Pitch Perception as Probabilistic Inference

Phillipp Hehrmann

Gatsby Computational Neuroscience Unit

University College London

London, United Kingdom

THESIS

Submitted for the degree of

Doctor of Philosophy, University of London

Declaration

I, Phillipp Hehrmann, 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 thesis.

Hannover, 2$^{\text{nd}}$ December 2011

Abstract

Pitch is a fundamental and salient perceptual attribute of many behaviourally important sounds, including animal calls, human speech and music. Human listeners perceive pitch without conscious effort or attention. These and similar observations have prompted a search for mappings from acoustic stimulus to percept that can be easily computed from peripheral neural responses at early stages of the central auditory pathway. This tenet however is not supported by physiological evidence: how the percept of pitch is encoded in neural firing patterns across the brain, and where – if at all – such a representation may be localised remain as yet unsolved questions.

Here, instead of seeking an explanation guided by putative mechanisms, we take a more abstract stance in developing a model by asking, what computational goal the auditory system is set up to achieve during pitch perception. Many natural pitch-evoking sounds are approximately periodic within short observation time windows. We posit that pitch reflects a near-optimal estimate of the underlying periodicity of sounds from noisy evoked responses in the auditory nerve, exploiting statistical knowledge about the regularities and irregularities occurring during sound generation and transduction. We compute (or approximate) the statistically optimal estimate using a Bayesian probabilistic framework.

Model predictions match the pitch reported by human listeners for a wide range of well-documented, pitch-evoking stimuli, both periodic and aperiodic. We then present new psychophysical data on octave biases and pitch-timbre interactions in human perception which further demonstrates the validity of our approach, while posing difficulties for alternative models based on autocorrelation analysis or simple spectral pattern matching.

Our model embodies the concept of perception as unconscious inference, originally proposed by von Helmholtz as an interface bridging optics and vision. Our 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.

Acknowledgements

Above all, my thanks and gratitude go to my supervisor Maneesh Sahani, whose insightful guidance, impeccable intuitions and never-resting analytic mind have saved my academic endeavours from stagnation time and time again.

I thank Peter Dayan for shaping the Gatsby Unit into the exceptionally stimulating environment that it is, and Peter Latham for re-hiring me so many times. My studies and travels were generously supported by the Gatsby Charitable Foundation. I have made many acquaintances with fellow students, postdocs and visitors, and I am deeply grateful to all of them: to David Barrett and Gabi Teodoru for their dedication to keeping the artistic spirit alive and hosting so many wonderful, memorable evenings filled with musical joy and pizza; to Ross Williamson above all for being Ross, but also for his helpful comments on parts of this thesis; to Vinayak and Charles for generously sharing their machine-learning expertise in hours of need; to Andriy for bringing art-house cinema to the seminar room; to Biljana for her caring advice and support; to Ulrik, Amy, Adam, Jakob, David, Loic and Ritwik for unforgettable travel memories; and to all the others, which together have made spending almost five years at the Gatsby Unit so very valuable and enjoyable.

I thank David McAlpine for sharing his inspiring passion for science with me, as well as Jennifer Linden, Björn, Lucy, Maria, Jannis and other members of the Ear Institute with whom I have had the privilege to interact. I have had many insightful discussions with Rich Turner that influenced the work presented in this thesis; similarly with Alain de Cheveigné, who I owe special thanks alongside Mark Plumbley for kindly agreeing to act as my examiners.

Outside the walls of academia, Georgia has been a truly wonderful friend and source of moral support. I also want to thank Race and Gerri for their generosity and kindness.

My parents, Kathi and Rainer, have always given me nothing but support and encouragement, and I cannot thank them enough for granting me so many exceptional opportunities in life. It has been an absolute delight to watch the young families of my brothers Johannes and Matthias grow, albeit from a distance, and I look forward to being able to spend more time with them in the near future.

Contents

Front matter

List of figures

List of tables

List of algorithms

Chapter 1

Introduction

1.1 Motivation, methodology and aims

1.1.1 The relevance of pitch

A common property of many behaviourally-significant sounds in our natural acoustic environment is the percept of pitch which they evoke in human listeners and, presumably, several animal species. Animals, even invertebrates, use pitched sounds as an integral part of their territorial and mating behaviour. The sophisticated social behaviour found in some mammalian species would be impossible without the use of shared systems of acoustic communication, human speech being the most richly structured amongst them. For most communication sounds, pitch — rather than being purely epiphenomenal — conveys information of interest to the listener, ranging from cues regarding the species, gender and size of the sound source to the intended semantic content. In all spoken languages, pitch carries prosodic information, i.e. additional semantic connotations beyond the written words (Bolinger, 1978). Tonal languages (such as Mandarin Chinese) use pitch also to distinguish lexical items and grammatical categories. In music, the definitions of musical scales, melody and harmony are unthinkable without reference to our percept of pitch. There is also evidence for a role of pitch in solving the "cocktail party problem", the ubiquitous challenge of separating sound sources in cluttered acoustic environments: differences in pitch can induce stream segregation between otherwise spectrally similar sounds (Vliegen and Oxenham, 1999), and they have been shown to substantially improve identification of simultaneously-presented vowels (de Cheveigné et al., 1997) and speech utterances (Brokx and Nooteboom, 1982). The fundamental nature of pitch perception is further highlighted by the fact that human listeners are seemingly able to hear the pitch of complex sounds without need for conscious effort or attention.

1.1.2 The puzzle about pitch

The phenomenon of pitch, fundamental as it may appear, nevertheless continues to evade a definitive explanation. What makes pitch difficult to study and grasp is the somewhat tautological observation that it does not constitute an objective, physical attribute of a sound, but a subjective aspect of the listener's perceptual experience. While this rules out a physical *definition* of pitch, one may nevertheless investigate the *relationship* between the physical attributes of sounds and the percept evoked in listeners. In order to do so, we first need to define an objective behavioural measure to quantify this elusive percept. Roughly speaking, the pitch of a target sound is commensurate to the frequency of a sinus tone which is judged as equal in pitch by a listener. Owing to its subjective nature, this judgement need not be the same across different listeners, or even across multiple stimulus repetitions for a single listener. The beginnings of the modern era in the psychophysical and physiological study of pitch can be traced back to the siren experiments of August Seebeck and the following debate with Georg Ohm in the mid-nineteenth century (Turner, 1977). Since then, considerable effort has been spent to uncover the relationship between acoustic stimuli and perceptual experience, as well as the underlying physiological processes in the ear and the brain. On the one hand, the resulting body of experimental results is almost overwhelming in its breadth and detail. On the other hand, scientific consensus regarding the interpretation of these results seems almost as far out of reach today as it did in the days of Ohm's and Seebeck's initial debate.

Even long before the nineteenth century, it had been known that the pitch of a periodic sound[1] typically corresponds to the inverse of its period. A continued source of both puzzlement and insight are artificially-designed sounds that evoke a pitch percept despite being highly *aperiodic*. Consider for example a white-noise signal — entirely aperiodic and unpitched — that is multiplied with a sinusoidal envelope: the resultant sound has a pitch, albeit weak, equal to the envelope frequency, even though no

[1] i.e. a sound consisting of exact repetitions of a single, short waveform segment

segment of the waveform ever repeats. Its envelope, however, is perfectly periodic by construction, just like the envelope of a truly periodic sound. Might pitch therefore be related to the periodicity of the waveform envelope, rather than the full waveform with all its fine structure? Firstly, this alone would not explain the faintness of the percept in the latter example compared to that of sinus tone or any other truly periodic sound. Secondly, if we multiply the same sinusoidal envelope with a high-frequency sinusoidal carrier instead of noise, the resultant sound again has a pitch, but it equals neither the frequency of the envelope nor that of the carrier. We will postpone further discussion until later — suffice it to say that there appears to be no single physical feature of sounds to which the percept of pitch is simply and consistently related. Accepting that a purely signal-based, "obvious" explanation is unlikely to exist, how else can or should one approach this difficult modelling problem? How does the central auditory system integrate signals arriving from a great number of peripheral sound receptors, driven by periodic or aperiodic sounds, over time into our unified percept of pitch?

1.1.3 Modelling methodologies

David Marr, in his seminal book on human vision, distinguishes three complementary levels at which information processing systems (such as the visual or auditory system) can be described and studied (Marr, 1982, chap. 1) :

1. *computational theory*, primarily concerned with identifying the goal or purpose of the system under study, and the strategy employed to achieve it;

2. *algorithm and representation*, studying what algorithms underlie the system's input-output transformation in order to achieve its goal, and the nature of their internal representation;

3. *hardware implementation*, the mechanisms by which these algorithms and representation are realised in the actual, physical system under study.

Considering the variety of proposed models of the human "pitch processor", there seems to be a bias for solutions that can be readily computed from peripheral neural responses at early stages of the central auditory pathway. This bias may stem from the apparent automaticity and ease with which human listeners can perceive pitch, or perhaps from the common use of pitched sounds for the purpose of acoustic communication even by much simpler, non-mammalian species lacking the advanced computational resources of the mammalian or human brain. However, this bias is not well supported by physiological evidence. Even today, investigations into the neural basis of pitch perception using electrophysiology, EEG/MEG or functional imaging techniques have not been able to establish conclusively, what biophysical mechanisms give rise to pitch, how it is encoded by neural firing patterns across the brain, and where — if at all — such a representation may be localised. Implementational and algorithmic considerations, in the sense of Marr's hierarchy, may therefore be of limited value as a starting point for developing models of human pitch perception. In this thesis, we present a model derived from computational principles instead: we define pitch as the optimal solution to a putative computational goal of the auditory system during listening.

1.1.4 Pitch as inference

What then is the goal of the auditory system? One of the most influential, computational theories of human perception even nowadays predates Marr's analysis by more than a century[2]. Hermann von Helmholtz proposed that perception reflects a process of unconscious inference about physical quantities of interest in the environment from imperfect and incomplete incoming sensory signals (von Helmholtz, 1867). Most natural, pitch-evoking sounds are approximately, though not perfectly, periodic within short observation time windows. Building on previous work by Goldstein (1973), we hypothesise 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. We formulate our model within the framework of Bayesian probabilistic inference (e.g. MacKay, 2003), which provides both the formal language to define this inference problem rigorously, and the algorithmic tools to compute (or approximate) its optimal solution. It is worth pointing out that our approach is not "blindly" computational in the sense that it disregards the algorithmic and physical levels in Marr's hierarchy entirely. By incorporating a statistical description of the peripheral neural response, on which inference in the model is based, our computational account does in fact explicitly obey fundamental, biophysically-warranted representa-

[2] Like Marr's theory, it was originally formulated in the context of vision.

tional constraints. We do, however, remain agnostic with regard to the algorithmic and biophysical implementation of the process of inference itself.

Pitch, as a correlate of periodicity estimation, is sometimes conceived of as a “primitive”, purely data-driven bottom-up cue that serves to support the more interesting and challenging “schema-driven” tasks of auditory scene analysis and beyond[3], which are influenced by learnt knowledge, prior expectations and other sources of top-down modulation (Bregman, 1990). Conversely, we will argue in this thesis that even a seemingly primitive auditory percept like pitch already reflects the outcome of a sophisticated inferential process, efficiently combining bottom-up sensory evidence with top-down expectations derived from long-term natural scene statistics.

1.2 Thesis overview

Chapter 2 establishes the background knowledge required for subsequent chapters. In particular, we will review and discuss key aspects of:

- the *psychophysics* of pitch perception,
- the *physiology* of the peripheral and central auditory system as pertains to the processing and representation of pitch, and
- existing *theories and models* of pitch perception.

In **chapter 3**, we present the formal definition of a generative, probabilistic model of near-periodic sounds and evoked responses in the auditory nerve. Two variants will be discussed that differ in their treatment of the relationship between periodicity and spectral envelope features of sounds. We will introduce the concept of pitch perception as Bayesian inference and present two algorithms for the approximate computation of optimal periodicity estimates based on the statistical assumptions embodied in our generative model.

Chapter 4 demonstrates the basic consistency of our model estimates with human psychophysics for a representative range of well-documented pitch-evoking sounds. We will show that the model accounts not only for the pitch of naturalistic, periodic sounds but also for that of unnatural, aperiodic sounds despite being poorly described by the

[3] such as segregation and identification of sound sources from mixtures, or semantic parsing

assumed generative process. We also discuss several phenomena which the model fails to capture. In particular, human pitch perception of harmonic sounds appears to be sensitive to spectral features other than their harmonicity, whereas no corresponding effects are evident in the model behaviour.

In **chapter 5**, we will review the evidence for a dependency between pitch and timbre, both in the statistics of natural, pitch-evoking sounds and in their perception by human listeners. We go on to extend the acoustic component of our generative model to account for this dependency and use a database of natural sounds to fit the parameters of this newly-introduced coupling. We then design and conduct a psychophysical experiment to test the predictions of the extended model regarding the influence of timbral features on the pitch of non-uniform pulse trains, a class of periodic sounds with a controllable degree of octave ambiguity. Behavioural effects observed in human listeners are well-predicted by our Bayesian model, while posing a difficult challenge to alternative, non-inferential models as well as our earlier, uncoupled model. Finally, we demonstrate that the extended model also provides a parsimonious solution to some of the issues raised in chapter 4.

Chapter 6 concludes this thesis with a discussion of the contributions achieved by our work so far and by pointing out several promising directions for future research.

Chapter 2

Background

2.1 Fundamentals of pitch perception

Pitch is a prominent subjective attribute of our perceptual experience of certain types of sounds, for example the voiced parts of human speech or those produced by many musical instruments. Despite being subjective, the relationship between physical stimulus and its evoked percept is nevertheless far from arbitrary. Otherwise, the core constructs of classical music theory[1] could not feasibly exist. The definitions of musical scales and intervals are based on the concept of pitch. Rules regarding the sequential and simultaneous arrangement of notes with different pitches (i.e. melody, harmony and counterpoint), meant to differentiate between what sounds pleasant or unpleasant to a listener, could hardly be effective if the pitches perceived during a musical performance varied arbitrarily from one listener to the next. But what are the fundamental, preserved features of pitch? First and foremost, pitch is most strongly evoked by periodic sounds and is largely determined by their *periodicity*. Second and corollary, pitch is *invariant* to a variety of substantial changes in the acoustic signal: two instruments can produce distinctly different sounds that nevertheless give rise to the same pitch despite gross differences in loudness and “tone colour”, or timbre[2] (for example the mellow, husky quality of a flute in its low register or the sharp, piercing sound of a distorted electric guitar). Not only can listeners tell if a sound has a pitch or not, and whether the pitches evoked by two sounds are the same or different: there also appears

[1] Western, Indian, Chinese, Arab-Persian and others

[2] Timbre is often treated as that perceptual attribute sounds which allows us to distinguish sounds of equal loudness, duration and pitch (ANSI, 1994). See also our discussion of timbre in section 5.1.1.

to be the general notion of an *ordering* amongst different pitches along some internal dimension. If we label, somewhat arbitrarily, the two extremes of this dimension as “low” and “high” (see e.g. Ashley (2004) for a discussion of alternative metaphors used in other cultural communities), then we can ask whether one pitch is higher than another. This is reflected in the definition of pitch according to the American National Standards Institute (ANSI, 1994):

> “Pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high.”

In this section, we will examine and review (rather descriptively) some key findings regarding the dependence of our subjective percept on physical parameters of the stimulus. Despite its strong dependence on periodicity, we will find that pitch is not simply related to a single, well-defined physical stimulus parameter. We will discuss past and present attempts to synthesise these diverse psychophysical findings into a coherent theory of the mapping between stimulus and percept later on in section 2.4.

The ANSI definition above makes no reference to the fundamental role of pitch in music. This was, in fact, made explicit in an earlier definition by the American Standards Association (ASA, 1960), whereby pitch “is that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale”. Of course, this is a problematic definition, considering that musical scales in turn are defined in terms of the pitches from which they are constructed. However, it points towards an important additional requirement for a sound-evoked sensation to classify as pitch, which is (tacitly or overtly) made by many psychoacousticians: that the sensation be able to support the recognition of musical intervals and melodies (e.g. Attneave and Olson, 1971; Burns and Viemeister, 1976; Semal and Demany, 1990; Pressnitzer et al., 2001). This requirement is of practical value as is sets an approximate lower limit for the accuracy of the perceptual ordering of sounds: the smallest musical interval in Western classical music is a semitone, corresponding to a frequency ratio of approximately $\sqrt[12]{2}$. If the accuracy of the ordering drops below this limit, interval and melody recognition are bound to be impaired.

Let us assume then, in agreement with both the ASA and ANSI definitions above, that the pitch of a sound can be assigned a position, or magnitude, along an internal scale. A major obstacle in the psychophysical study of pitch is that most listeners are unable to assign consistent verbal labels to the magnitude of their percept without external reference. This rare ability called “absolute pitch” is prevalent in as little as 0.01% of the general population[3] (Levitin and Rogers, 2005). The remainder of listeners may still be able to remember, recognise and produce absolute pitch magnitudes to some degree even without explicit labels (e.g. Bachem, 1937; Lockhead and Byrd, 1981), but their precision lacks greatly in comparison to possessors of genuine absolute pitch. Instead, they can report reliably only the *relative* change in pitch from one note to the other: whether the first pitch was higher or lower than the second. In addition, musically-trained listeners will typically be able to determine and label the magnitude of the change in terms of a musical interval, but we will not consider this ability of interval recognition any further for the remainder of this thesis.

The overwhelming majority of pitch-evoking sounds encountered naturally are periodic up to small deviations. Any periodic sound $x(t)$ with a period duration of Ω can be uniquely decomposed into a sum of constituent sinusoidal vibrations, $x(t) = \sum_{k=0}^{\infty} a_k \sin(\frac{2\pi k}{\Omega} t + \theta_k)$, according to Fourier’s famous theorem (e.g. Hartmann, 1997). Here, a_k is the amplitude of the k-th Fourier component, θ_k is its phase at $t = 0$ and $\frac{k}{\Omega}$ its frequency, which is the k-th integer multiple of the repetition rate $\frac{1}{\Omega} =: f_0$ of the original sound $x(t)$. In this Fourier representation, the simplest possible periodic sound is itself a sinusoid, as it cannot be further decomposed via Fourier analysis. If we fix its amplitude and phase to some arbitrary values, it is effectively parameterised only by its frequency. If one asks listeners to judge to relative pitches of two sinusoids with different frequencies, one will find their pitches are monotonically related to their frequencies: the sinusoid with the higher frequency will be judged higher in pitch. In this way, we can construct a frequency-labelled pitch scale, which is related to the listener’s internal scale via some unknown, monotonically increasing mapping function. With this frequency-labelled scale at hand[4], one can now determine the pitch of (almost) any other pitch-evoking target sound, again up to this unknown monotonic mapping, simply by matching the adjustable frequency of a pure tone to the target. The outcome of this matching is the “pure-tone equivalent” pitch of the target, measured in Hz. Variants of this procedure are amongst the most commonly used methods to quantify a listener’s pitch percept. The choice of using pure tones as the initial reference point is admittedly arbitrary. However, as we will soon discuss, the pitch of the large majority of periodic

[3] but seemingly higher amongst Asians

[4] as for example noted in an addendum to the ASA definition of pitch (ASA, 1960)

sounds with period Ω is in fact equal to $\frac{1}{\Omega}$. Thus, many periodic sounds can serve as a reference in the matching procedure almost equivalently. This can be extremely useful in cases where a gross difference in timbre between the target sound and a pure tone could otherwise impair the listener's matching performance. Furthermore, the use of complex stimuli helps to avoid covariation of other salient perceptual stimulus features with pitch that are unavoidable when using pure tones (see also de Cheveigné, 2010 for a more detailed discussion of this topic).

For all *our* intents and purposes, this matching-based definition of pitch will be sufficient. Nevertheless, several researchers have attempted to further quantify the implicit mapping between the listener's internal pitch scale and the pure-tone frequency scale in Hz. The results of these studies are not entirely consistent. Assuming that listeners can numerically compare values on their internal pitch scale, Stevens et al. (1937) had their subjects adjust the frequency of a pure tone, such that its *subjective pitch* was half that of a reference tone with a certain fixed frequency. This procedure was repeated for a range of reference frequencies and the resultant "fractionations" were averaged across subjects. From these average fractionations, a perceptual pitch scale in units of "mel" (as in "melody") was constructed: first, 1000 mel was simply defined as the pitch value of a 1000 Hz tone. Next, the frequency corresponding to a pitch of 500 mel was determined as the frequency that was deemed half as high in pitch as 1000 mel, and similarly for 250 mel and so on. The resultant curve of pitch in mel plotted against frequency in Hz is shown in Figure 2.1. Similar relationships between frequency and subjective pitch magnitude have been found using other methods, such as equisection (Stevens and Volkmann, 1940) and magnitude estimation (Beck and Shaw, 1961; see Stevens, 1971 for a methods overview). The derived scales, however, seem to be at odds with the commonly held notion that musical intervals correspond to certain, fixed distances between pitches. In music, the melodic distance or interval between two notes is determined by the ratio between the frequencies: a ratio of 1.5, for example, will almost always be heard as a "fifth", independent of the absolute frequencies. Two melodies will be recognised as the same, as long as the frequency ratios between successive notes remain preserved — independent of the starting note[5]. This suggests that a certain difference in subjective pitch corresponds to a fixed frequency ratio or equivalently: a logarithmic relationship between subjective pitch and frequency. Indeed, Attneave

[5] Possessors of absolute pitch will notice that it has been transposed, but can nevertheless identify it.

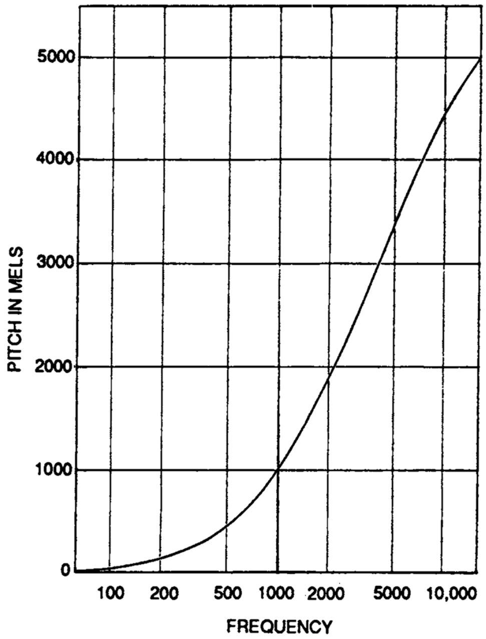


Figure 2.1: The mel scale of pitch (from Stevens et al., 1937).

and Olson (1971) obtained essentially this result in an experiment, where subjects were required to transpose pairs of pure tones, separated by a musical interval smaller than an octave, into a different frequency region by (continuous) adjustment of a frequency dial[6]. The reasons for this discrepancy between the mel-scale and the logarithmic scale of musical pitch are not entirely clear (see e.g. van Norden, 1982). The subjects in the latter experiment could have deliberately made their judgements so as to preserve the musical interval, instead of taking equidistant steps on their internal pitch scale. Alternatively, the subjects of Stevens et al. (1937), given no "melodic" context whatsoever, may have made their judgement not based on pitch but on the unavoidably covarying spectral centre of mass, which is regarded as a salient aspect of timbre. Whatever the resolution may be, we will content ourselves with the operational definition of pitch as the pitch-matched frequency of a pure tone, measured in units of Hz, throughout the remainder of this thesis unless explicitly stated otherwise.

In considering pitch as a percept along a single, monotonic dimension, we (and the ANSI) may have disregarded a further important aspect of pitch perception. Seemingly, pitch similarity is not simply determined by the magnitude of the pitch difference (be it on a linear, logarithmic or mel-scale). Instead, we perceive two pitches as highly

[6] Logarithmic scaling broke down for frequencies higher than approximately 5 kHz.

similar if their frequencies are related by one or even several octaves, i.e. when the frequency ratio is an integer power of 2. In pitch matching paradigms, subjects are often observed to make octave mistakes (e.g. Riker, 1946; cf. chapter 5), and even possessors of absolute pitch are prone to this kind of confusion (e.g. Bachem, 1937). Octave equivalence is a cross-culturally shared feature in many musical systems. The scales in Western classical music are based on subdivisions of the octave into smaller intervals up to approximately one twelfth of an octave (on a logarithmic scale, corresponding to a frequency ratio of $\sqrt[12]{2}$), and notes separated by one or several octaves are denoted by the same letter. Similarly, the scales of Indian, Arab-Persian and Chinese classical music (amongst others) are based on subdivisions of the octave, albeit different from those in classical Western scales (Burns, 1998). This has led to the concept of the position within an octave as a second, circular dimension called *tone chroma*, in addition to our first dimension, which scales monotonically with frequency and which is called *tone height* in contexts where this distinction is made (Bachem, 1950). Shepard (1982) proposed a simple spatial representation of pitch in these two dimensions in the shape of a helix, where the chroma dimension winds around a tone-height axis (see Figure 2.2).

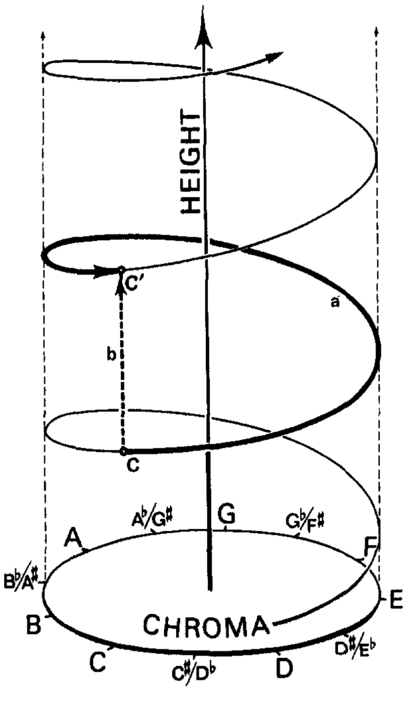


Figure 2.2: Helical representation of pitch height and chroma (from Shepard, 1982).

There are doubts as to whether perceived octave similarity is truly attributable to the pitch relation *per se*, or whether it might rather reflect our sense of musical consonance, i.e. the pleasantness of the sensation of two or more musical notes played simultaneously. The spectra of octave-related *natural* pitch-evoking sounds, composed of many near-harmonic frequency components, overlap to a higher degree than in any other interval relationship without creating perceptual "beats". Such beats — for lack of a better description: a "wha-wha"-like sensation of slow amplitude modulations due to the physical interaction of nearby spectral components in the peripheral auditory system (cf. section 2.2.3.2) — are held to be a major determinant of our sense of consonance (von Helmholtz, 1863; Plomp and Levelt, 1965; McDermott et al., 2010). Thus, we may perceive harmonic sounds in octave intervals as similar because they are equally *consonant* as sounds in unison. At face value, experiments by Riker (1946) seem to argue against this hypothesis: Riker observed that octave matching-mistakes were in fact more common for pure tones than for notes played on a piano, even though there is nothing supremely consonant about pure-tone spectra at octave intervals. However, the discussion is complicated by the fact that non-linear distortions in the middle or inner ear (see 2.2.3.2) can artificially introduce spectral components at multiples of the pure-tone frequencies, effectively rendering them into harmonic complex tones. Furthermore, our sense of octave equivalence could be *acquired* through repeated exposure to natural stimuli but subsequently influence our judgement of other pitch-evoking stimuli such as pure tones. In any case, even if tone chroma is a second dimension of pitch, it would appear that it is always uniquely determined by tone height (while the reverse is not true). We will consider pitch only as varying along a single monotonic dimension, as suggested by the ANSI definition and following the majority of psychophysical studies of pitch to date.

It should be pointed out that a stimulus, strictly speaking, does not "have" a pitch. Nevertheless, we will often use this imprecise but convenient terminology. The pitch evoked by one and the same physical stimulus may be different from one presentation to the next. What we typically mean when we say that a stimulus "has a certain pitch", is that the stimulus reliably evokes a near-identical percept over many trials and across many listeners. A sound which leads to a broad distribution of pitch estimates in each listener can be said to have weak pitch. Sounds which lead to a distinctly multimodal distribution of pitches can be said to have an ambiguous pitch. We will consider such a class of stimuli, namely non-uniform periodic click trains (Flanagan et al., 1962), later on in chapter 5. Shepard tones (Shepard, 1964) are another (much-studied) example of ambiguous pitch stimuli: complex tones with octave-spaced harmonics and a bell-shaped spectral envelope centred around some spectral centroid frequency, which have a clearly identifiable chroma and a highly ambiguous tone height. The source of perceptual inter-trial variability in these cases is not simply the stochastic nature of sensorineural transduction or differences in the peripheral gain due to either random fluctuations or loudness changes (cf. section 2.2). Instead, there is strong evidence that the pitch of a single such stimulus is substantially influenced *perceptually* by the sequence of preceding stimuli, presumably on the basis of their pitch (Dawe et al., 1998; Giangrand et al., 2003; Repp and Thompson, 2010; Chambers and Pressnitzer, 2011). Context effects are not limited solely to biases in octave judgements, but can also result in more accurate interval matching (Attneave and Olson, 1971) or pitch discrimination (Warrier and Zatorre, 2002) in the presence of an extended melodic context.

2.1.1 The pitch of periodic sounds

2.1.1.1 Pure tones

Given our matching-based procedure of measuring pitch above, this section may seem superfluous at first sight. Is not the pitch of a pure tone its own frequency *by definition*? Not quite. Firstly, there are limits to the range of frequencies over which pure tones can be frequency matched. Guttman and Pruzansky (1962) found that listeners were able to do so down to frequencies around 20 Hz, i.e. the lower limit of the human hearing range. Musical pitch, however, did not extend to such low frequencies. When asking listeners to transpose a reference tone up or down by an octave, Guttman and Pruzansky (1962) discovered that listeners' average deviation from the mathematical octave frequency increased to over a semitone for reference frequencies between approximately 39 and 46 Hz. More importantly, their response *variability* increased dramatically from around 60 Hz downwards. The upper limit of musical pure-tone pitch has been investigated in several studies (Ward, 1954; Attneave and Olson, 1971; Semal and Demany, 1990), all of them revealing a break-down of the percept in the range between 4–5 kHz. This coincides also with the limit of note-naming ability in possessors of absolute pitch found by Bachem (1948). Within these outer limits, the ability to discriminate small

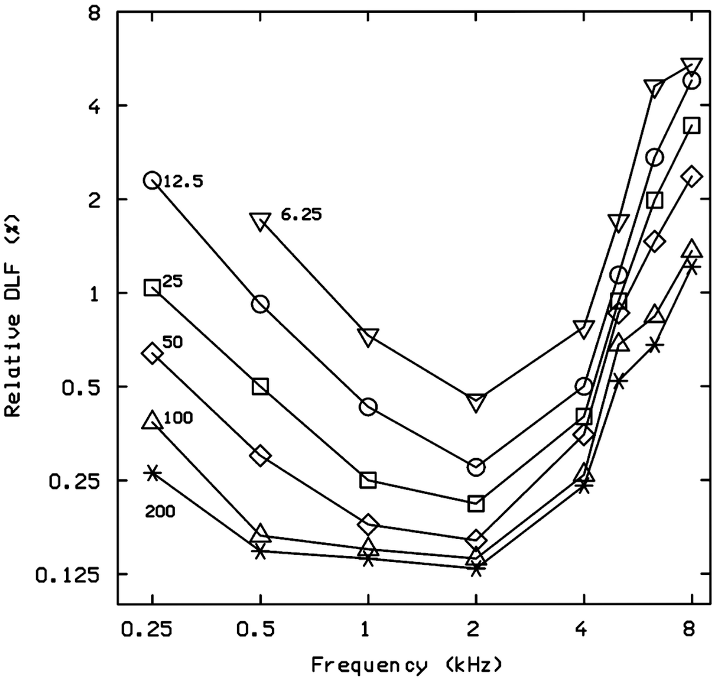


Figure 2.3: Pure-tone frequency difference limens for different stimulus frequencies and durations (different curves are labelled by duration in ms; data from Moore, 1973; figure adapted from Moore, 2003).

differences in frequency depends non-monotonically on frequency. Moore (1973) found that frequency difference limens — the smallest difference in frequency for which subjects can reliably determine (e.g. with 75% correct) which of two tones is higher — were smallest at around 2 kHz and increased monotonically for frequencies above and below[7] (see Figure 2.3). Discriminability also improved with tone duration.

The pitch of a pure tone depends to some degree on its loudness. When pitch-matching an 80 dB pure tone with an adjustable 40 dB tone, one would find that the two frequencies are not identical. While the level-dependence of pure-tone pitch is generally found to be negligible ($\leq$ 1%) in the range between 1–2 kHz, some studies (e.g. Stevens, 1935; Morgan et al., 1951), but not all, have found more substantial effects outside this range (see e.g. Moore, 2003 for a discussion). The direction of the pitch-shift is frequency-dependent. As a general rule, pitch below 2000 Hz decreases for higher levels and increases with level above 4000 Hz (ibid.). This effect calls for a specification of the reference level in matching-based measurement of pitch. To further complicate matters, the pitch evoked by the same stimulus can differ between the left and right ear of a single listener by several percent, a phenomenon known as *diplacusis binauralis*

[7] Stimuli were presented via loudness-calibrated headphones.

(Burns, 1982).

2.1.1.2 Harmonic complex tones

The vast majority of natural pitch-evoking sounds are approximately periodic, i.e. they can be (approximately) decomposed via Fourier analysis into a sum of sinusoids with frequencies equal to integer multiples of the sound periodicity rate, or *fundamental frequency* $f_0 = \frac{1}{\Omega}$, where Ω is the period duration (see above). Sounds of this kind are produced by the vocal tracts of humans and animals and by many musical instruments. By and large, the pitch of such a *harmonic complex tone* (HCT) with many harmonics is equal to its f_0.

The lower limit of the f_0-range over which HCTs evoke a pitch percept has been found to coincide with the limits of pure-tone pitch reasonably well. In a task where listeners were required to detect a semitone-difference between two four-tone melodies, Pressnitzer et al. (2001) determined the lower limit of melodic HCT pitch as somewhere within the range of 32–40 Hz. Krumbholz et al. (2000) found a similar limit in terms of f_0-discriminability, suggesting that the limits of matching-based and musical pitch are very similar for HCTs. The upper limit of melodic HCT pitch is commonly assumed to coincide with the limit of pure-tone pitch, but this is based on indirect evidence. The highest pitches on the piano and piccolo flute (the highest orchestral instrument) fall into the range of 4–5 kHz, and due to the physiological limitations of the peripheral auditory system, auditory nerve fibres can presumably maintain periodic, synchronised firing activity only up to comparable rates[8].

A particular type of HCT, which has attracted much attention and controversy in the related literature (cf. section 2.4), are HCTs with *missing fundamental*, i.e. in which the amplitude of the Fourier component with frequency f_0 is 0. Even in this case, a pitch equal to the missing fundamental is heard — at least within certain limits depending on the total number of Fourier components in the stimulus, their respective ranks (i.e. their factors with respect to f_0) and to some degree their phase relationships (see e.g. Ritsma, 1962, 1963; Houtsma and Goldstein, 1971; Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Renken et al., 2004). Various terms have been used to describe the pitch of the missing fundamental: tonal residue (Schouten, 1940;

[8] Note though, that the exact physiological limit in humans isn't known (cf. sections 2.2.3.2 and 3.2.2).

Schouten et al., 1962), periodicity pitch (Plomp, 1967), low pitch (Smoorenburg, 1970) or virtual pitch (Terhardt, 1974). As a general rule of thumb, missing-f_0 pitch is strongest when the HCT contains many low-rank harmonics. Nevertheless, Houtsma and Goldstein (1971, 1972) showed that even a HCT with only two components can elicit missing-f_0 pitch. The pitch of a HCT notably weakens when the lowest harmonic has a rank above approximately 8 to 10. When only few and high harmonics are present in the spectrum, the pitch corresponding to the f_0 vanishes and instead one or several pitches corresponding to the individual component frequencies can be heard. The pitch of HCTs with only high harmonics is susceptible to manipulations in the component phases, whereas phase manipulations amongst low-rank harmonics leave the pitch (both pitch frequency and strength) largely unaffected. This can be observed, for example, in the doubling of the perceived pitch of a high-harmonic HCT when the phases are adjusted so as to double the periodicity of the waveform *envelope* while leaving the periodicity of the actual, fine-structured waveform and its f_0 unchanged (Shackleton and Carlyon, 1994). It has previously been suggested (ibid.) that both the weakening of missing-f_0 pitch with increasing lowest harmonic number and its phase dependence are caused by the limited frequency resolution of the peripheral auditory system (cf. sections 2.2.3.2 and 3.2.1). The frequency separation between high harmonics (starting approximately around the 10th, somewhat dependent on f_0; see e.g. Moore and Gockel, 2011) is insufficient to evoke discernable peaks in the average firing-rate profile of the auditory nerve, eliminating a potentially important cue regarding the harmonicity (or periodicity) of the stimulus. At the same time, high harmonics exciting the same peripheral filter cause amplitude modulations in the filter output, the depth and (in extreme cases) the rate of which depend on the relative component phases. However, more recent evidence suggests that two different mechanisms underlie the weakening and increased phase-dependence of high-harmonic missing-f_0 pitch respectively. While peripheral resolution is indeed the likely cause for phase-related effects, it seems that the weakening of the pitch has a more central underlying mechanism and is more closely related to harmonic number *per se*, rather than peripheral resolvability: pitch discriminability of high-harmonic HCTs remains poor even when the harmonics are made resolvable by dichotic presentation of odd and even harmonics to opposite ears (Bernstein and Oxenham, 2003). The theoretical implications of these findings regarding the role of resolved, unresolved, low and high harmonics will be further discussed in section 2.4.

Experiments by Ritsma (1962) give us an indication regarding the limits of missing-f_0 pitch perception, albeit only for a restricted class of stimuli. Using sinusoidally amplitude-modulated (SAM) tones, i.e. complex tones with only three adjacent spectral components[9], he found that the pitch sensation generally deteriorated for f_0s of 800 Hz and above, independent of the frequency range of the partials. Furthermore, no pitch of f_0 could be heard if all component frequencies were above 5 to 6 kHz independent of the fundamental. Recently, however, Oxenham et al. (2011) showed that genuine, even musical missing-f_0 pitch can be evoked for f_0s up to at 2 kHz and partial frequencies wholly outside the traditionally-assumed existence region up to 6 kHz (using HCTs with up to 12 harmonics, considerably more than Ritsma (1962)).

If one adds a sufficient number N of harmonically-related sinusoids of equal amplitude in *cosine phase*, i.e. such that their peaks all coincide at the same point in time, the resultant waveform $x(t) = \sum_i a_i \cos(2\pi f_0 it)$ takes the shape of a periodic train of narrow, unipolar pulses or "clicks", spaced at intervals of $\Omega = 1/f_0$. As $N \to \infty$, $x(t)$ turns into sum of infinitely narrow Dirac pulses:

$$x(t) \to \sum_i \delta(t - i\Omega) \quad , \tag{2.1}$$

where

$$\delta(t) = \lim_{\Delta t \to 0} \begin{cases} \frac{1}{2\Delta t} & \text{if } t \in [-\Delta t; \Delta t] \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

is the Dirac δ-function[10] (e.g. Hartmann, 1997). Any periodic sound with period duration Ω can be written as the convolution of a Ω-spaced Dirac pulse train with a finite impulse response that equals the shape of single waveform-period between pulses. This is the basis for the source-filter model of human vocal production (Fant, 1960), and will similarly form the basis of our own statistical model of naturalistic sounds in chapter 3. The pitch of periodic pulse trains with finite pulse durations, produced by mechanical sirens, was the subject matter of a scientific debate between Ohm and Seebeck (Ohm, 1843; Seebeck, 1843) that shaped theories of pitch perception for almost a century (see

[9] Note that SAM tones are *not* harmonic in general. Two spectral side-band components are equally spaced in linear frequency around a central carrier frequency, the component spacing being determined to the modulation rate. A SAM tone is only harmonic, if the carrier frequency is an integer multiple of the modulation rate. The energy in the side bands is controlled by the modulation depth.

[10] or rather: one of its many representations

section 2.4). Seebeck doubled the periodicity of an isochronous pulse train by shifting the time of every other pulse by a small, constant amount. The perceived pitch halved, even though the sound contained only very little energy at the new, lower f_0, explicitly suggesting that the physical presence of the fundamental may be not be necessary in order to perceive a pitch at its frequency. Flanagan and Guttman (1960), Flanagan et al. (1962) and Guttman and Flanagan (1964) conducted related experiments using electronically generated pulse trains and investigated the dependence of pitch on stimulus parameters such as pulse timing, polarity and amplitude. In chapter 5, we will use trains of bipolar pulses with periodically alternating amplitudes in a psychophysical experiment regarding the dependence of pitch not only periodicity but also other spectral (or timbral) properties of sounds.

When listening to a complex sound, some listeners are more likely to "hear out" individual harmonics, each with a pitch equal to its frequency, rather than perceiving a single, compound sound with a pitch equal to the fundamental. von Helmholtz (1863) called these two modes of listening as "analytic" and "synthetic". Many listeners, especially those without musical training, find it difficult to listen analytically at all and listen synthetically by default. As Helmholtz points out, "a certain amount of undisturbed concentration is always necessary for analysing musical tones by ear alone" (ibid.). Aside from analytic listening being amenable to practice, listeners' preference for analytic or synthetic listening can to some degree be influenced by the experimenter by careful priming[11]. Furthermore, the presentation of stimuli in a noisy background appears to bias listeners to adopt a synthetic listening mode (see Moore, 2003 for a more detailed discussion). In our model (cf. chapter 3), only synthetic listening will be considered explicitly.

2.1.2 Non-periodic sounds

Sounds do not need to be periodic in order to elicit a pitch. We can, for example, add random noise to a periodic sound with period Ω — a pitch corresponding to Ω will still be heard, typically as soon as the HCT can be reliably detected within the noise background (Gockel et al., 2006). In this case, every successive stimulus segment of

[11] see e.g. Houtsma and Goldstein (1971) for a discussion of a study by D. Cross and H. Lane, *Attention to Single Stimulus Properties in the Identification of Complex Tones* in Experimental Analysis of the Control of Speech Production and Perception, ORA Report No. 05613-1-P, University of Michigan, Ann Arbor, 1963

duration Ω still bears some resemblance to one and the same underlying, average period "template". But even sounds where this is no longer the case can evoke a sensation of pitch, strong enough allow for both rate discrimination and recognition of musical intervals or melodies.

If a high-frequency pure tone of frequency f_c is sinusoidally amplitude modulated (SAM) with an envelope of lower frequency g, the resultant SAM tone has a spectrum with three discrete frequency components: one at f_c and two side-bands at $f_c \pm g$. The waveform may still be periodic, if f_c and g share a common integer submultiple, but even if this is not the case, a pitch can be heard as long as f_c and g fall within a certain range (Ritsma, 1962, 1963; cf. our discussion regarding the limits of missing-f_0 pitch above). Except in special cases however, its matched frequency equals neither f_c, g or the periodicity of the SAM waveform as a whole. The pitch of this type of sound has been extensively studied by Schouten (1938, 1940) and collaborators (Schouten et al., 1962). To a first approximation, the pitch of SAM tones can be described in terms of f_c and g by considering the spectral components as *shifted* harmonics of a missing-f_0 HCT with fundamental g (see Figure 2.4). Let n be the rank of the harmonic of g that is closest in frequency to f_c and Δf the frequency difference between the two. The pitch f_p of the SAM tone is approximately equal to

$$f_p = g + \frac{\Delta f}{n} \quad , \tag{2.3}$$

i.e. its frequency is approximately proportional to the amount of shift with respect to the harmonic stack $\{g, 2g, \ldots\}$. Interestingly, the pitch of SAM tones is ambiguous: as if listeners "misjudged" the value of n, i.e. the rank of the harmonic of g that is closest to f_c, they sometimes perceive the pitch as either higher or lower by one or

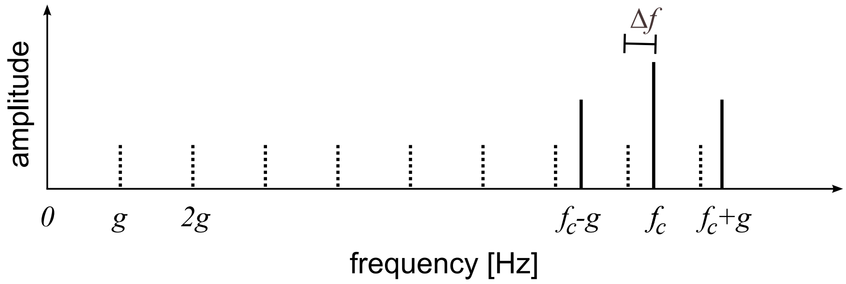


Figure 2.4: Amplitude spectrum of a sinusoidally amplitude-modulated tone. Dotted lines indicate multiples of the modulator frequency g (not present in the spectrum).

several discrete steps. For example, the pitch of a SAM tone with $f_c = 2040\,\text{Hz}$ and $g = 200\,\text{Hz}$ is typically heard as 204 Hz ($n = 10$), but sometimes also as 227 Hz ($n = 9$) or 185 Hz ($n = 11$) (Schouten, 1940). We will discuss the historic and theoretical importance of these findings in section 2.4. Quantitatively similar pitch shifts have also been observed for complex sounds with more than three (de Boer, 1956b; Patterson, 1973, 1976) or only two adjacent frequency components (Smoorenburg, 1970; Houtsma and Goldstein, 1972).

Rather than shifting all components of a HCT by the same amount simultaneously, a HCT can also be rendered inharmonic by mistuning a single frequency component. In this case, the "synthetic" pitch of the entire sound shifts to a small degree (up to about 1%) in the direction of the component mistuning (Moore et al., 1985). However, as the mistuning of the component exceeds approximately 3%, the pitch shift gradually vanishes and instead the percept segregates: the mistuned component can be heard as a separate, second pitch in addition to the fundamental pitch of the remaining harmonic components. Moore et al. (1985) found that harmonics between the second and fifth yielded the greatest overall effect, but inter-subject variability was high. Other studies, in which multiple harmonics were mistuned at a time, either congruently (Plomp, 1967; Ritsma, 1967) or incongruently (Dai, 2000), have generally come to the conclusion that mistunings of low-rank harmonics up to maximally the sixth have the greatest effect on pitch. This has led to the notion of *dominance* of some low-rank components over the pitch of the entire complex — qualitatively in agreement with our earlier observation that the pitch of missing-f_0 HCTs is strongest when they contain at least some low-rank partials (see section 2.1.1.2).

