associated with variance in instantaneous frequency and zero-crossing distances was small and merely achieved significance ($1.7 \pm 1.6\%$ on average across observers).

Realizing that the newly obtained observer models did assign only small weights

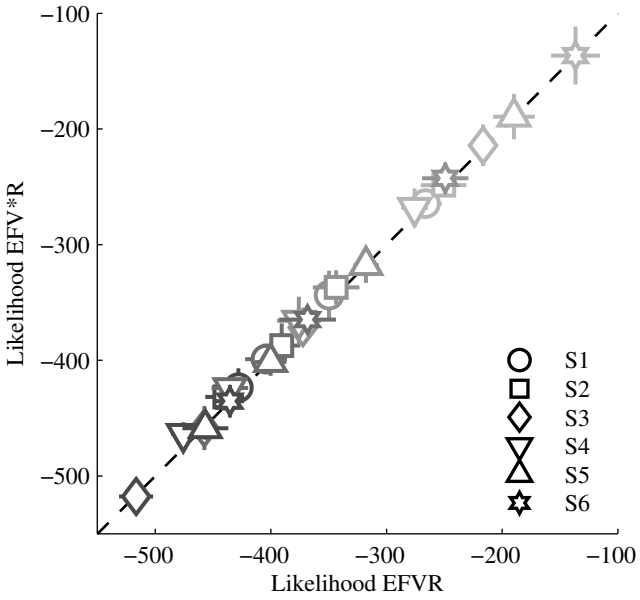


Figure 6.16: Comparison of model predictive power (in terms of model likelihood) when training with envelope power spectrum (V) or envelope extrema count and average slope (V*). Gray level depicts signal level as in previous figures.

to the envelope spectrum predictors as well as the classical fine structure predictors, the analysis was repeated again, this time entirely excluding these predictors. When comparing the predictive power of the models that rely either on the resulting reduced set of predictors (energy, fine structure spectrum, envelope slope and extrema count, previous responses, for a total of 29 predictors) or the earlier employed set of predictors (envelope power spectrum instead of slope and extrema count, resulting in 46 predictors), the previous set provided as good an explanation for observer behavior as the new much smaller set, as shown in Fig. 6.16 (the average likelihood difference across observers and signal levels amounted to only 3.4 ± 22.6). For every single listener and at all signal levels, the two alternative envelope predictors were able to capture the observers' dependence on the stimulus envelope just as well as a complete description of the envelope spectrum.

The set weights that resulted from model fitting with the reduced set of predictors are presented in Fig. 6.15, where the new set of envelope predictors V* consists of the average slope and the extrema count. In comparison to the previous results presented in Fig. 6.8, part of the weight on energy is transferred to the two alternative envelope predictors, so that envelope is now the most dominant cue for listeners S2 (46% for envelope against 18% for energy on average across signal levels) and S4 (47% against 18%) while being roughly in balance with energy for S1 (34% against 37%). Listener S6 that was formerly classified as not relying on the envelope, must now be considered envelope-dependent, but still energy-dominated (22% for envelope, 31% for energy). For these four observers, both the envelope slope and extrema count were weighted negatively, with the slope obtaining on average a five times (5.2 ± 3.8) larger weight. That means that mainly with an increase in average slope (and to a smaller extent in the number of extrema), observers responded "No" (perceived "noise" stimulus) with higher probability. This corresponds to a sensible decision strategy since average envelope slope and extrema count both grow at reduced signal levels.

For the observers S3 and S5, the new set of the envelope predictors slightly altered the relative weight on the energy (-3.6 ± 1.6 and $+1.7 \pm 1.5$ percent points, respectively)

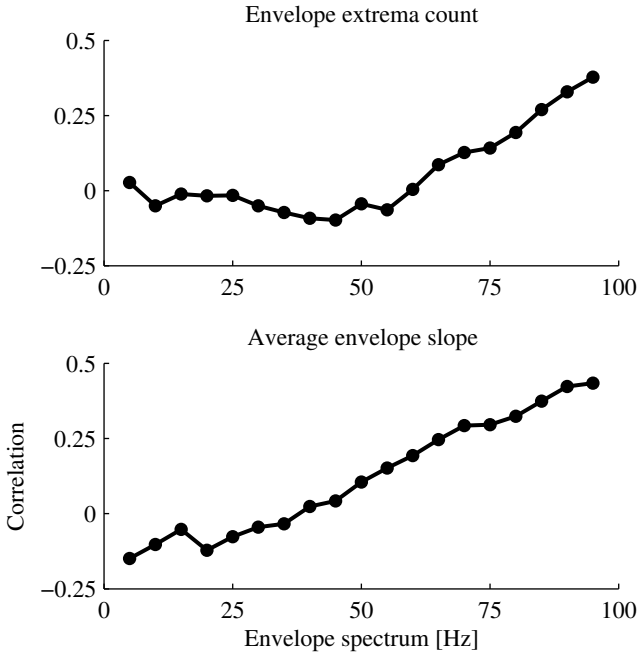


Figure 6.17: Linear correlation between components of the envelope power spectrum (x-axis) with envelope extrema count (top) and envelope average slope (bottom). The estimates were obtained with TiN maskers as used in the study (450-550 Hz wide, 200 ms long with 50 ms on/off-ramps).

and envelope ($+4.2 \pm 1.2$ and -1.9 ± 1.5 percent points) predictors. However, the earlier assessment still holds: With a relatively small set weight (12% and 7%, respectively), envelope characteristics of the stimulus only played a minor role compared to the other stimulus predictors (together about 71% and 87%, respectively). The relative set weight for the fine structure spectrum changed for two observers (S2: -3.3 ± 2.8% and S4: -6.9 ± 2.0%), while there was no significant difference for the other listeners. Differences in the set weight on earlier responses were well below 1% and insignificant for all observers (0.2 ± 1.1% on average). The weight distribution on individual previous responses was not effected, nor was the overall shape of the spectral filters.

6.5 Final Discussion

An additional analysis indicated that a large set of predictors (19 values defining the envelope power spectrum) could be replaced with only two simple envelope descriptors (average slope and extrema count) without reducing the predictive power of the individual observer models. Assuming that a model that relies on fewer predictors can be considered to be better, I must conclude that a model relying on envelope slope and extrema count instead of the earlier proposed envelope spectrum provides an even better description for the behavior of the observers. Nevertheless, both approaches provide a valid representation of observer behavior. In fact, they are directly related: Strong high frequency components in the envelope coincide with faster fluctuations which give rise to larger average slope values as well as an increased number of envelope extrema. In quantitative terms, for the presently used noise stimuli the linear correlation between components of the envelope spectrum and the average slope as well as extrema count

increases roughly linearly with envelope frequency (Fig. 6.17). Thus, for a listener that relied strictly on envelope slope or the number of extrema, the empirically observed envelope filters can be qualitatively expected. However, the increase in relative weight on the envelope predictors for some observers at the expense of energy or the fine structure suggests that the two alternative envelope predictors contain even more information related to the observer decision mechanism than the envelope spectrum.

Based on the modified set of predictors, the earlier classification of observer decision strategy in Table 6.1 had to be revised in two respects: One observer (S1) could not anymore be classified as energy-dominated, while a dependence on envelope characteristics was found for one additional observer (S6).

6.6 Summary and Conclusion

Using a large human behavioral data set collected in TiN detection experiments, this study applied a modern statistical analysis procedure to explain and predict trial-by-trial responses of individual listeners.

No simple answer was found as to which perceptual strategy observers rely on in a narrow-band TiN detection task—it appears there is no unique simple auditory feature that governs TiN detection. Instead, the importance attributed to particular auditory features was highly observer specific even for such a (nominally) simple psychophysical task. The only commonality across listeners was the use of multiple sound properties. Responses of all observers depended at varying proportions on sound energy and symmetrical or asymmetrical spectral detectors. The distribution of spectral weights can be interpreted as a multiple detector model, with one detector centered or slightly above signal frequency, and negatively weighted side-band filters above and below in order to compare information in different spectral bands (Gilkey and Robinson, 1986). An asymmetry in spectral weights has been observed before, but was characterized with much lower spectral resolution (Richards and Buss, 1996, their Figs. 1, 2).

The decisions of half of the listeners were also determined by envelope characteristics, although the estimated perceptual weights concentrated outside of the most informative regions of the envelope spectrum—assuming they relied on the envelope spectrum (Green *et al.*, 1992; Rosas and Wichmann, 2011), the listeners were thus not behaving "optimally". Instead, two simple envelope predictors, average slope and extrema count, were found to provide at least as good an explanation for observer decisions as a representation of the full envelope spectrum consisting of 19 predictors. Both findings do not support a trial-by-trial decision strategy that relies on the rich information contained in the envelope spectrum, even though corresponding models may successfully predict psychophysical TiN detection thresholds (Dau *et al.*, 1996b).

The strong idiosyncrasies that were observed emphasize the necessity for explaining behavior on an individual level instead of fitting models to data from multiple observers simultaneously—at least for the present data set. However, when large inter-individual differences appear for such a basic discrimination task, they are likely to also emerge in similar or more complex auditory tests.

Overall, the current study corroborates and consolidates previous findings on narrow-band TiN detection: Sound energy is an important determinant of observer decisions if present, though not necessarily dominant. In addition, all observers were found to rely on auditory filters which are presumably based on critical-band detectors. In addition, the behavior of some listeners was influenced by sound envelope. This dependence can be characterized either by a non-optimal envelope filter or, much simpler, by two scalar envelope descriptors.

The present study extended the perceptual weighting analysis used in earlier research to simultaneously quantify the influence of stimulus factors and the strength of sequential

dependencies in response behavior. It was shown that perceptual decisions in a fast paced auditory detection experiment are not purely stimulus-driven—even for highly-trained observers after thousands of trials. At least for some observers, responses in several immediately preceding trials had a substantial influence on the current decision, collectively gaining relative weights of up to 20%. The applied technique successfully incorporated Greens idea of not only taking external stimulus factors into account when explaining data on a "molecular" level but also internal factors in terms of the sequential effects of past responses.

Despite substantial differences in individual decision strategies, all observer models accurately predicted trial-by-trial responses, even though they were based on a simple linear weighting scheme (combined with nonlinearly transformed stimulus features). In terms of improving the agreement between predicted and empirically observed responses, I conclude that there is little room left for alternative models, e.g., relying on different sound features or more complex nonlinear decision mechanisms.

Finally, the additional investigation presented in section 6.4 that extended the analysis to a large and partially redundant set of stimulus predictors demonstrated the flexibility of the method proposed by Schönfelder and Wichmann (2012) for exploratory analysis of auditory psychophysical data. However, care should be taken when subjecting the same behavioral data to multiple analyses.[8] The supplementary investigation also emphasized the point that modeling results critically depend on the set of stimulus predictors that are used during data fitting—the relative weights must always be seen and interpreted in the light of the currently used set of model predictors. The more comprehensive the set of stimulus (and behavioral) predictors, the more likely an experimenter can infer an unbiased estimate of the perceptual decision scheme that underlies the observers' behavior. The sparse regularized regression procedure is particularly appropriate to such an investigation as it provides a way of probing large sets of predictors simultaneously.

The analysis method may be applied using other sets of features spanning different explanatory domains. Here, a set of abstract mathematical descriptors of the stimulus was chosen integrating information over the entire stimulus length. Alternatively, using spectrograms instead of Fourier transforms, additional information on a temporal scale may be useable similar to the method used by Shub and Richards (2009) to estimate psychophysical spectro-temporal weights. Such an approach may uncover strategies that rely on features that are confined to the beginning or end of the stimuli or that employ short-term cues as suggested by Davidson *et al.* (2009a). Correlations between predictors that result from overlapping spectro-temporal windows should not pose a problem for a sparse regularized regression. Furthermore, physiologically motivated descriptions of behavioral data from TiN detection could be obtained by relying on models of the auditory periphery that estimate spike trains in auditory nerve fibers (Heinz *et al.*, 2001; Carney *et al.*, 2002). The set of predictors could in this case be composed of a number of spike train metrics. Finally, following the observation that observer decisions depend on *responses* in preceding trials, descriptors of preceding *stimuli* may be included as model predictors to determine whether such dependencies (auditory "context effects") are present as well.[9]

All of the above suggestions have in common that the corresponding analysis could be based on data that has already been collected as long as single-trial stimulus waveforms and responses have been recorded.

