# Probabilistic models of contextual effects in Auditory Pitch Perception

*Vincent Adam*

A dissertation submitted in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

of

**University College London**.

Gatsby Computational Neuroscience Unit

University College London

June 5, 2018

I, Vincent Adam, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

# Abstract

Perception was recognised by Helmholtz as an inferential process whereby learned expectations about the environment combine with sensory experience to give rise to percepts. Expectations are flexible, built from past experiences over multiple time-scales. What is the nature of perceptual expectations? How are they learned? How do they affect perception? These are the questions I propose to address in this thesis. I focus on two important yet simple perceptual attributes of sounds whose perception is widely regarded as effortless and automatic : pitch and frequency. In a first study, I aim to propose a definition of pitch as the solution of a computational goal. Pitch is a fundamental and salient perceptual attribute of many behaviourally important sounds including speech and music. The effortless nature of its perception has led to the search for a direct physical correlate of pitch and for mechanisms to extract pitch from peripheral neural responses. I propose instead that pitch is the outcome of a probabilistic inference of an underlying periodicity in sounds given a learned statistical prior over naturally pitch-evoking sounds, explaining in a single model a wide range of psychophysical results.

In two other psychophysical studies I study how and at what time-scales recent sensory history affects the perception of frequency shifts and pitch shifts. (1) When subjects are presented with ambiguous pitch shifts (using octave ambiguous Shepard tone pairs), I show that sensory history is used to leverage the ambiguity in a way that reflects expectations of spectro-temporal continuity of auditory scenes. (2) In delayed 2 tone frequency discrimination tasks, I explore the contraction bias : when asked to report which of two tones separated by brief silence is higher, subjects behave as though they hear the earlier tone 'contracted' in frequency towards a

combination of recently presented stimulus frequencies, and the mean of the overall distribution of tones used in the experiment. I propose that expectations - the statistical learning of the sampled stimulus distribution - are built online and combined with sensory evidence in a statistically optimal fashion.

Models derived in the thesis embody the concept of perception as unconscious inference. The results support the view that even apparently primitive acoustic percepts may derive from subtle statistical inference, suggesting that such inferential processes operate at all levels across our sensory systems.

# Impact Statement

The technical work on additive non-linear regression presented in this thesis is very general and has broad applications in many fields where statistical data analysis is important. In this thesis it was successfully applied to psychophysical and neural data analysis. I documented and will share my code with the scientific community.

I report and explain psychophysical phenomena demonstrating how humans learn the statistics of the environment. My work provides a finer quantitative evaluation of statistical learning in humans. Hence it could be used to study differences in statistical learning across populations with or without learning disorders, paving the way to possible clinical applications for example early diagnosis of statistical learning deficit. My collaborator Merav Ahissar currently studies patterns of statistical learning in people with dyslexia (a learning disorder, once thought specific to language and reading but who is now believed to be more general [1]) and people with autism (a neurodevelopmental disorder).

Some deficits in statistical learning have been reported in dyslexia and autism. My work might thus have clinical applications and these are currently being explored.

Most of the work reported in this thesis has been presented at conferences. One study was published in an international journal.

# Acknowledgements

I would like to thank my supervisor Maneesh Sahani for his guidance and help in my explorations and maybe more than anything for his trust and the freedom he has given me over the years. His sharp critical eye combined with his soft British politeness were exactly what I needed to gently tame my meanderings.

My time at the Gatsby has been one of constant learning, exchange, discussion and stimulation and I am incredibly grateful to Peter Dayan for running and maintaining such an environment. I feel really indebted to the other faculty members: (gone) Yee Whye Teh, Arthur Gretton, Peter Latham and (freshly arrived) Aapo for their teachings, time and kindness.

I am grateful to have spent many years among fellow students, postdocs and visitors, some of which I share unforgettable memories with and have become close friends: to Arthur, Heiko, Joana for many shared climbing adventures and flat whites, to Mijung, Carlos for not taking themselves too seriously, to Arne, Gergo, Elena, Eszter, Sina, Kevin, Pedro, Wittawat, Balaji, Sofy and everyone who made my time here so valuable and enjoyable.

This thesis is the outcome of a very collaborative effort. I would like to thank Merav Ahissar and Itay for hosting me in Jerusalem and Claire Chambers and Daniel Pressnitzer for their trust.

Finally, thanks to my family for supporting and encouraging me despite not always understanding my long and technical endeavours in the academic world. A last warm thanks goes to Clare for her company and support in the last year of my PhD.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction: Bayesian inference, theories of perception and contextual effects

## 1.1 Bayesian inference

A cornerstone of this thesis is the use of the framework of Bayesian inference as a scientific model of human perception and to derive and develop tools for statistical data analysis.

Bayesian inference is a formal way of reasoning under uncertainty. Consider a 'prior' belief over some unknown variable $\theta$ taking the form of a probability distribution over its possible values $p(\theta)$ and a statistical model of how $\theta$ relates to observables $x$ in the form of a conditional probability $p(x|\theta)$. Bayesian inference describe the derivation of the posterior probability on $\theta$ as new data $\mathbf{x}$ becomes available following Bayes' rule $p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)\, p(\theta)$. This posterior constitutes an 'updated' belief combining the 'evidence' provided by the new dataset and the prior.

A simple example to illustrate the problem setting and the process of inference is the following. Imagine a biased coin, that lands heads with an unknown probability $\theta$. Now you observe the outcomes of tosses and aim to infer $\theta$. A first source of uncertainty in the problem lies in the inherently stochastic process of coin tossing

itself: Outcomes of tosses (the observables) are well described through a statistical model relating them to $\theta$ : $p(x = heads|\theta) = \theta$. In a Bayesian approach, one starts with a prior $p(\theta)$, a second source of uncertainty. Updating of beliefs follows as described.

In this framework, we will be interested in the computation of two objects: posterior density $p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)\, p(\theta)$ providing the full updated belief, and marginal likelihood $p(\mathbf{x}) = \int d\theta p(\mathbf{x}|\theta)\, p(\theta)$ or how the model explains the data. We will use marginal likelihood to compare possible *a priori* distribution $p(\theta;\lambda)$ when they are themselves indexed by hyperparameters $\lambda$.

## 1.2 Theories of perception

Historically, thinking about perception has focused on two separate important stages serving different computational goals, an encoding stage and the interpretation of the encoded information. Two key theories of perception have helped to understand these two stages: efficient coding [2, 3] and perception as inference [4, 5, 6]. Both theories propose that perception is highly contextual, with environment statistics framing either the encoding of sensory information or statistical priors used during inference.

No sensory stimulus is an island. Contextual effects (CE) are pervasive in perception whether temporal, as when stimuli presented earlier affect the current percept, or instantaneous, as when a variation in 'perceptual dimensions' other than that attended to, leads to perceptual biases. Both behavioral consequences and neural correlates of CEs have been the focus of investigation.

In this section, I review the two main theories of perception and their success in explaining neural coding and behavior.

### 1.2.1 Bayesian theories of perception

The idea of perception as inference dates back to Helmholtz [4]. He proposed that perception reflects a process of unconscious inference about physical quantities of interest in the environment from imperfect, incomplete or ambiguous incoming sensory signals.

Bayesian theories of perception focus on explaining behavior in perceptual task by defining the task at hand (the inference) and specifying the optimal solution to the task using in the language of probability theory. Indeed, since sensory transduction is inherently stochastic and the dynamics of the world are complex and often uncertain, the optimal inference is a probabilistic one and requires knowledge about their assumed statistical regularities.

A guiding principle of such theories is optimality. Human performance in perceptual tasks is compared to the optimal performance achievable in these tasks, with some degrees of freedom in the definition of the optimal performance used to achieve better fit (see [7] for a philosophical and methodological discussion).

The successes of this line of investigation lies in the fact that human behavior has been repeatedly shown to be close to optimal.

For example such theories have been successful in accounting for the role of context in resolving ambiguous stimuli [8, 9], for human cue combination [10], for the use of sensory uncertainty in sensory motor learning [11], and for a wide range of sensory illusions, understood as a mismatch between sensory evidence and expectations [12].

#### 1.2.1.1 Perceptual bias towards the prior

Formally, given a statistical generative model of sensation $x$ conditional on an unknown stimulus $y$ : $p(x|y)$ and a prior assumption on this stimulus $p(y)$, Bayesian theories describe perception as an outcome of the posterior computation $p(y|x) \propto p(x|y)\,p(y)$, for example the computation of its mean or median when subjects are asked to report the value of $y$, and the posterior variance when asked to report their degree of confidence in their judgement.

If the prior was non-informative $p(y) \propto 1$, the posterior would be proportional to the likelihood. When informative it generally leads to attractive perceptual biases: the posterior is 'shifted' toward the prior a sense that, according to some metric $d$ between distributions, the distance between the posterior and the prior is reduced compared to the distance from the normalized likelihood and the prior: $d(\text{prior}, \text{posterior}) < d(\text{prior}, \text{likelihood})$. Different metrics can be used such as dis-

tance in means, medians or modes (when unimodal), or the Kullback-Leibler divergence. For distributions in exponential family and the KL divergence as a metric, one can prove that the posterior is always shifted to the prior (although it might not be the case for the same distributions but using another metric). Examples of distributions and metrics where repulsion rather than attraction is obtained as a result of posterior inference are given in [13].

The Gaussian case is a stereotypical example. Assuming a generative model for $x$ where additive Gaussian noise with variance $\sigma_l^2$ corrupts the stimulus $y$ : $p(x|y) \propto \mathcal{N}\left(x; y, \sigma_l^2\right)$ and a Gaussian prior with variance $\sigma_p^2$ on the stimulus value: $p(y) \propto \mathcal{N}\left(y; \mu_p, \sigma_p^2\right)$ leads to a Gaussian posterior $p(y|x) = \mathcal{N}\left(x_i; \mu_{post}, \sigma_{post}^2\right)$ with reduced uncertainty $\sigma_{post}^2 = \left(\sigma_l^{-2} + \sigma_p^{-2}\right)^{-1} < \min\left(\sigma_l^2, \sigma_p^2\right)$ and precision weighted mean $\mu_{post} = \sigma_{post}^2 \left(\frac{x}{\sigma_l^2} + \frac{\mu_p}{\sigma_p^2}\right)$. For univariate Gaussian distributions, the precision refers to the inverse of the variance.

#### 1.2.1.2 Sensory Cue Combination

A success of Bayesian theories of perception is their ability to describe how humans combine different sources of information. In a series of work summarized in [10], it was shown that subject's behavior in tasks involving the combination of mulitple sensory cues could be well described as an inference process sensitive to the reliability of the sources to be combined. Although alternative explanations exists (for example [14]), ideal observer approaches have been useful in quantitatively describing this process.

Within this framework, a possible formulation of a cue combination task is the following: given a statistical generative model of two sensations $x_1, x_2$ conditionally independent given a shared unknown stimulus $y$: $[p(x_1, x_2|y) = p(x_1|y)\, p(x_2|y)]$ and an uninformative prior assumption on this stimulus $[p(y) \propto 1.]$, Bayes rules leads to a posterior density on the stimulus $p(y|x_1, x_2) \propto p(x_1|y)\, p(x_2|y)$. Assuming a Gaussian uncertainty for both sources $p(x_i|y) \propto \mathcal{N}\left(x_i; y, \sigma_i^2\right)$ leads to a Gaussian posterior $p(y|x_1, x_2) = \mathcal{N}\left(x_i; \mu_{post}, \sigma_{post}^2\right)$ with reduced uncertainty $\sigma_{post}^2 = \left(\sigma_1^{-2} + \sigma_2^{-2}\right)^{-1} < \min\left(\sigma_1^2, \sigma_2^2\right)$ and precision weighted mean $\mu_{post} = \sigma_{post}^2 \left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}\right)$.

Another interesting case of cue combination is when two underlying perceptual features $y_1, y_2$ are jointly assumed to give rise to a sensation $x$ through a generative model $p(x|y_1, y_2)$ and when these are assumed to be coupled with a joint density $p(y_1, y_2)$. The joint posterior is computed as $p(y_1, y_2|x) \propto p(x|y_1, y_2)\, p(y_1, y_2)$. If the information provided by $x$ about the first dimension was ambiguous assuming no coupling in the prior, the coupling may help to resolve this ambiguity. Hehrmann and Sahani explained timbral induced biases in the perception of ambiguous pitch using such a model where underlying pitch and timbre features parameterize a generative model of pitch evoking sounds, and priors derived from natural sound statistics encode a natural pitch-brightness correlation [15].

#### 1.2.1.3 Statistical learning in Bayesian theories of perception

The Bayesian framework for perception relies on the ability to learn statistical regularities of the environment. Our ability to acquire knowledge about the distribution of stimuli has been demonstrated repeatedly [16]. Several studies suggest that subjects may behave optimally: their priors match the task or natural statistics [17, 18, 19]. Short term distribution learning has also been reported. Chalk et al. [20] showed that expectations of simple stimulus features, e.g. direction of motion, can be developed implicitly through a fast statistical learning procedure. Moreover, it has been shown that prior distributions that are more complex than a simple Gaussian, such as skewed or bimodal, can be learned, in a relatively short amount of time [21, 11, 22].

### 1.2.2 Efficient Coding hypothesis

The efficient coding hypothesis provides a theoretical framework to understand the encoding stage of perception. It proposes that sensory systems have evolved to maximize the information transmitted to the brain about the environment, subject to the limitations of the resources available [2, 3]. Formally, denoting by $r(s)$ the possibly noisy encoding of a stimulus $s$ whose natural statistics are $p(s)$, a measure of its efficiency is the mutual information $I(s, r(s))$ which quantifies in expectation the transmitted information. Optimal codes are codes that maximize this objective

[23].

Optimal encoding theories have been helpful to understand many aspects of neural encoding. In the visual domain, it can explain codes in the retina [24] and the linear part of the structure of codes in early cortex [25]. In the auditory system it has been used to explain encoding in thalamic relays [26] and in early auditory cortical areas [27].

The efficient coding hypothesis makes relatively few predictions at a behavioral level. Apart from a few recent attempts [13, 28], it generally fails to account from most perceptual biases reported, which I review in the next section.

## 1.3 Contextual effects in perception

Sensory experience affects perception on timescales ranging from a few milliseconds to a life-time. Here, I focus on contextual effects on short timescales from a few milliseconds to tens of minutes, which includes the typical duration of psychophysical experiments or the time it takes to read a newspaper article. Sensory experience has two main effects on perception. First, it can enhance perceptual performance, as when it increases accuracy in a detection task [29]. Second, it can lead to perceptual biases in the sense that it shifts the percept reported from what it would have been had there been no context.

### 1.3.1 Attractive and Constrastive biases

In this section, I focus on on *temporal* contextual effects which describe the influence of past stimuli on current perception. In psychophysics, contextual effects are frequently classified according to whether the observed perceptual bias is attractive or contrastive. Contextual effects are labelled attractive when perception seems biased towards previous percepts or stimuli, as when a brief motion stimulus biases a subsequent ambiguous stimulus towards the same perceived direction of motion [30]. On the other hand, temporal contextual effects are labelled contrastive when perception is biased away from recent sensory history, as when the prolonged presentation of a motion stimulus in one direction creates the illusory perception of motion in the opposite direction for a subsequent stationary stimulus [31]. This

classification is not always obvious and some contextual effects might be both, as in the case of the biasing of ambiguous visual-motion [30] where it has been observed that attractive effects occurred for short time scales, whereas contrastive effects occurred for longer time scales.

### 1.3.2 Opposing explanations

These opposite effects have been explained based on different stages of perception. Attractive effects are often explained in the normative Bayesian framework as the biasing effect of prior information in probabilistic inference.

Prior information in this framework is specified as a distribution over the possible configurations of objects assumed to exist in the world. In probabilistic inference, information from the senses leads to an update of this distribution over configurations. The result of this update is an 'a posteriori' distribution that often favors 'a priori' likely configurations over less-likely ones even if both explain sensation equally well, hence the bias. A concrete example is given in 1.2.1.1.

These biases arise as a consequence of a computation whose other consequences are to reduce perceptual uncertainty (the range possible configurations can only shrink), to disambiguate ambiguous stimuli (by favoring some configuration over others) and to help stabilize perception [32].

Contrastive effects on the other hand are often described as the result of a mechanistic adaptation or habituation of the encoding stage, with a fixed decoding stage [33, 34] even in attempts to frame adaptation in the Bayesian framework [35]. Computationally, contrastive effects are thought to reflect the task of maximizing information transfer [36]. A recent attempt to merge efficient coding and Bayesian theories of perception in a single theory explains the contrastive and attractive effect of long term statistics in visual tasks [13]. This theory however does not explain how and on what timescales these statistics are learned.

### 1.3.3 The case of auditory low-level perception

In the auditory modality, contrastive context effects have been reported for basic low-level auditory features such as loudness or pitch. The subjective location of a

sound can be shifted away from that of a preceding context [37, 38]. The prolonged presentation of amplitude-modulation can elevate subsequent modulation detection thresholds [39]. Adaptation to short frequency glides affects subsequent temporal order judgement of brief tone pairs [40]. The prolonged presentation of frequency-shifts in spectral peaks or troughs produces a negative "afterimage" on the spectral motion of subsequent similar sounds [41]. More recently, it has been demonstrated that prolonged exposure was not always necessary: even very brief contexts were able to shift away the perception of spectral motion [42, 43].

There are comparatively fewer instances of attractive contextual effects in auditory perception. For frequency and pitch, a regression to the mean (or contraction bias, reviewed in more detail in Chapter 5) has been reported for successive pitch judgements [44]. When ambiguity is added to pitch judgements, hysteresis has been observed, one of the hallmarks of attractive effects [45, 46]. In auditory scene analysis, finally, the perceptual organization of ambiguous tone sequences is biased towards prior percepts [9]. Other contextual effects in auditory scene analysis may also be categorized as attractive, such as when a component tone is captured by a preceding context [47].

## 1.4 Modelling contextual effects in Perception: three case studies

In the following chapters of this thesis, I present models and interpretations of three sets of psychophysical studies involving 2 alternative forced choice (2AFC) discrimination tasks. In all studies, subjects' responses reveal strong attractive biases. Two different kinds of stimuli are used to reveal the bias.

In chapter 3, extending a model of human pitch perception by Hehrmann and Sahani [15], a psychophysical experiment was run in which subjects had to judge the direction of pitch shift between stimuli. The shift was designed to be ambiguous. In this study *instantaneous* timbral context had a biasing effect which we model as reflecting learned timbre-pitch statistical dependence observed in natural sounds.

In chapter 4, I report a second study by Chambers et al [48] in which context

also disambiguates an ambiguous pitch shift percept. Subjects had to report the direction of a half octave pitch shift, clear in magnitude (high signal to noise ratio) but ambiguous in direction. A model of pre-perceptual tracking of auditory scene is proposed to account for the reported effect.

In chapter 5, I report a third study by Lieder et al [49], in which subjects have to judge whether a frequency shift between a pair of pure tones is upward or downward. Stimuli are not ambiguous and contextual effects are most salient in more difficult trials (low signal to noise ratio). I analyzed the 'contraction bias' revealed by this task and proposed a model to account for the biasing effect of sensory history at multiple time-scales.

In chapter 2, prior to the presentation of these psychophysical studies and their modelling, I describe Gaussian Processes and recent technical innovations in their use for non-linear regression which I extend and use across multiple chapters.

## 1.5 Modelling methodology

David Marr distinguishes three complementary levels at which information processing systems (such as the visual or auditory system) can be described and studied [50]: (1) the computational level is primarily concerned with identifying the goal or purpose of the system under study, and the strategy employed to achieve it. (2) the algorithmic and representational level focuses on studying what algorithms underlie the system's input-output transformation in order to achieve its goal, and the nature of their internal representation. (3) the implementation level explains the mechanisms by which these algorithms and representation are realised in the actual, physical system under study. In this thesis, I report behavioral phenomena about the perception of basic auditory features whose perception is effortless and automatic: pitch and frequency shifts. Most modelling approaches to the perception of these features are mechanistic models and start with the assumption that correlates of percepts can readily be computed from peripheral neural responses at early stages of the auditory periphery. This tendency to favor explanations at an implementational level is not well supported by physiological evidence. For example, a physiological

correlate of the percept of pitch is yet to be found. In all my three projects, I follow computational principles instead: I developed computational models to account for these psychophysical phenomena all of which implement the hypothesis of perception as unconscious Bayesian inference. A common feature of these models is the assumption of a fixed noisy log-linear encoding of the frequency content of sounds as displayed in early auditory periphery [51] and as predicted by an efficient encoding based on natural sounds [27].

## 1.6 Summary of publications

Most of the work reported in this thesis has been published in international conferences and journals. My technical work on sparse additive Gaussian Process regression described in chapter 2 was presented and published in the proceedings of the IEEE International Workshop on Machine Learning for Signal Processing [52]. Its application to neural data analysis is to was presented at the 2017 edition of COSYNE [53]. My modelling of how context disambiguates the perception of ambiguous pitch shifts described in chapter 4 was presented at the 2016 edition of COSYNE [54] and was published in the journal Nature Communication [48]. My modelling of how context biases delayed 2 tones discrimination tasks described in chapter 5 was presented at the 2017 edition of COSYNE [49]. My extension of the model of human pitch perception developed by Hehrmann et al. described in chapter 3 was presented at the 2014 edition of COSYNE [55].

# Chapter 2

# Gaussian Processes and approximate inference

## 2.1 Introduction

A cornerstone of this thesis is the use of the framework of Bayesian inference as a scientific model of human perception and to derive and develop tools for statistical data analysis. In both cases, the need to describe rich, flexible and structured prior assumptions on functions has led me to use Gaussian Processes [56]. Inference using Gaussian Processes is often hard and prohibitively expensive when applied to large datasets and one needs to resort to approximations.

This chapter is organized as follows: I first review Gaussian Processes and their application to regression. I carry on describing approximations allowing them to scale to problems involving large datasets. Finally I present the problem of additive regression using Gaussian Processes and propose a new algorithm to solve it efficiently. Along the way, I describe applications of this algorithm and its use in this thesis.

## 2.2 Gaussian Processes

### 2.2.1 Definition

Gaussian Processes (GPs) are infinite collections of random variables, any finite subset of which follows a multivariate Gaussian distribution. They are defined by

a mean function $m$ and covariance function $k$. A draw from a GP defined on a index set $\mathscr{X}$ is a function on the domain $\mathscr{X}$. Given a list of points $X \in \mathscr{X}^N$ and a function draw $f \sim GP(m,k)$, the vector of function evaluations $\mathbf{f}(X)$ is a associated multivariate normal random variable such that $\mathbf{f}(X) \sim \mathscr{N}(\mathbf{m}(X), \mathbf{K}(X,X))$, where $\mathbf{m}$ is a vector of mean function evaluations and $\mathbf{K}$ is a matrix of covariance function values.

In this thesis, I will focus on the case $m = 0$ leaving the GP fully specified by the covariance function only. Also, in all applications, the input space is the one-dimensional real line ($\mathscr{X} = \mathbb{R}$)

### 2.2.2 Covariance functions

The covariance function (or kernel) of a GP defines the covariance between all pairs of function evaluations on $\mathscr{X}$. This covariance function is symmetric and positive-definite and captures a notion of similarity or nearness between pairs of function evaluations. For a pair of points $x, x' \in \mathscr{X}$, and $f \sim GP(0,k)$ the covariance between the two associated function evaluations is $cov(f(x), f(x')) = k(x,x')$. Sums of kernels (with positive weights) are kernels and so are finite products of kernels; this will be important for constructing new kernels from pre-existing ones in the following sections.

A covariance function $k(x,x')$ is said to be stationary if it is a function of $\tau = x - x'$, that is if is invariant to translations in the input space $\mathscr{X}$. The stationary covariances with domain $\mathbb{R}$ that we will consider have a spectral density $S(s)$ defined as its Fourier transform: $k(\tau) = \mathscr{F}^{-1}\{S\}(\tau) = \int_{\mathbb{R}} ds\, S(s) e^{2\pi i s \tau}$. The spectral density is the expected power spectrum, which is a familiar object in sound analysis.

In the next sections, I present some kernels capturing the prior assumptions made at various points in this thesis: smoothness and local periodicity.

#### 2.2.2.1 Smoothness

The Exponential Quadratic (EQ) kernel also called the Gaussian kernel has the form $k_{EQ}(\tau) = \exp\left(-\frac{\tau^2}{2l^2}\right)$. It is parameterized by $l$, the characteristic length-scale. Samples from a GP with a EQ kernel are smooth on a scale controlled by

$l$. This can be seen from the spectral density decaying with frequency as $S_{EQ}(s) \propto \exp\left(-2\pi^2 l^2 s^2\right)$. The limit of $l \to 0$, leads to the noise kernel $k_{noise}(\tau) = \delta(\tau = 0)$. The other limit $l \to \infty$ corresponds to the constant kernel $k_{const}(\tau) \propto 1$ (samples of which are constant functions).

#### 2.2.2.2 Periodicity

A function $f$ with period $\Omega$ is such that $\forall t, f(t+\Omega) = f(t)$. A stationary kernel $k$ defining a periodic GP must itself be periodic. This imposes equality (correlation equal to 1) between function evaluations at period spaced input points. A simple periodic kernel is the cosine $k_{cos}(\tau) = \cos(2\pi\tau\mu)$. Its spectral density is sparse and given by $S_{cos}(s) \propto \delta(s = \pm\mu)$. Samples from an associated GP are themselves cosine functions (with random amplitude and phase). Finite sums of harmonically related cosine kernels $k_{har}(\tau) = \sum_{j=1}^{J} w_j \cos(2\pi\tau\mu j)$ , $w_j > 0$ are periodic with spectral density $S_{har}(s) \propto \sum_{j=1}^{J} w_j \delta(s = \pm j\mu)$ . A classical periodic kernel inspired by the EQ kernel is $k_{per}(\tau) = \exp\left(-2\frac{sin^2(\pi\tau\mu)}{\gamma^2}\right)$. It has no closed form spectral density, although $\gamma$ controls the smoothness in a similar manner as $l$ in the EQ kernel: a EQ kernel with parameter $l$ and a periodic kernel with parameter $\gamma \approx \pi l\mu$ have equivalent smoothnesses.

#### 2.2.2.3 Almost periodic kernels

Relaxing the assumption of periodicity means relaxing the equality constraint of period-spaced function evaluations. Starting from a periodic kernel $k$, one may add noise to its sampled functions. In the case of additive independent Gaussian noise with mean 0 and variance $\sigma^2$, this corresponds to consider the kernel $\tilde{k}(\tau) = k(\tau) + \sigma^2\delta(\tau = 0)$. Another way of enforcing a more local periodicity is to weaken the correlation between function evaluation in a distance dependent manner, e.g. using a decaying kernel $h$ to build $\tilde{k}(\tau) = k(\tau)h(\tau)$.

## 2.3 Gaussian Processes for regression

In this section I briefly discuss the problem of Gaussian Process Regression and present sparse approximations to the problem, with a focus on the sparse variational approximation introduced by [57].

### 2.3.1 Setting

In GP regression, consider a data set $\mathscr{D} = \{x_i, y_i\}_{i=\{1..N\}}$, an observation model $y_i|f(x_i)$, and a GP prior on $f$. The aim is to compute the posterior $p(\mathbf{f}|\mathscr{D}) \propto p(\mathbf{f})\, p(\mathbf{y}|\mathbf{f})$ and the marginal likelihood $p(\mathbf{y}) = \int d\mathbf{f}\, p(\mathbf{y},\mathbf{f})$. Both the prior GP and the likelihood might have unknown parameters which may be selected by maximizing the marginal likelihood.

Two main difficulties arise when trying to compute the posterior. First, even in the simple case of conjugate likelihood (or Gaussian observation model) $y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathscr{N}\left(0, \sigma^2\right)$, where the marginal likelihood has an analytic form: $p(\mathbf{y}|\mathscr{D}) = \mathscr{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{nn} + \sigma^2\mathbf{I}\right)$ with $\mathbf{K}_{nn} = \mathbf{K}(X,X)$, computations require the expensive inversion of an $N \times N$ matrix. Second, when the likelihood is not conjugate, posterior estimates are not available in closed form and must be approximated, for example by expectation propagation [58] or variational inference [59]. Such approximations do not scale well with $N$.

### 2.3.2 Sparse approximations

Sparse approximations represent an approach for overcoming the aforementioned difficulties with GP regression, providing an attractive framework in settings with large datasets or non-conjugate likelihoods. A full review of sparse approximations is beyond the scope of this thesis, see [60, 61] for a review. Briefly, these approximations explicitly represent $m$ additional function evaluations of a GP $\mathbf{u}$ associated to pseudo-inputs $Z \in \mathscr{X}^m$ forming a prior $p\left(f_{\neq\mathbf{u}}, \mathbf{u}\right)$, where $f_{\neq\mathbf{u}}$ represents the function evaluations on $\mathscr{X}\backslash Z$, and modify this extended prior by introducing conditional independencies, for example $\tilde{p}(\mathbf{f},\mathbf{u}) = \prod_i p(f_i|\mathbf{u})p(\mathbf{u}) \neq p(\mathbf{f},\mathbf{u})$. Finally $\mathbf{u}$ are treated as parameters yielding a new parametric prior on $\mathbf{f}$. Denoting $\mathbf{K}_{mm} = \mathbf{K}(Z,Z)$, $\mathbf{K}_{nm} = \mathbf{K}(X,Z)$ and $\mathbf{K}_{mn} = \mathbf{K}(Z,X)$, this intuitively induces a low rank form to the matrix $\mathbf{K}_{nn} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$, with the rank governed by the size of the inducing set. An inconvenience of such approaches is that the prior is changed.

Another approach is to treat the inducing points as variational parameters in a variational framework by assuming the following form to approximate the posterior $p\left(f_{\neq\mathbf{u}}, \mathbf{u}|\mathbf{y}\right) \approx p\left(f_{\neq\mathbf{u}}|\mathbf{u}\right) q(\mathbf{u})$. This has been proposed by [57] for the conjugate

case and extended to the non-conjugate case by [62]. This approximation does not change the prior model and has appealing theoretical justifications [63].

In this variational framework, a lower bound on the log marginal likelihood is achieved using Jensen's inequality:

$$\begin{aligned}
\log p(\mathbf{y}) &\geq \mathbb{E}_{q(\mathbf{u},\mathbf{f})}\left[\log \frac{p(\mathbf{y}|\mathbf{u},\mathbf{f})\,p(\mathbf{f}|\mathbf{u})\,p(\mathbf{u})}{q(\mathbf{u},\mathbf{f})}\right] \\
&\geq \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})\,p(\mathbf{f}|\mathbf{u})\,p(\mathbf{u})}{p(\mathbf{f}|\mathbf{u})\,q(\mathbf{u})}\right] \\
&\geq \mathbb{E}_{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})}\left[\log \frac{p(\mathbf{y}|\mathbf{f})\,p(\mathbf{u})}{q(\mathbf{u})}\right] \\
&\geq \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y}|\mathbf{f})\right] + \underbrace{\mathbb{E}_{q(\mathbf{u})}\left[\log \frac{p(\mathbf{u})}{q(\mathbf{u})}\right]}_{-KL(q(\mathbf{u})||p(\mathbf{u}))} = \mathscr{L}(q) \qquad (2.1)
\end{aligned}$$

where $q(\mathbf{f}) = \int d\mathbf{u} q(\mathbf{u})\, p(\mathbf{f}|\mathbf{u})$, and $KL(q||p)$ is the Kullback-Leibler divergence between densities $q$ and $p$.

The left hand term of the bound is the expected log likelihood under the approximated posterior. The right hand term is the negative of a measure of nearness between the prior and the posterior distribution on the function evaluation at the inducing points. The density $q^*(\mathbf{u})$ maximizing this bound is $q^*(\mathbf{u}) \propto p(\mathbf{u}) \exp\left\{\mathbb{E}_{p(\mathbf{f}|\mathbf{u})} \log p(\mathbf{y}|\mathbf{f})\right\}$. It is intractable for most (non-conjugate) likelihoods. A further approximation followed by [62] is to restrict $q$ to be a multivariate Gaussian distribution $q(\mathbf{u}) = \mathscr{N}(\mathbf{u}|\mathbf{m},\mathbf{S})$ and to optimize for variational parameters $\mathbf{m},\mathbf{S}$. This leads $q(\mathbf{f})$ to be also Gaussian: denoting $\mathbf{A} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}$, we have $\mathbf{f}|\mathbf{u} \sim \mathscr{N}\left(\mathbf{f}; \mathbf{A}\mathbf{u}, \mathbf{K}_{nn} - \mathbf{A}\mathbf{K}_{mm}\mathbf{A}^T\right)$ so $q(\mathbf{f}) = \mathbb{E}_{q(\mathbf{u})}\left[p(\mathbf{f}|\mathbf{U})\right] = \mathscr{N}\left(\mathbf{f}; \mathbf{A}\mathbf{m}, \mathbf{K}_{nn} + \mathbf{A}(\mathbf{S} - \mathbf{K}_{mm})\mathbf{A}^T\right)$. When observations are assumed to be independent, this enables fast and accurate approximations to the expected likelihood (left hand term of the bound in Equation 2.1) using Gaussian Quadrature methods.