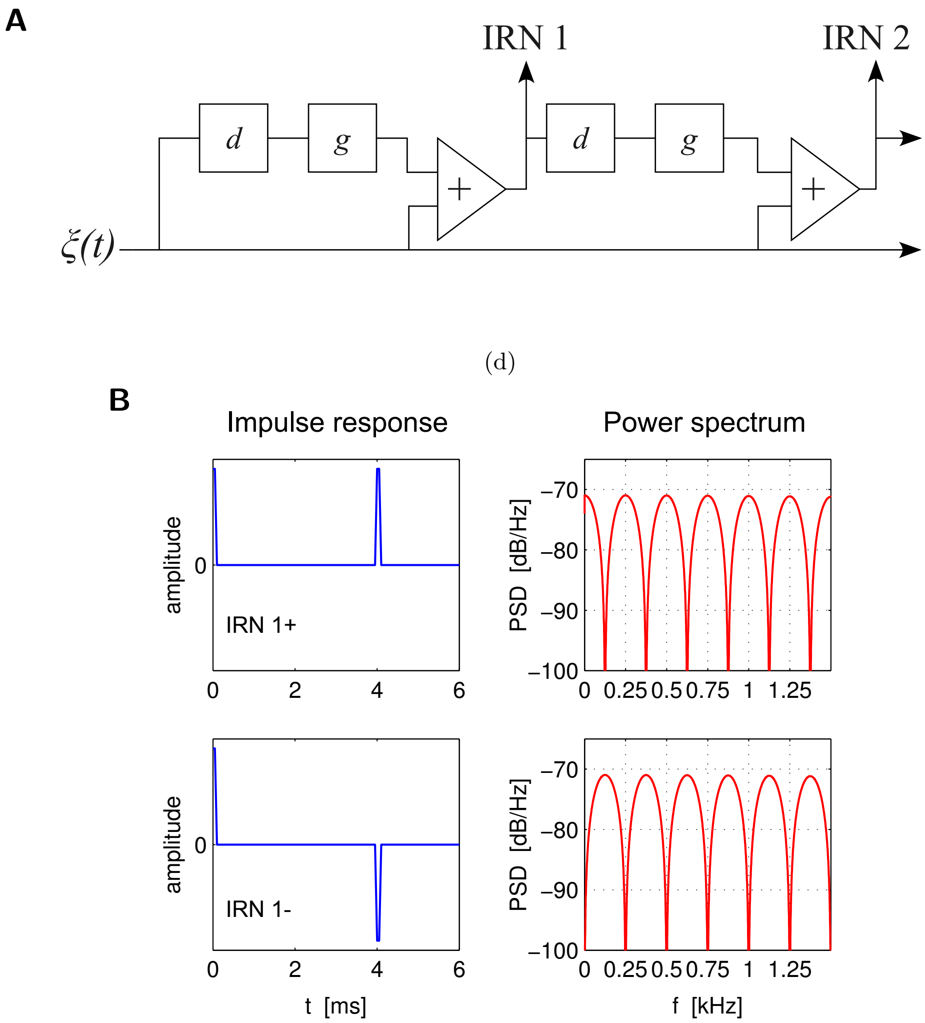
Amongst the most surprising examples of pitch-evoking, non-periodic sounds are those that are obtained not by simple manipulations of an originally periodic sound (such as the examples above), but instead by imposing weak spectral or temporal structure on otherwise perfectly aperiodic, uncorrelated noise. A much-studied example of this type of sound is *rippled noise* (Bilsen, 1966). Rippled noise is generated by delaying and adding a white noise signal back to itself with a time-delay of d and some multiplicative gain g. If this process is repeated more than once with identical delay and gain, *iterated rippled noise* (IRN) is obtained (Yost et al., 1996; Yost, 1996)[12]. We will denote IRN

[12] See Hartmann (1997) for an account of the discovery of an IRN-like pitch phenomenon by Dutch physicist Christiaan Huygens in 1693, generated by the sound of a water fountain being reflected off a flight of stairs in a court-yard.

with a positive gain of $g = 1$ as "IRN $n+$", where n is the number of iterations, and IRN with a negative gain $g = -1$ as "IRN $n-$". For IRN $n+$, a pitch is heard corresponding to the inverse delay $\frac{1}{d}$. This pitch is weak for $n = 1$ but becomes stronger as the number of delay-add iterations increases. The pitch of IRN $n-$ depends on n: for a single iteration, pitch matches are broadly distributed around two peaks at frequencies of approximately $\frac{1}{d} \pm 10\%$ with a distinct lack of matches at $\frac{1}{d}$. For higher number of iterations, the pitch drops to a frequency of $\frac{1}{2d}$, i.e. half the pitch of IRN $n+$, with a unimodal distribution around this mean (Yost, 1996). For $n > 1$, the pitch of IRN stimuli, as well as its increasing strength with n, is intuitively explicable by considering the delay-and-add/subtract procedure as a filtering operation. A single delay-and-add iteration can be expressed as a convolution of the original signal with an impulse response $h^+(t) = \delta(0) + \delta(d)$, the power spectrum of which is periodically modulated at a (spectral) rate of $\frac{1}{d}$ and peaks at frequencies $\frac{i}{d}, i = 0, 1, \ldots$ (Figure 2.5 B, top). Assuming that the spectral envelope of the source signal is flat, the spectral envelope of the filter output will have the same shape as the filter itself. Iterating the delay-and-add process as in Figure 2.5A introduces additional, shifted δ-peaks to $h(t)$, which leads to a progressive sharpening of the spectral modulations in the filter output. For $n \to \infty$, the spectrum is a perfect comb-spectrum, just like that of a Dirac pulse-train (see above). The delay-and-subtract operation can be implemented as a filter with impulse response $h^-(t) = \delta(0) - \delta(d)$, the power spectrum of which is also modulated at a rate of $\frac{1}{d}$, but with a first peak at $\frac{1}{2d}$ (Figure 2.5 B, bottom). Iterating this procedure also sharpens the spectrum, but it will more and more resemble that of an HCT with $f_0 = \frac{1}{2d}$ and odd-numbered harmonics only[13]. As the spectrum of IRN with many iterations becomes effectively harmonic, the pitch of IRN is most "interesting" in the low-n regime. In these cases, IRN provides only weak spectral cues (particularly after filtering by the peripheral auditory system). Furthermore, the dichotomy between the bimodal pitch of IRN $1-$ and the unimodal pitch of IRN $1+$ is not obvious from our simple spectral intuition above (cf. Figure 2.5).

Sinusoidally amplitude-modulated (SAM) noise, i.e. a white noise carrier signal multiplied with a sinusoidal envelope, is a pitch-evoking stimulus that lacks spectral features altogether. Even though the signal *envelope* is clearly periodic, the power spectrum of the signal altogether is perfectly flat in expectation, just as that of the noise car-

[13] The fundamental frequency is counted as the first harmonic



(d)

Figure 2.5: A: Schematic circuit for the generation of IRN up to $n = 2$. Uncorrelated Gaussian noise $\xi(t)$ is fed into two adders that receive progressively delayed and gain-modulated copies of $\xi(t)$ (after Yost, 1996). B: Impulse response (left) and power spectrum (right) of the filters corresponding to a single iteration in the generation of IRN 1+ (top) and IRN 1− (bottom).

rier itself. Nevertheless, a faint pitch corresponding to the envelope modulation rate can be heard for modulation rates up to 850-1000 Hz that allows for the recognition of rhythm-less musical melodies (closed-set and open-set), melody dictation and interval recognition and production (Burns and Viemeister, 1976, 1981). The accuracy of the pitch of SAM noise is on the order of one semitone (even with practice), i.e. barely sufficient for the use in a musical context.

High-pass or low-pass filtered white noise also gives rise to a weak pitch (Small and Daniloff, 1967; Fastl, 1971). When the filter slopes are steep, a pitch close to the cut-off frequency can be heard for low-pass frequencies below approximately 5-10 kHz (depending on the study) and high-pass frequencies above 500-600 Hz. When noise is band-pass filtered with a passband wider than about one fifth of an octave, two pitches corresponding to the two spectral edges can be heard. For narrower passbands, a single pitch is perceived around the centre frequency (Fastl, 1971; see also Fastl and Zwicker, 2007). Small and Daniloff (1967) speculated that the pitch in these cases may be caused by lateral, neural inhibition between adjacent peripheral frequency channels (cf. sections 2.2.3.2 and 2.3.1), creating an excitation peak around the noise edge frequency where inhibition can act only from one side, rather than both sides along the tonotopic axis (see also von Békésy, 1963 and Houtgast, 1972).

2.1.3 Binaural pitch

Pitch can even be evoked by dichotic stimuli, where the waveform at either ear alone has the statistics of pure white noise. Cramer and Huggins (1958) were the first to describe a binaural stimulus that is obtained from diotic white noise by introducing a 360° phase rotation between the left- and right-ear signal within a narrow frequency band (Figure 2.6, left). A pitch is evoked corresponding to the centre frequency of this phase-rotation band up to frequencies of about 1600 Hz (as determined by pitch-discriminability). Similarly, a pitch can be evoked by the introduction of a binaural "phase edge" between two otherwise identical tokens of white noise, such that all phases above a certain edge frequency are rotated by 180°(Klein and Hartmann, 1981). The authors reported a bimodal distribution of pitch matches around the edge frequency, but some subsequent studies have found a unimodal distribution instead (Culling et al., 1998; see also Akeroyd et al., 2001). Hartmann and McMillon (2001) found that even a binaural phase *coherence* edge is sufficient to generate a pitch (Figure 2.6, middle and right) slightly above the transition frequency above which the interaural phases switch from coherent to random. Akeroyd et al. (2001) showed that all three of these binaural stimuli can convey musical melodies, with Huggins pitch yielding the highest identification scores and binaural coherence-edge pitch the lowest. Astonishingly, Bilsen (1977) demonstrated that even missing-f_0 pitch can be evoked by a Huggins-like stimulus: when listening to dichotic noise with two 360° phase-rotation frequency bands with harmonically related centre frequencies (e.g. 600 and 800 Hz), a pitch equal to their f_0 (i.e. 200 Hz) can be heard. As all these (and more) phenomena evidently require the combination of signals from both ears, the earliest possible stage in the ascending auditory pathway, at which the formation of the percept may begin is the

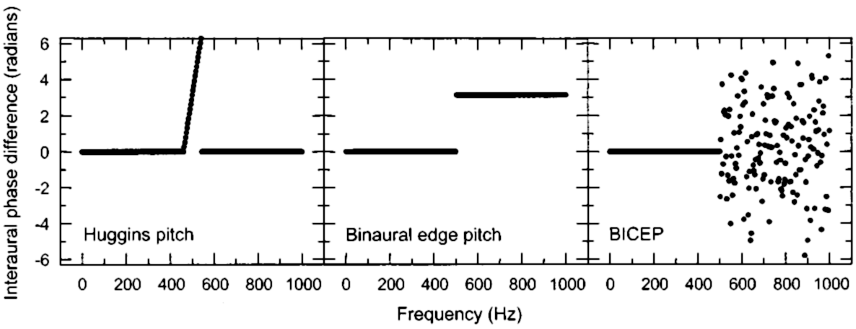


Figure 2.6: Binaural pitch-evoking stimuli. Interaural phase difference between two otherwise identical tokens of noise, presented to opposite ears, as a function of frequency: Huggins pitch (left), binaural edge pitch (middle) and binaural coherence edge pitch (right). From Akeroyd et al. (2001).

superior olivary complex in the brain stem (see section 2.3.1). The same is also true for the pitch of dichotically presented two-component HCTs (Houtsma and Goldstein, 1972) and dichotic variants of rippled noise (Bilsen, 1977). Fascinating and puzzling as these phenomena may be we will only consider monaural pitch effects when we develop our model in chapter 3.

2.2 The peripheral auditory system

2.2.1 The external ear

The external ear comprises the *pinna* and the external ear canal, or *meatus*, which is delimited medially by the ear drum, or *tympanum*, a flexible membrane that defines the boundary between external and middle ear (Figure 2.7). The *pinna* is made of soft, skin-covered cartilage tissue that is irregularly folded into a pattern of ridges and valleys. Owing to its large surface area in comparison to the diameter of the ear canal, it is well suited to amplify incoming air pressure waves prior to their mechanical transduction onto the middle and inner ear, improving our overall sensitivity to acoustic stimulation. The folds on the pinna surface cause air pressure oscillations of different frequency and direction to be differentially deflected, delayed and absorbed before entering the ear canal. Combined with the acoustic head-shadow, these effects give rise to an individually characteristic filtering pattern, referred to as the head-related transfer function (HRTF). Sounds, filtered in this manner, enter the ear canal through the *concha*, a cone-shaped depression of the pinna that acts as an acoustic funnel. In humans, the ear canal itself is approximately 2.5 cm long. Similar to an air-filled cylindrical tube with one closed end (such as a clarinet or stopped organ pipe), it is preferentially set into resonance by sound frequencies with corresponding wavelengths close to four times its own length (Gough, 2007), i.e. 10 cm in case of the (human) ear canal. Since sounds propagate through air at a speed of about 340 m s^{-1}, we can estimate the fundamental resonance frequency of the ear canal as approximately 3400 Hz. This enhanced resonance for frequencies around 3–4 kHz further shapes the overall frequency-response characteristics of the external ear, and presumably contributes to our enhanced perceptual sensitivity to frequencies in this range (e.g. Robinson and Dadson, 1956).

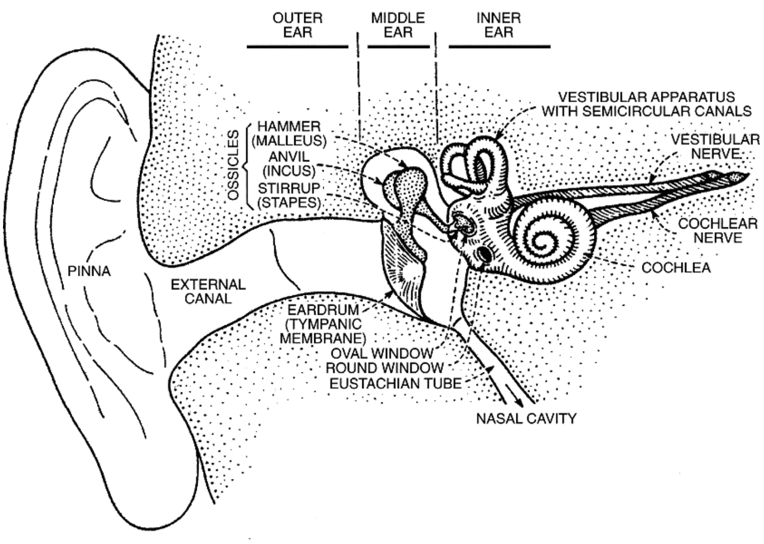


Figure 2.7: Anatomy of the human outer, middle and inner ear (from Flanagan, 1972).

2.2.2 The middle ear

The middle ear is an air-filled cavity, extending from the ear drum on the lateral side to the bony surface of the inner ear on the medial side. Its medial wall contains two membrane-covered openings, the *oval* and *round window*. Three minute ossicles – *malleus* (hammer), *incus* (anvil) and *stapes* (stirrup) – are delicately arranged as a lever system that mechanically transfers motions of the ear drum onto the oval window (and vice versa). Their purpose is to overcome the impedance mismatch between the air-filled outer and middle ear and the fluid-filled inner ear, i.e. to convert high-amplitude, low-pressure air vibrations into low-amplitude, high-pressure fluid vibrations. Akin to airborne sounds being largely reflected by a water surface, air vibrations in the inner ear would fail to transfer their energy onto the inner ear fluid if the two compartments were simply separated by a flexible membrane. Two factors contribute to the impedance-matching process performed by the middle ear ossicles. Firstly, their lever action reduces the motion amplitude by a factor of 1.3 in humans (Hemilä et al., 1995). Secondly, and more importantly, the area of the human ear drum is about 23-fold larger than that of the oval window (ibid.). Thus, the expected effective pressure gain at the oval window for an ideal transformer is close to 30-fold (approximately 30 dB). In reality however, the gain is generally lower and frequency-dependent, reaching a maximum of 20 dB for frequencies around 1 kHz (as measured in human cadavers). The ossicles are held in place by ligaments and the effectiveness of the entire lever mechanism is influenced by two muscles: the *stapedius* and *tensor tympani* which act to dynamically regulate the middle ear gain. During vocal production, jaw movement and in reaction to loud sound (exceeding 80 dB SPL), the *stapedius reflex* causes an involuntary muscle contraction that reduces the middle ear gain by up to 20 dB (Zakrisson, 1979). Effective mechanical transmission also requires the air pressure on both sides of the ear drum to be identical in the absence a sound. The *Eustachian tube*, a canal linking the middle ear to the pharynx, serves this purpose by equalising the barometric pressure in the middle ear cavity and the outside atmosphere (see Figure 2.7).

2.2.3 The inner ear

2.2.3.1 Anatomy

The inner ear is a complex bony structure containing both the vestibular system and the *cochlea* – the actual site of mechano-neural transduction of sound. The cochlea (lat. "snail") is a fluid-filled, helically coiled tube with an unrolled length of approximately 3.5cm in humans. Encased in a hard bone shell, the membranes of the oval and round window are the only flexible parts of the cochlear surface. Inside, the tube is tripartite along its entire length, containing three ducts separated by flexible membranes: the central *scala media* (or *cochlear duct*) including the actual cochlear sensory epithelium, surrounded by the *scala vestibuli* and *scala tympani* (see Figure 2.8). The scala media is separated from the scala vestibuli by *Reissner's membrane*, and separated from the scala tympani by the *basilar membrane*. Towards the tip (*apex*) of the cochlea, there is a direct connection, the *helicotrema*, between the scalae vestibuli and tympani, so that the two effectively form one long compartment that folds back onto itself, with the scala media enclosed in-between (Figure 2.8A). An important property of the basilar membrane with respect to its passive, mechanical response to stapes motion is its stiffness: it decreases, roughly exponentially, from the base to the apex by several orders of magnitude (von Békésy, 1960). While the scalae vestibuli and tympani are filled with *perilymph*, a fluid with a low concentration of potassium (K^+) ions, the scala media is filled with high-potassium *endolymph*. The ionic concentration gradient between these two fluids is a crucial driving-force for the active response of the cochlea under physiological conditions, as will be further detailed in section 2.2.3.2.

The sensory epithelium (*organ of Corti*) of the inner ear is located on the surface of the basilar membrane, containing some 15000 mechanoreceptor cells: 3000 *inner* (IHC) and 12000 *outer hair cells* (OHC). Attached to their cell bodies are *stereocilia*, fine bundles of filament to which the hair cells owe their name. Above their tips rests a glutinous sheet, the *tectorial membrane*. Displacement of the basilar membrane leads to a shearing between the organ of Corti and the tectorial membrane that deflects the hair bundles from their resting position. This causes either a de- or hyperpolarisation of the hair cells, depending on the direction of motion: deflections towards the lateral wall of the cochlear coil cause depolarisation, deflection towards the centre result in hyperpolarisation. Inner and outer hair cells are distinguished by their anatomical location, their physiological properties and their innervation patterns. A single row of IHCs extends along the central side (with respect to the coil) of the organ of Corti. Despite being fewer in numbers, IHCs are the major source of afferent signals to the central auditory system. They receive afferent innervation from so-called *type 1* fibres – dendritic neural processes of up to 30 cells situated in the *spiral ganglion* (Moser et al., 2006). The spiral ganglion is an aggregation of nerve cells that extends along the central axis of the cochlear helix. Axons of the spiral ganglion cells (SGCs) form the auditory nerve which projects onto neurons in the *cochlear nucleus* of the auditory brainstem. Depolarisation of an IHC following deflection of the stereocilia causes the initiation of action potentials in SGC dendrites. IHC afferent signals are modulated by the efferent fibres of the *lateral olivocochlear bundle* (LOC), which originate from the lateral *superior olivary complex* (SOC) in the ipsilateral midbrain and terminate directly onto SGC type 1 dendrites.

The outer hair cells are arranged in three rows along the lateral side of the organ of Corti. A key physiological property of the OHCs is their electromotility, owing to the

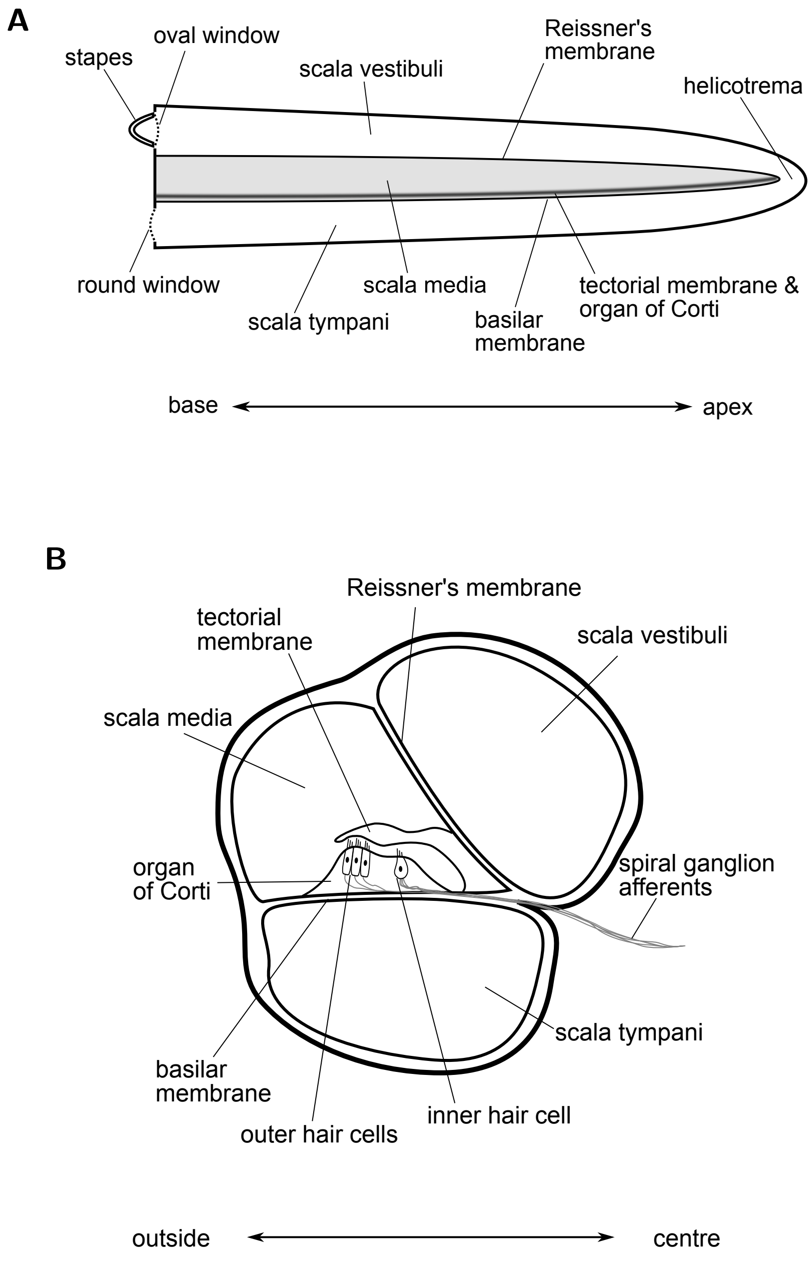


Figure 2.8: Schematic anatomy of an unrolled cochlea (after von Békésy, 1960). A: Longitudinal section. B: Radial section.

presence of the motor protein *prestin* in the plasma membrane (Brownell et al., 1985; Zheng et al., 2000): de- and hyperpolarisation of the OHC soma causes the cell body to rapidly stretch and contract, respectively, thereby interfering actively with the fluid-driven motions of the basilar membrane. OHCs receive afferent innervation from *type 2* SGC fibres. Despite the large number of OHCs, type 2 fibres constitute only 5% of all SGC afferents and their functional role within the ascending auditory system is not well understood. OHCs receive strong efferent innervation from fibers of the *medial olivocochlear bundle* (MOC), originating largely from midbrain nuclei surrounding the contralateral medial SOC and terminating onto the OHC bodies.

2.2.3.2 Function

As the footplate of the stapes, covering the oval window, moves inwards and outwards, the cochlear fluids themselves are set into motion. Since the fluids are almost incompressible and the walls of the labyrinth are rigid, an inward motion of the stapes must ultimately be compensated by an outward movement of the flexible membrane covering the round window at the base of the scala tympani, and vice versa. As part of this process, the membranes of the cochlear duct are being deflected up and down in a way that has led to the common notion of the basilar membrane as a spatial frequency analyser or "acoustic prism" (Zweig, 1976). When the stapes movement is sinusoidal, a travelling wave forms on the basilar membrane, starting at the base and propagating towards the apex. Initially, the propagation speed and wavelength are high, but as the membrane stiffness decreases away from the base, wavelength and speed drop while the wave amplitude builds up, until the wave reaches a critical point along the membrane. Beyond this point, frictional losses grow quickly as the wavelength decreases further, causing the wave to subside (see Figure 2.9). The point of maximal amplitude is determined by the frequency of the mechanical stimulus at the oval window, which sets the initial wavelength of the travelling wave. High-frequency stimulation generates a wave that peaks close to the base, whereas waves elicited by increasingly lower frequencies peak at increasingly apical positions on the basilar membrane (Figure 2.10). Conversely, each site along the BM has its own *characteristic frequency* (CF) for which its sensitivity is greatest. CFs decrease from the base to the apex.

Due to a shearing motion between the basilar and tectorial membrane, the stereocilia of the hair cells are alternately deflected towards and away from the centre of the cochlear

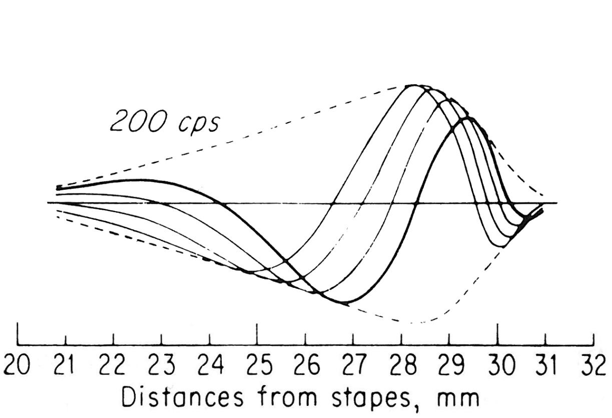


Figure 2.9: Cochlear travelling wave in response to a 200 Hz sinusoidal displacement of the stapes, depicted at four different stages of the phase cycle. From von Békésy (1960, chap. 12).

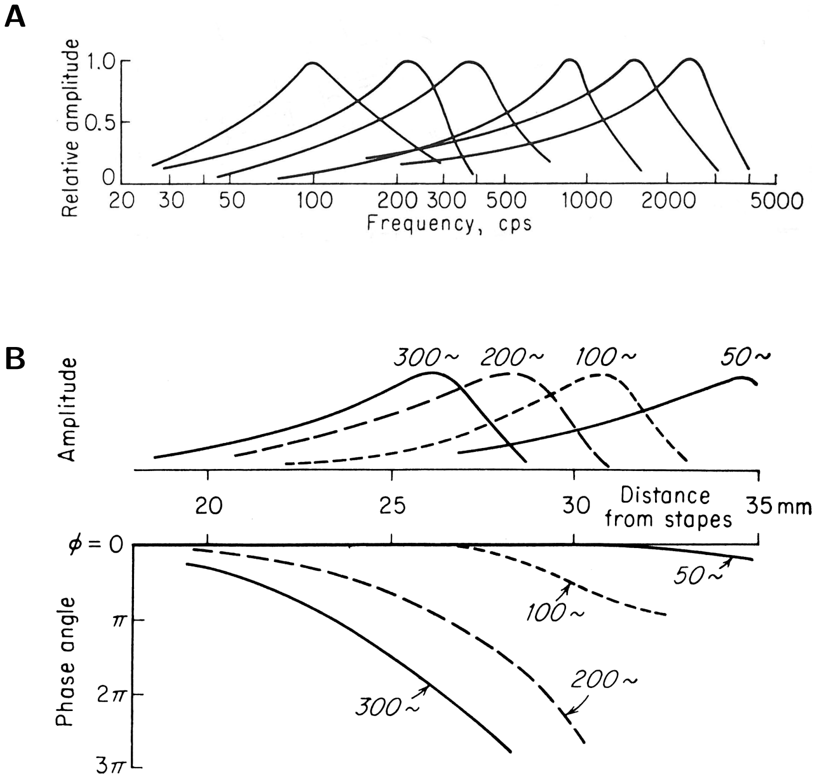


Figure 2.10: Mechanical frequency tuning of the cochlea. A: Vibration amplitude measured at six points on the basilar membrane (different curves), as a function of mechanical stimulation frequency. B: Vibration amplitude (top) and phase (bottom), measured along the length of the basilar membrane in response to stapes vibration at rates of 50, 100, 200 and 300 Hz. From (von Békésy, 1960, chap. 11).

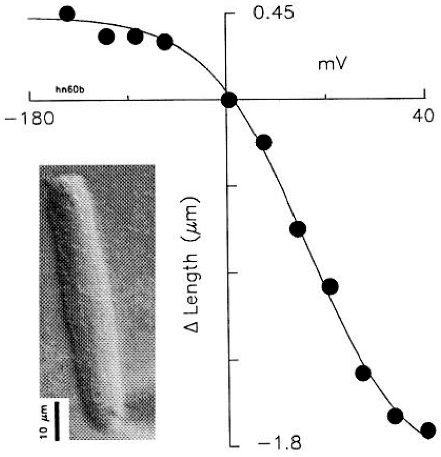


Figure 2.11: Electromotility of an outer hair cell: change in OHC length in response to voltage steps from a holding potential of -68.4 mV(from Santos-Sacchi, 1992)

coil, either by way of direct contact between their tips and the tectorial membrane or, for shorter bundles, through the viscosity of the surrounding medium. Deflections towards the centre open mechanically gated ion channels that enable the influx of potassium (K^+) ions from the K^+-rich endolymph of the scala media into the hair cell soma. During opposite deflection, these channels close, while at the same time, K^+ ions are excreted from the hair cell into the low-K^+ perilymphatic fluid of the scala tympani through voltage-gated K^+ channels. This results in an alternating de- and repolarisation of the hair cell.

IHCs and OHCs react differently to somatic depolarisation. In OHCs, the somatic potential induces a mechanical response of prestin molecules in the cell membrane: depolarisation results in a longitudinal contraction, hyperpolarisation in an elongation of the OHC body (see figure 2.11). The consequences of this electromechanical feedback in shaping the overall basilar membrane response will be discussed in further detail in section 2.2.3.3. In IHCs, the initial depolarisation effects the further, voltage-gated influx of calcium (Ca^{2+}) ions from the scala media endolymph. Increased Ca^{2+} triggers the calcium-dependent release of glutamate (and possibly other neurotransmitters) into the cleft between IHC and the afferent synapses of type 1 SGC fibres. Excitatory post-synaptic potentials (EPSCs) in the SGC dendrites are mediated by AMPA receptors and may result in the generation of action potentials that propagate along the SGC axons to the cochlear nucleus of the brainstem.

The influence of the efferent system (comprising the MOC and LOC bundles, cf. section

2.2.3.1) on the mechanical and neural response to sounds is far from fully understood, but experimental evidence points towards an inhibitory role. BM displacement amplitudes were found to be reduced around and above CF during electrical stimulation of efferent MOC fibres synapsing onto OHCs (Russell and Murugasu, 1997). Reductions were strongest for medium-intensity tones, qualitatively consistent with earlier observation made in the auditory nerve (Guinan and Stankovic, 1996). These effects are presumably mediated by a hyperpolarisation of the OHCs, thereby inhibiting their electromotility. The workings of the LOC efferent system is barely known at all, owing in large parts to the difficulty of selectively stimulating or recording from the fibres. The functional significance of the entire efferent system as a whole similarly remains speculative. Proposed benefits include dynamic range control, reduction of masking in noisy acoustic backgrounds and the protection of hair cells from cytotoxic acoustic overstimulation (Guinan, 2006).

On the whole, mechanical and neural tuning properties at any given point along the BM are closely matched. An important functional property that distinguishes mechanical from neural responses is the half-wave rectification that the IHCs effectively perform on the BM response. The IHC potential is depolarised only during half a phase of a BM oscillation. During the other half, it remains close to the cell's resting potential. Due to the slowness of the ion-channel dynamics, the IHC potential gradually ceases to follow every single peak of the BM oscillation, as oscillation rates increase above 1kHz. Instead, for high oscillation rates, the potential fluctuates only little around an elevated baseline. Thus, the IHC response approximately follows the (half-wave rectified) BM displacement waveform for slow oscillation rates, and its amplitude envelope for fast rates, performing a simple form of envelope demodulation (see Figure 2.12).

2.2.3.3 Non-linearities in the basilar membrane

Prior to the groundbreaking studies of von Békésy during the 1930s to 50s (von Békésy, 1960), the leading theory of cochlear mechanics had assumed that the stereocilia themselves acted as independent resonators tuned to different frequencies, similar to the strings of a harp or undamped piano (von Helmholtz, 1863). While a similar mechanism may indeed be found in the cochleae of some reptiles (e.g. Holton and Weiss, 1983), von Békésy experimentally established the now widely accepted travelling-wave mechanism as a major determinant of cochlear mechanics in mammals. However, work-

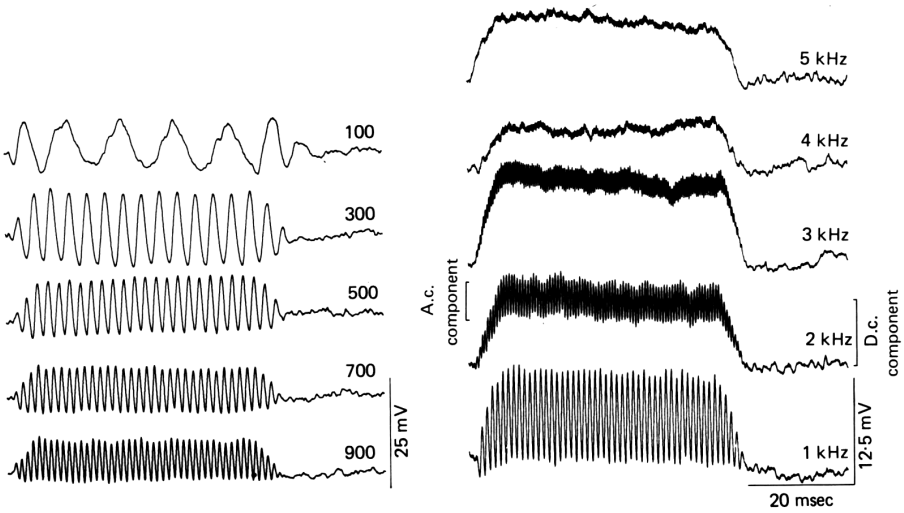


Figure 2.12: IHC receptor potentials in response to tones of different frequencies presented at 80 dB SPL, measured at the basal turn of a guinea pig cochlea (from Russell and Sellick, 1983).

ing solely on cadaver ears, he did not realise the significant role of the OHCs in further shaping the cochlear response to acoustic stimulation under physiological conditions (Rhode, 1971; Brownell et al., 1985). As the OHC somata are de- and hyperpolarised following stereocilia movement, prestin molecules in the cell membrane contract and expand, rapidly enough to follow frequencies far beyond the human range of hearing when electrically stimulated in vitro (Frank et al., 1999). This electromotile response acts to locally amplify the passive, fluid-driven vibration of the basilar membrane and thereby provides a positive feedback loop that can greatly enhance the amplitude of basilar membrane motion, particularly in the low-amplitude regime. As a consequence, the mechanical response of the active cochlea to sounds is highly non-linear. This can be seen, for example, in the intensity-dependence of cochlear tuning characteristics, when the intensity of acoustic stimuli is lowered from a moderately high level (e.g. 80 dB SPL) down to the response threshold while measuring the response amplitude at a fixed point on the basilar membrane across a range of stimulus frequencies. As the stimulus level decreases, the frequency eliciting the maximal response shifts upwards, while the frequency selectivity (or sharpness of tuning) around this frequency increases. Furthermore, the response amplitude for low-intensity stimuli at and near the characteristic frequency of a given site is amplified by as much as 50 dB above a linear prediction from high-intensity stimuli (Johnstone et al., 1986). Comparisons have been made between physiologically-intact preparations and those that were either dead (Rhode, 1973), pharmacologically lesioned (Ruggero and Rich, 1991) or lesioned by acoustic overstimulation (Ruggero et al., 1993). From these, it is clear that amplification effects are indeed due to active processes in the cochlea involving OHC electromotility. Further non-linear effects beyond the level-dependence of cochlear frequency selectivity have been linked to the active electromechanical OHC feedback, most prominently *compression* of responses to near-CF stimuli, *two-tone suppression* and *intermodulation distortions.*

Compression is a phenomenon observed predominantly at the basal (i.e. high frequency) end of the cochlea in many mammalian species, whereby the mechanical response to a tone near the CF of a given site increases sub-linearly (i.e. with a slope of less than 1 dB response amplitude per dB stimulus amplitude) as the stimulus level increases above 20 to 25 dB SPL. Figure 2.14 shows the velocity-intensity functions of BM responses to tones in the basal cochlea of a chinchilla. As can be seen, compression is confined to frequencies near CF (10 kHz). In several cases, response growth has been found to return to linear at high stimulus levels of 100 dB SPLand above, as seen in the responses to 9 kHz and 11 kHz tones. Since responses to off-CF tones (e.g. 4-7 kHz in Figure 2.14) continue to grow linearly at amplitudes where near-CF responses of equal amplitude are highly compressive, compression does not seem to reflect a mere ceiling effect of BM responses at high amplitudes. Compression is more pronounced at the base of the cochlea than near the apex, and is greatly diminished following trauma or death (Ruggero et al., 1996).

In two-tone suppression, the BM response to a tone (the probe) is reduced in the presence of another (the suppressor). Similar to amplification and compression, two-tone suppression is most pronounced for probe tones near the CF of a site, presented at low to medium intensities (Figure 2.15) and is reduced by hair cell trauma or death of the animal.

Intermodulation distortions are frequency components in the BM response to two (or more) tones which are not themselves contained in the stimulus and occur as a natural consequence of non-linearities in any system. In general, the response of a non-linear system to a signal consisting of two sinusoids with frequencies f_1 and $f_2, f_2 > f_1$ (*primaries*) may contain frequencies corresponding to arbitrary integer combinations $mf_1 + nf_2 : m, n \in \mathbb{Z}$ of the primaries (Figure 2.16). These distortions can sometimes be heard by the listener, most prominently so the *quadratic difference tone* (QDT) $f_2 - f_1$ and the *cubic difference tone* (CDT) $2f_1 - f_2$: their occurrence was described in

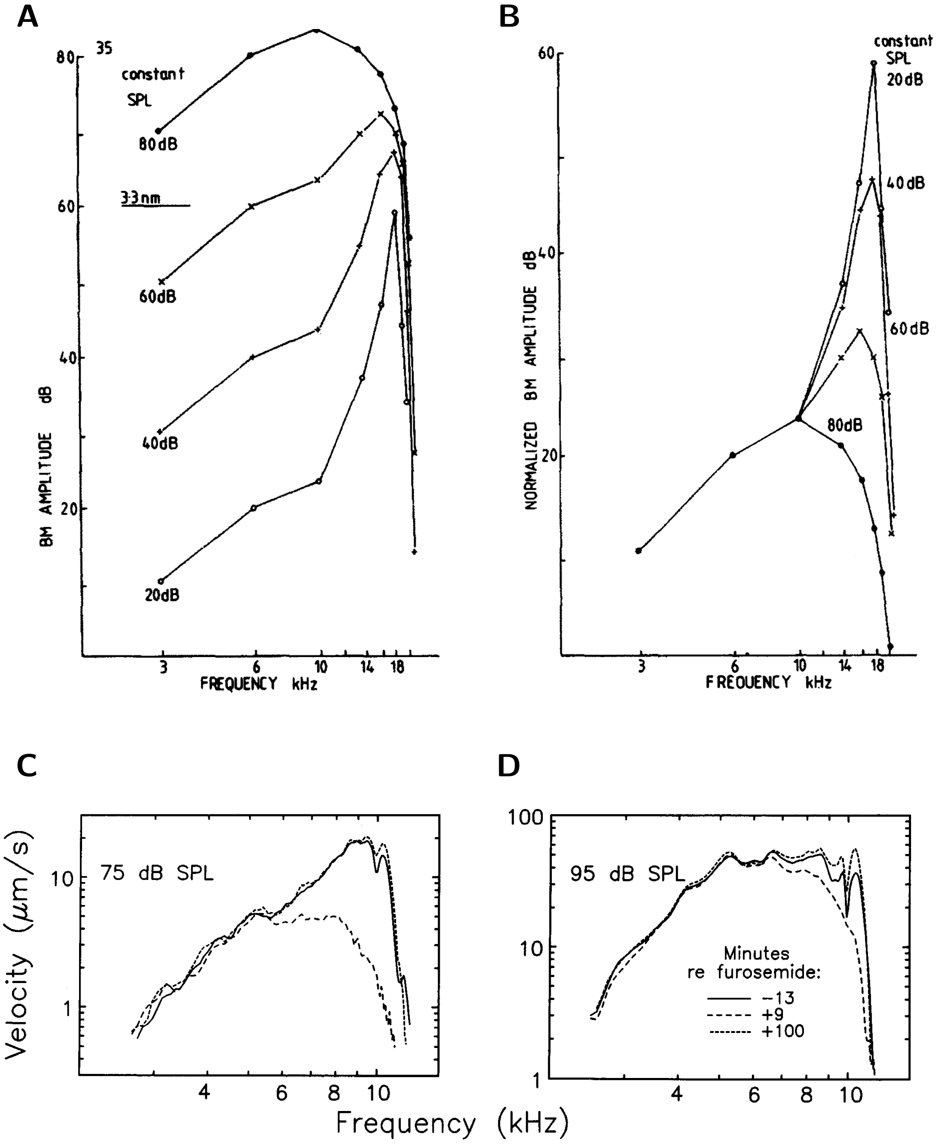


Figure 2.13: Active amplification of BM responses. A & B: Amplitude of the basal BM response in a guinea pig as a function of tone level and frequency on an absolute scale (A) and normalised by stimulus level (B), demonstrating super-linear response magnitude for low-intensity tones near CF. C & D : Level-specific effects of furosemide injection on BM responses at the basal turn of a chinchilla cochlea for stimuli presented at 75 dB (C) and 95 dB (D). Adapted from Ruggero and Rich (1991).

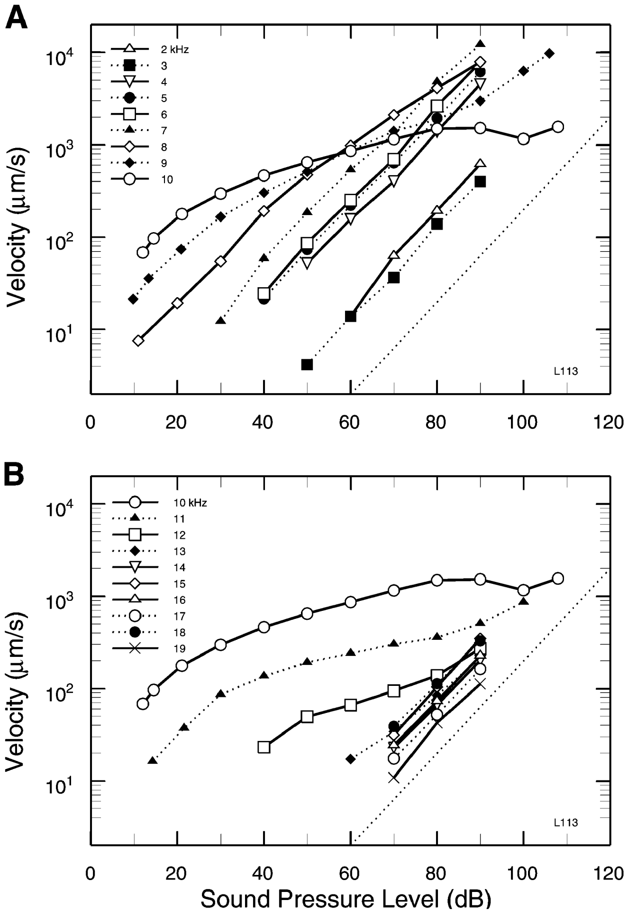


Figure 2.14: Velocity-intensity functions of BM responses to tones in the basal portion of a chinchilla cochlea, demonstrating compression near CF (10 kHz). A: Responses to tones at and below CF. B: Responses to tones at and above CF. From Ruggero et al. (1997).

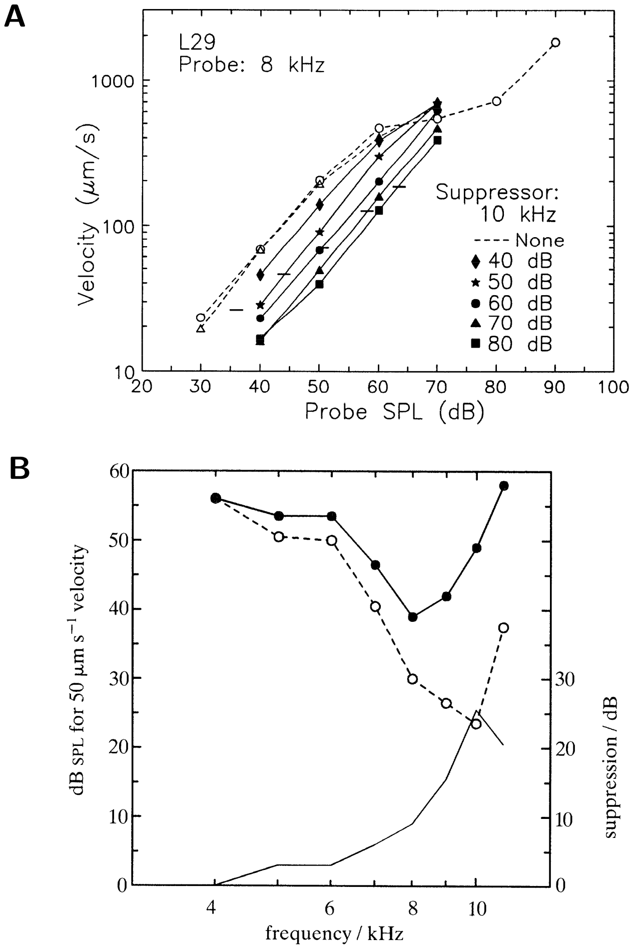


Figure 2.15: Examples of two-tone suppression. A: Velocity-intensity functions for probes and suppressors at different levels recorded at a basal site in a chinchilla, demonstrating highest suppression for low probe intensities (symbols denote different suppressor levels; from Ruggero et al., 1992a). B: Effect of a 12kHz suppressor at 63 dB on the response to probes at different frequencies, recorded at a basal site in a chinchilla. Open circles represent iso-velocity curves for the probe alone, filled circles for probe and suppressor. Suppression magnitude is indicated by the thin solid line, demonstrating CF specificity (from Ruggero et al., 1992b).

the early 18th century by Italian composer and violinist Giuseppe Tartini, and has long been used as a test for correct intonation by music practitioners (Mozart, 1756). The amplitude of the CDT on the BM depends on the absolute and relative amplitudes of the primaries as well as their frequency separation, with larger separations giving rise to lower distortion amplitudes (Cooper and Rhode, 1997). CDT amplitude, like the amplitudes of other intermodulation distortions, is highest at sites tuned to the distortion product frequency and can reach levels of up to -15 dB relative to the primaries (Robles et al., 1997). Furthermore, it is vulnerable to localised acoustic trauma at both the target and primary sites. This suggests that the CDT is generated at a site close to the primaries and subsequently propagated along the BM, where it is locally amplified at an appropriately tuned site (ibid.). Of particular relevance in the context of pitch perception is the *quadratic difference tone* (QDT), occurring at a frequency of $f_2 - f_1$. It is audible only at high stimulus levels and more challenging to measure on the BM, owing to the difficulty of maintaining a physiologically intact preparation while recording from apical (i.e. low-frequency) sites of the cochlea. Nevertheless, its existence has been clearly demonstrated (Figure 2.16B), with amplitudes as high as 23 dB below the primaries (Cooper and Rhode, 1997). Its relevance to pitch perception derives from the simple observation that in a harmonic complex sound with fundamental frequency f_0, any neighbouring pair of harmonics nf_0 and $(n+1)f_0$ gives rise to a QDT at f_0. Since the perceived pitch of such a sound coincides with the fundamental, the question arises whether the pitch of these and other sounds may reflect the immediate sensation of mechanical distortions on the BM rather than the result of higher-level signal processing. This issue will be further discussed in section 2.4.

Peripheral non-linearities, as described above, may also play an important role in psychophysical effects such as simultaneous masking (the elevation of tone detection thresholds in the presence of masking tones or noises, see e.g. Delgutte, 1990) or perceptual measures of frequency selectivity (Rosen et al., 1998). Such links, however, are very difficult to establish and remain tentative for the time being, largely due to a lack of human physiological data. As pointed out, the most important behavioural analogue from the perspective of pitch perception are combination tones perceived by listeners, which effectively augment the spectrum of the acoustic stimulus.

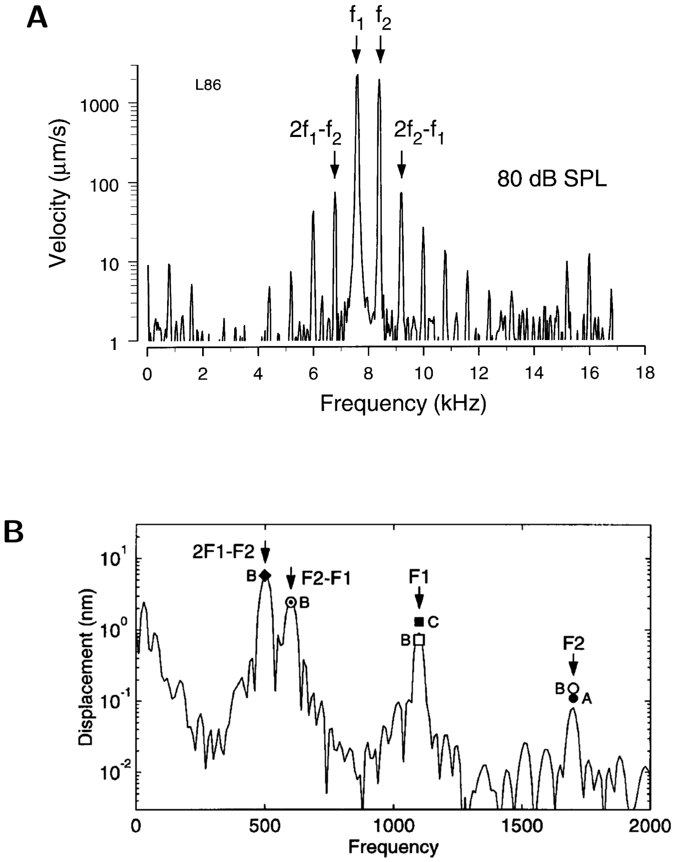


Figure 2.16: Intermodulation distortions. A: Frequency spectrum of the BM response to equal-amplitude tone pairs at the basal turn of a chinchilla cochlea (from Robles et al., 1991). B: Frequency spectrum of the BM response to equal-amplitude tone pairs at the apical turn of a chinchilla cochlea (from Cooper and Rhode, 1997).