[8] Subjecting a data set to multiple tests might lead to results that are spuriously significant, while just being the result of an uncontrolled "fishing expedition" until "something interesting" is found. The significance tests or confidence intervals need to be interpreted and corrected appropriately (Wagenmakers *et al.*, 2011).

[9] This proposition is closely related to the idea of dynamically changing templates that depend on a small number of previously presented stimuli (Davidson *et al.*, 2009a). An analogous effect was recently observed in visual psychophysics (Chopin and Mamassian, 2012).

Chapter 7

Overall Conclusion and Outlook

The present doctoral dissertation conjointly developed two strands: a methodological one that deals with the problem of extracting perceptual cues from behavioral data and an empirical one which pertains to the auditory paradigm of Tone-in-Noise (TiN) detection. In chapters 2 and 3 I laid out in depth the research history related to both aspects which provides the background and foundation of the present research. In this final chapter I again—but with a little more distance—review and discuss the methodological insights and empirical findings that I obtained in the process of preparing this dissertation. In a final outlook, I collect some ideas relating to possible applications of the proposed method and mention further research that the present work might inspire.

7.1 Conclusions

As regards the methodology, I discussed in section 2.2 that the classically employed techniques for observer modeling have significant difficulties when they are used to analyze behavioral data based on large sets of potentially interdependent stimulus features. To overcome this methodological limitation, I proposed to apply a recently developed algorithm from machine learning—a sparse L_1-regularized logistic regression—in combination with the classical linear observer model. In section 2.3, I also argued that a "sparseness" constraint represents a critical property of the analysis procedure given the aim of identifying the essential stimulus cues that best capture behavior. In addition, in chapter 2.5, I asserted that a "molecular", i.e., trial-by-trial, analysis of psychophysical data should take into account behavioral factors such as sequential dependencies in observer decisions which appear to contaminate a majority of behavioral data sets.

As regards the empirical aspect of the present work, the experimental paradigm of Tone-in-Noise detection, there still existed a substantial gap concerning the understanding of the perceptual mechanisms that govern listener decision in this seemingly simple task. As discussed in section 3.4, up until now none of the proposed models intended to explain observer behavior was able to explain the data on a trial-by-trial level. It had already become clear, though, that neither a single stimulus cue nor even a single strategy covering all observers exists. While applying a modern analysis procedure to a large psychophysical data set from TiN detection, I intended to simultaneously probe a large set of stimulus features for two purposes: first, to establish observer models that predict listener decisions on a single-trial level, and, second, to identify the perceptual cues that best explain individual behavior.

As a first step, I intended to convincingly demonstrate that the sparsely-regularized regression procedure indeed possessed the promised qualities. Therefore, I tested the method in computer-simulated behavioral experiments concluding that the procedure

worked surprisingly well: The perceptual cues underlying the decisions of several different simulated observers could be precisely identified from a large number of interdependent predictors. At the same time, the prediction of single-trial decisions was nearly optimal. Even in realistically noisy conditions and with limited amounts of data, the method was shown to hold substantial advantages over classical non-regularized weight estimation procedures which failed under these circumstances. This comprehensive test of the method provided the basis for its application to empirical data, which represented the next step in the preparation of the doctoral dissertation: In thoroughly prepared psychophysical experiments, I collected an extensive data set from multiple observers performing the classical TiN detection task. Subsequently the sparse regression procedure was applied to the data in order to train individual observer models and finally extract individual stimulus cues that underly the listeners TiN detection strategy.

As in several earlier studies (Richards *et al.*, 1991; Richards and Nekrich, 1993; Davidson *et al.*, 2006), the obtained decision mechanisms were highly idiosyncratic: Observers appeared to rely on diverse sets of different auditory features, including sound energy, critical filter-like spectral detectors and a small set of envelope descriptors. The spectral filters appeared strongly skewed with negative spectral weights below the signal frequency, which corroborated earlier—but less clear-cut—findings (Richards and Buss, 1996). In an additional data analysis using an extended set of stimulus descriptors, I further investigated the aspect of the perceptual strategy that was based on envelope characteristics. It was found that observers did not make use of the rich information present in the full envelope power spectrum as suggested by Green *et al.* (1992) and Dau *et al.* (1997), but were better captured by much simpler envelope descriptors as proposed by Richards (1992).

The majority of previous studies on the subject deliberately combined multiple signal levels during the data analysis without confirming beforehand whether this was an admissible approach. In the present study, data from different signal levels was separately analyzed with the purpose of identifying potential variations in observer strategy depending on the difficulty of the task. Notably, however, no substantial differences in terms of perceptual strategy were found for any observer. Listeners appeared to follow a remarkably stable decision rule, independent of whether the task was easy or hard. In this case then, combining different signal levels into a single analysis seems to be a viable procedure.

Based on Green (1964), I argued that a truly molecular analysis of psychophysical data needed to allow for both external and internal determinants of observer decisions. Consequently, the analysis of perceptual predictors was extended —in a natural, robust and efficient fashion—to take into account the effects of sequential dependencies (e.g., previous decisions). To my knowledge, this represents the first time that behavioral factors in terms of the sequential dependencies in observer responses are included in an analysis of perceptual cues. It was found that such dependencies were present for a substantial proportion of observers. For these, the precise properties of sequential dependencies were as diverse as the perceptual decision rules. Therefore, I conclude that any trial-by-trial analysis of behavioral data should generally take potential influences of sequential dependencies into account. Using several metrics for evaluating model predictive power, I finally demonstrated that—based on both perceptual and behavioral predictors—the trained linear observer models are indeed able to predict previously unseen data with high precision on a "molecular" level.

In conclusion, I gained new insights by combining classical psychophysical experiments and analysis techniques with recent machine learning tools. The theoretically predicted advantages of sparse regularization discussed in section 2.3 indeed translated into substantial practical benefits—an L_1-regularized multiple logistic regression represents a powerful approach to extract perceptual cues from psychophysical data. When applied to a classical auditory psychophysical paradigm, which is still not fully under-

stood, this procedure enabled the identification of perceptual cues and the estimation of "molecular" observer models that predict the behavior of individual listeners on a trial-by-trial level.

7.2 Outlook

The large data set collected in the completed study provides a rich source for further analysis. For example, the obtained linear decision models could be combined with drift diffusion models (Ratcliff, 1978; Ratcliff and Rouder, 1998) attempting to capture the relation between perceptual evidence and response latency. By further exploiting the rich information available through the recording of individual response times, observer behavior could then be captured in even more detail beyond the question which button they pressed. Alternatively, effects of learning could be studied by separately analyzing early and late experimental sessions. By combining data from multiple signal-to-noise ratios, limitations in the amount of available data could be counterbalanced. Such an analysis could, for example, investigate whether the improvement in performance during the learning phase can be attributed to a more efficient perceptual strategy or a general decrease in decision noise.

As an alternative to the presently used model predictors, which are based on mathematically inspired stimulus features, the behavioral data could be analyzed on different levels using other predictor sets. From the present analysis no inferences can be made on the actual neural mechanisms which implement the perceptual decision rules. Those represent a completely different level of analysis. Some insights regarding the question of neural mechanisms could be obtained, for example, by using dynamic physiological predictors as input to an observer model. For example, realistic models of auditory nerve fibers could be employed to estimate single or several parallel spike trains (Heinz *et al.*, 2001). A number of spike train metrics, such as spike count or synchronicity could then serve as predictors for behavior. A priori, however, the exact details of such an approach need to be well prepared as it is unclear what properties the potential critical features may have in this domain. Until now, the neural code remains a mystery and the selection of sensible spike train metrics is still under debate. For the presently employed stimuli, a whole range of auditory fibers are activated and it is an open question which of them carry the critical information—even binaural interactions in the auditory pathway may prove to be essential.

Davidson *et al.* (2009a) suggested that observers in a TiN detection task may rely on short-term cues. Relying on sound spectrograms as model predictors may be a valid approach to implement such a decision mechanism. Usually, neighboring sectors in a spectrogram exhibit correlation because both temporal and spectral analysis windows are overlapping. Using a sparse modeling procedure may help to sharpen the spectral-temporal weights which might otherwise be blurred due to these interdependencies. However, one potential difficulty with this approach is that the temporal location of the cue might change from trial to trial. Thus, the corresponding weights might "wash-out" when simultaneously analyzing the full data set. As mentioned by Davidson *et al.* (2009a), it is also not clear yet on which temporal time scale cues can be expected, as temporal integration time estimates vary from ten to hundreds of milliseconds.

Several aspects of the presently employed experimental design could be improved upon in future research projects. The experiments of the study at hand were set up to collect a single binary response to each auditory stimulus. There was a good reason for this: Originally, I intended to analyze this single-pass data with a support vector machine, which strictly focuses on binary classification. Only later, I decided to employ a sparse regularized logistic regression (for which efficient implementations had just been developed (Lee *et al.*, 2006; Park and Hastie, 2007)). By design, this procedure accurately

models response *probabilities*.

In contrast to the original intention and by a fortunate stroke of serendipity, I had presented a subset of stimuli for multiple times. The corresponding data were used to estimate observer consistency. In a future study, a controlled subset of stimuli in each session could be presented for multiple times. This would allow the detection of changes in observer consistency measures across the progress of the experiment which could then be related to changes in model predictive power.

One could even go one step further and make use of the central benefit of logistic regression: Instead of relying on single-pass stimulus-response data, all of the stimuli could be presented multiple times in a "frozen-noise" procedure, as has been done in earlier studies (Evilsizer *et al.*, 2002; Davidson *et al.*, 2009b; Macke and Wichmann, 2010). For each stimulus a precise response probability could then be estimated, which would be directly exploited while fitting a probabilistic model using logistic regression. This would obviate the need to relate model-observer agreement scores with observer consistency. Instead, a deviance score that compares empirical and predicted response probabilities on the level of individual, not pooled, stimuli would provide a more direct measure of model predictive power. Whether stimuli are presented once or multiple times, does not make any difference regarding model fitting. As long as there are enough stimuli so that the predictors roughly cover the entire permissible variable space, the reliability of the model estimates mainly depends on the overall number of trials. As noted by Macke and Wichmann (2010), *"for logistic regression, the two views of classification with repeated stimuli and regression onto probabilities are mathematically equivalent."*

Additional insights into the mechanism underlying TiN detection can be expected from the comparison of data collected in various experimental conditions. For example, it would be of significant interest how the reliance on the energy cue would be affected when roving level or equal-energy stimuli were used, which corrupt the informativeness of this particular stimulus feature. Another important aspect of TiN detection could be investigated by using wider or extremely narrow-band maskers. The question at which point the spectral information becomes uninformative to the observers tightly relates to the width of the presumed auditory filters as discussed in section 3.1. By switching between conditions from block to block, it could also be studied how flexibly observers are able to adjust the distribution of perceptual weights.

Due to its flexibility and generality, the proposed analysis procedure is not only applicable to the single detection paradigm tested here. Some researchers argued that a "free running" task, where short signals are presented in a continuous noise background, represents a more natural detection paradigm (Shub and Richards, 2009). Under such conditions, perceptual weights in the temporal as well as spectral domain might be very efficiently estimated using the proposed analysis method in combination with spectrograms. Moreover, the modeling approach could also be applied to more complex tasks. For example, the broad field of speech perception may substantially benefit from the ability to analyze and identify interdependent auditory features. In speech, many of the critical stimulus aspects, for example the spectro-temporal properties of formants, are tightly coupled. A sparse regression analysis may nevertheless allow researchers to identify the critical aspects of the speech sounds that govern perception. In visual psychophysics, when using stimuli sampled from natural images, which are highly structured, many stimulus characteristics are correlated and can not be decoupled without interfering with this "natural" structure.

In conclusion, L_1-regularization in combination with a logistic regression was demonstrated to be a powerful tool for the analysis of psychophysical data, in particular in conditions where one wishes to simultaneously consider a large number of potential stimulus features and corresponding predictors. It represents a flexible approach for reverse engineering the decision mechanism of individual observers from data obtained in purely behavioral paradigms. This enables experimenters to simultaneously test large sets

of potential perceptual cues. It can be used in a broad spectrum of experimental settings, including more complex tasks than the one examined in the present work. It is not limited to auditory perception, but can also be applied in other modalities such as vision, as long as the task at hand can be meaningfully described with a linear observer model. Even in neurophysiology, it could serve as a modern and more flexible replacement for classical reverse correlation procedures. I hope that the present work encourages the application of this modern data analysis procedure in other fields of experimental psychology and opens up new opportunities for understanding and quantitatively explaining perceptual processes.