### 2.3.3 Application: inferring pitch from auditory nerve activity.

This example will be described in depth in Chapter 3. In our model of human pitch perception, we have a generative model of auditory nerve activity $\mathbf{A}$ given a sound waveform $\mathbf{x}$. This provides us with a likelihood function $p(\mathbf{A}|x)$. We use a GP prior for the sound $p(\mathbf{x};\Omega)$ with $\Omega$ the period as a hyperparameter. This is a prototypical

regression setting with a rather complex likelihood. We are interested in finding the period maximizing the marginal likelihood $p(\mathbf{A};\Omega)$. Instead, we maximize the lower bound on $\log p(\mathbf{A};\Omega)$ using the described sparse variational approximation.

## 2.4 Additive regression

In this section, I present my work published in [52] where I extend the sparse variational approach to a class of regression models involving multiple latent unknown functions $f^{(1)},...,f^{(D)}$ on $D$ different dimensions $\mathscr{X}^{(1)},...,\mathscr{X}^{(D)}$, but where observations are conditionally independent given the sum of the functions, i.e $p\left(y|f^{(1)},...,f^{(D)}\right) = p\left(y|\sum_d f^{(d)}\right)$. The term $\sum_d f^{(d)}$ constitutes an additive predictor. This class of model is that of Generalized Additive Models (GAMs) and was introduced in [64] as a non parametric additive extension to Generalized Linear Models (GLM). The key motivation for the additivity is interpretability at the cost of generality (dimensions do not interact in the additive predictor). Another motivation is to avoid the curse of dimensionality: regression in higher dimensions requires increasingly more data or increasingly strong assumptions leading to poor or biased estimates.

### 2.4.1 Additive Gaussian Process regression

I now consider the case where $\mathscr{X} = \mathbb{R}^D$ for $D > 1$, and write $x = (x^{(1)},...,x^{(D)})$ for $x \in \mathscr{X}$. In the GP regression framework, using additive kernels [65, 66] imposes this desired structure to the prior. For example if $k(x,x') = \sum_d k^{(d)}(x^{(d)},x'^{(d)})$, the associated GP constrains the prior function to have additive structure $f(x) = \sum_d f^{(d)}(x^{(d)})$ where $f^{(d)} \sim GP(0,k^{(d)})$. A first approach to doing additive GP regression could be to carry out a classical single GP regression under additive kernel assumption and to use a sparse approximation.

In the variational inducing point framework, the posterior is a low dimensional process given by $p(\mathbf{f}|\mathbf{y}) \approx q(\mathbf{f}) = \int d\mathbf{u}\, p(\mathbf{f}|\mathbf{u})\, q(\mathbf{u})$. In the case where

$q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$, the posterior is a GP with non additive covariance structure

$$\begin{aligned} q(\mathbf{f}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\mu} &= \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{m} \\ \boldsymbol{\Sigma} &= \mathbf{K}_{nn} + \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\left[\mathbf{S}\mathbf{K}_{mm}^{-1} - \mathbf{I}\right]\mathbf{K}_{mn} \end{aligned}$$

Recovering the marginals requires the extra step of computing the joint multivariate posterior GP over the joint $q\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}\right) = \frac{q(\mathbf{f})}{p(\mathbf{f})} p\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}\right)$ and marginalizing. Denoting $q\left(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(d)}\right) = \mathcal{N}(\mathbf{v}, \mathbf{V})$ and indexing dimensions such that $q(\mathbf{f}^{(d)}) = \mathcal{N}(\mathbf{v}^{(d)}, \mathbf{V}^{(d)})$ and $cov(\mathbf{f}^{(d)}, \mathbf{f}^{(d')}) = \mathbf{V}^{(d,d')}$, we have

$$\begin{aligned} \mathbf{V}^{(d,d')} &= \mathbf{K}^{(d)} - \mathbf{K}^{(d)}\tilde{\mathbf{K}}^{-1}\mathbf{K}^{(d')} \\ \mathbf{v}^{(d)} &= \left(\mathbf{K}^{(d)} - \mathbf{K}^{(d)}\widetilde{\mathbf{K}}^{-1}\mathbf{K}_{sum}\right)\sigma^{-2}\mathbf{y} \end{aligned}$$

where $\mathbf{K}_{sum} = \sum_{d=1}^{D} \mathbf{K}^{(d)}$ and $\widetilde{\mathbf{K}} = \left[\boldsymbol{\Sigma}^{-1} - \mathbf{K}_{sum}^{-1}\right]^{-1} + \mathbf{K}_{sum}$.

This step is needed because the inducing points do not readily provide the information to reconstruct the individual components (they were indeed optimized to reconstruct the summed GP). One cannot compute this posterior joint for individual input points as this would split the mean and variance according to the relative prior variances of each underlying function at that point, rather than correctly taking into account the rest of both the prior and posterior GP structure. Instead the joint must be constructed over a large set of points (ideally the full initial dataset) and those data points should not share any coordinate. This final step is computationally expensive

### 2.4.2 Sparse approximation

Here I propose to extend the variational inducing point framework for additive GP models. The resulting algorithm was published in [52] and was used for statistical data analysis in chapter 5.

My extension involves directly constructing a parametric approximation to each GP. I do so by assuming a factorized approximation to the joint posterior $p(\mathbf{f}^{(1)}, ..., \mathbf{f}^{(D)}, \mathbf{u}^{(1)}, ..., \mathbf{u}^{(D)}|Y) \approx \prod_d q(\mathbf{u}^{(d)}) p(\mathbf{f}^{(d)}|\mathbf{u}^{(d)})$. A similar mean-field approach has been proposed independently by Saul et al [67] to allow the non-linear

combination of an arbitrary collection of GPs. However, they did not recognize the computational advantage induced by the additive structure.

The proposed approach provides some advantages over the previous studies. Each GP is conditioned on its own set of inducing points, allowing one to use fewer inducing points for less complex or lower-dimensional predictors (linear functions are well approximated by using linear kernels with two inducing points).

Inducing variables are readily interpretable as conditional variables for the prediction of each function. This comes however at the cost of losing the correlation structure across functions. Here, I use a mean field approximation that systematically underestimates the covariance of the individual components. Linear response methods may however be used to recover more complete estimates of the full covariance structure [68].

To simplify notation, I will write $\mathbf{F} = [\mathbf{f}^{(1)}, ..., \mathbf{f}^{(D)}]$ and $\mathbf{U} = [\mathbf{u}^{(1)}, ..., \mathbf{u}^{(D)}]$. I will consider an augmented model $p(\mathbf{F}, \mathbf{U}) = \prod_d p(\mathbf{f}^{(d)}, \mathbf{u}^{(d)})$ where $\mathbf{u}^{(d)}$ are associated to pseudo-input $Z^{(d)}$.

I will consider an approximation to the posterior of the form $p(\mathbf{F}, \mathbf{U}|\mathbf{y}) \approx \prod_d p(\mathbf{f}^{(d)}|\mathbf{u}^{(d)}) q(\mathbf{u}^{(d)})$.

The Titsias assumption [57] applied to the additive model leads to a lower bound on the marginal log likelihood

$$\begin{aligned} \log p(\mathbf{y}) &\geq \log \mathbb{E}_{p(\mathbf{F}|\mathbf{U})q(\mathbf{U})} \left[ \frac{p(\mathbf{y}|\mathbf{F},\mathbf{U}) p(\mathbf{F}|\mathbf{U}) p(\mathbf{U})}{p(\mathbf{F}|\mathbf{U}) q(\mathbf{U})} \right] \\ &\geq \mathbb{E}_{p(\mathbf{F}|\mathbf{U})q(\mathbf{U})} \left[ \log p(\mathbf{y}|\mathbf{F}) \right] - \mathtt{KL}(q) \\ &\geq \mathbb{E}_{q(\rho)} \left[ \log p(\mathbf{y}|\rho) \right] - \mathtt{KL}(q) \end{aligned} \tag{2.2}$$

with $\rho = \sum_d f^{(d)}$, the additive predictor and $\mathtt{KL}(q) = \sum_d KL\left( q(\mathbf{u}^{(d)}) || p(\mathbf{u}^{(d)}) \right)$. The predictive posterior is $q(\rho) = \mathbb{E}_{\prod_d q(\mathbf{u}^{(d)})} \left[ p(\rho|\mathbf{U}) \right]$.

### 2.4.2.1 Parametric assumptions for tractable bound

The posterior over the function evaluations at the inducing points $\mathbf{u}^{(d)}$ are approximated as multivariate Gaussian distributions: $q(\mathbf{u}^{(d)}) = \mathcal{N}(\mathbf{u}^{(d)}|\mathbf{m}^{(d)}, \mathbf{S}^{(d)})$

This leads $q(\rho_i)$ to take a univariate Gaussian form and enables fast and ac-

curate approximations to the expectations (left hand term of the bound in Equation 2.2) using Gaussian Quadrature methods. Writing $\mathbf{A}^{(d)} = \mathbf{K}_{nm}^{(d)}\mathbf{K}_{mm}^{(d)-1}$, I have

$$\mathbf{f}^{(d)}|\mathbf{u}^{(d)} \sim \mathcal{N}\left(\mathbf{f}^{(d)}; \mathbf{A}^{(d)}\mathbf{u}^{(d)}, \mathbf{K}_{nn}^{(d)} - \mathbf{A}^{(d)}\mathbf{K}_{mm}^{(d)}\mathbf{A}^{(d)T}\right)$$

$$\begin{aligned}\boldsymbol{\rho}|\mathbf{U} &= \sum_d \mathbf{f}^{(d)}|\mathbf{U} \\ &\sim \mathcal{N}\left(\boldsymbol{\rho}; \sum_d \mathbf{A}^{(d)}\mathbf{u}^{(d)}, \sum_d \mathbf{K}_{nn}^{(d)} - \mathbf{A}^{(d)}\mathbf{K}_{mm}^{(d)}\mathbf{A}^{(d)T}\right)\end{aligned}$$

so

$$\begin{aligned}q(\boldsymbol{\rho}) &= \mathbb{E}_{\prod_d q(\mathbf{u}^{(d)})}\left[p(\boldsymbol{\rho}|\mathbf{U})\right] \\ &= \mathcal{N}\left(\boldsymbol{\rho}; \sum \mu_{add}^{(d)}, \sum_d \Sigma_{add}^{(d)}\right)\end{aligned}$$

With

$$\begin{aligned}\boldsymbol{\mu}_{add}^{(d)} &= \mathbf{A}^{(d)}\mathbf{m}^{(d)} \\ \boldsymbol{\Sigma}_{add}^{(d)} &= \mathbf{K}_{nn}^{(d)} + \mathbf{A}^{(d)}\left(\mathbf{S}^{(d)} - \mathbf{K}_{mm}\right)\mathbf{A}^{(d)T}\end{aligned}$$

### 2.4.3 Applications to psychophysical data analysis

In Chapter 5, we use a non-linear extension of classical probit regression models used to model subject performance in discrimination tasks. Subjects were asked to report which of 2 pure tones presented successively was higher in frequency, in a task consisting of hundreds of such comparisons. Denoting by $\delta$ the frequency difference between the first and second tone in a trial, a classical model of subject responses is $p('f_1 \text{ higher}'|\delta) = \phi(\alpha\delta)$, where $\alpha$ is a free parameter that can be related to the subject's precision. In such models, only the two tones to be discriminated play a role in the decision (through their difference $\delta$). Instead I aimed to study the influence of stimuli presented in past trials on subjects' decisions. I constructed covariates from past trials $\mathbf{z}$ and extended the response model in an additive manner: $p('f_1 \text{ higher}'|\delta, \mathbf{z}) = \phi(\alpha\delta + \sum_d f_d(z_d))$, with $f_d$ the unknown functions to be learned.

### 2.4.4 Application to neural data analysis: Gaussian Process Factor Analysis (GPFA)

A natural extension of the additive framework is to further allow the weighting of individual functions to vary across observations. A recent analysis method for neural data analysis, GPFA, corresponds to this extension and was developed at the Gatsby Unit. Efficient approximations for the most general setting are however still lacking and this section aims to fill this gap. This work was made in collaboration with Lea Duncker, a PhD candidate at the Gatsby Unit and will be presented at Cosyne [53].

We consider the task of inferring smooth neural trajectories from single trials of simultaneously recorded neurons. GPFA is a method to solve this task under the assumptions that temporal correlations in the high-dimensional neural population are modelled via a lower number of shared latent processes, which linearly relate to conditionally Gaussian [69] or Poisson [70] observations in neural space. This can be seen as an extension of the additive setting with the introduction of a 'loading matrix' parameter $C$ specifying weights in the additive predictors.

Formally, we model $K$ latent processes as independent draws from Gaussian Processes (GP) with potentially different mean functions $m_k(t)$ and covariance kernels $\kappa_k(t,t')$. We model the intensity function, or firing rate, for each of $N$ neurons as a linear combination of these latent processes, together with a constant offset $\mathbf{d}$, and map this linear predictor through a pointwise non-linearity $g : \mathbb{R} \to \mathbb{R}^+$to obtain strictly positive firing rates:

$$\boldsymbol{\lambda}(t) \quad = \quad g\left(C\mathbf{x}(t)+\mathbf{d}\right), \mathbf{x}(t) = \left[x_1(t),...,x_K(t)\right]^T, \boldsymbol{\lambda}(t) = \left[\lambda_1(t),...,\lambda_N(t)\right]^T$$

We are interested in two extensions from previous work: (1) using a continuous-time point-process likelihood (to avoid the often arbitrary and information degrading binning process used in non-continuous-time methods), (2) to derive a fast approximate inference algorithm based on our additive sparse variational approximation.

**Continuous time GPFA**

Here, we model the number of spiking events in $\mathscr{T} \subset [0, \mathscr{T})$, $\Phi(n)$, for neuron $n$ according to a Poisson Process, such that $\Phi(n) \sim \text{Poisson}\left(\int_{\mathscr{T}} \lambda_n(t)dt\right)$. The intensity measure of the point process for neuron $n$ is thus given by $\Lambda_n(\mathscr{T}) = \int_{\mathscr{T}} \lambda_n(t)dt$ and drives the observed number of spiking events $\Phi(n)$. The model inputs are the observed spike times $\mathbf{t}^{(n)} \in \mathbb{R}_+^{\Phi(n)}$ for each neuron, which we will denote by $\mathscr{D} = \{\mathbf{t}^{(1)}, \ldots, \mathbf{t}^{(N)}\}$. The log-joint distribution of the observed spike times for all neurons and latent processes is hence

$$\log p(\mathscr{D}, \mathbf{x}(t)|C, \mathbf{d}, \theta) = -\sum_{n=1}^{N} \int_{\mathscr{T}} \lambda_n(t)dt + \sum_{n=1}^{N} \sum_{i_n=1}^{\Phi(n)} \lambda_n(t_{i_n}) + \sum_{k=1}^{K} \log p(x_k(t)|\theta_k)$$

Our goal is to learn the model parameters $C$ and $\mathbf{d}$, the kernel hyperparameters $\theta = \{\theta_1, \ldots, \theta_K\}$ and infer the latent trajectories $\mathbf{x}(t)$. This requires computing the posterior distribution $p(\mathbf{x}(t)|\mathscr{D}, C, \mathbf{d})$, which is intractable in this model.

**Sparse approximation**

In order to arrive at a scalable variational inference algorithm, we introduce a set of inducing points $U = [\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(K)}]$ for each latent process, which are evaluated on a set of "pseudo-spike-time" inputs $Z = [\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(K)}]$. We choose an approximating distribution of the form $q(U, X) = \prod_{k=1}^{K} q(\mathbf{u}^{(k)}, \mathbf{x}^{(k)}) = \prod_{k=1}^{K} p(\mathbf{x}^{(k)}|\mathbf{u}^{(k)})q(\mathbf{u}^{(k)})$. Using this approximation and an assumed Gaussian form $q(\mathbf{u}^{(k)}) = \mathscr{N}(\mathbf{m}^{(k)}, \mathbf{S}^{(k)})$, we derive a variational lower bound to the log-likelihood. Letting $h_n(t) = \sum_{k=1}^{K} c_{n,k} x_k(t) + d_n$ denote the linear predictor for the $n$-th neuron with marginal variational distribution $q(h_n(t)) = \mathscr{N}(\mu_n, {\sigma_n}^2)$ we have:

$$\begin{aligned} p(\mathscr{D}|C, \mathbf{d}, \theta) \quad \geq \quad & -\sum_{n=1}^{N} \int_{\mathscr{T}} \mathbb{E}_{q(h_n)}\Big[g(h_n(t))\Big]dt + \sum_{n=1}^{N} \sum_{i_n=1}^{\Phi(n)} \mathbb{E}_{q(h_n)}\Big[\log g(h_n(t_{i_n}))\Big] \\ & -\sum_{k=1}^{K} \text{KL}\left[q(\mathbf{u}^{(k)}) \| p(\mathbf{u}^{(k)})\right] \end{aligned}$$

Once again, this lower bound can be maximised directly with respect to the model parameters, hyperparameters, variational parameters and pseudo-spike times. The first term involves one-dimensional Gaussian integrals, which can be computed in closed form or using efficient numerical approximations depending on the choice of non-linearity $g(\cdot)$. The second term scales with the number of spiking events in

$\mathscr{D}$, while the Kullback-Leibler divergence in the last term scales with the number of inducing points. Thus, none of the terms scale with the total duration of the experiment, allowing for fine temporal resolution without increasing the computational burden of the algorithm.

**Application to simulated data**

We simulate neural spike trains for 50 neurons using three underlying latent processes (Fig. 2.1) and compare the inferred trajectories obtained from our continuous-time approach (PP-GPFA) with those obtained from the discretised-time Poisson approach in Zhao & Park, 2016 (vLGP)[70]. We initialize both methods with noise-corrupted versions of the generative parameters.

For PP-GPFA, we run a fixed number of inference-only steps and subsequently optimise the variational lower bound jointly with respect to the variational parameters, model parameters, hyperparameters and inducing points. Figure 2.2($1^{st}$ row) shows examples of the true firing rates and their estimates, showing that discretising time may lose important information, even at relatively small bin widths. Overall, the continuous time approach more accurately captures the high-frequency oscillations in the data than any of the discretisations and provides improved firing rate estimates (Figure 2.2, $2^{nd}$ row). This example illustrates that our point-process approach allows one to fully exploit the temporal resolution of the spike train and can capture underlying structure that may be missed when using binned spike-count observations.

## 2.5 Conclusion

This chapter provided a brief introduction to the framework of Bayesian Inference underlying computations in the modelling work of this thesis. I reviewed the problem of Gaussian Process regression and its sparse variational approximation. I presented my extension to this approximation scheme for additive models. These methods will be revisited in further chapters of this thesis.

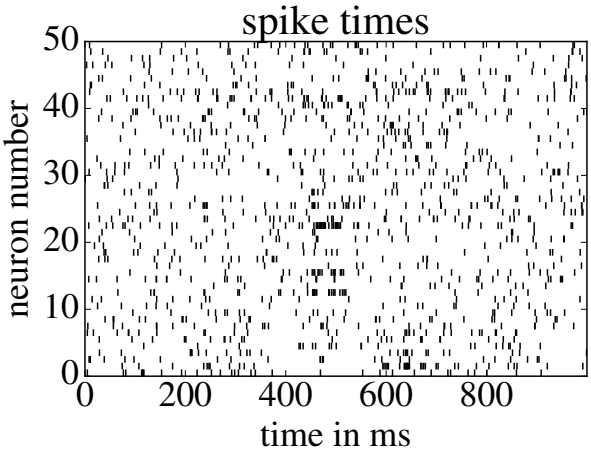


**Figure 2.1:** Simulated spike trains for 50 neurons

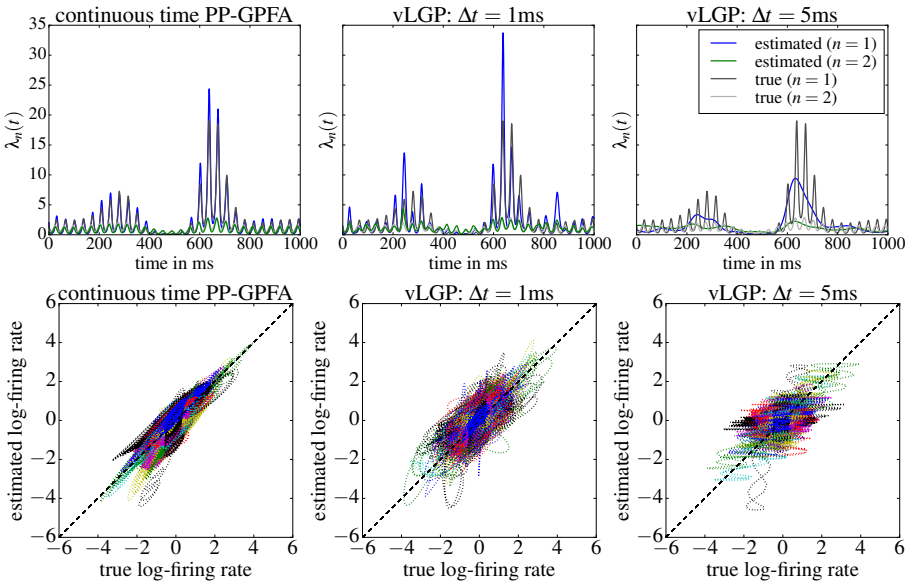


**Figure 2.2:** GPFA results: first row shows two example firing rates and their estimates under the different methods. Second row shows plots of the true log-firing rates against their estimates for each of the 50 neurons under the different methods.

# Chapter 3

# Pitch perception as probabilistic inference

## 3.1 Collaboration statement

The work presented here is an extension of the unpublished PhD work of Phillipp Hehrmann. A large portion of the chapter is devoted to presenting and summarizing this work. This presentation starts with the necessary yet brief literary review of the psychophysics of pitch perception and the neuro-physiology of the auditory periphery necessary to justify the initial work. I will only introduce the material necessary to understand the approach. Some visual material was borrowed from this earlier work and this is acknowledged in all cases. Heiko Strathman and Dino Sejdinovic also helped in early technical discussion on this project.

## 3.2 Introduction

Four properties - duration, loudness, pitch and timbre - are commonly used to describe the perceptual quality of a sound. Pitch and timbre remain vaguely defined. Pitch is described as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high" [71]. For the broad class of periodic sounds, the pitch of a sound is well described by its fundamental frequency. However this definition is ambiguous and fails to account for the pitch of a large range of sounds for which there is no fundamental frequency or the perceived pitch departs from it. More importantly, the perception of pitch is highly dependent

on the context or recent stimulation history. Timbre is even less well-defined. A typical approach is to define timbre as the ensemble of all those qualities that distinguish sounds of equal perceived pitch, loudness and duration [72]. The perceptual space of timbre was once termed the "psychoacoustician's multidimensional waste-basket" because of its definition in terms of negatives [73]. Among its dimensions that have been studied [74, 75, 76] are the spectral center of mass – referred to as "brightness", the noisiness of sounds, the rise and decay rate of the waveform and the related spectro-temporal modulations.

The perception of these sound qualities, which occurs with no conscious effort or attention, is still poorly understood and models describing perception as mechanistic feature extraction often fail to account for many reported effects in the psychoacoustic literature. In the case of pitch, this has lead to the proposal that pitch perception might be an inferential process combining sensory evidence to prior expectations [15]. Those prior expectations were described in the form of a probabilistic model of pitch-evoking sounds that parametrically captured notions of underlying periodicity, timbre and noisiness. Sensory evidence was given through a neural stochastic transduction model capturing the principal features of the dynamical sound encoding in the auditory periphery.

In this chapter, after summarizing this approach and presenting some of its successes, I present my extensions to it. These extensions are both scientific and technical. On the scientific side, the model is extended and a novel psychophysical experiment is reported. On the technical side, a novel algorithm is derived and tested to perform the - necessarily approximate - inference underlying the proposed model of pitch perception. This algorithm was described in section 2.3.2.

## 3.3 Previous work: a probabilistic model of pitch perception

### 3.3.1 Basics of pitch perception

Pitch is an important perceptual quality of many natural sounds. In all spoken languages, pitch carries prosodic information. In music, the definitions of musical

scales, melody and harmony are unthinkable without reference to our perception of pitch. As a stable attribute of sound sources, pitch plays a crucial role in scene analysis. For example, knowledge of the slow variation of pitch in speech helps source separation in a cocktail party context. As a perceptual quality, it has resisted attempts at defining it as a physical property of sounds alone. As an example, past sensory experience affects the perception of pitch. So a same sound can be perceived in different ways depending on the context in which it is presented.

Psychoacoustics as a field was initiated as a means to uncover the relationship between the subjective experience of the perception of sounds and the acoustic stimuli, and to understand the neural underpinnings supporting auditory perception.

Reviewing definitions of pitch, the psychoacoustic literature on pitch perception and proposed models of human pitch perception is out of the scope of this thesis. I refer the reader to the book chapters of Alain de Cheveigne [77, 78], or to the recent textbook by Schnupp, Nelken and King [79]. Such a review was performed in the original thesis work of Philipp Hehrman [15]. Instead, I here summarize the key ideas necessary to understand the following work.

1. To a first approximation, pitch may be treated as a unidimensional perceptual dimension related to periodicity of sounds. Many physically different sounds may evoke the same pitch and different pitches can be ordered on a scale from low to high. As a perceptual dimension it can only be indirectly studied, through psychophysical experiments such as discrimination or matching tasks.
2. The percept of pitch is constructed from the transduced neural activity in the nervous system. Properties of the transduction have a strong influence on pitch perception.
3. Throughout the years, many non periodic sounds, such as non harmonic complexes or amplitude modulated white noise samples, have been found to nonetheless evoke a sense of pitch. The striking tolerance of the sense of pitch to strong departures from periodicity for this 'zoo' of pitch evoking sounds has fuelled and constrained the development of models of pitch perception.

### 3.3.2 Pitch perception as Bayesian inference

Most natural, pitch-evoking sounds are approximately, though not perfectly, periodic within short observation time windows. Building on previous work by Goldstein [80], Hehrmann et al [15] hypothesised that the auditory system is trying to estimate their periodicity, based only on indirect observations through the noisy, evoked neural response in the auditory nerve. Since the physical process of sound generation, transmission and sensorineural transduction is inherently stochastic, optimal inference requires knowledge about the underlying statistical regularities and irregularities. They formulated their model within the framework of Bayesian probabilistic inference [81], which provides both the formal language to define this inference problem rigorously, and the algorithmic tools to compute (or approximate) its optimal solution.

Formally let $\Omega$ be the unknown period of a sound, $\theta$ be some additional unknown features of the same sound (here timbral ones) and let $p(\Omega,\theta)$ be their known natural frequencies of occurrence in the environment. A generative model of a pitch evoking sound $s(t)$ was assumed conditional on those sound features $p(s(t)|\Omega,\theta)$. Finally a generative model of auditory nerve activity $A(t)$ in response to a sound input was assumed $p(A(t)|s(t))$. Inferring the underlying features from auditory nerve activity correspond to computing the posterior

$$\begin{aligned} p(\Omega,\theta|A(t)) &\propto p(A(t)|\Omega,\theta)\, p(\Omega,\theta) \\ &\propto \int ds(t)\, p(A(t)|s(t))\, p(s(t)|\Omega,\theta)\, p(\Omega,\theta) \end{aligned}$$

Inferring pitch corresponds to the further marginalization of the posterior over the other features.

$$p(\Omega|A) = \int d\theta\, p(\Omega,\theta|A)$$

In the following sections I will describe the different elements of this model: the sound model and the transduction model. Following this exposition, I will present my extension to the original model and its ability to explain some psychophysical observations.

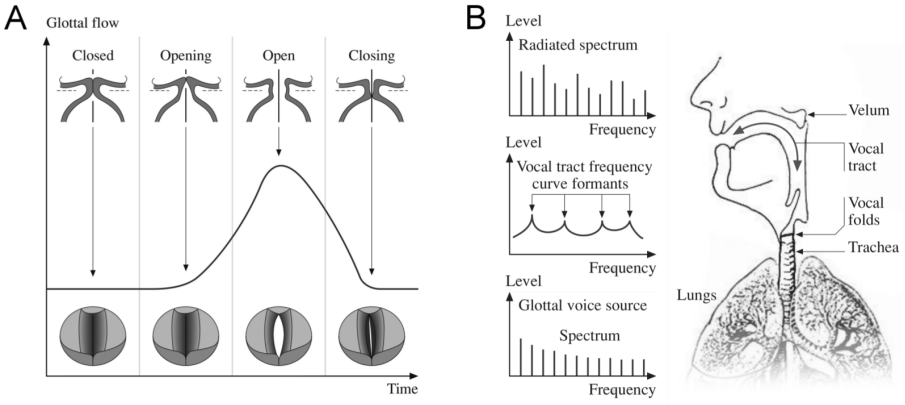


**Figure 3.1:** Schematic of human vocal production. A: Glottal air-pressure wave- form generated by the periodic opening and closing of the vocal folds. B: The spectrum of the emitted waveform (top) is the product of the glottal source spec- trum (bottom) and the vocal tract resonance spectrum (middle). Taken from [15], originally adapted from Lindblom and Sundberg [82]

### 3.3.3 Sound model: prior on pitch evoking sound

The source-filter theory of speech production [83] is essentially a model of periodic sounds. In this model, the vocal tract is assumed to act as a linear filter on the broad spectrum periodic glottal source, as depicted in Figure 3.1. As such it can be regarded more broadly as a model of general pitch evoking sound production. The authors constructed a generative model of such sounds by adding a distribution over the filter. More formally, it was assumed that the soundwave of pitch evoking sounds $x(t)$ could be expressed as the convolution (denoted $*$) of a $\Omega$-period dirac impulse train $III_\Omega(t)$ with a fixed filter $f(t)$. A generative model of stationary periodic sounds was thus constructed by providing a prior on the filter $p(f)$. Pitch perception being robust to reasonable amounts of noise, additive noise of variance $v^2$ was added to the soundwave.

$$\begin{aligned} f &\sim p(f|\Omega) \\ \eta &\sim p(\eta) \\ x(t)|f,\eta &= (III_\Omega * f)(t) + v\eta(t) \end{aligned} \tag{3.1}$$

The prior on the impulse response $f(t)$ was designed to allow for variability in the periodic pattern to be repeated, to capture in a parametric form two timbral notions. The first one is brightness. A filter was sampled by smoothing out a white noise

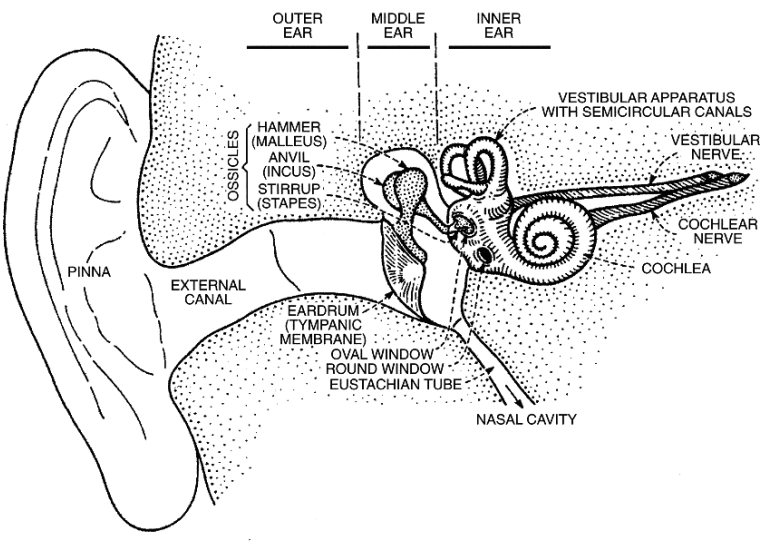


**Figure 3.2:** Anatomy of the human outer, middle and inner ear (from [84])

token $n(t)$ with a Gaussian filter $k_\sigma(t) = \exp\left\{-\frac{t^2}{2\sigma^2}\right\}$ of width $\sigma$. This imposed a Gaussian shape to the spectral density of the filter and resulting sound. The second timbral parameterization was to impose an explicit decay to the filter by applying a decaying envelope $e(t)$ to white noise token $n(t)$ prior to its smoothing. This can be thought of applying an extra amplitude modulation to the signal. Width and phase of this envelope $e(t)$ were also randomized. This generative model defines a distribution over a wide range of pitch evoking sounds.