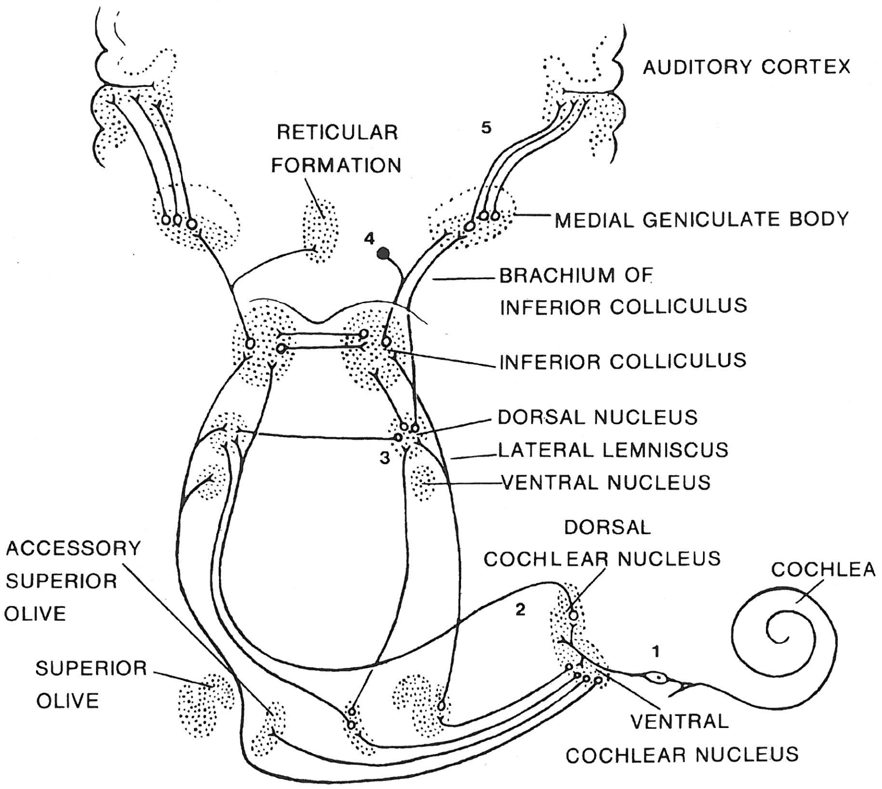


Figure 2.17: Major structures of the ascending mammalian central auditory pathway (from Gacek, 1972).

2.3 Processing of pitch in the central auditory system

In this section, we will give a brief overview of the neural representation of stimulus periodicity and — possibly — pitch along the ascending auditory pathway, covering the brainstem, midbrain and auditory cortex (see Figure 2.17). It is generally believed that neurons in the brainstem and midbrain represent selected physical features of the sound, all of them likely to contribute to our percept of pitch, but none of them by themselves a general representation across the entire range of pitch-evoking sounds. Physiological data is necessarily restricted to non-human species, making direct comparisons impossible. In the cortex, non-invasive physiological measurements are available from humans. In our review of the cortical substrate of pitch, we will limit ourselves to these, in addition to electrophysiological data from primates, whose auditory cortical architecture appears to be closely homologous to humans. A recent, more general overview including a wealth of recent findings made in ferrets (e.g. Nelken et al., 2008; Bizley et al., 2009, 2010) can be found in Walker et al. (2011).

2.3.1 Brainstem and midbrain

In the auditory nerve, two types of representations of sound periodicity have been observed and studied: one is based on a tonotopic rate-place representation of the stimulus spectrum, the other one on the timing of neural events in individual peripheral frequency channels. Spectral peaks in the stimulus give rise to peaks in the response-amplitude profile of the basilar membrane (BM) along its tonotopic axis, if their frequencies are sufficiently separated. As the BM amplitude is a major determinant of the inner-hair-cell (IHC) potential and auditory nerve (AN) firing response, peaks in the BM amplitude profile are preserved in the average firing-rate profile of the auditory nerve fibres (if one orders them by their characteristic frequencies). For a harmonic sound, the separation between peaks is equal to its fundamental frequency, which may serve as the basis for its determination. However, as the frequency bandwidth of the BM increases roughly linearly with characteristic frequency (CF), such peaks are only discernible for low-rank harmonics up to approximately the 10th[14]. In addition to this coarse rate-place representation — limited to resolved, approximately harmonic spectra — periodicity information is also contained in the temporal pattern of activity in single auditory nerve fibres (or groups of them). Activity in a single auditory nerve fibre is approximately phase-locked to peaks in the stimulus waveform, band-passed within a frequency range determined by the place of the innervating IHCs along the BM axis. Hence, stimulus periodicity within this frequency range is preserved in the periodicity of the evoked AN response. Cariani and Delgutte (1996a,b) measured auditory nerve responses in cats to a large variety of (human-)pitch-evoking sounds: artificial vowels, sinusoidally amplitude-modulated (SAM) tones, SAM noise and click trains. They computed first-order and all-order inter-spike-interval (ISI) histograms, pooled across many fibres (the latter being equivalent to the fibres' autocorrelation function), and found that the peak of the all-order ISI histogram provided an excellent match with the "typical" human pitch percept in all cases, independent of the stimulus level[15]. Furthermore, sounds typically associated with salient pitch percepts produced larger peaks in the ISI histogram than weakly-pitched sounds. This form of representation is expected to deteriorate as the periodicity rates approach the phase-locking limit

[14] The exact boundary between resolved and unresolved harmonics is a matter of definition and debate, and depends to some degree on f_0 (see e.g. Moore and Gockel, 2011 for a recent summary).

[15] Histogram peaks based on the first-order ISI were less robust in comparison, e.g. jumping between fundamental and formant frequencies depending on sound level

of the AN fibres. Cedolin and Delgutte (2010) investigated the periodicity-limit of the representation, again in cats, using complex harmonic sounds as stimuli. They found a sharp increase in estimation errors ($>$ 10% deviation) as the fundametal frequencies exceeded approximately 1300 Hz (despite the fact that AN fibres in cats have been shown to phase-lock almost perfectly up to 2 kHz, and to some degree up to 4.5 kHz (Johnson, 1980)). Conversely, the estimation of the fundamental from average-rate profiles was possible from 400-500 Hz *upwards* (up to 3.5 kHz, the highest f_0 used). These results demonstrate that neither all-order ISIs nor rate-place profiles alone seem to carry sufficient information to support human pitch perception across its entire pitch-height range (if one accepts the cat AN as a legitimate proxy).

Auditory nerve fibres synapse onto neurons in the *cochlear nucleus* (CN). The CN can be divided into three sub-nuclei (dorsal, anteroventral and posteroventral), each of them tonotopically innervated and organised (e.g Malmierca and Hacket, 2010). Out of the many different cell types in the cochlear nucleus, some have gathered considerable interest with regard to their role in periodicity processing (see e.g. Winter, 2005). Bushy cells, or "primary-like" neurons, in the ventral CN have response characteristics much like their AN inputs: they phase-lock to stimulus peaks up to high frequencies, so that information about the fine-structure periodicity is preserved for potential read-out by their postsynaptic targets (Winter et al., 2001), for example in the form of all-order ISI histogram as discussed above. So-called "chopper neurons" in the ventral CN show a sharp modulatory pattern in their firing rates, at cell-intrinsic modulation rates between about 100 to 500 Hz in cats (Kim et al., 1990). Within this range, modulation depth in the response is enhanced when the stimulus periodicity coincides with the intrinsic chopping rate. This has led to the hypothesis that chopper neurons may help to transform the all-order ISI representation of periodicity in the AN into a "temporal place representation", which would be more convenient to read out subsequently (Kim et al., 1990; Frisina et al., 1990; Winter et al., 2001; Wiegrebe and Meddis, 2004): given a linear array of chopper neurons, spanning the entire range of natural chopping rates for *each* peripheral frequency channel, stimulus periodicity would manifest itself as a peak in the modulation-depth profile along this array that is consistent across channels. The existence of such an arrangement of cells remains speculative however, and the range of periodicities that can be represented in this fashion would be strictly limited by the range of intrinsic chopping frequencies. Aside for their hypothesised role in transformation of temporal periodicity cues, Blackburn and Sachs (1990) observed that the rate-place profile in chopper neurons along the tonotopic gradient was considerably less dependent on sound level than that of primary-like cells (and their AN inputs), thus providing an easy-to-read spectral representation of the stimulus. Lateral suppression further acts to sharpen the rate-place profile (Rhode and Greenberg, 1994). "Octopus cells" in the posteroventral CN receive convergent input across a wide range of frequencies. While their frequency tuning is therefore less specific than in other CN cell types, their temporal precision and tendency to phase-lock to amplitude modulations in the stimulus is conversely much higher (Rhode, 1994; Oertel et al., 2000). Overall, the CN preserves both spectral and temporal fine-structure information in its varied response patterns. More than just a passive relay, however, some cell types seem to facilitate the further use of one of these cues or the other for downstream processing stages.

Following the ascending pathway leads us to the superior olivary complex (SOC), the first stage of binaural convergence. Its primary functional role in hearing is typically considered to be sound localisation (see e.g. Grothe et al., 2010). Two major nuclei, the lateral and medial superior olive (LSO and MSO), are highly sensitive to interaural differences in sound intensity and timing respectively. Within the ascending pathway, the SOC seems to have gathered only little attention with regard to its role in periodicity processing and pitch perception. However, due to the specialised function of the MSO in the processing of interaural time and phase, it may be involved in the generation of binaural pitch percepts such as the Huggins pitch (however weak and rare the phenomenon may be; see section 2.1.3). As discussed in section 2.2.3.1, the SOC is the origin of efferent connections to the inner and outer hair cells.

Neurons in the central nucleus of the inferior colliculus (ICc) receive innervation from both the SOC (directly and indirectly via the lateral lemniscus) and the CN. Again, neurons are ordered along a tonotopic gradient: from low frequencies dorsally to high frequencies ventrally. Interestingly from our point of view, IC neurons respond in a band-pass manner not only to pure tones frequency: Langner and Schreiner (1988) demonstrated band-pass tuning also to the *modulation* frequency of SAM tones for modulation rates up to 1 kHz in anaesthetised cats (where a neuron's CF was used as carrier frequency). Unlike in CN chopper neurons, this manifests itself not only in the modulation depth of the IC responses, but also in their mean firing rates. Furthermore, Schreiner and Langner (1988) reported that the modulation-tuning characteristics followed a gradient *orthogonal* to the dorsal-ventral tonotopic axis, which they called a "periodotopic map". Similar results were also obtained from awake chinchillas (Langner et al., 2002). The physiological mechanisms that underlie periodicity tuning in the IC are not well understood. McAlpine (2004) showed that quadratic distortions (cf. section 2.2.3.3) on the basilar membrane can evoke neural responses in the IC when the *modulation* rate of an amplitude-modulated tone matches the CF of a neuron (while the carrier frequency is consequently much higher). As cells in Langner and Schreiner (1988) and Langner et al. (2002) had best modulation frequencies (BMFs) typically far below their CFs (which were used as *carrier* frequencies), the results cited above are not readily explained as tonotopic responses to quadratic cochlear distortions[16]. Hewitt and Meddis (1994) proposed a mechanism, by which IC neurons act as coincidence detectors of synchronised spikes in their input, originating from a population of CN chopper units with identical intrinsic chopping rates. High modulation gain in the output of the chopper neurons (see above) is accompanied by increased spike synchronicity between neurons, which drives the IC coincidence detector to higher firing-rates when stimulus modulation rate matches the CN chopping rate. However appealing, there is little direct evidence for such a mechanism and the model cannot readily explain a discrepancy between the range of chopping rates in the CN and BMFs in the IC found in vivo (see Krishna and Semple, 2000 for a thorough discussion). Independent of the underlying mechanism, it remains an open question to what degree (or whether at all) the periodicity map in the IC is related to pitch perception. The simplicity of the read-out based on average firing rate, without the need for high temporal resolution makes it an attractive candidate basis for periodicity estimation in the thalamus or cortex, where neurons become increasingly insensitive to fast-rate fluctuations in their inputs (Wallace et al., 2002, 2007). Furthermore, given the degree of convergence in the IC, the observed periodicity-tuned responses themselves could reflect a sophisticated, combined estimate based on the entire variety of spectral, temporal and binaural cues available. However, BMFs in the IC extend well below the lower limit of pitch (30-40 Hz) on the one hand, while falling well short of its upper limit (4-5 kHz) on the other hand. In addition, modulation frequency, periodicity and pitch are not one and and the same thing. Thus, the generality and robustness of this putative periodotopic representation across a much larger variety of sounds remains to be tested.

[16] Reports of modulation tuning to AM tones with high carrier frequencies, far outside the neuron's pure tone response range (Biebel and Langner, 2002), however, must be treated with caution for this reason.

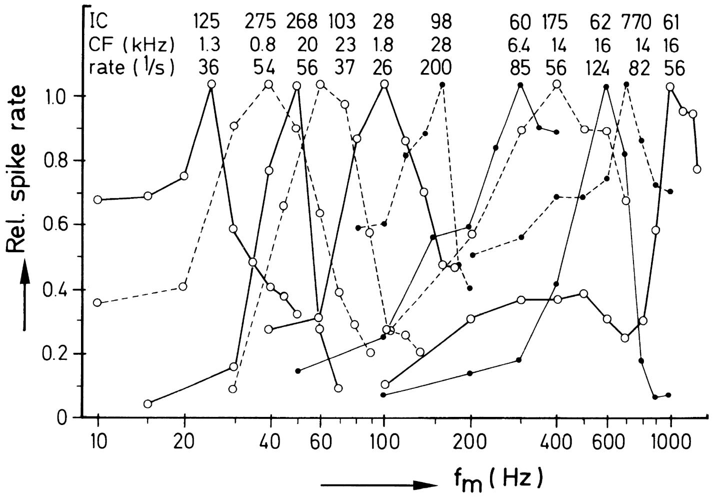


Figure 2.18: Modulation frequency tuning in the central nucleus of the inferior colliculus in cats (open circles: single units, filled circles: multi-unit clusters). Peak firing rate and unit CF are indicated above the tip of each curve (from Langner and Schreiner, 1988).

The medial geniculate body (MGB) in the thalamus is the target of afferent IC fibres. Of its three subdivisions (the ventral, medial and dorsal MGB), only the ventral is tonotopically organised (see e.g. Anderson et al., 2007). Preuss and Müller-Preuss (1990) demonstrated the existence of band-pass tuning to modulation rates of amplitude-modulated tones. As with modulation-rate tuning in the IC, it is currently unknown to what degree these responses in the MGB encode a particular physical feature of the sound waveform and to what degree they reflect the subjective perceptual experience of the subject, even in cases when the two may be at odds. Bartlett and Wang (2007) recently studied the responses of MGB neurons in awake marmoset monkeys using regular click trains. Their main finding, which qualitatively differentiated MGB from IC responses, was that a substantial fraction of MGB neurons with band-pass tuning for click rate responded with wholly stimulus-desynchronised firing patterns. The functional significance of these findings, however, remains yet to be determined (see e.g. Wang et al., 2008 for a review and discussion).

2.3.2 Cortex

The auditory cortex is a collection of morphologically and physiologically heterogeneous subfields, located on the temporal lobe in either brain hemisphere. While the architecture of the subcortical ascending pathway appears to be largely preserved across many mammalian species, the organisation of auditory cortex is more heavily species-dependent (see e.g. Malmierca and Hacket, 2010). One widely shared architectural feature is the distinction between primary and primary-like "core" areas, that are surrounded by several secondary "belt" areas. The core areas receive ascending, thalamic innervation primarily from the tonotopically organised ventral MGB and typically show some form of tonotopy themselves. The non-primary belt areas receive inputs also from the tonotopically unstructured dorsal and medial MGB, as well as strong intracortical input from the core areas. Multiple core regions can be distinguished physiologically in many species by a reversal of the tonotopic gradient at their mutual boundaries. In humans, the surface of Heschl's gyrus (HG) is the anatomical site of the core areas of the auditory cortex (e.g. Hackett et al., 2001). Aside from approximate tonotopic gradients of preferred spectral tuning observed across species, the functional characterisation of auditory cortical neurons remains a largely unsolved problem. A major obstacle in this endeavour is the seeming lability of cortical response properties in the face of changes in the long-term acoustic or even behavioural context in which stimuli are presented (e.g. Ulanovsky et al., 2003; Fritz et al., 2003, 2005). Hence, it would appear that auditory cortex, rather than representing physical stimulus features itself, might be a key stage in constructing the mental representation of the subjectively-experienced "auditory scene" from physical-like features represented in the midbrain. If that were the case, auditory cortex (primary or secondary) would be a promising candidate area to contain the neural substrate of a perceptual auditory attribute such as pitch. Lesion studies in humans and animals also suggest that an intact auditory cortex is necessary to perform pitch-related behavioural tasks. Whitfield (1980), for example, showed that cats lost their ability to perform and re-learn a missing-fundamental discrimination task following bilateral ablation of the auditory cortex. In humans, Zatorre (1988) found the discrimination performance of patients with unilateral temporal-lobe lesions — particularly those with damage to the right HG — to be significantly impaired when the task required discrimination of missing-fundamental tones, but not when the fundamental was present.

Experimentally, one can approach the search for a possible representation of pitch from many different angles. One could, for example, commit to some specific assumptions regarding the semantics of a neuronal representation of pitch and search for neurons (or groups and networks of neurons) in auditory cortex that meet these assumptions. One particular such form of representation, which experimenters have been keen to find, is an explicit rate-place code for pitch along a frequency-labelled axis. In this putative code, pitch (specifically: pitch height) would be determined by the place of maximum activity along the axis, and the height of the peak (relative or absolute) would reflect the perceptual saliency of the percept. All pitch-evoking sounds, and only those, should elicit a discriminable peak of activity corresponding to their pitch height. Finally — and critically, in order to classify as a representation of perceived pitch — the peak should reflect the subjective pitch height on each single trial, rather than one of several, possibly conflicting, physical cues, for example[17]. One hallmark of such a representation would be the existence of neurons with band-pass tuning for pitch height. Several studies in the primary auditory cortex (A1) of monkeys have failed to identify such neurons. Schwarz and Tomlinson (1990), for example, recorded responses to pure tones and harmonic complex tones (with and without fundamental) in A1 of awake macaques, which had previously been trained to perform a missing-f_0 pitch discrimination task. However, all neurons essentially responded to the spectral content of the stimulus within their respective excitatory and inhibitory receptive fields, rather than to the missing fundamental. Fishman et al. (1998) came to similar conclusions based on multi-unit activity (MUA) recorded in a different macaque species. In a study of A1 in marmoset monkeys, Kadia and Wang (2003) discovered that about 20% of their units showed multi-peaked frequency responses to pure tones with peaks often occurring at "harmonically-related" frequencies (which they defined as related by simple integer ratios). Whilst one can imagine that such neurons might be very useful for determining the f_0 of a complex sound via some form of spectral pattern matching (cf. section 2.4), they clearly do not constitute an explicit representation of pitch by themselves. To date, the strongest evidence *for* the existence of an explicit rate-place code for pitch is based on responses recorded in the core region of auditory cortex in marmoset monkeys (Bendor and Wang, 2005). A subset of neurons near the boundary between A1 and primary-like field R were found to be bandpass-tuned not only to pure tones but also to the missing fundamental of harmonic complex tones with partial frequencies entirely outside the neurons' pure-tone response range (see Figure 2.19 A & B). The neurons also fulfilled several other "neccessary conditions" for an explicit representaion. They reponded to IRN and click-train stimuli with time delays

[17] Note, that the existence of such an explicit representation is by no means a necessity, and *a priori* no more likely than an "implicit" representation, whereby pitch is encoded by the joint activity of neurons in a widely distributed network, not a single one of which exhibits any obvious kind of tuning to experimenter-defined stimulus parameters (see e.g. Bizley et al., 2009).

and click intervals matched to their preferred frequency. Furthermore, response firing rates were higher for complex tones with low harmonics (IRN with many iterations; regular click-trains) than for high-harmonic complex tones (IRN with few iterations; highly jittered click trains), in agreement with expected psychophysical changes in pitch salience (Figure 2.19 C). Doubts remain, however, whether the control for harmonic distortion products was effective. Thus, the data is suggestive but not conclusive as evidence for an explicit representation of pitch in field R. Another reason for concern is the fact that characteristic frequencies of pitch-selective neurons did not exceed 800 Hz, indicating that an additional representation is necessary to account for the perception of pitch above this range.

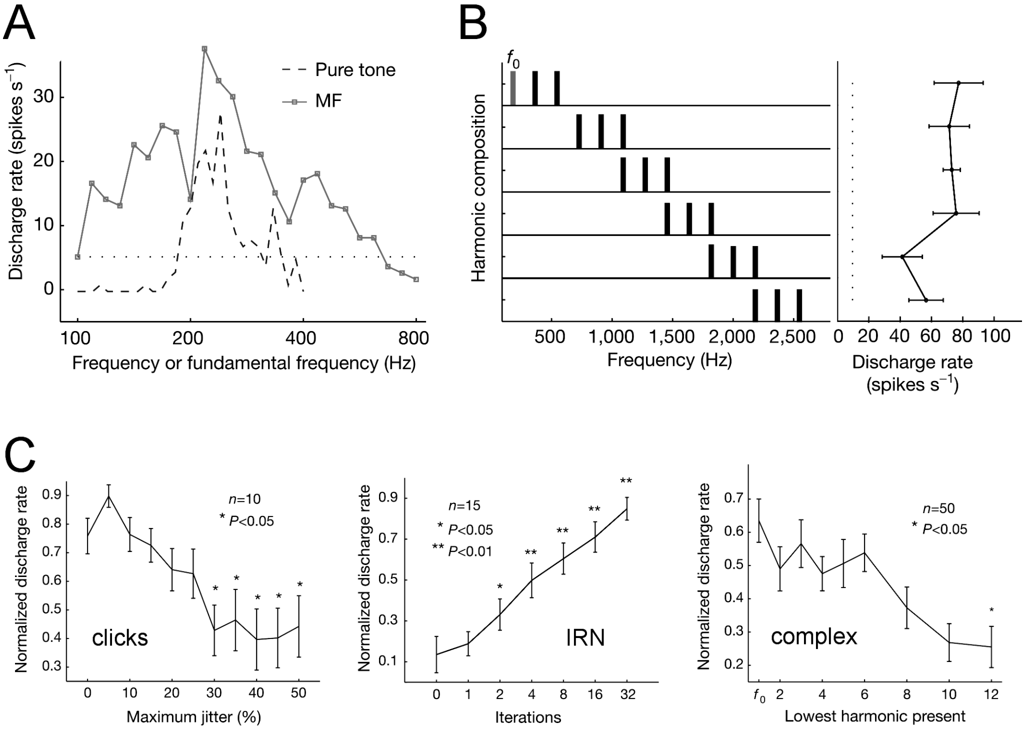


Figure 2.19: Neural responses to different pitch-evoking sounds in marmoset field R. A: Tuning of a single neuron to pure-tone frequency and missing f_0. B: Firing rate of a single neuron during presentation of missing-f_0 complex tones. Harmonic numbers range from 1-3 (top) to 12-14 (bottom). Dotted line indicates +2 SEM from spontaneous rate C: Population-averaged, normalised discharge rates for click trains (left), IRN (middle) and complex tone stimuli (right). Stimulus parameters (f_0 or Δt) were chosen for each cell to match its CF. Adapted from Bendor and Wang (2005).

Regardless of the exact semantics of the representation of pitch, one can ask whether pitch-related processing in cortex is spatially confined to specific brain regions at all. A number of studies have attempted to answer this question using functional magnetic resonance imaging (fMRI). fMRI provides a measure of the blood-oxygenation-level-dependent (BOLD) magnetic spin density in the tissue and is thought to provide a reasonable proxy for local metabolic demand (Logothetis, 2003). Its spatial resolution is on the order of millimetres but its temporal resolution is low — on the order of several seconds — due to the slow reaction time of the brain vasculature to changes in oxygen demand. A locally confined "pitch processor" might be expected to cause a stronger BOLD signal when listening to a pitch-evoking sound than in the presence of a non-pitched sound. Patterson et al. (2002) contrasted fMRI responses to band-pass-filtered iterated rippled noise (IRN) and white noise, the former evoking a pitch percept despite the absence of prominent, synchronised envelope modulations and sharp rate-place peaks in the AN response. A selective increase in the fMRI signal for the pitched sounds over noise was found bilaterally in lateral HG, most likely outside A1, but still within the core region (homologous to fields R and RT in primates). Penagos et al. (2004) provided additional support for the role of lateral HG as a pitch centre by showing that the strength of the fMRI signal in response to missing-fundamental sounds covaried with its expected pitch strength (but not with fundamental frequency or spectral band-pass region). Experiments by Hall and Plack (2009), however, question the generality of a pitch-related activity increase in HG using a pure tone, different harmonic complex tones and a binaural Huggins-pitch stimulus in addition to IRN. They showed that *only* IRN evoked strong responses in HG, while also activating other regions, including the *planum temporale* (PT). The other stimuli did non reliably activate HG, but like IRN, they also activated the PT (which Patterson et al. (2002) had identified as a region sensitive to time-varying as opposed to static pitch). Activation of HG was also not found to covary with differences in pitch discriminability, measured in the subjects during the experiment. In conflict with some of the findings by Hall and Plack (2009), Puschmann et al. (2010) recently demonstrated the activation of HG in response to two types of binaural pitch stimuli, including Huggins pitch. For the time being, it seems, the debate surrounding the roles of HG and the PT in pitch perception — as measurable in fMRI experiments — remains unresolved.

Other studies have investigated the cortical substrates of pitch using magnetoencephalography (MEG), measuring changes in the magnetic field above the skull surface effected by the flow of electrical charges during neural activity[18]. Pantev et al.

[18] MEG provides much better temporal resolution that fMRI (milliseconds), but the location of the current-generating sources underlying the MEG response are not uniquely determined by the measure-

(1989) recorded MEG responses to pure tones and harmonic complex tones in humans and found that pure tones of a certain frequency would cause activation in the same region as harmonic complex tones with matched f_0. They concluded that the known tonotopy of auditory cortex is in fact a periodotopy. Langner et al. (1997) took similar measurements but in conflict with the previous results, reported a periodotopic gradient *orthogonal* to the tonotopic axis, i.e. similar to orthogonal layout of frequency and periodicity previously found in the IC (Schreiner and Langner, 1988; see section 2.3.1). Both studies, however, suffer from different potential confounds due to peripheral distortion products (McAlpine, 2004; see also Walker et al., 2011 for a more thorough discussion), in addition to the very limited range of stimuli used — leaving us once again in the dark with regards to the cortical representation of pitch.

To summarise, we know very little about the cortical substrate of pitch perception in spite of long-lasting and intense experimental efforts. Furthermore, no matter what the ultimate representation may be, there is virtually no empirically-backed indication of the types of processing steps and computations in cortex that might give rise to our percept. We have *some* reason to believe that neurons in the lateral Heschl's gyrus may play an important role in these computations, possibly in addition to a wider range of non-primary areas located in the temporal lobe, including for example the *planum temporale*. This may give us a vague sense of where to place pitch perception within the hierarchy of auditory computational tasks — ranging from the mere detection of a sound at the one end to complex tasks such as auditory scene analysis, object recognition and semantic processing at the other end (e.g. Nelken, 2004). The putative involvement of "higher" cortical areas suggests, if anything, that pitch should perhaps be thought of as the outcome of a rather high-level computation reflecting not only instantaneous sensory evidence but also subjective prior knowledge and top-down expectations. At the level of the brainstem and midbrain, we have a better understanding of the relationship between stimulus and evoked neural activity than we do in cortex. However, it appears that these are preliminary processing steps, operating on different periodicity-related stimulus features in parallel, and that they are only combined into a unified percept further upstream. In terms of their value in constraining models of pitch perception, neural response properties in the early auditory pathway may be suggestive of the kind of "interim results" we may want to arrive at at some point during our computation.

ments, making definitive source localisation difficult

By themselves, however, they do not tell us much about the computations that build up on them.

2.4 Theories and models of pitch perception

As we have seen in section 2.1, pitch is a percept that can be evoked by a large variety of different sounds, from elementary sinusoidal vibrations, through complex-shaped periodic vibrations, all the way to highly irregular, aperiodic sounds. In this section, we will explore whether or to what degree we can bring order to and explain these diverse findings in terms sound properties and our current understanding of the physiological processes involved in the transformation of the physical signal in the peripheral auditory system. What is the uniting factor amongst all those various sounds that evoke the same pitch? As we will see, a satisfactory answer to this question is currently not at hand. At the same time, however, we will discover a number of promising leads, each of which can account for a substantial fraction of the observed phenomena. We will then attempt to formulate a model that is capable of incorporating these various leads in a concise, principled fashion over the course of the remaining chapters.

We are exposed to a great variety of approximately periodic sounds in our environment, and their periodicity often carries behaviourally relevant information. Furthermore, the majority of periodic sounds evoke a pitch equal to their period. It therefore seems only reasonable to assume that pitch perception reflects a process of periodicity estimation, developed to extract this information. One way to approach the problem of modelling pitch might therefore be to find a reliable method of periodicity estimation and test whether its predictions also hold for all the special cases of periodic sounds that do not evoke a pitch equal to their period (or perhaps none at all) and those of non-periodic sounds that nevertheless evoke a pitch as discussed.

A seemingly straightforward method for determining the period of a perfectly periodic sound $x(t)$ with $x(t) = x(t + \Omega)\ \forall t$ is to determine its autocorrelation function

$$R(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) \cdot x(t-\tau)\,\mathrm{d}t \quad , \tag{2.4}$$

which gives a measure of the temporal self-similarity of the signal with itself at all possible time-lags τ (assuming that we can somehow determine the integral in the expression above). One of the fundamental properties of R is that it takes its maximum at $\tau = 0$ (e.g. Hartmann, 1997), where its value equals the signal power $\frac{1}{T}\int_{-T/2}^{T/2} x(t)^2$. Since our sound is periodic, i.e. $x(t) = x(t + \Omega)$, we also know that $R(\Omega), R(2\Omega), \ldots$ must take the same, maximum value. All we need to do to determine Ω in this idealised case is to find the smallest non-zero value of τ at which $R(\tau) = R(0)$.

Alternatively, we could compute the Fourier transform $\hat{x}(f)$ of $x(t)$:

$$\hat{x}(f) = \int_{-\infty}^{\infty} x(t) \cdot \exp(2\pi f i t)\,\mathrm{d}t \quad . \tag{2.5}$$

According to Fourier's theorem, the Fourier amplitudes $|\hat{x}(f)|$ can only be non-zero at frequencies $f_k = \frac{k}{\Omega}, j = 1, \ldots, \infty$, i.e. integer multiples of the inverse sound period. Therefore, in order to find Ω we have to determine the greatest common divisor of all non-zero Fourier components. In reality, of course, our signals are only finite in duration and sampled at discrete points in time in the case of digital systems. We may still compute discretised variants of equations (2.4) and (2.5) and compute R and $\hat{x}$ numerically for any given signal. However, as soon as the signal is no longer perfectly periodic, R does not necessarily peak equally and maximally at $\tau = \Omega$ and its multiples, and the non-zero Fourier components of $|\hat{x}|$ may no longer have a greatest common divisor (greater than the sampling interval) at all. Nevertheless, heuristics can be used to determine Ω based on either the autocorrelation function or the Fourier amplitude- or power-spectrum (see e.g. de Cheveigné, 2005). For autocorrelation-based estimates of Ω, we need a rule to pick one of possibly many local peaks in $R(\tau)$ of approximately equal height. For spectral-based estimates of Ω, an appropriate metric is required that allows us to determine the fundamental frequency of the harmonic frequency stack that best fits the observed spectrum — a process commonly referred to as "spectral pattern matching" or "template matching".

While these two general approaches will reappear in many theories of pitch perception in *modified* form, we can also see why they are by themselves not suitable as models of pitch perception, when applied directly to the unprocessed waveform of the acoustic stimulus. One point of concern is the marked degradation of the pitch of HCTs as the rank of the lowest harmonic increases above 8 to 10. It is not evident *a priori*, why it should be much harder to match harmonics 12 to 15 against a harmonic template than it is to match harmonics 5 to 8. Of even greater concern, however, is the pitch of aperiodic sounds such as SAM noise or interrupted noise. The spectra of these sounds are flat in expectation, and the sound waveform is uncorrelated in time: an impossible task for pattern matching or autocorrelation. One might be tempted to perform periodicity analysis not on the raw waveform, but on its periodic *envelope*, which can be readily extracted from the signal using a variety of methods (e.g. Turner and Sahani, 2011; see also section 3.2.2). Unfortunately, this brings an entire host of new problems with it (aside from not solving the issue of high-harmonic HCTs). Firstly, the pitch of SAM tones is in general not equal to their envelope periodicity, but shifted away from it. Secondly, the envelope periodicity and modulation-depth of HCTs depends greatly on their phase relationships, whereas listeners are largely insensitive to phase relationships between low-rank harmonics.

In order to explain these subtle effects, it may be necessary to take the physiological properties and constraints of the peripheral auditory system into account. Von Helmholtz (1863) developed a comprehensive theory of "physiological acoustics", including, amongst others aspects, a theory of pitch now commonly referred to as the "place theory". Its central tenet is that the pitch of a sound is determined by the place of maximum excitation of the basilar membrane in the inner ear, which Helmholtz thought of as a mechanical Fourier analyser (cf. section 2.2.3.3). Helmholtz essentially adopted an earlier theory by Ohm (1843), known as Ohm's acoustic law. Ohm had postulated, that in order to hear a pitch, the Fourier spectrum must contain a component at this frequency, and that the pitch of a harmonic complex tone is determined by the lowest of these[19]. An intense debate between Ohm and Seebeck ensued (e.g. Seebeck, 1843; see also Turner, 1977 for a detailed historic account), who in turn had found contradictory evidence that sounds could evoke a pitch equal to their periodicity rate, even when the corresponding Fourier component was missing or extremely weak[20] (see section 2.1.1.2). Helmholtz, in embedding Ohm's acoustic law into a physiological context, provided what seemed like a resolution to this debate (Helmholtz, 1856): he realised that non-linearities in the peripheral transduction process (cf. section 2.2.3.3) can introduce harmonic distortion products (DPs) at the fundamental frequency of a missing-f_0 HCT, such that the *effective* stimulus exciting the auditory nerve (AN)

[19] Its key sentence, translated as literally as possible, reads as follows: "The pulses required to produce a tone of frequency m must follow each other in intervals of length $\frac{1}{m}$, and in each of these intervals they must continuously contain the shape $a.sin 2\pi(mt+p)$, either purely or such that this form can at least be segregated as a real, constituent part."

[20] In other words: Seebeck had discovered the pitch of the missing fundamental.

fibres does no longer lack energy at the fundamental[21].

Helmholtz' place theory became the dominant theory of pitch perception for many decades. Experiments by Schouten (1938, 1940), however, proved it wrong. Schouten, using equipment which provided a new degree of control over the acoustic stimulus, devised a scheme of eliminating the DP at the fundamental of a missing-f_0 HCT by *adding* a pure tone at f_0, carefully calibrated in amplitude and phase so as to cancel the BM excitation caused by the DP itself. The pitch of the missing fundamental remained unchanged. Furthermore, he generated SAM tones as described in section 2.1.2. The pitch of a SAM tone is generally not equal to its envelope modulation frequency. A quadratic DP however, invoked by Helmholtz as the explanation for missing-f_0 pitch, is expected to occur precisely at the modulation frequency, not the perceived pitch. Hence, Schouten's experiments effectively ruled out harmonic distortions, and thus Helmholtz' place theory, as the sole (or dominant) mechanisms in determining the pitch of a sound that lacks a tonal component at the pitch frequency. Schouten proposed a different mechanism (see Figure 2.20). Consider a peripheral bandpass filter that is centred around the carrier frequency f_c of the SAM tone and wider than the modulation frequency g (i.e. the spectral components are unresolved). Its output will resemble the waveform of the SAM tone itself. Action potentials in the associated AN fibre (or bundles thereof) will occur phase-locked to peaks in the filter output (cf. also section 2.3.1). Since the peaks in the temporal fine structure (TFS) of the filter output *shift* progressively relative to the phase of the envelope modulation, the time-span between the two highest peaks of two successive envelope periods is somewhat shorter than the envelope period itself (τ_1 in Figure 2.20). In fact, it matches the perceived pitch very closely (e.g. 204 Hz for $f_c = 2040\,\text{Hz}$ and $g = 200\,\text{Hz}$). The ambiguity of the percept, and the occasional matching to frequencies around 185 and 227 Hz can similarly be explained by the timing of spikes that occur not at the two highest TFS peaks but at peaks slightly before or after (e.g. τ_2 and τ_3 in Figure 2.20). Thus, the pitch of SAM tones according to Schouten's explanation is based on the analysis of peak times in the output of AN fibres that are stimulated by unresolved harmonics (for example by computing a histogram of inter-spike-intervals). As such, it cannot constitute a general theory of pitch by itself: missing-f_0 exists, and is strongest, for resolved HCTs

[21] Helmholtz speculated that the source of the non-linearity was the asymmetry of the middle-ear ossicular chain. He was wrong, in that the the active mechanical response of the basilar membrane has now been identified as the predominant source of non-linearity in the transduction process. Nevertheless, his observation regarding the generation of distortion products remains valid in principle.

(Shackleton and Carlyon, 1994, cf. section 2.1.1.2), even though the peripheral channel outputs are effectively unmodulated pure tones of different frequencies in this case. The same argument holds for the pitch shift of resolved SAM tones (or shifted HCTs with a greater number of components, e.g. de Boer, 1956b; Patterson, 1973, 1976).

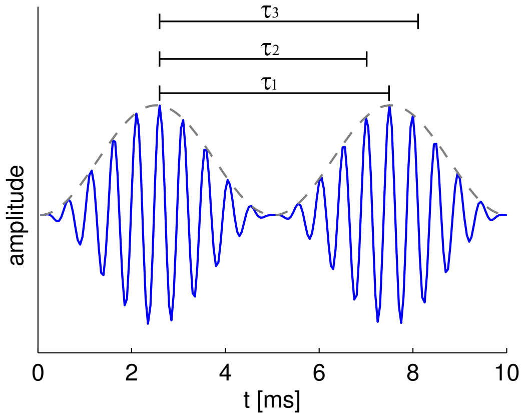


Figure 2.20: Sinusoidally amplitude-modulated tone with $f_c = 2040\,\mathrm{Hz}$ and $g = 200\,\mathrm{Hz}$ (blue; grey line indicates the 200 Hz envelope). Timing of the peaks in the fine structure underneath the envelope shifts in phase from one envelope peak to the next. $\tau_1 \approx 4.9\,\mathrm{ms}$ indicates the time-span between the two-largest fine-structure peaks within two consecutive peaks of the envelope, corresponding closely to the dominant pitch of 204 Hz. τ_2 and τ_3 match the alternative pitches of approximately 185 and 227 Hz. After Schouten et al. (1962).

Modern theories of pitch perception can be broadly divided into those that are based on refinements of Helmholtz' place theory, and those that are based on a temporal analysis of peripheral filter outputs similar to Schouten's idea (see Figure 2.21). The former class of "pattern matching" theories extends Helmholtz' in that pitch is derived from the *joint* pattern of excitation along the entire BM (or AN), rather than just its peak. The latter class of "temporal" models extends Schouten's idea by combining the outcome of the channel-by-channel periodicity analysis into a joint estimate across channels. Periodicity analysis in these models is often based on computing the autocorrelation function of each channel (see above). We will discuss several of these models along with their respective merits and shortcomings in the following.

2.4.1 Examples of spectral pattern-matching models

Pattern matching models work on the assumption that the pitch of a sound is determined by comparing the peripheral representation of the stimulus spectrum against harmonic stacks with different fundamental frequencies until the best match is found.

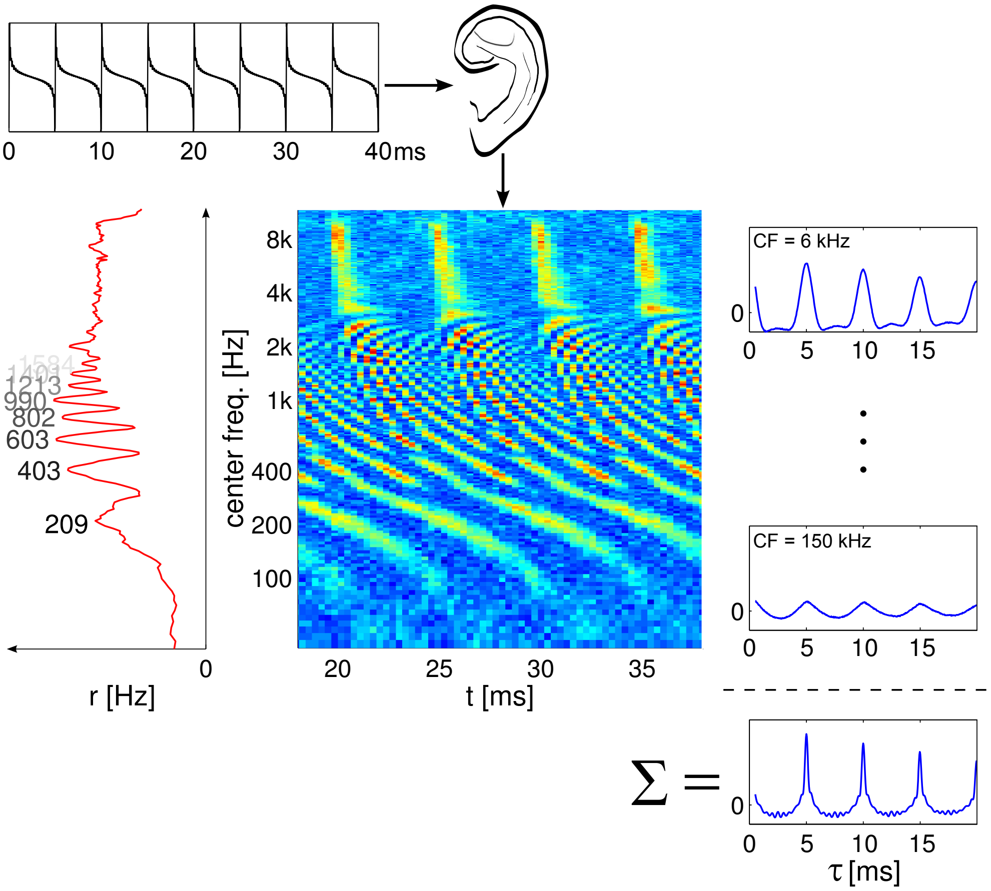


Figure 2.21: Spectral and temporal cues in models of pitch. Centre: A sound (200 Hz HCT) evokes a time-varying neural firing pattern in the auditory nerve. Left: The average-rate profile along a tonotopic axis gives a coarse representation of the stimulus spectrum. Local firing-rate maxima in the profile occur in fibres with CFs close multiples of 200 Hz. Right: Autocorrelation analysis in each peripheral channel yields a measure of temporal stimulus self-similarity. High harmonics generate envelope modulations at 200 Hz in high frequency channels due to the widening of filters with increasing CF. Summation across channels is a commonly-used strategy to obtain an aggregate measure across channels.

Put in a different way, they try to find an approximate greatest common divisor of the observed spectral components above the lower limit of pitch perception ($> 30 - 40\,\text{Hz}$). Initially suggested by de Boer (1956b), three influential models of this kind were published in close succession by Goldstein (1973), Wightman (1973) and Terhardt (1974, 1979), and more related work has been published since: Duifhuis et al. (1982) proposed a model in the spirit of Goldstein, that contains (in the authors words) "elements that are virtually identical" to elements in Terhardt (1979). Hermes (1988) developed a method that similarly resembles that of Terhardt. Cohen et al. (1995) provided a neural-network implementation of spectral template matching, which Grossberg et al. (2004) incorporated into a hierarchical model of auditory scene analysis. All these models are fundamentally limited in their scope by the assumption that pitch is based on aurally resolved spectral peaks: as such, they cannot explain the pitch of unresolved HCTs and SAM tones, or that of spectrally white sounds such as SAM noise. Several points regarding these limitation deserve mention at this point.

Firstly, Wiegrebe and Patterson (1999) have shown that the amplitude modulations in *band-limited* SAM noise create a distortion product on the basilar membrane at the modulation frequency (see also Strickland and Viemeister, 1997). Furthermore, when the distortions were cancelled by adding a pure-tone in anti-phase (as introduced by Schouten (1938); cf. section 2.4 above), listeners could no longer discriminate between different modulation rates. This indicates that the pitch of narrow-band SAM noise may indeed critically depend on peripheral distortions, and thus spectral models could be sensitive to it provided that they are based on peripheral front-ends with sufficient physiological realism. However, Wiegrebe and Patterson (1999) also found that AM-rate discrimination of SAM noise was impaired but *still* possible despite the cancellation tone, when the noise bandwidth was wider than twice the modulation frequency. Thus, the pitch of SAM noise in general does not appear to be explicable in terms of distortion products alone[22].

Secondly, with regard to the limitations of spectral pattern-matching models in explaining the pitch of unresolved HCTs, it is perfectly possible in principle to extract spectral information about unresolved components of the stimulus from a spectral analysis of the response of high-frequency AN fibres (limited of course to their respective

[22] Note also that the original experiments by Burns and Viemeister (1976, 1981) were performed using both wide-band SAM noise *and* band-limited SAM noise with added band-reject masking noise.

response bandwidths). Thus, a spectral-based, central pitch processor performing pattern matching on this "peripherally-derived" spectrum could also take unresolved harmonics into account during periodicity estimation. However, this possibility is typically not exploited in pattern-matching models. A notable exception is the central spectrum model by Srulovicz and Goldstein (1983) which implements precisely this idea, generating an internal representation of the stimulus spectrum from spike-interval histograms in the peripheral channels. However, to the best of our knowledge, no attempt was subsequently made to combine this potentially powerful spectral representation with a corresponding, pattern-matching-based periodicity estimator.

2.4.1.1 Wightman's pattern transformation model

In terms of its categorisation, Wightman's pattern transformation model of pitch (Wightman, 1973) is an interesting case: pitch predictions are based on the profile of average firing rates of auditory nerve fibres ordered by their CFs and the model is hence legitimately classified as an example of spectral pattern-matching. On the other hand, what Wightman tries to achieve with his model is to compute an approximate stimulus autocorrelation function and estimate the stimulus periodicity by finding its maximum non-zero peak (cf. equation (2.4)). Wightman himself describes the model therefore as "essentially an autocorrelation model" (ibid.). The simple, appealing logic behind the model is the following: According to the Wiener-Khinchin theorem, the autocorrelation function $R(\tau)$ is related to the Fourier power spectrum $|\hat{x}(f)|^2$ via the Fourier transform (Hartmann, 1997). Since the central auditory system has access to a degraded but nevertheless recognisable representation of the stimulus spectrum in the form of time-averaged AN firing rates, it might attempt to apply the Fourier transform to his surrogate spectrum to obtain a surrogate autocorrelation function and perform periodicity estimation based on the latter.

In our implementation of Wightman's model for use in chapter 5 (cf. section 5.4.3), we emulate the peripheral transduction process as a linear gammatone filter bank bank followed by half-wave rectification and low-pass filtering in each channel. A more detailed description of this peripheral front-end is given in section 3.2. For a given acoustic stimulus, we perform the following steps to obtain a prediction for its pitch:

1. Compute the average firing rate a_i of each peripheral channel with CF f_i ($i =$

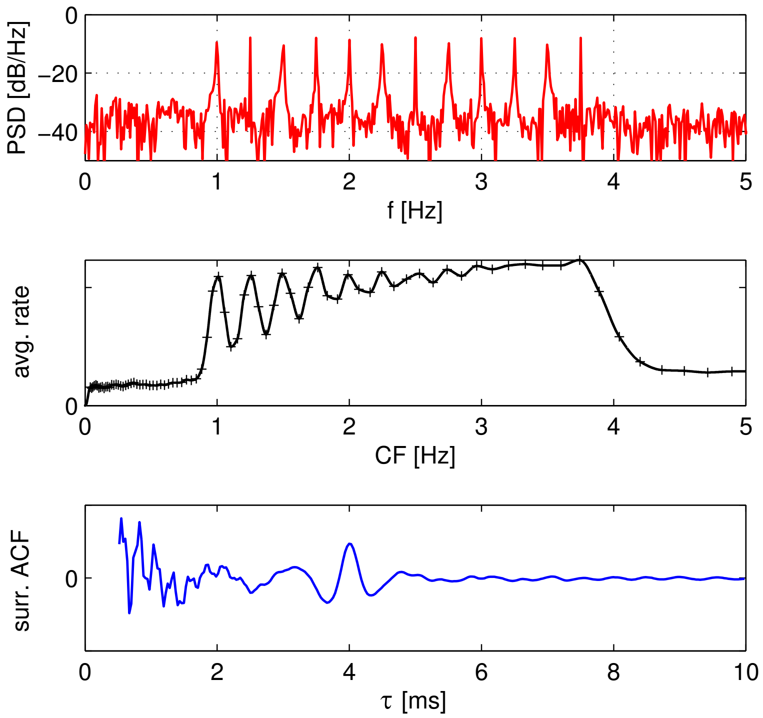


Figure 2.22: Missing-f_0 pitch in Wightman's pattern transformation model. Top: Power spectrum of a 250 Hz missing-f_0 HCT with harmonics 4 to 15 in a background of low-intensity white noise. Middle: Average firing rate of auditory nerve fibres as a function of their CF. Crosses indicate CFs in the model, the curve is a cubic Hermite interpolation. Bottom: Inverse Fourier transform of the firing rate profile. Note the peak at $\tau = 4$ ms, corresponding to the f_0 of 250 Hz.