Appendices

Appendix A

Setup, Headphones and Calibration

This chapter details the experimental setup used for collecting the auditory psychophysical data. Generally, such a setup consists of a device for generating and presenting sounds and a response device for recording observer decisions. In addition, a monitor was employed for displaying instruction and feedback.

A graphical overview over the experimental setup used in the present study is provided in Fig. A.1. Digital signal generation and output, as well as response registration and feedback display were controlled from an *Apple Mac Pro* desktop computer running the scientific computing software *Matlab* (The Mathworks, Inc., 2010) with the *Psychtoolbox-3* extension, a toolbox developed for precisely controlled presentation of both visual and auditory stimuli as well as response recording (Kleiner *et al.*, 2007).

The digital sound signal was fed to an external Firewire sound card *RME Fireface 400* (Audio AG, Germany) which converted the digital signal into an analog voltage output at 96 kHz and with 24 bit precision. The sound card has a specified signal-to-noise ratio of 110 dB. As an externally connected device it avoids any risk of electro-magnetic interference from internal computer circuits. This output signal was then amplified in fanless and thus perfectly quiet headphone amplifiers *NAIM Headline* (NAIM Audio Ltd., UK) with a strictly linear response curve. Their output is able to drive low-impedance headphones ($< 10\ \Omega$) as were used in the setup.

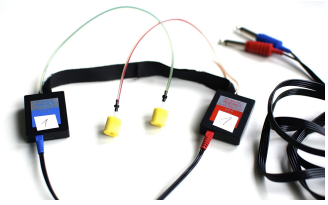
The experiments were performed in a quiet laboratory room with thick black curtains and black walls in order to minimize distraction for the listeners. They were seated at a desk which was empty except for the response box and a visual display. Two observers could perform the experiment in parallel in order to save experimental time. They were sitting in separate compartments in the lab with curtains obstructing visual contact. Both listeners were simultaneously but independently presented with the same stimuli while individual responses were recorded. However, this procedure was only realized regularly for two observers that were friends ("AB" and "AJ"). After attempts to combine other listeners, they reported to prefer being alone in the laboratory because they felt distracted. In pairs, listeners had to agree on the length of the pauses between blocks which seemed to be more difficult when observers were not acquainted.

Because experiments were not performed in a professional sound-insulated chamber, special measures had to be taken to ensure that external noise that entered through the walls or door did not effect data collection.

In particular, highly insulating in-ear headphones *Etymotic ER-2* (Etymotic Research, Inc., USA) were employed which provided a $-30dB$ external sound attenuation according to the manufacturer. In addition, sounds were presented at an average level of 70 dB

RME Fireface 400 external sound card. The display was covered with black tape to not distract observers.

NAIM headphone amplifier (top) with external power supply (bottom) and RCA signal input plugs.



A pair of *Etymotic ER-2* in-ear headphones with black neck-strap and two separate mono TRS connectors. The sound is generated inside the black boxes and transmitted to the ear through elastic plastic tubes.

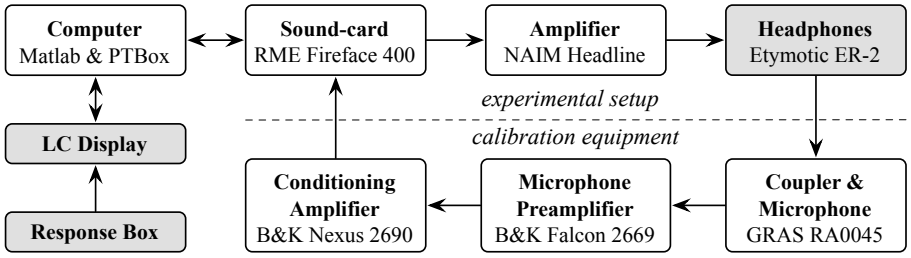


Figure A.1: Schema of the technical setup that was installed for the experiment. The bottom part on the right was only used during calibration of the headphones and not during the main experiment. The items to which the study participants are directly exposed are marked in gray.

SPL, far above the external sounds that might have reached the eardrum despite the insulation. Extensive hearing tests (see Appendix B) demonstrated that threshold levels in a wide range of frequencies did not differ significantly between the laboratory room and a professional sound insulating chamber. In conclusion, the data are very unlikely to be contaminated with influences from external noise sources.

Temporal measurements determined from the arrival of a USB signal to a computer, such as the participants reaction time recorded from an ordinary USB response box, are notoriously imprecise due to strong fluctuations in the latency inherent in the USB interface protocol. To circumvent this problem, I used special hardware for recording listener responses: the *Response Time Box* developed by Xiangrui Li.[1] Just as an ordinary response box, the *Response Time Box* is connected via the USB interface. But instead of relying on the arrival time of the USB signal at the computer, the *Response Time Box* records high-precision response times *internally* whenever a button is pressed and transmits those to the computer.[2] At the beginning of an experimental session the internal clock of the device is precisely calibrated to be in sync with the clock of the presentation computer, so that button presses on the response box can be precisely related to events of stimulus presentation. The *Psychtoolbox-3* extension contains special functions to control this particular device.

[1] Department of Psychology, University of Southern California, xiangrui.li@usc.edu, http://lobes.usc.edu/RTbox (last accessed on 06/29/12)

[2] The developer claims to achieve a precision of 0.5 ms (personal communication with Mario Kleiner).

The *Response Time Box* with four buttons and USB connector (left).

Visual information was provided on a standard desktop LC-display using *Psychtoolbox-3* graphics functions, including general instructions at the beginning of the experiment as well as feedback concerning the listener's performance (see section 6.1.2 for details).

In an auditory experiment it is essential to know the quality, and in particular the level of the sounds that are presented to the listener. Therefore, the employed headphones as well as the entire signal generation and amplification/attenuation chain need to be precisely calibrated.

Although in theory one could rely on the specification of the experimental devices to estimate the relationship between the digital signal generated in the computer and the strength of the acoustic signal entering the ear, this is difficult or may even be impossible in practice. In fact, already the digital-to-analog translation performed in the sound card depends on a large number of potentially uncontrolled factors: the logical sound driver, the state of the digital mixer and other settings on the sound card. In addition, the subsequent stages of signal manipulation (preamplifier, attenuator) may not offer direct and explicit control over their input-output properties while the resulting signal strength also depends on the usually unknown internal resistances of these devices.

Thus, in order to reliably establish the relationship between digital signal input and acoustic output, it is generally advisable to measure both values at the respective ends of the signal processing chain while essentially ignoring the intermediate processes. In

practice, this means that a calibration microphone measures the output of the headphones while the computer is configured to generate a signal at known RMS(root-mean-square)-power.

Unfortunately, the same problem that applies to signal generation applies to the case of signal recording, i.e., the output of the microphone can not be directly interpreted as an absolute measure of sound level. This ill-fated circle can be broken with a pistonphone, a device that reliable generates a specified sound level in a defined frequency range. First, the recording chain including the microphone, its pre- and main amplifier as well as the A/D-converter are jointly calibrated by adjusting the software used for recording and measuring in such a way that it reproduces the specified sound level while the microphone is exposed to the active pistonphone.

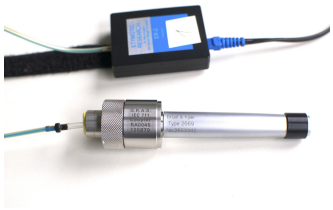
In a next step, the microphone can then be coupled to the headphones in order to determine the acoustic output with respect to the digital output signal from the sound generation software. Generally, this relationship has to be measured across the entire range of frequencies using pure tones. Assuming a linear transduction in the signal chain, the acoustic output for any kind of input signal can then be directly predicted from the individual measurements. In case that the output of a headphone is not equal across the range of frequencies used during experiments, the signal needs to be equalized, i.e., the signal in different frequency regions needs to be amplified or attenuated to correct for the variation in input-output transduction.

Ideally, a researcher in basic audition would like to control the signal that arrives at the ear canal and ultimately the tympanic membrane. Therefore, the headphone-microphone coupling represents a critical interface as it has to mimic the transfer function from the headphones to the tympanic membrane. For on-ear headphones, so-called artificial ears are commonly used for calibration, whose resonant properties resemble those of a human ear. Similar devices consisting of metallic tubes can be used for in-ear headphones as used in the present study (*Etymotic ER-2*).

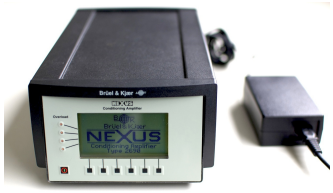
In the present experimental setup, the recording chain consisted of a *G.R.A.S. RA0045* (G.R.A.S. Sound&Vibration A/S, Denmark) in-ear headphone coupler-microphone, as well as a *B&K Falcon 2669* microphone preamplifier and *B&K Nexus 2690* conditioning amplifier (Brüel & Kjær A/S, Denmark). For sound recording the same sound card and software was used as for stimulus generation (*RME Fireface 400* and *Matlab* including the *Psychtoolbox-extension*).

After the recording chain was calibrated using a *B&K 4220* pistonphone, both pairs of headphones used in the experiment were calibrated by presenting pure tones at frequencies ranging from 100 to 20, 000 Hz (see Fig. A.2). For each individual headphone the sound level varied less than ±2.5 dB from 100 – 1000 Hz.[3] At the narrow frequency band used in the main experiment (450 – 550 Hz), the variation amounted to less than ±1 dB. Given these small deviations, I decided not to equalize the headphones across frequencies during the experiments. Such a filtering procedure would only have complicated the stimulus presentation without any noticeable advantage. In any case the headphone coupler merely mimics the resonant properties of an *average* human ear canal. Individual physiological variations, which correspond to changes in the resonant properties and thus, the effective transduction, most likely result in level deviations stronger than 1 dB. These were not taken into account in the procedure discussed here. In theory, a more precise calibration could be obtained by a sophisticated measurement of the sound level *inside the ear canal* while headphones are inserted. However, studies such as the one presented here do not commonly follow such a procedure—it is comparatively expensive and does not provide any significant benefits, unless the sound level needs to be determined very precisely. After all, it was necessary to merely confirm that the overall sound level presented to the listeners was in the appropriate range (within a few decibels) and that there were no steep variations in signal transduction which may have qualitatively altered the stimuli.

A *Brüel & Kjær 4220* pistonphone (top) including a barometer (bottom) to correct for deviations in sound level and standard frequency due to the ambient air pressure.



A *G.R.A.S. RA0045* in-ear headphone coupler connected to a *B&K Falcon 2669* microphone preamplifier (bottom) is coupled to an *Etymotic* headphone (top, bottom left).



A *B&K Nexus 2690* conditioning amplifier with power supply.

[3] These values are slightly larger than those indicated on the manufacturer specification sheet (Fig. A.3). The difference may be attributable to the use of a "Zwislocki coupler" in those measurements.

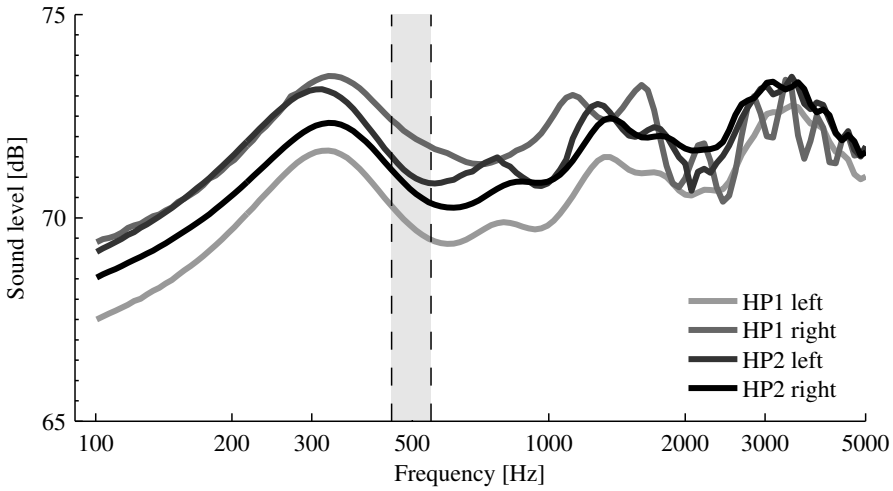


Figure A.2: Headphone Calibration I. Sound levels measured at the output of two *Etymotic ER-2* headphone sets (HP1 and HP2, respectively) across frequencies from 100-5000 Hz. The gray shaded area represents the range of frequencies occurring in the main Tone-in-Noise detection experiment.