### 3.3.4 Transduction model

A simplified model of auditory transduction was derived to capture its key features known to affect and limit pitch perception. I here describe the transformation steps from soundwave $s(t)$ entering the ear to tonotopic activity in the auditory nerve $A(t)$, along the auditory pathway depicted in Figure 3.2, and how they were modelled.

The auditory periphery starts with the external ear, which corresponds to the pinna and the ear canal. Sound $s(t)$ propagating through the air passes through the ear canal and hits the ear drum, the entry to the air-filled middle ear. The ear drum transmits the mechanical vibration it receives via three ossicles to the inner ear, through one of the two flexible parts of cochlea situated at its base: the oval window. The cochlea is a fluid-filled, helically coiled tube encased in a hard shell

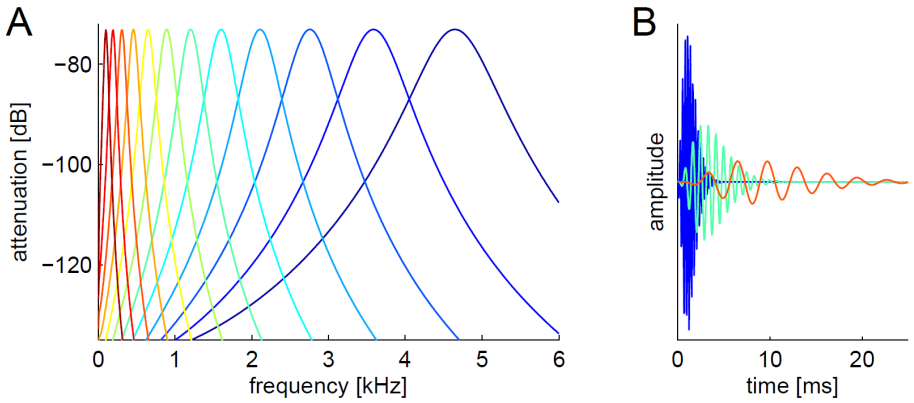


**Figure 3.3:** Gammatone Filter bank. A: Spectral magnitude response of 12 gammatone filters as used in the model. B: Impulse response of three filters (from [15]).

bone. There is where the transduction happens.

#### 3.3.4.1 Basilar membrane motion

The basilar membrane runs along the tube, from base to apex (see cross-section in Figure 3.4). It is deflected by the incoming vibration of the oval windows in positions depending on the frequency content of the incoming mechanical vibration in a tonotopic manner, with a log mapping between position and frequency from $20Hz$ at the apex to $20kHz$ at the base. A classical description of the oscillating movement of the basilar membrane $m(x)$ at a position $x$ is as the output of the convolution with a Gammatone filter with central frequency $f_x$: $m_x(t) \propto t^{n-1}e^{-2\pi\beta_x t}\cos(2\pi f_x t)$. The Gammatone filterbank is a discretized model of the motion of the full membrane with center frequencies log-homogeneously spaced and pass-band width growing with center frequency. In our model, the filterbank parameters were chosen following the implementation of Patterson et al. [85] and Slaney [86] so as to best match observations of human basilar membrane motion. A depiction of the filterbank used in [15] and in my own experiments is shown in Figure 3.3.

#### 3.3.4.2 Hair cells and auditory nerve

Hair cells populate one side of the basilar membrane as depicted in Figure 3.4. I only consider the feed-forward role of inner hair cells (IHC) in conveying sound information to the brain. The motion of the basilar membrane $m(t)$ opens ion channels on the IHCs giving rise to a ion current flow $i(t)$, later generating spikes in the

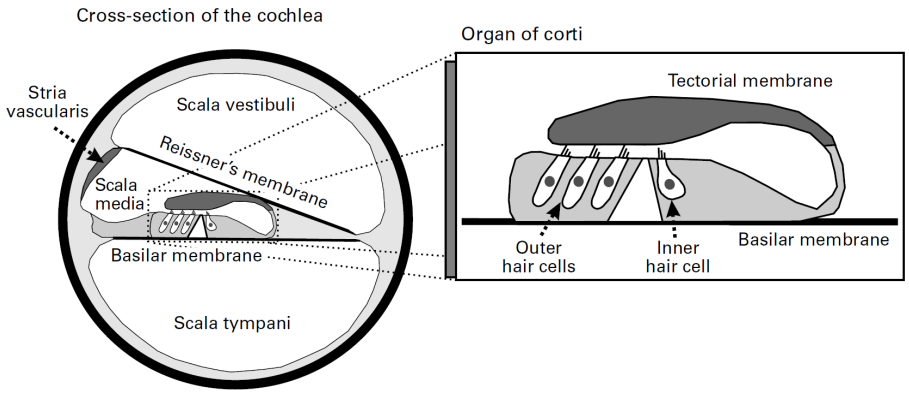


**Figure 3.4:** Cross-section of the cochlea, and schematic view of the hair cells in the organ of Corti (from [79]).

auditory nerve after a few synaptic relays. This ion channel opening is asymmetric with the oscillation sign and for slow oscillation is well modelled as a half-wave rectification of the oscillation: $i(t) \propto r(m(t))$. For differentiability reasons, a soft rectifier was used: $r(z) = \frac{\log(1+\exp(\alpha z))}{\alpha}$. The ion current created as a result of motion has an inertia and can only "track" the oscillation that gave rise to it for frequencies up to 4 to 5 kHz [87]. Higher frequencies are still transmitted, but the filtering of the rectified oscillation leads to a demodulated envelope signal only (See Figure 3.5). This final inertia was modelled as the action of a low-pass linear filter $l(t)$: $i(t) = (r(m) * l)(t)$. Finally the fibres of the auditory nerve are organized in a similar tonotopic manner, and transmit spike trains with rates corresponding to the IHC potentials fluctuations.

A graphical depiction of the full model (sound and transduction) is given in Figure 3.6.

#### 3.3.4.3 Features affecting pitch perception

The proposed model of auditory transduction accounts for two main limitations of this transduction affecting the perception of pitch.

1. Resolvability limit: the bandwidth of the filters describing basilar membrane motion scales with their central frequency. For high frequencies, this bandwidth is large. Hence, multiple harmonics of harmonic complexes of low pitched sounds may pass through a broad filter and later interact due to the

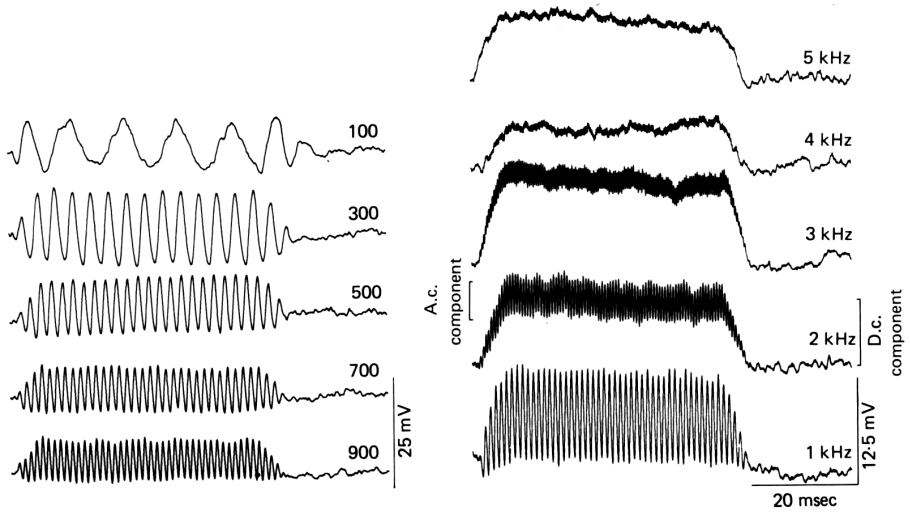


**Figure 3.5:** IHC receptor potentials in response to tones of different frequencies presented at 80 dB SPL, measured at the basal turn of a guinea pig cochlea (from [87]).

non linearity of the rectification. This means individual high harmonics cannot be read out from individual channels; they are not resolvable

2. Phase-locking limit: the IHC potential in response to high frequency tones do not oscillate like the basilar membrane at the corresponding position. Information about the phase of the oscillation is lost.

These limitations have fuelled a debate on the nature of the information carried about pitch in the auditory nerve activity and how it is read out later on in cortical areas. On the one hand, the irresolvability of high harmonics prevents a "spectral" only representation of pitch: there is no encoded spectrogram from which to clearly read out the harmonics present in a harmonic complex. On the other hand, the lack of temporal information in the output of high-frequency tuned IHC prevents a purely "temporal" approach : there is no periodicity information in a flat signal. This debate is explained in great detail in [77].

### 3.3.5 Inference

The inferential problem of interest is to compute the marginal likelihood or evidence $l(\Omega) = p(A|\Omega)$ of a pitch candidate value $\Omega$, where $A$ is the transduced auditory nerve activity due to a sound $s \in \mathbb{R}^D$ . The calculation of the evidence requires marginalizing over a sound as follows: $l(\Omega) = \int ds\, p(A,s|\Omega)$. Hehrmann et al [15]

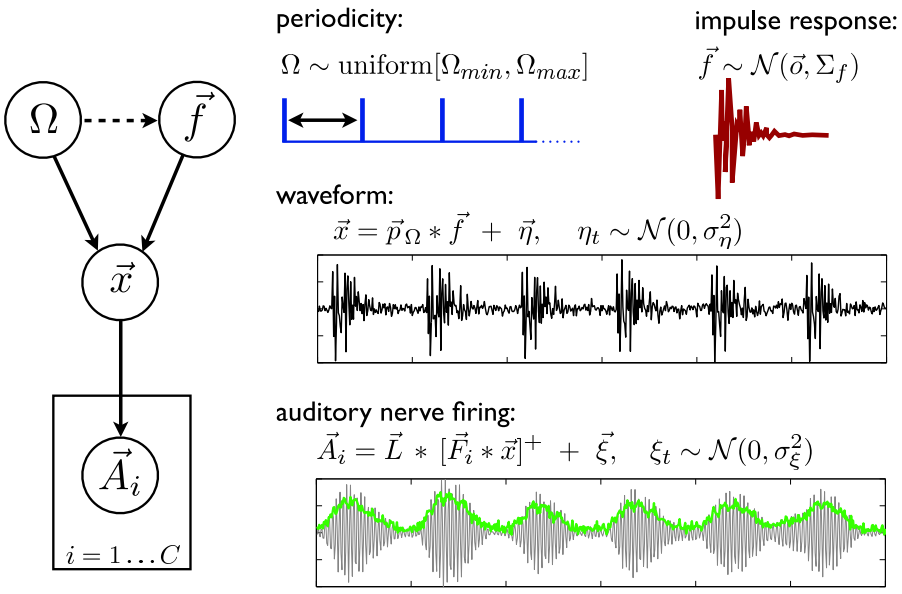


**Figure 3.6:** A generative model of naturalistic, approximately periodic sounds and evoked auditory nerve responses. A pulse train with period $\Omega$ is convolved with a randomly-generated acoustic impulse response $f$ and corrupted by additive noise to obtain an acoustic waveform $x$. $x$ evokes responses in auditory nerve fibres $i = 1...C$ as follows: in each channel, the waveform is filtered by a linear bandpass filter with impulse response $b_i$ . Its output is half-wave rectified and low-pass filtered before further noise is added, resulting in a demodulation of the filter outputs for oscillation rates above the low-pass cutoff frequency.

proposed two algorithms to approximate the evidence.

### 3.3.5.1 Laplace approximation

The first one was the Laplace approximation [81]: Writing $s^* = \arg\max p(A, s|\Omega)$, the sound that maximizes the posterior, the log-joint distribution over $A$ and $s$ can be Taylor-expanded around $s^*$:

$$\log p(A, s|\Omega) \approx \log p(A, s^*|\Omega) - \frac{1}{2}(s - s^*)^T H_{s^*}(s - s^*)$$

where $H_{s^*}$ is negative Hessian of the log joint at $s^*$

$$H_{s^*} = -\nabla_s^2 \log p(A, s|\Omega)|_{s^*}$$

Intuitively, this corresponds to approximating the posterior over the unobserved sound as a multivariate Gaussian distribution around its mode, with the covariance

matching the curvature at the mode. The evidence can then be approximated as

$$\begin{aligned} l(\Omega) &= \int ds\, p(A,s|\Omega) \\ &\approx p(A,s^*|\Omega)\sqrt{\frac{(2\pi)^D}{\det H_{s^*}}} \end{aligned}$$

#### 3.3.5.2 Sampling based approximation

The evidence can be seen as an expectation $l(\Omega) = \mathbb{E}_{p(s|\Omega)}[p(A|s)]$. Provided with independent and identically distributed (iid) samples from the prior, $s_i \underset{iid}{\sim} p(s|\Omega)$, the evidence can be approximated as $l(\Omega) \underset{N\to\infty}{\approx} \frac{1}{N}\sum_i^N p(A|s_i)$ which is a particular case of Importance Sampling. This fails in high dimensions because most of the mass in the integral $\int ds\, p(A,s|\Omega)$ is associated with values of $s$ likely under the prior but unlikely to explain the observations $A$ well. Thus the time required to obtain even few representative samples becomes prohibitively large. In [15] the authors proposed a sampling based scheme known to provide accurate approximations in high dimensions: Annealed Importance sampling [88] using Hamiltonian Monte Carlo sampling [89].

#### 3.3.5.3 Limitations

The Laplace approximation though fast often performs poorly in high dimensions. Finding the posterior mode might be hard and the Gaussian approximation to the posterior around the mode might be a poor one. On the other hand, the sampling algorithm proposed is better behaved but prohibitively expansive to run; for this reason, it was tested but not used in [88]. The algorithm I used based on sparse variational Gaussian Process approximations provided the needed intermediate technical solution that is both fast and accurate.

## 3.4 Extension: Simpler yet richer prior model

### 3.4.1 Another timbral dimension: pattern variability

The generative model of sound proposed earlier is one of noisy periodic sounds. Noise is the only deviation from periodicity modelled. Another deviation observed in most natural pitch evoking sounds, including voiced speech is the slow variation

in time of the pattern that is exactly repeated in periodic sounds. This variation is a critical aspect of the timbre. When lacking, sounds sound somehow unnatural or artificial. Keeping the source-filter model inspiration, a different filter could be used in the convolution with each impulse of the impulse train. A simple probabilistic model of this discrete variation, that preserves the stationarity of the model is of a variance preserving autoregressive model (a particular instance of a Markov chain):

$$f_{i+1}|f_i = e^{-\Omega/\rho} f_i + \sqrt{1-e^{-\Omega/\rho}}\varepsilon_i$$

where both the initial filter $f_0$ and $\varepsilon_i$ are independently sampled from the same shared filter model $p(f)$. The constructed signal can no longer be written as a convolution. Instead, it is given by: $x|\underline{\mathrm{F}} = \sum_i f_i * \delta(t - i\Omega)$, where $\underline{\mathrm{F}} = (f_0, f_1 ...)$ represents the consecutive filters of the chain. In the filter definition, $\rho$ parameterize the temporal locality. This new generative model defines a distribution over pitch evoking sounds and has more "mass" on voiced speech sounds whose repeated pattern vary in time. I showed in previous work that speech sounds are more likely under this model than under a white noise model of matched variances [90]. This was not true under the original model of [15].

### 3.4.2 A Gaussian process formulation

Chosen $p(f)$ to be smoothed Gaussian white noise with a Gaussian filter $k_\sigma$ as in the original model without the decaying envelope, this generative model can be shown to result in audio samples that correspond to draws from a Gaussian Process with a stationary covariance function given by

$$\mathscr{K}_\theta(\tau) = \sum_{l=-\infty}^{\infty} e^{-\Omega|l|/\rho} k_\sigma(\tau - l\Omega) + \nu^2\delta(\tau). \tag{3.2}$$

This stationary kernel is associated to a spectral density $S(s)$, through the Fourier transform

$$S(s) = \mathscr{F}\{\mathscr{K}_\theta\}(s) = \left(\sum_{l=-\infty}^{\infty} e^{-\Omega|l|/\rho} e^{-il\Omega}\right) \mathscr{F}\{k_\sigma\}(s) + \nu^2$$

The Fourier series has a clearer interpretation if we replace the discrete weights in the sum $e^{-\Omega|l|/\rho}$ by the continuous $k_{1,\rho}(\tau) = e^{-|\tau|/\rho}$ in the kernel expression

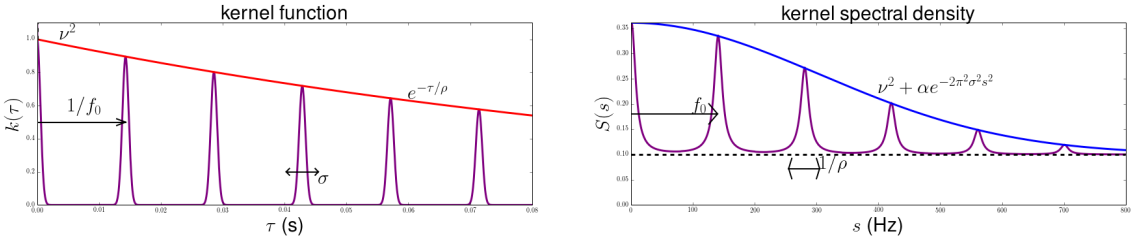


**Figure 3.7:** Kernel function of Equation 3.3 and its spectral density.

(Laplacian covariance function), resulting in:

$$\begin{aligned}\tilde{\mathscr{K}_\theta}(\tau) &= k_{1,\rho}(\tau)\sum_{l=-\infty}^{\infty} k_\sigma(\tau - l\Omega) + \nu^2\delta(\tau) \\ &= k_{1,\rho}(\tau)\left(\mathrm{III}_\Omega * k_\sigma\right)(\tau) + \nu^2\delta(\tau),\end{aligned}$$

In this expression, the kernel is expressed as a product of a periodic kernel made of $\Omega$-spaced repetitions of the Gaussian kernel $\left(\mathrm{III}_\Omega * k_\sigma\right)(\tau)$ and of a Laplacian kernel $k_{1,\rho}(\tau)$ with additional white noise. The spectral density of this kernel is

$$\tilde{S}(s) = \mathscr{F}\{\tilde{\mathscr{K}_\theta}\}(s) = \mathscr{F}\{k_\rho\} * \left(\mathscr{F}\{k_\sigma\}(s)\mathrm{III}_\mathfrak{f}(s)\right) + \nu^2,$$

which is a Gaussian-modulated Dirac comb, convolved by $\mathscr{F}\{k_\rho\}(s) = \frac{2\rho}{1+\rho^2 s^2}$. Another interpretation is as a blurred harmonic spectrum, where the blur is controlled by parameter $\rho$.

An alternative kernel function that has very similar properties to (3.2) is given by

$$K_\theta(\tau) = \exp\left(-\frac{2}{\gamma^2}\sin\left(\pi\mathfrak{f}\tau\right)^2\right)\exp\left(-\frac{|\tau|}{\rho}\right) + \nu^2\delta(\tau). \tag{3.3}$$

This kernel is obtained from standard kernels: periodic and Laplacian, which have been used extensively in the literature on Gaussian Processes – cf. [91] and references therein. This kernel function constitutes a good approximation to (3.2) while preserving the interpretation of its parameters: fundamental frequency $\mathfrak{f} = 1/\Omega$, smoothness $\gamma$, correlation decay $\rho$, and the additive noise parameter $\nu$. Smoothness in this formulation is approximately related to that in 3.2 as $\gamma \approx \sigma\pi/\Omega$. Figure (3.7) illustrates the interpretation of these parameters in both the time and frequency domain.

The formulation as a Gaussian Process has the following advantages: Scien-

tifically it provides a structured prior on sounds in a simple and elegant manner. Technically it makes it possible to use the wide range of recent technical advances in probabilistic inference using Gaussian Processes, and in particular, the sparse variational inducing point approximation.

### 3.4.3 Unifying temporal and spectral methods

Let us consider the task of periodicity estimation directly from the sound wave, ignoring the transduction part of our model.

A sound $x$ obtained from the model on time samples $t = (t_1 \dots t_s)$ is simply a normal vector with mean 0 and covariance matrix $\mathbf{K}_\theta$, where $(\mathbf{K}_\theta)_{ij} = K_\theta(t_i - t_j)$. Thus, the log-likelihood function of parameters $\theta$ can be written as

$$L(\theta) = \log\{\mathcal{N}(x; 0, \mathbf{K}_\theta)\} = -\frac{1}{2}\log|\mathbf{K}_\theta| - \frac{1}{2}x^T\mathbf{K}_\theta^{-1}x.$$

Let us now restrict attention to the fully periodic case, i.e., we set $\rho = \infty$. When the sound duration and the sampling frequency $f_s$ are respectively multiples of $\Omega = 1/\mathfrak{f}$ and $\mathfrak{f}$, $\mathbf{K}_\theta$ is a circulant matrix. It can thus be diagonalized by the Fourier matrix $F$ as $\mathbf{K}_\theta = F^*\Lambda_\theta F$. The eigenvalues in $\Lambda_\theta$ are positive and correspond to the expected spectral energy under the generative model. By denoting $s = Fx$, the likelihood function can be rewritten as

$$\begin{aligned} L(\theta) &= -\frac{1}{2}\log|\mathbf{K}_\theta| - \frac{1}{2}x^T\mathbf{K}_\theta^{-1}x \\ &= \log\left\{\mathcal{N}\left(|s|; 0, \Lambda_{\theta,i}\right)\right\} \\ &= -\frac{1}{2}\sum_i \log\Lambda_{\theta,i} - \frac{1}{2}\sum_i \frac{|s|_i^2}{\Lambda_{\theta,i}}. \end{aligned}$$

Thus, the likelihood of sound $x$ corresponds to the likelihood of the modulus of its spectrum $|s|$ under the centred Gaussian density of variance corresponding to the expected spectral density $\Lambda_\theta$.

Pitch estimation through likelihood maximization has a direct interpretation as probabilistic spectral matching [92] where the patterns to be matched to are the expected spectrum under our statistical model. The strobing method for pitch estimation [93] corresponds to maximum likelihood under a GP model with a Dirac comb kernel. This model is the limiting case of $\sigma, \nu \to 0$ and $\rho = \infty$ in Equation

3.2. It is interesting to note that phase information is irrelevant. This is a direct consequence of the choice of a stationary kernel. Psychophysically this is consistent with the fact that humans are insensitive to the relative phase of harmonics in a harmonic stack so long as the harmonics are resolved.

When the kernel is no longer periodic, the covariance matrix of the associated Gaussian is no longer circulant and its eigenvectors are no longer simple discrete sines. However, for small deviations from periodicity, the interpretation as spectral matching remain a valid approximation. In this view, the different parameters of the generative model have the following interpretation: $\nu$ is a tolerance for spectral power homogeneously on the whole spectral domain, $\rho$ is a tolerance on the precise spectral location of the peaks of harmonics in a signal, $\sigma$ imposes an expected the decay of spectral energy at high frequencies.

The autocorrelation method for pitch estimation correspond to likelihood maximization with a non periodic kernel: the cropped dirac comb ($\delta_{-\Omega} + \delta_0 + \delta_\Omega$), where $\delta_x$ is the $x$-shifted dirac function. The underlying generative assumption is a 'local periodicity' and non-smoothed white noise filters.

In this view the generative approach I propose generalizes both spectral and temporal methods methods when the specificities of transduction are ignored.

## 3.5 Model evaluation

The evaluation presented here is not meant to be exhaustive. A full attempt at reproducing a large range of psychophysical observation relating to pitch perception was performed by Hehrmann et al [15]. Here, I show that the new version of the model along with the new inference method can replicate a subset of critical observations: I demonstrate the ability of our model to predict the pitch of both periodic and non-periodic sounds.

### 3.5.1 Evaluation methodology

- **Sounds**: All sounds were designed to evoke a pitch around $f_0 = 250Hz$, have a duration of $40ms$ and to have a waveform of unit variance. They were sampled at a rate of $16kHz$.

- **Transduction model**: The filterbank in the transduction model had 12 channels. The noise level was set to achieve a signal to noise ratio of 10dB. Auditory nerve activity was binned at $5kHz$.
- **Sound model**: $\rho$ was set to $40ms$, $\sigma$ was set to $0.1ms$.

For all sounds $s$, auditory nerve activity $A(s)$ was sampled and likelihood of the full model was computed on a grid of candidate pitch values depending on the predicted pitch.

### 3.5.2 Sounds

Sounds tested are :

- **HCT**: a missing fundamental harmonic complex with equal amplitude harmonics from rank 3 to 7, evoking a clear pitch at $f_0$
- **IRN**: an Iterated rippled noise token [94], consisting of white noise repeatedly delayed by $\Omega$, and added back to itself 4 times. This sound evokes a clear though weak pitch at $f_0 = \Omega^{-1}$.
- **AMT**: Amplitude modulated tones, a non-harmonic triplet of pure tones at frequencies $fc-g$,$f_c$,$f_c+g$, with $f_c = 1920$ and $g = 200$. The individual tones of this example are unresolved. Denoting by $n$ the rank of the harmonic of $g$ that is closest in frequency to $f_c$ and $\Delta f$ the frequency difference between the two, the perceived pitch of the triplet has been shown to be approximately equal to $f_p = g + \frac{\Delta f}{n}$. In this example, the closest harmonic of $g$ close to $f_c$ is at $n = 10$ and is equal to $2000Hz$, hence $f_p = 192Hz$ [95]. Subjects often report alternative pitch percepts at values corresponding to an under or over estimation of $n$ by $\pm 1$. In this case, alternative reports are at $212Hz$ and $172Hz$.

Example samples of the tested sounds are given in Figure 3.8.

### 3.5.3 Results

The likelihoods for an array of pitch candidate values are reported in Figure 3.9 for 20 repetitions of the sound generation, transduction, and inference. For the IRN and HCT a clear peak at $f_0 = 250Hz$ is obtained. the inferred pitch for the inharmonic

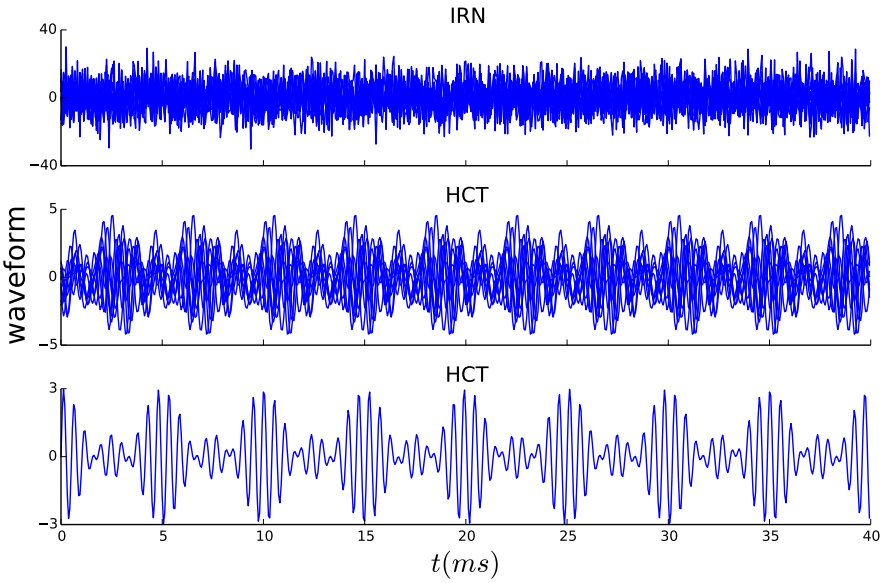


**Figure 3.8:** Sounds used for model evaluation: a IRN, a harmonic complex with missing fundamental, a non harmonic tone triplet.

tone triplet is shifted to $267Hz$ consistent with human reports, although other pitch candidates at lower values are also plausible. This demonstrates the ability of the proposed model of human pitch perception to predict human pitch percept for a variety of pitch evoking sounds.

### 3.5.4 Discussion

On the examples reported, the model predicts the most important qualities of perceptual reports across many different sounds. Although other models also explain those qualities, explaining the pitch of sounds who either lack temporal or spectral pitch for cues altogether with a single model is already a success [78].

## 3.6 Psychophysical experiment: timbre influences pitch perception

### 3.6.1 Motivation

#### 3.6.1.1 Demonstrating the effect of brightness of pitch perception

The most common type of errors in human pitch judgements, aside from small deviations around the 'true' pitch due to limited discriminability, are octave mistakes,

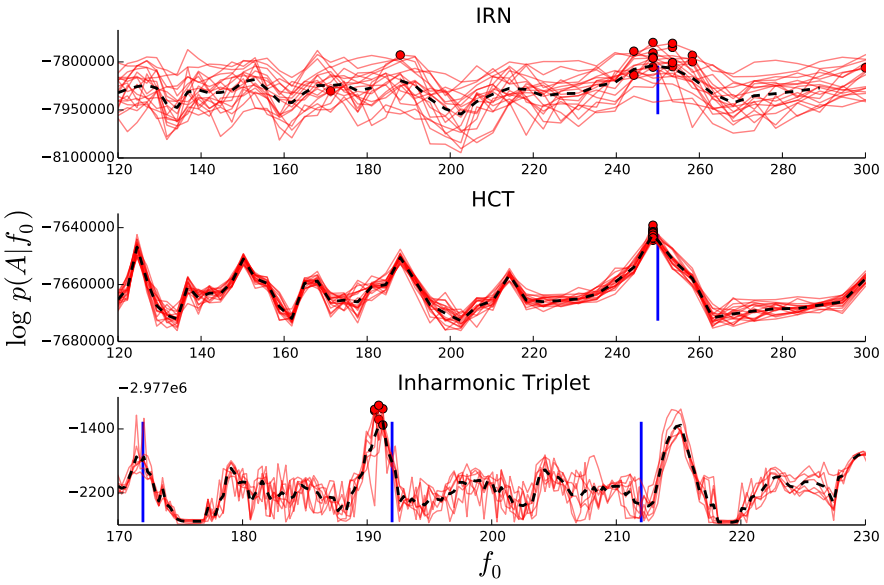


**Figure 3.9:** Likelihood profile for the tested sounds. In red, likelihoods (lines) and maximum likelihood (circles) for several repetitions of the experiment. In dashed black, average over the repetitions. Vertical blue lines mark predicted dominant and secondary pitches.

or the tendency to report octave-related values for pitch evoking sounds [96]. This is because octave related sounds are perceived as perceptually close. This has led to the concept of the position within an octave as a second, circular dimension called tone chroma, in addition to the first dimension, which scales monotonically with frequency and which is called the tone height. A simple spatial representation of pitch in these two dimensions is in the shape of a helix, where the chroma dimension winds around a tone-height axis [97]. Hehrmann et al [15] showed that these errors are strongly influenced by the brightness of sounds. To do so, convex mixtures of two click trains sharing the same chroma, one with a 'low' pitch of $f_0 = 250Hz$ and one with a 'high' pitch of $2f_0 = 500Hz$, were designed with $\kappa \in [0,1]$ controlling the relative weight of each click train: $\kappa = 0$ corresponds to the low pitch sound, $\kappa = 1$ corresponds to the high pitch sound, intermediate values correspond to alternating click trains. At some intermediate value of interpolation $\kappa^*$, the mixture is ambiguous in pitch, in the sense that when indirectly asked to report which of the two pitches -low or high- is played, subjects report either of the two with equal probability. A representation of some mixtures in the helicloidal perceptual space

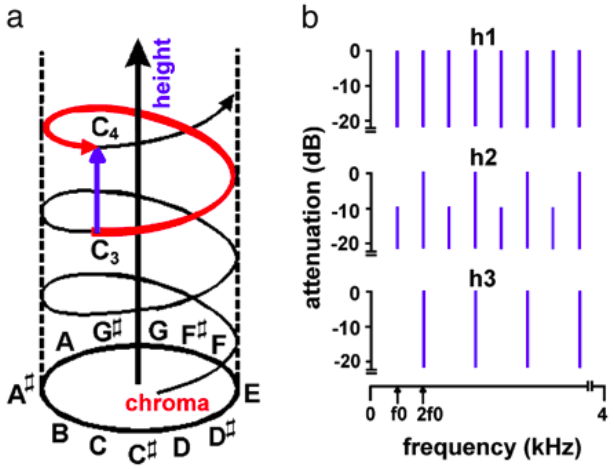


**Figure 3.10:** (a) The pitch helix. The musical scale is wrapped around so that each circuit (red) is an octave. (b) Spectrum of the individual click trains (h1,h3) and one mixture (h1). (from [98] where similar sounds are used).

and of their spectrum is given in Figure 3.10. Smoothing the mixture (i.e. manipulating its brightness) affected the point of subjective ambiguity in the following way: the darker (resp. brighter) the sound the higher (resp. lower) the proportion of 'high' pitch sound was needed in the mixture for it to be perceived has high.