$1 \ldots C$).

2. Interpolate between the ERB-spaced CFs f_i using cubic Hermite splines to obtain a firing-rate profile $a(f)$ that is linearly sampled in frequency.

3. Compute the discrete inverse Fourier transform of $a(f)$ to obtain the surrogate autocorrelation function $\widetilde{R}(\tau)$.

4. Choose the highest peak of $\widetilde{R}(\tau)$ (above 1 ms) as the pitch period.

Figure 2.22 demonstrates that the model is capable of finding the missing fundamental of a resolved HCT. As the harmonics of the sound above approximately the 10th do not produce discernable peaks in the $a(f)$ the model is fundamentally limited to pitches evoked by resolved frequency components.

2.4.1.2 Terhardt's theory of virtual pitch

Terhardt (1974, 1979) proposed a different approach to pattern matching. In his model of virtual pitch, a spectral analyser determines the frequency and amplitude of significant local peaks in the Fourier power spectrum of the stimulus. In the next step, the perceptually-relevant component amplitudes (and to a small degree even the frequencies) are determined based on an overall spectral gain function, masking effects between nearby spectral peaks as well as the local level of background noise. In the final stage, the adjusted and thresholded spectral component enter a process of "subharmonic summation"[23]. Each component adds to the evidence for the presence of a harmonic stack with fundamental frequency equal to either the component frequency itself, or its integer submultiples (or subharmonics) up to some maximum rank (e.g. 12). A virtual pitch profile is obtained, the peaks of which are considered possible pitch matches to the stimulus, in order of their magnitude. Subharmonic summation had been used for f_0 estimation already prior to Terhardt (e.g. Schroeder, 1968; see also Hermes, 1988 for a subsequent example), but not in combination with a front-end that modifies the raw stimulus spectrum in order to take masking and resolvability into account. Terhardt distinguishes the candidate virtual pitches from "spectral pitches", which simply correspond to individual, audible frequency components, thus differentiating between synthetic and analytic listening modes (cf. section 2.1.1.2). Terhardt further posited that the association of component frequencies and their subharmonics be learnt through the extensive exposure to (near-harmonic) speech sounds during development. Shamma and Klein (2000) proposed an alternative mechanism, by which harmonic templates emerge from cross-correlated firing of AN fibres with harmonically-related CFs even when stimulated with noise or other broad-band stimuli such as clicks. In any case, the origin of the knowledge about harmonic relationships is largely irrelevant when using Terhardt's model to predict pitch once the learning phase is completed. Terhardt et al. (1982) published a detailed algorithmic description of the model, where the outcome of the learning phase, i.e. the knowledge about harmonic frequency ratios, is fully hard-coded into the algorithm.

In our implementation of the virtual pitch model (cf. section 5.4.2), we precisely followed the specifications of Terhardt et al. (1982) with two small modifications. Firstly, we did not take interaction effects between spectral components into account that modify their effective frequencies by small amounts. According to the authors, these can be safely ignored in many cases, except for very specific minute pitch-shift phenom-

[23] We have borrowed this term from a related, subsequent model by Hermes (1988), not from Terhardt himself.

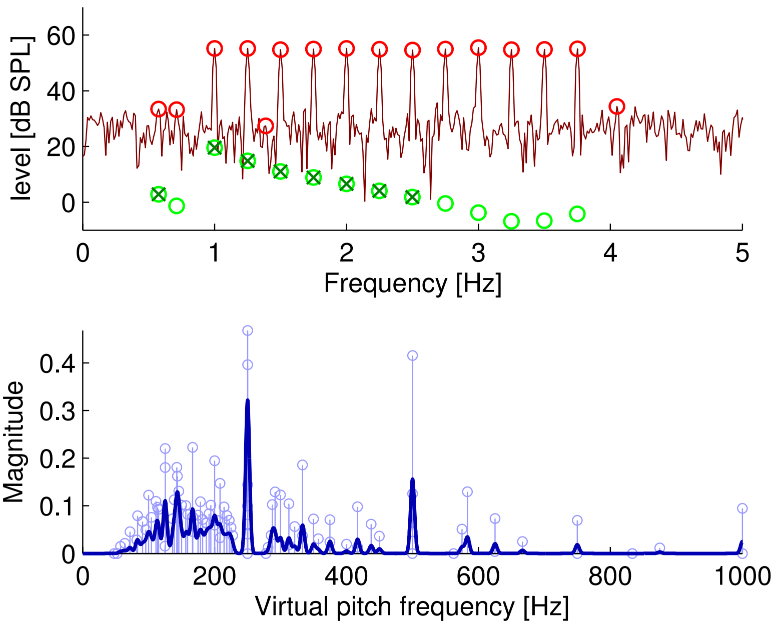


Figure 2.23: Missing-f_0 pitch in Terhardt's virtual pitch theory. Top: Power spectrum (up to 5 kHz) of a missing-f_0 HCT with harmonics 4-15 (dark red). Steep local peaks are extracted (red circles) and their effective amplitudes adjusted according to a spectral weighting, masking between nearby frequencies and local masking due to noise (green circles). Components with effective amplitudes above 0 dB (green crosses) are entered into a process of subharmonic summation. Bottom: Extracted virtual pitch candidates (light blue) and smoothed virtual pitch profile (dark blue). The model correctly identifies the missing fundamental of 250 Hz.

ena. Secondly, we convert the final set of candidate virtual pitches from a collection of discrete pairs of frequencies and magnitudes (i.e. a sum of delta-functions) into a continuous virtual pitch profile by applying a narrow (2 Hz) Gaussian smoothing filter. This was found to be useful to avoid multiple virtual pitch candidates at near-identical frequencies in the case of noisy stimuli (a similar effect could have been achieved by a discrete binning of virtual pitches).

2.4.1.3 Goldstein's optimum processor model

Goldstein developed a model to explain the pitch matching behaviour for two-component complex tones reported by Houtsma and Goldstein (1972) (cf. section 2.1.2). Out of the three pattern matching models discussed in this section, Goldstein's "optimum processor" model is the most abstract in its treatment of the peripheral auditory system. It is assumed that a peripheral front-end simply returns the frequencies of the two stimulus components. Therefore, not only information about component phases is lost (as in all models based on the stimulus power spectrum), but also their respective amplitudes. Given these highly reduced inputs, however, Goldstein develops a statistical framework whereby the listener assumes that the two observed spectral components are two successive harmonics of the same, missing f_0 — as was the case in the experiments by Houtsma and Goldstein (1972) — and forms an *optimal* estimate of its frequency, exploiting statistical knowledge about the imperfections of the peripheral spectral pre-processor.

In particular, Goldstein assumed that the two frequencies f_1 and f_2 available to the central pitch processor are independent, noisy samples drawn from two Gaussian distributions with means equal to the true, harmonic stimulus frequencies $n_1 f_0$ and $n_2 f_0$ and frequency-dependent variances $\sigma^2(n_i f_0)$:

$$\begin{aligned} \mathrm{P}(f_i \,|\, n_i, f_0) &= \mathcal{N}(f_i \,|\, n_i f_0,\, \sigma(n_i f_0))\,, \quad i = 1, 2 && (2.6)\\ &= \frac{1}{\sqrt{2\pi\,\sigma^2(n_i f_0)}} \exp\left(-\frac{(f_i - n_i f_0)^2}{2\,\sigma^2(n_i f_0)}\right) \quad . && (2.7) \end{aligned}$$

The rank of the lowest component n_1 is unknown, but it is assumed that $n_2 = n_1 + 1$ (or when extended to cases with more than two components: $n_i = n_1 + i - 1$). Furthermore, the central processor has full knowledge about the frequency dependence of the noise variance $\sigma^2(f)$. According to Goldstein's theory, pitch reflects the *maximum likelihood* estimate (e.g MacKay, 2003) of f_0, i.e. the central processor determines those values of the two unknown variables f_0 and n_1 that make the observations f_1 and f_2 seem most likely[24]:

$$(f_0^*, n_1^*) = \underset{(f_0, n_1)}{\operatorname{argmax}} \prod_i \mathrm{P}(f_i \,|\, n_i, f_0) \quad . \tag{2.8}$$

Note that Goldstein's model for all its lack of physiological detail is not a purely signal-based pitch-determination algorithm: the frequency-dependent noise variance of the spectral preprocessor reflects an inherent property of the peripheral auditory system.

The restrictive assumptions regarding the input representation limit the scope of Goldstein's model rather severely. It is not readily applicable to stimuli with noisy and continuous spectra, and where applicable its predictions do not generalise to situations where amplitude or phase differences between different spectral components become important in determining the pitch (see e.g. chapter 5). Nevertheless, the general

[24] Note that the inferred rank n_1^* is irrelevant.

framework of statistical estimation that underlies Goldstein's model can be extended to more complex scenarios. Our own model, which we will develop in chapter 3 can be seen as one possible such extension. In particular, we will replace Goldstein's highly abstract spectral preprocessor with a more realistic auditory-nerve model (albeit still schematic). This will allow our model to process sounds with arbitrary spectra, exploiting whatever information regarding the stimulus periodicity is preserved in the AN response, be it harmonic peaks in the rate-place profile or its more fine-grained temporal properties.

De Boer (1977) noted an interesting relationship between the models of Wightman, Terhardt and Goldstein: he showed that Goldstein's model is approximately equivalent to Terhardt's in the limit of $\sigma(f)^2 \rightarrow 0$, i.e. in the zero-noise case of Goldstein's spectral preprocessor. Wightman's model, conversely, can be approximately obtain from Goldstein's when $\sigma(f)^2$ is assumed to grow large. As a result, the three models are sometimes regarded as near-equivalent. As we will see in chapter 5 (section 5.4), the behaviour of Terhardt's and Wightman's models can nevertheless be markedly different, despite their constituting different limiting cases of the same underlying model.

2.4.2 Temporal models and summary autocorrelation

Licklider (1951) proposed the "duplex theory of pitch perception", the direct ancestor of many modern, timing-based pitch theories. Licklider thought of pitch as a percept along two intrinsic dimensions: chroma and height (cf. section 2.1). He hypothesised that the mechanism for determining pitch along these two dimensions was similarly "duplex" in nature. Pitch height, according to his theory, is determined by the place of maximum excitation along the BM, while chroma is determined by a time-domain analysis of the peripheral frequency-channel outputs. Modern descendants of his theory typically do not make a categorical distinction between pitch height and chroma, and regard the temporal component of Licklider's duplex model as a way to determine pitch along a single dimension that implicitly determines both these attributes. For this temporal component, Licklider imagined that the auditory system continuously computes a running, time-windowed autocorrelation function (ACF) $h_i(t, \tau)$ in each frequency channel i, where t denotes the time of evaluation, and τ the lag at which the ACF is evaluated. If we assume that the time-window over which the peripheral channel-output $a_i(t)$ is integrated in the computation of the ACF is exponentially shaped with time constant λ, we can write h_i as

$$h_i(t,\tau) = \int_0^t a_i(t') \cdot a_i(t'-\tau) \cdot \exp((t'-t)/\lambda)\,\mathrm{d}t' \quad , \tag{2.9}$$

i.e. as a leaky integration of the product $a_i(t') \cdot a_i(t'-\tau)$. This functional form seemed particularly appealing to Licklider, as he considered it to map directly onto a possible neural substrate. For every value of τ in each channel i, he speculated that a neuron received two copies of a_i as its input, one of them delayed by τ with respect to the other by means of a neural delay line. The recipient neuron fires when spikes in its two inputs coincide, thereby effectively calculating their point-wise product. A further neuron integrates the output of the coincidence detector, but its excitation "dissipates itself spontaneously, perhaps at a rate proportional to the amount accumulated" (Licklider, 1951), giving rise to a time-varying membrane potential exactly of the form of equation (2.9). If one imagines these neurons to be arranged on a two-dimensional lattice (extending in dimensions of peripheral CF and τ), a HCT would give rise to a ridge along the CF-Dimension for the value of τ that corresponds to the inverse of its f_0. For fibres stimulated by a single, resolved harmonic nf_0, h_i peaks at all integer multiples of $\frac{1}{nf_0}$, including $\frac{1}{f_0}$ itself. For high-frequency fibres stimulated by a number unresolved harmonics, the AN activity a_i follows essentially the envelope of the BM vibration (cf. sections 2.2.3.2 and 3.2.2), which is periodic at rate f_0 and therefore causes a peak in h_i at $\tau = \frac{1}{f_0}$. Critically, this autocorrelation-ridge is preserved for high-CF fibres independent of the presence of low-rank harmonics, providing a basis for unresolved missing-f_0 pitch that is missing from virtually all spectral models. Such a ridge is similarly generated by non-periodic stimuli such SAM noise, which also evokes periodic envelope modulations in high-frequency fibres (see Figure 2.24).

Licklider's original duplex theory does not specify *how* the ridge in $h_i(t,\tau)$ is to be identified by a subsequent pitch processor. Slaney and Lyon (1990) presented an implementation of Licklider's theory, where $h_i(t,\tau)$ (after an additional step of edge-sharpening) was summed across channels. Summation as a means for combining information across channels had already been proposed by van Norden (1982), with the slight difference that h_i in his model was based on first-order inter-spike-intervals (ISIs) rather than autocorrelation (which for a neural spike train is equivalent to computing the all-order ISI histogram; see also Cariani and Delgutte, 1996a). The summation approach was

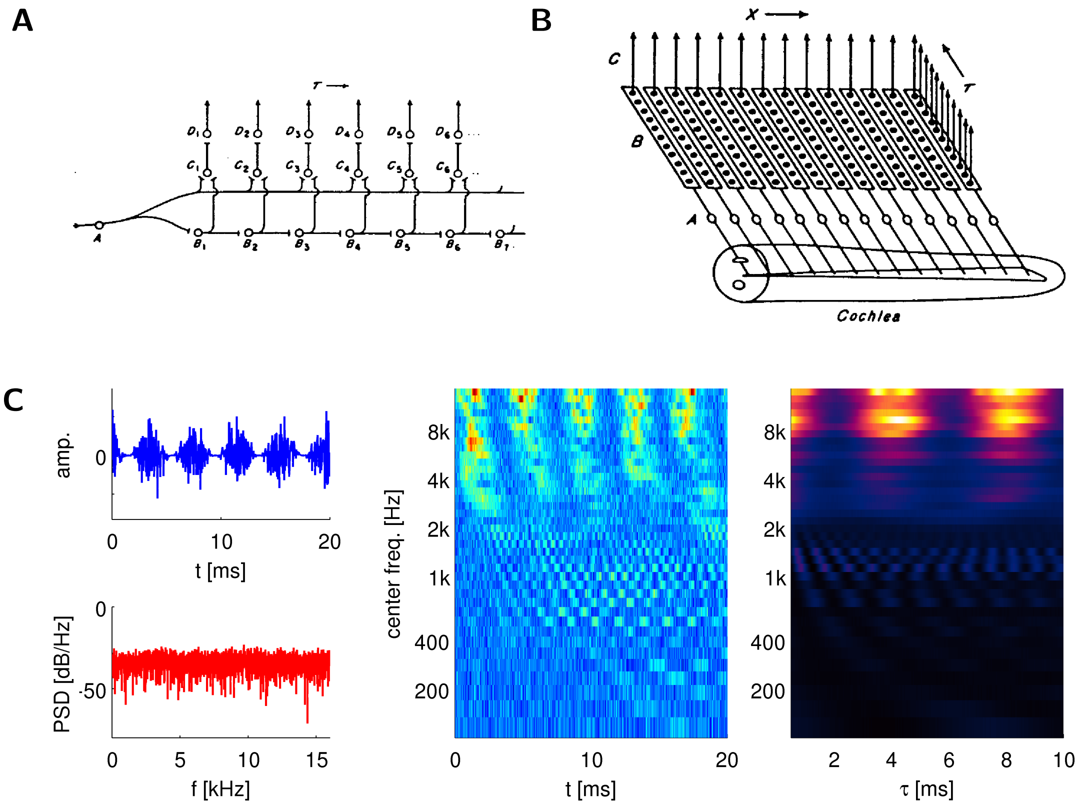


Figure 2.24: Licklider's duplex theory of pitch. A: Putative delay-line mechanism underlying the autocorrelation computation. Synaptic relays at neurons B_k cause increasing delays relative to the output of neuron A. Neurons C_k act as coincidence detectors and neurons D_k perform a leaky integration of their outputs (from Licklider, 1951). B: Spatial array of neurons along dimensions of peripheral CF ("x") and delay τ (ibid.). C: AN response (centre) to a 250 Hz SAM noise stimulus (top left: waveform, bottom left: power spectrum) and output of the autocorrelators at the end of the stimulus (right), showing peak in high-CF autocorrelators at multiples of the modulator periodicity of 4 ms.

adopted and popularised by Meddis and Hewitt (1991), who coined the term "summary autocorrelation function" (SACF) for the function

$$s(t,\tau) = \sum_i h_i(t,\tau) \quad . \tag{2.10}$$

There are numerous possible ways to "read out" the SACF, and defining an appropriate strategy is not a trivial task. One difficulty arises from the fact that s varies over time. To simplify the problem, one might decide to base the decision regarding the pitch of the sound solely on its shape at the time T of stimulus offset. Needless to say, information regarding the periodicity of the stimulus is lost by disregarding the time-course of s. Assuming that pitch is determined by the evaluation of s at a single point in time, we still need a decide which lag τ^* to report. A straightforward strategy is to pick the peak

$\tau^* = \operatorname{argmax}_\tau s(T, \tau)$ (constrained within certain bounds in order to avoid reporting a pitch close to $\tau = 0\,\mathrm{ms}$). Peak-picking disregards potentially valuable additional cues, such as the reoccurence of peaks in intervals of $\frac{1}{f_0}$ in the case of HCTs. For situations where pitch is measured by direct matching against a reference stimulus (rather than interval or melody recognition, for example), Meddis and Hewitt (1991) suggested to use the squared Euclidean distance

$$D^2 = \int (s_t(T_t, \tau) - s_r(T_r, \tau))^2 \, \mathrm{d}\tau \tag{2.11}$$

between the SACFs s_t and s_r of the target sound and the reference as a measure of their pitch similarity. Pitch matching is then achieved by finding the reference sound that minimises D^2. This method is not without pitfalls, however, as D^2 is highly sensitive not only to variations in the shape of s due to differences in pitch, but also to those effected by differences in level or spectral envelope between the two sounds.

Despite these concerns, Meddis and Hewitt (1991) demonstrated that the SACF model can, in principle, account for a substantial range of pitch-related phenomena. We have already discussed how missing-f_0 HCTs (including unresolved ones) and SAM noise can create ridges in Licklider's 2D-lattice of autocorrelators, and these ridges are preserved when summing across different fibres. In the case of SAM tones and shifted HCTs, unresolved components below the phase-locking limit in the auditory nerve give rise to multiple peaks in s in close correspondence to the pitch matches measured psychophysically (e.g. Schouten et al., 1962; Patterson and Wightman, 1976; see also section 2.1.2). SACF also explains the pitch of IRN, including the dependence of the pitch of IRN $n-$ on the number of iterations n (cf. section 2.1.2), which is not trivially explained by its power spectrum. For IRN $1-$, the delay-and-subtract process creates an *anti-correlation* between waveform amplitudes separated by the delay interval d. This anti-correlation is preserved in the temporal fine-structure of the lower-frequency AN fibres, causing a local minimum to occur in $s(T, \tau)$ at $\tau = d$. Meddis and O'Mard (1997) further showed that the SACF model is to some degree phase-sensitive. While the phases of resolved harmonics do not influence the position of the peak in s, phase relationships between unresolved harmonics that double the envelope modulation rate of their respective filter outputs can cause a doubling of the reported pitch (cf. Shackleton and Carlyon (1994) and section 2.1.1.2).

Based on these results alone, SACF explains a wider range of phenomena than any of the pattern matching models discussed in the previous section. Yet, the SACF model is not without issues either. As Houtsma and Smurzynski (1990) have shown (and others such as Shackleton and Carlyon (1994) and Bernstein and Oxenham (2003) have subsequently confirmed), the pitch of missing-f_0 HCTs is markedly stronger (subjectively and in terms of f_0-discriminability) when low-rank harmonics are present in the stimulus. Carlyon (1998) demonstrated that no such systematic effect is found in the SACF model (in its implementation by Meddis and O'Mard (1997)). The model showed very little degradation even when all harmonics were unresolved. The authors interpreted these results as evidence for the existence of two separate pitch processors: one that performs pattern-matching on the resolved spectral components of the stimulus and gives rise to a strong percept of pitch, and one based on periodicity analysis in the output of fibres stimulated by unresolved harmonics that provides a much weaker sense of pitch. We have already discussed one weakness of this argument (cf. section 2.1.1.2), namely that peripheral resolvability does not appear to be the primary cause for the weakening of the pitch percept. Furthermore, Bernstein and Oxenham (2005) presented a modification of the SACF model which addresses the criticism by Carlyon (1998) at least qualitatively. Instead of simply summing the individual channel-ACFs, a channel- and lag-dependent weight-matrix $W = \{w_{i,\tau}\}$ is introduced, and equation (2.10) is replaced by the weighted sum

$$\tilde{s}(t,\tau) = \sum_i w_{i,\tau}\, h_i(t,\tau) \tag{2.12}$$

Bernstein and Oxenham (2005) chose $w_{i,\tau}$ such that high-CF autocorrelators h_i would contribute little or nothing to $\tilde{s}$ at long time-lags, thereby lowering or abolishing the SACF peak at the fundamental of an unresolved missing-f_0 HCT.

Numerous other modifications have been proposed regarding details of the original SACF model by Meddis and Hewitt (1991). One recurrent issue concerns the choice of λ, the integration time constant of the autocorrelators. Licklider proposed values of $\lambda \approx 2\,\text{ms}$, which would makes the model insensitive to periodicity rates substantially lower than 500 Hz and allow for rapid fluctuation of the periodicity estimate. Meddis and O'Mard (1997) used a value of $\lambda = 10\,\text{ms}$ which is still somewhat short of the lower limit of pitch around 30 Hz (Pressnitzer et al., 2001). Wiegrebe (2001) proposed the use of CF-dependent integration time constants λ_i, such that low-frequency autocorrelators integrate over longer time windows. Balaguer-Ballester et al. (2008) recently presented a substantial extension, whereby the short-term SACF s, computed with $\lambda \approx 10\,\mathrm{ms}$, is itself integrated over time. Integration is leaky, as in the autocorrelators, but with considerably longer time constant λ_l (e.g. hundreds of milliseconds). The outcome of this integration is the low-pass filtered SACF (LP-SACF)

$$l(t,\tau) = \int_0^t s(t',\tau) \cdot \exp((t'-t)/\lambda_l) \quad . \tag{2.13}$$

Its benefit over the the unfiltered SACF s is that spurious, short-term peaks in s due to random fluctuation are evened out in l over longer durations, while at the same time maintaining the lower limit of period-sensitivity imposed by λ (i.e. the time constant of the autocorrelators). In addition to changes to the "central" component of the model, refinements have also been made to the peripheral front-end that feeds into the autocorrelators. The peripheral front-end used by Balaguer-Ballester et al. (2008), for example, is based on a biophysically detailed model of inner-hair cell (IHC) and AN responses (Sumner et al., 2002), including a non-linear, compressive cochlear filter bank, calcium-dynamics in the IHC soma and stochastic, calcium-dependent neurotransmitter-trafficking at the IHC synapse.

Licklider thought of autocorrelation as a physiological mechanism. While coincidence detection and integration are indeed two fundamental modes of neural computation, direct anatomical or physiological evidence for the type of autocorrelators required for a mechanistic interpretation of his theory has not been found. In particular, the existence of precisely-timed neural delay lines with delays up to 30 ms (in order to account for the lower limit of pitch) appears doubtful. De Cheveigné and Pressnitzer (2006) recently proposed an alternative potential mechanism whereby "synthetic" delays are generated through cross-channel phase interactions. This abolishes the need for neural delay lines, at least for resolved harmonics. Positive evidence for the use of such a mechanism by the auditory system however is also lacking.

Various models based on the temporal analysis of peripheral filters by means other than summary autocorrelation have been developed. Patterson et al. (1992) proposed the "auditory image model" (AIM) as an alternative central stimulus representation (see also Patterson, 2000). Like Licklider's lattice of correlators, the auditory image extends in dimensions of peripheral CF and time. Rather than tracking the short-term ACF however, running averages of "snapshots" of the AN activity within a certain time window are computed. This running average is not computed continuously (in which case the model would simply compute the peripherally resolved spectrum), but in a "strobed" fashion: snapshots are taken at discrete points in time, triggered independently in each channel by prominent peaks in the channel output. When the AN activity pattern across several channels is periodic, and strobed integration is repeatedly triggered in an approximately phase-locked manner, ridges along the CF dimension appear in the auditory image at times corresponding to multiples of the common periodicity. Just like Licklider's lattice by itself, however, the AIM is not a model of pitch perception but instead a particular representation of the stimulus and its recent past that may lend itself well to subsequent periodicity detection. A model that is formally closely related to autocorrelation is the "cancellation model" of pitch (de Cheveigné, 1993, 1998). Instead of the point-wise product $a_i(t) \cdot a_i(t-\tau)$, the squared difference $(a_i(t) - a_i(t-\tau))^2$ is integrated over time similar to equation (2.4). While the resultant integral is equivalent to h_i up to sign-reversal and a constant, the model also generates an output in which the stimulus periodicity is suppressed. Thus, it provides not only a means for estimating the stimulus periodicity, but also implements a harmonic cancellation filter which may provide a basis for segregating concurrent, periodic sounds. Cariani and Delgutte (1996a,b) used real data recorded from cat AN in place of a simulated, peripheral front-end as the basis for computing an aggregate all-order ISI histogram, closely related to the SACF (cf. also section 2.3.1). The authors found that the ISI histograms allowed for the reliable prediction of the pitch of missing-f_0 HCTs, SAM tones (ambiguity and shifts), SAM noise and others.

None of the models discussed in this section make use of the frequency label associated with each peripheral channel during periodicity estimation: the channel identity is entirely irrelevant during summation, except to a limited degree in the extensions by Wiegrebe (2001) and Bernstein and Oxenham (2005). Oxenham et al. (2004) designed a stimulus to specifically test whether the same might be the case in human listeners. They imposed three harmonically-related low-frequency modulators of 300, 400 and 500 Hz (i.e. with a missing modulator f_0 of 100 Hz) on high-frequency sinusoidal carriers, thereby generating what they called a "transposed complex tone" (cf. section 4.6). In the SACF model, these modulator give rise to a 100 Hz peak in the SACF during summation of the single-channel ACFs, similar to a missing-f_0 complex tone composed of the modulator frequencies. In contrast, human listeners did not perceive a pitch at 100 Hz, suggesting that the human auditory system may indeed be sensitive to channel identity when combining periodicity information across peripheral channels. Recently, Balaguer-Ballester et al. (2008) showed that the LP-SACF model (see above) correctly *fails* to predict at 100 Hz at sound intensities comparable to those used in the psychophysical experiments. For low stimulus intensities, however, the model still predicts a pitch of 100 Hz, suggesting that the reason for the absence of a pitch in the former case is likely due to their elaborate peripheral model (which may have been driven to saturation by the louder stimuli) rather than changes to the central pitch processor. So far, the prediction of level-dependence has not been tested psychophysically, and it is therefore not clear yet whether the effect reported by Oxenham et al. (2004) is indicative of a fundamental failure mode of the SACF and related models.

Chapter 3

A generative model of near-periodic sounds and auditory nerve responses

We have argued in the previous chapters, that near-periodic sounds are ubiquitous in our environment, and that their periodicity carries behaviourally relevant information: it aids in distinguishing predator from prey and in establishing the identity of objects and agents in our environment. During acoustic communication, periodicity is purposefully used as a means to convey information from sender to listener. Since periodicity and pitch are so closely related in the case of most natural, near-periodic sounds, it seems only appropriate to think of pitch perception as a process of periodicity estimation. Periodic stimuli evoke physiological responses in the peripheral auditory system which are themselves highly regular. These regularities manifest themselves in a number of different statistics commonly used to describe the peripheral response. Measuring the average firing rates of fibres in the auditory nerve, one hallmark of many periodic sounds is the occurrence of firing rate maxima in AN fibres with CFs harmonically related to the periodicity rate of the stimulus. Temporal fluctuations around these mean rates in each channel are also highly periodic, at rates indicative of the stimulus periodicity. Models of pitch perception typically operate on one or the other of these two cues, but not both. Spectral pattern matching models are clearly too limited in their scope to provide a general account of pitch perception on their own. Temporal models, such as summary autocorrelation, explain a wider range of phenomena, but have difficulty explaining why, on the whole, pitch is stronger for stimuli that pattern matching *can* account for. Some authors have therefore called for a "dual mechanism" explanation, whereby the output of two pitch processors operating on different periodicity cues is combined at a final stage that determines the overall percept (e.g. Carlyon and Shackleton, 1994; Carlyon, 1998; see de Cheveigné, 2005 for a discussion of potential pitfalls).

From a normative standpoint, we agree that disregarding one or the other of two seemingly non-redundant cues is wasteful and suboptimal, if accurate estimation is to be achieved. Our aim here is to develop a model which uses information about the stimulus periodicity, contained in the evoked peripheral response, as efficiently as possible. The problem of determining the stimulus periodicity based only on indirect observations seems fundamentally ill-posed: as sound waves travel through air from their source to the ear of a listener, they are perturbed by noise, mixed with other sounds, filtered and reflected in ways that can only appear as stochastic to an observer. A further source of variability is the process of sensorineural transduction itself. Hence, an observer cannot establish the exact identity of the stimulus or its periodicity with absolute certainty. Von Helmholtz (1867), considering similarly ill-posed problems in human vision, developed an influential theory of perception based around this premise. Evolving a school of thought that dates back to 11th century astronomer and mathematician Alhazen, he posited that our percepts arise from a process of *unconscious inference*, whereby incoming nervous sensations are combined with prior knowledge and memories in order to form an estimate of their most likely underlying cause in the external world: "The general rule according to which visual representations determine themselves is that we always find present in the visual field such objects as would have to exist in order for them to produce the same impression on the neural apparatus under the usual normal conditions of the use of our eyes" (von Helmholtz, 1867, in a translation by Westheimer, 2008).

Bayesian probability theory (e.g. Jaynes, 2003; MacKay, 2003) provides a formal framework that allows for optimal statistical inference about the hidden causes underlying a set of stochastic observations, provided that the observer is given knowledge about their statistical regularities and interdependencies. "Ideal observer" models of this kind have become valuable tools in the study of human perception (see e.g. Knill and Richards, 1996; Kersten et al., 2004; Shams and Beierholm, 2010) and have been successful in explaining psychophysical performance limits and perceptual illusions not only in the visual domain (e.g. Weiss et al., 2002), but also in cross-model perception (Ernst and Banks, 2002; Wozny et al., 2010) and the perception of time (Ahrens and Sahani, 2011). In this chapter, we will present an ideal observer model of pitch as an optimal estimate of the unobserved periodicity of a noise-corrupted periodic sound observed indirectly through noisy evoked responses in the auditory nerve.

The statistical dependencies between the unobserved periodicity and the observed AN response is expressed in terms of a generative, statistical model (Figure 3.1). The generative process is divided into two phases, sound generation (section 3.1) and sensorineural transduction (section 3.2). The outcome of the first phase is a waveform $\boldsymbol{x}$, drawn from a distribution over all possible waveforms with period duration Ω and corrupted by noise. In the second phase, the sound evokes a time-varying neural response $A = \{a_i(t)\}$ in peripheral frequency channels $i = 1 \dots C$. Combined, the two stages of the model define a probability distribution $\mathrm{P}(A \mid \Omega)$ over auditory nerve responses to sounds with periodicity Ω. Using approximate Bayesian inference techniques (section 3.3), we can evaluate the posterior distribution $\mathrm{P}(\Omega \mid A^*)$ over periodicities given a particular, observed auditory nerve response pattern A^*, on the basis of which an estimate of Ω is formed.

3.1 Sound generation

Our generative model of acoustic waveforms takes inspiration from the true generative mechanism underlying a ubiquitous type of pitch-evoking sound in our natural environment: voiced speech. According to the source-filter theory of speech production (Fant, 1960), vowel sounds are generated acoustically through the excitation of resonances in the vocal tract (comprising the laryngeal, pharyngeal, nasal and oral cavities) by sharp, periodic air puffs emitted from the vocal source, the larynx. During vocal production, muscles inside the larynx contract, causing the vocal folds to obstruct the tracheal air flow. Air pressure builds up at the vocal folds until reaching a threshold, upon which the folds open and the pressure is released eruptively, following which the cycle repeats (see Figure 3.2).

To a first approximation, the air pressure waveform emitted by the vocal source can be characterised as a regular pulse train $\boldsymbol{\delta}_\Omega$ with an inter-pulse interval Ω, while the vocal

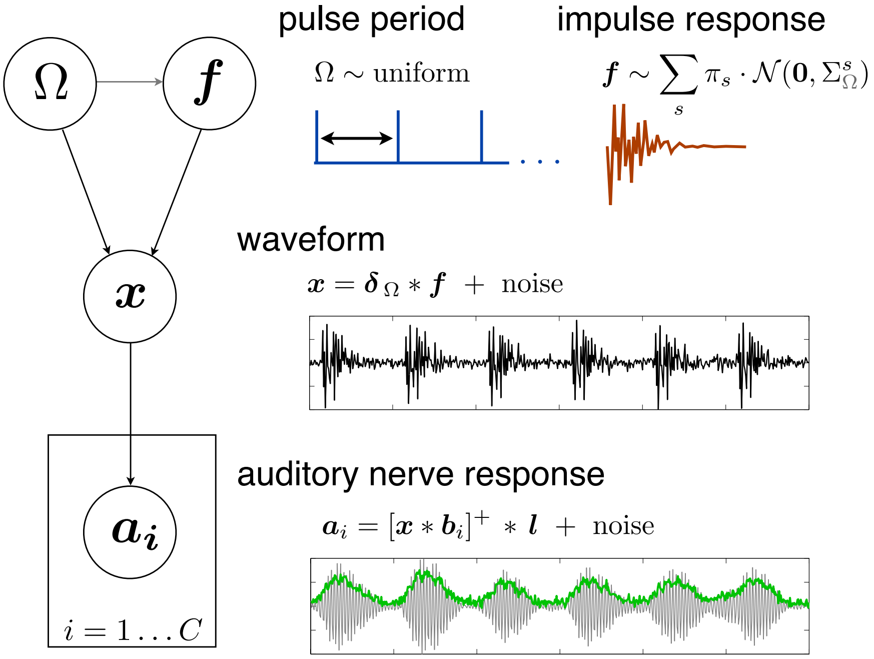

pulse period
$\Omega \sim$ uniform
impulse response
$\boldsymbol{f} \sim \sum_s \pi_s \cdot \mathcal{N}(\mathbf{0}, \Sigma_\Omega^s)$
Ω
$\boldsymbol{f}$
waveform
$\boldsymbol{x} = \boldsymbol{\delta}_\Omega * \boldsymbol{f}\ +$ noise
$\boldsymbol{x}$
auditory nerve response
$\boldsymbol{a}_i = [\boldsymbol{x} * \boldsymbol{b}_i]^+ * \boldsymbol{l}\ +$ noise
$\boldsymbol{a_i}$
$i = 1 \ldots C$

Figure 3.1: A generative model of naturalistic, approximately periodic sounds and evoked auditory nerve responses. A pulse train with period Ω is convolved with a randomly-generated acoustic impulse response $\boldsymbol{f}$ and corrupted by additive noise to obtain an acoustic waveform $\boldsymbol{x}$. Two variants of the model are distinguished by the presence or absence of a dependency between Ω and $\boldsymbol{f}$. x evokes responses in auditory nerve fibres $i = 1 \ldots C$ as follows: in each channel, the waveform is filtered by a linear bandpass filter with impulse response $\boldsymbol{b}_i$. 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.

tract acts on this pulse train as a linear filter, the impulse response of which we will denote as $\boldsymbol{f}$. Different vowel sounds are produced by changing the shape of the vocal tract, thereby affecting its preferred resonance frequencies and hence the shape of $\boldsymbol{f}$, while the stereotypical pulse-shape of the vocal source remains unaffected. Comparable mechanisms underlie vocal production not only in other mammals and birds, but even in certain amphibians and fish (Patterson et al., 2007), even though the organs used and sound sources and resonance chambers will of course differ. The sounds of many musical instruments are similarly produced following the pulse-resonance principle (van Dinther and Patterson, 2006; Gough, 2007). In brass instruments, for example, the pulses are generated through the regularly intermittent obstruction of the air flow into the instrument by the player's tensed lips. In many wind instruments, wooden reeds fulfil the exact same purpose, whereas in string instruments, the hairs of the bow pull a string away from its straight resting position until the tension becomes so great that it snaps back in an abrupt, saw-tooth like motion (Gough, 2007). Thus, the

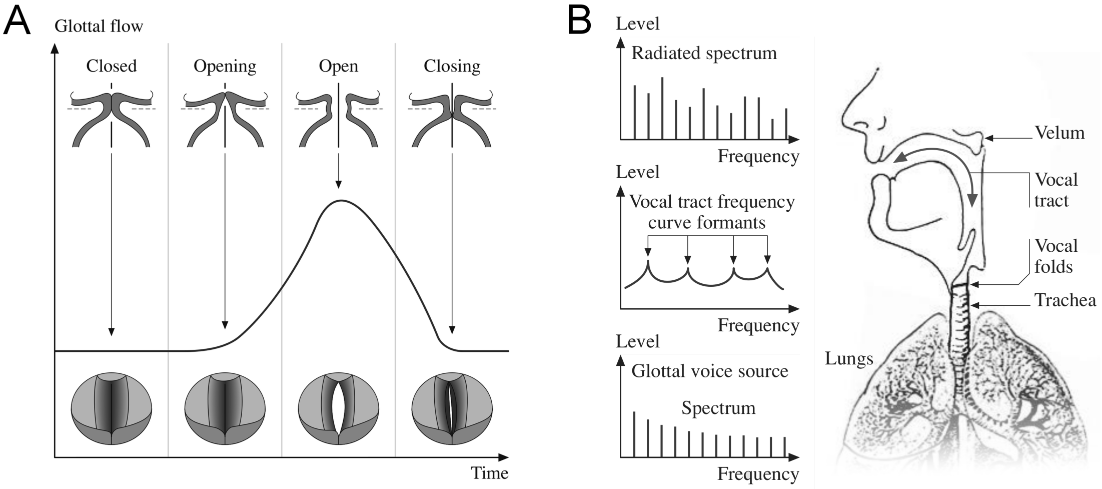


Figure 3.2: Schematic of human vocal production. A: Glottal air-pressure waveform generated by the periodic opening and closing of the vocal folds. B: The spectrum of the emitted waveform (top) is the product of the glottal source spectrum (bottom) and the vocal tract resonance spectrum (middle). Adapted from Lindblom and Sundberg (2007).

pulse-resonance principle of sound generation underlies a wide range of natural, pitched sounds. But more than that: *any* periodic waveform can be decomposed into a series of pulses and a stereotypical impulse-response, regardless of whether it was physically generated by a pulse-resonance-type mechanism or not.

In our generative model, we treat $\boldsymbol{\delta}_\Omega$ as a sum of δ-pulses, regularly spaced at a time interval Ω:

$$\delta_\Omega(t) = \sum_{k=0}^{\infty} \delta(t - k \cdot \Omega) \tag{3.1}$$

The Fourier transform of this equation is a flat, even-spaced comb-spectrum with a lowest spectral peak at $f_0 = \Omega^{-1}$, the fundamental frequency. Any truly periodic sound $\boldsymbol{x}$ can thus be formed by the convolution of $\boldsymbol{\delta}_\Omega$ with an appropriately chosen impulse response $\boldsymbol{f}$. As natural sounds are hardly ever perfectly periodic, we include an uncorrelated, isotropic Gaussian noise term in our generative equation to allow for random perturbations, resulting in an approximately periodic waveform

$$\boldsymbol{x} = \boldsymbol{f} * \boldsymbol{\delta}_\Omega + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma_x^2 \, \mathbb{I}^{T \times T}) \tag{3.2}$$

of length T. Here and throughout the following, $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$, covariance matrix Σ and probability density function $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \det^{-1}(2\pi\Sigma) \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^{\intercal}\right)$. $\mathbb{I}^{T \times T}$ denotes the T-by-T identity matrix.

Equation (3.2) defines a probability distribution over sound waveforms, *given* an underlying pulse-rate Ω and acoustic impulse response $\boldsymbol{f}$. In order to obtain a full model of the *marginal* distribution over sounds, considering all possible values of Ω and $\boldsymbol{f}$, we will need to specify prior distributions over these two variables.

In our model, Ω is drawn uniformly from some range $[\Omega_{\text{min}}; \Omega_{\text{max}}]$. This is arguably an unrealistic simplification, when compared to the distribution of fundamental frequencies f_0 encountered in our environment. For example, the natural distribution of f_0s in human speech is clustered broadly around approximately 100 Hz and 200 Hz, corresponding to the natural speaking ranges of adult male and female speakers (Simpson, 2009). As we will see later, our prior distribution in the model would have to vary over many orders of magnitude in order to substantially affect the outcome of its pitch estimation. Hence, while it would be conceptually simple to replace the uniform prior with a more naturalistic distribution, we think our choice is not crucial to the model's success or failure.

Regarding the prior distribution over impulse responses $\boldsymbol{f}$, we will discuss two alternatives in the following. First, we will present a basic formulation of the model, in the following referred to as the "uncoupled model", in which the spectro-temporal characteristics of $\boldsymbol{f}$ are *independent* of the periodicity Ω (section 3.1.1). We will refer to this as the "uncoupled" model in paragraphs and chapters to come. In section 3.1.2, we will present a general framework for introducing a statistical dependency between Ω and $\boldsymbol{f}$ into the model. We will investigate the nature of this dependency in the "coupled model" later on in chapter 5.

3.1.1 Uncoupled model

The intuition behind our prior distribution over $\boldsymbol{f}$ in the uncoupled model is simple: we require the amplitude envelope of a typical draw of $\boldsymbol{f}$ to decay monotonically after initial excitation with some arbitrary, unknown rate, while making no explicit assumption about the temporal fine-structure of $\boldsymbol{f}$ underneath this envelope. In the model, this is implemented as a multivariate mixture of Gaussian distributions (MoG), in which each mixture component (labelled by s) is associated with a particular time scale τ_s over which the impulse envelope decays. The mean of each mixture component is a vector of zeros: we do not expect the *average* impulse response — across all possible sounds —

to have a particular positive or negative amplitude at any point in time. What remains yet to be specified, is a covariance matrix Ψ^s for each component. Its diagonal elements, $\Psi^s(t,t)$ control the average *squared* amplitude of $\boldsymbol{f}$ at any time-point t. Hence, we can enforce the decay of the impulse envelope by letting the diagonal elements of Ψ^s drop off — in our specific case with a squared-exponential time course. The non-diagonal elements of $\Psi^s(t,t)$ control the mutual dependency of pairs of the elements of $\boldsymbol{f}$ at different points in time, such as for example the degree of smoothness in $\boldsymbol{f}$. At this point, we will limit ourselves to diagonal covariance matrices Ψ^s: the elements of $\boldsymbol{f}$ are entirely uncorrelated in the generative process, allowing for arbitrarily irregular shapes of the temporal fine structure within the decaying envelope. We will, however, consider an extended, more realistic model that allows for the flexible control over the covariance structure of $\boldsymbol{f}$ in section 3.1.2. As a final detail, we allow for a temporal delay ϕ_s in the onset of each impulse response. This detail may not seem particularly relevant for sound generation itself. It will turn out to be rather important when we invert the model to perform inference about the periodicity Ω of a sound (i.e. when we estimate its pitch), the waveform of which does not necessarily peak at $t = 0$: up to the resolution set by the modeller in choosing the range of ϕ_s in the mixture distribution, this allows for an optimal alignment of the pulse train $\boldsymbol{f}$ to the sound waveform. The entries of the mixture covariance matrices are thus given by

$$\Psi^s(t,t') = 0 \quad \text{if } t \neq t' \tag{3.3}$$

$$\Psi^s(t,t) = \begin{cases} 0 & \text{if } t < \phi_s \\ \exp(-\frac{(t-\phi_s)^2}{2\tau_s^2}) & \text{otherwise} \end{cases} . \tag{3.4}$$

By setting the non-diagonal elements of Ψ^s to zero, the elements of $\boldsymbol{f}$ are temporally uncorrelated. Hence, the average power spectrum of repeated draws of $\boldsymbol{f}$ is therefore white, i.e. broadband and flat (cf. Figure 3.4 for examples).

Following our description above, the full set of generative equations for the uncoupled model reads as follows:

$$\Omega \sim \text{uniform}([\Omega_{\min}; \Omega_{\max}]) \tag{3.5}$$

$$\boldsymbol{f} \sim \sum_{s=1}^{S} \pi_s \cdot \mathcal{N}(\mathbf{0}, \Psi^s) \tag{3.6}$$

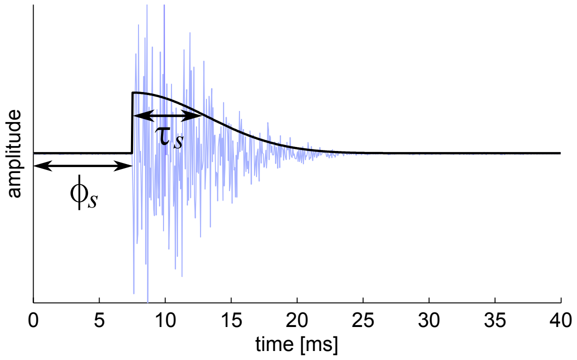


Figure 3.3: Diagonal elements of the covariance matrix Ψ^s of a single mixture component (black) and representative draw $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \Psi^s)$ (blue).

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \, \mathbb{I}^{T \times T}) \tag{3.7}$$

$$\boldsymbol{x} = \boldsymbol{\delta}_\Omega * \boldsymbol{f} + \boldsymbol{\eta} \tag{3.8}$$

with $\boldsymbol{f} \in \mathbb{R}^{1 \times M}$, $\Psi^s \in \mathbb{R}^{M \times M}$ and $\boldsymbol{x} \in \mathbb{R}^{1 \times T}$. For the sake of simplicity, we let $\pi_s = \frac{1}{S}$, i.e. all combinations of time-scales τ_s and delays ϕ_s are *a priori* equally likely.