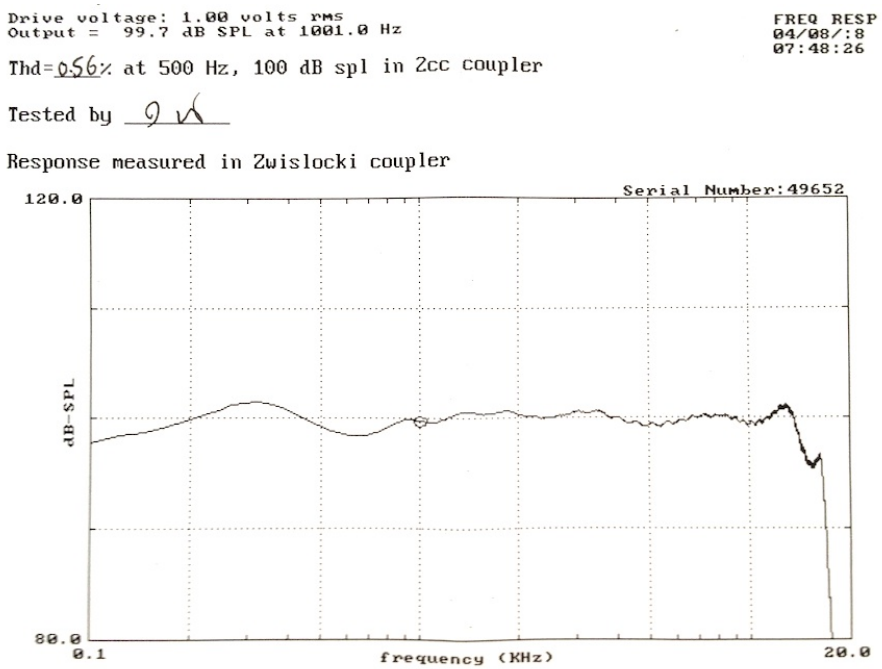


Figure A.3: Headphone Calibration II. A sample manufacturer calibration chart for one headphone. These specification sheets are prepared during production and provided with individual headphones. The y-scale is much larger as compared to Fig. A.2.

Across the four individual headphones (two left/right pairs), the overall sound output differed by 2 – 3 dB. As this was easy to correct by digitally scaling the overall stimulus amplitude, the signal to individual headphones was adjusted so that each would produce the same sound level at 500 Hz.

Appendix B

Observer Hearing Test

Listeners in psychophysical experiments concerned with normal hearing abilities are generally required to be tested for "normal hearing". The hearing threshold depends on sound frequency. A listener is commonly considered to have "normal hearing" when her threshold is no more than 10–20 dB above the average population hearing curve (Quigley and Paul, 1984; Fastl and Zwicker, 2007; ANSI S3.6).

Before listeners started the main experiment, their hearing was measured in the range of 100 Hz to 10 kHz in 20 logarithmic steps and in both ears consecutively. Pure tones were presented as short beeps repeating at 20 Hz, to be more easily discernible against spurious subjective auditory percepts. The hearing threshold was measured in a simplified staircase procedure—listeners indicated by button press whether they heard the tone. If they indicated to perceive it, the signal level was decreased by 3 dB, otherwise increased by the same amount. After three reversals the step size was lowered to 1.5 dB, and after another three reversals, the average level of the last two presentations was determined as the hearing threshold. The next frequency tested was randomly chosen from the remaining values. The same procedure was repeated for the frequency range of the main experiment, 450–550 Hz in 10 logarithmic steps.

The hearing test was performed in the hearing laboratory of my research group, a quiet and darkened room, though without professional sound insulation, as well as a large sound-insulated double-walled room manufactured by *IAC* (Industrial Acoustics Company GmbH, Niederkrüchten, Germany) and located at the Institute of Biology of the Humboldt University of Berlin, Germany. I did not find any systematic differences between the two locations—the difference in measured thresholds averaged across frequencies was smaller than 2 dB for all observers (on average 0.6 ± 1.1 dB)—except for one subject reporting a monaural Tinnitus-like percept during the hearing test in our lab (but not the main experiment), which appeared as a strong increase in threshold of $\approx$ 20 dB at 3 – 4 kHz.

In order to determine whether subjects exhibited normal hearing, I compared the measured thresholds with a standard auditory threshold curve for insert earphones combined with an occluded ear simulator (ANSI S3.6, Table 7). As depicted in Fig. B.1, across frequencies all audiograms fell below a +15 dB limit (except for the "Tinnitus"-anomaly). For the critical frequency range, 450–550 Hz, the audiograms fell within a ±10 dB range. Consequently, all observers were considered normal hearing and should not have any difficulties in discriminating sounds presented at 70 dB SPL.

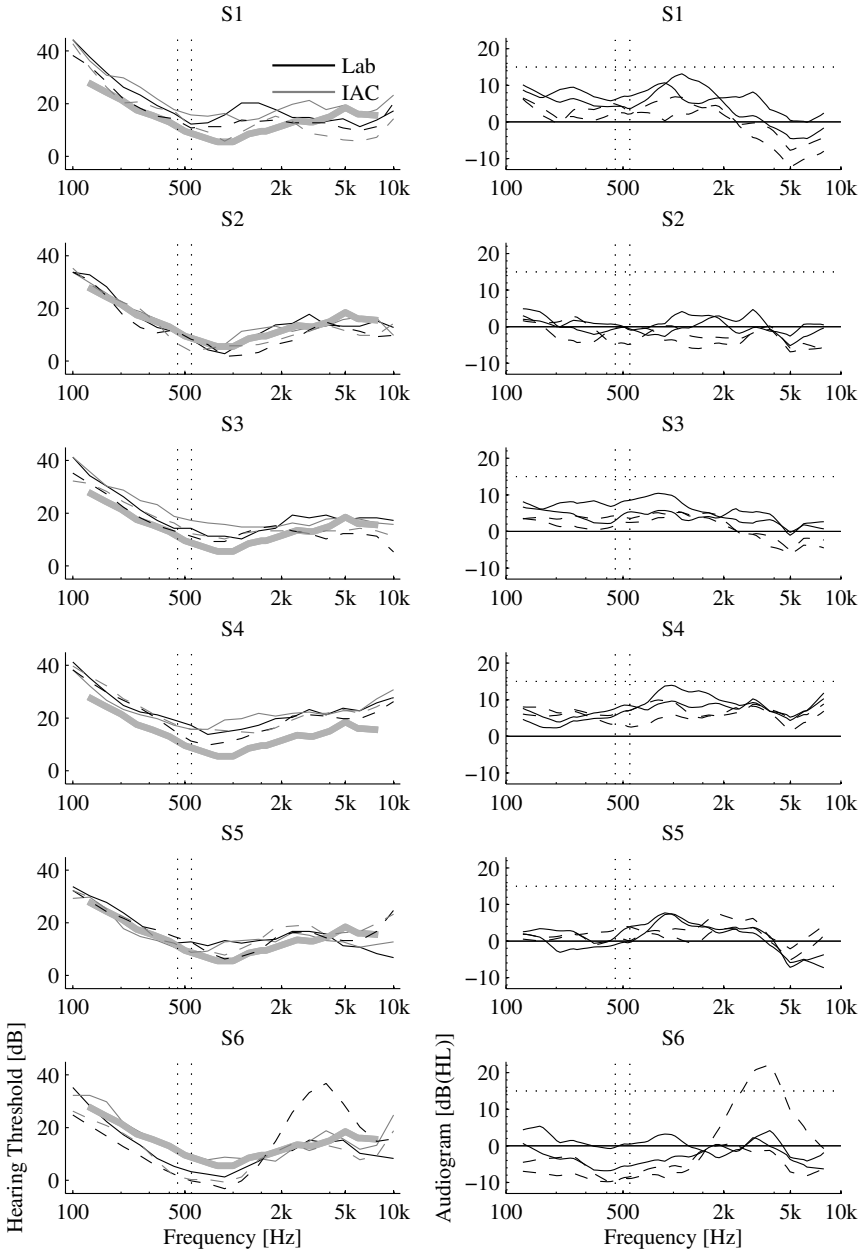


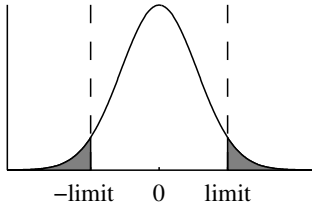
Figure B.1: Hearing Test. **Left:** Hearing thresholds with standard hearing curves (thick gray line according to (ANSI S3.6, Table 7). **Right:** Audiograms (difference of individual threshold to average population hearing threshold) for all listeners as measured in the auditory lab ("Lab", black) and sound-insulated chamber ("IAC", gray). Results from the left and right ear are plotted in dashed and clean lines, respectively. The dotted horizontal line in the audiograms indicates the upper bound within which a listener was considered as "hearing normally". The vertical lines in all plots designate the frequency region where stimuli were presented in the main experiment.

Appendix C

Clipping of Random Noise Signals

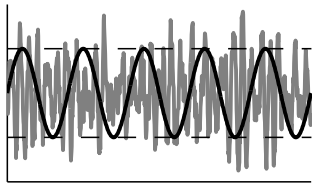
As a random statistical process, Gaussian noise generates values across an infinite range, even though results far beyond the standard error from the mean are very rare. On the other hand, a digitally stored signal, as is used for computer-generated stimuli in a perceptual experiment, can only represent a strictly limited range of values. In particular, the output to a D/A-converter—such as a sound or graphics card—only accepts a certain maximum input value and all signals above are simply clipped at that ceiling.

As a result, a signal strictly generated as Gaussian noise is modified through digital representation and D/A-conversion. The magnitude of these effects depends both on the experimental system and the statistical properties of the stimuli. In the following, I discuss the frequency at which a random signal is expected to be clipped because its value lies beyond the threshold that the experimental system can represent.



A signal generated from a Gaussian noise process always contains a certain proportion of samples that cross any finite threshold (gray area).

The percentage of samples of a normally distributed random signal that is clipped corresponds to the surface of the Gaussian density function that extends beyond the clipping threshold (see top side plot) and can be computed from the cumulative normal distribution. In order to obtain a well-controlled signal with a small amount of clipped samples, the sound level needs to be adjusted so that only a small percentage of sound samples falls beyond the clipping limit. In general, a noise signal needs to be fainter than a pure tone to be processed by the system without significant distortion (see bottom side plot).


A pure tone signal and a random noise signal with equal average sound level. The pure tone never crosses the threshold, unlike the noise signal.

In Fig. C.1, I plotted the relationship between the fraction of Gaussian noise samples that cross the threshold and the sound level of the noise signal. The sound level is specified as the level difference between the noise and the loudest stationary pure tone that can be transmitted without clipping (a sine wave that oscillates between ±threshold). For example, to have less than one in every thousand sound samples cross the threshold, the level must be -10 dB below the pure tone. These "sound levels" only refer to the stimulus as represented in the computer and sound card. A decrease in "digital" sound level can be fully compensated by increasing, e.g., the overall gain of the sound card or headphone amplifier.