A possible confound of the use of alternating click trains is that the sharp attack of dominant clicks might mask the following lower amplitude clicks explaining part or all of the effect of smoothing as a decrease of the sharpness of the clicks and hence of their masking effect. Here, I partly reproduce this experiment changing the stimuli mixed together to generate mixtures ambiguous in pitch. A click train is a harmonic complex containing all harmonics with equal amplitudes and zero phases. I randomized the phases of these harmonics which preserved the pitch and spread the energy in time preventing the undesired masking effect.

#### 3.6.1.2 Learned pitch-brightness dependence in natural sounds explains the effect

Hehrmann et al [15] revealed a statistical dependence between pitch and brightness studying large databases of natural pitch-evoking sounds. More precisely, the spectral centroid $f_c$ defined as the center of mass of the spectrum was shown, on average, to covary with pitch as $f_c(f_0) = 56.8\sqrt{f_0}$ (see Figure 3.11). In the generative model

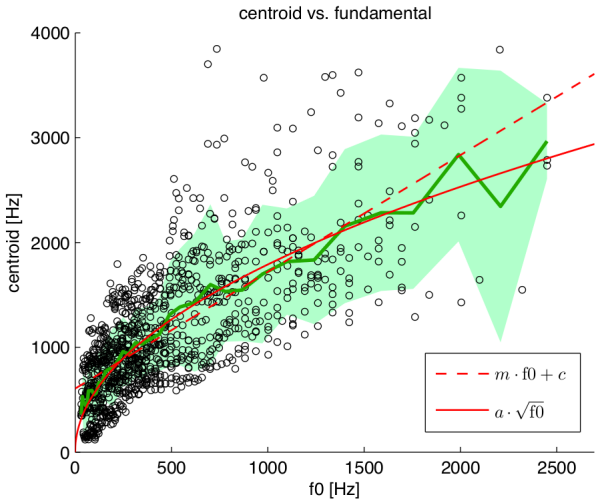


**Figure 3.11:** Fundamental frequency ($f_0$) and spectral centroid ($f_c$) for sounds from a collection of 20 musical instruments and two voices. Black circles represent $f_0$ and $f_c$ for each tone from the ensemble. Red line represent the best fitting parametric coupling (from [15]).

they described, and in my extension to it, the spectral centroid is inversely proportional to the smoothness parameter $f_c \propto 1/\sigma$. As a result, the statistical knowledge of the pitch-brightness dependence is readily encoded by forcing a deterministic coupling between $\sigma$ and $f_0$. The authors showed that this *a priori* coupling in the model could reproduce the effect of brightness on pitch. In the following section, we test the hypothesis that the same coupling explains the biasing effect of brightness on pitch on my modified version of his experiment.

### 3.6.2 Task description

A fundamental frequency was set to $f_0 = 250Hz$ for the whole experiment. 'Target' sounds (T) were constructed to ambiguously evoke pitch at either $f_0$ or $2f_0 = 500Hz$. Ambiguous sounds were created as an interpolation of non-ambiguous sounds at $f_0$ and $2f_0$ with interpolation coefficient $\kappa \in [0,1]$. $\kappa = 0,1$ both correspond to non ambiguous sounds at the two extremes of the octave range. For intermediate values of $\kappa$, the percept progressively shifts reaching a point of subjective equality (PSE). these target (T) sounds were presented to subjects in between two flankers (F) designed to have a non ambiguous pitch at frequency $f_i = f_0\sqrt{2} \approx 353Hz$, which is midway between $f_0$ and $2f_0$ on a log scale. Sub-

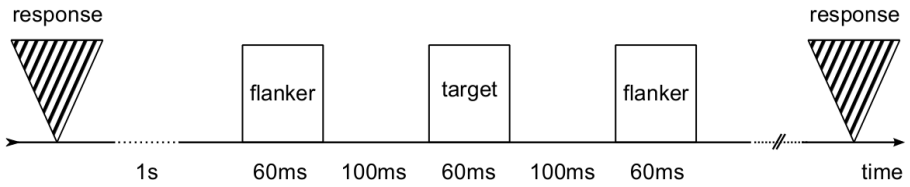


**Figure 3.12:** Task design: after an inter-trial interval of 1$s$, a sound triplet consisting of a target surrounded by two identical flankers is presented, following which subjects have to judge the melodic contour as either "rising-falling" or "falling-rising".(from [15]).

| | |
|---|---|
| brightness $\sigma$ | $\sigma_{dark} = 0.2$ms, $\sigma_{medium} = 0.12$ms, $\sigma_{broad} = 0.08$ms |
| ambiguity $\kappa$ | 0, 0.2, 0.4, 0.6 , 0.8, 1 |

**Table 3.1:** Target parameters

jects were presented triplets of sounds FTF played consecutively. Intuitively if the ambiguous target is perceived as closer to $2f_0$, the perceived sequence undergoes a melodic contour constituted of an upward shift followed by a downward shift ↗↘. If the target is perceived as closer to $f_0$, then the perceived melodic contour is ↘↗. Subjects were asked to classify the melodic contour as either ↗↘ or ↘↗. The manipulation of interest is to vary the target brightness.

Target sounds for both periods were constructed as follows. First the two periodic sounds to be mixed were click trains in which the phases of all harmonics were randomized. The resulting signals were smoothed using a Gaussian kernel $k_\sigma$ with width $\sigma \in \{\sigma_{dark}, \sigma_{mid}, \sigma_{broad}\}$ applying one of 3 desired brightness level to the click trains. Targets were constructed by mixing these sounds with mixing factor $\kappa \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ (see Table3.1). Flankers were missing fundamental harmonic complex tones with brightness level corresponding to the $\sigma_{mid}$ conditions.

### 3.6.3 Experimental setup and results

Five subjects (including two of the authors), aged between 25 and 30, took part in the experiment. After instruction and a short training period, each were presented 18 different stimulus classes, 30 times each. These 18 classes correspond to 3 different brightness levels (broad, medium,dark) $\times 6$ ambiguity levels (See Table 3.1). The precise timing of the presentation of stimuli in each trial is given in Figure 3.12.

For each target condition and subject, I computed the fraction of trials in which

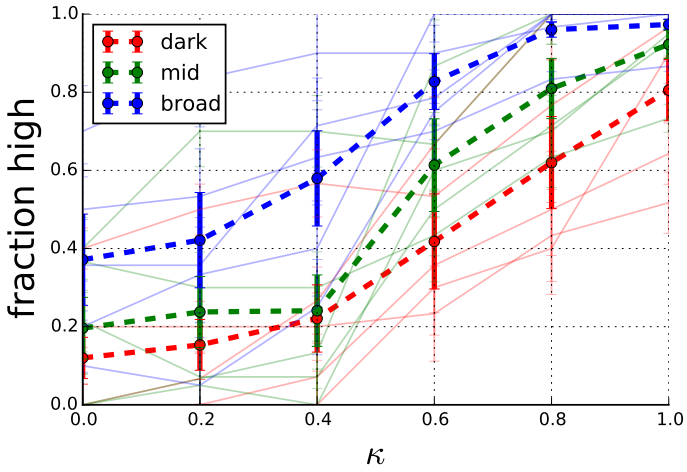


**Figure 3.13:** Psychophysical results: individual performance (lines) and average over subjects (dashed) for the 18 conditions. Colors correspond to the brightness factor.

the contour was perceived as ↗↘. A ↗↘ response implies that the pitch of the target was perceived as higher than the flankers, a ↘↗ response implies that it was perceived as lower. Furthermore, the likely pitches for any target, predicted from any pitch perception model, are either 250 or 500 Hz. Hence, I assume throughout the following that I can equate a ↗↘ responses with a pitch percept of 500 Hz ("high"), and a ↗↘ response with a pitch of 250 Hz ("low"), which are the two possible pitches reported when the ambiguous test sounds are played in isolation. I generated a set of three psychometric curves for each subject, showing the fraction of high-pitched targets as a function of the ambiguity parameter for each of the three timbre conditions (broad, medium dark). These curves were averaged across subjects and are presented in Figure 3.13.

The results qualitatively match our expectations. As the target signal get broader, the responses are biased towards "high" as revealed by the shifts in the point of subjective equality. The two extremes of the ambiguity were expected to be non-ambiguous whatever the brightness of the target. This is the case for three subjects. The two others show either permanent bias towards responding 'high', or still perceives the sounds generated from the two extremes of the octave range as ambiguous. This explains the shape of the averaged results.

## 3.7 Conclusion

In this chapter, I summarized the proposal by Phillipp Hehrmann that pitch perception could regarded as the outcome of a computational problem: inferring the periodicity of an underlying sound from its transduced activity in the auditory periphery. I presented two technical extensions. First the model was extended to provide a richer description of pitch evoking sounds by introducing the timbral notion of local periodicity. Second, a novel algorithm was derived to solve the inference problem. I demonstrated the validity of these extensions by testing the predictions of the new model on a subset of sounds for which it reproduced the pitch percept reported by human subjects.

A new psychophysical experiment demonstrated the effect of timbral brightness on pitch perception. A limitation of this work is that we restricted our analysis to stationary sounds of fixed underlying periodicity. In natural sounds such as speech, pitch varies smoothly with time. This knowledge could be added to the model and could allow to account for temporal contextual effects in pitch perception.

# Chapter 4

# Temporal contextual effects in the perception of ambiguous pitch shift

## Collaboration Statement

The work presented here is the result of a collaboration between Claire Chambers, Daniel Pressnitzer at ENS (Paris) and Maneesh Sahani and myself at the Gatsby Unit (UCL, London). Experiments were carried out by the French team. Models and analysis reported here all started after the behavioral data were collected. They were developed and carried out in London. Results of this collaboration are jointly published in [48].

## 4.1 Introduction

Chambers et al. reported a perceptual phenomenon, where prior acoustic context has a large, rapid, and long-lasting effect on a basic auditory judgment [48]. Pairs of complex tones were constructed to include ambiguous transitions between frequency components, such that listeners were equally likely to report an upward or downward "pitch" shift between tones. It was then observed that presenting context tones before the ambiguous pair could almost fully determine the perceived direction of shift. This context effect generalized to a wide range of temporal and spectral scales, encompassing the characteristics of most realistic auditory scenes. My contribution to this work is to have constructed a computational model that quantitatively explained and reproduced the behavioral results. The model proposes that

the reported bias is the side product of an underlying computation: a pre-perceptual tracking of spectrally local and temporally continuous components assumed to be the building blocks of auditory scenes. The function implements a simple constraint of spectro-temporal continuity and leads to the binding of successive sound elements in a probabilistic manner. Similar tracking mechanisms might underlie many human scene analysis computations in natural perception.

## 4.2 Contextual resolution of perceptual ambiguity: a psychophysical study

In this section, I report recent psychophysical results [48] from the PhD work of Claire Chambers. These results reveal a strong effect of prior acoustic context on the perception of ambiguous stimuli. Details of the statistical analysis of the psychophysical results may be found in [48] and are omitted from this description.

### 4.2.1 Shepard tones and Ambiguity

Shepard tones [99] are constant-interval chords parameterized by a base frequency $f_b$ and a log-normal spectral envelope $e(f)$. Given $f_b$ and $e$, a Shepard tone is constructed as a mixture of pure tones at frequencies corresponding to all powers of 2 of the base frequency $f_b$ and with amplitude given by the spectral envelope $e$. That is the waveform is $x(t) = \sum_{i\in\mathbb{Z}} \cos\left(2\pi 2^i f_b t + \phi_i\right) e\left(2^i f_b\right)$, where $\phi_i$ corresponds to the phase of each component, which were always randomized.

A fixed envelope covering all audible octaves (Fig. 4.2a, tone $T_1$) is considered throughout all experiments. All Shepard tones are described by considering a single octave of base frequencies. This is because octave-related base frequencies result in physically identical Shepard tones.

Perceptually, a sequence of two Shepard tones with a small increase (resp. decrease) in base frequency is perceived as an upward (resp. downward) step in pitch. However, a half octave shift is perceived as ambiguous, that is, on average listeners report hearing an upward shift 50% of the time.

Throughout all the following experiments, Shepard tones were used.

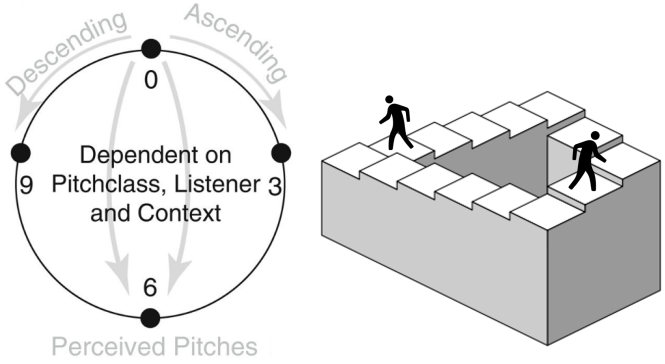


**Figure 4.1:** Pitch Chroma Circle of Shepard tones and Escher stairs. Small upward (resp. downward) shifts in pitch chroma for otherwise similar Shepard tones are perceived as such whereas $1/2$ octave shifts are ambiguous. On Escher stairs, one can try to judge whether one character is climbing upward or downward toward the other one. Ambiguous situations arise when the 2 characters are opposed on the stairs.

#### 4.2.1.1 Ambiguity

As a baseline, two Shepard tones ($T_1$ and $T_2$) were presented in close succession. The duration of each tone was 125ms. The inter-tone interval (ITI) between $T_1$ and $T_2$ was 125ms. The frequency interval between $T_1$ and $T_2$ was varied randomly across trials. Listeners reported which tone, $T_1$ or $T_2$, was higher in pitch. Replicating previous findings [99], listeners tended to report the direction of pitch shift corresponding to the shorter log-frequency distance between successive components (Fig. 4.2b). A special case occurred when the $T_1$-$T_2$ interval was exactly half an octave (six semitones): there was no shorter path favoring either upward or downward shifts. Accordingly, perceptual reports were evenly split between "up" and "down" responses [100, 101]. Although strong idiosyncratic biases across listeners have been reported previously for such ambiguous stimuli [100], the randomization of absolute base frequency across trials cancels out such biases.

### 4.2.2 Contextual effects: Resolving ambiguity

Acoustic contexts consisting in Shepard tones were introduced before the pitch-shift judgements on the test pair, $T_1$-$T_2$. The $T_1$-$T_2$ interval was held fixed at a half-octave, the ambiguous case.

#### 4.2.2.1 Single contextual Shepard tone

In its simplest form, a single context tone, C, was played immediately before the test pair. If context did not matter, responses should be evenly split between "up" and "down". In contrast, it was found that there was a strong influence of the context tone on perceptual reports. Listeners tended to report the shift encompassing the frequency components of the context tone, with maximal bias for a context tone located halfway in between $T_1$ and $T_2$ (Fig. 4.2d).

#### 4.2.2.2 Multiple contextual Shepard tones

In the same experiment, the number of context tones was randomly varied between 0 and 10 context tones. The relative chroma of context tones were randomly drawn, uniformly, from one of two half-octave frequency regions: only positive intervals or only negative intervals relative to $T_1$ (Fig. 4.2e). Perceptual responses were summarized by computing $P(Bias)$, the proportion of time listeners responded with a bias in the direction expected from Fig. 4.2d., the single tone context case. $P(Bias) = 1$ would correspond to listeners always reporting pitch shifts encompassing the frequency region of the context tones, whereas $P(Bias) = 0$ would correspond to listeners always reporting the opposite direction of pitch shift. An absence of context effect, that is, a response probability unaffected by the context, would correspond to $P(Bias) = 0.5$.

The strength of the context effect increased with the number of context tones (Fig. 4.2f). Remarkably, after about five context tones were presented, almost all perceptual reports were fully determined by the preceding context. For the exact same ambiguous test pair, listeners went from randomly reporting up or down shifts to almost invariably reporting the same direction of shift with context.

It is important to note that the sequence of "up" and "down" pitch shifts during the context was randomly varied across trials (see Fig. 4.2e for one instance). Rather, the frequency content of the context (more precisely its pitch chroma) relative to the test was the experimental variable. This resolves previously conflicting reports concerning contextual processing with Shepard tones. Attempts to adapt-out the direction of frequency-shifts only resulted in weak and unreliable contrastive

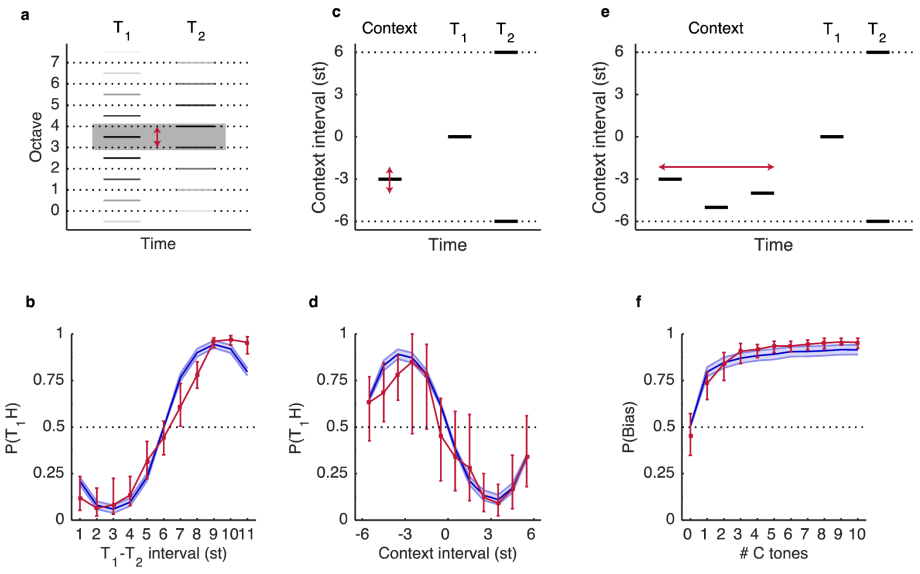


**Figure 4.2:** Ambiguous test pairs and context effects. **(a)** Schematic spectrogram of the $T_1$-$T_2$ test pair (amplitude coded as gray scale). $T_1$ and $T_2$ are "Shepard tones" with octave-related components. For the interval of half-an-octave (6 st) represented here, all components of $T_1$ are exactly halfway in between two components of $T_2$ on a log-frequency scale. Here and in remaining stimulus plots, red arrows indicate the variable manipulated in the experiment. **(b)** Perceptual judgements without context. Here and elsewhere behavioral results are shown in red and are overlaid with simulated results shown in blue. The proportion of "$T_1$ higher" responses, P($T_1$H), is plotted as a function of interval between $T_1$ and $T_2$ (n=11 listeners). Here and in all subsequent figures, error bars show 95% confidence intervals. The dotted line at 0.5 indicates the point of subjective indifference, with as many "up" and "down" responses for the same physical stimulus. **(c)** Example trial with a single context tone preceding the ambiguous test pair. For clarity, the illustration is restricted to a one-octave range (gray patch of panel A) but actual stimuli included many frequency components, with the same arrangement in all audible octaves. **(d)** The P($T_1$H) is shown as a function of the interval between C and $T_1$ (n=11 listeners). Without context effects, all responses would be at 0.5. **(e)** Example trial with multiple context tones. Listeners reported the perceived shift between $T_1$ and $T_2$ only. **(f)** The proportion of reporting a shift encompassing the frequency region of the context tones, P(Bias), is shown as a function of the number of context tones (n=11 listeners, red). In the baseline condition with no context tone, P($T_1$H) is displayed

context effects [102, 103], whereas studies which, retrospectively, can be understood as having manipulated the frequency relationship between context and test, found strong context effects [45, 46, 101].

#### 4.2.2.3 Generalization over time-scales

It was then examined whether the context effect observed could generalize to longer and shorter time scales. In the case of the biasing of ambiguous visual-motion [30], it has been observed that attractive effects occurred for short time scales, whereas contrastive effects occurred for longer time scales. It was thus possible that the auditory context effect reverses, or disappears, depending on the time scale of the stimulus. How fast can the context effect be established? A single context tone was used and its duration was varied between 5$ms$ and 320$ms$, with the duration of the test tones $T_1$ and $T_2$ maintained at 125$ms$ (Fig. 4.4a). The same direction of perceptual bias was observed over a broad range of context tone durations (Fig4.4b). The bias increased with context duration and saturated between context durations of 160ms to 320ms and I expect longer durations to even more strongly determine the direction of subjective reports.

The authors also addressed the complementary question of how long the bias persisted, once established. Five context tones were presented, each 125$ms$ long, followed by a silent gap and then a test pair (Fig. 4.4c). The gap was varied between 0.5$s$ to 64$s$, during which attention was not controlled. Predictably, the bias decreased with increasing gap duration (Fig. 4.4d). In the individual data, for some listeners, there was in fact very little decrease in bias between 0.5$s$ and 64$s$. Thus the present context effects covers a wide range of time scales; it is induced with a context as short as 20$ms$ but persists over interruptions up to 64$s$ for some listeners. For comparison, contrastive effects for speech recognition have been demonstrated for a context as brief as about 300$ms$ [104] and for interruptions of up to 10$s$ [105, 106]. Shorter time constants have been found for the spectral motion aftereffect, which can be observed with inducers as short as 100$ms$ [42, 43]. The long time-constants found here are in fact reminiscent of what has been termed “storage” for visual aftereffects [107], but not demonstrated in audition. The present auditory context effect thus covered an unusually broad range of temporal parameters. Importantly, the fact that context effects can be both rapidly established and persist for a long time shows that their underlying mechanisms may operate for most everyday audi-

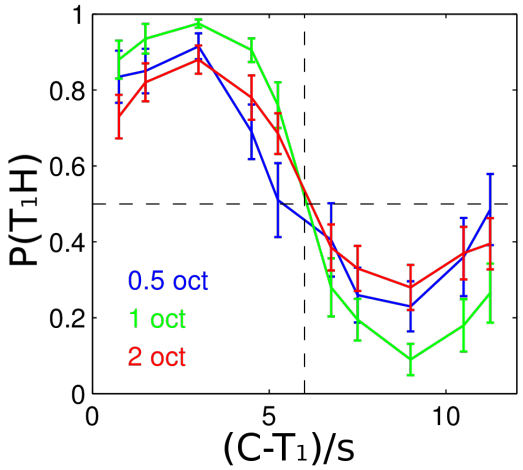


**Figure 4.3:** Generalization over frequency scale. P($T_1$H) is shown as a function of the octave-scaled interval between C and $T_1$.

tory scenes. For instance, the median duration of short segmental cues in natural speech, such as unstressed vowels and consonants, is about 70*ms* [108]. This is longer than the minimal duration of 20*ms* required for observing context effects. Conversely, a persistence of 32*s* as observed is enough to accommodate prosodic cues [108] and even pauses between sentences. Similar time scales may be found for music [109]. The time scales covered by the context effects thus encompass those typical of speech and music.

#### 4.2.2.4 Generalization over frequency scales

The individual pure tones constitutive of a Shepard tone are octave-related: ratios of their frequencies are powers of 2. The octave interval is fundamental to the definition of the pitch chroma [110] but Shepard tones can be generalized to arbitrary intervals. Is the context effect tied to the octave interval? To answer this question, the first contextual experiment with a single context tone (as described in sec. 4.2.2.1) was reproduced using Shepard-like tones with intervals of $\frac{1}{2}$,1 and 2 octaves. The shape of the bias was preserved irrespective of the scale as shown in Fig. 4.3, demonstrating the invariance of the context effect to the interval. Results of this experiments were separately reported in [111].

So far, highly-constrained Shepard-like complex tones were used as a tool to probe and highlight contextual processing, but it was important to test whether the

observed effects could generalize to other types of sounds. In another experiment, sounds with as little constraints as possible with respect to their frequency content were used: sounds with completely random spectral components. The density of the random spectra was systematically varied, from very dense with approximately 8 tones per octave, to very sparse with approximately 0.3 tones per octave. The only constraint was to enforce ambiguous frequency shifts in the final test pair, as ambiguity in this test pair was critical to the paradigm being sensitive to context effects.

The stimuli are illustrated in Fig. 4.4e. Context sequences were constructed so that frequency components of the context would be expected to favor only one direction of shift within the test pair (this expected direction of shift encompasses all the context tones). Results showed that all random spectral stimuli produced a large bias, from the sparsest to the densest, with a decrease for the densest condition (Fig. 4.4f).

The context effect was observed for random spectra stimuli, which were completely different form the Shepard tones used so far. Moreover, the effect was robust when tested over different frequency scales, from very sparse sounds to very dense sounds. It is likely that in the densest case of 8 components per octave, which showed a decline of the magnitude of effect, the overall spectral pattern was starting to become unresolved within auditory cortex [112]. The only limit for the context effect in terms of frequency content thus seems to be to stimulate non-overlapping frequency regions. A similar robustness to spectral scale was found in the standard speech contrast effect [113].

This generalization to arbitrary sounds again suggests that the mechanisms underlying the context effect could apply to typical natural auditory scenes. Given that the context effect was just as strong for random spectra as for Shepard tones, the spectral shape of the sounds should not matter at all. The limiting factor seems to be spectral density, and these results show that at least the first 8 harmonics of any periodic sound, such as the voiced parts of speech or musical instrument sounds, would fall within the existence region of the context effect.

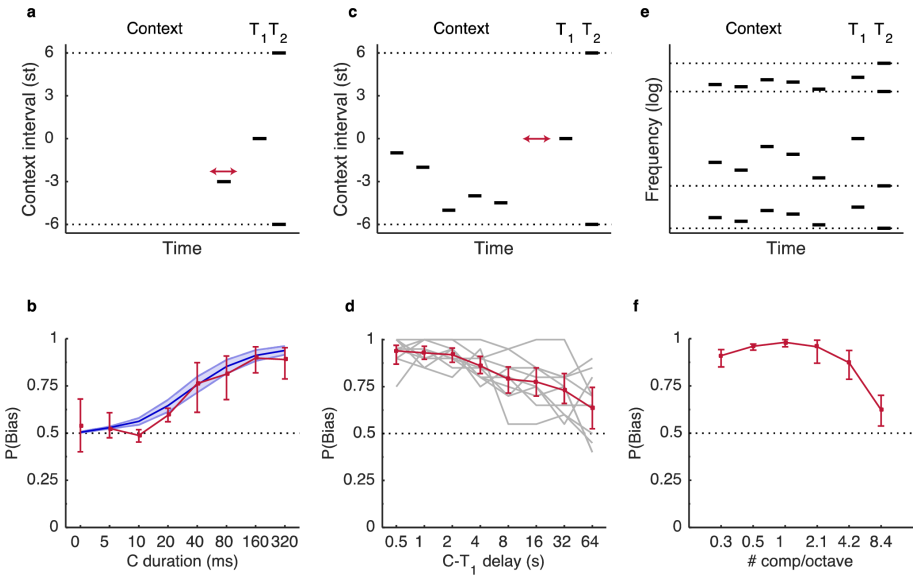


**Figure 4.4:** Generalization over time and frequency scales. **(a)** Example trial testing the effect of the duration of a single context tone. **(b)** The mean $P(Bias)$ is displayed as a function of the context tone duration ($n = 10$ listeners). In the baseline condition without a context tone, P($T_1$H) is displayed. **(c)** Example trial testing the effect of a silent gap between context and test. **(d)** The mean $P(Bias)$ is displayed as a function of the C-$T_1$ silent gap ($n = 10$ listeners). Individual listeners are shown by gray lines. **(e)** Example trial testing random spectra. Components of $T_2$ were distributed randomly, but each exactly in-between two components of $T_1$. The context tones components were restricted to favor only one possible direction of shift between $T_1$ and $T_2$. Dotted lines represent one frequency "cycle" of the stimulus, they would be equally spaced at one octave for Shepard tones. **(f)** The mean $P(Bias)$ is shown as a function of the number of components per octave ($n = 10$ listeners).

#### 4.2.2.5 Hysteresis

When Shepard tones are presented in ordered sequence of increasing or decreasing intervals, a strong hysteresis effect was reported by the same authors [45]. The task involved the sequential presentation of pairs of Shepard tones with a fixed standard reference tone and a tone of either increasing or decreasing interval relative to the reference tone, with a $1st$ increase (resp. decrease) at each appearance from 1st to $11st$. The order of presentation within the tone was randomized so that the hysteresis could not be confounded with a response bias. It was found that in the upward (resp. downward) sequence, the sequential presentation shifted the interval of subjective ambiguity from $6st$ to $10st$ (resp. $2st$), a classical signature of hysteresis. A condition where increasing tones were shuffled was also conducted as a compari-

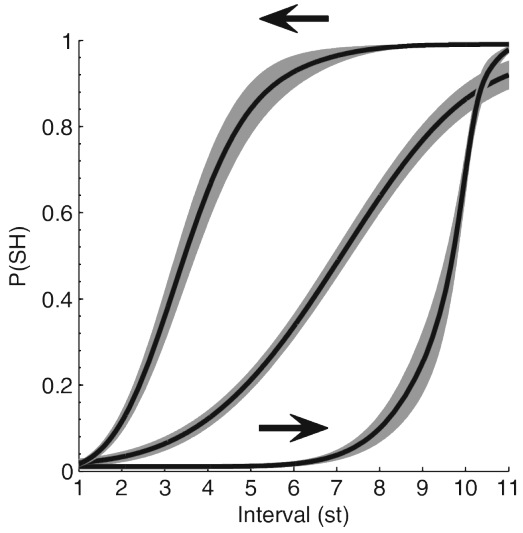


**Figure 4.5:** Results for random, increasing, and decreasing presentation orders from [45]. Probability standard higher ($P(SH)$) as a function of the interval between the standard and comparison tones. Fitted curves are shown for the random (middle curve), increasing (rightward-facing arrow), and decreasing (leftward-facing arrow) conditions, with the shaded areas displaying the standard errors of the means.

son. No hysteresis was found in this random condition. A summary of subject's behavior in all 3 conditions is shown in Figure 4.5.

### 4.2.3 Previous attempts at explaining the contextual effect

Two studies attempted to provide a causal understanding of the contextual effects reported here.

#### 4.2.3.1 Neural decoding

A first study explored if the biasing effect of context could be read out from early sensory cortex in ferrets [114]. The authors recorded neural activity in auditory cortex while they listened to isolated Shepard tones of various chroma. They then regressed the pitch chroma on the recorded neural activity. The key manipulation was then to predict the pitch chroma from the neural activity resulting from the sensory experience of Shepard tones closely following a context as in the single context tone psychophysical experiment. Given the attractive nature of the behavioral result in humans, and the working hypothesis that the shift decisions in human were the result of a comparison of encoded chroma, the authors expected this contraction to

be reflected in the predicted pitch chroma. Their prediction was that the decoded chroma would be attracted towards the chroma of the context. The opposite turned out to be true: the decoded pitch chroma was pushed away from the chroma of the context. This left the link between psychophysical phenomenon and its neural correlates unexplained.

#### 4.2.3.2 Neuro-mechanistic (bottom-up) model

A second study proposed a "neuro-mechanistic" model of the biasing effect of context [115]. The model is based on the assumption of the existence of a neural population encoding of frequency shifts and of its role in shift perception. This assumption is motivated by psychophysical studies revealing the existence and properties of "frequency shift detectors"[116, 117, 40] and physiological evidence for tonotopic maps of direction selective neurons in auditory cortex [118].

The model describes the dynamical evolution of the firing rates of a recurrent neural network - composed of excitatory and inhibitory populations - as a system of first order non-linear differential equations in the spirit of the Wilson and Cowan model [119]. The network contains two tonotopically organized, excitatory populations, $E_{up}$ and $E_{down}$ , that respond preferentially to ascending or descending stimuli in pitch, respectively. These preferences are generated by a third inhibitory population $I$ (also tonotopically organized) that provides inhibition asymmetric in frequency to the two populations (see Fig. 4.6). The three populations have the same tonotopic external input: input tones are modelled as symmetric bump of excitation centered on the tone frequencies on the log-frequency axis.