We can write down an analytical expression for $\mathrm{P}(\boldsymbol{x} \,|\, \Omega)$, the distribution over waveforms with periodicity Ω, marginalised across all possible draws of $\boldsymbol{f}$. In order to arrive at this expression, we observe that we can rewrite the convolution of $\boldsymbol{x}$ with $\boldsymbol{\delta}_\Omega$ equivalently as a matrix multiplication,

$$\boldsymbol{\delta}_\Omega * \boldsymbol{f} = \boldsymbol{f} \cdot \Delta_\Omega \quad , \tag{3.9}$$

where Δ_Ω is the convolution matrix of $\boldsymbol{\delta}_\Omega$, i.e. an $M \times T$ Toeplitz matrix with progressively time-shifted copies of $\boldsymbol{\delta}_\Omega$ as its rows. Since $\mathrm{P}(\boldsymbol{f} \,|\, \Omega)$ is a MoG distribution, and since the Gaussian family of distributions is closed under linear transformation and addition, $\mathrm{P}(\boldsymbol{x} \,|\, \Omega)$ is similarly a MoG distribution. The covariance matrices Σ^s of its mixture components are obtained from $\mathrm{P}(\boldsymbol{f} \,|\, \Omega)$ by linear transformation with Δ_Ω and subsequent addition of isotropic Gaussian noise:

$$\boldsymbol{x} \sim \sum_{s=1}^{S} \pi_s \cdot \mathcal{N}(\mathbf{0}, \Sigma^s) \tag{3.10}$$

$$\Sigma^s = \Delta_\Omega^\intercal \Psi^s \Delta_\Omega + \sigma_x^2 \cdot \mathbb{I} \tag{3.11}$$

Figure 3.4 shows draws from the model for different settings of τ_s. The temporal extent

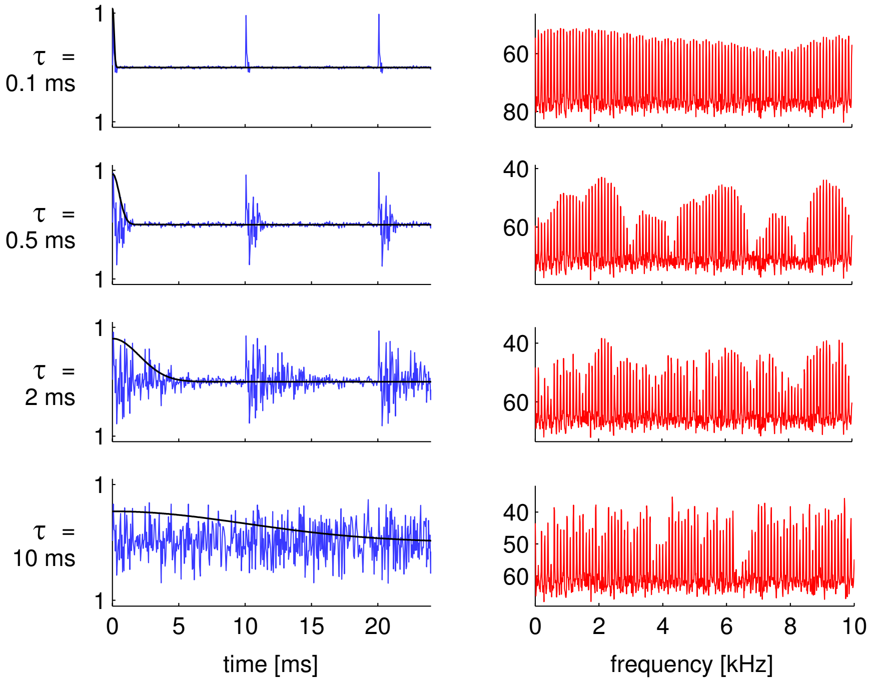


Figure 3.4: Influence of the envelope time constant τ_s on draws from the model. Left: Sound waveforms drawn from the model for different values of τ_s (Ω = 10 ms, identical random seed, low noise); black curves depict the diagonal entries of Ψ^s. Right: Corresponding amplitude spectra on logarithmic scale with arbitrary reference. Note the effect of τ_s on spectral scale of smoothness.

of the impulse response increases from a sharp pulse for $\tau_s = 0.1$ ms (top left) to almost flat within a single pulse interval for $\tau_s = 10$ ms (bottom left). Its amplitude envelope is jittered around the diagonal entries of Ψ^s (black curves). The spectral envelope of $\boldsymbol{\delta}_\Omega$ is flat, as is the expected spectrum of $\boldsymbol{f}$ (owing to the mutual statistical independence of elements of $\boldsymbol{f}$). Therefore, the spectral envelope of $\boldsymbol{x}$, being the product of the two, must also be flat in expectation. Nevertheless, the choice of τ_s has an interesting effect on the spectra of *individual* draws of $\boldsymbol{x}$ (Figure 3.4, right): as τ_s increases, the spectral scale of smoothness decreases. This phenomenon is easily understood for the extreme corner-cases $\tau_s \to 0$ and $\tau_s \to \infty$ (cf. top and bottom panels). As $\tau_s \to 0$, the shape of the temporal envelope of $\boldsymbol{f}$ approaches a perfect δ pulse, the spectrum of which is both deterministic and perfectly flat. Hence, $\boldsymbol{x}$ will similarly have a perfectly flat, i.e. smooth, envelope. At the other extreme, as $\tau_s \to \infty$, draws of $\boldsymbol{f}$ approach the statistics of stationary, white Gaussian noise. The Fourier amplitudes of such noise are individually Raleigh-distributed, and mutually independent (e.g. Hartmann, 1997) — in other terms: maximally unsmooth. For intermediate values of τ_s, the smoothness of the power spectrum of $\boldsymbol{f}$ is likewise intermediate between these two extremes (cf.

panels for $\tau_s = 0.5$ and 2 ms in Figure 3.4). We can also think of this as a reflection of the duality of the Fourier transform: multiplication with a fast-decaying envelope in the time domain corresponds to a convolution of the spectrum with a broad (i.e. low-pass) filter. As our model contains mixture components across a wide range of τ_s, the model is equally likely to produce sounds of any degree of spectral smoothness during sound generation, and indifferent towards this spectral feature during inference.

3.1.2 Coupled model

The model presented in the previous section features independent priors over periodicity and impulse response, the latter of which effectively governs the perceived timbre of a sound. As we will discuss at length in chapter 5, it is reasonable to assume that these two variables should in fact be linked, based on both acoustic and psychophysical evidence. We will demonstrate at this point how our present formalism can be easily extended to capture such statistical dependencies while retaining the structure of the generative equations (3.5) . . . (3.8).

The key step is to replace the unconditional prior (3.6) over $\boldsymbol{f}$ by the conditional distribution

$$\boldsymbol{f} \,|\, \Omega \sim \sum_{s=1}^{S} \pi_s \cdot \mathcal{N}(\mathbf{0}, \Psi^s_\Omega) \quad , \tag{3.12}$$

where the covariance matrices Ψ^s_Ω now depend on the periodicity Ω, rather than using a fixed set of matrices Ψ^s for all periodicities as was the case in the previous section. While these covariance matrices Ψ^s_Ω could depend on Ω in any arbitrary way, we will consider here a special case, albeit a powerful and flexible one. Specifically, we restrict Ψ^s_Ω to the class of covariance matrices that correspond to a linear filter $\boldsymbol{h}_\Omega$ being applied to a draw from $\mathcal{N}(\mathbf{0}, \Psi^s)$, where Ψ^s is a diagonal matrix as defined in (3.4). In terms of the source-filter model of speech production, this corresponds to allowing for a systematic variation of the vocal tract shape (and thus its filtering properties) with fundamental frequency. If we let $\widetilde{\boldsymbol{f}}$ denote the initially drawn impulse response prior to filtering, $\boldsymbol{h}_\Omega$ the filter kernel, and $\boldsymbol{f}$ the final impulse response, we can describe the generative process in two steps as

$$\widetilde{\boldsymbol{f}} \sim \mathcal{N}(\mathbf{0}, \Psi^s) \tag{3.13}$$

$$\boldsymbol{f} = \widetilde{\boldsymbol{f}} * \boldsymbol{h}_\Omega \tag{3.14}$$

Alternatively, we do not need to introduce the auxiliary variable $\widetilde{\boldsymbol{f}}$ at all. We can rewrite the convolution $\widetilde{\boldsymbol{f}} * \boldsymbol{h}_\Omega$ as a multiplication of $\widetilde{\boldsymbol{f}}$ with a Toeplitz matrix H_Ω which implements the exact same linear operation (cf. equation (3.9), where we used the same trick):

$$\boldsymbol{f} = \widetilde{\boldsymbol{f}} \cdot H_\Omega \tag{3.15}$$

As the family of Gaussian distributions is closed under linear transformation, $\boldsymbol{f}$ is also Gaussian-distributed and its covariance matrix Ψ^s_Ω is given by

$$\Psi^s_\Omega = H^\mathsf{T}_\Omega \Psi^s H_\Omega \tag{3.16}$$

More convenient still, as the two-sided multiplication with H_Ω corresponds to sequentially filtering the rows and columns of Ψ^s with $\boldsymbol{h}_\Omega$, equation (3.16) is equivalent to a convolution of Ψ^s with a two-dimensional kernel, which is given by the outer product of $\boldsymbol{h}_\Omega$ with itself: $\Psi^s_\Omega = \Psi^s \underset{\text{2d}}{*} (\boldsymbol{h}^\mathsf{T}_\Omega \cdot \boldsymbol{h}_\Omega)$.

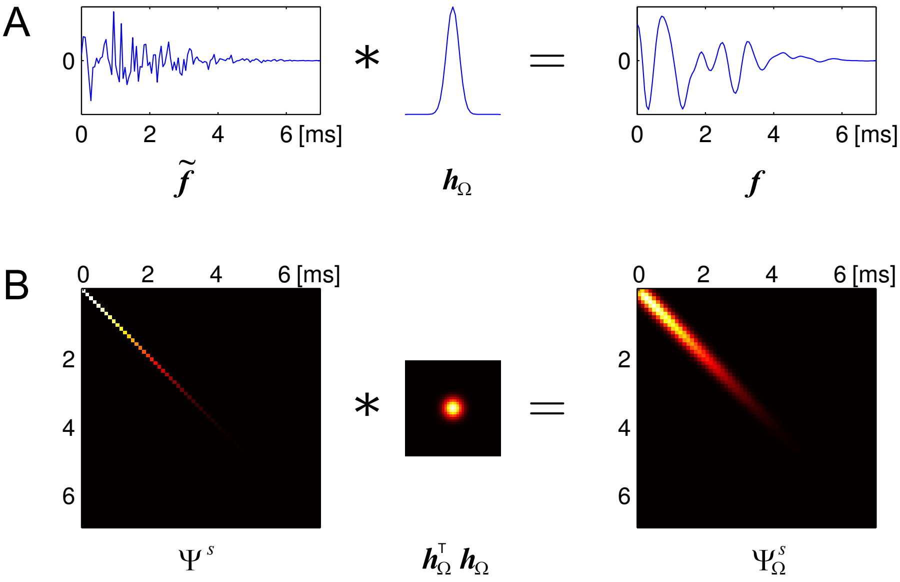


Figure 3.5: Coupling of Ω and $\boldsymbol{f}$ through a linear filter $\boldsymbol{h}_\Omega$. A: $\widetilde{\boldsymbol{f}}$ is drawn from a Gaussian distribution with diagonal covariance Ψ^s and convolved with $\boldsymbol{h}_\Omega$ to obtain $\boldsymbol{f}$. B: The effect of this convolution on the covariance structure of $\boldsymbol{f}$, compared to that of $\widetilde{\boldsymbol{f}}$, is equivalent to convolving the Ψ^s with the outer product of $\boldsymbol{h}_\Omega$ with itself.

In Figure 3.5, we have chosen a Gaussian-shaped kernel for illustrative purposes, but note that $\boldsymbol{h}_\Omega$ could in principle have any arbitrary shape. At this point we will leave the specific shape of $\boldsymbol{h}_\Omega$ and its dependence on Ω unspecified. We will revisit the issue of defining a suitable coupling between Ω and $\boldsymbol{f}$ in chapter 5.

Whatever the exact dependence of Ψ^s_Ω on Ω may be, we can still write down the marginal distribution of the acoustic waveform $\boldsymbol{x}$ given a particular value Ω in general terms (where we marginalise with respect to $\boldsymbol{f}$), according to our derivation of equation (3.10):

$$\boldsymbol{x} \,|\, \Omega \sim \sum_{s=1}^{S} \pi_s \cdot \mathcal{N}(\mathbf{0}, \Sigma^s_\Omega) \tag{3.17}$$

$$\Sigma^s_\Omega = \Delta_\Omega \Psi^s_\Omega \Delta^\intercal_\Omega + \sigma^2_x \cdot \mathbb{I} \tag{3.18}$$

3.2 Transduction

The process of sensorineural transduction was described in considerable detail earlier (cf. 2.2), and will be summarised here only briefly. During acoustic stimulation, the inner-ear fluids are set into motion by the middle ear ossicles, causing the basilar membrane (BM) to vibrate. To a first approximation, the basilar membrane performs a spatial frequency analysis of the sound by means of a gradient of preferred resonance frequencies along its axis. The passive, approximately linear response of the BM is considerably modulated by an active component, largely due to the electromotile feedback of the outer hair cells (OHCs). Motion of the basilar membrane at a certain point along its axis results in an alternating de- and repolarisation of the inner hair cells (IHCs), which in turn initiate the generation of action potentials in the fibres of the auditory nerve (AN) during phases of depolarisation. For high oscillation rates of the BM, the IHC potential ceases to follow individual phases of the BM response. Instead, it fluctuates only little around an elevated baseline. As a result, the IHC potential (and consequently the firing rates in the auditory nerve) reflects the amplitude *envelope* of the basilar membrane response for high frequencies (progressively above 1-2 kHz), rather than the raw amplitude waveform (cf. Figure 2.12).

In the model, we restrict ourself to a rather schematic, functional characterisation of these highly complex biophysical processes. The basilar membrane will be modelled as a linear system, much like envisioned by von Helmholtz (1863), von Békésy (1960) and others before the discovery of the active contribution of the OHCs to the BM response. We will also not attempt to model the progressive loss of temporal fine structure in the IHCs in biophysical detail: instead, amplitude demodulation of the BM response will be achieved using a simple but effective signal-processing technique for envelope extraction.

3.2.1 Basilar membrane response

We want to model the BM response as the outputs $\boldsymbol{c}_i$ of a bank of linear band-pass filters $\boldsymbol{b}_i$, where each filter corresponds to a different place along its axis:

$$\boldsymbol{c}_i = \boldsymbol{x} * \boldsymbol{b}_i \quad \forall\, i = 1 \ldots C \quad . \tag{3.19}$$

A widely-applied type of filter for linear characterisations of the basilar membrane is the gammatone filter (de Boer and de Jongh, 1978). The impulse response of a gammatone filter is the product of a sine-wave carrier with an envelope shaped like a the integrand in the definition of the gamma function. Several parameters govern the shape of a gammatone filter: the carrier has an amplitude a_i, frequency f_i and phase θ_i, while the gamma envelope is characterised by a filter order n (integer and typically constant across all filters) and its bandwidth β_i. It impulse response is given by

$$b_i(t) = a_i\, t^{n-1} e^{-2\pi\beta_i t} \;\cos(2\pi f_i\, t + \theta_i) \quad . \tag{3.20}$$

Johannesma (1972) proposed the gammatone impulse response, albeit not under its present-day name, as an analytic description for linear filters he derived from the responses of neurons in cat cochlear nucleus using reverse correlation (de Boer and Kuyper, 1968; see also de Boer and de Jongh, 1978; Aertsen and Johannesma, 1980; de Boer and Kruidenier, 1990 for subsequent uses of the gammatone filter to describe physiological reverse-correlation filters). Importantly, gammatone filters are not only suitable for the characterisation of peripheral neural responses in animals, but have also been found to provide excellent fits to psychophysical estimates of human peripheral auditory filters (Patterson and Moore, 1986; Glasberg and Moore, 1990), given an appropriate choice of the filter order n and the relationship between frequency f_i and bandwidth β_i (Patterson et al., 1992). The "equivalent rectangular bandwidth"

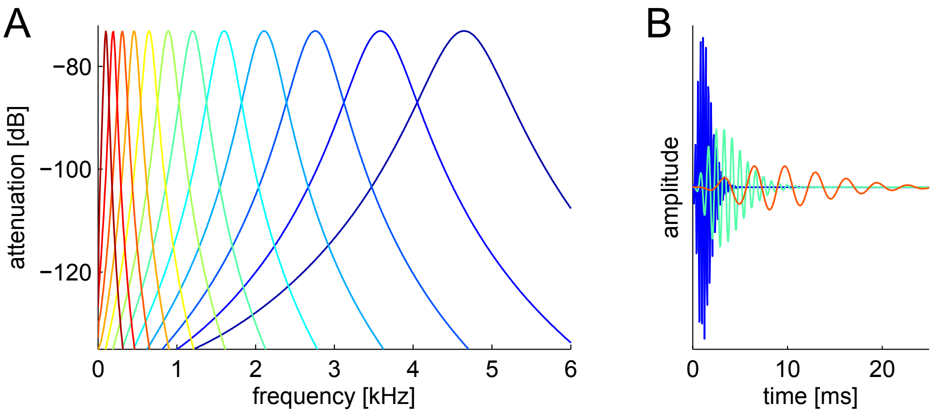


Figure 3.6: Gammatone filter bank. A: Spectral magnitude response of 12 gammatone filters as used in the model. CF spacing and filter bandwidths grow in proportion to the human ERB scale (cf. equation (3.21)) B: Impulse response of three filters (colours match A).

(ERB) of human auditory filters approximately depends on their centre frequencies f as summarised by Glasberg and Moore (1990):

$$\mathrm{ERB}(f) = 24.7\,(0.00437\,f + 1) \quad . \tag{3.21}$$

Following Patterson et al. (1992) and Slaney (1993) in our implementation, we choose $\beta_i = 1.019\,\mathrm{ERB}(f_i)$, $n = 4$, $\theta_i = 0$, and a spacing of neighbouring carrier frequencies f_i that is proportional to their bandwidths. The gains are chosen so as to give the same peak attenuation around f_i for all filters.

Gamma-tone filter banks of this kind are commonly used as the initial peripheral processing stage in models of human auditory perception (e.g. Meddis and Hewitt, 1991; Patterson et al., 1995; Meddis and O'Mard, 1997; Wiegrebe, 2001; Bernstein and Oxenham, 2003). Interestingly, gammatone filter banks seem to be particularly well-suited for the efficient coding of natural sounds. Smith and Lewicki (2006) used a sparse-coding model to learn an optimal basis for the representation of natural sounds as well as speech (Smith and Lewicki, 2005). They reported that the learnt basis vectors bore a striking similarity to the impulse responses of gammatone filters, suggesting that the process of human vocal production on the one hand, and the peripheral neural encoding of speech and other natural sounds on the other are almost ideally co-adapted to each other.

There is an ongoing debate as to whether the ERB scale according to (Glasberg and

Moore, 1990) (equation (3.21)) provides an adequate estimate of human peripheral filter bandwidths. As physiological data regarding the tuning properties of the active cochlea or single auditory nerve fibres is not available for humans, bandwidth estimates are based on psychophysical measurements instead. In a classical paradigm, the effect of a notched-noise masker on the detectability of a simultaneous target tone at the centre of the spectral notch in measured (Patterson, 1976; Glasberg and Moore, 1990). Central to this kind of bandwidth estimation is the assumption that energy in the noise masker impairs the detection of the target only if it falls within the bandwidth of the same peripheral auditory filter, but not outside. Shifting the masker passbands symmetrically away from the target tone, thereby widening the notch while keeping the masker energy constant, one will find that the detection threshold of the tone improves and then saturates, once the notch exceeds a certain bandwidth. From the progression of these threshold improvements, approximate filter shapes and bandwidths can be estimated. The human ERB scale summarises these bandwidth estimates, which are thought to reflect the frequency selectivity of the cochlea. Recently, Shera et al. (2002) and Oxenham and Shera (2003) have challenged the established ERB scale: using a forward-masking paradigm instead of simultaneous masking, and low sound intensities in order to maximise the occurrence of active, non-linear effects in the basilar membrane response (cf. section 2.2.3.3), Oxenham and Shera (2003) arrived at bandwidth estimates up to two times sharper than previously reported. However, Ruggero and Temchin (2005) showed that a behavioural forward-masking paradigm lead to an *overestimation* of the sharpness of cochlear tuning in several species, for which direct physiological measurements of cochlear frequency selectivity are available, suggesting that the same may be the case in humans. Notwithstanding the possibility of future revisions, we chose to adopt the traditional ERB scale in our model. In our own psychophysical experiments (chapter 5), presentation levels around 75 dB SPL were used, at which point active amplification and sharpening of the basilar membrane response can be expected to be substantially reduced compared to the near-threshold levels used in Oxenham and Shera (2003), even if the concerns of Ruggero and Temchin (2005) turn out to be unfounded.

3.2.2 Auditory nerve response

We emulate the demodulation-like behaviour of the IHCs by a simple envelope extraction method, consisting of the half-wave rectification (HWR) of the BM response, followed by low-pass filtering at a frequency where phase-locking of the IHC declines (HWR-LP). If we let $\mathbf{r}(\cdot)$ denote the element-wise rectification of a vector, and $\boldsymbol{l}$ the impulse response of our low-pass filter, we obtain the output firing rate $\boldsymbol{a}_i$ of a single auditory nerve fibre in the model as:

$$\boldsymbol{a}_i = \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l} + \boldsymbol{\xi}_i, \quad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \sigma_A^2\, \mathbb{I}) \qquad . \tag{3.22}$$

Owing to the addition of the isotropic Gaussian noise term $\boldsymbol{\xi}_i$ to every channel, the distribution over activity patterns $A = \{\boldsymbol{a}_i\}_{i=1\ldots C}$, conditioned on the output of the BM filters, is itself Gaussian with a non-linear dependency between the BM output and its mean according to equation (3.22):

$$\mathrm{P}(A \,|\, \boldsymbol{x}) = \prod_{i=1}^{C} \mathcal{N}\big(\boldsymbol{a}_i \,|\, \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l},\, \sigma_A^2\; \mathbb{I}\big) \tag{3.23}$$

Strict half-wave rectification ($\mathrm{r}\,(z) = \max(0, z)$) introduces discontinuities in the derivatives of r at $z = 0$. In order to enable us to draw on the family of gradient- and Hessian-based methods for inference (cf. section 3.3), we avoid these discontinuities by applying a “soft” rectification function r :

$$\mathrm{r}\,(z) = \frac{\log(1 + \exp(\alpha z))}{\alpha} \qquad . \tag{3.24}$$

The strictness of r is controlled by the parameter α. For $\alpha \to \infty$, $\mathrm{r}\,(z) \to \max(0, z)$.

From a technical point of view, several heuristics other than our HWR-LP approach are in common use for the extraction of a slow-moving envelope from an amplitude-modulated signal (see e.g. Turner and Sahani, 2011). For example, full-wave rectification ($\mathrm{r}\,(z) = |z|$) or squaring can similarly be used prior to low-pass filtering. Another common approach, mathematically more principled, is to compute the amplitude of the analytic signal, obtained via a Hilbert transform of the waveform (e.g. Hartmann, 1997). All these methods have their specific drawbacks, and it is hence *a priori* difficult to choose one over the others from a signal-processing standpoint alone. Considering the underlying physiological process in the IHCs (cf. section 2.2.3.2) and the HWR-like response behaviour that is visible in their responses to low-frequency BM vibration (Figure 2.12), however, the HWR-LP method appears to be preferable to the alternatives outlined above in the context of our model.

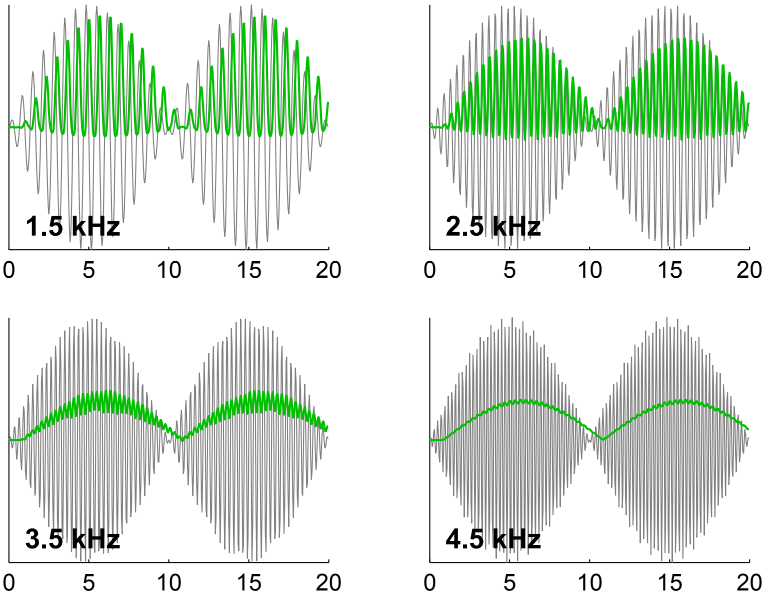


Figure 3.7: Progressive amplitude demodulation of high-frequency oscillations by the auditory nerve model. Waveforms prior to demodulation (grey) are pure tones, sinusoidally ring-modulated at 50 Hz. After half-wave rectification and low-pass filtering (green), fine structure is is progressively lost while the shape of the waveform envelope remains recognisable. The magnitude response of the low-pass filter dropped by 80 dB between 1.5 and 4.5 kHz, no noise was added to the output.

The frequency range over which AN fibres gradually cease to follow fine-structure peaks of the BM response varies from species to species and is not known from physiological measurements in humans. In guinea pigs, phase locking (measured as the ratio between modulation depth and average excitation level in the evoked IHC potential) starts to decline at around 600 Hz and is no longer detectable at around 3.5 kHz (Russell and Sellick, 1983). Other mammalian species, such as cats and squirrel monkeys, show near-constant phase locking for frequencies up to 2 kHz, which then declines over a range of 1 to 1.5 octaves (Rose et al., 1967; Johnson, 1980). In humans, the marked loss of pitch discriminability and musical interval and melody identification for tones with fundamental frequencies above 4 to 5 kHz has been interpreted as a reflection of the physiological upper limits of phase locking, but direct evidence is lacking. For the purpose of demodulation in our model, we truncated the impulse response of an ideal low-pass filter to obtain a FIR low-pass filter with a magnitude drop-off of 80 dB between 1.5 and 4.5 kHz using a Kaiser window (see Oppenheim et al., 1999, chapt. 7) without specific fine tuning of the filter characteristics.

In summary, the sensorineural transduction stage of our model is highly schematic compared to the complexity of the true biophysical mechanisms, but nevertheless captures two essentials aspects of the human peripheral auditory system. Mimicking cochlear frequency analysis, sounds are filtered into different frequency bands, the bandwidths of which grow approximately linearly with their centre frequencies. The ability of model auditory nerve fibres to follow the output of these filters phase-by-phase gradually declines between 1.5 kHz and 4.5 kHz, giving rise to envelop demodulation in frequency channels within and above this range. In using a linear filter bank, we ignore the active, OHC-driven component of the physiological BM response, which may be an acceptable simplification when considering its response to moderately high sound levels (cf. section 2.2.3.3), where OHC electromotility is greatly reduced compared to low levels. We also assume that the magnitude of the IHC potential is in essence linearly related to the BM response magnitude. We could, in principle, include some form of instantaneous compression into the model, for example by letting r saturate for high input value, but did not explore this possibility[1].

3.3 Inference

We have, so far, defined a generative statistical model of approximately periodic sounds and the resultant evoked firing rate patterns in the auditory nerve. At its core, the model comprises equations (3.10) and (3.22) in the uncoupled case, where sound periodicity and acoustic impulse response are treated as independent variables, and equations (3.17) and (3.22) in the coupled case, were a statistical dependence between the two variables is introduced by means of a periodicity-dependent filter $\boldsymbol{h}_\Omega$, modifying the shape of the acoustic impulse response $\boldsymbol{f}$ prior to its convolution with the pulse train $\boldsymbol{\delta}_\Omega$. Acoustic waveforms generated in this way are passed through a simple model of the peripheral auditory system up to the auditory nerve (AN), the response of which is again stochastic.

For every value of Ω our generative model implicitly defines a marginal distribution

[1] Since our algorithms for performing inference in the generative model (see section 3.3) require the computation of the first and second derivatives of r, these derivatives need to remain well-behaved unless new inference schemes were to be developed in addition.

over A:

$$\mathrm{P}(A \mid \Omega) = \int \mathrm{P}(A, \boldsymbol{x} \mid \Omega)\, \mathrm{d}\boldsymbol{x} \tag{3.25}$$

$$= \int \mathrm{P}(A|\boldsymbol{x}) \cdot \mathrm{P}(\boldsymbol{x}|\Omega)\, \mathrm{d}\boldsymbol{x} \tag{3.26}$$

Computing $\mathrm{P}(A \mid \Omega)$ for an *observed* AN activity pattern A is the basis for performing inference about Ω our model. In this section, we will discuss ways of performing this non-trivial operation of Bayesian model inversion.

The essential obstacle in computing $\mathrm{P}(A \mid \Omega)$ according to equation (3.25) is the high dimensionality of the unobserved latent variable $\boldsymbol{x}$, the acoustic waveform, in conjunction with its non-linear transformation during the transduction-stage of the model (equation (3.22)). Under our prior, $\boldsymbol{x}$ itself is distributed according to a mixture of high-dimensional Gaussian distributions (equations (3.10) and (3.17)). If our peripheral model were fully linear, $\mathrm{P}(A \mid \Omega)$ would similarly be MoG distributed. Albeit computationally challenging, a wide range of practical inference techniques has been developed for this case, based on the framework of Gaussian Processes (GPs; Rasmussen and Williams, 2006; Rasmussen and Nickisch, 2010). Non-linear generalisations of GPs exist, for which inference is also tractable. In a Warped Gaussian Process (Snelson et al., 2004) for example, observations are assumed to be non-linearly transformed draws from a GP. However, while Warped GPs allow for an arbitrarily high-dimensional latent variable space, they are restricted to one-dimensional observation spaces and furthermore require the warping function to map onto the *entire* real line, which is at odds with our requirements for $\mathrm{r}(\cdot)$ to perform some form of rectification on its inputs (in order for amplitude modulation to occur).

As we could not readily apply existing inference procedures "off the shelf" for use with our model, we developed two different algorithms of our own. One is based on the Laplace approximation (e.g. MacKay, 2003), a deterministic Gaussian approximation to the integral $\int \mathrm{P}(A, \boldsymbol{x} \mid \Omega)\, \mathrm{d}\boldsymbol{x}$ around its mode with respect to $\boldsymbol{x}$ (section 3.3.1). The second uses a combination of Markov Chain Monte Carlo (MCMC) sampling techniques (Neal, 1993, 1998, 2010) that allow us to approximate the same integral stochastically, albeit at a much higher computational cost (section 3.3.1). Both approaches rely on the gradient $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$ of the log-joint distribution over observed and latent variables, and the Laplace approximation furthermore requires the computation of its

Hessian matrix, $\nabla^2_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$. The — somewhat tedious — formal derivation of the gradient and Hessian is given in Appendix A.

Assuming that we have a practicable method of computing $\mathrm{P}(A \,|\, \Omega)$, how can we use it to perform inference about Ω, in other words: how can we predict pitch? Suppose we are given an observation A: the firing rate pattern across the entire auditory nerve in response to an arbitrary, unknown stimulus. We assume that the auditory system, and particular the "pitch processor", treats A *as if* evoked by an approximately periodic sound as described by our generative equations, even though the true, physical generative process may have been different.

Ideally then, we would like to compute the full *posterior* distribution

$$\mathrm{P}(\Omega \,|\, A) = \frac{\mathrm{P}(A \,|\, \Omega)\, \mathrm{P}(\Omega)}{\int \mathrm{P}(A \,|\, \Omega')\, \mathrm{P}(\Omega')\, \mathrm{d}\Omega'}, \tag{3.27}$$

where $\mathrm{P}(A \,|\, \Omega) =: \mathcal{L}(\Omega)$ is called the *likelihood* of Ω and $\mathrm{P}(\Omega)$ is the prior distribution over periodicities in our generative model. Given the posterior, a reasonable and intuitive estimate of the periodicity underlying A is the MAP (*maximum a posteriori*) estimate

$$\Omega^* = \underset{\Omega}{\operatorname{argmax}}\ \mathrm{P}(\Omega \,|\, A) \quad , \tag{3.28}$$

i.e. the most likely periodicity to have caused the observed AN activity pattern A by means of some unobserved acoustic waveform.

It is worth noting at this point, that what constitutes the optimal, or "rational", choice in the more general context of Bayesian decision theory (rather than that of pure perceptual inference) depends on the subjective cost function of the listener. This cost function may vary across individuals and depend on the task at hand. The MAP estimate is the optimal choice when the cost function is "all-or-none", i.e. when any failure to infer the true underlying periodicity is deemed equally unfavourable. For an arbitrary cost function $C(\widetilde{\Omega}; \Omega)$ that specifies the subjective cost of the observer for reporting a periodicity of $\widetilde{\Omega}$ when the true periodicity is Ω, the rational estimate (in the formal sense) is obtained by minimising the expected cost under the posterior distribution: $\Omega^* = \operatorname{argmin}_{\widetilde{\Omega}} \int C(\widetilde{\Omega}; \Omega)\, \mathrm{P}(\Omega \,|\, A)\, \mathrm{d}\Omega$ (see e.g. Mamassian et al., 2002; Shams and Beierholm, 2010). Another widely-used cost function is the squared error

$C(\tilde{x}; x) = (\tilde{x} - x)^2$, which is minimised in expectation by the *posterior* mean $\Omega \cdot \mathrm{P}(\Omega \,|\, A)$. A different strategy, which seems to provide a good description of human behaviour in some perceptual decision making tasks such as audio-visual target localisation (Pick et al., 1969; Wozny et al., 2010), is to draw a random sample from the posterior distribution on each trial rather than making a deterministic choice on the basis of the posterior distribution (such as the MAP or posterior mean). Following the sampling strategy, a subject's average decision (i.e. the mean response over many repeats of the same stimulus) will minimise the squared-error cost. Nevertheless, the sampled decisions will in general be suboptimal on any given single trial and vary across multiple repeats of the same stimulus. For our purpose of predicting pitch in the absence of an experimenter-controlled cost function, we will use the MAP estimate (equation (3.28)) throughout the remainder of this thesis, i.e. the single most probable period duration given all sensory evidence and prior expectations. Owing to the uniform prior over Ω, this is equivalent to the *maximum likelihood* (ML) estimate in our particular case. In principle, however, our Bayesian framework allows not only for optimal perceptual inference about Ω, but also for the prediction of optimal behaviour in more complicated situations, where for example the cost function may be shaped by the experimenter by rewarding or punishing listeners (human and animal) depending on their response to the stimulus. In this case, one may expect to find behavioural biases in the listeners' decisions, which are not inherent in the posterior distribution itself.

In practice, we will evaluate the posterior only at a set of selected candidate periodicities $\{\Omega_1, \ldots, \Omega_N\}$, the range and sampling density of which will depend on the stimulus at hand and the question we want to ask. Rather than covering the entire range of human hearing, for example, we may choose $\{\Omega_1, \ldots, \Omega_N\}$ to lie within a 1- to 2-octave range around the true fundamental of a harmonic sound, with quarter-tone intervals between neighbouring periodicities (i.e. 24 values of Ω_n per octave in that case). As we have chosen a flat prior $\mathrm{P}(\Omega)$ in the model (cf. equation (3.5)), we can thus simply return the ML (*maximum likelihood*) estimate Ω_{n^*} amongst the candidates, where

$$n^* = \operatorname*{argmax}_n \mathcal{L}(\Omega_n). \tag{3.29}$$

We will show examples of the log-likelihood profiles $\{\ln \mathcal{L}(\Omega_1), \ldots, \ln \mathcal{L}(\Omega_N)\}$ and ML periodicity estimates for a variety of pitch-evoking sounds in chapter 4. For the remainder of the current chapter, we will investigate two alternative methods for approximating the likelihood $\ln \mathcal{L}(\Omega) = \mathrm{P}(A \mid \Omega)$.

3.3.1 Laplace approximation

As laid out in the section above, we need to compute the integral

$$\mathrm{P}(A \mid \Omega) = \int \mathrm{P}(A, \boldsymbol{x} \mid \Omega)\, \mathrm{d}\boldsymbol{x} \tag{3.30}$$

$$\tag{3.31}$$

where the dimensionality T of the integrand $\boldsymbol{x} \in \mathbb{R}^T$ is high (for example, $T = 1000$ for a 50 ms waveform segment, sampled at a resolution of 20 kHz). Following Laplace's method (MacKay, 2003), the intractable integral $\int \mathrm{P}(A, \boldsymbol{x} \mid \Omega)\, \mathrm{d}\boldsymbol{x}$ is approximated by a Gaussian integral instead. To find the (unnormalised) approximating distribution $\mathrm{Q}(\boldsymbol{x})$, we Taylor-expand the *log*-joint distribution $\ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$ up to second order around its mode $\boldsymbol{x}^* = \mathrm{argmax}_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$, yielding

$$\ln \mathrm{Q}(\boldsymbol{x}) = \ln \mathrm{P}(A, \boldsymbol{x}^* \mid \Omega) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^{\top} H_{\boldsymbol{x}^*} (\boldsymbol{x} - \boldsymbol{x}^*). \tag{3.32}$$

Here, $H_{\boldsymbol{x}^*}$ denotes the Hessian matrix of $-\ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$, evaluated at $\boldsymbol{x}^*$:

$$H_{\boldsymbol{x}^*} = -\nabla^2 \ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)\Big|_{\boldsymbol{x}=\boldsymbol{x}^*} \quad . \tag{3.33}$$

In other words, the covariance matrix of the approximating Gaussian Q is given by the negative Hessian of the log-joint distribution at its mode. Finally, our estimate of $\mathcal{L}(\Omega)$ is obtained as the integral of Q over $\boldsymbol{x}$:

$$\int \mathrm{Q}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathrm{P}(A, \boldsymbol{x}^* \mid \Omega) \sqrt{\frac{(2\pi)^T}{\det H_{\boldsymbol{x}^*}}} \quad . \tag{3.34}$$

Note that in practice, we will compute the log-likelihood $\ln \mathcal{L}(\Omega)$ instead due, to numerical scaling issues.

In order to compute the Laplace approximation, we need to find the modal waveform $\boldsymbol{x}^* = \mathrm{argmax}_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$, which we achieve by performing gradient ascent in $\boldsymbol{x}$ using the L-BFGS method (Limited-memory Broyden-Fletcher-Goldfarb-Shanno; see

Nocedal and Wright, 1999)[2]. A derivation of the required gradient, as well as the Hessian, of $\ln \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$ is given in Appendix A.

Algorithm 3.1 Laplace approximation for $\ln \mathrm{P}(A \,|\, \Omega)$

Input: observed AN activity A; periodicity Ω; initial condition $\boldsymbol{x}_0$

```
    $\boldsymbol{x}^* \leftarrow \mathrm{LBFGS}(A, \Omega, \boldsymbol{x}_0)$
    $H_{\boldsymbol{x}^*} \leftarrow -\nabla^2 \ln \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)\Big|_{\boldsymbol{x}=\boldsymbol{x}^*}$
    return $\ln \mathrm{P}(A, \boldsymbol{x}^* \,|\, \Omega) + \frac{T}{2}\ln(2\pi) - \frac{1}{2}\operatorname{logdet}(H_{vx^*})$
```

3.3.2 Hamiltonian Annealed Importance Sampling

Naively, one might attempt to evaluate $\mathcal{L}(\Omega)$ by simple Monte Carlo sampling (e.g MacKay, 2003), drawing samples of $\boldsymbol{x}$ from the model prior $\mathrm{P}(\boldsymbol{x} \,|\, \Omega)$ and evaluating them under $\mathrm{P}(A \,|\, \boldsymbol{x})$:

$$\begin{aligned}
\mathcal{L}(\Omega) &= \mathrm{P}(A \,|\, \Omega) && (3.35)\\
&= \int \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)\,\mathrm{d}\boldsymbol{x} && (3.36)\\
&= \int \mathrm{P}(A \,|\, \boldsymbol{x})\,\mathrm{P}(\boldsymbol{x} \,|\, \Omega)\,\mathrm{d}\boldsymbol{x} && (3.37)\\
&\underset{N\to\infty}{\approx} \frac{1}{N}\sum_{i=1}^{N} P(A \,|\, \boldsymbol{x}_i)\ , \quad \boldsymbol{x}_i \sim \mathrm{P}(\boldsymbol{x} \,|\, \Omega) && (3.38)
\end{aligned}$$

Despite being unbiased, however, this estimator is not useful in practice. We can view equation (3.38) above as a particular case of Importance Sampling. In order to simplify notation, we will denote $p(\boldsymbol{x}) := \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$ in the following, and view p as an unnormalised distribution over $\boldsymbol{x}$: the "target distribution". Its normalising constant $\mathcal{Z}_p = \int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is thus equal to $L(\Omega)$, i.e. the quantity we want to estimate. Importance Sampling allows us, in principle, to estimate $\mathcal{Z}_p$ by sampling from some tractable "proposal distribution" $q(\boldsymbol{x})$:

$$\begin{aligned}
\mathcal{Z}_p &= \int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} && (3.39)\\
&= \int \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x}) && (3.40)
\end{aligned}$$

[2] We used a Matlab implementation by Mark Schmid, minFunc.m, available from `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`.

$$\approx \frac{1}{N}\sum_{i=1}^{N}\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}\ , \quad \boldsymbol{x}_i \sim q(\boldsymbol{x}) \tag{3.41}$$

Equation (3.38) is the special case where the proposal distribution equals our model prior: $q(\boldsymbol{x}) = \mathrm{P}(\boldsymbol{x}\,|\,\Omega)$. Unfortunately, Importance Sampling suffers from a well-known deficiency when applied to high-dimensional variable spaces, despite being asymptotically unbiased (e.g Neal, 1993, 1998; MacKay, 2003). Most of the mass of the integral $\int p(\boldsymbol{x})\mathrm{d}x$ is associated with values of $\boldsymbol{x}$ that are typical under the *target* distribution p. If the overlapping volume between the proposal and target distribution is small, the time required to obtain even few representative samples (if any) is prohibitively large, unless the dimensionality of $\boldsymbol{x}$ is low, and p and q concentrate their mass in the same region of the variable space (see also Minka, 2005). As a consequence, Importance Sampling estimates of $\mathcal{Z}_p$ (i.e. $\mathcal{L}(\Omega)$) would suffer from unacceptably large variance in our case. Other methods (e.g. Rejection Sampling) based on drawing independent samples from q fail for the same reasons.

Markov Chain Monte Carlo methods (Neal, 1993; Murray, 2007) — such as Metropolis-Hastings, Gibbs sampling, or Hamiltonian Monte Carlo — provide an alternative way of obtaining samples from p (albeit not independent). A stochastic transition operator $T(\boldsymbol{x} \to \boldsymbol{x}')$ is repeatedly applied to an initial sample $\boldsymbol{x}_1 \sim q$, yielding a sequence of further samples $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. T is carefully designed such that it has p as its stationary equilibrium distribution:

$$\int T(\boldsymbol{x} \to \boldsymbol{x}')p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \ = p(\boldsymbol{x}') \quad . \tag{3.42}$$

After an initial "burn-in" phase, this procedure will generate dependent samples from the target distribution: in effect, T gradually bridges the distributions q and p, despite the fact that their divergence may be large (assuming that they share at least some support). Much current effort is being spent on finding efficient transition operators T, that require only short burn-in periods and produce sequential samples as independently as possible. Frustratingly however, even estimators of $\mathcal{Z}_p$ based on true samples from p (e.g. Newton and Raftery, 1994), obtained for example via MCMC sampling, are often badly behaved (see Murray, 2007 for discussion).

Annealed Importance Sampling (AIS; Neal, 1998) is one of only a small number of techniques, that have been shown to allow the estimation of $\mathcal{Z}_p$ in high-dimensional variable spaces for intractable distributions p, while keeping the variance of the estimates under control at least asymptotically. In AIS, the transition operator T is not fixed during a sampling run, but instead forms *itself* a chain $T_1, \ldots, T_N$. By convention, the $\boldsymbol{x}_n$ in AIS are sampled "backwards", starting with $\boldsymbol{x}_N \sim q$ and then according to

$$\boldsymbol{x}_i \,|\, \boldsymbol{x}_{i+1} \sim T_i(\boldsymbol{x}_{i+1} \to \cdot) \quad i = N-1, \ldots, 0 \tag{3.43}$$

Each T_i has a different equilibrium distribution q_i, that morphs between $q = q_N$ and $p = q_0$ to a steadily increasing degree:

$$q_i = q^{1-\beta_i}\, p^{\beta_i} \quad 1 = \beta_0 > \beta_1 > \ldots > \beta_N = 0 \tag{3.44}$$

At the end of a single AIS run comprising samples $\boldsymbol{x}_N, \ldots, \boldsymbol{x}_0$, we obtain an "importance weight"

$$w = \prod_{i=1}^{N} \frac{q_{i-1}(x_{i-1})}{q_i(x_{i-1})} \quad . \tag{3.45}$$

If we repeat this procedure multiple times, each iteration j yielding an importance weight w_j, we finally obtain an estimate of $\mathcal{Z}_p$:

$$\widetilde{\mathcal{Z}}_p = \frac{1}{S} \sum_{j=1}^{S} w_j \quad , \tag{3.46}$$

which will provably converge to the true $\mathcal{Z}_p$ as $S \to \infty$ (see Neal, 1998).

Aside from requiring a potentially large number of intermediate steps N in each run, the success or failure of AIS (as with any other MCMC method) depends largely the ability of transition operators T_i to mix quickly, i.e. to generate near-independent samples after only a short time. We implemented each T_i as Hamiltonian Monte Carlo (HMC) sampler, an MCMC method that uses gradient information about q_i in order to explore the variable space more efficiently than simple, random-walk-like procedure. (Neal, 1993, 2010). Hamiltonian Monte Carlo generates samples from a target distribution (in our case, $q_i(\boldsymbol{x})$ as in equation (3.44)) by simulating a Hamiltonian dynamical system with random initial conditions: a particle with initial position $\boldsymbol{x}^0$ and initial momentum $\boldsymbol{u}^0$ moves through a landscape defined by the potential-energy function $E(\boldsymbol{x}) = -\ln(q_i(\boldsymbol{x}))$ while maintaining a constant total energy $H(\boldsymbol{x}, \boldsymbol{u}) = E(\boldsymbol{x}) + K(\boldsymbol{u})$, where $K(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}\boldsymbol{u}^\intercal$ is its kinetic energy[3]. At any point in (continuous) time, the instantaneous change in position of the particle is given by $\nabla_{\boldsymbol{x}} E$, i.e. the negative gradient of the log-target distribution, while the momentum $\boldsymbol{u}$ changes as $\nabla_{\boldsymbol{u}} K = \boldsymbol{u}$. In the computer simulation of this system, the dynamics are discretised using the "leap-frogging" procedure, whereby the variables are updated in each step as follows (note the half-steps in $\boldsymbol{u}$):

$$\boldsymbol{u}^{t+1/2} = \boldsymbol{u}^t - \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} E(\boldsymbol{x}^t) \tag{3.47}$$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \epsilon \boldsymbol{u}^{t+1/2} \tag{3.48}$$

$$\boldsymbol{u}^{t+1} = \boldsymbol{u}^{t+1/2} - \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} E(\boldsymbol{x}^{t+1}) \tag{3.49}$$

Leap-frogging yields smaller discretisation errors than the simple Euler method, resulting in more stable dynamics and ultimately more efficient sampling (see below). After letting the dynamical system evolve for some pre-determined number of steps L, the final position $\boldsymbol{x}^L$ is stochastically accepted or rejected as a new sample in the Markov chain, depending on a draw of $\xi \sim$ uniform([0; 1]):

$$\boldsymbol{x}^* = \begin{cases} \boldsymbol{x}^L & \text{if } \ln \xi < H(\boldsymbol{x}^0) - H(\boldsymbol{x}^L) \quad \text{(accept)} \\ \boldsymbol{x}^0 & \text{otherwise} \quad \text{(reject)} \end{cases} \tag{3.50}$$

The final acceptance/rejection-step, called a Metropolis-Hastings update (Hastings, 1970), ensures that q_i is indeed the stationary distribution of the HMC sampler, and is necessary to compensate for numerical errors due to the discretised dynamics. Observing that the position variable $\boldsymbol{x}$ tends to get repeatedly drawn towards *low* points in the potential-energy landscape in this dynamical system[4], we get an intuition for why the imaginary particle preferentially explores high-probability regions of our target distribution $q_i(\boldsymbol{x}) = \exp(-E(\boldsymbol{x}))$, and hence why HMC is likely to return representative samples of q_i.

In order to use this approach then, we need to be able to evaluate $\nabla_{\boldsymbol{x}} E = -\nabla_{\boldsymbol{x}} \log q_i$.

[3] For $\boldsymbol{x} \in \mathbb{R}^2$ for example, this corresponds to the idealised path of a ball on a frictionless surface, the height profile of which is defined by $-\ln q_i$

[4] Note though, that the particle will never *rest* there, owing to its kinetic energy that increases by the same amount by which its potential energy decreases.

Recalling equation (3.44), we see that

$$\log q_i(\boldsymbol{x}) = (1-\beta_i)\ln q(\boldsymbol{x}) + \beta_i \ln p(\boldsymbol{x}) \tag{3.51}$$

$$= (1-\beta_i)\ln \mathrm{P}(\boldsymbol{x}\,|\,\Omega) + \beta_i \ln \mathrm{P}(A,\boldsymbol{x}\,|\,\Omega) \tag{3.52}$$

$$= \ln \mathrm{P}(\boldsymbol{x}\,|\,\Omega) + \beta_i \ln \mathrm{P}(A\,|\,\boldsymbol{x}) \quad , \tag{3.53}$$

the gradients of which we have already derived for use in our Laplace-based inference scheme (cf. section 3.3.1 and Appendix A). The entire combined HMC/AIS sampling algorithm (HAIS) is given below (Algorithm 3.2). Following a recommendation by Neal (1998), we chose the annealing schedule $\beta_1 > \ldots > \beta_N = 0$ to be essentially log-linearly spaced, with a short linearly ramping segment from $\beta_N = 0$ to $\beta_{0.9N} = 0.01$, i.e. for the first 10% of AIS samples (see also Berkes et al., 2008).