In a TiN detection experiment, both noise and signal+noise stimuli are presented, the latter having higher amplitudes on average depending on signal level. On the other hand, temporal windowing of stimuli (see Appendix D) reduces the risk of clipping. For the present experiments, the level of the stimuli was adjusted so that none of the *noise stimuli* showed any clipping. For the highest signal level employed (15 dB), i.e., the worst condition in terms of clipping, about one in 2000 *signal+noise stimuli* exhibited clipping. This corresponds to less than one stimulus for every other experimental session or to less

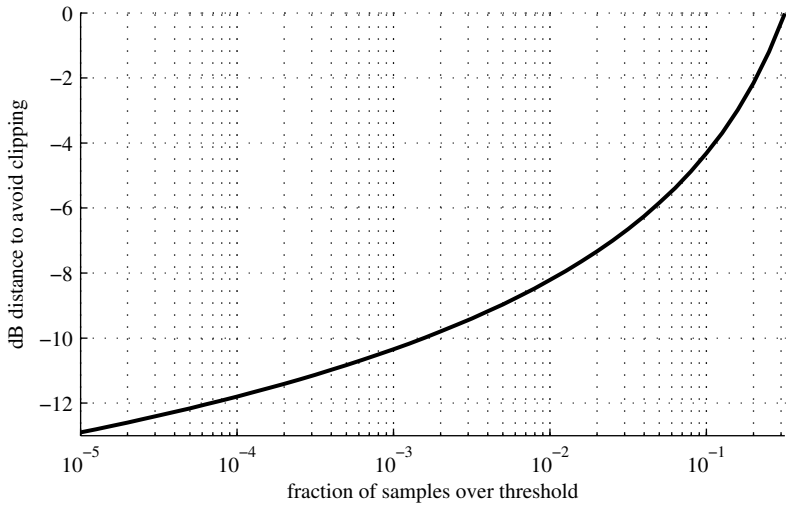
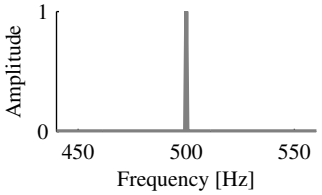


Figure C.1: Relationship between sound level and clipping. Depending on the overall sound level, a certain amount of sound samples of Gaussian random noise is clipped. In order to decrease the number of clipped samples (x-axis), the overall sound level of the noise needs to be attenuated to be significantly more faint than the loudest pure tone that the system can process without clipping (y-axis).

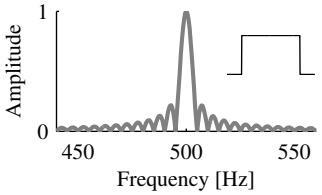
than two out of the approximately 2500 signal+noise trials collected in each condition. Neither should this influence observer behavior nor have any significant influence on analysis results. For lower signal levels, the clipping rate was even smaller.

Appendix D

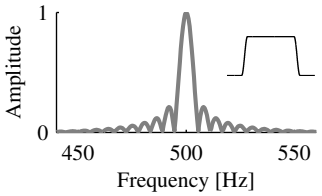
Spectral Leakage and Rise/Fall-Times



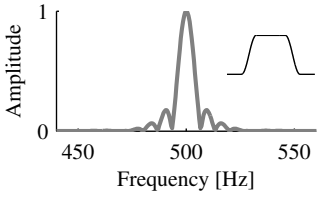
Fourier spectrum of a 500 Hz tone.



A 500 Hz tone that lasts 200 msec with a rectangular windowing function (the small inset depicts the sound level over time).



Spectrum after applying a 20 msec modified Hann window (black line in Fig. D.1).



Spectrum after applying a 50 msec modified Hann window (white line in Fig. D.1).

Generally, a Fourier transformation (FT) of finite signals assumes cyclic boundary conditions, i.e., the signal is expected to repeat at infinity. The FT of a pure tone has a single peak only in the case of a tone being presented forever. The FT of a pure tone of limited duration contains additional components that spread laterally to both sides of the frequency of the tone, and are stronger for shorter tones.

Short duration signals, such as presented during many auditory experiments, result in such *"spectral leakage"* especially when rectangular on/off-gating functions are employed. The resulting signal contains frequency components that are different from the pure tones the signal was originally composed off.

The resulting FT can be derived by transforming the unit rectangle. The FT of the unit rectangle corresponds to a sinc-function $sinc(x) = \sin(\pi x)/(\pi x)$, which has long tails. Consequently, the FT of a pure tone gated with a unit rectangle function has long tails as well. Following the *"uncertainty principle"*, the length of the signal is inversely proportional to the width of the leakage in the frequency domain (Hartmann, 1998, chap. 8).

In practice, with temporally finite stimuli, the effect of spectral leakage can never be entirely avoided. In addition, technical equipment such as D/A converters, amplifiers and headphones may contribute to this effect, e.g., by introducing nonlinear transformations. However, the use of slowly raising on- and off-ramps at the beginning and end of the sound signal can significantly reduce spectral spread. The experimenter needs to find a good trade-off for the length of the ramp: longer ramps decrease spread but also shorten the steady-state portion of the signal.

The stimuli used in the present study had a length of 200 ms and a spectral width of 100 Hz. For signal windowing I relied on a standard cosine-squared gating function (or "Hann window"):

$$w = 1 - \cos^2 \pi t = \frac{1}{2}\left(1 - \cos 2\pi t\right) \tag{D.1}$$

which was modified to contain a centered steady state portion. As the optimal trade-off, a rise/fall-time of ¼ of the stimulus length, or 50 ms, was chosen resulting in a 100 ms steady-state length. As can be derived from the graph in Fig. D.1 (white dashed line), the spectral leakage extends about 25 Hz beyond the original signal (corresponding to five Fourier bins). Outside, the spectral side lobes are attenuated by -50 dB or below and are very unlikely to interfere with the observers perception.

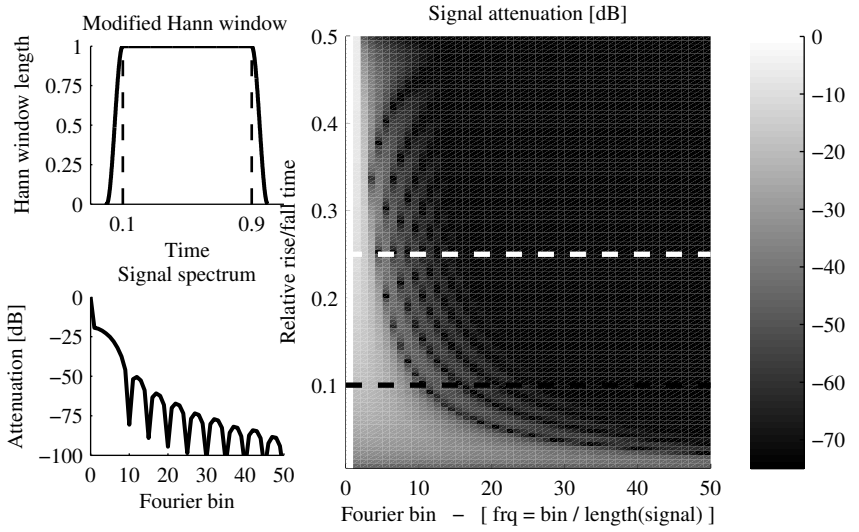


Figure D.1: Spectral properties of a finite signal gated with a modified Hann window. **Left:** Sample temporal envelope of a pure tone signal gated with a modified Hann window of ⅒ of the signal length (top) and corresponding frequency spectrum on a decibel scale (bottom). **Right:** Fourier spectrum (x-axis and color) of a signal gated with modified Hann windows of different widths (y-axis). The black line represents the sample signal on the left (and the third plot in the margin on the previous page), the white line represents the Hann window width used in the experiments (as depicted in the fourth plot in the margin of the previous page). The frequency is represented as the Fourier bin. To determine the respective frequency, the bin has to be divided by the overall length of the signal (e.g., for a 200 ms signal each Fourier bin has a width of 5 Hz and bin #10 corresponds to 50 Hz.). The rise/fall-time of the Hann window is given as a number relative to the overall signal length. A window of length 0.5 corresponds to a signal where the rise- and fall-time jointly span the entire signal length and there is no steady-state portion.

Appendix E

Digital Filters

The strategy of some of the artificial observers described in section 5.1.1 depends on digital filters in the spectral as well as the envelope-spectral domain. A precise definition of these frequency filters is provided in the present section.

In general, digital filters can be used to remove certain portions of the spectrum of a time-dependent signal. Ordinary filters are linear, i.e., different frequency components do not interact, and time-invariant, i.e., response properties are constant and independent of the input signal. Since filtering is simply described by a multiplication in the frequency domain, the general filtering procedure usually follows these steps (Hartmann, 1998):

1. Transform the signal from a time to a frequency representation using Fourier transformation.

2. In Fourier space, filter the signal by multiplying the individual components with a filter function $H(s)$, where s represents the complex frequency variable: $s = if$.

3. Transform the filtered signal back to the time domain with the inverse Fourier transform.

In the following, the simplest and most common second order filters are defined (Hartmann, 1998).

E.1 Band-Pass Filter

The simplest band-pass filter is characterized by a gain, a center frequency and a width. Its transfer function—which determines the change in amplitude (factor $|H|$) and phase (angle $\angle(H)$) of the individual frequency components—can be characterized as

$$H(s) = \frac{KWs}{s^2 + Ws + F_o^2} = \frac{KWif}{-f^2 + Wif + F_o^2} = \frac{Kif F_o/Q}{-f^2 + if F_o/Q + F_o^2} \tag{E.1a}$$

$$|H| = \frac{KWf}{\sqrt{(F_o^2 - f^2)^2 + W^2 f^2}} \tag{E.1b}$$

$$\angle(H) = \tan^{-1} \frac{F_o^2 - f^2}{Wf} \tag{E.1c}$$

with K = gain, W = bandwidth, F_o = center frequency. The quality factor is defined as $Q = F_o/W$. The rise time t_R, the time the response takes to go from 10% to 90%, is directly related to the bandwidth W: $Wt_R = 0.7$.

E.2 Low-Pass Filter

A low-pass filter removes high frequencies from a signal and has the following transfer function:

$$H(s) = \frac{K}{(s/F_c)^2 + 2s/F_c + 1} = \frac{K}{1 - (f/F_c)^2 + \sqrt{2}if/F_c} \tag{E.2a}$$

$$|H| = \frac{K}{\sqrt{1 + (f/F_c)^4}} \tag{E.2b}$$

$$\angle(H) = \tan^{-1}\frac{\sqrt{2}f/F_c}{(f/F_c)^2 - 1} \tag{E.2c}$$

with K = gain, F_c = cut-off frequency and quality factor of $Q = 1/\sqrt{2}$. The asymptotic gain amounts to $-12dB/octave$.

E.3 High-Pass Filter

A high-pass filter removes low frequencies from a signal and has the following transfer function:

$$H(s) = \frac{K}{(F_c/s)^2 + \sqrt{2}F_c/s + 1} = \frac{K}{1 - (F_c/f)^2 - \sqrt{2}iF_c/f} \tag{E.3a}$$

$$|H| = \frac{K}{\sqrt{1 + (F_c/f)^4}} \tag{E.3b}$$

$$\angle(H) = \tan^{-1}\frac{\sqrt{2}F_c/f}{(F_c/f)^2 - 1} \tag{E.3c}$$

with K = gain, F_c = cut-off frequency and quality factor set to $Q = 1/\sqrt{2}$, while the asymptotic gain is again $-12dB/octave$. The cut-off frequency F_c or bandwidth W correspond to $-3dB$ in power, thus $1/2$ in power or $1/\sqrt{2}$ in amplitude.

Appendix F

Rescaling of Weights to Linear Filters

I intend to interpret the weights from the linear model that refer to spectral predictors as a filter that the observer relies on when making a decision. As a preprocessing step, individual predictors had been standardized, i.e., they were individually scaled. The model weights refer to these scaled inputs. In contrast, the observer filters are applied in the original non-scaled input domain. In order to be interpreted as filter gains, the model weights therefore need to be rescaled to the range of the original (unscaled) predictors. This section provides a derivation of the scaling that must be applied to the predictor weights.

An observer with a linear filter decision mechanism is described as

$$O_{obs} = \sum_{i}^{pred.} f_i \tilde{p}_i + f_0 \tag{F.1}$$

with filter $\vec{f}$, bias f_0, and *non-normalized* predictor p. After model fitting, the linear model output to an individual input sample p is computed as

$$O_{model} = \sum_{i}^{pred.} w_i p_i + b \tag{F.2}$$

with predictor weight $\vec{w}$, bias b and the individual *normalized* predictor sample p_i.

$$p_i = \frac{\tilde{p}_i - < \tilde{p}_i >}{\sigma(\tilde{p}_i)} \tag{F.3}$$

where mean and standard deviation are determined over a sufficiently large input data set, so that $< p_i >= 0$ and $\sigma(p_i) = 1$ over the presented samples and for all predictors p_i.