As an example to understand the direction selectivity, input from the second tone of an ascending tone pair will fall in a region inhibited by the first tone in the downward population, but left unchanged by the first tone in the upward population. To model the long-term effects of contextual tones, a slow facilitation of inhibitory synaptic weights is added: tonotopic regions that were previously inhibited are more inhibited later on by similar inputs. Essentially the model relies on the separation of two timescales: the fast inhibition leads to a direction-selective population for pairs of closely followed tones while the slow facilitation enhances and biases the

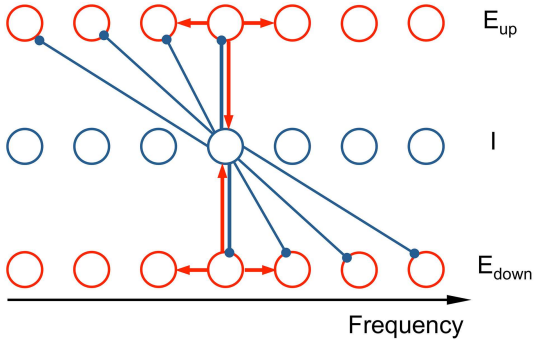


**Figure 4.6:** Neuro-mechanistic model (Huang et al)[115]. The network model consists of two excitatory populations ( $E_{up}$ and $E_{down}$ ) and an inhibitory population ($I$), tonotopically organized. The asymmetric inhibitory feedback leads to an ascending/descending frequency change preference for the $E_{up}$ and $E_{down}$ populations, respectively. Each unit is a local subpopulation, positioned at its characteristic frequency (CF). Red arrows signify recurrent excitation and blue arrows inhibition. The subset of the connections shown illustrates the architecture's qualitative nature: the synaptic footprints from $E$ to $E$ and from $E$ to $I$ are narrow and symmetric; from $I$ to $E$ the footprint is broad and asymmetric.

fast inhibition.

This model can reproduce many aspect of the psychophysical results presented so far. These predictions are described in details in [115, 111]. Here I describe two predictions where its predictions do not match the observed psychophysical results, or where the predictions do not appear satisfying.

- Using fixed tuning of frequency shift detector cannot explain the generalization over frequency scales. Indeed, when the interval between 2 consecutive frequency components gets large, inhibition following the first tone may have a local effect on the network around that does not extend as far as to bias the activation due to the presentation of the first tone. This model would predict no bias for the conditions with large intervals (bigger than octave) of the experiment described in 4.2.2.4.
- The model can explain the effect of past stimuli on perceptual judgements up to durations set by a parameter controlling the decay of the synaptic facilitation of inhibition which to my knowledge has not been observed in neural systems. In [115], it is set to 2 seconds which would not explain the persistence of the bias for up to more than 30 seconds as reported in [48]. This

parameter could be set to a larger value but it is then unclear which biological mechanisms could support such a long facilitation. However, timescales of learning and adaptation in biological neural networks and their underlying mechanisms are still largely unknown so the long time-scale used in [48] might well turn out to be realistic.

In the next section, I will explain the motivation and the modelling approach I took to explain the perceptual phenomena. Instead of a bottom-up approach, I took a normative approach explaining the bias as the consequence of the resolution of a computational problem: the pre-perceptual tracking of components of auditory scenes.

## 4.3 Computational model of pre-perceptual grouping

### 4.3.1 Motivation

What computational function might be served by the kind of contextual processing revealed by the behavioral biasing effects? Many otherwise surprising perceptual phenomena may arise from computational principles that reflect expectations derived from learned statistical properties of the natural world [6]. I constructed and simulated an inferential model to ask whether the same might be true of the context effects documented earlier.

The feature that listeners reported in those experiments – the direction of pitch shifts – necessarily depends on the comparison of sounds over time. Because the context effects were observed for random-spectra stimuli (Fig. 4.4f), which do not produce a unitary pitch percept, it is likely that listeners reported frequency shifts between successive frequency components. The core idea of my model was that prior context may inform which successive frequency components were bound together, and therefore, which successive components were compared to estimate perceptual features such as pitch-shifts. The way prior context informed temporal binding was by assuming some degree of spectro-temporal continuity in the acoustics of sound sources: current frequency components are likely to be followed by

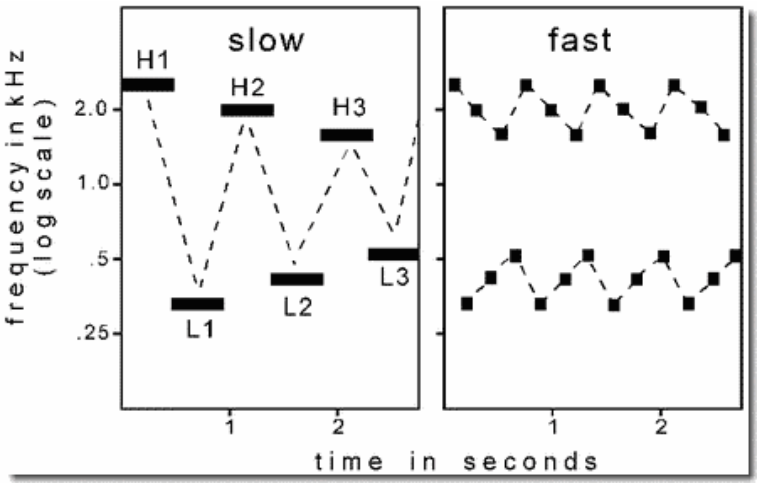


**Figure 4.7:** Auditory Stream segregation in a cycle of six tones. From Bregman [120] The sequence used in this example consists of three high and three low tones in a six-tone repeating cycle. Each of the high tones (H1, H2, H3) and the low tones (L1, L2, L3) has a slightly different pitch. The order is H1, L1, H2, L2, H3, L3, .... An alternation of high and low tones is heard when the cycle is played slowly. Two 'streams' of sound, one formed of high tones and the other of low ones, each with its own melody, as if two instruments, a high and a low one, were playing along together is heard when the cycle is played fast. In both case, dashed lined reflect the perceptual organisation of the auditory scene.

future components at the same or nearby frequencies, because of the persistence in the characteristics of sound sources. This assumption is reminiscent of a qualitative explanation of streaming in auditory scene analysis [120, 121]: nearby elements in time and frequency are bound together to form auditory objects. This can be thought of as analogous to the formation of a visual contour by connecting edges. My assumption is that a similar process might be happening at a pre-perceptual level and for simple auditory features such as pure tones.

It can already be seen that this simple idea qualitatively accounts for the context effects, as the pitch direction reported by listeners always maximized spectro-temporal continuity (Fig. 4.2b)[45, 46, 101].

### 4.3.2 Model description

This hypothesized process was implemented as inference within a probabilistic generative model of auditory scenes (Fig 4.9). Inference in this model took as input a set of frequencies at different times, and assigned each observed frequency to what I termed a "track". In the generative framework, the log-frequencies of individ-

ual components were taken to be normally distributed with variance $\sigma^2_{track}$ around the center frequency associated with a track. The center frequency of each track could evolve slowly through time. I also assumed that the internal representation of the frequency of each component was independently corrupted by sensory noise, with variance $\sigma_{sens}(d) = \sigma_{sens}\sqrt{d_0/d}$ inversely proportional to the duration $d$ of the tone, with $d_0$ an arbitrary constant (See appendix A for a justification). The two variances, $\sigma^2_{track}$ and $\sigma^2_{sens}$, were the two free parameters of the model.

Responses to experimental stimuli were simulated by inferring the evolution of the model tracks, starting from the first context tone and ending with the final test tone. Tracks were initialized for each component of the first context tone, with a normal posterior belief about each center frequency, the mean of which was set to the (noise-corrupted) component frequency and the variance of which was set to $\sigma^2_c = \sigma^2_{sens} + \sigma^2_{track}$.

Inference for each subsequent tone then followed two steps: (1) the noisy component tones were probabilistically assigned to the tracks, in proportion to their probabilities of generation from them; (2) beliefs about the track center frequencies were updated, according to the evidence provided by the assigned components weighted by their probabilities of assignment. Formally, this approach corresponded to mean-field filtering in a factorial hidden Markov model [122]. These filtering steps were iterated for each tone in stimulus sequence. To compare the model with behavioral data, I finally predicted the perceived pitch shift between test tones by summing the shifts between every possible combination of test tone components, weighted by the inferred probability that the combination originated within the same track. The predicted pitch shifts thus reflected track assignments carried over from the first tone of the context sequence, and favored pitch shifts within tracks (Fig 4.9).

The model processed sound sequences in three stages: initialization, tracking, and construction of the overall shift percept. I describe each of these steps more formally below along with the fitting procedure used.

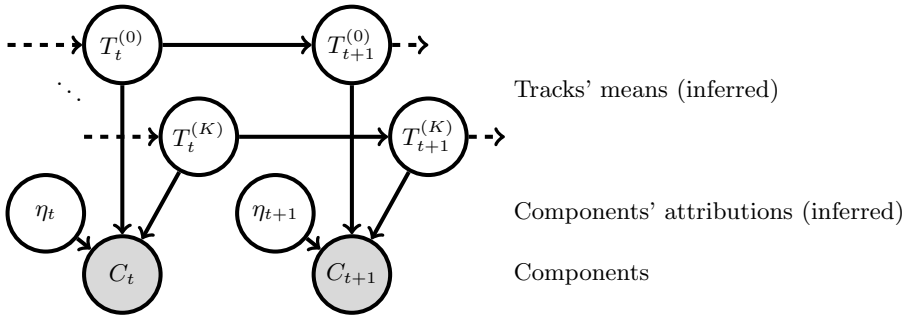


**Figure 4.8:** Graphical model of the generative process assumed to underlie auditory scenes is displayed. Arrows denote statistical conditional dependencies. Horizontal arrows describe the temporal continuity of tracks (T) that are assumed to generate the spectro-temporal components (C) constituting the scene. Each component belongs to a single track whose identity ($\eta$) is inferred.

#### 4.3.2.1 Generative model

Assumptions of spectro-temporal continuity of tracks are captured in the structure and parameterization of a generative model for auditory scenes. The model is a factorial hidden Markov model [122]. Variables in the model are (1) track means $T_t^{(i)}$ where $t$ indexes time and $i$ the identity of the track, (2) chords $C_t$ whose individual components $C_t^{(j)}$ are pure tones, (3) for each tone $C_t^{(j)}$, its originating track $\eta_t^{(j)}$. In the model, the track means evolve independently in time according to a random walk. Tones are independently drawn according to a Gaussian distribution centered at their randomly sampled group mean and with a fixed standard deviation corresponding to the typical group spread.

More formally, the hierarchical generative model of the scene can be written as

- $p(T_0^{(i)}) \sim \mathcal{N}\left(T_0^{(i)};.,\sigma_{track}^2+\sigma_{sens}^2\right)$: the prior on track mean, where the mean depends on tone that initiates the track, as described in the next section.
- $\eta_t^{(i)} \sim Discrete(\pi)$ : prior on track attribution $\pi_i \propto 1$
- $C_t^{(j)}|T_t^{(1..K)}, \eta_t = k \sim \mathcal{N}\left(C_t^{(j)};T_t^{(k)},\sigma_{track}^2+\sigma_{sens}^2\right)$: tone observation conditional on track identity
- $T_{t+T}^{(k)}|T_t^{(k)} \sim \mathcal{N}\left(T_{t+T}^{(k)};T_t^{(k)},\gamma^2\right)$: random drift in time of the track mean

A graphical representation of this generative is given in Figure 4.8.

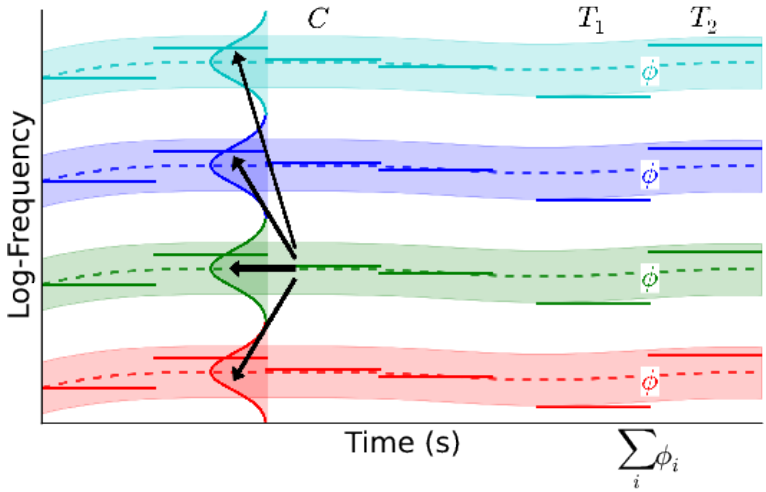


**Figure 4.9: Probabilistic model**. A track (color coded) is instantiated for each of the component of the first context tone. Track means (dashed lines) are updated after each attribution for subsequent context tones. The colored patch indicates the standard deviation of the underlying generative process. An attribution step is illustrated for the 3rd tone of the context sequence (black arrows); arrow weights indicate the probability of attribution for each stream. Using the same procedure, the final test tones components are attributed. Perceptual features (here, pitch shifts $\Phi_i$) are finally computed within tracks. Green arrows represent all possible bindings from one component tone of $T_1$ to component tones of $T_2$; arrows thickness represents the likelihood that this binding was generated by the green track. The most likely binding correspond to a pitch shifts encompassing the context tone components, consistent with the perceptual bias.

#### 4.3.2.2 Parameters and Initialization

When a chord consisting of several components is given as input to the model, it initiates one track per component tone, but maintains uncertainty in the form of a Gaussian distribution about the central frequency of the track. The distribution is centered on the observed tone frequency and has a variance that is equal to the total variance that would be expected in the sensed frequencies of tones associated with that track. This variance is the sum of two parts $\sigma_c^2 = \sigma_{track}^2 + \sigma_{sens}^2(d)$; the variance of acoustic frequencies associated with any one track ($\sigma_{track}^2$) and the variance of the sensory noise that corrupts the sensed frequency ($\sigma_{sens}^2(d)$), depending on the tone duration, as described earlier.

#### 4.3.2.3 Tracking

As new chords are input into the model, the component tones are attributed to the different tracks. This attribution process corresponds to inferring the underlying track assumed to be associated with each tone. It requires the ability to predict the continuation of each track and to assess the likelihood of new observation to belong to this track. Such tracking mechanisms and how there dependence on distance between tracks have been reported in studies of attentional tracking of a source in sound mixture [123]. Here I assume that multiple tracks are inferred in parallel in a process akin to multi-target tracking with multifocal attention in vision [124]. In the auditory domain, there is evidence for the ability to track multiple sound source in music but less so in more challenging natural situations [125] The belief about the mean of each track is then updated to incorporate the new observed frequencies. Tone attribution is a "soft" process, each tone is partly attributed to all tracks with a probabilistic weighting called a "responsibility". Tracks closer in frequency to a given component tone assume greater responsibility for it. Specifically, the responsibility is given by the probability under the model that a given tone with frequency $g^{(j)}$ might have arisen from the distribution of frequencies associated with track $i$. I introduce an attribution label $\eta^{(j)}$ which is the (unknowable) identity of the track which actually generated tone $j$. Then the responsibility is just the probability that $\eta^{(j)} = i$. That is, if ongoing beliefs concerning the mean frequency are defined by $\left\{\left\{\mu_t^{(1)}, \sigma_t^{(1)}\right\}, \ldots, \left\{\mu_t^{(K)}, \sigma_t^{(K)}\right\}\right\}$ and the chord $c$ is presented, I compute for each tone $j$ and each track $i$, the responsibility $r_j^i$ as follows.

$$\begin{aligned} r_j^i &= p\left(\eta^{(j)} = i | g^{(j)}, \left\{\mu_t^{(i)}, \sigma_t^{(i)}\right\}\right) \\ &\propto \mathcal{N}\left(g_j; \mu_t^{(i)}, \sigma_c^2\right) \exp\left(-\frac{1}{2}\frac{\sigma_t^{(i)2}}{\sigma_c^2}\right) \end{aligned}$$

Once attribution has taken place, the mean frequency of each track is updated with the frequencies of all tones, weighted by the responsibilities. I update the beliefs about the ongoing tracks as follows: for each track $i$, I compute the effective number of tones attributed to that track, $n^{(i)} = \sum_k r_k^i$ and the weighted mean frequency of the

tones attributed to that track, $v^{(i)}$

$$v^{(i)} = \frac{1}{n_i}\sum_k r_k^i g_k$$

Mean and variance of the belief about track $i$ are updated as

$$\sigma_t^{(i)2} \leftarrow \left(\frac{1}{\sigma_t^{(i)2}} + \frac{n^{(i)}}{\sigma_c^2}\right)^{-1}$$

$$\mu_t^{(i)} \leftarrow \left(\frac{1}{\sigma_t^{(i)2}} + \frac{n_i}{\sigma_c^2}\right)^{-1}\left(\frac{\mu_t^{(i)}}{\sigma_t^{(i)2}} + \frac{v^{(i)}}{\sigma_c^2}\right)$$

Finally, since the prior belief about the dynamics of track means is that of a Wiener Process, variances of the beliefs about each tracks is incremented by $\gamma^2\delta t$ where $\delta t$ is the inter-onset interval between Shepard tones and $\gamma$ is the assumed rate of change of track means. A wide range of values led to quantitatively similar fits for slow time-scales of variation up to approximately a $10^{th}$ of an octave per second. These correspond to slow variations relative to the overall duration of the context and the test pair. For this reason, I set this parameter to zero for all the results reported.

This process of attribution and updating is repeated for all remaining stimuli in the trial. Full detail of the derivation of the filtering updates can be found in Appendix B.

#### 4.3.2.4 Shift percept construction

Finally, I modelled the behavioral response of the listener when judging the direction of pitch shift between a pair of consecutive chords. Frequency shifts were computed locally within each track, and these local shifts were then combined across tracks to build a global percept of pitch change. The local frequency shift within a track was taken to be the sum over all possible oriented shifts between pairs of consecutive tones in the two chords, weighted by how likely they were to both belong to that track. For track i,

$$\phi_i = \sum_{j_1,j_2} r_{j_1}^i r_{j_2}^i \left(g^{(j_2)} - g^{(j_1)}\right)$$

The global shift percept was then simply the sum of the track-local shifts:

$$\phi = \sum_i \phi_i$$

A binary percept was constructed by thresholding $\phi$ at 0. When $\phi$ is positive, a rising pattern is predicted, and when $\phi$ is negative, a falling pattern is predicted.

### 4.3.3 Results

#### 4.3.3.1 Fitting procedure

I generated model predictions for psychophysical experiments, where Shepard tone pairs were presented without context (Experiment 1), where one context tone was presented before the ambiguous tone pair (Experiment 2), where several context tones were presented (Experiment 3) and where the duration of one context tone was varied (Experiment 4). In order to assess the performance of the model relative to the behavioral data, I estimated the maximum-likelihood parameters of the model. Participants performed different subsets of the full set of experiments, with four participants having completed Experiments 1, 2, and 4; seven having completed Experiments 1 and 2, and six participants having completed only Experiment 4. I took full advantage of the data available by estimating one set of the parameters, $\sigma_{sens}$ and $\sigma_{track}$, for each individual using the data from all the experiments that the listener completed. The log likelihood provided an estimate of the fit of the model predictions to the psychophysical data. The parameter $\sigma_{track}$ was estimated in the range of 0.1-2 (up to two octaves) and $\sigma_{sens}$ was estimated in the range of 0.01-0.5 (up to half an octave). As can be seen in Figure 4.10, the value of the parameter $\sigma_{track}$ has little effect on the fit to the data up to 0.5 octaves, which is the frequency span of the context region. Therefore, to generate the simulations reported here, an arbitrary value in this range was selected for $\sigma_{track}$, which was fixed at 0.16 across participants, and the parameter $\sigma_{sens}$ was selected to maximize the likelihood of each individual response for all tasks and conditions. An estimate of the uncertainty in the parameter value for $\sigma_{sens}$ was provided by the 95% confidence interval around the mode of the normalized likelihood, displayed in Figure 4.11.

#### 4.3.3.2 Results and interpretation

The resulting model predictions are shown superimposed on the behavioral data in Figs. 4.2 and 4.4, and in most cases the confidence intervals between behavior and model overlap. The model thus provides a single interpretative framework for most of the behavioral data, which I now detail. Without context, tracks were initialized at the components of the first test tone $T_1$, and the highest probability was for each track to bind the component in the following tone $T_2$ that was closest in log-frequency, hence favoring the pitch shift over the smallest frequency distance between $T_1$ and $T_2$ (Fig. 4.2b). In the ambiguous case corresponding to an interval of 6 st, each component of $T_2$ became equidistant on average from two tracks originating from $T_1$; the symmetry was broken randomly by the simulated sensory noise, and so either shift direction was favored equally often (Fig. 4.2b). Introducing a context tone made it more likely that the tracks initiated by the context would capture their neighboring $T_1$ and $T_2$ components, and thus favor a pitch shift encompassing the context frequency region. The predicted bias was strongest when the context components, and thus prior track centers, fell halfway between the components of $T_1$ and $T_2$ as then the probability that the sensory noise would disrupt the context-induced tracks was smallest (Fig. 4.2d). Adding context tones increased the confidence in the track mean value (and their influence was more likely to average around the half-way point), thus increasing confidence in the assignment of $T_1$ and $T_2$ tones, consistent with the build-up of the effect with the number of context tones (Fig. 4.2f). The decrease in assumed sensory noise associated with longer tones was finally consistent with the effect of tone duration on context (Fig. 4.4b).

#### 4.3.3.3 Model Predictions for the other experimental conditions

Four of the experiments presented in 4.2 were left out of my model based analysis and discussion: (1) the case of increasingly long delays between the context and the test pair (2) the case of uneven and complex tones used to build both the context and the test pair. (3) the case of even but non octave Shepard-like tones used to build both the context and the test pair. (4) the hysteresis effect.

My model was derived at a computational level and it's parameters constitutes

beliefs about statistical properties of the environment. In some of the conditions listed above, these believes are too simple, in the sense that they no longer reflect the statistical properties of the stimuli well enough.

Here, I discuss how my computational model can either readily account for these results or how it would need to be extended to account them. In all these conditions, the stimuli are organised in 'tracks' of bands or well separated bands of energy in the spectro-temporal domain. Intuitively, my account of the behavior in these conditions is similar to that given for the main experiment: the model successfully performs tracking and the expectations build from track bias the judgment of ambiguous (or close to ambiguous) test pairs. All the extensions described here include the initial model as a separate case. A single of extension of the original model could in theory explain all conditions at once, but this possibility is not explored here.

## (1) Long delays

The only source of stochasticity in the model in the sensory noise. Longer delays after the context lead to more uncertain ('broader') tracks at the time the test pair is presented. What drives the bias in the model is the relative closeness of $T_1$ to the tracks infered from the context. This is left unchange by long delay. Essentially, the model predicts a bias that would persist for very long time, until the tracks are so broad (covering many octaves) they almost completely overlap.

A first possible change to the model would be to continuously inject an independent memory noise in the model's beliefs about tracks. This would make the model slowly 'forget' about the context as noise is added (with long context-test delays) while maintaining unbiased performance unchanged. A second possible change to the model would be to add a mechanism whereby uncertain tracks would have a probability to disappear. Indeed in the model, neither creation nor deletion of tracks is directly addressed although this is a critical component of any serious scence analysis algorithm. With such a time dependent deletion mechanism, the bias would decrease with longer context-test delays)

### (2) Uneven complex tones and (3) Shepard-like tones

For simplicity, the model presented assumes a shared fixed generative width for tracks $\sigma_{track}$. This was convenient because that is indeed how the 'underlying' tracks are in most of the contexts I designed with Shepard tones. When using uneven complex tones (with different consecutive intervals instead of octave ones between consecutive pure tones in the mix), the 'underlying' tracks would each have a different typical width. An extension that would accomodate for these different widths and unknown would be for the model to also infer the width in an online fashion. Such a model has been explored but is not included in the thesis. It can account for the observed biases in both the uneven and the even condition. It also naturally accounts for the biases reported in the case of pairs of ambiguous non-octave Shepard tones.

### (1) Hysteresis

The results for the hysteresis experiment are short term effects where the last few tones set the bias. One can see the 'contexts' in this task as multiple 'underlying' tracks with gradually increasing means. My model would readily perform tracking for these contexts and explain the observed response patterns. One parameter $\gamma$, the assumed rate of change of track means, might need to be adjusted to match the faster rate of change of the 'underlying' track means in this task.

#### 4.3.3.4 Discussion

Using a single pair of parameters per subject, I achieved excellent quantitative agreement across the range of basic experiments. To maintain focus on the core mechanism, I did not attempt to model the data for the gap duration and random spectrum experiments, which would have required extensions to the computational model as described in 4.3.3.3. The success of the simple version of the model nevertheless supports the hypothesis that the context effect was based on temporal binding processes, enforcing simple statistical constraints of spectro-temporal continuity. Models with similar structure have been suggested before for auditory scene analysis [126, 127]. Indeed most auditory scenes are mixture of auditory ob-

jects separated in some feature space and in these models, online inference of the objects amounts to tracking. The present framework is different in what it aims to represent: the tracks identified by the model here do not need to correspond to perceptually separated streams, which can be attended to at will [128]. Instead, many parallel tracks may be formed, implicitly, to guide the estimation of task-relevant perceptual features. The model I proposed here is computational and agnostic about the implementation. Its main purpose is to propose an abstract understanding of the function that causes the observed patterns of biases. This understanding allows to create bridges with other fields (auditory scene analysis), to form and explore alternative predictions to these that would stem from a model of a different nature, for example a mechanistic one.

## 4.4 Conclusion

Behavioral data showing that prior context can have a profound influence on a simple auditory judgement of pitch shift was reported. Perceptual decisions could be fully swayed one way or another depending on prior context, for physically identical sounds and for pitch shift values far from threshold. The existence region of the context effect, for time scales and spectral scales, was shown to encompass the prevalent statistics of natural auditory scenes. A probabilistic model provided a functional interpretation for the context effects, in the form of temporal binding under the constraint of spectro-temporal continuity.

My functional interpretation of the context effect is cast in terms of binding, and not auditory streaming, in spite of the apparent resemblance between the two notions [120, 129]. This is because the behavioral effect was observed in conditions not usually associated with subjective streaming. An auditory stream takes some time to build up [130], and it breaks apart for long pauses [120] or large frequency discontinuities [121]. In contrast, contextual effects in very short sequences, after very long silent intervals, and in the perception of very large frequency shifts were observed. The current results thus reveal a form of binding process outside of the parameter range of subjective streaming. This generalization is needed when

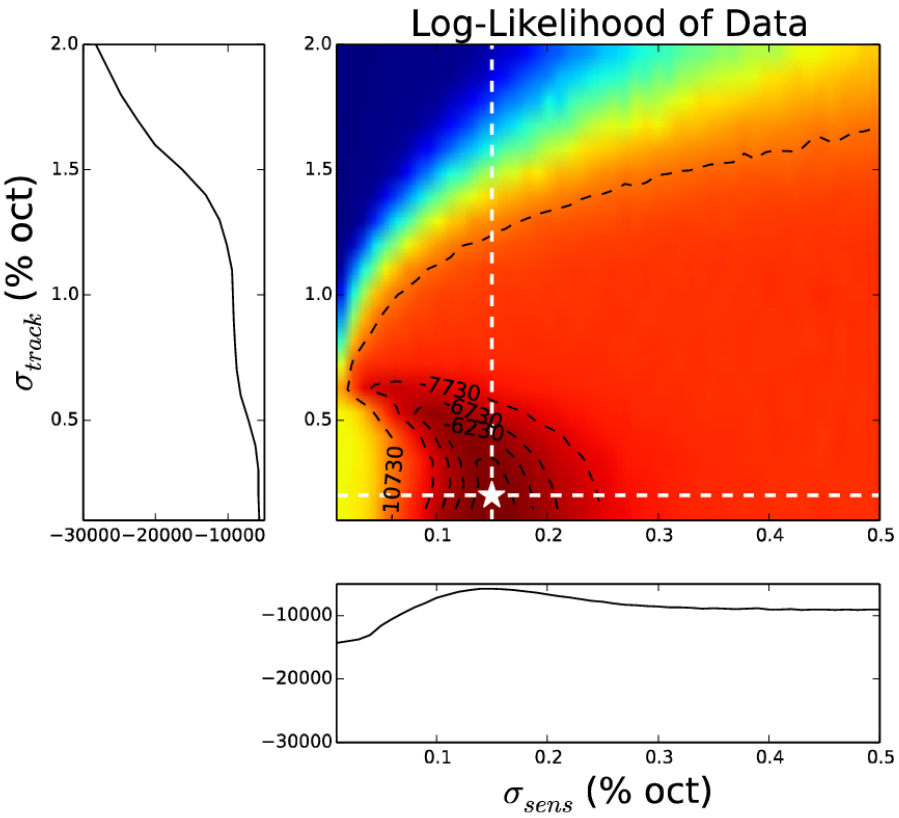


**Figure 4.10:** Log-likelihood of the psychophysical data as a function of model parameters, summed across participants. A broad range of parameters provides a good fit to the data. Dashed lines in black represent iso-value curves of the log-likelihood. The side panels represent the log-likelihood along the two white dashed lines in the main panel. In the left panel, the parameter $\sigma_{sens}$ is fixed to the best-fitting value. For the parameter $\sigma_{track}$, values up to 0.5 provided a similar fit. Therefore, $\sigma_{track}$ was set to an arbitrary value in this range (0.16) in the bottom panel.

one considers the situation faced by auditory perception. Even in the simplest case of a single auditory source, frequency components will be encoded by independent neural populations at the peripheral level. Because sound production is by nature dynamic, there will also be temporal gaps between those components. This introduces an inherent ambiguity: which component should be compared to which to estimate perceptual features? I suggest that the auditory system keeps track of the parallel evolution of frequency components in a way that maximizes continuity. This idea has strong similarities with what has been termed "serial dependency" in vision [32], and is motivated by the same *a priori* continuity constraints on object persistence in the real world [131].

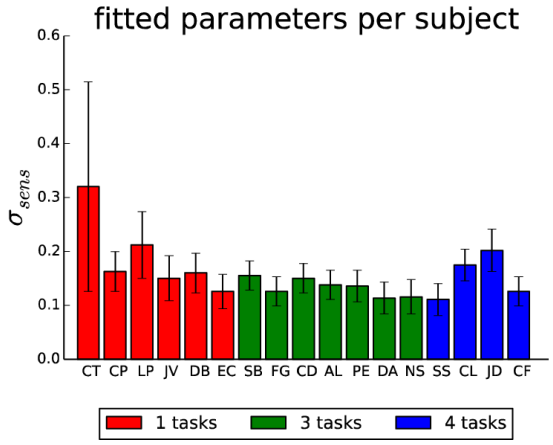


**Figure 4.11:** A bar chart displays best-fitting parameter for individual subjects. The error bars display the narrowest 95% confidence interval around the mode of the normalized likelihood, which provide an estimate of the estimation uncertainty.

# Chapter 5

# Sensory history affects perception through online updating of prior expectations

## Collaboration statement

The work presented in this chapter is the result of a collaboration between, on one hand, Itay Lieder and Merav Ahissar from the Hebrew University (Jerusalem, Israel) and on the other hand Vincent Adam and Maneesh Sahani from the Gatsby Unit, UCL (London, UK). Psychophysical experiments were carried out by the Israeli team. Models and model-based analysis were carried out by the London team.

## 5.1 Introduction

As previously discussed in the general introduction 1.3 and in Chapter 4, perception may be influenced by past stimuli over timescales that range from milliseconds to hours. Here, we studied the impact of recent and long-term stimulus history on the "contraction bias". When asked to report which of two tones separated by brief silence is higher, subjects behave as though they hear the earlier tone "contracted" in frequency towards a combination of recently presented stimulus frequencies, and the mean of the overall distribution of tones used in the experiment [44]. It has been proposed that the long-term contraction bias arises normatively, through the

combination of a noisy memory trace left by the first tone with a prior belief based on the experimental stimulus distribution; a suggestion consistent with increased bias with delay and cognitive load [132, 133]. How detailed is this prior belief, how is it formed, and how does it interact with the effects of recent stimuli?

We measured two-tone frequency discriminations made against different background distributions of stimuli. Using a novel non-linear regression framework, we found that while one component of the bias reflected the overall stimulus distribution in detail, a second component revealed a non-linear influence of recent stimuli that depended on subjects' sensitivity in the task, but not on the sampling distribution used in the task. We reconciled these findings within a single coherent model, wherein subjects' noisy percepts of tone frequency combine with a single prior that they construct online based on their own variable and uncertain experience, eventually leading to an approximation to the experimental distribution. This suggests that both short- and long-term biases arise through a single mechanism, which is tuned to optimise perception in uncertain conditions.

## 5.2 Contraction bias and recency effects

One of the most commonly reported types of bias in psychophysical tasks is the "contraction bias". This bias observed in delayed matching or discrimination tasks describes the tendency of a subject to overestimate (resp. underestimate) a stimulus when it is in the upper end (resp. lower end) of the overall stimulus distribution in effect contracting its value towards the center. It was first reported in a delayed matching task [134], where a first card had to be matched in size, after a delay, to one of many cards of different sizes presented together. In this experiment, small cards were matched to too big cards while large cards were matched to too small ones. This tendency was later found to accurately describe behavior in simple perceptual tasks, such as identification [135, 136, 20], detection [22, 20, 137] and classification [138]. Its effects are also observed in many two alternative delayed discrimination tasks [139, 140] where the same under or over estimation processes qualitatively explains the direction of the observed biases. This form of the bias has been shown

to be tied to the sampling distribution of the stimuli, with narrower (uniform) distributions leading to smaller biases and the 'center of attraction' closely following the mean of the distribution [139].