Algorithm 3.2 Hamiltonian AIS for $\ln \mathrm{P}(A\,|\,\Omega)$

```
Input: observation $A$; periodicity $\Omega$; annealing schedule $\beta_N, \ldots, \beta_0$
  let $q := \ln \mathrm{P}(\boldsymbol{x}\,|\,\Omega)$,  $p := \ln \mathrm{P}(A,\boldsymbol{x}\,|\,\Omega)$,  $q_i := q^{1-\beta_i}\, p^{\beta_i}$
  for $j = 1$ to $S$ do
      draw $\boldsymbol{x}_N \sim q_N$
      for $k = N-1$ down to $0$ do
          draw $\boldsymbol{u}^o \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$
          $\boldsymbol{x}^o \leftarrow \boldsymbol{x}_{k+1}$
          for $l = 1$ to $L$ do
              $\boldsymbol{u}^l \leftarrow \boldsymbol{u}^{l-1} + \frac{\epsilon}{2}\nabla \log q_k(\boldsymbol{x}^{l-1})$
              $\boldsymbol{x}^l \leftarrow \boldsymbol{x}^{l-1} + \epsilon \boldsymbol{u}^l$
              $\boldsymbol{u}^l \leftarrow \boldsymbol{u}^l + \frac{\epsilon}{2}\nabla \log q_k(\boldsymbol{x}^l)$
          end for $\{l\}$
          draw $\xi \sim$ uniform([0; 1])
          if $\log(\xi) < \log q_k(\boldsymbol{x}^L) - \log q_k(\boldsymbol{x}^o) - \frac{1}{2}(\langle \boldsymbol{u}^L, \boldsymbol{u}^L\rangle - \langle \boldsymbol{u}^o, \boldsymbol{u}^o\rangle)$ then
              $\boldsymbol{x}_k \leftarrow \boldsymbol{x}^L$ {accept update}
          else
              $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k+1}$ {reject update}
          end if
      end for $\{k\}$
      $\hat{w}_j \leftarrow \sum_{i=1}^{N}(\beta_{i-1} - \beta_i)\log p(A|\boldsymbol{x}_{i-1})$
  end for $\{j\}$
  $\log \hat{\mathcal{Z}}_p \leftarrow \log \sum_j \exp(\hat{w}_j) - \log S$
  return  $\log \hat{\mathcal{Z}}_f$
```

3.4 Summary

In this chapter, we have presented a generative, statistical model of approximately periodic sounds and subsequently evoked firing rates in the auditory nerve. The acoustic component of our model is a probabilistic extension of the source-filter model of speech production (Fant, 1960). A key variable is the unobserved periodicity Ω of the sound prior to it being corrupted by noise (and further transformed into neural activity). In short, our model defines probability distributions over observed auditory nerve responses for each possible such periodicity. The second, essential variable for this purpose is the acoustic impulse response $\boldsymbol{f}$ of the sound source, which determines the shape of the sound waveform within a period of Ω (note though that the impulse response can last longer than the period duration). We presented two alternatives for characterising the distribution over impulse responses $\boldsymbol{f}$. Both assume that the envelope shape of $\boldsymbol{f}$ can be described as an initial excitation which subsequently decays with an unknown, arbitrary time constant. This assumption is implemented by drawing $\boldsymbol{f}$ from a mixture distribution, where each mixture component corresponds to one of many different time constants. The two alternatives models differ in their assumptions regarding the fine structure of $\boldsymbol{f}$ underneath the decaying envelope. In the *uncoupled* model (section 3.1.1), $\boldsymbol{f}$ is temporally uncorrelated and independent of the sound periodicity. In the more complex *coupled* model (section 3.1.2), temporal correlations are introduced by means of specifying the covariance structure of $\boldsymbol{f}$. Furthermore, the covariance structure, which also governs the expected overall spectral envelope, is allowed to depend on the periodicity Ω.

Having specified probability distributions of over auditory nerve responses for each possible periodicity Ω, we can use the framework of Bayesian probabilistic inference to infer the most likely sound periodicity underlying some given, observed pattern of auditory nerve activity. Due to the complexity of the generative model, exact inference is computationally intractable. Two approximate inference algorithms were therefore presented. The *Laplace* algorithm (section 3.3.1) is based on a Gaussian approximation to the posterior distribution over unobserved waveform around its mode. *Hamiltonian Annealed Importance Sampling* (HAIS; section 3.3.2) in contrast is a Markov Chain Monte Carlo method that attempts to estimate the volume of the intractable posterior distribution stochastically by drawing iterative samples.

Chapter 4

Basic model evaluation

In this chapter, we will perform a first evaluation of the basic, uncoupled model as defined by the generative equations (3.5) - (3.8) and (3.23). Two questions are of immediate interest here. The acoustic component of our generative model is a model of periodic sounds, corrupted by additive Gaussian noise. Therefore, a first test should be whether our approximate inference schemes are capable of estimating the true periodicity of sounds of this type at the very least. Beyond this, we want to investigate to what degree our model is suitable as a explanation for psychophysical phenomena in human pitch perception, i.e. to what degree it can explain the pitch of non-periodic sounds or the pitch of periodic sounds in cases where it does not coincide with the periodicity rate of the stimulus.

In order to estimate the periodicity of an arbitrary sound, we first generate an AN response A according to our forward model (cf. section 3.2), and use this self-generated response as input for one of our inference algorithms. We then evaluate the log-likelihood function $\ln \mathcal{L}(\Omega) = \ln \mathrm{P}(A \,|\, \Omega)$ for an appropriate set of possible candidate periodicities.

A number of choices have to be made regarding the setting of stimulus- and model parameters. For all following examples, we arbitrarily scale the acoustic stimulus to have mean squared amplitude (i.e. power) of 20. The covariance matrices of the generative model are similarly scaled to yield the same expected power, i.e. the model knows the energy of the stimulus. Likewise, when performing inference about periodic sounds presented in a background of noise at a particular signal-to-noise ratio (SNR), we typically scale the signal and noise components in the generative model (i.e. $\frac{1}{T}$ trace(Σ^s)

and σ_x^2 in equation (3.6)) to obey the same ratio. This may seem like a potentially unfair advantage for our model, since a human listener does not know the true acoustic SNR *a priori* when listening to a stimulus. However, we found this potential advantage to be of little practical concern: when comparing the likelihood values across a range of models with SNRs set to different values, we obtain the highest likelihood values for assumed SNRs in the model close to the true SNR in the stimulus. Thus, the outcome of the periodicity estimation would essentially be the same if the model were to *infer* (or integrate out) the stimulus SNR as a further unobserved variable instead, while increasing the computational cost substantially. This was verified for a range of stimuli. Furthermore, we would argue that it is not unreasonable to assume that a human listener can form a relatively reliable estimate of the stimulus SNR over the course of a prolonged psychophysical experiment.

Our peripheral gammatone filter-bank has 80 channels with CFs extending from 40 Hz to maximally 16 kHz (or lower, when we reduce the sampling rate below 32 kHz for the sake of computational efficiency). The level of AN noise (σ_A^2 in equation (3.22)) was chosen so as to yield high SNRs (on the order of 10 dB) in channels with high evoked activity levels, equally during the generation of simulated AN responses and during inference.

A final consideration concerns the choice of envelope time-constants τ_s and pulse-train offset ϕ_s that define the mixture covariance matrices Ψ^s according to equation (3.4), and which we introduced to the model partially as a means to integrate out these unknown parameters of the waveform during inference. For the sake of computational efficiency, we want to avoid integrating over values of τ_s which contribute only negligeable probability mass to $\mathcal{L}(\Omega)$. Conversely, it is important to ensure that the range of timescales τ_s considered is wide enough such that a reasonable explanation of the waveform envelope can be provided for every setting of Ω. Hence, when evaluating $\ln \mathcal{L}(\Omega)$ for a range of Ω spanning several octaves, we balance these two constraints by choosing the timescales τ_s relative to (and differently for) each Ω from a range between 0.1Ω and Ω. As Figure 3.4 demonstrates, this allows for sharp pulses as well as for waveforms that remain approximately flat within the duration of one period Ω.

In section 3.3, we presented two alternative schemes – Laplace and Hamiltonian Annealed Importance Sampling (HAIS) – to estimate the unobserved stimulus periodicity Ω from an observed pattern $A = \{\boldsymbol{a}_i\}$ of time-varying auditory nerve (AN) activity in peripheral frequency channels $i = 1 \ldots C$. In either case, our goal is to compute the log-likelihood function $\ln \mathcal{L}(\Omega) = \ln \mathrm{P}(A \,|\, \Omega) = \int \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)\, \mathrm{d}\boldsymbol{x}$. In the Laplace scheme, we approximate $\ln \mathcal{L}(\Omega)$ as the Laplace integral of $\ln \mathrm{P}(A \,|\, \Omega)$ around the most likely, unobserved waveform $\boldsymbol{x}^*$. In the second scheme (HAIS), we use Annealed Importance Sampling (with the Markov transition operator implemented as a Hamiltonian Monte Carlo sampler) to approximate the integral over unobserved waveforms stochastically. Figure 4.1 shows the estimated log-likelihood functions for three different sounds (pure tone, HCT and SAM tone) under the two different approximation schemes around their pitch frequency ($\ln \mathcal{L}^{Lap}$ and $\ln \mathcal{L}^{HAIS}$) . While there is an overall offset between $\ln \mathcal{L}^{Lap}$ and $\ln \mathcal{L}^{HAIS}$ in each of the three examples (i.e. a multiplicative scaling difference between $\mathcal{L}^{Lap}$ and $\mathcal{L}^{HAIS}$), the shape of the estimated log-likelihood functions is otherwise similar for the two approximations. We do not have a solid analytic understanding for this seemingly consistent discrepancy, whereby the Laplace approximation yields higher likelihoods than our sampling-based method. On the one hand, it is clear that non-Gaussianity of the distribution $\mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$ can in principle give rise to biases in the Laplace integral, as can local optima during the initial gradient-ascent phase of the algorithm. On the other hand, convergence of MCMC methods in high-dimensional spaces may be slow and is generally hard to establish. Even though this was not considered further for the purpose of this thesis, one might attempt to identify the true cause of the discrepancy by reducing the peripheral processing step to a fully linear operation, in which case $P(A \,|\, \Omega)$ would become a mixture of Gaussian distributions (cf. discussion in section 3.3) with fully-known parameters. Linearising the peripheral model, however, has significant functional consequences by making envelope demodulation invariably impossible, and it is unclear whether the insights gained by this analysis would even hold for the actual non-linear model. In the absence of a established ground truth for $\ln \mathcal{L}(\Omega)$, all likelihood estimates in the following sections and chapter will be computed using the Laplace approximation due to its much lower computational demand[1].

[1] To give a rough sense of scale: In the Laplace approximation, a few hundred gradient steps in $\boldsymbol{x}$ are typically needed to find a maximum of $\ln \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$, plus a single evaluation of its Hessian. If we perform 500 annealing steps for each HAIS sample, where each annealing step itself involves 5 steps of simulating the Hamiltonian dynamics, we need 2500 gradient calculations for a *single* sample. 400 samples were used to estimate the likelihoods shown in Figure 4.1, requiring more than 1000 times the number of gradient steps required for the Laplace-based algorithm in total. A general, more precise comparison of the computational cost is problematic since both the number of gradient steps in the Laplace method and the number of MCMC samples required for convergence of th HAIS method depend on the stimulus. From our rough comparison above, it is nevertheless clear that the increased effort of HAIS over Laplace can easily become prohibitively large.

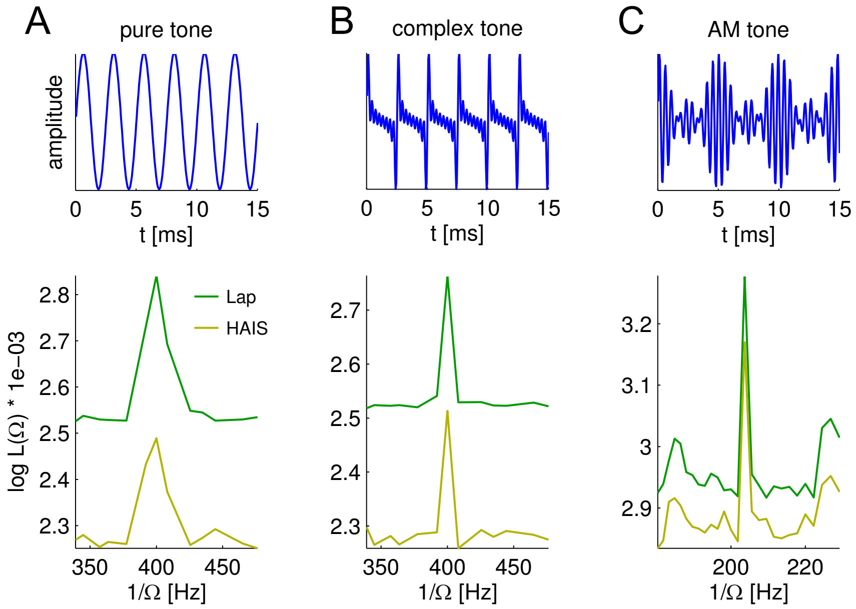


Figure 4.1: Inference: Examples of periodicity estimates obtained using either a Laplace approximation or Hamiltonian Annealed Importance Sampling (HAIS). Stimulus waveforms (top) and estimated log-likelihood profiles $\ln \mathcal{L}(\Omega)$ (bottom).

4.1 Pure tones

As a first example, we will show model estimates of the periodicity of pure tones. We generated 80 ms pure tones at a sampling rate of 32 kHz with frequencies ranging from 50 Hz to 6.4 kHz. Gaussian white noise was added at a SNR of 6 dB. We computed $\ln \mathcal{L}(\Omega)$ for a three-octave range around the true stimulus frequency in up to 48 steps per octave (limited by the sampling rate for higher frequencies) as shown in Figure 4.2 (top). The true stimulus frequency was inferred in every case. Notably, there is a strong octave ambiguity between the true pure tone frequency and the frequency one octave below, but not above[2]. This is consistent with features of the SACF, where peaks occur at integer multiples of Ω, and with spectral pattern matching models where the presence of a spectral component is regarded as evidence for an f_0 at subharmonic frequencies (not however with a simple, maximum-excitation place model). A noteworthy feature of the estimated profiles is a broadening of the peaks (on a logarithmic scale) at the low-frequency end, which is likely due to a combination of factors. Firstly, the bandwidth of low-CF gammatone filters relative to their centre frequency, as well as their spacing, is wider than that of high-frequency fibres, owing to their linear relation to the ERB scale

[2] Similar, still lower peaks occur also at higher-order submultiples of the tone frequency (not shown)

(cf. equation (3.21) in section 3.2.1). Hence, the same small difference in log-frequency may be harder to discriminate for low frequencies than for medium-to-high frequencies. Peaks also broaden for the highest frequencies: discriminability in the model is limited for technical reasons as Ω approaches the sampling rate limit. For a tone at 6.4 kHz and a sampling rate of 32 kHz, the smallest detectable difference is approximately 20%. Another likely reason for the broadening of $\ln \mathcal{L}(\Omega)$ at low frequencies is the small number of cycles that can be observed over the 80 ms duration of the stimulus. This is further amplified by the fact that we evaluate $\ln \mathcal{L}(\Omega)$ based on the AN activity only up to 80 ms, i.e. before ringing of the peripheral filters has subsided. As low-CF filters are slower to build-up their response, more information is lost in this way for low-frequency stimuli. In the bottom panel of Figure 4.2, we compare the shape of $\ln \mathcal{L}(\Omega)$ for a 400 Hz pure tone and different stimulus durations ranging from 20 to 80 ms, evaluated at a rate of 24 steps per octave (the overall difference in scale is due to the lower acoustic SNR of 0 dB is this example). As one would expect, the model accumulates evidence over time and the likelihood function becomes increasingly peaked for longer stimulus durations. We did not systematically explore longer durations, but there is no reason to believe that this trend should subside. Data on pure-tone frequency difference limens by Moore (1973) (see Figure 2.3) for example suggests an upper limit to the integration time window in human perception between 100 and 200 ms[3].

4.2 Harmonic complex tones

Next, we will consider the pitch of harmonic complex tones (cf. section 2.1.1.2). We generated missing-f_0 complex HCTs with 11 consecutive harmonics of 250 Hz each. The rank n of the lowest harmonic was varied from 4 to 22 in steps of three, i.e. ranging from HCTs with many resolved harmonics to entire unresolved sounds. Stimuli were 80 ms long, sampled at 32 kHz and presented in white noise at 0 dB SNR. Houtsma and Smurzynski (1990) used a very similar set of stimuli to measure f_0-discriminability as a function of n. While subjects were able to perform the discrimination task even for the entirely unresolved HCTs in principle, the authors reported a steep decrease in discriminability between $n = 7$ and $n = 13$ which then flattened off for values of $n = 16$

[3] At the same time, however, human listeners are sensitive to changes in pitch at a much faster rate, indicating that the time scale of integration may be dependent on the stimuli, stimulus context or task (see e.g. Balaguer-Ballester et al., 2009).

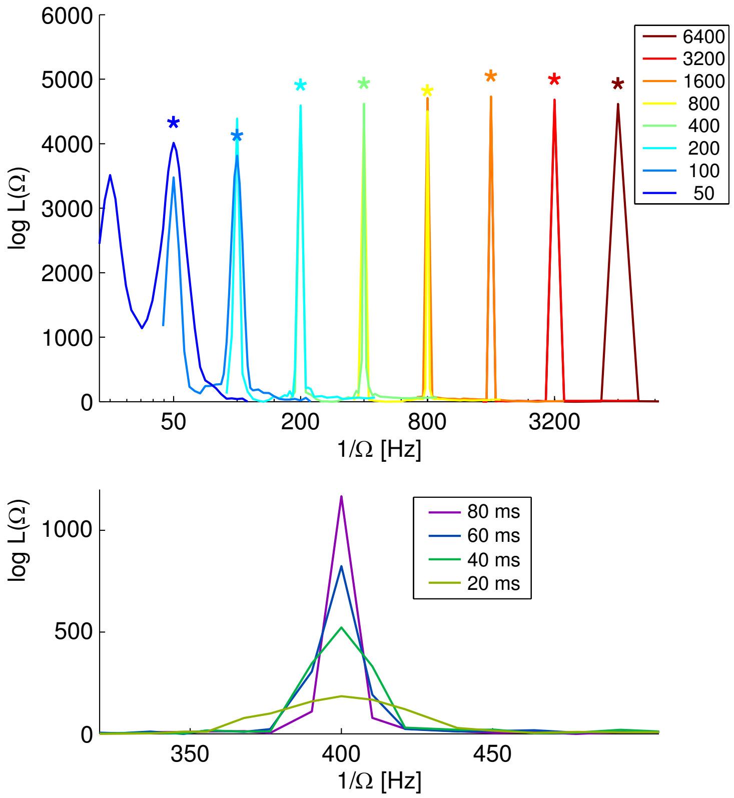


Figure 4.2: Top: Log-likelihood profiles for pure tones in the frequency range of 50 to 6400 Hz. Stars indicate the ML estimate for each curve. Bottom: Dependence of $\ln \mathcal{L}(\Omega)$ on stimulus duration.

and higher. These are important results for two reasons. Firstly, they demonstrate (along with many other sets of experiments) that pitch perception based entirely on unresolved spectral components is possible. This rules out models based solely on pattern-matching of the peripherally resolved spectrum (but not pattern matching of spectra derived from the time course of neural activity as proposed by Srulovicz and Goldstein (1983); cf. section 2.4.1). Secondly, the data of Houtsma and Smurzynski (1990) are also challenging for models based on the summary autocorrelation function (SACF; cf. section 2.4.2), since the the SACF does not naturally account for the marked drop in pitch strength around $n = 10$ (even though this challenge has been addressed in a recent modification by Bernstein and Oxenham (2005)). Therefore, this set of stimuli is an interesting test case for the Bayesian model.

As Figure 4.3 shows, the Bayesian model behaves qualitatively much like the SACF model. On the one hand, the stimulus periodicity is correctly inferred for resolved and unresolved HCTs. However, we find no indication of a "weakened" reliability of the model estimates for stimuli with $n > 10$. We did not evaluate $\ln \mathcal{L}(\Omega)$ densely enough around 250 Hz, or with sufficient number of stimulus samples, in order to compute f_0 difference limens that would allow us to compare our model estimates directly to the discriminability results of Houtsma and Smurzynski (1990). We can, however, interpret the height of the local peak in $\ln \mathcal{L}(\Omega)$ at 250 Hz as an indicator of the subjective single-trial certainty of the model about its inference (within a narrow region). Using this as a measure of “pitch strength”, we must conclude that pitch strength in the Bayesian model is unaffected by the rank of the lowest harmonic. We will not consider ways of addressing this undesirable property of the model at this point, but instead revisit the issue in section 5.5 after introducing an extension of our generative model targeted at capturing human perceptual pitch-timbre interactions.

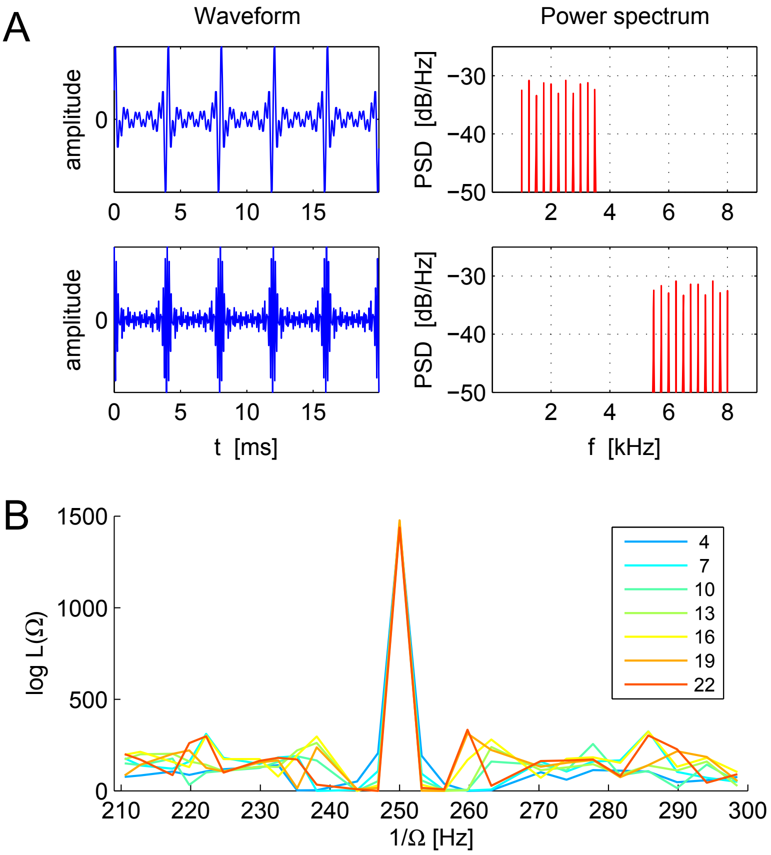


Figure 4.3: Missing-f_0 complex tones. A: Waveforms and spectra of two 250 Hz missing-f_0 HCTs with lowest harmonic rank $n = 4$ (top) and $n = 22$ (bottom). B: $\ln \mathcal{L}(\Omega)$ for HCTs with n varying from 4 to 22.

A different aspect of missing-f_0 pitch perception is the differential effect of phase manipulations on the pitch of HCTs with resolved and unresolved harmonics. Shackleton and Carlyon (1994) showed that phase relationships between spectral components that cause a doubling of the *envelope* periodicity of the sound from f_0 to $2f_0$ results in a doubling of pitch for unresolved, but not resolved HCTs (cf. section 2.1.1.2). There is a plausible mechanism for this effect: high-frequency AN fibres, stimulated by two or more unresolved harmonics, follow the envelope rather than the temporal fine-structure of the BM vibrations. Therefore, they cannot distinguish whether a doubling of envelope periodicity is caused by a doubling of f_0 or by component phase-shifts. Given our choice of peripheral front-end which explicitly comprises a stage of high-frequency envelope demodulation (cf. section 3.2.2), we expected to observe similar effects in our Bayesian model.

Following Shackleton and Carlyon (1994), we generated four different 250 Hz missing-f_0 HCT stimuli. Wide-band HCTs with equal-amplitude partials in either "sine phase" or "alternating phase" were filtered into two different frequency bands: a "low" one extending up to 625 Hz, and a "high" one extending from 3.9 to 5.4 kHz. Figure 4.4A shows the two high-filtered stimuli with a clearly-visible doubling of the envelope modulation rate for the alternating-phase HCT relative to the sine-phase HCT (a similar doubling occurs for the low-filtered stimuli). Estimation results (evaluating $\ln \mathcal{L}(\Omega)$ only at 250 and 500 Hz in this instance) are shown in 4.4B: while the inferred pitch of the two low-filtered HCTs is 250 Hz irrespective of the phase relationship, the pitch of the high-filtered HCTs is phase-dependent: the pitch of the alternating-phase HCT is estimated as 500 Hz, while that of the sine-phase HCT remains at the true f_0 of 250 Hz. Note that by the exact same mechanism, the SACF model achieves an identical effect (Meddis and O'Mard, 1997).

4.3 Iterated rippled noise

Iterated rippled noise (IRN) is generated by a process of repeatedly delaying a noise token by an interval d and adding it back to itself after multiplication with a gain factor g (cf. section 2.1.2, Figure 2.5). The mean reported pitch for IRN with $g = 1$ is equal to $\frac{1}{d}$ (250 Hz for $d = 4\,\text{ms}$), independent of the number of iterations. For $g = -1$, the percept is different. Pitch matches to IRN1$-$ (i.e. IRN with a single delay and subtraction) are bimodally distributed around approximately $1/d \pm 10\%$. For higher numbers of iterations n, the percept drops to $\frac{1}{2d}$ (Yost et al., 1978; Yost, 1996; see Figure 4.6C). The pitch of IRN $n+$, as well as that of IRN $n-$ for $n \gtrapprox 2$, is explicable both by the (broad) peaks in their resolved spectra and in terms of the temporal correlations

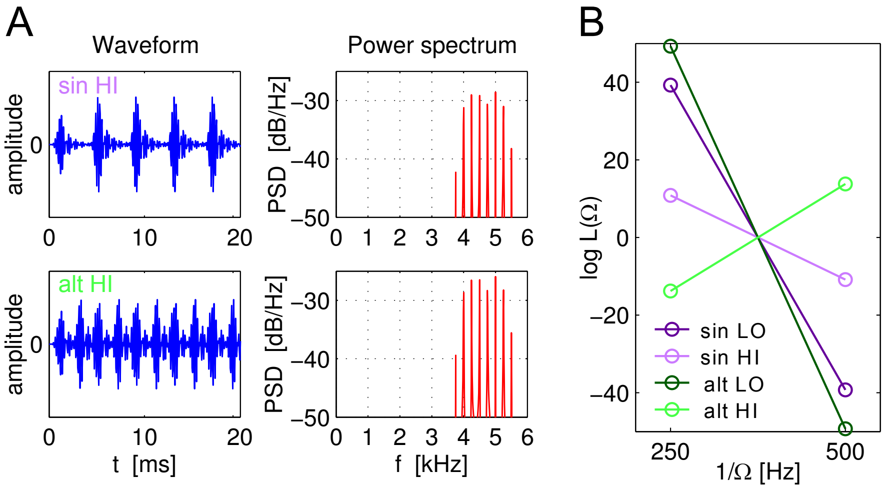


Figure 4.4: Phase-dependence of unresolved missing-f_0 pitch. A: Waveforms and spectra of two 250 Hz HCTs with components in sine (top) of alternating phase (bottom) filtered into a high, unresolved frequency region. B: log(Ω) evaluated at 250 and 500 Hz for low- and high-filtered 250 Hz HCTs in sine and alternating phase. Note the octave-increase in pitch for the high-filtered sound in alternating phase (light green).

induced by the delay-and-add process.

We generated IRN with a delay of 4 ms, $g = \pm 1$ and one or four delay-add iterations, denoted as IRN 1+, IRN 1−, IRN 4+ and IRN 4− (Figure 4.5). The stimulus duration was 100 ms at a sampling rate of 20 kHz. Following Yost (1996), stimuli were low-pass filtered at 4 kHz (6th order Butterworth filter).

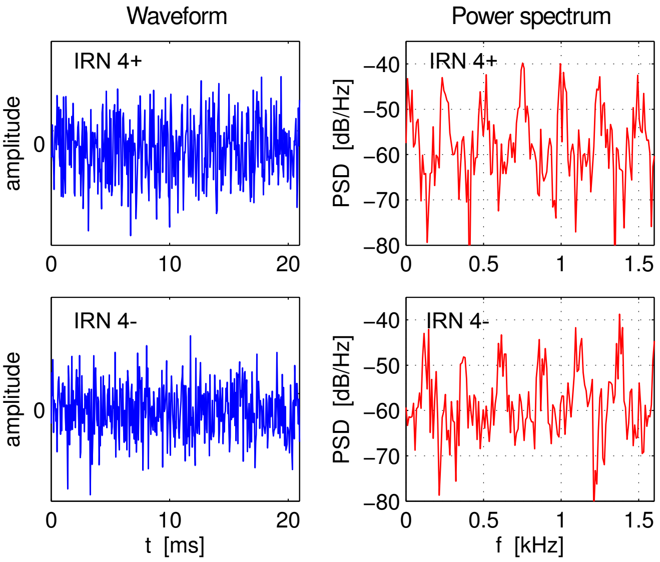


Figure 4.5: Waveforms and spectra of IRN 4+ and IRN 4−.

Figure 4.6A shows 10 log-likelihood profiles for each stimulus condition and their respective means. The mean profiles for IRN 1+ and IRN 4+ (orange and red) both peak at $1/\Omega = 250\,\mathrm{Hz}$ as expected from the psychophysics literature. The mean profiles for IRN 1− and IRN 4− both have a local minimum around that same value. The profile of IRN 4−, but not that of IRN 1− peaks visibly at $1/\Omega = 125\,\mathrm{Hz}$. We computed histograms of the ML estimates of Ω in semitone bins (Figure 4.6B). All 10 estimates for IRN 1+ and IRN 4+ fell within one semitone around 250 Hz ($\ln \mathcal{L}(\Omega)$ itself was evaluated at 48 values per octave). The distribution of ML estimates for IRN 1− was very broad and we generated an additional 15 samples to obtain a clearer picture. Peaks appear in the histogram around both 250 Hz and 125 Hz, but no estimate fell within one semitone of either of these two values. The clustering of pitch matches slightly above and below 250 Hz with an absence of matches at 250 Hz itself is in good agreement with psychophysical results reported by Yost (1996) (Figure 4.6C, top left and right). However, the model has an overall greater tendency to infer periodicities around 125 Hz than apparent in subjects' matching behaviour. For IRN 4−, model results and human psychophysics are in close agreement, with pitch matches clustered unimodally around 125 Hz in both cases (Figures 4.6B and 4.6C, bottom left).

Meddis and Hewitt (1991) argued that the SACF model accounts for the pitch of IRN 1+ and IRN 1−, but the the periodicity estimates for IRN 1− were restricted to within ±20% of the inverse delay $\frac{1}{d}$ in order to avoid spurious pitches caused by random peaks in the SACF. Owing to the noisy, broadly modulated spectrum of IRN with a low number of iterations, the pitch of such stimuli is difficult (if not impossible) to account for with spectral pattern-matching models that require a line-spectrum representation of the stimulus, e.g. the models by Goldstein (1973) and Terhardt (1974). Even assuming that some spectral pre-processor was able to identify the peaks in the spectrum, it still remains unclear how those models would account for the drop in pitch between IRN 1− and IRN 4− as the peak frequencies do not change when the number of iterations is increased. The pattern transformation model by Wightman (1973) is capable of processing continuous spectra such as those of IRN. Yost and Hill (1979) demonstrated that the model is able to account for bimodal distribution of pitches around $\frac{1}{d}$ in the case of IRN 1− at least qualitatively.

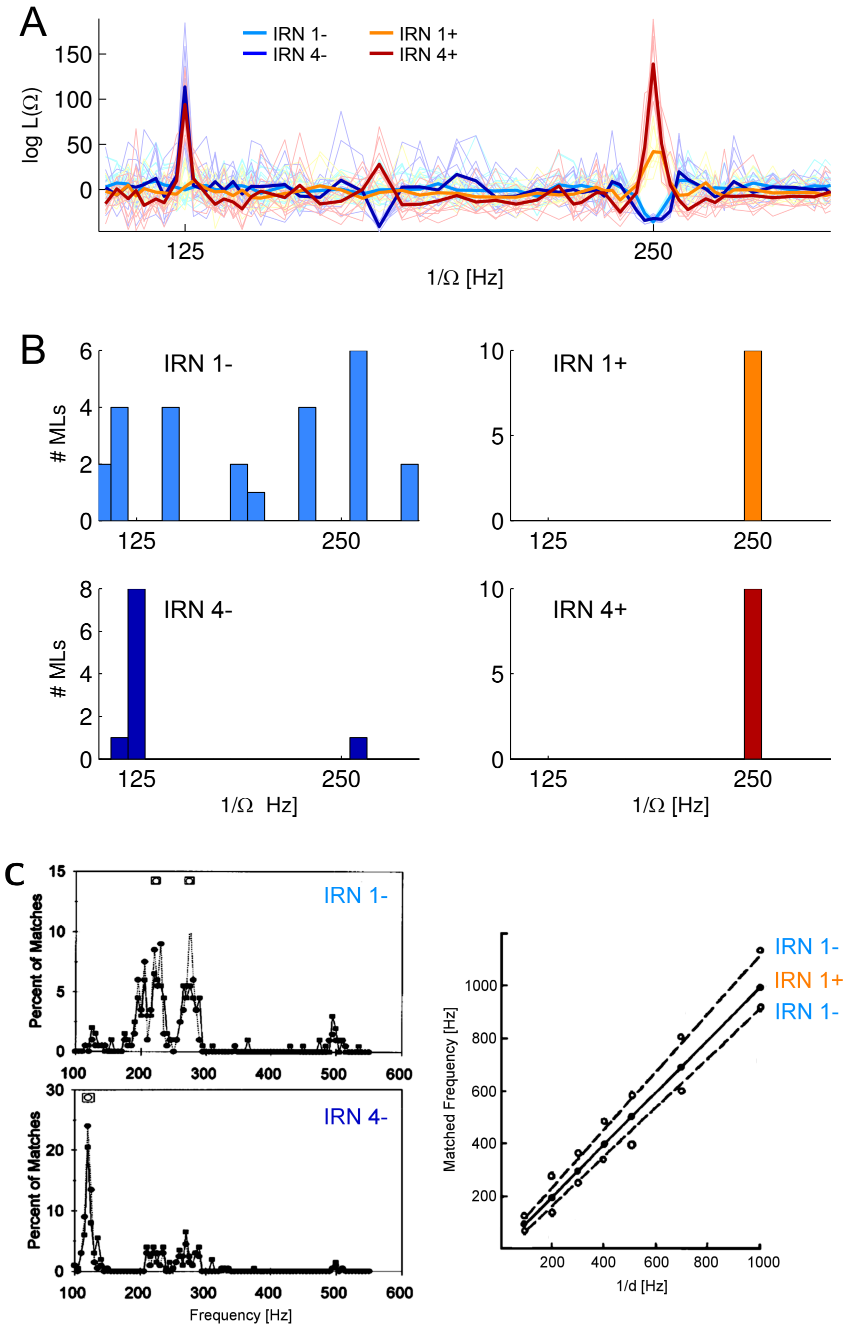


Figure 4.6: Model estimates and psychophysical data for IRN stimuli with a delay of 4 ms, positive and negative gain, 1 and 4 iterations. A: $\ln \mathcal{L}(\Omega)$ for 10 samples of each of the four stimulus conditions (thin lines) and their respective means (thick lines). B: Histograms of ML estimates of Ω in semitone bins (25 samples for IRN 1−, 10 each for the IRN 1+, IRN4+ and IRN 4−) C: Psychophysical data. Left: histograms of pitch matches for IRN 1− and IRN 4− (adapted from Yost, 1996). Right: Modes of the distribution of pitch matches for IRN 1− and IRN 1+ as a function of delay, showing a lack of matches at $f_0 = \frac{1}{d}$ in the case of IRN 1− (adapted from Yost et al., 1978).

4.4 Pitch shift of amplitude-modulated tones

Sinusoidally amplitude-modulated (SAM) tones with carrier frequency f_c and modulation rate g have the line spectrum of a frequency-shifted HCT with frequencies f_c-g, f_c and f_c+g. Schouten (1940) approximated the dominant perceived pitch as $f_p = g + \frac{\Delta f}{n}$, where n is the rank of the harmonic of g closest to f_c, and $\Delta f = f_c - ng$ is the amount of shift (cf. equation (2.3) section 2.1.2). The pitch of SAM tones is explicable both by their SACF, as well as by spectral pattern matching, provided of course that the components are peripherally resolved in the latter case. Needless to say, the pitch of unresolved SAM tones is not explained by pattern matching models.

We tested the model behaviour for shifted, three-component complex sounds much like SAM tones. We varied the centre frequency f_c between 1900 Hz and 2100 Hz, with a constant component spacing of $g = 200$ Hz around f_c (see Figure 4.7). We chose a high sampling rate of 44 kHz in order to be able to evaluate $\ln \mathcal{L}(\Omega)$ with high resolution around g. The stimulus duration was 60 ms. We observed that the peak of $\ln \mathcal{L}(\Omega)$ shifted away from 200 Hz for $f_c \neq 2000$ Hz (Figure 4.7B shows three examples), exactly as described by Schouten's approximation (4.7C). $\ln \mathcal{L}(\Omega)$ in the model has several, clearly distinguishable modes other than the ML estimate, the two next-highest of which are also shown in Figure 4.7C. Their positions are well described by Schouten's approximation where n is chosen higher or lower than the true rank of the harmonic of g closest to f_c. Schouten made the same observation psychophysically: subjects, when encouraged or cued, reported additional pitch matches at corresponding frequencies (cf. section 2.1.2). In the model, a discontinuity occurs around $f_c = 2100$ Hz, at which point the spectral components of the stimulus are better explained as harmonics $10-12$ of a harmonic sound with $f_0 \approx 190$ Hz than as harmonics $9-11$ of an $f_0 \approx 210$ Hz. Similar discontinuities around the point where f_c is half-way between two harmonics of g (e.g. 2000 and 2200 Hz) have been observed psychophysically for shifted harmonic complex tones (e.g. de Boer, 1956a).

4.5 Amplitude-modulated noise

Sinusoidally amplitude-modulated (SAM) noise gives rise to a weak pitch corresponding to the modulation frequency, for modulation frequencies up to approximately 1 kHz

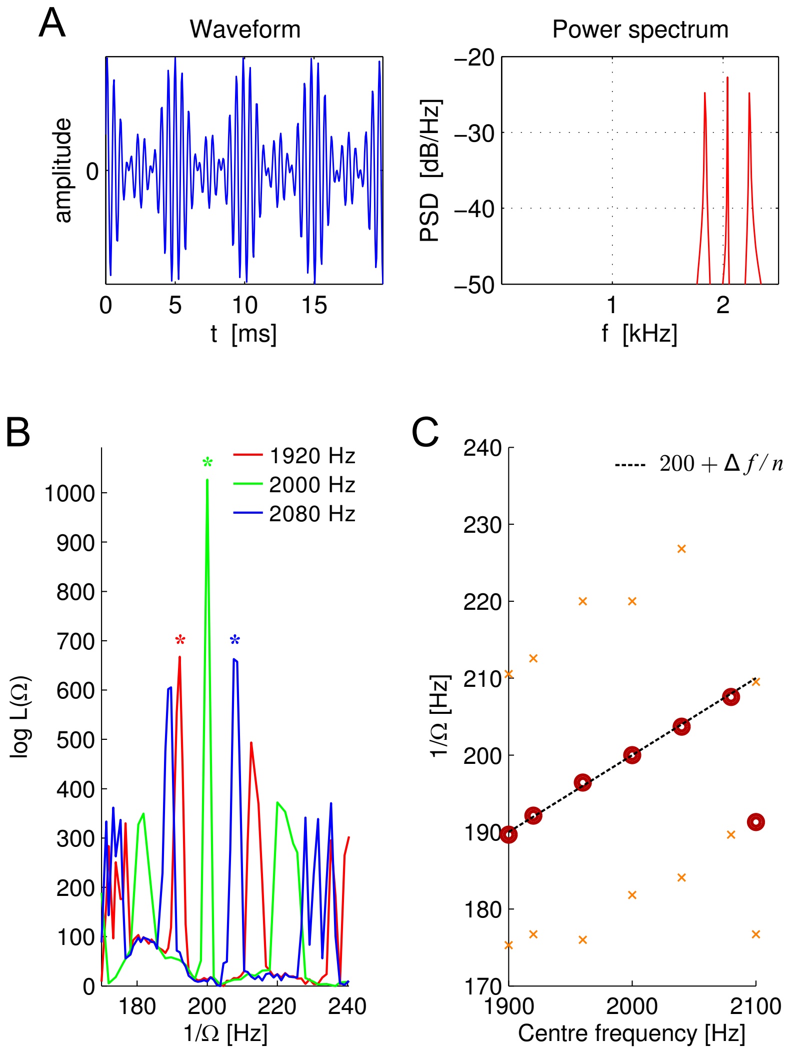


Figure 4.7: 200 Hz amplitude-modulated tones. A: Waveform and power spectrum of a shifted HCT with frequency components 1840, 2040 and 2240 Hz, similar to a 2040 Hz pure tone sinusoidally modulated at 200 Hz. B: $\ln \mathcal{L}(\Omega)$ for three SAM tones with carrier frequencies f_c of 1920, 2000 and 2080 Hz, and a constant modulation rate of $g = 200$ Hz. Stars indicate the ML estimate of Ω, showing a pitch shift for the tones with $f_c \neq 2$ kHz. C: ML estimates (circles) and two highest side-peaks in $\ln \mathcal{L}(\Omega)$ (crosses) for SAM tones with f_c ranging from 1900 to 2100 Hz. Note the discontinuity at $f_c = 2100$ Hz. The dotted line indicates a linear approximation to subjects' matching behaviour derived by Schouten (1940) from psychophysical data (cf. section 2.1.2).

(Burns and Viemeister, 1976, 1981). The variability of this pitch is on the order of one semitone. Even though the pitch of *narrow-band* SAM noise may be mediated by mechanical distortions in the inner ear, modulation rates of SAM noise remain discriminable even when distortions at the modulation frequency are actively cancelled (Wiegrebe and Patterson, 1999; cf. section 2.1.2). Since SAM noise has flat, featureless power spectrum, pattern matching fails to predict its pitch, while SACF and related temporal models are sensitive to the amplitude modulations generated in high-CF fibres.

We tested the sensitivity of our model to SAM white noise with a modulation rate of 200 Hz (Figure 4.8A). The stimulus duration was 80 ms at a sampling rate of 24 kHz, no unmodulated noise was added. Figure 4.8B shows the estimated log-likelihood profiles for 20 different draws of white noise (blue curves) as well as the averaged profile (black). While $\ln \mathcal{L}(\Omega)$ tends to peak around 200 Hz on average, the true modulator frequency was not inferred as the most likely stimulus periodicity for each individual sample. The histogram of the ML estimates (24 bins per octave) is shown in 4.8B. 19 out of 20 estimates fell within approximately one semitone of 200 Hz (4.8C), indicating that the inferred periodicity is more labile than that of most other pitch-evoking sounds discussed in this chapter (with the exception of IRN 1−). There is another qualitative difference between the pitch of SAM noise and that of other pitch-evoking stimuli in the model. For the purpose of inferring Ω, we initially set the assumed acoustic SNR in the model to an arbitrary low value (-9 dB in the data shown). While the setting of this value was found to have little influence on the outcome of the periodicity estimation and its reliability, we observed that the likelihoods were overall higher for lower settings of the SNR. In fact, we found that the single most likely explanation of the observed AN response was obtained when the model assumed the stimulus to be pure noise (in which case, of course, all periodicities Ω are equally likely). An example for one SAM noise stimulus is shown in Figure 4.8D: the dotted line represents the (flat) likelihood profile under the “noise only” model. It is persistently higher than the likelihood profile under the model assuming -9 dB SNR. This means in turn that the model, when forced to decide whether the stimulus contains a periodic element at all, would prefer to regard the stimulus as entirely aperiodic. In order to explain the pitch of SAM noise, we could assume that the model integrates over SNRs in the stimulus (effectively treating it as another unobserved nuisance variable), in which case a peak is expected to persist at 200 Hz, albeit weak. Alternatively, we could accept that a listener may be able to wilfully override his own point-estimate of the SNR and listen to the sound under the *prior* assumption that a periodic component must be present in the stimulus. The ability to do so might, for example, depend on listening practice and training with SAM noise, or musical expertise in general.

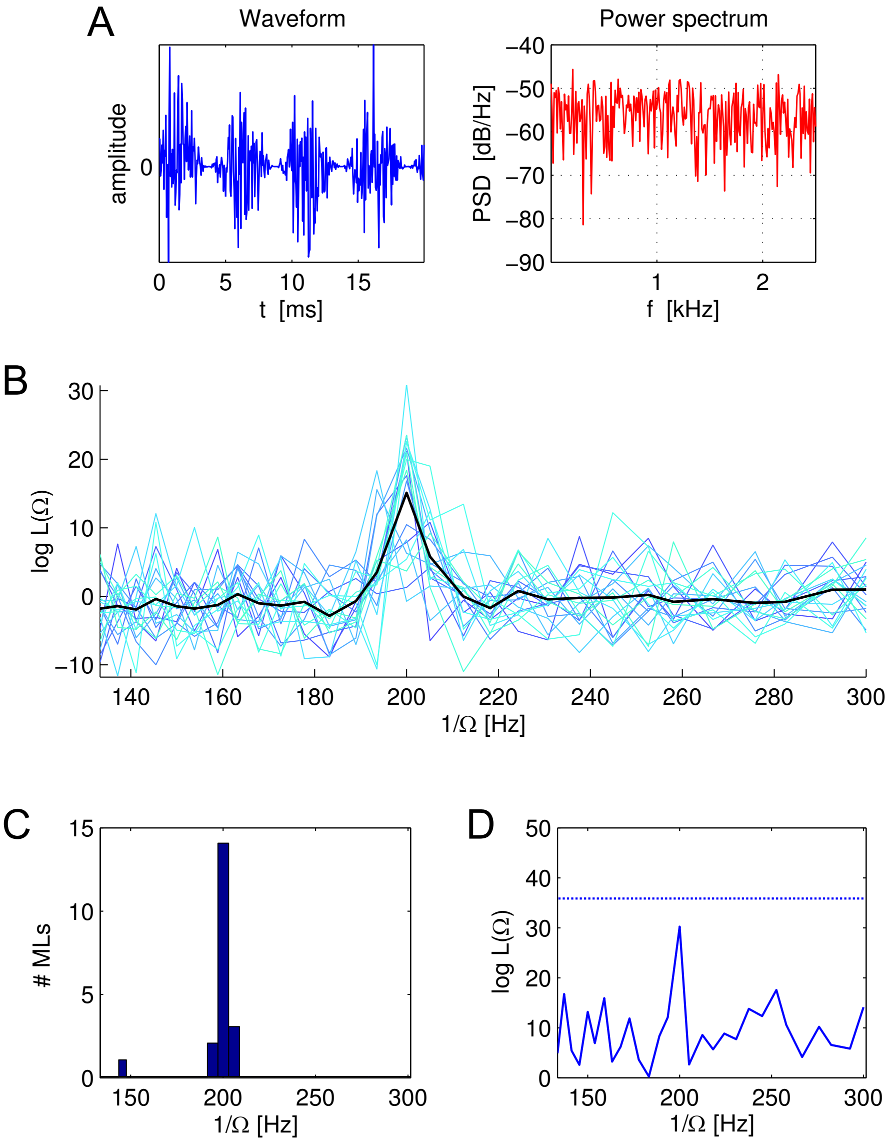


Figure 4.8: The pitch of 200 Hz sinusoidally amplitude-modulated noise. A: Example stimulus waveform and spectrum B: Log-likelihood profiles for 20 samples of SAM noise (blue) and their mean (black). Individual profiles were shifted to have zero-mean for ease of visualisation. C: Histogram of the ML estimates of Ω for the samples shown in B (24 bins / octave). D: $\ln \mathcal{L}(\Omega)$ for a SAM noise stimulus, evaluated under two different models. Assuming pure noise in the stimulus (dotted line) yields overall higher likelihoods than assuming a low, but finite SNR of -9dB (solid curve).

4.6 Transposed complex tones

Transposed complex tones (TCTs) are the summed products of half-wave rectified sinusoidal modulators with high-frequency sinusoidal carriers, where the modulators share a common fundamental frequency f_0^m (Figure 4.9A). The carrier frequencies are chosen such that the different modulation bands do not overlap in their resolved, peripheral excitation pattern. The stimulus modulation-rates are reflected by periodic modulations of the peripheral filter outputs in the respective carrier frequency bands (Figure 4.9B, right and centre). Oxenham et al. (2004) designed this type of stimulus to test whether periodicity information is combined across different peripheral frequency channels irrespective of their spectral identity in terms of CF, as suggested by the summary autocorrelation model (see section 2.4.2). The stimulus frequencies used in their study were 300, 400, and 500 Hz for the modulators, and 4, 6.35 and 10.08 kHz for the carriers. As shown in Figure 4.9B (right), the periodic modulation of AN activity in the carrier bands gives rise to peaks in the autocorrelation function (ACF) of the corresponding channels, which coincide at a lag of $\tau = 10\,\text{ms}$. Following summation, the SACF (bottom) peaks at $\tau = 10\,\text{ms}$, resulting in a pitch prediction of 100 Hz. With TCT embedded in a background of pink noise, human listeners were found to be unable to match f_0^m, the fundamental frequency of the modulators. Furthermore, three out of four listeners were unable to perform f_0^m discrimination at all, while a fourth subject had a f_0^m difference limen of approximately 8%. At the same time, f_0 discrimination and matching was possible for three-component missing-f_0 HCTs (with harmonic component frequencies equal to the TCT modulator frequencies) under the same noise conditions. These results contradict the predictions of the SACF model. Not surprisingly, however, there is no basis for reporting a low pitch of 100 Hz in spectral pattern matching models.