As a simplification, I omitted here the logistic function that maps the linear decision variable to the response probability. Nevertheless, I was able to identify the decision mechanism of the linear model with an observer employing a linear filter. Thus, the output of both must be equal $O_{obs} = O_{model}$:

$$\sum_i f_i \tilde{p}_i + f_0 = \sum_i w_i p_i + b \tag{F.4a}$$

$$= \sum_i \frac{\tilde{p}_i - < \tilde{p}_i >}{\sigma(\tilde{p}_i)} w_i + b \tag{F.4b}$$

$$= \sum_i \left[\frac{w_i}{\sigma(\tilde{p}_i)} \right] \tilde{p}_i + \left(- \sum_i \frac{< \tilde{p}_i >}{\sigma(\tilde{p}_i)} w_i + b \right) \tag{F.4c}$$

In principle, the $\tilde{p}_i$'s can have any value, therefore each individual summand inside the first sum (in square brackets) must be equal to the corresponding f_i. Similarly, the second term in round brackets must be equal to f_o. Consequently, the necessary rescaling of the model weights to observer filters can be directly derived.

$$f_i = \frac{w_i}{\sigma(\tilde{p}_i)} \tag{F.5}$$

while the bias corresponds to

$$f_o = b - \sum_i \frac{<\tilde{p}_i>}{\sigma(\tilde{p}_i)} w_i. \tag{F.6}$$

The ensemble of filter components f_i then represents the filter that is supposedly applied by the observer to the original non-scaled predictor.

Appendix G

Observer Consistency and Predictability

As discussed in section 2.5, in order to interpret the predictive power of an observer model, the reliability of the observer needed to be taken into account. Essentially, the limited reliability of the listener sets a strict upper bound for any model to predict the behavior. In the following, I attempt to derive this upper bound regarding both model-observer agreement as well as model likelihood.

G.1 Upper Bound for Model-Observer Agreement Depending on Observer Consistency

Given observer consistency measured in two-pass experiment, there exists a maximum "percent agreement" (between observer and model) that the best model can achieve, as derived in the Appendix of Neri and Levi (2006). However, there is an even simpler way of establishing the relationship between model-observer agreement and observer consistency.

As the first step, I consider one stimulus for which the observer has a certain probability $\tilde{p}_{yes}$ of responding "Yes". If this stimulus is presented twice, then the expected consistency regarding this stimulus is identical with the probability that the observer gives two identical responses:

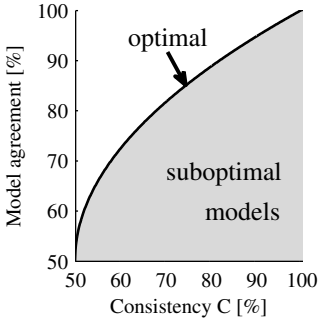
$$C = \underbrace{\tilde{p}_{yes}^2}_{\text{Two Yes-responses}} + \underbrace{\tilde{p}_{no}^2}_{\text{Two No-responses}} \tag{G.1a}$$

$$= \tilde{p}_{yes}^2 + (1 - \tilde{p}_{yes})^2 \tag{G.1b}$$

$$= 1 + 2(\tilde{p}_{yes}^2 - \tilde{p}_{yes}) \tag{G.1c}$$

A perfectly fitted binary model *always* gives the answer with a higher associated probability. For the sake of simplicity, I assume that $\tilde{p}_{yes} > 0.5$ (otherwise $\tilde{p}_{yes}$ could be replaced with $\tilde{p}_{no} = 1 - \tilde{p}_{yes}$), then the model *always* responds "Yes". This response necessarily agrees with the observer with probability $\tilde{p}_{agree} = \tilde{p}_{yes}$. With these results, the relationship between observer consistency C and model-observer agreement $\tilde{p}_{agree}$ can be directly derived:

$$\tilde{p}_{agree} = \frac{1}{2} + \sqrt{\frac{C}{2} - \frac{1}{4}} \tag{G.2}$$



Relationship between observer consistency C and model-observer agreement $\tilde{p}_{agree}$. Given observer consistency, no model can reliably achieve a prediction agreement above the black curve.

Every imperfect model achieves an agreement below this upper bound while the results of a perfect model precisely follow the bound.

G.2 Upper Bound for Model Likelihood

In the following, I attempt to establish a similar bound for the likelihood that a model can achieve. Likelihood and percent agreement are tightly related. While percent agreement simply compares the binary responses of the observer with the binary responses of the model, likelihood takes the probability into account, with which the model predicts a certain outcome, usually ranging from 0% (always "No") to 100% (always "Yes"). Predicting the correct/wrong response with large confidence is rewarded/punished more strongly, than uncertain predictions closer to 50%. In general, it is a more sensitive measure for the quality of the model. It is preferably used during cross-validation as it varies less across different subsets of the data. Being a graded measure, it allows a more fine grained model comparison: models that achieve the same percent agreement score, may still differ significantly in terms of the likelihood.

In order to find the maximum achievable likelihood given a measured consistency, I use the Lagrange method. The log-likelihood is to be maximized, assuming a different value of p for each trial:

$$L = \sum_{i=1}^{k} r_i \log \tilde{p}_i + (1 - r_i) \log(1 - \tilde{p}_i) \tag{G.3}$$

The consistency results in a constraint:

$$C = \frac{1}{k} \sum_{i=1}^{k} \tilde{p}_i^2 + (1 - \tilde{p}_i)^2 \tag{G.4}$$

The Lagrangian needs to be extremized:

$$\Lambda = L + \lambda C \tag{G.5}$$

Setting the derivative for $\tilde{p}_i$ to zero

$$\partial_{\tilde{p}_i} \Lambda = 0 \tag{G.6}$$

leads to a set of equations, each of which depend on each $\tilde{p}_i$ separately, i.e., there are no interdependencies for the different $\tilde{p}_i$'s. Thus, instead of using a probability $\tilde{p}_i$ for each trial, only one p can be used, which applies for the Yes-trials as well as the No-trials with $1 - p$. Then, the log-likelihood and consistency become

$$L = n_y \log p + n_n \log p = k \log p \tag{G.7}$$

$$C = p^2 + (1 - p)^2 \tag{G.8}$$

The derivative of the Lagrangian

$$\Lambda = k \log p + \lambda (p^2 + (1 - p)^2) \tag{G.9}$$

simplifies to

$$\partial_p \Lambda = \frac{k}{p} + \lambda (4p - 2) \tag{G.10}$$

Setting the derivative to zero results in the quadratic form

$$p^2 - \frac{p}{2} + \frac{k}{4\lambda} = 0 \tag{G.11}$$

which is solved by

$$p = \frac{1}{4} \pm \sqrt{\frac{1}{16} - \frac{k}{4\lambda}} \tag{G.12}$$

These p's are entered into the constraint function to

$$C = \frac{3}{4} - \frac{k}{2\lambda} \mp \sqrt{\frac{1}{16} - \frac{k}{4\lambda}} \tag{G.13}$$

This equation can be solved for the Lagrange multiplier λ/k

$$\frac{\lambda}{k} = \frac{1 - 2c \mp \sqrt{2c^2 - 1}}{2(c-1)(2c-1)} = \frac{-1 \mp 1/\sqrt{2c-1}}{2(c-1)} \tag{G.14}$$

This multiplier is entered into the quadratic equation for p, which then gives rise to the probability of responses that is consistent with the constraint and leads to the maximum likelihood

$$L_{max} = k \log \tilde{p}_{max}. \tag{G.15}$$

In fact, it suffices to only consider the λ/k and p with the +-sign, they give the maximum likelihood that conforms with the consistency.

However, these theoretical results have no true relevance in practice. When generating arbitrary data with randomly sampled p's for each trial, the likelihood that the best predicting model achieves is indeed bounded by the computed maximum, but the values are mostly distributed in an area far away from the bound. This is because the maximum bound is only achieved when all p's are equal in all the trials, which in practice never happens. When there are very few trials, i.e., with a small k around 2–16, the upper bound is more often reached by the data, but with larger k's, it merely is a theoretical, but not a practical limit.

G.3 Expected Value and Variance of Model Likelihood

I continue by computing the expected value (and the variance) of the likelihood distribution given a fixed number of trials k and consistency C, assuming the ideal model, which predicts the correct p's and the constraint derived from the consistency, is known. First, X and Y (the number of Yes- and No-responses) are assumed to be binomially distributed:

$$X \sim Binom(n_y, p) \text{ and } Y \sim Binom(n_n, 1-p)$$

Now I compute how often Yes- or No-responses are again observed for previous Yes- or No-trials during the second round, the test set. The corresponding likelihood amounts to:

$$\begin{aligned} L \;=\; & \underbrace{X \log p}_{\text{Yes-responses in Yes-trials}} + \underbrace{(n_n - Y)\log p}_{\text{Yes-responses in No-trials}} \\ & + \underbrace{(n_y - X)\log(1-p)}_{\text{No-responses in Yes-trials}} + \underbrace{Y \log(1-p)}_{\text{No-responses in No-trials}} && \text{(G.16a)} \\ =\; & [n_n + X - Y]\log p + [n_y - (X-Y)]\log(1-p) && \text{(G.16b)} \\ =\; & n_n \log p + n_y \log(1-p) + (X-Y)[\log p - \log(1-p)] && \text{(G.16c)} \end{aligned}$$

The expected value of the likelihood is

$$\mathbb{E}(L) = [\mathbb{E}(X) - \mathbb{E}(Y) + n_n]\log p + [\mathbb{E}(Y) - \mathbb{E}(X) + n_y]\log(1-p) \tag{G.17}$$

With $k = n_y + n_n$, $\mathbb{E}(X) = n_y p$ and $\mathbb{E}(Y) = n_n(1-p)$, I arrive at the simple form

$$\mathbb{E}(L) = kp\log p + k(1-p)\log(1-p) \tag{G.18}$$

The variance is more complicated, and is derived from:

$$\mathbb{V}(L) = \mathbb{E}(L^2) - \mathbb{E}(L)^2 \tag{G.19}$$

For this, one would need to know $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$. For the binomially distributed variable X it is true that

$$\mathbb{V}(X) = n_y p(1-p) \tag{G.20}$$

On the other hand, it is also known that

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 && \text{(G.21a)} \\ &= \mathbb{E}(X^2) - n_y^2 p^2 && \text{(G.21b)} \end{aligned}$$

From that, I obtain:

$$\begin{aligned} \mathbb{E}(X^2) &= n_y p + n_y p^2(n_y - 1) && \text{(G.22)} \\ \mathbb{E}(Y^2) &= n_n(1-p) + n_n(1-p)^2(n_n - 1) && \text{(G.23)} \end{aligned}$$

In addition, there are cross terms, which easily decompose into $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, since X and Y are independent. In the following, I focus on L^2, with $Z = X - Y$ for simplification

$$\begin{aligned} L^2 &= [(n_n + Z)\log p + (n_y - Z)\log(1-p)]^2 && \text{(G.24a)} \\ &= [\underbrace{n_n \log p + n_y \log(1-p)}_{a} + Z\underbrace{(\log p - \log(1-p))}_{b}]^2 && \text{(G.24b)} \\ &= a^2 + b^2 Z^2 + 2abZ && \text{(G.24c)} \end{aligned}$$

With $\mathbb{E}(Z) = \mathbb{E}(X) - \mathbb{E}(Y)$ and $\mathbb{E}(Z^2) = \mathbb{E}(X^2) + \mathbb{E}(Y^2) - 2\mathbb{E}(X)\mathbb{E}(Y)$ I obtain

$$\mathbb{E}(L^2) = a^2 + b^2[\mathbb{E}(X^2) + \mathbb{E}(Y^2) - 2\mathbb{E}(X)\mathbb{E}(Y)] + 2ab[\mathbb{E}(X) - \mathbb{E}(Y)] \tag{G.25}$$

and after including all that is known about $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ etc., I arrive at

$$\begin{aligned} \mathbb{E}(L^2) = {} & (n_n + n_y)[(n_n(p-1) + n_y(p-1) - p](p-1)\log^2(1-p) \\ & -2(n_n + n_y - 1)(p-1)p\log(1-p)\log p \\ & +p[1 + (n_n + n_y - 1)p]\log^2 p) \end{aligned} \tag{G.26}$$

which can be further simplified using $k = n_y + n_n$

$$\begin{aligned} \mathbb{E}(L^2) = {} & k[(p-1)[(p-1)(k-1) - 1]\log^2(1-p) \\ & +p[p(k-1) + 1]\log^2 p - 2p(p-1)(k-1)\log p\log(1-p)] \end{aligned} \tag{G.27}$$

To make it a little more clear, with $\hat{p} = 1 - p$ this can be written as

$$\begin{aligned} \mathbb{E}(L^2) = {} & k[\hat{p}[\hat{p}(k-1) + 1]\log^2\hat{p} + p[p(k-1) + 1]\log^2 p \\ & +2p\hat{p}(k-1)\log p\log\hat{p}] \end{aligned}$$

Now, together with $\mathbb{E}(L)$ (Eq. G.18), I know all ingredients for $\mathbb{V}(L)$ (Eq. G.19). I computed both the expected value as well as the variance of the distribution of L that is achieved under the assumption of one fixed p- and $(1-p)$-value for Yes- and No-responses, respectively. When comparing these values to simulated data sampled with different trial numbers, it is found the more trials there are the more the empirical distribution deviates. The reason is, that in reality, p fluctuates from trial to trial affecting the likelihood L accordingly—the more trials there are, the more the p varies.