Evidence for contraction of the first stimulus in perceptual tasks has been found across domains and species. In the visual domain it was demonstrated for color discrimination [141], bar-length matching [133], duration discrimination [142], frequency discrimination [143] and bar-length discrimination [132]. In the auditory domain, it was demonstrated for intensity discrimination [139], frequency discrimination [44, 144, 145], duration discrimination [142, 146]. It has also been observed in tactile discrimination [147], and in non-human subjects such as rats [147] and monkeys [148].

Aside from stimulus distribution alone, several factors have been shown to influence the contraction bias. For fixed durations of stimulus presentation, the bias increases in magnitude with the delay between the first and the second stimuli [141, 139, 146] or when a competitive task is to be solved during the delay [132].

Beyond the effect of overall stimulus statistics, many studies have suggested that the most recent trials affect decision more strongly than more remote ones [44, 32, 136, 149, 42, 142].

### 5.2.1 Models of the contraction bias

In this section, I present two approaches to the quantitative study of the bias in delayed discrimination tasks. I consider the case of delayed 2-tone discrimination tasks consisting in several trials where a first pure tone of log-frequency $f_1$ has to be compared to a second tone of log-frequency $f_2$ presented after a short delay. Subjects participating in such task are asked to report whether or not "$f_1 > f_2$".

Regression studies aim to understand and quantify which aspects of sensory history are predictive of subjects' decisions. Ideal observer (IO) models explicitly represent noise and uncertainty due to assumed sensory and memory processes and their role in the information processing from stimuli to decisions.

#### 5.2.1.1 Regression models

Previous work explained the role of sensory history in perceptual tasks using regression models. This approach builds on the classical models of signal detection theory where a subject's decision probability is modelled as $p(y=1|\delta)=\phi(\alpha\delta)$ where the only covariate $\delta=f_1-f_2$ is a signed measure of difficulty of the trial, and the parameter $\alpha$ reflects a subject's precision, or discrimination ability, and $\phi$ is a monotonous function from $\mathbb{R}$ to $[0,1]$ such as the logistic function and it may include lapse rates [150].

The effect of past sensory history on decisions has been introduced into this framework by incorporating a linear history dependent term into the argument: $p(y=1|\delta,h)=\phi\left(\alpha\delta+\beta^T\psi(h)\right)$. Here $\psi$ denotes a set of hand-crafted features assumed to be relevant (i.e. predictive of subjects decisions)[151, 44], and $\beta$ denoting their associated weight in the regressor.

Such models have been successful in revealing the overall magnitude and timescale of the effect of sensory history. However, when no strong a priori guides the search of the relevant features $\psi(h)$ that best explains observed behavioral data, the space of features to explore is very large and a manual exploration of this space is prohibitively costly.

One approach to alleviate the costly search could be to include a large number of such features into the regression and impose sparsity constraint onto the solution [152]. However, to be efficient in our context where the number of covariates is small, this approach requires relatively strong assumptions about the solution to guide the design of the features: (1) features need to be not too correlated and (2) a sparse combination of these features need to approximate well the 'true' solution. We did not follow this approach.

In Section 5.4, having no a priori intuition about exactly what and how previous sensory history biased perception, we instead take a non-parametric approach to learn history-dependent features directly from subjects' responses.

#### 5.2.1.2 Ideal Observer models

We focus on the normative model of Ashourian and Lowenstein [132] predicting a contraction bias. In addition to a noisy transduction shared with models of discrimination from signal detection theory [150], the authors assume a noisy working memory that corrupts the representation of the first stimulus $f_1$ to memorise. This leads, at the time of decision, to different level of noise corrupting the two elements to be discriminated: $f_1$ and $f_2$. Provided with the additional knowledge of the sampling distribution $p(f)$ and given the noisy representation $\tilde{f}$ of tone $f$, the ideal decision combines information from the form of the noise $p\left(\tilde{f}|f\right)$ with the sampling distribution $p(f)$ used as a prior.

We describe this approach in greater detail in section 5.5. Put simply, the optimal solution involves for both tones, computing the posterior over the true stimulus $f$ given the noisy version $\tilde{f}$ and the prior $p(f)$. In mathematical terms, the posterior over the stimulus value is

$$p\left(f|\tilde{f}\right) \propto \hat{p}\left(\tilde{f}|f\right) p(f)$$

that is, the product of the prior $p(f)$ and the likelihood $\hat{p}\left(\tilde{f}|f\right)$ quantifying the model representation of its uncertainty, reflecting a knowledge of the statistics of its noise (see Fig. 5.10). For the example of a uniform prior, the effect of the prior is to 'crop' the likelihood, hence biasing the posterior towards the center of the prior. Both tones' posteriors are biased in the same direction, but more so in the case of the first one given its higher uncertainty arising from the combination of both memory and sensory noise. This difference leads to the perceptual bias.

Importantly, both the uncertainty $p\left(\tilde{f}|f\right)$ and the sampling distribution $p(f)$ can be manipulated by the experimenter allowing for detailed predictions. Indeed, the authors in [132] found that an increase in the memory load during the interval between the stimuli can be mapped onto an increase in the memory noise. The bias is also directly dependent on the sampling distribution chosen by the experimenter. The same authors ran multiple 2-length discrimination experiments where they varied the position (but not the shape) of a uniform sampling distribution over lengths and found that the pattern of contraction is translated as much as the sampling dis-

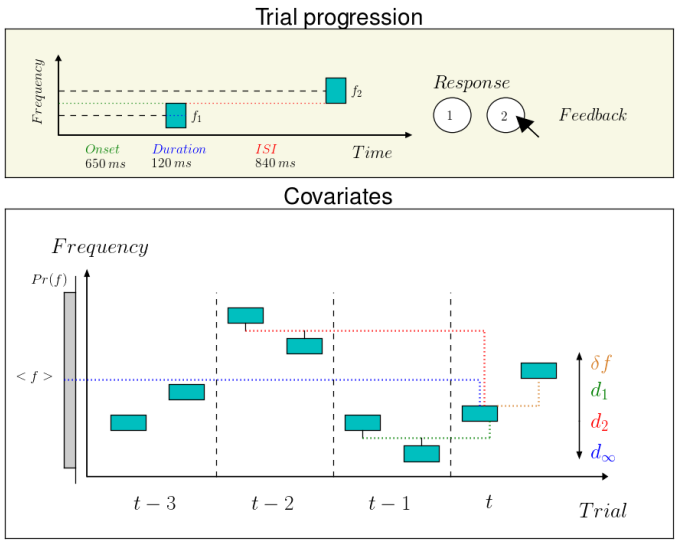


**Figure 5.1:** Task design and covariates for regression analysis. **Top**: detail of a trial. **Bottom**: At trial $t$, covariates capturing recent sensory history $d_\tau$ are defined as the frequency distance between the average of the tone frequencies at time $t-\tau$ and the frequency of the first tone of the current trial. $d_\infty$ is the centered absolute frequency of the first tone in the trial.

tribution.

As the authors of the study noted, "the extent to which the shape of the prior distribution can be learned and utilized in Bayesian reasoning, however, awaits future studies". In section 5.5, we explore the question of how the prior distribution is learned and show how the learning process we propose can explain recency biases revealed in Section 5.4.

## 5.3 Task

### 5.3.1 Stimuli & feedback

Three 2-alternatives forced choice (2AFC) pure tone discrimination experiments were conducted. These experiments were coded mainly in JavaScript and run using a web-browser, with tones loaded and played using HTML 5.

In all experiments, the interstimulus interval (ISI) was set to 850$ms$, tone duration was set to 120$ms$. The delay between a response and the presentation of the first tone of the following trial, the trial onset, was set to 650$ms$. Following their responses, participants were given feedback, presented as a happy (resp. sad) smiley

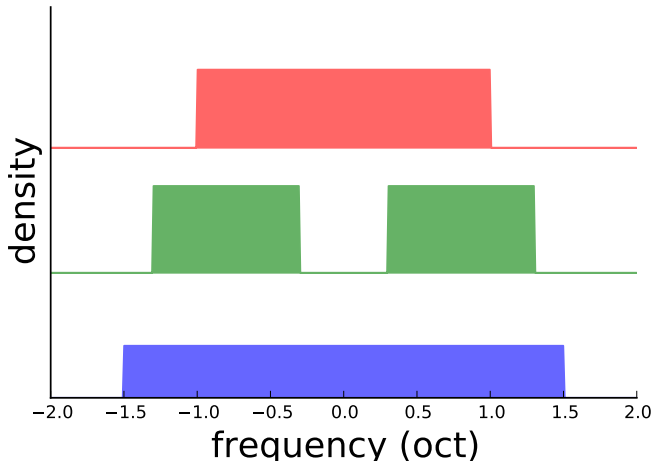


**Figure 5.2:** The mean centered stimulus distributions used in our three experiments: (red) uniform with 2 octaves width, experiment 1, (green) bimodal, experiment 3, (blue) uniform with 3 octaves width, experiment 2.

cartoon after a correct (resp. incorrect) answer. Furthermore, after the completion of each block, they received information about their mean accuracy performance on that block. This shared task design is depicted in Figure 5.1.

The absolute difference between the two tones log-frequencies $|f_1 - f_2|$ was sampled uniformly within the range of $0.5 - 10\%$. This difference was positive or negative with equal probability. For 58 participants, tones were sampled log-uniformly between 500 to 2000 Hz. For the remaining 72 participants, tones were sampled log-uniformly between 400 and 1600 Hz (both ranges are exactly 2 octaves wide). The second experiment was identical except for that tones were log-uniformly sampled from 283-2263 Hz (3 octaves range). In the third experiment, tones were sampled in a bimodal fashion. Tones were sampled from a mixture of two non-overlapping log-uniform distributions: one with a range of 440-800 Hz and the second one with a range of 1228-2295 Hz. On a given trial, one of the modes was selected at random, with equal probability for both modes, and the first tone was then sampled log-uniformly from within that mode. The stimulus distributions used in the three experiments are shown in Figure 5.2.

### 5.3.2 Participants

All participants in this study participated voluntarily via Amazon's "Mechanical Turk" web framework. They participated from the United States of America. Participants with either an approval rate below 95% or a total number of HITs[1] (human intelligence task) of less than 1000 were not allowed. We emphasized that the experiments must be performed: (1) using headphones and in a quiet environment, (2) with either a laptop or a desktop computer and (3) only by people with good hearing and ages ranging from 20 to 50 years old. They were given a payment of $2 for 300 performed trials (corresponding to a duration of approximately 15-20 minutes). Each participant could only participate once. Respectively 130, 156 and 158 subjects participated in the first, second and third experiments.

### 5.3.3 Instructions and training

Subjects first had to give their age, gender and musical experience (in years of practice). They were required to have headphones to participate or to abort the task otherwise. To assert familiarty with the concept of pitch, subjects were given the opportunity to freely engage with four clickable buttons ordered by the height of the pitch they invoke. The pitch of the 4 buttons were respectively 660, 780, 1150 and 1520Hz covering a total of 1.2 octaves. They were asked to set the loudness to a comfortable level during the training session. Then instructions were given about the actual task as follows: "On each trial, two tones will be played consecutively: tone 1 $\rightarrow$ tone 2. Then I ask: which of the two tones had the higher pitch?" Subjects then performed the task for 30 training trials. The training was followed by a small break, before the main task started.

### 5.3.4 Participants' performance-based exclusion

We included subjects in our analyses depending on their performance on the tasks. Mean accuracy was used as an indirect measure of the amout of bias information in subjects' responses. Intuitively, at one extreme of the performance axis, sub-

[1] On Amazon Mechanical Turk, a Human Intelligence Task, or HIT, is a question that needs an answer. Account owners on the platform answer HITs and have a track record specifying the number of HITs they have answered so far.

jects who are too good make no mistake so the biasing effect of sensory history is never revealed. At the other extreme, subjects who perform the worst are very noisy and their response variability hides their biases. We excluded all subjects who had a mean accuracy $< 60\%$ correct and $\geq 90\%$, excluding 62 (48%), 60 (38%), 71 (45%) participants for the 2-octaves uniform, 3-octaves uniform and Bimodal distributions respectively. These exclusion thresholds can be better understood after reading our theoretical understanding of the sensory biases and their dependences on performance (see Section 5.5.3.1). Inclusion statistics are reported in 5.1.

In our analyses, we binned subjects depending on their mean accuracy. Given our stationary task design, this only makes sense if subjects are themselves stationary in their performance. To assess this directly from data, we quantified the variability of performance at slow time scales by computing for each subject $s$ a variability score $v_s$: the standard deviation of mean accuracies in consecutive windows of 30 trials. This score is non zero both because subjects typically get better during the task (mean accuracy difference for the first and last 100 trials is 0.05) and because subjects are noisy. We considered an unusually high score as reflecting a temporary drop of engagement in the task. Hence, we excluded subjects whose score was four standard deviations higher than the population mean (across subjects) resulting in the exclusions of 2 additional subjects for the 3-octaves distributions.

Because we did not record the precise timing of responses of subjects, we are unable to exclude subjects based on single trial performance (for example, it would be desirable to exclude trials following an unintended break).

### 5.3.5 Additional notes

#### Notes on the influence of training

Subject's tone discrimination abilities are known to vary across people with musical training leading to lower discrimination threshold [158]. However, we are not interested in absolute performance. Instead we are interested in the systematic biases and study the dependence of the bias on performance.

### Notes on equipment in online experimentation

For our data acquisition method to be valid, on each system our experiment is ran, two quantities must have a low variability: (1) the latency between executing the code to present a sound and that sound being presented, (2) the difference between the played duration of a sound and its intended duration. Previous research suggests that both these quantities have standard deviations below 1ms in a wide range of common hardware configurations and web browsers [153, 154]. Hardware or software configurations were not use as filters to exclude participants.

Subjects were asked to set the sound loundness to a confortable level throughout the experiment. For subjects with normal hearing and for pures tones within the range considered, confortable level are know to be in a range allowing good performance [155]. There are know pitch-loudness interactions in pitch perception [156, 157]. However, for fixed levels, these interactions are small for the levels and frequency ranges considered.

## 5.4 A descriptive analysis using GAMs

We model the biasing effect of sensory history $h$ on subjects' decisions using a non parametric model:

$$p(\text{“}f_1 > f_2\text{”}|\delta,\alpha,h) \quad = \quad \phi(\alpha\delta + b(h))$$

This model falls in the family of Generalized Additive Models (GAM) [64]. In all our GAM regression analyses we included a constant offset term to the additive predictor. We do not show these offsets in the equations describing our models. In all our analyses, the offset is never inferred as significantly different from 0, and our additive predictors have zero mean.

### 5.4.1 Covariate description

From [44] we knew that both the very recent trials and the absolute position of the first tone were important covariates to consider. We define the main covariate in trial $t$ to be $\delta^t = f_1^t - f_2^t$, the difference between the log-frequencies of the two tones of that trial, accounting for the trial difficulty if there was no bias. To account for the

effect of history, we use as additional covariates the frequency distance (in log Hz) between the contracted tone $f_1^t$ and the arithmetic average of both tones at trial $t-\tau$, that is we define $d_\tau = f_t - \frac{1}{2}\left(f_1^{t-\tau} + f_2^{t-\tau}\right)$. Considering the average effect of the two tones of past trials rather than the individual effect of these tones is justified by the fact that, because the task is set to be hard (small $\delta$) and the sampling ranges wide, distances $f_t - f_1^{t-\tau}$ and $f_t - f_2^{t-\tau}$) are highly correlated. These covariates are illustrated in Figure 5.1.

In addition to defining and using the covariates accounting for the influence of recent trials, we also defined a special covariate to reflect the influence of absolute position of the first tone. We center the covariate to yield $d_\infty = f_t - \langle f \rangle$, where $\langle f \rangle$ is the mean of distribution of the tone log-frequencies. Covariates $d_1, ..., d_\tau, d_\infty$ are illustrated in Figure 5.1.

### 5.4.2 Additivity assumption

Both technical and interpretability considerations led us to further assume an additive structure the bias: $b(h) = \sum_i b_i(d_i)$

On the technical side, inferring a joint function over the covariates $b(d_1, ..., d_\infty)$ requires a very large number of data points (scaling as a power of the number of covariates) which we could not be gathered. The additive assumptions also makes it possible to visualise and interpret the effect of the different covariates. This becomes harder if the functions act on more than two inputs. Interpretability comes at the cost of restricting the expressivity of the model. With the assumed additive structure, any interactions between variables are lost. In our analysis we systematically checked that the additivity assumption was not substantially wrong by inferring pairwise interactions when possible.

### 5.4.3 Results

#### 5.4.3.1 Subjects are biased by sensory history in a non-linear way

The most salient inter-individual difference is in accuracy. In our probit regression setting, this maps into heterogenous values for the fitted parameters $\alpha$. As a first approximation, we set our model to have a different value of $\alpha$ for each subject

(indexed $s$) and assumed the bias terms to be equal for all subjects.

$$p(y=1|\alpha^s,\delta,\mathbf{b},\mathbf{d}) \quad = \quad \phi\left(\alpha^s\delta + b_1(d_1) + b_\infty(d_\infty)\right)$$

We fit this regression model to the subject responses of Experiment 2 (3-octave sampling distribution), including subjects whose overall accuracy fell within $[60\%, 90\%[$. Inclusion statistics are given in table 5.1.

We focused on this experiment both because it is the task in which covariates $d_1$ and $d_\infty$ are the least correlated, and because we have the largest number of participants in this case. Results shown in Figure 5.3 reveal that the biasing functions are non-linear and antisymmetric.

To assess the prediction gain achieved by using the different covariates and by using non-linear functions rather than linear ones, we compute the relative 10-fold cross-validated AUC (Area Under the ROC curve [150]) for 5 different biases: no bias, bias described by linear functions ($w_1 d_1 + w_\infty d_\infty$), 3 biases described by non-linear functions ($b_1(d_1)$, $b_\infty(d_\infty)$ and $b_1(d_1) + b_\infty(d_\infty)$ ).

Because the magnitude of biases are small and have no effect on most trials, the predictive gain between a biased and unbiased model is small if predictive performance is assessed on all trials. This is because both models explain most trials equally well.

To further highlight the predictive gain of the different biases, we now focus on the trials for which we expect a predictive gain. We select trials where the magnitude of the history bias is larger than the magnitude of the predictor with no bias $|b(h)| > |\alpha\delta|$, and where both favor opposite decisions $b(h)\alpha\delta < 0$.

Absolute and relative cross-validation results on both all and selected trials are shown in Figure 5.4.

We found that predictive power was increased when using non-linear bias functions and when both covariates were uses. In that case there was no loss in predictive power when an additive structure was imposed. The predictive gains were also larger by almost one order of magnitude when the AUC was calculated on subset of trials where the gain was expected to be the most salient.

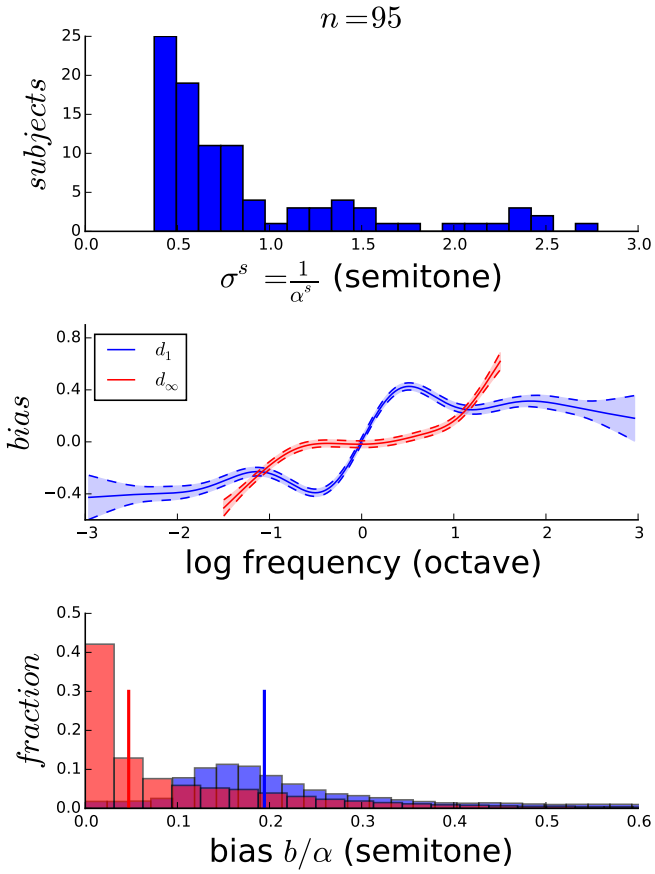


**Figure 5.3:** Effective Precision and shared bias for Experiment 2 (3-octave sampling range). **Top**: histogram of precisions $\sigma_s = 1/\alpha^s$. **Middle**: inferred biases and standard error. $b_1\,(d_1)$ (blue), $b_\infty\,(d_\infty)$ (red). **Bottom**: Histogram of the absolute magnitude of the induced biases in semitones, and median values.

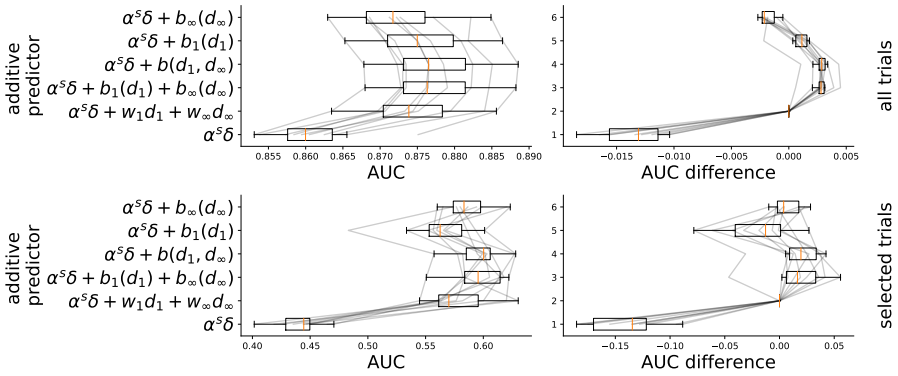


**Figure 5.4:** Absolute (left) and relative (right) cross validated AUC of GAM regression models with different sensory history biases evaluated on either all (top) or a subselection (bottom) of trials where the sensory biases are expected to drive the decision most (Experiment 2).

#### 5.4.3.2 Sensory bias is stronger than response bias

Apart from stimulus history, response and feedback history might affect the discriminations [159, 138, 160]. Unpublished observations suggested that a negative feedback could promote a switching response, while a positive response might encourage perseveration with the last response. Irrespective of the feedback, subjects might tend to choose the same response as in the previous trial, a phenomenon often referred to as response inertia. We sought to quantify the relative contribution of past responses and feedback by explicitly adding terms to our additive predictor capturing the intuitions described above.

Let $y^{t-1}$ be the previous response and $e^{t-1}$ be the previous feedback. Both variables take values in $\{-1,1\}$. We add the following 2 covariates to our predictor: $x_r = y^{t-1}$ for the response bias, $x_f = y^{t-1}e^{t-1}$ for the feedback bias. Both covariates were centered.

Following the same methodology as in the previous section, we compare different models with additive predictors including either or both response history covariates $(x_f, x_r)$ or sensory history covariates ($d_1$, $d_\infty$). For each model, a separate fit is done for each subject (all trials included). Having one model per subject means there are fewer data points per model. To reduce the complexity of the models, sensory biases are set to be linear ($b_x(d_x) = w_x d_x$ for $x \in \{1,\infty\}$). It was shown in the previous section that such linear biases have higher predictive accuracy that unbiased model for the dataset of Experiment 2)

All models compared that have a linear predictor are nested, with the most complex model being:

$$p(y^t = 1|\alpha, \delta^t, \mathbf{w}, \mathbf{b}, \mathbf{d}) = \phi\left(\alpha\delta^t + w_1 d_1 + w_\infty d_\infty + w_f x_f + w_r x_r\right)$$

Still focusing on Experiment 2, results shown in Figure 5.5 reveals that, sensory history biases have more predictive power than response or feedback history biases (with the latter having little to no contribution to subject responses). The key force driving the bias is therefore more the sensory history than the response and feedback histories.

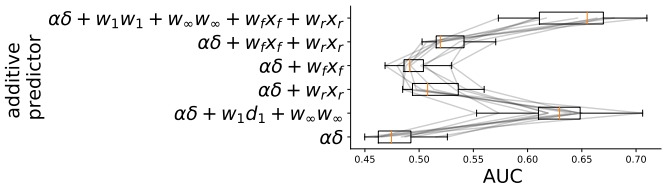


**Figure 5.5:** Cross-validated likelihood, comparing linear models with sensory, response and feedback history dependent biases on data from Experiment 2. All trials are included and one model is fitted per subject.

#### 5.4.3.3 Most recent sensory experience has the strongest influence

We have so far restricted our covariates to $d_1$ and $d_\infty$. Previous results have reported that preceding trials beyond the most recent might also influence responses [44]. Thus we considered terms for additional covariates related to sensory history up to a lag of 5, leading to the following GAM model:

$$p\left(y^t = 1|\mathbf{h}\right) \quad = \quad \phi\left(\alpha^s \delta^t + b_1\left(d_1^t\right) + b_2\left(d_2^t\right) + b_3\left(d_3^t\right) + ... + b_\infty\left(d_\infty^t\right)\right)$$

Figure 5.6(right) shows the inferred biasing function of each covariate. Histogram of the function values taken on all trials are reported in Figure 5.6(middle). The normalized variance over trials $v_d = \frac{var[b_d]}{\sum_{d'} var[b_{d'}]}$ for each bias term in the additive predictor is reported in 5.6(left). This measure was used in [151] and corresponds to the normalized variances of the bias distributions shown in Figure 5.6(middle).

We found that the proportion of the variance explained by the recent trials decays as a function of the lag, with only the 3 most recent trials having a significant contribution to the bias. The bias term for $d_\infty$ is preserved in its shape and magnitude when the additional recent trial covariates are added (see Fig. 5.6(right) and Fig. 5.3).

An important observation is the persistence of the $b_\infty$ bias term when covariates corresponding to trials further in the past are added to the regression. At this point, we can only conclude that the biasing effect of past trials cannot be explained by a simple additive combination (otherwise, $b_\infty$ term would vanish as more lags are included in the regression).

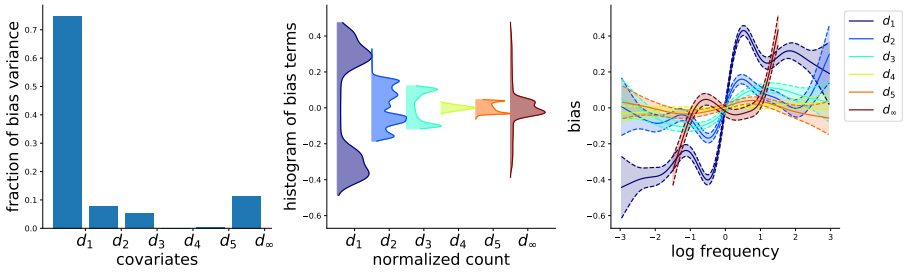


**Figure 5.6: Left**: Normalized fraction of bias terms in additive predictor. **Middle**: Histogram of bias values. **Right**: Inferred biasing functions and standard error.

| Experiment | Total | included | excluded | $\leq 60\%$ | $< 90\%$ |
|---|---|---|---|---|---|
| 1 (unimodal 2 octaves) | 130 | 68 | 62 | 24 | 38 |
| 2 (unimodal 3 octaves ) | 156 | 96 | 60 | 19 | 41 |
| 3 (bimodal) | 158 | 87 | 71 | 35 | 36 |

**Table 5.1:** Inclusion table per experiment and per accuracy bounds

#### 5.4.3.4 Bias for different distributions

We now applied the same analysis to 3 different datasets corresponding to 3 versions of the same task differing only in regard to the sampling distribution of the stimuli (unimodal 3oct, unimodal 2oct, bimodal), selecting subjects whose overall accuracy fell between 60% and 90%. Inclusion statistics are given in table 5.1.

First we show that for all 3 distributions, the non-linear additive predictor $b_1(d_1)+b_\infty(d_\infty)$ has the best predictive power. For each distribution, relative cross-validated AUCs for different bias models are reported in Figure 5.9.

Inferred biases are shown in Figure 5.7. The bias $b_\infty$ varies the most across distributions, both quantitatively and qualitatively. Within the $[-1oct, 1oct]$ range, $b_1$ functions were essentially equal (note that beyond this range, posterior uncertainty is larger). Thus our observations reveal that the recency bias $b_1$ is not sensitive to changes in the sampling distribution while the long term bias $b_\infty$ is.

Confirming observations of Ashourian et al [132], subjects' long term bias appears to be attractive and 'follows' the distribution for the 2 uniform sampling distributions of frequencies. The case of the bimodal sampling distribution reveals local contraction toward the mean of each component of the mixture, superimposed on a global contraction towards the overall mean of the distribution.

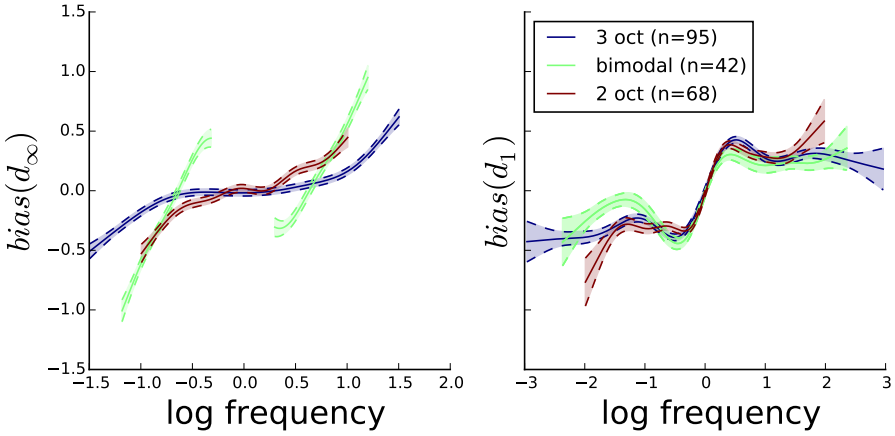


**Figure 5.7:** Inferred sensory history bias functions for the 3 different experiments.

| Experiment | [60-70%[ | [70-80%[ | [80-90%[ | $\leq 60\%$ | $> 90\%$ |
|---|---|---|---|---|---|
| 1 (unimodal 2 octaves) | 21 | 21 | 26 | 24 | 38 |
| 2 (unimodal 3 octaves ) | 20 | 25 | 51 | 19 | 41 |
| 3 (bimodal) | 23 | 23 | 41 | 35 | 36 |

**Table 5.2:** Inclusion table per experiment for regression analysis for accuracy bounds [60-70%[,[70-80%[,[80-90%[

Finally, we report the performance-dependence of the bias. For each of the 3 experiments we grouped the subjects in 3 accuracy groups ([60-70%[,[70-80%[,[80-90%[) and fit a GAM with a joint bias over $d_1$ and $d_\infty$, and individual $\alpha$ per subject. Inclusion statistics are reported in Table 5.2. Biases are shown in Figure 5.8. They demonstrate the validity of the additive assumption (a separate bias function for $d_1$ and $d_\infty$) that is inferred even when this assumption is not made a priori.

It is interesting to note at this point that the range of reported biases does not vary much across accuracy groups or across distributions. The bias $b$ of a predictor $\alpha\delta + b$ is dimensionless. This predictor can be rewritten as $\alpha\left(\delta + \frac{b}{\alpha}\right)$ which reveals that the 'effective' bias in log Hertz is $\frac{b}{\alpha}$. In our analysis, the fact that $b$ is invariant across accuracy groups does not mean the bias has the same magnitude across groups. Instead, the better the subject (large $\alpha$) the smaller the 'effective' bias $\frac{b}{\alpha}$ (the smaller the contraction in log Hertz).