We generated TCTs as described in Oxenham et al. (2004). Stimuli were 80 ms long, sampled at a rate of 32 kHz (the same rate that was used by the authors of the study) and embedded in pink noise at a total SNR of 0 dB. Figure 4.10 (top) shows the log-likelihood profiles for three different draws of background noise. In each instance, the model infers a stimulus periodicity corresponding to the high-frequency carrier bands, rather than the low fundamental of 100 Hz. At the same time, missing-f_0 HCTs with component frequencies 300, 400 and 500 Hz in the same noise background give rise to a ML estimate at the fundamental (bottom). Thus, the (correct) failure to report a pitch of 100 Hz in case of the TCT is not simply due to an overwhelming noise masker. Instead, we can understand the differential behaviour of the Bayesian model (which is lacking in the SACF model) intuitively by considering the statistics of naturally-evoked AN responses to periodic sounds. The observation of amplitude modulations of 300, 400 and 500 Hz in frequency channels with *matching* CFs is perfectly reasonable for an approximately harmonic sound with $f_0 = 100\,\text{Hz}$ as these are indicative of harmonics

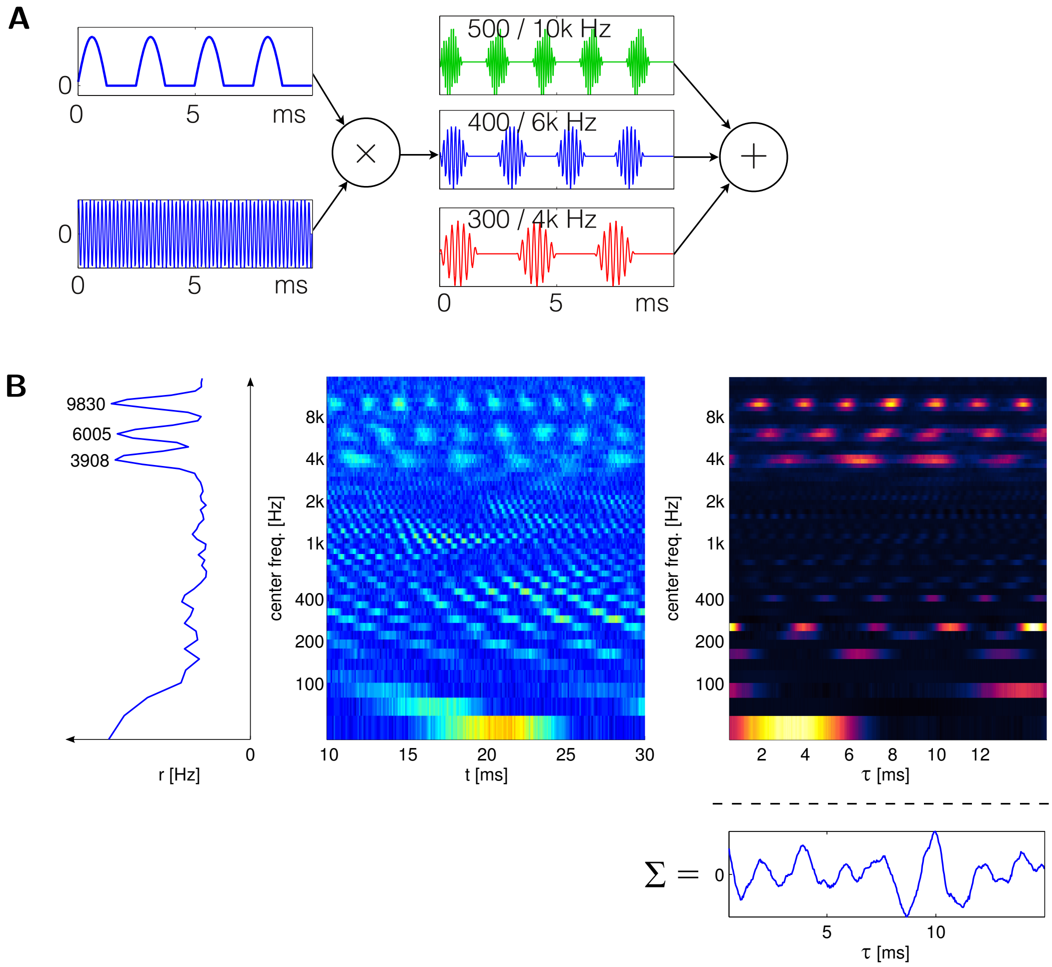


Figure 4.9: A: Generation of a transposed complex tone. Rectified sinusoidal envelopes with a common fundamental frequency are used to modulate high-frequency carriers, and the modulated carriers are summed. B: Simulated auditory nerve response to a transposed complex (centre; modulator and carrier frequencies as in A). The carriers generate peaks in the average AN firing rates, independent of the modulator frequencies (left). The modulators cause regular peaks in each of the three carrier frequency-band, which coincide at lag $\tau = 10\,\text{ms}$ and cause a peak in the SACF (right).

3 to 5 in the stimulus. Contrarily, the observation of the same set of modulation rates transposed into high-frequency channels is not as typical for harmonic sounds with $f_0 = 100\,\text{Hz}$. Since the harmonics of f_0 are unresolved in this high frequency range, they should evoke modulations of 100 Hz in each channel, unless the phase relationships between harmonics happen to be very specifically set — differently in each carrier band — so as to generate envelope modulation rates of 300, 400 and 500 Hz in the filter outputs. To the degree that our generative model captures this simple intuition, it should therefore down-weight the evidence for a low, common fundamental provided by the high-CF channels in case of the TCT compared to the correctly-matched low-CF

channels in case of a missing-f_0 HCT.

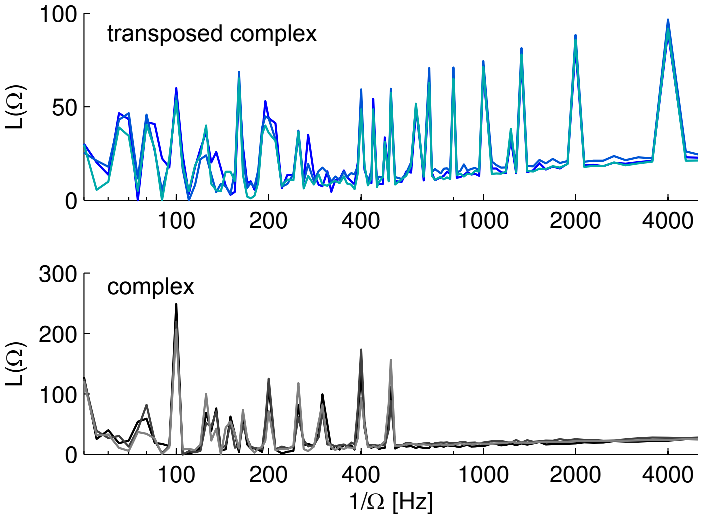


Figure 4.10: Top: $\ln \mathcal{L}(\Omega)$ for a transposed complex tone with $f_0^m = 100\,\text{Hz}$ in three different backgrounds of pink noise. Carriers of 4, 6.35, 10.08 kHz were modulated at rates of 300, 400 and 500 Hz respectively. Bottom: $\ln \mathcal{L}(\Omega)$ of three missing-f_0 HCTs with component frequencies equal to the modulator frequencies.

As discussed in section 2.4.2, an recent extension of the SACF model by Balaguer-Ballester et al. (2008) leads to the (desirable) failure of the model to identify f_0^m as the pitch of the TCT, except when the stimulus intensity in the simulation was lowered considerably. As the central, low-pass SACF (LP-SACF) pitch processor does still not take peripheral CF into account during the SACF computation, the change in model behaviour was most likely effected by the use of a more complex model of the peripheral response (Sumner et al., 2002) compared to earlier versions of the SACF model (e.g Meddis and O'Mard, 1997). Without further psychophysical evidence regarding level dependence on the one hand, and a demonstration of the phenomenological validity of the peripheral model-response to TCT stimuli in noise on the other hand, it is therefore impossible to draw strong conclusions regarding the value of the results by Oxenham et al. (2004) as evidence for or against SACF-like models and pitch-processing strategies that disregard the coherence or incoherence of filter CFs and their output firing patterns in general. In any case, the Bayesian model responds differentially to HCTs and TCTs despite its simple peripheral model which yields no distinction between these stimuli when used with a SACF-like read-out.

4.7 Discussion

The evaluations in this chapter have shown that the Bayesian model is able to infer the true periodicity of periodic sounds such as pure tones and HCTs. These sounds are well-described by the acoustic component of the generative model. In addition, the AN activity pattern used as input to our inference algorithm was generated according to the same peripheral model that we then assumed during inference. Thus, these stimuli provide in many ways ideal conditions for the model, and any failure to infer the correct periodicity would have been indicative of a severe deficiency of our approximate inference schemes. While the model is able to infer the true period duration and thus to predict the pitch *frequency*, we have observed that it fails to account for the degradation of pitch *strength* as a function of the lowest-ranking harmonic of a HCT. A similar failure was previously observed for the SACF model and interpreted as evidence against its suitability as a unified model of pitch.

The Bayesian model produces periodicity estimates for a variety of non-periodic sounds that are in close accordance with human pitch-matching behaviour, even though the generative model assumptions are substantially violated. The model predicts the pitch shift of SAM tones, the pitch of IRN+ and IRN- (including the dependence of the latter on the number of iterations), and the pitch of SAM noise. We have seen that the pitch of SAM noise is particularly weak in the model, in that highest likelihoods overall are achieved by assuming an SNR of 0 in the model, at which point periodicity is no longer detectable. Integration over SNRs or a strong, subjective prior in favour of the presence of a periodic sound could explain the existence of the percept despite pure noise being the single most likely interpretation of the evoked AN response. For all these stimuli, the SACF model makes virtually identical predictions. This demonstrates that periodicity estimates based on AN autocorrelation can be optimal for a considerable range of pitch-evoking stimuli.

Transposed harmonic tones (TCTs) were presented as an example where the prediction of SACF and the Bayesian model differ. While SACF predicts a low pitch at the missing fundamental $f_0^m = 100\,\mathrm{Hz}$ of the modulators, the Bayesian model predicts a high pitch around one of the carrier frequencies (4 kHz), explaining why subjects are unable to discriminate the pitch of TCTs with different f_0^m (but equal carriers).

Several of the phenomena discussed in section 2.1 are *not* captured by the model. We found no tendency to infer periodicities close to the edges of steeply filtered low- or high-pass noise in the model, contrary to the pitch reported by human listeners (Small and Daniloff, 1967; Fastl, 1971). Lateral inhibition has been suggested as a potential explanation, as it could give rise to a peak in peripheral (or central) representation of the stimulus spectrum around the edge frequency, and physiological evidence for lateral inhibition has been found in chopper neurons of the cochlear nucleus (Rhode and Greenberg, 1994; cf. section 2.3.1). To the best of our knowledge, neither SACF-like nor spectral models have been reported to predict spectral edge pitch, and our own implementation of the SACF model (cf. section 5.4) did not show such behaviour either. Note though that the spectral model by Cohen et al. (1995) does include a stage of centre-surround lateral interaction which could in principle give rise to the desired effect. A further phenomenon which the Bayesian model fails to predict are the small pitch shifts ($\approx 1\%$) of HCTs with a mistuned low-rank partial (Moore et al., 1985). In our model, the estimated periodicity simply remained at the fundamental of harmonic stimulus components. Meddis and O'Mard (1997) reported pitch shifts in the SACF model when pitch was determined by matching of the Euclidean distance D^2 between target and reference SACF (cf. section 2.4.2), but even then the amount of shift was underestimated by a factor of approximately two.

Out of the unexplained phenomena above, we consider the failure to account for the weakened pitch of high-rank missing-f_0 HCTs the most concerning, as it subjects the Bayesian model to the same fundamental criticism regarding its suitability as a unified model of pitch as the SACF model. As we will see in section 5.5, an extension of the generative model aimed at explaining pitch-timbre interactions in human perception turns out to provide a promising solution for the problem of pitch strength. Notably, the proposed model extension is rooted in the statistical properties of natural, pitch-evoking sounds, providing functional insight in addition to an improved phenomenological description of human pitch perception.

Chapter 5

Octave biases and timbral effects in the perception of non-uniform periodic pulse trains

5.1 Motivation

All results presented thus far were obtained from the basic model described in section 3.1.1, in which the stimulus period Ω and the acoustic impulse response $\boldsymbol{f}$ are treated as independent in the generative process. As $\boldsymbol{f}$ determines the shape of the spectral envelope of the sound, and Ω the spacing of spectral peaks underneath this envelope (the lowest of which is the fundamental frequency $f_0 = \Omega^{-1}$), sounds generated according to this model will contain, on average, the same amount of energy in any given range of frequencies, independent of the fundamental being low or high. Even more: since the generative distribution over impulse responses is temporally uncorrelated, the average spectrum of $\boldsymbol{f}$ is essentially white (cf. Figure 3.4). Hence, sounds of all periodicities are *a priori* assumed to have flat, broadband spectral envelopes under the model. Perceptually, the spectral envelope of a sound is a major determinant of its timbre, while the periodicity determines – by and large – its pitch. Thus, one might say that our basic, uncoupled model assumes pitch and timbre to be independent.

For natural pitch-evoking sounds, however, this assumption does not seem to hold, both acoustically and perceptually. In this chapter, we will review evidence for a systematic relationship between the fundamental frequency and spectral envelope in the acoustic properties of natural pitched sounds (such as produced by human voices and musical instruments) on the one hand, and for interactions between the perceptual attributes of pitch and timbre on the other. Based on this evidence, we propose an extended model that incorporates a coupling between these two quantities in the generative process, and fit the coupling parameters to a collection of instrumental and vocal sounds. Finally, we present psychophysical data that demonstrates the ability of our extended model to accurately predict effects of timbre on octave biases in the perception of acoustic stimuli with inherent octave ambiguities. While these effects arise naturally in our model as a consequence of Bayesian cue combination, they prove to be a difficult challenge for existing, more heuristic models based on either spectral pattern matching or autocorrelation analysis of the auditory nerve firing pattern. Finally, we go on to show that our extended model now captures a perceptual phenomenon which was previously unexplained by the simpler, uncoupled model. A dependence of the pitch strength of missing-f_0 harmonic complex tones on the rank (i.e. harmonic number) of the lowest harmonic in the spectrum arises through the coupling, which matches a similar dependence observed in human interval identification performance and pitch discriminability.

5.1.1 Timbre, brightness and spectral centroid

Four properties — duration, loudness, pitch and timbre — are commonly used to describe the perceptual quality of a sound. Out of those four, timbre is arguably the least well-defined. In fact, a typical approach is to define timbre as the ensemble of all those qualities which distinguish sounds of equal perceived pitch, loudness and duration[1] (Plomp, 1970; ANSI, 1994; see Houtsma, 1997 for a review): like for example the same musical note played on a violin and an oboe. Owing to its vague definition, solely in terms of negatives, timbre has somewhat disrespectfully been described as the "psychoacoustician's multidimensional waste-basket"(McAdams and Bregman, 1979). Several studies have attempted to identify and disentangle the underlying dimensions of this complex perceptual space (e.g. Lichte, 1941; Grey, 1977; McAdams et al., 1995), based on similarity ratings made by listeners between pairs of sounds. One dimension which has been identified consistently across studies correlates physically with the

[1] Spatial location appears to be an often-neglected aspect in this list of attributes.

spectral centre of mass (centroid), and is commonly described verbally as “brightness”. While other physical parameters (such as rise and decay rates of the waveform envelope, noisiness, or the presence of amplitude and frequency modulations) also contribute to the timbre of a sound, we will focus on brightness in the following. Brightness has not only been studied more carefully than any other aspect of timbre in its effect on pitch perception in the past (see section 5.1.3). It is also the aspect which is most readily incorporated into our existing statistical framework without the need for severe structural changes to the model, as we will show in section 5.2.

5.1.2 Relationship between fundamental frequency and spectral centroid in natural sounds

Over the broad range of pitch-evoking sounds that listeners (animal or human) are likely to encounter in their environment, there is *a priori* good reason to assume that their fundamental frequencies and spectral centroids should in fact depend on each other. Vocalisations, arguably the most common and relevant type of pitched sounds, are generated by the excitation of a resonator – the vocal tract, encompassing the laryngeal, pharyngeal, oral and nasal cavities in mammals – by the periodically modulated air flow originating from the primary sound source, namely the vocals folds (mammals) or syrinx (birds). It is an almost trivial observation, that small animals do not only tend to modulate their air flow at higher rates than larger animals, but that they also have shorter vocal tracts with correspondingly higher resonance frequencies[2]: compare for example the roar of a lion to the squeak of a rat. The same holds true not only across, but also within species: the vocal folds of women and children open and close at higher rates than those of men when speaking at their natural pitch, while at the same time the speech formants, i.e. dominant peaks in the spectral envelope, are shifted towards higher frequencies (Peterson and Barney, 1952), owing to their shorter vocal tracts (Patterson et al., 2008). Measuring the formant frequencies in the vocal tracts of soprano singers as they sang scales on different vowels, Joliveau et al. (2004) found that a similar, positive correlation between fundamental and centroid appears to exist even within a single individual, as the pitch moves from the bottom the the top of the

[2] Even though the resonance frequencies of the vocal tract are difficult to determine exactly due its complex shape, an inverse relationship between size (in any dimension) and resonance frequency exists for analytically tractable resonator shapes, such as cylindrical tubes, spheres or cuboids.

vocal range[3]. Another common source of pitched sounds are musical instruments. As with vocalisations, it is similarly true that low-pitched sounds are typically produced by large instruments with low resonance frequencies, while high-pitched instruments are small and resonate best at high frequencies (van Dinther and Patterson, 2006). The variation of the spectral centroid with f_0, however, depends on the instrument type: while they appear to co-vary strongly in some instruments, the spectral envelope remains more or less fixed in others (see for example Figure 5.1). We will quantify this relationship for a large ensemble of instrumental and vocal sounds more carefully in section 5.2.2.

5.1.3 Psychophysical effects of timbre on pitch

The previous section was concerned with the relationship of fundamental frequency and spectral centroid in the statistics of natural sounds, but what about their (approximate) perceptual counterparts, pitch and timbral brightness? In a study by Plomp and Steeneken (1971), subjects rated the perceptual dissimilarity between pairs of tones that differed in their fundamental frequency, spectral envelope, or both at the same time. From the structure of these dissimilarity rating, Plomp concluded that pitch and timbre are effectively processed as independent dimensions: the dissimilarity of tone pairs that differed in both fundamental and spectral envelope was found to be close to the summed dissimilarities along either dimension alone. Such a “city block” metric, as opposed to a Euclidean metric, in the judgement of dissimilarities along multiple dimensions is often regarded as evidence for perceptual separability (Garner, 1974). Several recent studies (Marozeau et al., 2001, 2003; Schubert and Wolfe, 2006; Marozeau and de Cheveigné, 2007) have indeed confirmed that differences in pitch have only rather limited influence on the perception of simultaneous differences in timbre (brightness in particular), and furthermore suggest that overall pitch-dependent shifts of sounds in timbre space (leaving timbre *differences* invariant) are also small.

While the effect of pitch on timbre may be small, there is a substantial body of evidence suggesting that timbre can affect the perception of pitch more severely. According to Hesse (1982), compositional practise since the nineteenth century has been aware of the need to compensate for gross timbral differences of certain instruments, alternating

[3] Note though, that the effect is strongest for fundamental frequencies above the natural speaking range.

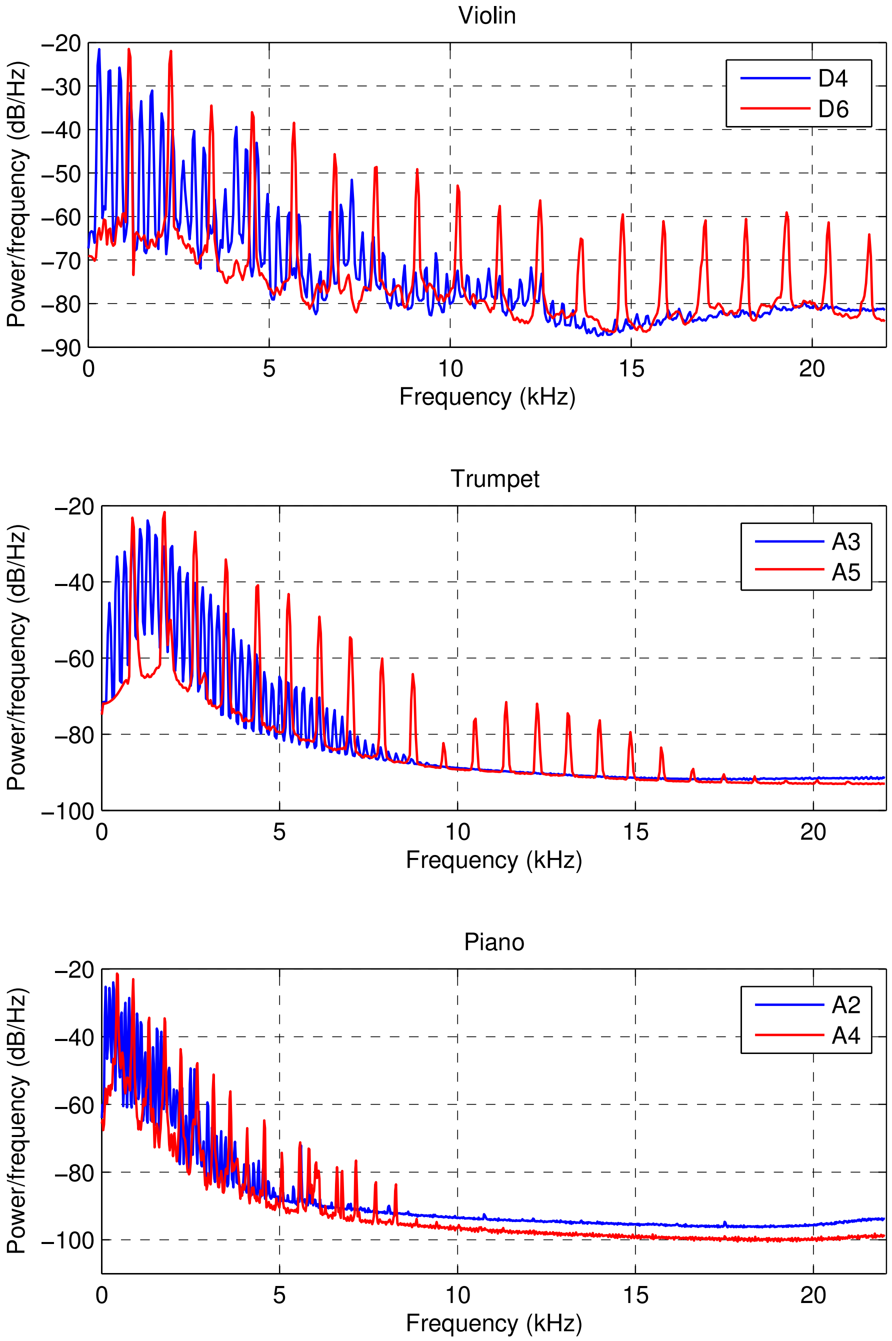


Figure 5.1: Spectra of a violin, a trumpet and a piano, each playing two notes separated by two octaves (lower-pitched note in blue, higher-pitched note in red). Samples taken from the University of Iowa Musical Instrument Samples database.

in the execution of a continuous melody, by octave transpositions in order to avoid perceived breaks in the melodic contour[4]. Similarly, von Helmholtz (1863) stated that "although the pitch of a compound tone is, for musical purposes, determined by that of its prime, the influence of the upper partial tones is by no means unfelt. They give the compound tone a brighter and higher effect". Hesse (1982) himself conducted experiments in which he had music students transcribe melodies that were played with synthesised sounds of various degrees of spectral brightness. When two notes of the same fundamental frequency were played in sequence, while the timbre was considerably brightened at the same time, students tended to transcribe the interval as an octave, rather than a unison. Similarly, a downward leap (in fundamental frequency) of a fourth would often be transcribed as an upward leap of a fifth, again differing from the true interval between the two fundamentals by an octave. Hesse concluded not only that brightness and pitch are dependent upon one another, but also that brightness is "a component or dimension of pitch". This interpretation was shared by van Norden (1982), who suggested that the perception of pitch height (i.e. that aspect of the pitch that distinguishes between different notes separated by one or several octaves) and brightness are mediated by the same underlying process. Robinson (1993) contested this view: in his experiments, subjects had to judge either whether the pitch interval between two notes was a unison or octave, or whether the instruments playing the two notes were the same or different[5]. Subjects, musically trained and untrained alike, were able correctly discriminate instrument identity (presumably based on their brightness) in the presence of differences in pitch height. Like Hesse (1982), however, he found that pitch height judgements were strongly impaired when the two notes were played on different instruments, i.e. with different brightness (more so, and for longer durations, in musically untrained subjects than in musicians). From this asymmetry, Robinson concluded that the perception of pitch height and brightness are based on separate processes, but with a unidirectional, partial dependency of pitch height on timbre.

Aside from the rather large-scale effect of brightness on the perception of octave position within the same pitch class (chroma), interactions between pitch and timbre have also been documented in the discrimination of nearby fundamental frequencies. Singh and Hirsh (1992) showed that an increase in brightness can perceptually compensate for a

[4] A mention of this rule is seemingly found in F. A. Gevaert's handbook on instrumentation, "Traité général d'instrumentation", from 1863 (Hesse, 1982).

[5] Two pitches, 130 and 260 Hz, and two synthetic instruments with different spectral brightness were used.

decrease in the fundamental frequency of up to 2% (i.e. less than half a semitone), such that the pitch is perceived as identical. Vurma et al. (2010) recently confirmed these results using natural vocal and instrumental sounds, indicating that timbre-induced, perceptual pitch shifts are likely to affect the subjective quality of intonation in everyday music performance. Measuring reaction times for pitch and timbre discrimination, Melara and Marks (1990) found an interference effect between timbre and pitch in a Garner speeded classification task: both attributes were varied simultaneously in order to assess the influence of changes in the unattended dimension on the processing speed of the attended one (Garner, 1974). The pitch differences were of a similar relative magnitude as in the study by Singh and Hirsh (1992) (close to 2%). This finding was recently confirmed by Silbert et al. (2009), who speculated that "the locus of interactive effects between f_0 and spectral shape is, in some respect, postsensory", occurring *after* f_0 and spectral shape have been extracted, more or less independently, from the spacing between spectral components (f_0) and their relative amplitudes (shape), respectively.

What about interval sizes other than octaves and microtones, which would be indicative of an effect of timbre not only on pitch height, but also on pitch class? Here, the evidence is more scarce. Krumhansl and Iverson (1992) demonstrated that reaction time interactions, akin to those found by Singh and Hirsh (1992), occur also for fundamentals separated by a tritone (i.e. half an octave). Pitt (1994) extended these results, again for a tritone interval, comparing the performance of musicians and non-musicians. While reaction time effects were observed in both groups, non-musicians (but not musicians) were also prone to pitch misclassifications (same/different) when the timbre changed simultaneously. Somewhat worryingly though, subjects did not only report changes in pitch when the fundamental remained unchanged, they also *failed* to report a true difference in fundamental frequency in 40% of all cases. Considering that the tritone is typically regarded as the most strongly contrasting inverval in Western tonal music, this raises the issue whether the non-musical listeners were simply overburdened with the experimental task, which required them to detect and report changes in both timbre and pitch at the same time for each trial. For example, a strong difference in timbre may have diverted their attention from changes in pitch, rather than exactly compensating for the tritone difference in f_0. This explanation seems particularly likely, as there is no indication that the two different timbres used in the experiment were chosen deliberately to produce an effect on pitch of approximately a tritone.

There is the well-documented phenomenon of a transition from missing-fundamental pitch to a purely spectral pitch as the rank of the lowest harmonic of a missing-fundamental complex tone is increased, especially when the total number of harmonics is low (e.g. Smoorenburg, 1970; Houtsma and Fleuren, 1991; Renken et al., 2004). For natural sounds, however, the fundamental frequency is typically present in the spectrum, and brightness is determined by the degree to which a continuous spectral envelope extends into high-frequency regions, rather than by small isolated groups of high-rank harmonics.

In summary, there is solid evidence for an effect of timbre, and in particular brightness, on the perception of pitch height within the same pitch class under naturalistic stimulus and listening conditions, as well as for small timbre-induced biases within approximately a quarter-tone range around the true fundamental. Evidence for systematic effects of timbre on pitch chroma is less extensive and less convincing in comparison. From a methodological point of view, octave-scale pitch differences would appear to be more readily measurable than microtonal differences close to the limit of discriminability, both experimentally and in a model. We therefore choose to focus on the influence of brightness on the perception of within-chroma octave position throughout the remainder of this chapter.

5.2 Incorporating f_0-dependent timbral characteristics into the Bayesian model

In section 5.1.2, we argued for a dependence between fundamental frequency and spectral centroids in the statistics of natural sounds. Even though we have described an abstract formalism, capable of linking these two quantities in our generative model in principle (cf. section 3.1.2), all results in chapter 4 were obtained using a model in which they were *a priori* uncoupled. The primary goal of this section is to describe, in more practical terms, how to link fundamental and centroid frequency in the model so as to better capture their dependency as measured in a database natural pitched sounds. Several qualitative predictions regarding the influence of brightness on pitch arise from the extended model. In section 5.3, we present new psychophysical data of human listeners in task specifically designed to test these predictions. We will go on to demonstrate that these data are quantitatively well-described by our extended

Bayesian model, and compare its performance to several established models of pitch perception, based on more heuristic computations such as pattern matching of the peripherally-resolved spectrum (Terhardt, 1974; Wightman, 1973) or autocorrelation analysis of evoked auditory nerve responses (Meddis and Hewitt, 1991; Bernstein and Oxenham, 2005; Balaguer-Ballester et al., 2008).

5.2.1 General parametric form

Our initial specification of the Bayesian model (section 3.1.2) allows for a generative coupling between the periodicity Ω and spectral properties of the impulse response $\boldsymbol{f}$ by means of the following two-step procedure:

1. First, an impulse response $\boldsymbol{f}$ is drawn from the distribution $\mathrm{P}(\boldsymbol{f}) = \sum_s \pi_s \, \mathcal{N}(\mathbf{0}, \Psi^s)$, a mixture of Gaussians with diagonal covariance matrices Ψ^s (cf. equation (3.6)).
2. Next, $\boldsymbol{f}$ is convolved with a filter kernel $\boldsymbol{h}_\Omega$, the shape of which can depend on Ω in an arbitrary way. In expectation, $\boldsymbol{f}$ is spectrally white prior to filtering, and hence $\boldsymbol{h}_\Omega$ fully determines its spectral envelope.

We also showed that this procedure is formally equivalent to drawing $\boldsymbol{f}$ from a different distribution $\mathrm{P}_\Omega(\boldsymbol{f}) = \sum_s \pi_s \, \mathcal{N}(\mathbf{0}, \Psi^s_\Omega)$, where each covariance matrix Ψ^s_Ω is obtained by multiplying a diagonal matrix Ψ^s on both sides with H_Ω, the convolution matrix implementing the filter $\boldsymbol{h}_\Omega$: $\Psi^s_\Omega = H^\intercal_\Omega \Psi^s H_\Omega$ (cf. equation (3.14)).

In order to apply this general framework in our model, we need to choose a workable parametrisation of $\boldsymbol{h}_\Omega$: one that is simple enough to fit yet sufficiently powerful to capture those aspects of the statistical dependency of Ω and $\boldsymbol{f}$ in natural sounds that matter the most from a perceptual point of view. We will approach this issue in two stages. First, we choose a family of kernels that is parametrised, for each value of Ω, by only a single shape parameter $\lambda(\Omega)$. Following that, we will show how to fit the scalar function $\lambda(\Omega)$ to a database of instrumental and vocal sounds in section 5.2.2.

We can use the spectra depicted in Figure 5.1 as guiding examples in order to define a set of desiderata for a suitable parametrisation of $\boldsymbol{h}_\Omega$. Firstly, like the spectra of many natural pitched sounds, they are skewed: from a peak in the frequency region around the fundamental, or perhaps a low-ranking harmonic, the spectral envelope decays rather

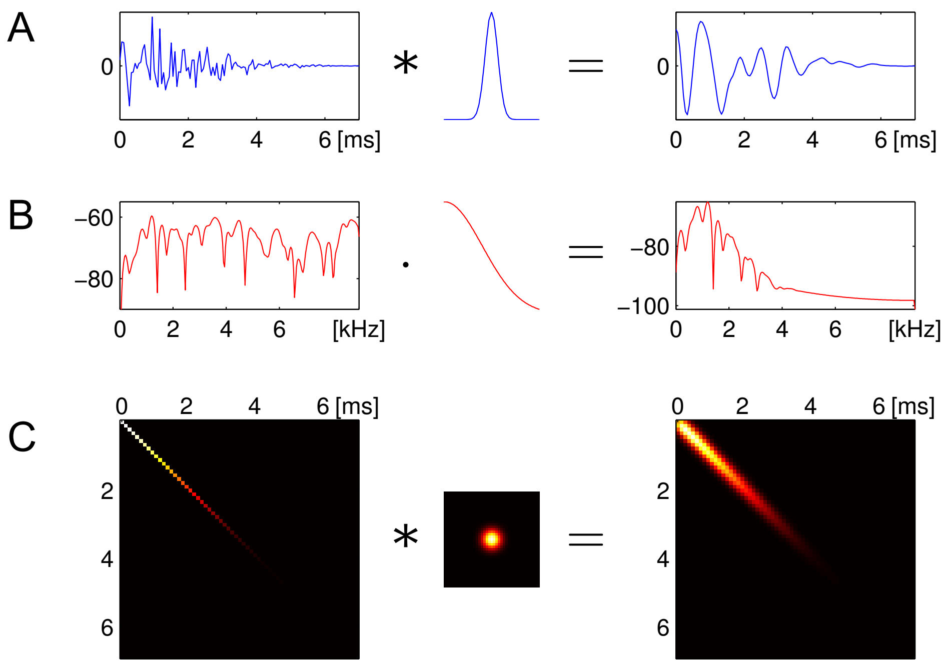


Figure 5.2: Effect of smoothing the model impulse response with a Gaussian kernel. A: Waveform representation of a draw from $\mathrm{P}(\boldsymbol{f})$ before (left) and after smoothing (right) with a Gaussian kernel (centre). B: Spectral representation of the same draw and kernel. C: Effect of smoothing on the covariance matrix of $\mathrm{P}(\boldsymbol{f})$.

smoothly as frequency increases. Secondly, as the pitch increases (5.1, red versus blue curves), the envelopes appear to stretch, rather than shift, while largely retaining their characteristic, skewed shapes. Based on these two observations, we chose to model $\boldsymbol{h}_\Omega$ simply as Gaussians kernels, each with a single shape parameter $\lambda(\Omega)$ corresponding to its width:

$$\boldsymbol{h}_\Omega(t) \propto \exp\left(-\frac{1}{2}\frac{t^2}{\lambda(\Omega)^2}\right) \tag{5.1}$$

The Fourier amplitude spectrum of such a kernel is given by the positive half of Gaussian function, centred at 0 Hz with its spectral width $\hat{\lambda}(\Omega)$ inversely proportional to $\lambda(\Omega)$:

$$\mathcal{F}\{\boldsymbol{h}\}(f) \propto \exp\left(-\frac{1}{2}\frac{f^2}{\hat{\lambda}(\Omega)^2}\right) \tag{5.2}$$

with

$$\hat{\lambda}(\Omega) = \frac{1}{2\pi\lambda(\Omega)} \quad . \tag{5.3}$$

The effect of such a Gaussian-shaped kernel $\boldsymbol{h}_\Omega$ — with an arbitrarily chosen width for now — on a draw from $\mathrm{P}(\boldsymbol{f})$ (i.e. a mixture of Gaussians with diagonal covariances) can be seen in Figure 5.2 (panels A and B). Figure 5.2C demonstrates how the covariance structure of $\boldsymbol{f}$ is changed by this filtering operation. Prior to filtering, the covariance matrix is diagonal (with variances decaying along the diagonal according to equation (3.4)). After filtering with $\boldsymbol{h}_\Omega$, there is a band of positive covariances around the diagonal, dropping off with increasing distance from it, that enforces local smoothness in filtered draws of $\boldsymbol{f}$. Our choice of kernel, as Figure 5.2 goes to show, already meets two of our desiderata: firstly, its spectral envelope peaks at at 0 Hz. When multiplied with a harmonic comb spectrum, such as that of the pulse train $\boldsymbol{\delta}_\Omega$ in our model, the resultant sound will on average have the highest spectral peak at the fundamental and decay with a smooth Gaussian-shaped envelope towards higher frequencies. Our final requirement — that the spectra of higher-pitched sounds should essentially be stretched versions of the spectra of lower-pitched sounds — imposes constraints not on the general kernel shape as such, but on the relationship of its width $\lambda(\Omega)$ on the period duration of the pulse train. In order to meet this requirement, we need to specify $\lambda(\Omega)$ such that it increases with Ω. Since $f_0 = \Omega^{-1}$ and $\hat{\lambda} \propto \lambda^{-1}$, this will cause the spectral width $\hat{\lambda}$ to increase with f_0 and result in the desired stretching effect. From visual inspection of the spectra in Figure 5.1 we can perhaps make the educated guess that $\lambda(\Omega)$ should grow sub-linearly, as a two-octave increase in f_0 appears to result in a stretching of less than two octaves in the envelope in all three cases. The exact relationship between Ω and λ will be determined quantitatively in the following section.

5.2.2 Fitting the timbral f_0-dependence to natural pitched sounds

We used the sounds of 20 orchestral instruments over their entire playing range, provided by the University of Iowa Musical Instrument Samples database[6], as well as recordings of two-octave scales sung on different vowels[7]. For each note, the fundamental frequency (f_0) and the width of the spectral spectral envelope (f_c) were estimated automatically (see Appendix B for the algorithm). We restricted the range of pitches to below 2500 Hz (approximately an E♭7). Firstly, the f_0-estimates became increas-

[6] The instruments contained in this data set are: piano, flute (concert, alto, bass), oboe, clarinet (E♭, B♭, bass), bassoon, saxophone (soprano, alto), french horn, trumpet (B♭), Trombone (tenor, bass), tuba, violin, viola, cello, double bass. Out of three loudness levels available for each instrument, only the medium-loudness samples, without vibrato where applicable, were used. All samples were recorded in anechoic conditions using high-fidelity recording equipment. The samples are freely available at `http://theremin.music.uiowa.edu/MIS.html`.

[7] Scales on the vowels [e], [ɑ], and [ɔ] (closed *"eh"*, open *"ah"* and open *"oh"*), sung by the author and one professional female singer, were recorded in an anechoic room.

ingly prone to octave mistakes at higher frequencies. Secondly, the few samples in this highest f_0 region all originated from only a single instrument (piano).

Figure 5.3 shows a scatter-plot of the f_0 and f_c estimates across our entire ensemble of sounds. A positive correlation between f_0 and f_c is evident from the raw data. In addition, we binned the f_0s into successive whole-tone intervals and computed $\langle f_c(f_0) \rangle$, the mean centroid frequency for each bin (green line in Figure 5.3). We found that the resulting f_0-f_c relationship was well fit by a scaled square-root function,

$$\langle f_c(f_0) \rangle = a \cdot \sqrt{f_0} \quad . \tag{5.4}$$

We fitted the scaling parameter a to our data set, using generalised linear regression to minimise $\sum_i (f_{c,i} - a\sqrt{f_{0i}})^2$, the summed squared error between predicted and observed f_c across all data points (Figure 5.3, solid red line). This approach yielded an estimate of $a = 56.8$. Since this estimate might be unduly dominated by the great number of samples with f_0 below 500 Hz, we also fit a to the f_0-binned means (green line) directly, thereby compensating for the varying density of samples along the f_0-axis. This second estimate agreed with the first within 2%, i.e. almost perfectly. Even though the space of possible functional relationships was not explored systematically any further, we found that a linear relationship, $f_c = m \cdot f_0 + c$ (Figure 5.3, dashed red line) provided a slightly worse fit to the data despite having an additional free parameter (for both for the raw and binned estimate). Considering the visually convincing agreement with our empirical binned centroids, we commited to the square-root relationship of equation (5.4).

Having described the relationship between f_0 and f_c in our ensemble of musical instrument sounds, how should we choose the smoothing time scale $\lambda(\Omega)$ in our model, so as to match these natural statistics? A scaled half-normal distribution of width $\hat{\lambda}(\Omega)$, such as the average spectral envelope of the impulse response $\boldsymbol{f}$ in our model after smoothing, has a mean (i.e. centre of mass) of

$$f_c(\Omega) = \sqrt{2/\pi}\, \hat{\lambda}(\Omega) \quad . \tag{5.5}$$

If we re-express f_c in equation (5.4) as a function of $\Omega = \frac{1}{f_0}$, substitute it into equation

(5.5) and solve for $\hat{\lambda}(\Omega)$, the spectral width, we obtain

$$\hat{\lambda}(\Omega) = a \cdot \sqrt{\frac{\pi}{2\,\Omega}} \quad . \tag{5.6}$$

Using equation (5.3), we can finally solve for $\lambda(\Omega)$, the temporal width of the Gaussian smoothing filter as a function of the period Ω:

$$\lambda(\Omega) = \frac{\sqrt{\Omega}}{a \cdot \pi^{3/2}\,\sqrt{2}} \quad . \tag{5.7}$$

Using the previously determined value of $a = 56.8$, draws from our generative model will show the same average relationship between period (or fundamental frequency) and spectral centroid as observed in our natural sound ensemble. In the following, Ω is typically expressed in milliseconds, rather than seconds: an additional factor of $\sqrt{1000}$ is to be applied to the right-hand side of equation (5.7) in this case, yielding a combined factor of approximately 0.07069. Thus, within 1% of our true estimate of the scaling relationship, we obtain the model fit used throughout the remaining experiments:

$$\lambda(\Omega^{[\,\mathrm{ms}]}) = 0.07\,\sqrt{\Omega^{[\,\mathrm{ms}]}} \quad . \tag{5.8}$$

5.2.3 Qualitative predictions

Based on the nature of the coupling between periodicity and spectral envelope in our generative model, we can make some qualitative predictions regarding the dependence of pitch perception on timbral stimulus features that we would expect to find in the behaviour of a Bayesian ideal observer. Owing to the periodicity-dependence of spectral brightness during sound generation, brightness becomes informative about periodicity during inference. Since the expected spectral centroid increases monotonically with fundamental frequency, the presence of strong high-frequency components in the spectrum is broadly indicative of high pitches, whereas a concentration of spectral energy only at low frequencies points towards low pitches. During inference, these timbral cues need to be weighted against evidence arising from the temporal periodicity of the auditory nerve responses.

It is one of the fundamental properties of Bayesian cue combination that different cues are weighted according to their reliability (e.g. Ernst and Banks, 2002). For periodic

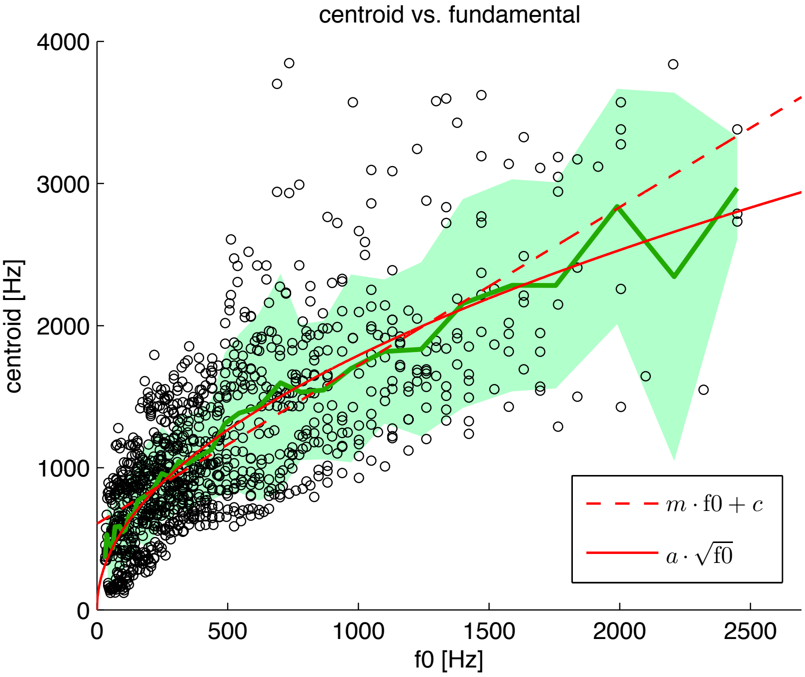


Figure 5.3: Fundamental frequency (f_0) and spectral centroid (f_c) for sounds from a collection of 20 musical instruments and two voices. Black circles represent f_0 and f_c for each tone from the ensemble. The green line shows the mean f_c for each whole-tone bin along the abscissa, the shaded area represents ±1 standard deviation around the mean. Shown in red are the best linear (dotted) and square-root (solid) fits to the raw data set.

sounds that evoke strongly peaked, periodic envelope modulations across many peripheral frequency channels, the temporal cue alone will determine the sound periodicity almost unambiguously and with much greater certainty than the spectral envelope by itself. In cases like this we would expect the combined periodicity estimate to be dominated by the temporal cue and largely independent of timbre. However, we do expect to see a stronger influence of timbre on pitch in cases where a periodicity estimate based on the temporal cue alone is ambiguous: in this case, the centroid of the spectral envelope will be useful in distinguishing between alternative interpretations of the observed data.

Therefore, we should be able to bias a listener's pitch percept of temporally ambiguous sounds (but not that of temporally unambiguous sounds) by manipulating their spectral brightness, provided that the listener judges pitch like an ideal observer that assumes a dependency between periodicity and spectral envelope similar to the one embodied in our coupled model. In the next section, we will present an experiment designed specifically to test this prediction.

5.3 Octave biases in the perception of non-uniform periodic pulse trains

The most common type of errors in pitch judgements, aside from small deviations around the true pitch due to limited discriminability, are octave mistakes. As we discussed in section 5.1.3, these errors are strongly influenced by the brightness of stimulus timbre. Hence, in order to test our hypothesis and model, we sought a stimulus with clearly defined pitch chroma on the one hand, but with high octave ambiguity on the other. One simple class of stimuli which has exactly this property are periodic pulse trains with amplitudes alternating between two values a_1 and a_2. For an inter-pulse interval of duration Δt and (even) pulse amplitudes $a_1 = a_2$, the pulse train is clearly periodic at a rate of $\frac{1}{\Delta t}$. For $a_2 = 0$, the train is obviously periodic at half the rate $\frac{1}{2\Delta t}$. For any other combination of amplitude $a_1 \neq a_2$, the pulse train is also periodic at this slower rate, but it might be mistaken for the high rate perceptually, especially if a_1 and a_2 are almost identical, and in the presence of added noise. Thus, by varying the relative amplitude $r = \frac{a_1}{a_2}$ from $1 \rightarrow \infty$, one would expect subjects to change their pitch judgements from $\frac{1}{\Delta t}$ (high) initially to $\frac{1}{2\Delta t}$ (low) once r becomes sufficiently large, moving through a region of high ambiguity in-between. An experiment just like this was performed by Flanagan et al. (1962), who found that the pitch percept dropped by an octave for values of $\frac{a_1}{a_2}$ in excess of 6 to 10 dB (see Figure 5.4). Importantly, this initial plateau of the psychometric curves is unlikely to result from a basic failure to detect the amplitude alternations: Pollack (1971) measured jitter detection threshold for square waves with amplitude jitter added to each half-wave, showing that relative jitter amplitudes of few percent were detectable for stimuli of comparable duration and period length.

We adapted Flanagan's stimulus paradigm in order to test our hypothesis, that stimulus timbre should influence the perceptual octave-boundary between high- and low pitch, and that it should do so specifically in cases where ambiguity regarding the stimulus period is high. Listeners had to judge the pitch of alternating-amplitude pulse trains and harmonic complex tones with different timbral characteristics as either high or low compared to flanker sounds which had a pitch half-way in-between.