G.4 Conclusion

Unfortunately, as long as the distribution of the probability p across trials is unknown, there is no way to make a precise prediction about the distribution of the likelihood L. Consistency only provides some vague idea of the range of the probabilities p, and dictates a strict upper bound for L. One option to circumvent this obstacle would be to assume a β-distribution for p and compute its parameters from response bias and consistency. Unfortunately, the distribution of p is likely to be more complex. Thus, in order to make reliable predictions about L, one would need to learn the true distribution by presenting each stimulus a number of times. But at that point, the task of finding the maximum likelihood that the best model could achieve becomes rather uninteresting—instead, one would evaluate the quality of the model by directly comparing the measured and predicted probabilities.

Bibliography

Abbey, C. K. and Eckstein, M. P. (**2002**). "Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments.", Journal of Vision **2**, 66–78.

Abbey, C. K. and Eckstein, M. P. (**2006**). "Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer.", Journal of Vision **6**, 335–355.

Ahumada, A. (**1996**). "Perceptual classification images from Vernier acuity masked by noise", Perception **25**, ECVP Abstract Supplement.

Ahumada, A. (**2002**). "Classification image weights and internal noise level estimation", Journal of Vision **2**, 121–131.

Ahumada, A. and Lovell, J. (**1971**). "Stimulus features in signal detection", The Journal of the Acoustical Society of America **49**, 1751–1756.

Ahumada, A. J. (**1967**). "Detection of tones masked by noise: A comparison of human observers with digital-computer-simulated energy detectors of varying bandwidths", Technical Report, University of California, Los Angeles, CA, USA, doctoral dissertation, available as Technical Report No. 29, Human Communications Laboratory, Department of Psychology, University of California, Los Angeles.

Alexander, J. M. and Lutfi, R. A. (**2004**). "Informational masking in hearing-impaired and normal-hearing listeners: Sensation level and decision weights", The Journal of the Acoustical Society of America **116**, 2234–2247.

ANSI S3.6 (**2004**). *ANSI S3.6-2004, Specifications for Audiometers*, American National Standards Institute.

Berg, B. G. (**1989**). "Analysis of weights in multiple observation tasks", The Journal of the Acoustical Society of America **86**, 1743–1746.

Berg, B. G. (**2004**). "A temporal model of level-invariant, tone-in-noise detection", Psychological Review **111**, 914–930.

Berg, B. G., Nguyen, Q. T., and Green, D. M. (**1992**). "Discrimination of narrow-band spectra. I: Spectral weights and pitch cues", The Journal of the Acoustical Society of America **92**, 1911–1918.

Bishop, C. M. (**2006**). *Pattern Recognition and machine learning*, 738 pages (Springer Science+Business Media, LLC, New York, USA).

Blackwell, H. R. (**1952**). "Studies of Psychophysical Methods for Measuring Visual Thresholds", Journal of the Optical Society of America **42**, 606–614.

Browne, M. W. (**2000**). "Cross-Validation Methods.", Journal of Mathematical Psychology **44**, 108–132.

Busse, L., Ayaz, A., Dhruv, N. T., Katzner, S., Saleem, A. B., Schölvinck, M. L., Zaharia, A. D., and Carandini, M. (**2011**). "The detection of visual contrast in the behaving mouse.", The Journal of neuroscience : the official journal of the Society for Neuroscience **31**, 11351–11361.

Candes, E. and Tao, T. (**2005**). "Decoding by linear programming", arXiv **math.MG**.

Carney, L. H., Heinz, M. G., Evilsizer, M. E., Gilkey, R. H., and Colburn, H. S. (**2002**). "Auditory phase opponency: A temporal model for masked detection at low frequencies", Acta Acustica United With Acustica **88**, 334–347.

Chopin, A. and Mamassian, P. (**2012**). "Predictive properties of visual adaptation.", Current biology **22**, 622–626.

Claerbout, J. F. and Muir, F. (**1973**). "Robust modeling with erratic data", Geophysics **38**, 826–844.

Collett, D. (**2003**). *Modelling Binary Data*, 408 pages, 2nd edition (Chapman & Hall/CRC, Boca Raton, Florida, USA).

Cox, D. R. (**1958**). "The regression analysis of binary sequences", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **20**, 215–242.

Cramer, J. S. (**2003**). *Logit Models from Economics and Other Fields*, chapter 9, 2nd edition (Cambridge University Press).

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers", The Journal of the Acoustical Society of America **102**, 2892–2905.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**a). "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure", The Journal of the Acoustical Society of America **99**, 3615–3622.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**b). "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements.", The Journal of the Acoustical Society of America **99**, 3623–3631.

Davidson, S. A., Gilkey, R. H., Colburn, H. S., and Carney, L. H. (**2006**). "Binaural detection with narrowband and wideband reproducible noise maskers. III. Monaural and diotic detection and model results", The Journal of the Acoustical Society of America **119**, 2258–2275.

Davidson, S. A., Gilkey, R. H., Colburn, H. S., and Carney, L. H. (**2009**a). "An evaluation of models for diotic and dichotic detection in reproducible noises", The Journal of the Acoustical Society of America **126**, 1906–1925.

Davidson, S. A., Gilkey, R. H., Colburn, H. S., and Carney, L. H. (**2009**b). "Diotic and dichotic detection with reproducible chimeric stimuli", The Journal of the Acoustical Society of America **126**, 1889–1905.

De Boer, E. and Kuyper, P. (**1968**). "Triggered correlation", IEEE Transactions on Biomedical Engineering **BME-15**, 169–179.

Dobson, A. J. and Barnett, A. G. (**2008**). *An Introduction to Generalized Linear Models*, 3rd edition (Chapman and Hall/CRC).

Donoho, D. L. (**2004**). "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution", Communications on Pure & Applied Mathematics **59**, 797–829.

Donoho, D. L., Elad, M., and Temlyakov, V. (**2006**). "Stable recovery of sparse overcomplete representations in the presence of noise", IEEE Transactions on Information Theory **52**, 6–18.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, chapter 6.11, 2nd edition (John Wiley & Sons, Inc., New York).

Dye, R. H., Stellmack, M. A., and Jurcin, N. F. (**2005**). "Observer weighting strategies in interaural time-difference discrimination and monaural level discrimination for a multi-tone complex.", The Journal of the Acoustical Society of America **117**, 3079–3090.

Evilsizer, M. E., Gilkey, R. H., Mason, C. R., Colburn, H. S., and Carney, L. H. (**2002**). "Binaural detection with narrowband and wideband reproducible noise maskers: I. Results for human.", The Journal of the Acoustical Society of America **111**, 336–345.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin., C.-J. (**2008**). "Liblinear: A library for large linear classification", Journal of Machine Learning Research **9**, 1871–1874, URL `http://www.csie.ntu.edu.tw/~cjlin/liblinear`, (date last accessed 06/29/2012).

Fastl, H. and Zwicker, E. (**2007**). *Psychoacoustics – Facts and Models*, 3rd edition (Springer-Verlag Berlin Heidelberg, Germany).

Fechner, G. T. (**1860**). *Elemente der Psychophysik* (Breitkopf und Härtel, Leipzig, Germany).

Fletcher, H. (**1938**). "Loudness, masking and their relation to the hearing process and the problem of noise measurement", The Journal of the Acoustical Society of America **9**, 275–293.

Fletcher, H. (**1940**). "Auditory patterns", Reviews of Modern Physics **12**, 47–65.

Forster, M. R. (**2000**). "Key Concepts in Model Selection: Performance and Generalizability", Journal of Mathematical Psychology **44**, 205–231.

Franz, M. O. and Schölkopf, B. (**2006**). "A unifying view of wiener and volterra theory and polynomial kernel regression", Neural computation **18**, 3097–3118.

French, N. R. and Steinberg, J. C. (**1947**). "Factors Governing the Intelligibility of Speech Sounds", The Journal of the Acoustical Society of America **19**, 90–119.

Fründ, I., Haenel, N. V., and Wichmann, F. A. (**2011**a). "Inference for psychometric functions in the presence of nonstationary behavior.", Journal of Vision **11**.

Fründ, I., Wichmann, F. A., and Macke, J. H. (**2011**b). "Sequential dependencies in perceptual decisions", in *European Mathematical Psychology Group Meeting* (Paris, France).

Fründ, I., Wichmann, F. A., and Macke, J. H. (**2012**). "Dealing with sequential dependencies in psychophysical data", in *COSYNE* (Salt Lake City, USA).

Gilkey, R. H. and Robinson, D. E. (**1986**). "Models of auditory masking: A molecular psychophysical approach", The Journal of the Acoustical Society of America **79**, 1499–1510.

Gold, T. (**1948**). "Hearing. II. The Physical Basis of the Action of the Cochlea", Proceedings of the Royal Society B: Biological Sciences **135**, 492–498.

Goris, R. L. T., Wagemans, J., and Wichmann, F. A. (**2008**). "Modelling contrast discrimination data suggest both the pedestal effect and stochastic resonance to be caused by the same mechanism.", Journal of Vision **8**, 17.1–21.

Green, D. M. (**1960**). "Psychoacoustics and Detection Theory", The Journal of the Acoustical Society of America **32**, 1189–1203.

Green, D. M. (**1964**). "Consistency of auditory detection judgments", Psychological Review **71**, 392–407.

Green, D. M., Berg, B. G., Dai, H., Eddins, D. A., Onsan, Z., and Nguyen, Q. T. (**1992**). "Spectral shape discrimination of narrow-band sounds", The Journal of the Acoustical Society of America **92**, 2586–2597.

Green, D. M., Luce, R. D., and Duncan, J. E. (**1977**). "Variability and sequential effects in magnitude production and estimation of auditory intensity", Perception & Psychophysics **22**, 450–456.

Green, D. M. and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics*, 455 pages (John Wiley & Sons, Inc., New York).

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis", The Journal of the Acoustical Society of America **76**, 50–56.

Hartmann, W. M. (**1998**). *Signals, Sound, and Sensation (Modern Acoustics and Signal Processing)* (Springer Science+Business Media, LLC, New York).

Hastie, T., Tibshirani, R., and Friedman, J. (**2009**). *The Elements of Statistical Learning*, 745 pages, 2nd edition (Springer Science+Business Media, LLC, New York).

Hawkins, J. E. and Stevens, S. S. (**1950**). "The Masking of Pure Tones and of Speech by White Noise", The Journal of the Acoustical Society of America **22**, 6–13.

Heinz, M. G., Zhang, X., Bruce, I. C., and Carney, L. H. (**2001**). "Auditory nerve model for predicting performance limits of normal and impaired listeners", Acoustics Research Letters Online **2**, 91–96.

Howarth, C. I. and Bulmer, M. G. (**1956**). "Non-random sequences in visual threshold experiments", Quarterly Journal of Experimental Psychology **8**, 163–171.

Isabelle, S. K. and Colburn, H. S. (**1991**). "Detection of tones in reproducible narrow-band noise.", The Journal of the Acoustical Society of America **89**, 352–359.