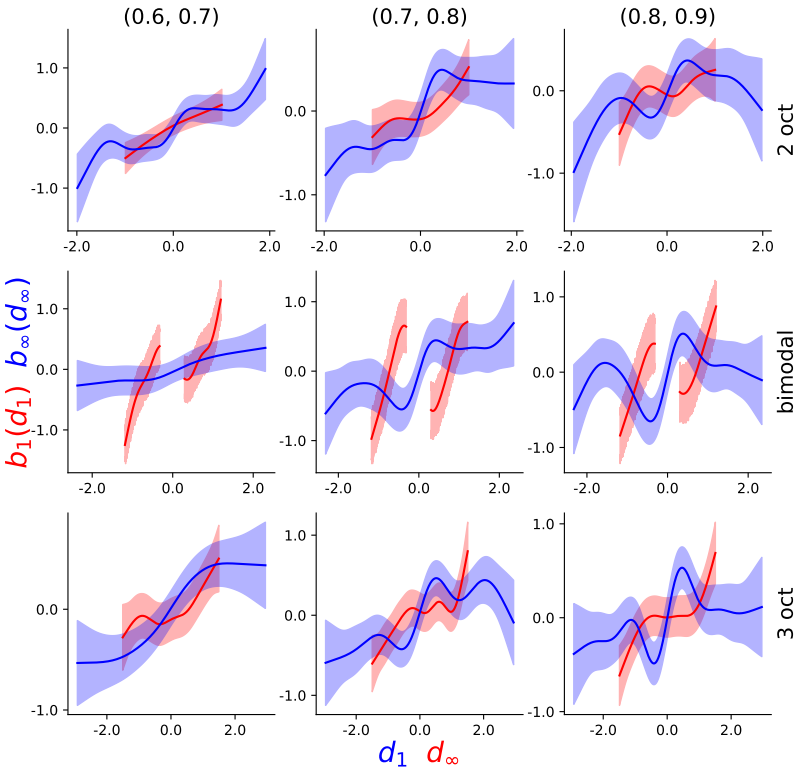


**Figure 5.8:** Bias $b_1(d_1)+b_\infty(d_\infty)$ for the 3 stimulus distributions and for 3 accuracy groups.

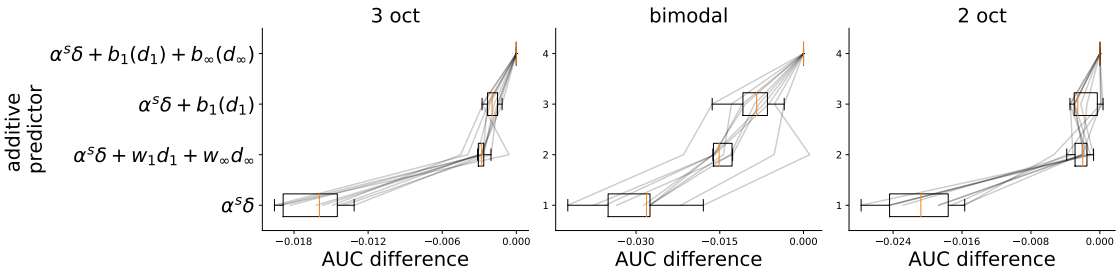


**Figure 5.9:** Relative Cross Validated AUC for 3 distributions and different sensory biases

## 5.5 Ideal observer models

### 5.5.1 Aims and goals

Our GAM analysis of subjects' responses across sampling distributions revealed a long term bias induced by sensory history that depended on the stimulus distribution. This distribution dependence of the long term bias is a characteristic of Ideal observer models [132] which encouraged us to study the predictions of similar IO models for the broader variety of distributions we used.

Our analysis also revealed a biasing effect of the most recent sensory history, which is not predicted by the IO approach in [132].

Our aim here is twofold: (1) assess if this IO approach can predict the observed long term biases revealed by our regression analysis, (2) attempt to explain both long term and short term effects within a single framework.

### 5.5.2 Theory

Following Ashourian et al [132], we derive an IO model for the delayed discrimination task. For each tone $f$, subjects are assumed to observe a noisy version $\tilde{f}$ of $f$ with, at decision time, more noise on the first tone due to a noisy memory retention mechanism. We denote by $p(f)$ the true stimulus distribution, $p\left(\tilde{f}_i|f_i\right)$ the noise model, and $\hat{p}(f)$ and $\hat{p}\left(\tilde{f}_i|f_i\right)$ subjects' internal representation of these distributions. These internal representations of the environmental statistics might differ from the true ones, especially for the stimulus distribution $\hat{p}(f)$ which is only indirectly experienced through a limited number of samples during the task.

We set the noise model to be Gaussian centered on the true log-frequencies, that is $p\left(\tilde{f}_i|f_i\right) = \mathcal{N}\left(\tilde{f}_i; f_i, \sigma_i^2\right)$. Noise is exclusively sensory for the second tone $\sigma_2^2 = \sigma_s^2$ . The first tone has an additional memory noise $\sigma_1^2 = \sigma_s^2 + \sigma_m^2$. We will assume that subjects have an accurate knowledge of their own noise, that is their internal uncertainty matches the actual level of noise: $p\left(\tilde{f}_i|f_i\right) = \hat{p}\left(\tilde{f}_i|f_i\right)$. This is motivated a priori by both the aims to deviate as little as possible from optimality (the starting point of Bayesian rational analyses like the one we conduct here) given our IO model and to minimize the complexity of the model.

#### 5.5.2.1 Optimal decision and likelihood

Given two noisy tones $\tilde{f}_1, \tilde{f}_2$, the optimal decision on whether "$f_1 > f_2$" is $1\left[\hat{P}\left(f_1 - f_2 > 0 | \tilde{f}_1, \tilde{f}_2\right) > \frac{1}{2}\right]$, where $1[x] = \begin{cases} 0, & \text{for } x \leq 0 \\ x, & \text{for } x > 0 \end{cases}$.
Averaging over noise realisations leads to the optimal decision probability:

$$P(\text{“}f_1 > f_2\text{”} | f_1, f_2) \quad = \quad E_{\tilde{f}_1, \tilde{f}_2 \sim p\left(\tilde{f}_1, \tilde{f}_2 | f_1, f_2\right)} 1\left[\hat{P}\left(f_1 - f_2 > 0 | \tilde{f}_1, \tilde{f}_2\right) > \frac{1}{2}\right]$$

At this point, we are left to specify the subjects' internal representation of the stimulus distribution $\hat{p}(f)$.

In the first section, I will consider the case where subjects' prior matches the true distribution as in Ashourian et al [132], further exploring the predictions of this approach. In the second section, I introduce a prior learning rule that enables one to capture both recent and long term biases in the same unifying framework.

#### 5.5.2.2 Approximation

In most cases, the decision probability has no closed form and needs to be approximated. The decision probability can be rewritten as

$$P(\text{“}f_1 > f_2\text{”} | f_1, f_2) \quad = \quad E_{\tilde{f}_1, \tilde{f}_2 \sim p\left(\tilde{f}_1, \tilde{f}_2 | f_1, f_2\right)} 1\left[\hat{m}\left(\tilde{f}_1, \tilde{f}_2\right) > 0\right]$$

with $\hat{m}\left(\tilde{f}_1, \tilde{f}_2\right) = \text{median}\ \left[\hat{P}\left(f_1 - f_2 | \tilde{f}_1, \tilde{f}_2\right)\right]$.

First, we approximate the median of the posterior over the difference $f_1 - f_2$ by the difference of the medians of the individual posteriors:

$$\text{median}\ \left[\hat{P}\left(f_1 - f_2 | \tilde{f}_1, \tilde{f}_2\right)\right] \quad \approx \quad \underbrace{\text{median}\ \left[\hat{P}\left(f_1 | \tilde{f}_1\right)\right]}_{\hat{m}_1\left(\tilde{f}_1\right)} - \underbrace{\text{median}\ \left[\hat{P}\left(f_2 | \tilde{f}_2\right)\right]}_{\hat{m}_2\left(\tilde{f}_2\right)}$$

Unreported simulations demonstrate the accuracy of this approximation for the distributions we consider.

Second, we approximate the expectation via Monte Carlo samples

$$E_{\tilde{f}_1, \tilde{f}_2 \sim p\left(\tilde{f}_1, \tilde{f}_2 | f_1, f_2\right)} 1\left[\hat{m}\left(\tilde{f}_1, \tilde{f}_2\right) > 0\right] \quad \approx \quad \sum_i \sum_j 1\left[\hat{m}\left(\tilde{f}_1^{(i)}, \tilde{f}_2^{(j)}\right) > 0\right]$$

Where possible, the median and the expectations are computed analytically.

Another option to approximate the expectation over noise is to approximate each conditional noise distribution by a discrete distribution. For example $p(\tilde{f}_j | f_j)$

is approximated by discrete outcomes and weights $\{\tilde{w}_j^s, \tilde{f}_j^s\}$ (for example, $\tilde{f}_j^s$ uniformly around the mean $f_j$ and unnormalized weights proportional to the density function $\tilde{w}_j^s \propto p(\tilde{f}_j^s|f_j)$ ) . The posterior median distribution $p(\hat{m}_j|f_j)$ is approximated by the discrete distribution $\{\tilde{w}_j^s, \hat{m}_j(\tilde{f}_j^s)\}$. The difference of the posterior median $p(\hat{m}_1 - \hat{m}_2|f_1, f_2)$ is also a discrete distribution $\{w^s, dm^s\}$. Hence the decision probability correspond to its cumulative density function evaluated at 0 is easily computable as

$$E_{\tilde{f}_1,\tilde{f}_2\sim p\left(\tilde{f}_1,\tilde{f}_2|f_1,f_2\right)} 1\left[\hat{m}(\tilde{f}_1,\tilde{f}_2) > 0\right] \approx \sum_s 1[dm^s > 0]w^s$$

An advantage of this approximation is that it leads to deterministic decision probabilities. An additional relaxation of the hard threshold $x \to 1[x > 0]$ to a soft threshold $x \to \phi(\beta x)$, with $\phi$ the sigmoid function and $\beta$ controling the relaxation, leads to a differentiable decision probability, allowing for gradient based deterministic optimisation of the likelihood of subject responses with respect to the parameters of our IO models.

### 5.5.3 The Fixed Prior case

In this section I focus on the setting where $\hat{p}(f)$ is fixed to $p(f)$ throughout the experiment. The model is fully specified by the memory and noise variances $\sigma_s^2, \sigma_m^2$ and $p(f)$.

#### 5.5.3.1 Theoretical biases

Given noisy tones $\tilde{f}_1, \tilde{f}_2$ the optimal decision on whether "$f_1 > f_2$" can be rewritten as $1\left[\hat{m}\left(\tilde{f}_1, \tilde{f}_2\right) > 0\right]$ with $\hat{m}\left(\tilde{f}_1, \tilde{f}_2\right) = \text{median}\ \left[\hat{p}\left(f_1 - f_2|\tilde{f}_1, \tilde{f}_2\right)\right]$. In other words, the ideal decision is to threshold the median of the posterior on the tones difference at 0. For a given pair of tones $f_1, f_2$ and considering all possible conditional noise realisations, we have a random threshold $\hat{m}|f_1, f_2$ and hence a decision probability $P(\text{“}f_1 > f_2\text{”}|f_1, f_2) = E_{\hat{m}|f_1,f_2} 1[\hat{m} > 0|f_1, f_2]$.

Writing $\hat{\mu}(f_1, f_2) = E[\hat{m}|f_1, f_2]$ and $\hat{\sigma}^2(f_1, f_2) = Var[\hat{m}|f_1, f_2]$ and approximating the median distribution by the moment-matched Normal distribution

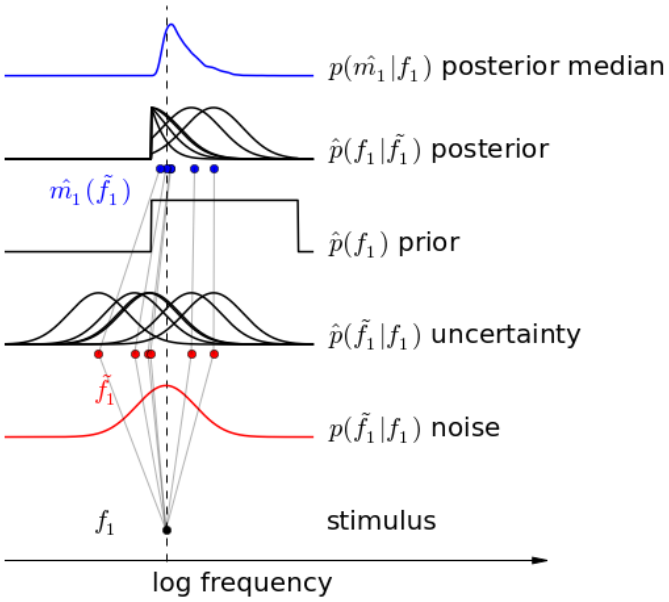


**Figure 5.10:** Illustration of the implementation of the IO model. Here the prior is set to be a uniform distribution. Given a true stimulus $f_1$, noisy encoding leads to representation $\tilde{f}_1$. For a fixed representation $\tilde{f}_1$, combining sensory uncertainty $p\left(\tilde{f}_1|f_1\right)$ with prior expectations $p\left(f_1\right)$ on frequency leads to the posterior $\hat{p}\left(f_1|\tilde{f}_1\right)$ and its median $\hat{m}_1\left(\tilde{f}_1\right)$ used as a threshold to form a decision. Considering all possible noisy representations leads to a distribution over the posterior median $p\left(\hat{m}_1|f_1\right)$. Relative noise standard deviation and prior width are here arbitrary and chosen to best illustrate the implementation.

$\hat{m}|f_1,f_2 \sim \mathcal{N}\left(\hat{\mu}\left(f_1,f_2\right),\hat{\sigma}^2\left(f_1,f_2\right)\right)$ leads to the decision probability

$$\begin{aligned} P\left(\text{“}f_1 > f_2\text{”}|f_1,f_2\right) &\approx \phi\left(\frac{\hat{\mu}\left(f_1,f_2\right)}{\hat{\sigma}\left(f_1,f_2\right)}\right) \\ &\approx \phi\left(\frac{f_1-f_2}{\hat{\sigma}\left(f_1,f_2\right)}+\hat{b}\left(f_1,f_2\right)\right) \end{aligned}$$

Given this formulation, a subject's performance is fully summarized by a bias function $\hat{b}\left(f_1,f_2\right)$ and the variability $\hat{\sigma}\left(f_1,f_2\right)$. An intuitive way to report the bias is to compute the decision probability of the model for a pair of similar tones: $P\left(\text{“}f_1 > f_2\text{”}|f_1=f_2\right) = \phi\left(\hat{b}\left(f_1,f_2=f_1\right)\right)$

An illustration of the IO model in this fixed prior case is given is Figure 5.10.

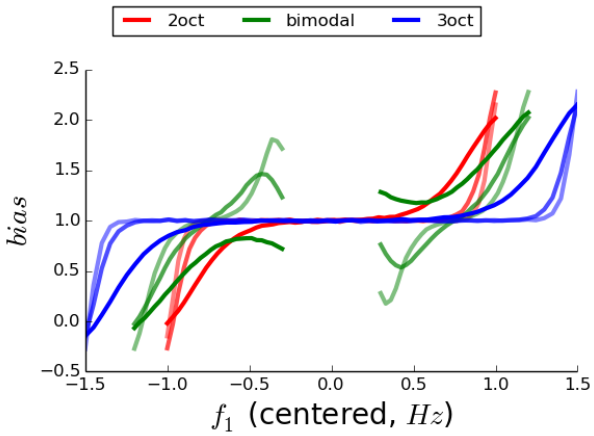


**Figure 5.11:** Theoretical biases for the normative model across distributions and accuracy range. Colors correspond to different distributions, Line contrast corresponds to different accuracy ranges (darker corresponds to lower accuracy)

### 5.5.3.2 Predicting biases across distributions and subjects

For each of the three stimulus distributions described in section 5.3, I grouped subjects into three groups depending on their overall accuracy ([60-70%],[70-80%],[80-90%]) leading to a total of 9 conditions as in section 5.4.3.4. Likelihood maps for parameters $\rho = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}$ and $\sigma_{tot}^2 = \sigma_1^2 + \sigma_2^2$ are reported for each condition in Figure 5.12 revealing clear optima. For each condition, I repeatedly randomly split the trials into training and test sets (10% test, 10 repeats) and fit the IO model, maximizing the likelihood. Fitted parameters $\rho$ and $\sigma_{tot}$ are reported in Figure 5.13 with errorbars depicting standard deviations across the random training sets. For all 3 sampling distributions, $\sigma_{tot}$ decreases with the accuracy range, while $\rho$ remains mainly unchanged across accuracy ranges and distribution. Theoretical biases reported in Figure 5.11 demonstrate the differential effects of the sampling distribution and the accuracy of subjects. The sampling distribution sets the overall bias range and shape. The effect of the accuracy range is more subtle. For uniform sampling distributions, the bias is observed further away from the distribution edges as accuracy decreases.

As a further validation of the model, the predicted accuracies of the fitted models on task are reported in Figure 5.14 and show that for all sampling distributions and groups, predicted accuracy falls into the corresponding accuracy range.

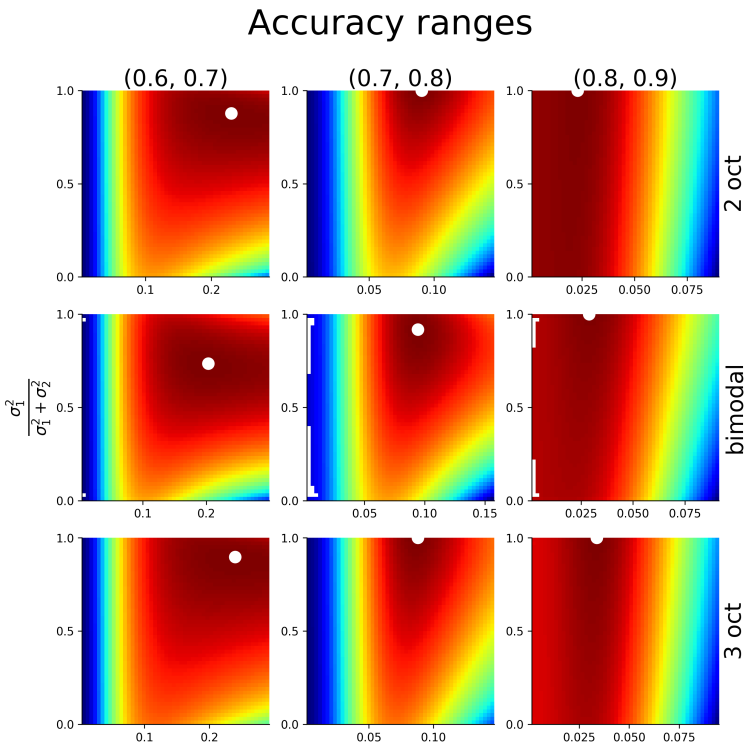


**Figure 5.12:** Likelihood map for the IO model with true prior for 3 accuracy ranges and 3 experiments. White dots mark the parameter set maximizing the likelihood.

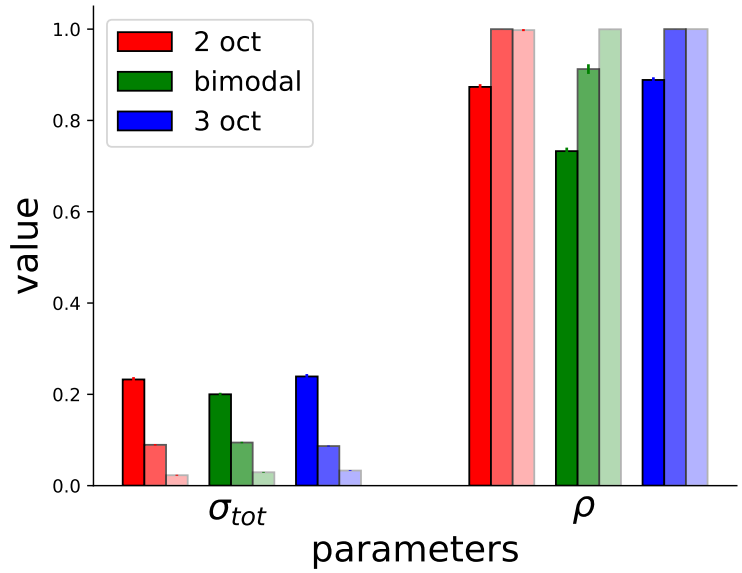


**Figure 5.13:** Fitted parameters for the IO model with true prior, for 3 accuracy ranges and 3 sampling distributions. Color contrast correspond to different accuracy ranges (darker corresponds to lower accuracy).

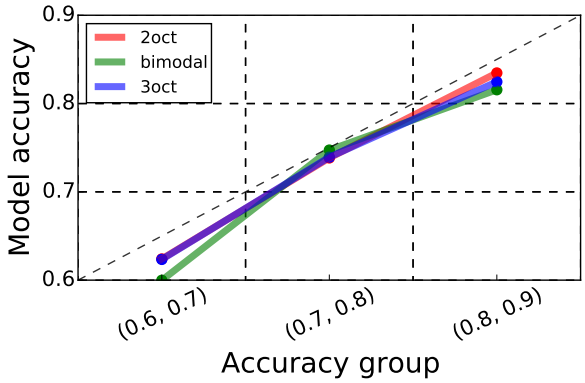


**Figure 5.14:** Mean accuracy of fitted model for each group and experiment.

This model predicts no effect of recent sensory history on decisions. However, the theoretical long term biases look like the ones inferred in our GAM analysis. This qualitative yet precise resemblance across sampling distributions and subject accuracy groups suggests that the long term biases may be the mark of an uncertainty dependent combination of learned prior expectations and sensory information.

## 5.5.4 Learning the prior

### 5.5.4.1 Model specification

We now formalize the assumption that subjects learn the stimulus distribution based on their own variable and uncertain experience. In the previous section we used the stimulus distribution as a prior and revealed that a range of complex distributions may be learned. Our GAM analysis further revealed that the tones of recent trials have a separate effect from that of the more distant past sensory history. We choose to model the learned prior at time $t$ as a mixture of Gaussians $\hat{p}_t(f) = \sum_\tau w_\tau \mathcal{N}\left(f; \tilde{f}_{t-\tau}, \sigma_\tau^2\right)$. This form allows one to learn and approximate various stimulus distributions and to account for relatively different contributions of past trials in the prior, depending on the lag $\tau$.

The choice of the lag-dependence on both weights $w_\tau$ and variances $\sigma_\tau^2$ is constrained by empirical observations: First, both recent and long-term sensory history affect subjects' perceptual decisions. The recency effect suggests that recent trials should be dominant while the long term effect suggests that trials further back

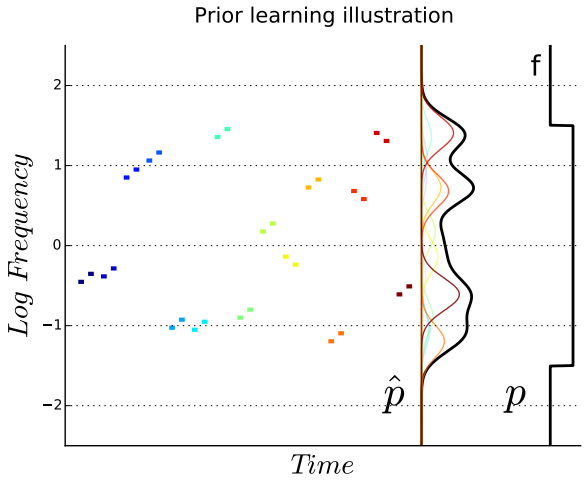


**Figure 5.15:** Illustration of the prior learning model. Sampling distribution $p$ is approximated as a mixture of Gaussians $\hat{p}$, with components centered on past stimuli and stronger weights for recent trials.

also contribute to the mixture. Second, the model should be able to capture inter-individual differences in subjects' contextual bias.

Beyond these desiderata, the actual parametric choice for the learning is somewhat arbitrary.

We choose the following parameterization:

- $\sigma_\tau^2 = \sigma_p^2$ (independent of $\tau$)
- $w_\tau \propto o + (1-o).e^{-\tau/\eta}$

The variance of mixture components is independent of the lag and left as a free parameter. Weights decay exponentially with rate $\eta$ and with an offset $o$. The full model has a total of 5 parameters: the noise parameters $\sigma_s, \sigma_m$ and the prior learning parameters $\tau, \sigma_p, o$. A cartoon description of the prior learning mechanism is shown in Figure 5.15.

### 5.5.4.2 Results

For the same accuracy groups and for the 3 sampling distributions as in section 5.4.3.4, I repeatedly randomly split the trials into training and test sets (10% test, 10 repeats). I fit the prior learning model to the training set by maximizing the likelihood and report the likelihood of the fitted model on the test set.

Results are reported in Figure 5.17. For all distributions and accuracy ranges,

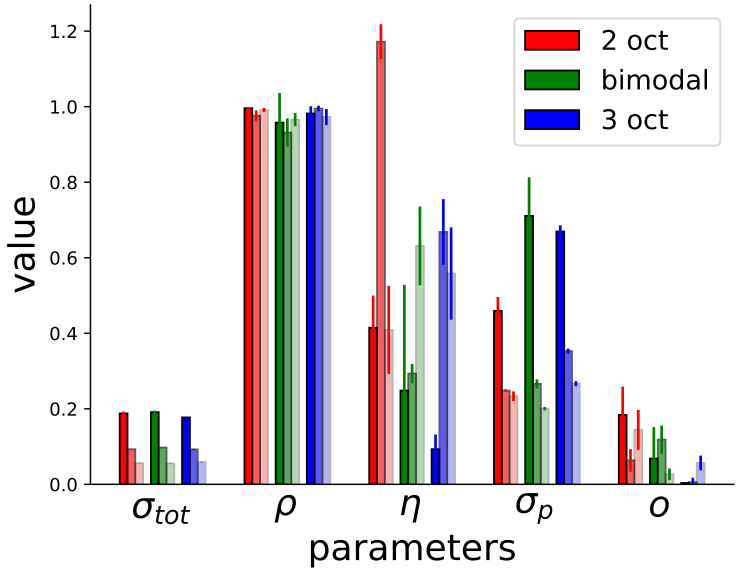


**Figure 5.16:** Fitted parameters for the prior learning IO model, for 3 accuracy ranges and 3 sampling distributions. Color contrast correspond to different accuracy ranges (darker corresponds to lower accuracy)

the prior learning model achieves a better predictive likelihood than the model using the true sampling distribution as a prior. However, the fitted models consistently predict lower accuracies on the task (Figure. 5.18). Overall, these results demonstrate the better predictive abilities of our model including a parametric distribution learning mechanism. Our method however, does not match the prediction abilities of the GAM analysis. This quantitative failure shown in Figure 5.17 could be explained by the strong parametric assumptions of the pre-existing IO (log encoding, additive sensory and memory noise) and the somewhat arbitrary form of our distribution learning mechanism. Fitted parameters are reported in Figure 5.16. Fitted parameters $\sigma_{tot}$ and $\rho$ have the same qualitative and quantitative variation with sampling distribution and accuracy range as in the initial IO model with a fixed prior. In all cases the decay rate $\eta$ is below $\approx 1$, meaning the fitted decay is fast. The parameter $\sigma_p$ covaries with $\sigma_{tot}$, matching the intuition that the noisier the perception the less accurate is the learned prior.

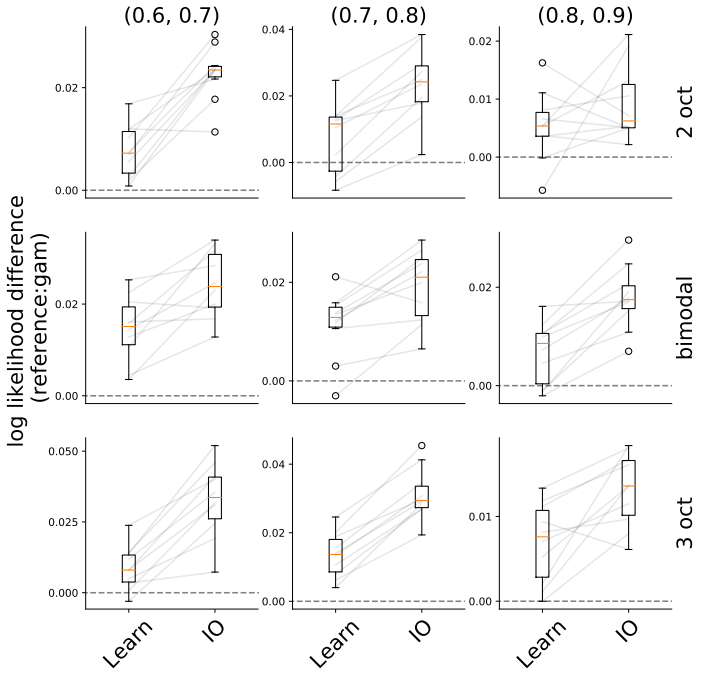


**Figure 5.17:** Cross-validated likelihoods difference for the IO model with fixed and learned prior. These are substracted to the likelihood of a GAM model with bias terms $b_1(d_1) + b_\infty(d_\infty)$. Results are shown for 3 accuracy ranges and for the 3 sampling distributions.

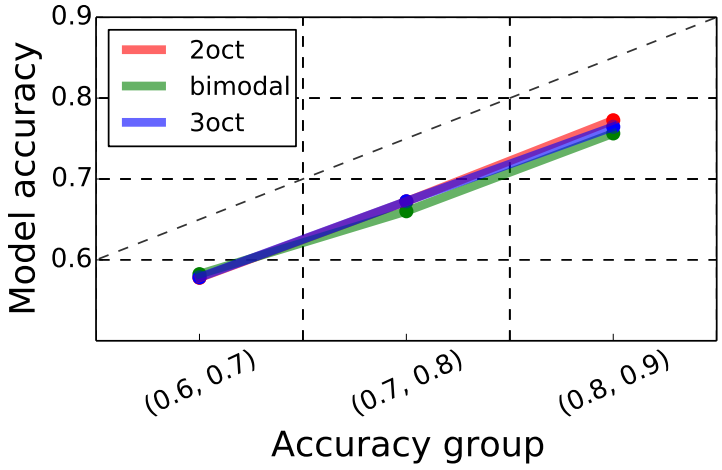


**Figure 5.18:** Fitted model predicted accuracy for each sampling distribution and accuracy range.

## 5.6 Discussion

In this chapter, our GAM analysis revealed how both recent and long-term sensory history affect decisions in a delayed discrimination task. We found that subjects' decisions are biased by long term sensory history in a way that depends on the sampling distribution used in the task, while the effect of the more recent sensory history is independent of the sampling distribution. This analysis revealed that the bias cannot be explained as a linear contraction towards a weighted sum of past tone frequencies. Instead these past frequencies bias decisions in a non-linear manner.

A pre-existing IO model could explain qualitative properties of the long-term bias. This bias reflected a detailed statistical knowledge of the sampling distribution for 3 distributions. The predictions also extended to the case of a wide Gaussian distribution of standard deviation equal to 0.4 octaves for which the predicted bias is linear (not shown). However, this IO model predicted no recency effect. Introducing a distribution-learning mechanism into the IO model made it possible to qualitatively reproduce bias at both time scales and quantitatively better predict subject responses in all all conditions.

Although we achieved a good qualitative fit, our model did not match the predictive abilities of the descriptive GAM analysis we carried out. This is probably a consequence of the parametric form of the model, that is too simplistic or erroneous in its description of the memory, encoding, statistical learning or decision processes. Further work could help refine the model and provide a better quantitative match. Additional experimental manipulations could also help constrain the model. Manipulating the delay between the discriminated tones could help constrain our model of working memory. Manipulating the tone duration could help constrain our model of noise and uncertainty.

An important choice we made in this study is to assume that subjects uncertainty matched their memory and sensory noise. This is a strong assumption and subjects might instead over or underestimate the noise level in their sensory and memory processes. Since uncertainty and learned prior shape the magnitude and shape of the bias, allowing uncertainty to differ from noise could lead to different

conclusions about the statistical learning processes underlying the recency effect and could provide a better match to the data.

This work was done in collaboration with Merav Ahissar who plans to use this more detailed characterisation of behavior in simple psychophysical tasks as a way to understand statistical learning in psychiatric populations such as dyslexics and members of the autistic spectrum.

# Chapter 6

# General Conclusions

Three sets of psychophysical experiments demonstrating contextual effects in auditory perception have been reported along with computational models to explain their results. These models are based on the framework of perception as inference. The objects of study in these three experiments and the proposed explanations are closely related yet different.