Jäkel, F. and Wichmann, F. A. (**2006**). "Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers", Journal of Vision **6**, 1307–1322.

Jeffress, L. A. (**1964**). "Stimulus-Oriented Approach to Detection", The Journal of the Acoustical Society of America **36**, 766–774.

Jeffress, L. A. (**1967**). "Stimulus-Oriented Approach to Detection Re-Examined", The Journal of the Acoustical Society of America **41**, 480.

Kemp, D. T. (**1978**). "Stimulated acoustic emissions from within the human auditory system", The Journal of the Acoustical Society of America **64**, 1386.

Kidd, G., Mason, C. R., Brantley, M. A., and Owen, G. A. (**1989**). "Roving-level tone-in-noise detection", The Journal of the Acoustical Society of America **86**, 1310–1317.

Kienzle, W., Franz, M. O., Schölkopf, B., and Wichmann, F. A. (**2009**). "Center-surround patterns emerge as optimal predictors for human saccade targets", Journal of Vision **9**, 7.1–15.

Kleiner, M., Brainard, D., and Pelli, D. (**2007**). "What's new in psychtoolbox-3?", in *Perception 36 ECVP Abstract Supplement*.

Koh, K., Kim, S., and Boyd, S. (**2007**). "An interior-point method for large-scale l1-regularized logistic regression", The Journal of Machine Learning Research **8**, 1519–1555.

Krantz, D. H. (**1969**). "Threshold theories of signal detection.", Psychological Review **76**, 308–324.

Kwok, J. T.-y. and Tsang, I. W.-h. (**2004**). "The pre-image problem in kernel methods.", IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council **15**, 1517–1525.

Lages, M. and Treisman, M. (**1998**). "Spatial frequency discrimination: visual long-term memory or criterion setting?", Vision Research **38**, 557–572.

le Cun, Y. (**1988**). "A theoretical framework for Back-Propagation", in *Proceedings of the 1988 Connectionist Models Summer School*, 21–28 (Pittsburgh, Pa).

Lee, S. I., Lee, H., Abbeel, P., and Ng, A. Y. (**2006**). "Efficient L1 Regularized Logistic Regression", AAAI Conference on Artificial Intelligence .

Lutfi, R. A. (**1995**). "Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks", The Journal of the Acoustical Society of America **97**, 1333–1334.

Macke, J. H. and Wichmann, F. A. (**2010**). "Estimating predictive stimulus features from psychophysical data: The decision image technique applied to human faces.", Journal of Vision **10**, 22.1–24.

Marill, T. (**1956**). "Detection theory and psychophysics", Technical Report, Massachusetts Institue of Technology, Cambdrige, MA, USA, doctoral dissertation, available as Technical Report No. 319, Research Laboratory Of Electronics, Massachusetts Institue of Technology, Cambdrige, MA, USA.

Matthew, R., Banjevic, M., Chan, A. S., Myers, L., Wolkowicz, R., Haberer, J., and Singer, J. (**2005**). "Use of the l1-norm for selection of sparse parameter sets that accurately predict drug response phenotype from viral genetic sequences", in *AMIA Annual Symposium Proceedings*, 505–509.

Murray, R. F. (**2011**). "Classification images: A review.", Journal of Vision **11**.

Nelder, J. and Wedderburn, R. (**1972**). "Generalized linear models", Journal Of The Royal Statistical Society: Series A **135**, 370–384.

Neri, P. and Levi, D. M. (**2006**). "Receptive versus perceptive fields from the reverse-correlation viewpoint", Vision Research **46**, 2465–2474.

Noble, W. S. (**2006**). "What is a support vector machine?", Nature Biotechnology **24**, 1565–1567.

Park, M. Y. and Hastie, T. (**2007**). "L1,Äêregularization path algorithm for generalized linear models", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69**, 659–677.

Pashler, H. and Yantis, S. (**2002**). *Stevens,Äô Handbook Of Experimental Psychology – Volume 1: Sensation and Perception*, 3rd edition (John Wiley & Sons, Inc., New York, USA).

Patterson, R., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1991**). *The Auditory Filter Bank* (MRC-APU Report 2341, Cambridge, England).

Pedersen, B. and Ellermeier, W. (**2008**). "Temporal weights in the level discrimination of time-varying sounds", The Journal of the Acoustical Society of America **123**, 963–972.

Peterson, W. W. and Birdsall, T. G. (**1953**). "The theory of signal detectability", Technical Report, University of Michigan, Ann Arbor, MI, USA, technical Report No. 13, Electronic Defense Group, Departement of Electrical Engineering, University of Michigan, Ann Arbor, MI, USA.

Peterson, W. W., Birdsall, T. G., and Fox, W. C. (**1954**). "The Theory of Signal Detectability", Ire Transactions On Information Theory 171–212.

Pfafflin, S. M. and Mathews, M. V. (**1966**). "Detection of auditory signals in reproducible noise", The Journal of the Acoustical Society of America **39**, 340–345.

Pièron, H. (**1914**). "Recherches sur les lois de variation des temps de latence sensorielle en fonction des intensités excitatrices", L'Année Psychologique **20**, 17–96.

Quigley, S. P. and Paul, P. V. (**1984**). *Language and Deafness* (College-Hill Press, San Diego, CA, USA).

Ratcliff, R. (**1978**). "A theory of memory retrieval.", Psychological Review **85**, 59–108.

Ratcliff, R. and Rouder, J. (**1998**). "Modeling response times for two-choice decisions", Psychological science : a journal of the American Psychological Society / APS **9**, 347–356.

Richards, V. M. (**1992**). "The detectability of a tone added to narrow bands of equal-energy noise", The Journal of the Acoustical Society of America **91**, 3424–3435.

Richards, V. M. (**2002**). "Varying feedback to evaluate detection strategies: the detection of a tone added to noise", Journal of the Association for Research in Otolaryngology : JARO **3**, 209–221.

Richards, V. M. and Buss, E. (**1996**). "Frequency correlation functions for the detection of a tone added to modulated noise maskers", The Journal of the Acoustical Society of America **99**, 1645–1652.

Richards, V. M., Heller, L. M., and Green, D. M. (**1991**). "The detection of a tone added to a narrow band of noise: The energy model revisited", The Quarterly Journal Of Experimental Psychology A, Human Experimental Psychology **43**, 481–501.

Richards, V. M. and Nekrich, R. D. (**1993**). "The incorporation of level and level-invariant cues for the detection of a tone added to noise", The Journal of the Acoustical Society of America **94**, 2560–2574.

Richards, V. M. and Tang, Z. (**2006**). "Estimates of effective frequency selectivity based on the detection of a tone added to complex maskers", The Journal of the Acoustical Society of America **119**, 1574–1584.

Richards, V. M. and Zhu, S. (**1994**). "Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients", The Journal of the Acoustical Society of America **95**, 423–434.

Ringach, D. and Shapley, R. M. (**2004**). "Reverse correlation in neurophysiology", Cognitive Science **28**, 147–166.

Rosas, P. and Wichmann, F. A. (**2011**). "Cue Combination: Beyond Optimality", in *Sensory Cue Integration*, edited by J. Trommershäuser, K. Körding, and M. S. Landy, 144–152 (Oxford University Press, Oxford, UK).

Rose, J. E., Brugge, J. F., Anderson, D. J., and Hind, J. E. (**1967**). "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey", Journal of Neurophysiology **30**, 769–793.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (**1986**). "Learning representations by back-propagating errors", Nature **323**, 533–536.

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (**2010**). "Sparse logistic regression for whole-brain classification of fMRI data", NeuroImage **51**, 752–764.

Sachs, M. B. (**1974**). "Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli", The Journal of the Acoustical Society of America **56**, 1835.

Schafer, T. H., Gales, R. S., Shewmaker, C. A., and Thompson, P. O. (**1950**). "The Frequency Selectivity of the Ear as Determined by Masking Experiments", The Journal of the Acoustical Society of America **22**, 490–496.

Schölkopf, B. and Smola, A. J. (**2001**). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 626 pages (The MIT Press, Cambridge, MA).

Schönfelder, V. H. and Wichmann, F. A. (**2012**). "Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data", The Journal of the Acoustical Society of America **131**, 3953–3969.

Sherwin, C. W., Kodman, F. J., Kovaly, J. J., Prothe, W. C., and Melrose, J. (**1956**). "Detection of Signals in Noise: A Comparison between the Human Detector and an Electronic Detector", The Journal of the Acoustical Society of America **28**, 617–622.

Shub, D. E. and Richards, V. M. (**2009**). "Psychophysical spectro-temporal receptive fields in an auditory task", Hearing research **251**, 1–9.

Snyder, R. L., Leake, P. A., Rebscher, S. J., and Beitel, R. E. (**1995**). "Temporal resolution of neurons in cat inferior colliculus to intracochlear electrical stimulation: effects of neonatal deafening and chronic stimulation.", Journal of Neurophysiology **73**, 449–467.

Swets, J. A. (**1961**). "Is there a sensory threshold?", Science (New York, NY) **134**, 168–177.

Swets, J. A., Green, D. M., and Tanner, W. P. (**1962**). "On the Width of Critical Bands", The Journal of the Acoustical Society of America **34**, 108–113.

Tanner, W. P. and Swets, J. A. (**1953**). "A new theory of visual detection", Technical Report, University of Michigan, Ann Arbor, MI, USA, technical Report No. 18, Electronic Defense Group, Departement of Electrical Engineering, University of Michigan, Ann Arbor, MI, USA.

Tanner, W. P. and Swets, J. A. (**1954**). "A decision-making theory of visual detection", Psychological Review **61**, 401–409.

Thorburn, W. M. (**1918**). "The myth of occam's razor", Mind **XXVII**, 345–353.

Tibshirani, R. (**1996**). "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **58**, 267–288.

Tikhonov, A. N. (**1963**). "On solving ill-posed problems and the method of regularization", Doklady Akademii Nauk USSR **153**, 501–504.

Treisman, M. and Williams, T. C. (**1984**). "A theory of criterion setting with an application to sequential dependencies.", Psychological Review **91**, 68–111.

Tropp, J. A. (**2006**). "Just relax: Convex programming methods for identifying sparse signals in noise", IEEE Transactions on Information Theory **52**, 1030–1051.

Vapnik, V. N. (**2000**). *The Nature of Statistical Learning Theory*, 314 pages, 2nd edition (Springer-Verlag, Inc., New York).

Verhey, J. L., Dau, T., and Kollmeier, B. (**1999**). "Within-channel cues in comodulation masking release (CMR): experiments and model predictions using a modulation-filterbank model", The Journal of the Acoustical Society of America **106**, 2733–2745.

Verhulst, P.-F. (**1838**). "Notice sur la loi que la population suit dans son accroissement", Correspondance Mathematique et Physique **10**, 113, published by A. Quetelet.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (**2011**). "Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).", Journal of personality and social psychology **100**, 426–432.

Watson, C. S. (**1964**). "Measurement of Individual Stimuli in a Signal-Detection Task", The Journal of the Acoustical Society of America **36**, 1042–1042.

Wichmann, F. A. (**1999**). "Some aspects of modelling visual perception: contrast discrimination", Ph.D. thesis, University of Oxford, UK, unpublished.

Wichmann, F. A., Graf, A., Simoncelli, E. P., and Bülthoff, H. (**2005**). "Machine learning applied to perception: Decision-images for gender classification", Advances in Neural Information Processing Systems **17**, 1489–1496.

Wichmann, F. A. and Hill, N. J. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit", Perception & Psychophysics **63**, 1293–1313.

Yeshurun, Y., Carrasco, M., and Maloney, L. T. (**2008**). "Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model", Vision Research **48**, 1837–1851.

Yovel, Y., Franz, M. O., Stilz, P., and Schnitzler, H.-U. (**2008**). "Plant classification from bat-like echolocation signals", PLoS Computational Biology **4**, e1000032.

Yovel, Y., Melcon, M., Franz, M. O., Denzinger, A., and Schnitzler, H.-U. (**2009**). "The voice of bats: How greater mouse-eared bats recognize individuals based on their echolocation calls", PLoS Computational Biology **5**, e1000400.

Zucchini, W. (**2000**). "An Introduction to Model Selection.", Journal of Mathematical Psychology **44**, 41–61.