## 6.1 Pitch and Frequency Shifts

In all three experiments, subjects were asked to report the direction of a shift in a dimension related to pitch. In chapter 5, pure tones were compared and frequency shift equates to pitch shift. In chapter 3, tones were designed to be ambiguous between two possible octave related values of pitch. In this study I assumed that the pitches extracted from each sound are the quantities being compared in the decision process. In one experiment of chapter 4, non harmonic complexes were compared and it is less clear what subjects reported. It is not pitch given that the non-harmonic complexes have no clear pitch. Instead, it appears to be a combination of local frequency shifts and this same the intuition underlies our proposed tracking model to account for the results. Two different mechanisms seem to underlie the perception of frequency and pitch shifts. The perception of the non-ambiguous pitch of harmonic complexes can be thought of as one of 'vertical' or 'synchronous' grouping. Tones in harmonic ratio are perceived as a whole, no single harmonic stands out. Pairs of harmonic complexes can be constructed with upward perceived change of

pitch but overall decrease of frequency content. This is to be opposed to the temporal or 'horizontal' grouping or binding of tones. When pairs of tones evoking a clear pitch are compared, subjects behave as if they first extracted the pitch of the two tones and then compared those. When the tones have no clear pitch, local frequency shifts are what drives subjects reports. The interaction of these two mechanisms is largely unknown but their existence can be justified by the fact that they support different functions of interest in auditory scene analysis.

## 6.2 A shared contraction bias explanation?

The two experiments reported in chapters 4 and 5 have striking similarities in their results and explanations. In both cases the percept seems to reflect the contraction of a stimulus towards a combination of those recently experienced. Our explanations however differ. In the case of pure tone discrimination, we were interested in the precise magnitude of the contraction of a tone represented in memory. This was possible due to the simplicity of the task design with simple pure tones presented separately. We derived predictions on how this contraction scales with assumed levels of noise and uncertainty in a normative model of the task. In the case of the biasing of ambiguous shifts, an implicit contraction breaks the symmetry causing the ambiguity of the task but in this analysis (that we performed chronologically earlier), it was not the main focus of the analysis. In principle, one could attempt to explain the disambiguation using the model we derived for the pure tone discrimination case. Indeed, given the nature of the stimuli used, statistical learning would lead to learn a mixture model from context tones with components aligned with the biasing regions. Contraction of the first tones towards the nearest 'bump' would then break the ambiguity in the desired direction. A shared explanation is however unlikely. In the Shepard tone experiment, a single tone context 3 semitones above $T_1$ almost completely resolves the ambiguity and leads to 90% of upward responses. Such a bias could only be explained by a contraction of the first tone by 2 to 3 semitones which would lead to very large biases even for test pairs with clear non-ambiguous shifts. This is clearly not the case, non ambiguous pairs of Shepard

$T_1, T_2$ with 1 semitone interval are hardly biased at all by the same context tone 3 semitones above $T_1$.

In this thesis, we suggest different mechanisms may be at play in these two experiments which makes a possible unification difficult and speculative. More work needs to be conducted to validate (or falsify) the models we proposed as guides to understand perceptual processes.

## 6.3 Summary of contributions to auditory neuroscience and future directions

The work presented in this thesis provides a better understanding of the dynamics of perception. Experiments of chapter 5 is the most detailed demonstration of the contraction and exhibits very clearly the different time-scales the integration of sensory history into perception. I revisited the decades old experimental paradigm of tone discriminations and showed how it can be used to study the dynamics of statistical learning, hence extending the range of its applications in auditory psychophysics. Our statistical learning model needs to be further validated. A key experiment would consist in observing behavior of subjects when the stimulus distribition is changed in the middle experiment. Would subject notice the change? How would it affect their learning of the stimulus statistics?

The work presented in chapter 4 provides further evidence that sensory history biases perception. Our proposed computational account, although rather tailored to the particular stimuli used in the experiments proposes an abstract way of understanding the effect of sensory history that is simpler and leads to a better explanation of the psychophysical data gathered so far than alternative neuro-mechanistic models. It also provides another way of reasoning about auditory scenes and their structure that might prove useful to design the next iteration of experiments.

Finally, the use of GAMs to study psychophysical data is not novel but proved to be crucial in these studies. The development of algorithms to perform complex regression analyses on large datasets is a methodological contribution that has

applications way beyond the scope of the particular experiments reported in this thesis. In future work, I plan to more widely share software implementing these methods to make them more readily available to the research community.

## Appendix A

# Duration dependence of tone Likelihood

The particular choice of the temporal dependence on tone precision is not arbitrary. Assume $\sigma^2$is the variance of a one second-long tone of log-frequency $f$. The likelihood of this tone given a group mean $\mu$ is $\mathcal{N}(f;\mu,\sigma^2)$.

$$\begin{aligned} p(f|\mu,\sigma^2) &= \mathcal{N}(f;\mu,\sigma^2) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(f-\mu)^2\right) \end{aligned}$$

Now we split this tone into $n$ consecutive subtones $(f_1,...,f_n)$ of duration $d(n) = 1/n$ and value $f$, and write $\sigma_n^2$ for the variance of the likelihood of each subtone. The new joint likelihood for this group of subtones is

$$\begin{aligned} p(f_1,...,f_n|\mu,\sigma^2) &= \prod_i \mathcal{N}(f_i;\mu,\sigma_n^2) \\ &\propto \exp\left(-\frac{1}{2{\sigma^2}_n}\sum_i(f_i-\mu)^2\right) \\ &\propto \exp\left(-\frac{n}{2{\sigma^2}_n}(f-\mu)^2\right) \end{aligned}$$

Equating the precisions for the single or group likelihood leads to $\frac{1}{{\sigma^2}_n} = \frac{n}{{\sigma^2}_n}$ or ${\sigma^2}_n = n\sigma^2 = \frac{\sigma^2}{d(n)}$ .

From this result, any duration $d$ can be approximated as $d \approx k.d(n)$, with $k = \lfloor d/d(n) \rfloor$, so the sensory variance for a tone of duration $d$ is $\frac{{\sigma_n}^2}{k} = \frac{1}{\lfloor d/d(n) \rfloor}\frac{\sigma^2}{d(n)} \approx \frac{\sigma^2}{d}$

## Appendix B

# Details of the derivations of the filtering procedure

Notation

- $p\left(T_t^{(k)}\right)$ prior at time $t$ on $k^{th}$ track mean
- $p(\eta_t)$ prior at all times on tone assignment
- $C_t^{(n)}$ : $n^{th}$-tone of chord $\mathbf{C}_t$ observed at time $t$

The joint at over variables at time $t$ is

$$\begin{aligned}\log p\left(T_t^{(1..K)}, \eta_t^{(1..N)}, C_t^{(1..N)}\right) &= \sum_n \sum_k \delta(\eta_t^{(n)} = k) \log p\left(C_t^{(n)} | T_t^{(k)}, \eta_t^{(n)} = k\right) \\ &+ \sum_n \sum_k \delta(\eta_t^{(n)} = k) \log p(\eta_t^{(n)} = k) \\ &+ \sum_k \log p(T_t^{(k)})\end{aligned}$$

The assumed factorization of the posterior over latent variables after the observation of a chord at time $t$ is

$$p\left(T_t^{(1..K)}, \eta_t^{(1..N)} | C_t^{(1..N)}\right) \approx \prod_k q\left(T_t^{(k)}\right) \prod_n q(\eta_t^{(n)})$$

We approximate the posterior by minimizing the Kullback Leibler divergence between the true posterior and the assumed factored form (fully factored variational expectation maximization [161, 162])

$$Q(T_t, \eta_t) = \arg\min_Q KL(Q(T_t, \eta_t) | p(X_t, \eta_t | C_t))$$

The variational updates for attribution are

$$\begin{aligned} q\left(\eta_t^{(n)}=k\right) &\propto \exp\left(\langle \log p\left(C_t^{(n)}|T_t^{(k)},\eta_t^{(n)}=k\right)\rangle_{q\left(T_t^{(k)}\right)} + \log p(\eta_t^{(n)}=k)\right) \\ &\propto \pi_k \exp\left(-\frac{1}{2\sigma_c^2}\langle\left(C_t^{(n)}-T_t^{(k)}\right)^2\rangle_{q\left(T_t^{(k)}\right)}\right) \\ &\propto \pi_k \exp\left(-\frac{1}{2\sigma_c^2}\left(\left(C_t^{(n)}-\mu_t^{(k)}\right)^2+\sigma_t^{(k)2}\right)\right) \\ &\propto \pi_k \mathcal{N}\left(C_t^{(n)};\mu^{(k)},\sigma_c^2\right)\exp\left(-\frac{1}{2}\sigma_t^{(k)2}/\sigma_c^2\right) \end{aligned}$$

Since tone attributions sum to one, we have

$$q\left(\eta_t^{(n)}=k\right) = \frac{\pi_k \mathcal{N}\left(C_t^{(n)};\mu^{(k)},\sigma_c^2\right) e^{-\frac{1}{2}\sigma_t^{(k)2}/\sigma_c^2}}{\sum_{k'}\pi_{k'}\mathcal{N}\left(C_t^{(n)};\mu^{(k')},\sigma_c^2\right)e^{-\frac{1}{2}\sigma_t^{(k')2}/\sigma_c^2}}$$

The variational updates for track statistics are

$$\begin{aligned} q\left(T_t^{(k)}\right) &\propto \exp\left(\log p(T_t^{(k)}) + \sum_n \langle\delta(\eta_t^{(n)}=k)\rangle_{q\left(\eta_t^{(n)}=k\right)} \log p\left(C_t^{(n)}|T_t^{(k)},\eta_t^{(n)}=k\right)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_t^{(k)2}}\left(T_t^{(k)}-\mu_t^{(k)}\right)^2 - \frac{1}{2\sigma_c^2}\sum_n r_n^{(k)}\left(T_t^{(k)}-C_t^{(n)}\right)^2\right) \\ &\propto \exp -\frac{1}{2}\left(T_t^{(k)2}\left[\frac{1}{\sigma_t^{(k)2}}+\frac{\sum_n r_n^{(k)}}{\sigma_c^2}\right] - 2X^{(k)}\left[\frac{\mu_t^{(k)}}{\sigma_t^{(k)2}}+\frac{\sum_n r_n^{(k)}C_t^{(n)}}{\sigma_c^2}\right]\right) \\ &= \mathcal{N}\left(X^{(k)};\mu_{t+}^{(k)},\sigma_{t+}^{(k)2}\right) \end{aligned}$$

with $r_n^{(k)} = \langle\delta(\eta_t^{(n)}=k)\rangle_{q\left(\eta_t^{(n)}=k\right)}$.

This leads to the track statistics updates:

$$\begin{aligned} \sigma_{t+}^{(k)2} &= \left(\frac{1}{\sigma_t^{(k)2}}+\frac{\sum_n r_n^{(k)}}{\sigma_c^2}\right)^{-1} \\ \mu_{t+}^{(k)} &= \left(\frac{\mu_t^{(k)}}{\sigma_t^{(k)2}}+\frac{\sum_n r_n^{(k)}C_t^{(n)}}{\sigma_c^2}\right) / \left(\frac{1}{\sigma_t^{(k)2}}+\frac{\sum_n r_n^{(k)}}{\sigma_c^2}\right) \end{aligned}$$

# Bibliography


[1] Heida Maria Sigurdardottir, Hilda Bjork Danielsdottir, Margret Gudmundsdottir, Kristjan Helgi Hjartarson, Elin Astros Thorarinsdottir, and Árni Kristjánsson. Problems with visual statistical learning in developmental dyslexia. *Scientific Reports*, 7(1):606, 2017.

[2] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

[3] Horace B Barlow. Possible principles underlying the transformations of sensory messages. 1961.

[4] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*, volume 9. Voss, 1867.

[5] Renwick E Curry. A bayesian model for visual space perception. In *Seventh Annual Conference on Manual Control*, volume 281, page 187, 1972.

[6] David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[7] Carlos Zednik and Frank Jäkel. Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, 193(12):3951–3985, 2016.

[8] PC Klink, RJA Van Wezel, and Raymond van Ee. United we sense, divided we fail: context-driven perception of ambiguous visual stimuli. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):932–941, 2012.

[9] Joel S Snyder, Olivia L Carter, Erin E Hannon, and Claude Alain. Adaptation reveals multiple levels of representation in auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4):1232, 2009.

[10] Marc O Ernst. A bayesian view on multimodal cue integration. *Human body perception from the inside out*, 131:105–131, 2006.

[11] Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

[12] Wilson S Geisler and Daniel Kersten. Illusions, perception and bayes. *Nature neuroscience*, 5(6):508–510, 2002.

[13] Xue-Xin Wei and Alan A Stocker. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.

[14] Pedro Rosas and Felix A Wichmann. Cue combination: Beyond optimality. *Sensory cue integration*, pages 144–152, 2011.

[15] Phillipp Hehrmann. *Pitch Perception as Probabilistic Inference*. PhD thesis, University of London, 2011.

[16] Peggy Seriès and Aaron R Seitz. Learning what to expect (in visual perception). *Frontiers in human neuroscience*, 7, 2013.

[17] Ahna R Girshick, Michael S Landy, and Eero P Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14(7):926–932, 2011.

[18] Ruyuan Zhang, Oh-Sang Kwon, and Duje Tadin. Illusory movement of stationary stimuli in the visual periphery: Evidence for a strong centrifugal prior in motion processing. *The Journal of Neuroscience*, 33(10):4415–4423, 2013.

[19] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604, 2002.

[20] Matthew Chalk, Aaron R Seitz, and Peggy Seriès. Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, 10(8):2–2, 2010.

[21] Luigi Acerbi, Sethu Vijayakumar, and Daniel M Wolpert. On the origins of suboptimality in human probabilistic inference. *PLoS Comput Biol*, 10(6):e1003661, 2014.

[22] Nikos Gekas, Matthew Chalk, Aaron R Seitz, and Peggy Seriès. Complexity and specificity of experimentally induced expectations in motion perception. *J Vis*, 13:1–18, 2013.

[23] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[24] Simon Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10):910–912, 1981.

[25] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

[26] Yang Dan, Joseph J Atick, and R Clay Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *The Journal of Neuroscience*, 16(10):3351–3362, 1996.

[27] Michael S Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363, 2002.

[28] Deep Ganguli and Eero P Simoncelli. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural computation*, 2014.

[29] Cathryn J Downing. Expectancy and visual-spatial attention: effects on perceptual quality. *Journal of Experimental Psychology: Human perception and performance*, 14(2):188, 1988.

[30] Ryota Kanai and Frans AJ Verstraten. Perceptual manifestations of fast neural plasticity: Motion priming, rapid motion aftereffect and perceptual sensitization. *Vision research*, 45(25):3109–3116, 2005.

[31] Stuart Anstis, Frans AJ Verstraten, and George Mather. The motion aftereffect. *Trends in cognitive sciences*, 2(3):111–117, 1998.

[32] Jason Fischer and David Whitney. Serial dependence in visual perception. *Nature Neuroscience*, 17(5):738–743, 2014.

[33] Peggy Seriès, Alan A Stocker, and Eero P Simoncelli. Is the homunculus aware of sensory adaptation? *Neural Computation*, 21(12):3271–3304, 2009.

[34] Neel T Dhruv and Matteo Carandini. Cascaded effects of spatial adaptation in the early visual system. *Neuron*, 81(3):529–535, 2014.

[35] Alan Stocker and Eero P Simoncelli. Sensory adaptation within a bayesian framework for perception. *Advances in neural information processing systems*, 18:1289, 2006.

[36] Caspar M Schwiedrzik, Christian C Ruff, Andreea Lazar, Frauke C Leitner, Wolf Singer, and Lucia Melloni. Untangling perceptual memory: Hysteresis and adaptation map into separate cortical networks. *Cerebral Cortex*, 24(5):1152–1164, 2014.

[37] Johannes C Dahmen, Peter Keating, Fernando R Nodal, Andreas L Schulz, and Andrew J King. Adaptation to stimulus statistics in the perception and neural representation of auditory space. *Neuron*, 66(6):937–948, 2010.

[38] Makio Kashino and Shinya Nishida. Adaptation in the processing of interaural time differences revealed by the auditory localization aftereffect. *The Journal of the Acoustical Society of America*, 103(6):3597–3604, 1998.

[39] Alexander Gutschalk, Christophe Micheyl, and Andrew J Oxenham. The pulse-train auditory aftereffect and the perception of rapid amplitude modulations. *The Journal of the Acoustical Society of America*, 123(2):935–945, 2008.

[40] Makio Kashino and Minae Okada. The role of spectral change detectors in sequential grouping of tones. In *Auditory signal processing*, pages 195–201. Springer, 2005.

[41] ZJ Shu, NV Swindale, and MS Cynader. Spectral motion produces an auditory after-effect. *Nature*, 364(6439):721–723, 1993.

[42] David Alais, Emily Orchard-Mills, and Erik Van der Burg. Auditory frequency perception adapts rapidly to the immediate past. *Attention, Perception, & Psychophysics*, 77(3):896–906, 2015.

[43] Ningyuan Wang and Andrew J Oxenham. Spectral motion contrast as a speech context effect. *The Journal of the Acoustical Society of America*, 136(3):1237–1245, 2014.

[44] Ofri Raviv, Merav Ahissar, and Yonatan Loewenstein. How recent history affects perception: the normative approach and its heuristic approximation. *PLoS Comput Biol*, 8(10):e1002731, 2012.

[45] Claire Chambers and Daniel Pressnitzer. Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, & Psychophysics*, 76(5):1271–1279, 2014.

[46] J Giangrand, Betty Tuller, and JA Scott Kelso. Perceptual dynamics of circular pitch. *Music Perception: An Interdisciplinary Journal*, 20(3):241–262, 2003.

[47] CJ Darwin and NS Sutherland. Grouping frequency components of vowels: When is a harmonic not a harmonic? *The Quarterly Journal of Experimental Psychology*, 36(2):193–208, 1984.

[48] Claire Chambers, Sahar Akram, Vincent Adam, Claire Pelofi, Maneesh Sahani, Shihab Shamma, and Daniel Pressnitzer. Prior context in audition informs binding and shapes simple features. *Nature Communications*, 8, 2017.

[49] Itay Lieder, Vincent Adam, Maneesh Sahani, and Merav Ahissar. Sensory history affects perception through online updating of prior expectations. *Cosyne Abstracts, Salt Lake City USA*, 2017.

[50] David Marr. A computational investigation into the human representation and processing of visual information. *Vision*, pages 125–126, 1982.

[51] Toshio Irino and Roy D Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America*, 101(1):412–419, 1997.

[52] Vincent Adam, James Hensman, and Maneesh Sahani. Scalable transformed additive signal decomposition by non-conjugate gaussian process inference. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, pages 1–6. IEEE, 2016.

[53] Vincent Adam, Lea Duncker, and Maneesh Sahani. Continuous-time point-process gpfa using sparse variational methods. *Cosyne Abstracts, Salt Lake City USA*, 2017.

[54] Vincent Adam, Claire Chambers, Maneesh Sahani, and Daniel Pressnitzer. Pre-perceptual grouping accounts for contextual dependence in the perception of frequency shift. *Cosyne Abstracts, Salt Lake City USA*, 2016.

[55] Vincent Adam and Maneesh Sahani. Bayesian perception of the pitch of non-stationary natural sounds. *Cosyne Abstracts, Salt Lake City USA*, 2014.

[56] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[57] Michalis K Titsias. Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

[58] Matthias Seeger. Expectation propagation for exponential families. Technical report, 2005.

[59] Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural Comput.*, pages 786–792, 2009.

[60] Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(10), 2008.

[61] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[62] James Hensman, Alex Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. *arXiv preprint arXiv:1411.2005*, 2014.

[63] Alexander G de G Matthews, James Hensman, Richard E Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, 2016.

[64] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[65] David K Duvenaud, Hannes Nickisch, and Carl E Rasmussen. Additive gaussian processes. In *Advances in neural information processing systems*, pages 226–234, 2011.

[66] Nicolas Durrande, David Ginsbourger, Olivier Roustant, and Laurent Carraro. Additive covariance kernels for high-dimensional gaussian process modeling. *arXiv preprint arXiv:1111.6233*, 2011.

[67] Alan Saul, James Hensman, Aki Vehtai, and Neil D Lawrence. Chained Gaussian processes. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, 2016.

[68] Ryan J Giordano, Tamara Broderick, and Michael I Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015.

[69] M Yu Byron, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in neural information processing systems*, pages 1881–1888, 2009.

[70] Yuan Zhao and Il Memming Park. Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *arXiv preprint arXiv:1604.03053*, 2016.

[71] American National Standards Institute. ANSI s1.1-1994 (r2004): Acoustical terminology, 1994.

[72] Adrianus JM Houtsma. Pitch and timbre: Definition, meaning and use. *Journal of New Music Research*, 26(2):104–115, 1997.

[73] Stephen McAdams and Albert Bregman. Hearing musical streams. *Computer Music Journal*, pages 26–60, 1979.

[74] William Heil Lichte. Attributes of complex tones. *Journal of Experimental Psychology*, 28(6):455, 1941.

[75] John M Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[76] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres:

Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3):177–192, 1995.

[77] Alain De Cheveigne. Pitch perception models. In *Pitch*, pages 169–233. Springer, 2005.

[78] Alain de Cheveigné. Pitch perception. *The oxford handbook of auditory science: Hearing*, pages 71–104, 2010.

[79] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory neuroscience: Making sense of sound*. MIT press, 2011.

[80] Julius L Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *The Journal of the Acoustical Society of America*, 54(6):1496–1516, 2005.

[81] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[82] Björn Lindblom and Johan Sundberg. The human voice in speech and singing. In *Springer handbook of acoustics*, pages 669–712. Springer, 2007.

[83] G Fant. Acoustic theory of speech production. the hague, the netherlands: Mouton & co, 1960.

[84] James L Flanagan. *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media, 2013.

[85] Roy D Patterson, K Robinson, J Holdsworth, D McKeown, C Zhang, and M Allerhand. Complex sounds and auditory images. *Auditory physiology and perception*, 83:429–446, 1992.

[86] Malcolm Slaney et al. An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 35:8, 1993.

[87] IJ Russell and PM Sellick. Low-frequency characteristics of intracellularly recorded receptor potentials in guinea-pig cochlear hair cells. *The Journal of Physiology*, 338:179, 1983.

[88] Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[89] R Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011.

[90] Vincent Adam and Maneesh Sahani. Bayesian perception of the pitch of non-stationary natural sounds. *Cosyne Abstracts, Salt Lake City USA*, 2014.

[91] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[92] Ernst Terhardt, Gerhard Stoll, and Manfred Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, 71(3):679–688, 1982.

[93] Roy D Patterson, Mike H Allerhand, and Christian Giguere. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894, 1995.

[94] William A Yost. Pitch of iterated rippled noise. *The Journal of the Acoustical Society of America*, 100(1):511–518, 1996.

[95] Jan Frederik Schouten. The perception of subjective tones. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, 41:1086–1093, 1938.

[96] BL Riker. The ability to judge pitch. *Journal of Experimental Psychology*, 36(4):331, 1946.

[97] Roger N Shepard. Geometrical approximations to the structure of musical pitch. *Psychological review*, 89(4):305–333, 1982.

[98] JD Warren, Stefan Uppenkamp, Roy D Patterson, and Timothy D Griffiths. Separating pitch chroma and pitch height in the human brain. *Proceedings of the National Academy of Sciences*, 100(17):10038–10042, 2003.

[99] Roger N Shepard. Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, 36(12):2346–2353, 1964.

[100] Diana Deutsch, F Richard Moore, and Mark Dolson. The perceived height of octave-related complexes. *The Journal of the Acoustical Society of America*, 80(5):1346–1353, 1986.

[101] Bruno H Repp. Spectral envelope and context effects in the tritone paradox. *Perception*, 26(5):645–665, 1997.

[102] Lloyd A Dawe, John R Platt, and Eydra Welsh. Spectral-motion aftereffects and the tritone paradox among canadian subjects. *Perception & psychophysics*, 60(2):209–220, 1998.

[103] Bruno H Repp and Jacqueline M Thompson. Context sensitivity and invariance in perception of octave-ambiguous tones. *Psychological research*, 74(5):437–456, 2010.

[104] Matthias J Sjerps, James M McQueen, and Holger Mitterer. Evidence for precategorical extrinsic vowel normalization. *Attention, Perception, & Psychophysics*, 75(3):576–587, 2013.

[105] Lori L Holt. Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4):305–312, 2005.

[106] Peter Ladefoged and Donald Eric Broadbent. Information conveyed by vowels. *The Journal of the Acoustical Society of America*, 29(1):98–104, 1957.

[107] Peter G Thompson and J Anthony Movshon. Storage of spatially specific threshold elevation. *Perception*, 7(1):65–73, 1978.

[108] Dennis H Klatt. Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5):1208–1221, 1976.

[109] Aniruddh D Patel, John R Iversen, and Jason C Rosenberg. Comparing the rhythm and melody of speech and music: The case of british english and french. *The Journal of the Acoustical Society of America*, 119(5):3034–3047, 2006.

[110] Albert Bachem. Tone height and tone chroma as two different pitch qualities. *Acta Psychologica*, 7:80–88, 1950.

[111] Huang Chengcheng. *Neuromechanistic models for auditory perception*. PhD thesis, New York University, 2015.

[112] Yael Bitterman, Roy Mukamel, Rafael Malach, Itzhak Fried, and Israel Nelken. Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature*, 451(7175):197–201, 2008.

[113] Christian E Stilp, Paul W Anderson, and Matthew B Winn. Predicting contrast effects following reliable spectral properties in speech perception. *The Journal of the Acoustical Society of America*, 137(6):3466–3476, 2015.

[114] Bernhard Englitz, S Akram, SV David, C Chambers, Daniel Pressnitzer, D Depireux, JB Fritz, and Shihab A Shamma. Putting the tritone paradox into context: insights from neural population decoding and human psychophysics. In *Basic Aspects of Hearing*, pages 157–164. Springer, 2013.

[115] Chengcheng Huang, Bernhard Englitz, Shihab Shamma, and John Rinzel. A neuronal network model for context-dependence of pitch change perception. *Frontiers in Computational Neuroscience*, 9:101, 2015.

[116] Laurent Demany and Christophe Ramos. On the binding of successive sounds: Perceiving shifts in nonperceived pitches. *The Journal of the Acoustical Society of America*, 117(2):833–841, 2005.

[117] Laurent Demany, Daniel Pressnitzer, and Catherine Semal. Tuning properties of the auditory frequency-shift detectors. *The Journal of the Acoustical Society of America*, 126(3):1342–1348, 2009.

[118] Li I Zhang, Andrew YY Tan, Christoph E Schreiner, and Michael M Merzenich. Topography and synaptic shaping of direction selectivity in primary auditory cortex. *Nature*, 424(6945):201–205, 2003.

[119] Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1, 1972.

[120] Albert S Bregman. Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3):380, 1978.

[121] Leo Paulus Antonie Servatius van Noorden et al. *Temporal coherence in the perception of tone sequences*. PhD thesis, Technische Hogeschool Eindhoven, 1975.

[122] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

[123] Kevin JP Woods and Josh H McDermott. Attentive tracking of sound sources. *Current Biology*, 25(17):2238–2246, 2015.

[124] Patrick Cavanagh and George A Alvarez. Tracking multiple targets with multifocal attention. *Trends in cognitive sciences*, 9(7):349–354, 2005.

[125] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[126] Mounya Elhilali and Shihab A Shamma. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6):3751–3771, 2008.

[127] Gautham Mysore and Maneesh Sahani. Variational inference in non-negative factorial hidden markov models for efficient audio source separation. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1887–1894, 2012.

[128] Shihab A Shamma, Mounya Elhilali, and Christophe Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences*, 34(3):114–123, 2011.

[129] Brian CJ Moore and Hedwig E Gockel. Properties of auditory stream formation. *Phil. Trans. R. Soc. B*, 367(1591):919–931, 2012.

[130] Stuart M Anstis and Shinya Saida. Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance*, 11(3):257, 1985.

[131] David Burr and Guido Marco Cicchini. Vision: Efficient adaptive coding. *Current Biology*, 24(22):R1096–R1098, 2014.

[132] Paymon Ashourian and Yonatan Loewenstein. Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one*, 6(5):e19551, 2011.

[133] Sarah R Allred, L Elizabeth Crawford, Sean Duffy, and John Smith. Working memory and spatial judgments: Cognitive load increases the central tendency bias. *Psychonomic bulletin & review*, pages 1–7, 2016.

[134] Harry Levi Hollingworth. The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17):461–469, 1910.

[135] S Chase, P Bugnacki, LD Braida, and NI Durlach. Intensity perception. xii. effect of presentation probability on absolute identification. *The Journal of the Acoustical Society of America*, 73(1):279–284, 1983.

[136] SR Purks, DJ Callahan, LD Braida, and NI Durlach. Intensity perception. x. effect of preceding stimulus on identification performance. *The Journal of the Acoustical Society of America*, 67(2):634–637, 1980.

[137] Nikos Gekas, Aaron R Seitz, and Peggy Seriès. Expectations developed over multiple timescales facilitate visual search performance. *Journal of vision*, 15(9):10–10, 2015.

[138] Martin Wiener, James C Thompson, and H Branch Coslett. Continuous carryover of temporal context dissociates response bias from perceptual influence for duration. *PloS one*, 9(6):e100803, 2014.

[139] JE Berliner, NI Durlach, and LD Braida. Intensity perception. vii. further data on roving-level discrimination and the resolution and bias edge effects. *The Journal of the Acoustical Society of America*, 61(6):1577–1585, 1977.

[140] Eustace Christopher Poulton. *Bias in quantifying judgements*. Taylor & Francis, 1989.

[141] Maria Olkkonen, Patrice F McCarthy, and Sarah R Allred. The central tendency bias in color perception: Effects of internal and external noise. *Journal of vision*, 14(11):5–5, 2014.

[142] Oliver Dyjas, Karin M Bausenhart, and Rolf Ulrich. Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics*, 74(8):1819–1841, 2012.

[143] Martin Lages and Michel Treisman. Spatial frequency discrimination: visual long-term memory or criterion setting? *Vision research*, 38(4):557–572, 1998.

[144] Sagi Jaffe-Dax, Ofri Raviv, Nori Jacoby, Yonatan Loewenstein, and Merav Ahissar. A computational model of implicit memory captures dyslexics' perceptual deficits. *The Journal of Neuroscience*, 35(35):12116–12126, 2015.

[145] Ofri Raviv, Itay Lieder, Yonatan Loewenstein, and Merav Ahissar. Contradictory behavioral biases result from the influence of past stimuli on perception. *PLOS Comput Biol*, 10(12):e1003948, 2014.

[146] Karin M Bausenhart, Oliver Dyjas, and Rolf Ulrich. Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, 77(4):1033–1043, 2015.

[147] Arash Fassihi, Athena Akrami, Vahid Esmaeili, and Mathew E Diamond. Tactile perception and working memory in rats and humans. *Proceedings of the National Academy of Sciences*, 111(6):2331–2336, 2014.

[148] E Schwartz, R Romo, and Y Loewenstein. The computational principles and neural mechanisms underlying contraction bias, 2008.

[149] Walt Jesteadt, R Duncan Luce, and David M Green. Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1):92, 1977.

[150] DM Green and JA Swets. Signal detection theory and psychophysics. *Society*, 1:521, 1966.

[151] Ingo Fründ, Felix A Wichmann, and Jakob H Macke. Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of vision*, 14(7):9–9, 2014.

[152] Vinzenz H Schönfelder and Felix A Wichmann. Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *The Journal of the Acoustical Society of America*, 131(5):3953–3969, 2012.

[153] Stian Reimers and Neil Stewart. Auditory presentation and synchronization in adobe flash and html5/javascript web experiments. *Behavior Research Methods*, 48(3):897–908, 2016.

[154] Destiny L Babjack, Brandon Cernicky, Andrew J Sobotka, Lee Basler, Devon Struthers, Richard Kisic, Kimberly Barone, and Anthony P Zuccolotto. Reducing audio stimulus presentation latencies across studies, laboratories, and hardware and operating system configurations. *Behavior research methods*, 47(3):649–665, 2015.

[155] Irwin Pollack. Comfortable listening levels for pure tones in quiet and noise. *The Journal of the Acoustical Society of America*, 24(2):158–162, 1952.

[156] J Verschuure and AA Van Meeteren. The effect of intensity on pitch. *Acta Acustica united with Acustica*, 32(1):33–44, 1975.

[157] Stanley S Stevens. The relation of pitch to intensity. *The Journal of the Acoustical Society of America*, 6(3):150–154, 1935.

[158] Murray F Spiegel and Charles S Watson. Performance on frequency-discrimination tasks by musicians and nonmusicians. *The Journal of the Acoustical Society of America*, 76(6):1690–1695, 1984.

[159] Arman Abrahamyan, Laura Luz Silva, Steven C Dakin, Matteo Carandini, and Justin L Gardner. Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences*, page 201518786, 2016.

[160] Pete R Jones, David R Moore, Daniel E Shub, and Sygal Amitay. The role of response bias in perceptual learning. *Journal of experimental psychology: learning, memory, and cognition*, 41(5):1456, 2015.

[161] Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

[162] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.