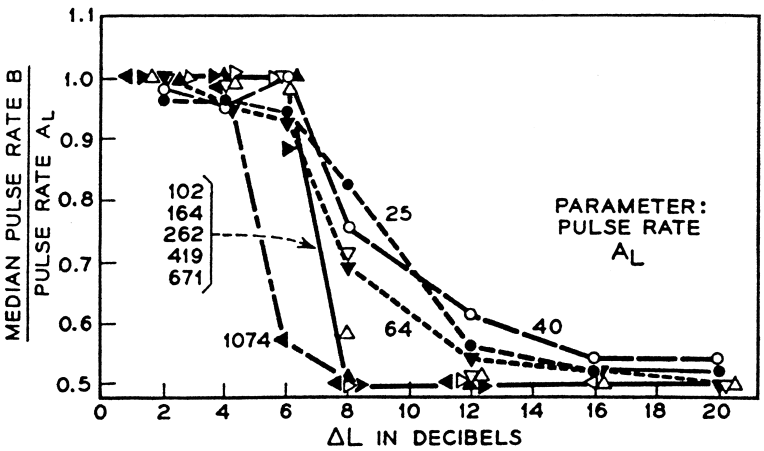


Figure 5.4: Results of matching the pitch of a regular, uniform pulse train (pulse-rate B) to that of train of pulses with alternating amplitude (pulse rate A_L). Different curves represent different pulse rates A_L. For most pulse rates used in the experiment, the matched pulse rate B drops by an octave from A_L to $A_L/2$ as the amplitude ratio reaches approximately 8 dB. Figure from Flanagan (1972), results originally published in abstract form in Flanagan et al. (1962).

5.3.1 Experimental methods

5.3.1.1 Participants

Seven subjects, including the author, took part in the experiment. All subjects were graduate students at University College London, between 23 to 30 years old and male. Participation was voluntary and without monetary reward. No subject had previously been diagnosed with a hearing disorder or reported any subjective hearing impairments. Three subjects played a musical instrument around the time of the experiment, a fourth subject had had musical training in the past. Excluding the author, subjects had not previously participated in psychoacoustic experiments, and were fully naive with regards to the purpose and expected outcome of the task.

5.3.1.2 Stimuli

In each trial, the stimulus consisted of a target sound, preceded and followed by two identical flankers. Target sounds were either alternating click trains (ACTs) or harmonic complex tones (HCTs). The ACT targets had an inter-click interval of 2 ms. Individual clicks were bipolar with a duration of 0.3 ms. The click amplitudes alternated periodically between two values a_1 and a_2. Thus, for values $a_1 = a_2$, the ACTs were periodic at a rate of 500 Hz. For values $a_1 \neq a_2$, they were periodic at a rate of

250 Hz, i.e. one octave lower. Independent of the click amplitude ratio $r = \frac{a_1}{a_2}$, the click trains had a similar, broad-band spectral envelope, which we refer to as having a "broad" timbre (Figure 5.5A). In order to manipulate their timbral brightness, some ACT stimuli were low-pass filtered at 3 kHz (36 dB/octave, 6th order Butterworth filter), in the following referred to as having a "dark" timbre (Figure 5.5B). Harmonic complex tones had fundamental frequencies of either 250 or 500 Hz with equal-amplitude partials in sine phase. In order to match the two different timbres of the ACT stimuli, even-amplitude partials extended from the fundamental up to either 20 kHz (broad) or 3.5 kHz (dark).

Flankers were always missing-fundamental HCTs with a fixed f_0 of 353 Hz, exactly half an octave in between the two possible target fundamental frequencies of 250 and 500 Hz. Flankers comprised the partials 2 to 20, in sine phase and with equal amplitude. Containing many peripherally resolved partials, the flankers were expected to evoke a strong, unambiguous pitch of 353 Hz.

Both targets and flankers were presented in a background of low-pass filtered noise with a cutoff frequency near 1 kHz, in order to mask possible mechanical distortion products at either 250, 353 or 500 Hz (see section 2.2.3.3). The level of the masking noise in relation to the target and flanker level was chosen such that the overall signal-to-noise ratio (SNR) was either 0 or -6 dB (Figure 5.5C). In any given trial, the SNR and absolute levels were identical for the target and flankers.

Each subject was presented 30 samples from each of 19 different target stimulus conditions, as summarised in Table 5.1. 15 different ACT conditions were used, comprising five different amplitude ratios (1, 1.69, 2.25, 2.89 and 4) for each of three combinations of noise level and timbre (0 dB SNR, broad; -6 dB SNR, broad; 0 dB SNR, dark). The amplitude ratios were based on previous pilot data (from the author and a different group of subjects), which indicated a transition of the perceived pitch for the ACT stimuli from 500 to 250 Hz within this range. The remaining four stimulus conditions comprised HCTs with a pitch of either 250 or 500 Hz and broad (harmonics up to 20 kHz) or dark timbre (harmonics up to 3.5 kHz). Overall presentation level was kept constant across trials. Target and flankers had a duration of 60 ms, separated by 100 ms of silence.

2 ms alternating click train						
SNR	timbre	amplitude ratio				
0dB	broad	1	1.69	2.25	2.89	4
-6dB	broad	1	1.69	2.25	2.89	4
0dB	dark	1	1.69	2.25	2.89	4

harmonic complex tone			
SNR	timbre	f_0	
0dB	broad	250 Hz	500 Hz
0dB	dark	250 Hz	500 Hz

Table 5.1: Stimulus conditions used in the experiment. The 15 ACT conditions comprised five different amplitude ratio for each of three different combinations of SNR and timbre. Four further conditions comprised harmonic complex tones with either low or high f_0 and broad or dark timbre.

Stimuli were generated digitally in Matlab[8] with a sampling rate of 44.1 kHz and presented, using the Psychophysics Toolbox extension (Brainard, 1997; Kleiner et al., 2007), in a quiet room via circumaural headphones with high passive noise attenuation (Sennheiser HDA 200) at a comfortable listening level. Listening levels chosen by the subjects ranged from 70 to 78 dB SPL.

5.3.1.3 Task

In each trial, subjects were asked to judge the shape of the melodic contour of an A-B-A sound triplet as either rising-falling or falling-rising, where A is flanker and B a target stimulus as described above (see Figure 5.6). Trials for all 19 target conditions were intermixed randomly. Subjects listened to the sounds via headphones and responded either by clicking one of two buttons depicting the two melodic contours on a computer screen with a mouse, or by pressing one of two corresponding keys on the keyboard. Reaction times were unconstrained, the intertrial interval between response and onset of the next stimulus triplet was 1 s. No feedback was given during the experiment. Subjects were told that the purpose of this experiment was to probe their subjective perceptual experience and that there was no strictly right or wrong response in any trial. They were further informed that they were likely to experience ambiguous stimuli, in which case they should make a quick decision according to their best judgement. Furthermore, subjects were informed that the stimuli were independent from trial to trial and not influenced by their previous choices in any way, and that long sequences of identically shaped triplets may occur simply by chance. Before the start of the main task, subjects were given the opportunity to familiarise themselves with the user interface in two short test runs. The first of these contained only broadband ACT

[8] MATLAB® version 7.10.0 (R2010a), The Mathworks Inc., Natick, Massachusetts

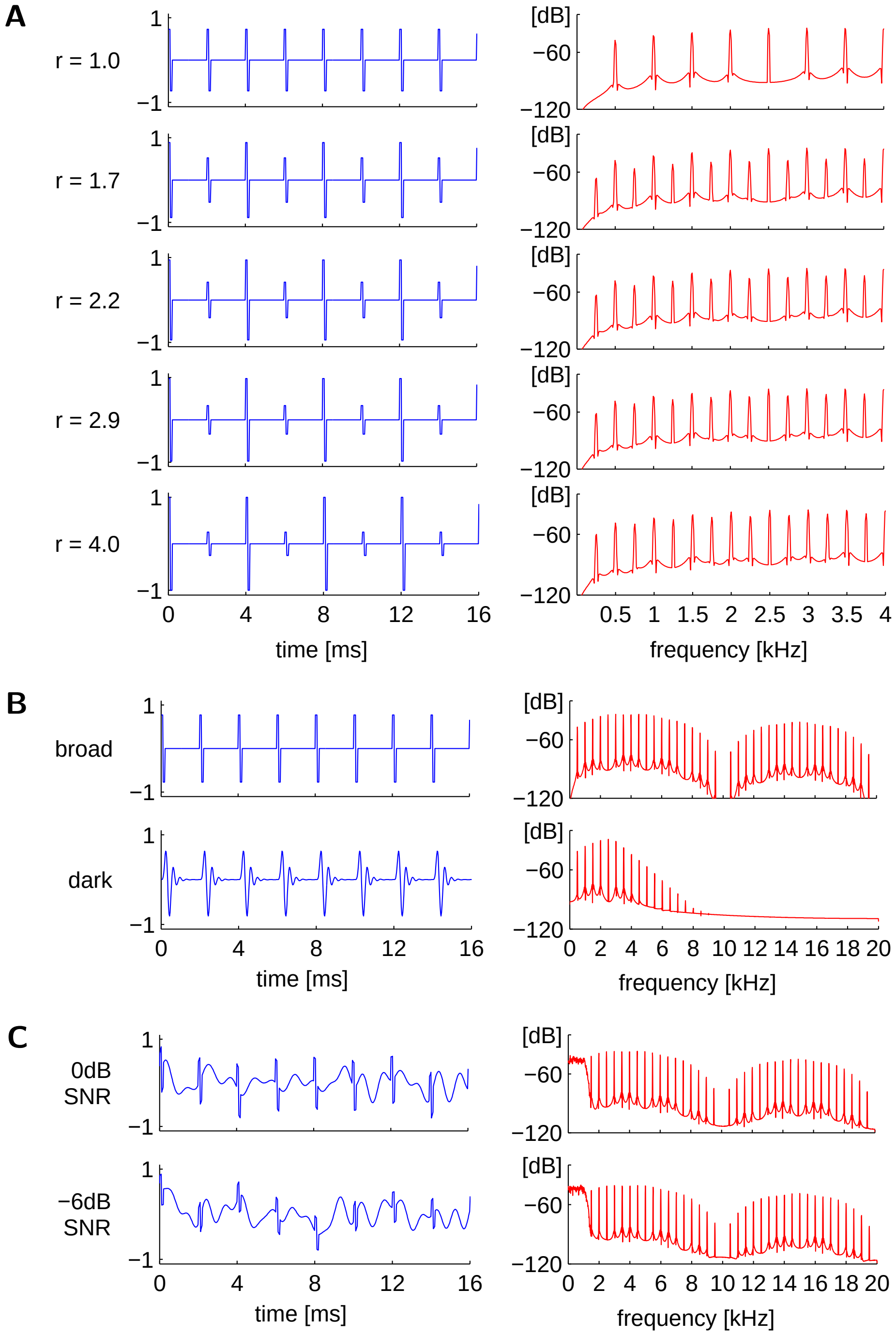


Figure 5.5: Effect of manipulating amplitude ratios (A), spectral envelope (B) and noise level (C) on the stimulus waveform (left column) and power spectral density (right column). A: Pulse trains with the five amplitude ratios used in the experiment, shown without noise (the spectra are shown only for the range 0 – 4 kHz) B: Even-amplitude pulse train ($r = 1$) with (top) and without (bottom) low-pass filtering at 3 kHz, shown without noise. C: Unfiltered, even-amplitude pulse train ($r = 1$) with noise added at SNRs of 0 dB (top) and -6 dB (bottom).

targets (0 dB SNR) with amplitude ratios of either 1 or 9 . These were expected to give rise to relatively unambiguous pitches of 500 and 250 Hz, respectively. The second test run contained broadband ACT with (more ambiguous) amplitude ratios 1, 2.25 and 4, again present with 0 dB SNR. Both test runs together lasted approximately 10 minutes. Before the start of the main task, subjects were informed that they would experience be a greater variety of different sound colours and qualities than during the test runs.

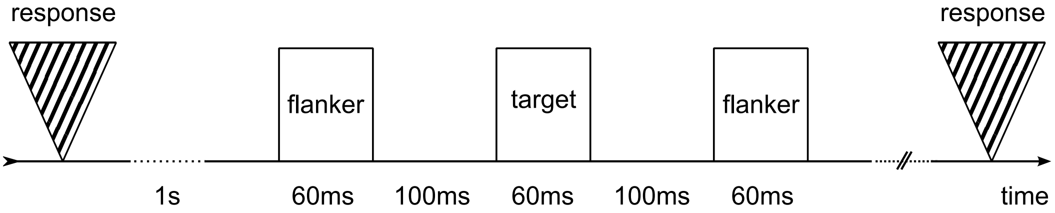


Figure 5.6: Task design: after an inter-trial interval of 1s, a sound triplet consisting of a target surrounded by two identical flankers is presented, following which subjects have to judge the melodic contour as either "rising-falling" or "falling-rising".

5.3.2 Psychophysical results

For each target condition and subject, we computed the fraction of trials in which the contour was perceived as rising-falling. A "rising-falling" response implies that the pitch of the target was perceived as higher that the flankers, a "falling-rising" response implies that it was perceived as lower. Furthermore, the likely pitches for any target, predicted from virtually any theory or model, are either 250 or 500 Hz. Hence, we assume throughout the following that we can equate a "rising-falling" responses with a pitch percept of 500 Hz ("high"), and a "falling-rising" response with a pitch of 250 Hz ("low").

For the ACT targets, we generated a set of three psychometric curves for each subject, showing the fraction of high-pitched targets as a function of the click amplitude ratio r for each of the three timbre conditions (0dB broad, -6dB broad, 0dB dark). These curves were averaged across subjects.

Two subjects – one with previous musical training and one without – were excluded from this average, based on their responses to the four HCT targets conditions: responses were close to chance in each case (as well as for the ACT stimuli), even though both targets and flankers were expected to evoke a clear pitch percept (Figure 5.7). This indicates that they were generally unable to reliably distinguish between pitch octaves in our task setting, even under favourable stimulus conditions. The five remaining subjects (three with musical training, two without) judged the pitch of the HCTs consistently (within and across individual).

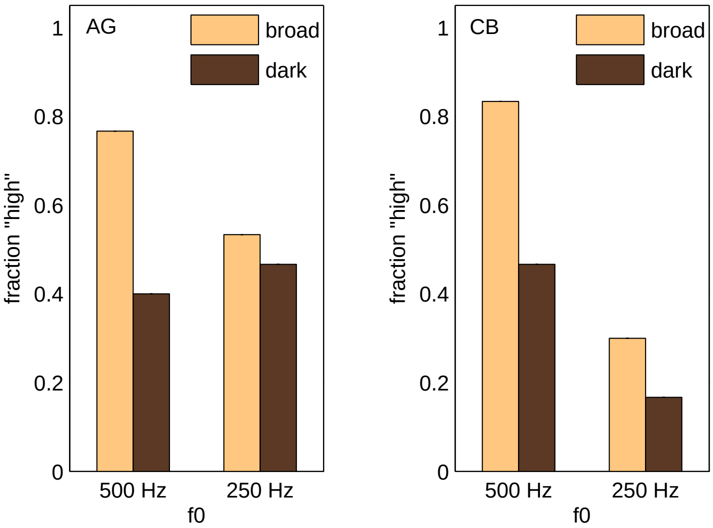


Figure 5.7: Two subjects were excluded from further analysis based on their low reliability in identifying the octave of the complex harmonic tone stimuli (fundamentals of 250 and 500 Hz, broad and dark timbre). Bars represent the fraction of targets judged as high-pitched (chance level 0.5).

5.3.2.1 Timbral effects on the pitch of alternating click trains and harmonic complex tones

Psychometric curves, averaged across the five included subjects, are shown in Figure 5.8A. Starting with the broadband ACT stimuli at 0 dB SNR (blue curve), we observe that subjects judge the pitch of the target as predominantly high for click amplitude ratios up to 2.25 (cf. Figure 5.5A), reaching the 50% level at around 2.89 and finally dropping to below 10% for our maximum ratio of 4. Comparing our psychometric curves with the results reported by Flanagan et al. (1962) and Flanagan (1972), we find that the two datasets are in good agreement, despite several differences in the details of the stimuli employed. While our click trains were bipolar and presented in a background of low-pass noise, Flanagan used unipolar click trains, presumably without added noise[9].

[9] To the best of our knowledge, the most complete published description of Flanagan et al.'s results is a figure in Flanagan (1972) (reproduced in Figure 5.4), the other source being a conference abstract (Flanagan et al., 1962). Both sources lack methodological detail, but we can assume the absence of noise based on other publication by the authors made around the same time (e.g. Flanagan and Guttman, 1960).

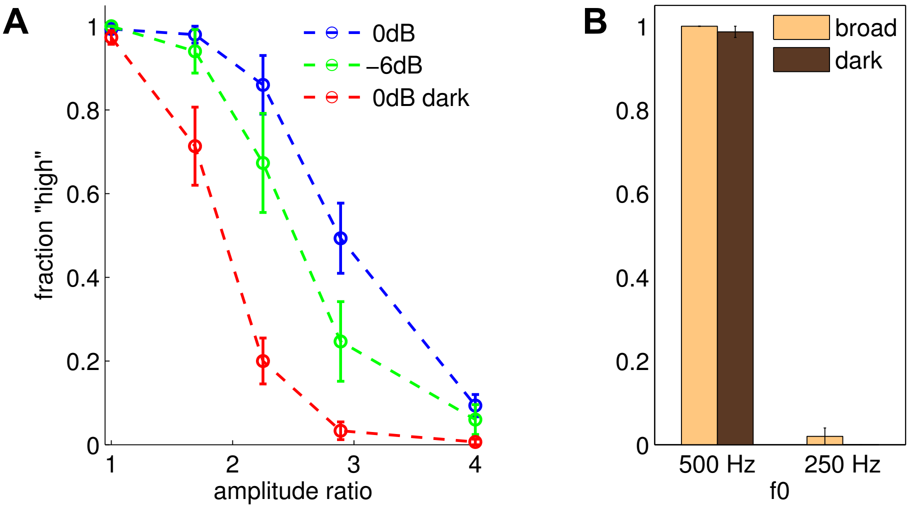


Figure 5.8: High-low pitch judgements, averaged across five subjects: mean fraction of target judged as "high" (melodic contour "rising-falling") ± 1 SEM. A: Responses to ACT targets with varying click amplitude ratios (abscissa). Stimulus conditions were: broadband timbre with low noise level (blue), broadband timbre with high noise level (green) and dark timbre with low noise level (red). B: Responses to complex harmonic tone target stimuli with matched fundamental frequencies (250 / 500 Hz) and timbres (broadband / dark).

In Flanagan's study, subjects reported their percept by matching an even-amplitude pulse train to the target sound. The median matched pitch, depicted in Figure 5.4, dropped by a factor of two for amplitude ratios of 8 dB (approximately 2.5) and greater at comparable pulse rates. Note that, while the *fraction* of "high" responses in our data (individuals and average) drops off more gradually than Flanagan's median matched pulse rates, we would obtain a similarly sharp transitions for median reported pitch in our data.

In comparison to the broadband stimulus, the reported pitch for low-pass filtered click trains drops from high to low for considerably lower amplitude ratios (5.8A, red curve). While the pitch of the even-amplitude, 2 ms pulse train ($r = 1$) consistently remains at 500 Hz despite the change in timbre, the fraction of "high" responses has dropped by almost 30% at the next-highest value tested ($r = 1.69$) and the percept is predominantly low for all remaining $r \geq 2.25$. Decreasing the power of the broadband click stimulus relative to the low-pass masking noise from 0 dB to -6 dB (green curve), also has a biasing effect towards the lower pitch (compared to the 0 dB broadband stimulus) but the effect is overall less pronounced than that of darkening the timbre by low-pass filtering the click train.

Figure 5.5B shows the averaged pitch judgements for the complex harmonic tone stimuli. Presented in identical 0 dB low-pass noise, the same timbral manipulation that induced a significant change in the perceptual octave bias for click train stimuli had virtually no effect on subjects' pitch judgements: the perceived pitch equals the fundamental frequency (250 and 500 Hz) irrespective of the frequency of the highest harmonic present in the spectrum (20 kHz and 3.5 kHz).

Our psychophysical results are in general agreement with previous studies that reported timbre-induced octave-biases in the perception of pitch (Hesse, 1982; Robinson, 1993, cf. also von Helmholtz, 1863). They are also in good qualitative agreement with our more specific predictions based on the idea of an ideal observer, that optimally combines periodicity- and spectral envelope-based pitch cues, assuming a statistical dependency of form described in section 5.1.2 between the two. Not only is there an overall greater tendency for stimuli with dark timbre to be perceived as lower in pitch: crucially from the standpoint of an ideal observer, this timbral biasing becomes apparent *especially* when the periodicity of the stimulus itself is inherently ambiguous. In our data, the effect of timbre is strongest for click train stimuli with intermediate click amplitude ratios, and wholly absent in both the even-amplitude click trains ($r = 1$) and the complex harmonic tones. The consistent, timbre-independent pitch percept of the harmonic tones argues against a *general* perceptual confusion between periodicity and timbral characteristics in our subjects. While we cannot strictly rule the interpretation that a confusion occurred selectively for some stimuli but not others, we would argue instead that subjects were still trying to infer periodicity, but were doing so using information contained in the spectral envelope shape – owing to its correlation with periodicity in the statistics of natural sounds – where periodicity judgements based on stimulus self-similarity alone were in-determinate. This latter interpretation of the results, based until now merely on a plausible but somewhat vague intuition, will be strengthened by the good quantitative agreement between our psychophysical data and the periodicity estimates of our Bayesian model, that instantiates these intuitions in a formal way.

5.4 Pitch-timbre interactions in models of pitch

5.4.1 Bayesian model: coupled and uncoupled

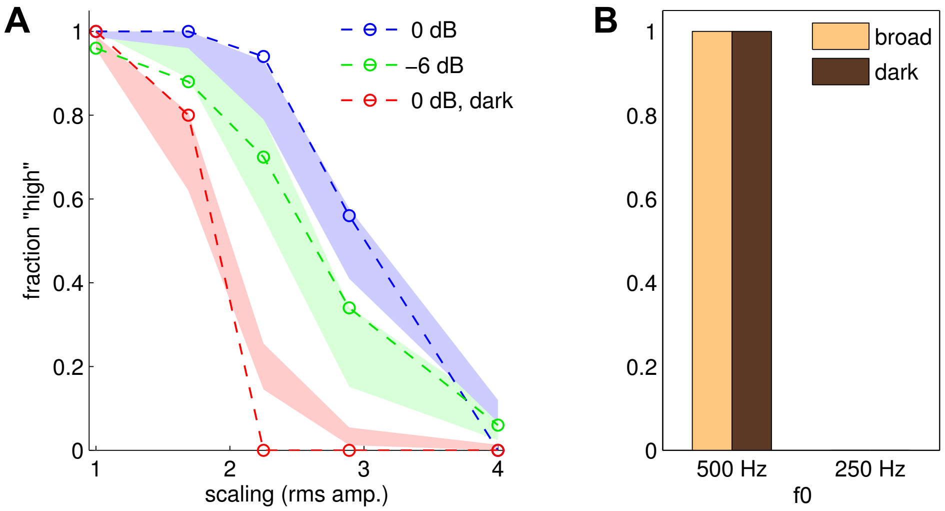


Figure 5.9: Periodicity estimates of the Bayesian model with prior dependency between periodicity and spectral envelope. A: Click train stimuli; dotted lines represent model predictions, shaded areas indicate subjects' mean psychometric curves ±1 SEM (cf. Figure 5.8). B: Harmonic complex tones.

We used the same set of stimuli as in the psychophysical experiments to compute periodicity estimates under the Bayesian model. Periodicity Ω and spectral extent of the acoustic impulse response $\boldsymbol{f}$ were linked as described in section 5.1.2: the time-domain width $\lambda(\Omega)$ of the low-pass filter applied to the initial draw of $\boldsymbol{f}$ increased proportional to the square root of Ω: $\lambda(\Omega) = 0.07\sqrt{\Omega}$ (Ω and λ in ms). The possible time-scales of the envelope of $\boldsymbol{f}$ were chosen in 8 steps between 0.1 ms (shorter than a single impulse) and 4 ms (the period length for $\Omega = 250$ Hz). The peripheral filter bank had 80 frequency channels, spaced between 40 and 16 kHz. We used a sampling rate of 44 kHz, rather than 44.1 kHz, to ensure that both a periodicity of 250 and 500 Hz would align exactly with the sampling grid while changing the temporal and spectral characteristics of the stimulus as little as possible. The stimulus duration was 60 ms, as in the psychophysical experiment. For every stimulus, we computed the log-likelihoods $\ln \mathcal{L}(\Omega = 2\,\text{ms})$ and $\ln \mathcal{L}(\Omega = 4\,\text{ms})$ (using our Laplace-approximation based inference scheme, cf. section 3.3.1), and classified the inferred periodicity as "high" when $\ln \mathcal{L}(\Omega = 2\,\text{ms}) > \ln \mathcal{L}(\Omega = 4\,\text{ms})$. 50 repetitions of each stimulus condition were used to compute the fraction of "high" responses.

Figure 5.9A shows the outcome of our estimation for the click train stimuli. We observe a close agreement between psychophysical data (shaded areas) and model estimates for all three timbral conditions. Like human listeners, the model judges the pitch of the 0 dB broadband stimuli (blue) to be predominantly high for click amplitude ratios of $r \leq 2.25$. For $r = 2.89$, the model reports high and low pitch with almost equal probability. For $r = 4$, where human subjects report the low pitch with $\approx 90\%$ probability, the model reports the pitch as exclusively low. Low-pass filtering of the click trains (red) has a strong effect on octave preference in the model for intermediate amplitude ratios ($1.69 \leq r \leq 2.89$), but none for the even-amplitude stimulus ($r = 1$) or the highest amplitude ratio, $r = 4$ (where a small effect remains for human subjects). The greatest deviation between model and human behaviour occurs for $r = 2.25$, where the model's preference for the high pitch drops to 0, whereas humans still report approximately one in five stimuli as high-pitched. As in human subjects, an increase of the noise level (green) results in a noticable bias towards hearing the lower pitch, though less pronounced than the bias due to low-pass filtering the click trains.

The Bayesian model is also consistent with human pitch judgements of the complex harmonic tone stimuli: the true stimulus periodicity is inferred correctly irrespective of changes in timbre (Figure 5.9B). In summary, the modelling results match our qualitative expectations regarding the effects of timbre on pitch in an ideal observer, as well as providing an accurate account of human psychophysical performance.

A natural question to ask is to what degree the success of the model depends on the coupling of pitch and timbre in the generative process, which we introduced in this chapter. Figure 5.10 shows the results of the Bayesian model with pitch and timbre uncoupled (i.e. $\lambda(\Omega) \equiv 0$). We can immediately see that the fit between model and psychophysical data deteriorates in the absence of this coupling. The model is now noticeably biased towards reporting the low pitch for intermediate values of r, while the additional biasing effect of low-pass filtering the click stimulus is small compared to human behaviour and the coupled model (5.10A, blue and red curves). Furthermore, increasing the low-pass noise while keeping the timbre broad has a qualitatively different effect in the uncoupled model than in subjects and the coupled model: pitch judgement are more biased towards the *high* pitch than in the low-noise broadband condition (green and blue curves). Pitch judgements of the model for the complex tone stimuli are unaffected by the change of the model prior: pitch ratings are independent of our timbral manipulations – as one would expect given their prior independence in the model. One more aspect of the model behaviour is worth pointing out: even though there no longer is a spectrally induced bias for reporting the high pitch in case of the alternating click trains, the model still favours the high pitch for values of r up to 1.69, rather than its true periodicity (i.e. the low pitch). Since we have been arguing, that our model infers the stimulus periodicity *optimally*, does this not constitute a worrisome failure? The explanation for this somewhat paradoxical behaviour lies in another feature of our prior distribution over the acoustic impulse response $\boldsymbol{f}$. In our model, $\boldsymbol{f}$ is expected to decay at some unknown, but finite rate τ. In order to explain a waveform with intermediate amplitude ratio, e.g. $r = 1.69$, with a periodicity of 4 ms (low), the model needs to assume a biphasic envelope for $\boldsymbol{f}$ (i.e. one with two distinct local maxima), whereby each of the two pulses corresponds to a peak in the envelope. This conflicts with the prior favouring impulse responses with monotonic, and hence monophasic, envelopes. Conversely, there is no such conflict when we assume an underlying periodicity of 2 ms. This bias for monophasic responses, inherent in the model, gives rise to a slight preference for reporting the high pitch despite stronger evidence for the low periodicity in the self-similarity pattern of auditory nerve response itself. Of course, this holds only up to some value of r, beyond which where this evidence becomes overpowering (while at the same time the cost for having the second peak under prior decreases, as it becomes lower for increasing values of r). The exact point at which this change in preference occurs, depends to some degree on the amount of noise — acoustic and neural — assumed in generative process. As in the coupled model before, we set the acoustic SNR in the model to the mean SNR of our stimulus ensemble, while the noise variance in the auditory nerve was low compared to the average firing rate of the active fibres. The results of neither the coupled nor the uncoupled model depended critically on the fine-tuning of these parameters.

5.4.2 Pattern matching: Terhardt

We implemented Terhardt's spectral pattern matching model of virtual pitch (Terhardt, 1974; Terhardt et al., 1982) as described in section 2.4.1.2. In short, the model extracts peaks in the Fourier spectrum, adjusting their relative hights by an overall spectral weighting function as well as local competition between nearby peaks. Each peaks then contributes evidence for a pitch not only at its own frequency but also at its subharmonics. For harmonic spectra, this process results in the greatest overall

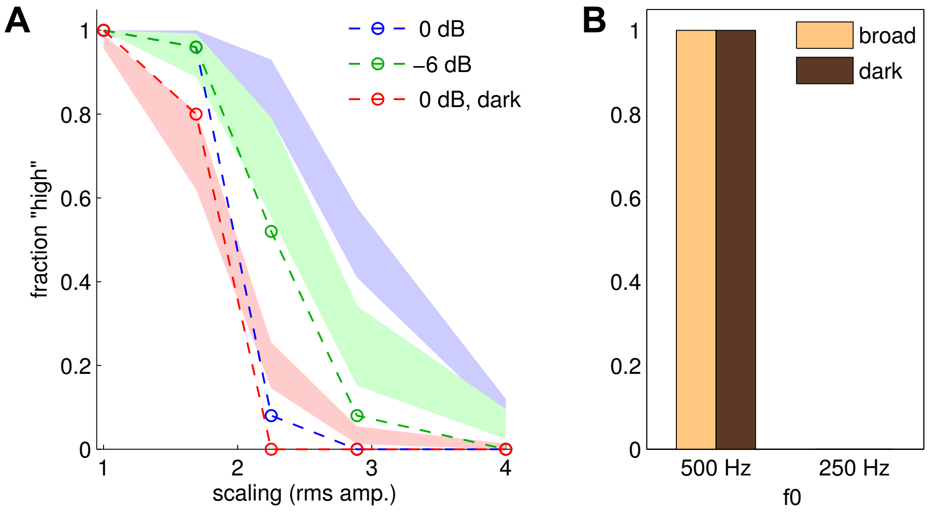


Figure 5.10: Periodicity estimates of the Bayesian model *without* prior dependency between periodicity and spectral envelope. A: Click train stimuli; dotted lines represent model predictions, shaded areas indicate subjects' mean psychometric curves ±1 SEM (cf. Figure 5.8). B: Harmonic complex tones.

accumulation of evidence at the fundamental frequency of the harmonic series.

Applied to our stimulus set, Terhardt's model succeeds in predicting the pitch of the complex tone stimuli and its timbre independence (Figure 5.11B). We found, however, that it doesn't match human behaviour for the ACT stimuli well: for amplitude ratios $r > 1$, the model immediately predicts the pitch to drop to 250 Hz, irrespective of changes in timbre or noise level (5.11A). Thus, the model appears to reflect the true periodicity of the acoustic stimulus, which drops from 2 ms to 4 ms as soon as r deviates from 1, rather than human perception. Since model predictions already drop to the lower pitch for all $r \neq 1$ in case of the 0 dB broadband stimulus (blue), we wouldn't be able to measure a timbral biasing effect in our stimulus ensemble even if it did exist in the model. In principle, one might expect such effects to occur in Terhardt's model: each spectral peak contributes only to a limited range of subharmonics during the process of subharmonic summation, and thus a range of high-frequency peaks should contribute evidence only to the high pitch interpretation, but not the low, of the spectral profile at hand. However, we did not perform a more fine-grained analysis of the model prediction in the range of $1 < r < 1.69$ to test for this. Furthermore, as Terhardt limited the initial spectral analysis in the model to frequencies up to 5 kHz, the model arguably is effectively blind towards our low-pass filtering of the ACTs at 3 kHz. We therefore repeated the predictions with a frequency range extending up to 10 kHz, but found that the model behaviour remained exactly the same.

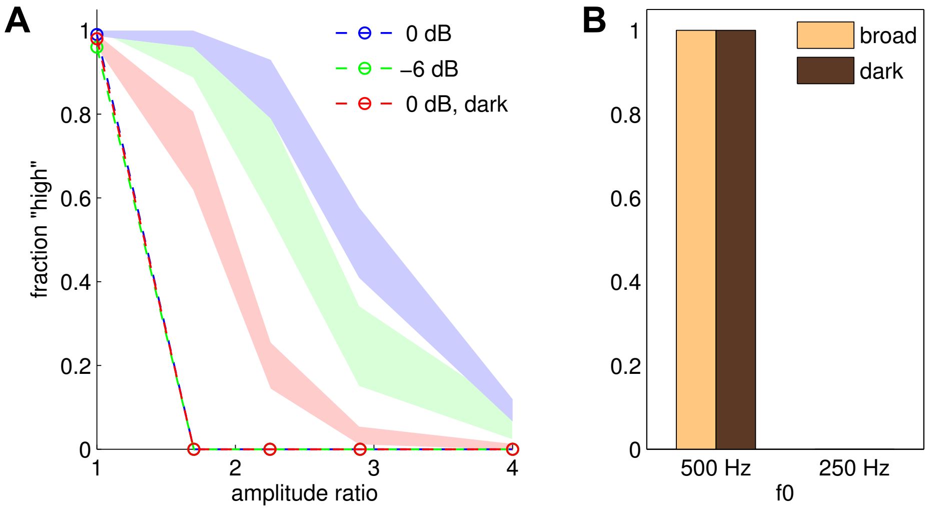


Figure 5.11: Virtual pitch estimates from Terhardt's pattern matching model. A: Click train stimuli; dotted lines represent model predictions, shaded areas indicate subjects' mean psychometric curves ±1 SEM (cf. Figure 5.8). B: Harmonic complex tones.

5.4.3 Pattern transformation: Wightman

Wightman's pattern-transformation model of pitch (Wightman, 1973) uses a peripheral, spectral representation of the stimulus as the basis for its pitch estimates (cf. section 2.4.1.1). It interprets the time-averaged peripheral firing rate profile along a tonotopic axis as a substitute for the true Fourier power spectrum of the acoustic stimulus. Applying a Fourier transform to this peripheral spectrum, a degraded estimate of the stimulus autocorrelation function is obtained[10]. The estimated stimulus periodicity, or pitch, is determined by the highest peak in this surrogate autocorrelation function.

When we applied the pattern-transformation model to the output of our simple peripheral model, we were surprised to find that the model reported exclusively the high pitch for *all* click train stimulus conditions (data not shown). Closer inspection revealed the source for this mode of failure. Due to the tuning width of the peripheral filters, the spectral peaks that correspond to the odd harmonics of the 250 Hz fundamental are too low in amplitude compared to the even harmonics (i.e. multiples of 500 Hz) in order to evoke discernable peaks in the average firing-rate profile. They are effectively unresolved by the peripheral filter bank, even in the low-frequency range – hence the aberrant model behaviour. It has been argued that peripheral frequency resolution in humans may be considerably higher than classical psychophysical estimates, de-

[10] The estimate would be exact if it was based on the true power spectrum, according to the Wiener-Khinchin theorem

rived from notched-noise masking experiments (e.g. Glasberg and Moore, 1990), would suggest. Bandwidth estimates based on otoacoustic emissions as well as forward masking effects at low sound intensities (Shera et al., 2002; Oxenham and Shera, 2003) indicate that human frequency resolution may in fact be up to twice as sharp as typically assumed, even though these claims have subsequently been disputed (Ruggero and Temchin 2005; see also Moore and Gockel 2011 for a recent review on peripheral resolvability). Since our peripheral model has been based on the wider bandwidth estimates of Glasberg and Moore (1990) throughout, we repeated our modelling attempts with peripheral bandwidths reduced by a factor of two while increasing the number of channels to 200, in order to test whether Wightman's pattern transformation approach could explain our psychophysical findings when applied to a more sharply defined peripheral spectral profile. Figure 5.12 shows that even in this setting, the model retains its strong bias to judge the pitch as of the click train stimuli as high across all stimulus conditions. We can observe an effect of timbre on pitch that follows the same ordering as that observed in human listeners for $r = 4$, but overall the quantitative discrepancy between model and human data remains high.

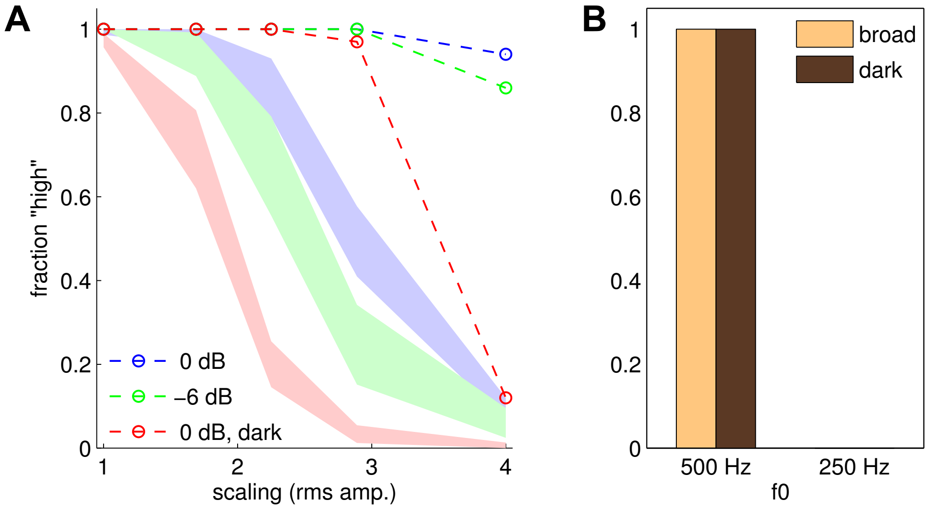


Figure 5.12: Pitch estimates from Wightman's pattern transformation model. A: Click train stimuli; dotted lines represent model predictions, shaded areas indicate subjects' mean psychometric curves ±1 SEM (cf. Figure 5.8). B: Harmonic complex tones.

Comparing the models of (Terhardt, 1974) and (Wightman, 1973), we observe that they mark almost opposite poles of a spectrum, with a strong bias towards the low pitch in the former and an equally extreme bias towards the high pitch in the latter. This may seem surprising at first sight, as both models are often treated as closely related. This notion seems to be largely due to de Boer (1977), who compared the three spectral pitch models of Terhardt, Wightman and Goldstein (1973). He showed that the former two can be seen as two different limiting cases of the same approximation to Goldstein's less tractable "optimum processor" model (see section 2.4.1.3). Despite this unifying theoretical framework, the different limiting assumptions inherent in Terhardt's and Wightman's models can apparently lead to markedly different predictions: this is evidenced by our modelling results and seems to confirm Terhardt's own reservations against de Boer's unifying treatment of these models (Terhardt, 1977).

5.4.4 Summary autocorrelation

"Summary autocorrelation" is a general principle underlying a whole family of related models (e.g. Meddis and Hewitt, 1991; Meddis and O'Mard, 1997; Bernstein and Oxenham, 2005; Balaguer-Ballester et al., 2008). They all involve the computation of some form of temporal autocorrelation function (ACF) of the neural firing pattern in each channel of a peripheral band-pass filter bank. A "summary autocorrelation function" (SACF) is obtained, essentially by summing the individual ACFs across channels, and the final periodicity judgement is based on this (see section 2.4.2). Instantiations of this principle can vary in the implementation of the peripheral neural response to the acoustic stimulus, the ACF computation and summation stage as well the read-out mechanism operating on the final SACF. It is therefore difficult to prove or disprove the principle of summary autocorrelation in general. We chose two variations of this model, one at the simplistic and one at the complex end of the spectrum of possible implementations, to investigate whether our psychophysical data might be in accord with the principles of summary autocorrelation.

Our first implementation is based on the same peripheral model that we assumed in the generative process underlying our Bayesian estimation (cf. section 3.2): a linear gamma-tone filter bank (80 channels from 40 Hz to 16 kHz) followed by envelope demodulation. For each channel, we compute the autocorrelation function across the entire duration of the stimulus ($\lambda = \infty$ in equation (2.9)) after removing the channel mean. Individual ACFs are summed with equal summation weight for each frequency channel and time-lag of the SACF. We determine the model response by comparing the maximum heights of the SACF in two narrow ranges around 2 ms ("high") and 4 ms ("low"). This may seem like an overly simplistic use of the SACF at first sight: in the context of pitch *matching* paradigms, the periodicity decision is sometimes made by finding the sound whose SACF is most similar to that of the target stimulus[11]. In our psychophysical paradigm, however, there is no comparison stimulus against which the target SACF could be matched and thus this mode of read-out is not readily applicable. Instead, we essentially ask whether the SACF peak occurring at a lag of $\frac{1}{353}$ s during the flanker stimulus (owing to its periodicity being half an octave between 250 and 500 Hz) shifts towards a shorter or longer lag during target presentation. This form of read-out is in agreement with Meddis and Hewitt, 1991 and in our opinion reflects a reasonable choice, given our paradigm based on judging the pitch contour of a three-tone "melody" in the absence of direct reference stimuli (250 and 500 Hz) for comparison of the second tone.

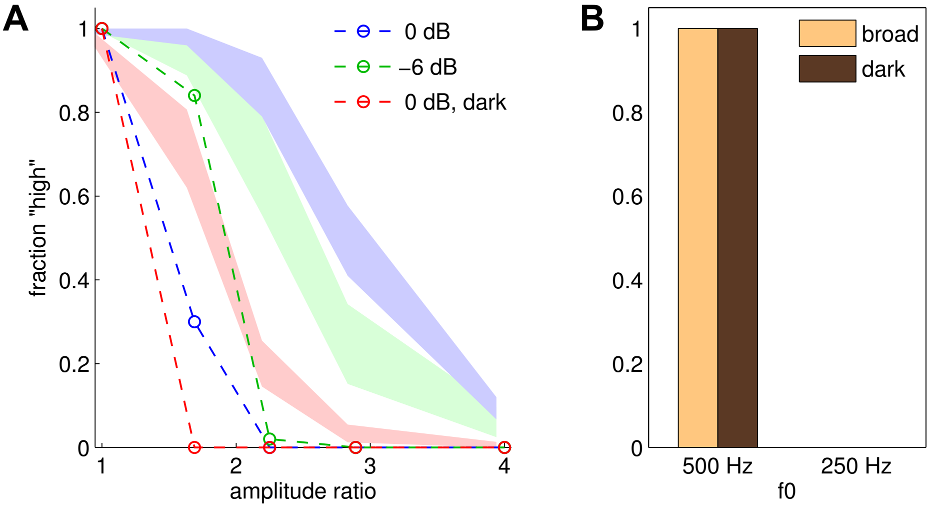


Figure 5.13: Pitch estimates from a simple summary autocorrelation model; peripheral front-end as in the Bayesian model, unweighted summation across all channels for all time lags (SACF). A: Click train stimuli; dotted lines represent model predictions, shaded areas indicate subjects' mean psychometric curves ±1 SEM (cf. Figure 5.8). B: Harmonic complex tones.

The outcome of this procedure is shown in Figure 5.13. Not unlike Terhardt's model, our first implementation of summary autocorrelation is unsuitably biased towards reporting the low pitch for all click train stimuli with $r > 1$ (5.13A), reflecting true stimulus periodicity rather than listeners' subjective percept of pitch. Darkening the stimulus spectrum by removing high-frequency components leads to a small, further biasing for click trains stimuli (blue and red curves) but not the complex harmonic tones (5.13B), qualitatively in agreement with human behaviour. Contrary to our human

[11] Note though, that this form our read-out is also problematic by itself, as the SACF is sensitive to changes in the stimulus other than its periodicity and the matching might thus reflect an entirely different feature of the stimulus in theory (cf. section 2.4.2).

data however, an increase in the masking noise results in an *increase* of the fraction of high pitch judgements (blue and green curves), rather than a decrease. Leaving the peripheral model and read-out mechanism unchanged, we verified that this model response pattern was robust towards towards several changes in the ACF and SACF computation stages. Instead of computing the exact ACF across the entire stimulus, we used a short-term ACF (no prior mean-removal) with either a single integration time constant of 10 ms for all channels (Meddis and O'Mard, 1997) or a different CF-dependent time constant for each channel (Wiegrebe 2001; cf. section 2.4.2). The fit between model and psychophysical data was also not notably improved by applying channel- and lag-dependent summation weights, gradually limiting the influence of high-frequency channels on long SACF lags (Bernstein and Oxenham, 2005). The behaviour of our simple autocorrelation model is in many ways comparable to that of the Bayesian model without pitch-timbre dependence: note in particular the effect of increasing the masking noise in both cases (cf. Figure 5.10). The greatest difference between the two is an overall greater tendency of the uncoupled Bayesian to report the high pitch for $r = 1.69$. This difference is consistent with our previous explanation (cf. section 5.4.1): at low values of $r > 1$, the Bayesian model prior favours monophasic impulse response shapes over biphasic ones, overruling periodicity violations that should favour the low-pitch interpretation. The summary autocorrelation model has no such biasing mechanism — not for high periodicities *per se*, but for certain waveform shapes over others *given* an assumed periodicity.

A serious and possibly unfair limitation to the performance of our first implementation may have been imposed by our choice of the peripheral model. Starting with Meddis and Hewitt (1991), proponents of the SACF model have typically applied the summary autocorrelation principle to the output of a far more biophysically-detailed peripheral processing stage than we have done so far. In order to make a fair comparison, we used a recent extension of the original summary autocorrelation model (Balaguer-Ballester et al., 2008)[12]: an initial, time-varying SACF function is computed based on a physiologically detailed peripheral model (Meddis, 2006), including a human-like middle-ear transfer function, calcium and neurotransmitter trafficking in the outer hair cells and firing-rate adaptation in the auditory nerve. Finally, a low-pass filtered SACF (LP-SACF) is generated by leaky integration of the initial SACF over time. Even with

[12] A Matlab code package encompassing all stages of the model is available from Ray Meddis' website, `http://www.essex.ac.uk/psychology/department/people/Meddis.html`

this considerably more complex (and tunable) model, we did not succeed in capturing the essential characteristics of pitch-timbre interactions in our human listeners. Our best attempt is shown in Figure 5.14. The pitch estimates are representative of low- to medium-spontaneous rate auditory nerve fibres, when the stimulus is presented at an intensity level around 45 dB SPL (±5 dB tolerance, approximately), using lag- and channel dependent summation weights (Bernstein and Oxenham, 2005). Our relative success depended crucially on this particular combination of settings, while the choice of integration time constants, both for within-channel ACFs and during computation of the LP-SACF, had little influence. Note though, that we are allowing for a 20-30 dB level difference between experiment and model: the stimulus presentation level in our psychophysical experiment was around 75 dB SPL, at which point the LP-SACF model fails completely.

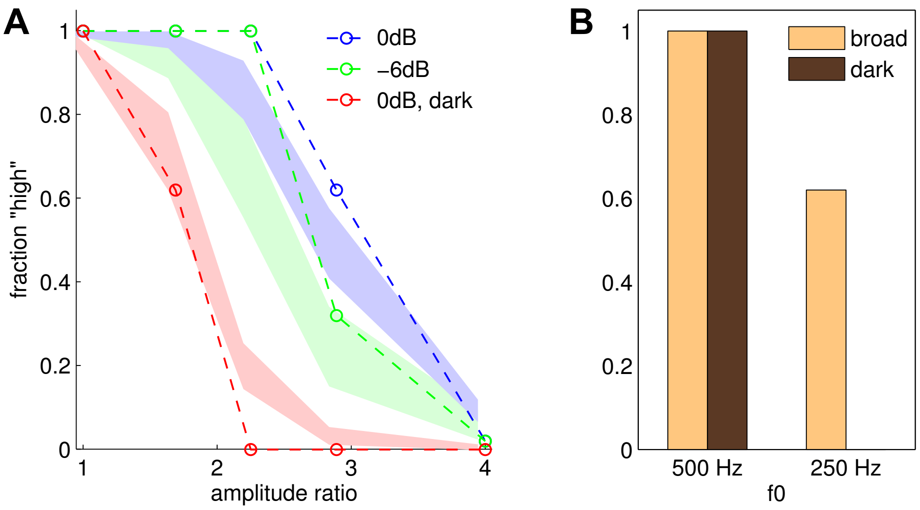


Figure 5.14: Pitch estimates from an extended summary autocorrelation model, using physiologically plausible peripheral model, channel- and lag-dependent summation weights and low-pass filtering of the summary autocorrelation function over time (LP-SACF). 45 dB SPLstimulus presentation level. A: Click train stimuli; dotted lines represent model predictions, shaded areas indicate subjects' mean psychometric curves ±1 SEM (cf. Figure 5.8). B: Harmonic complex tones.

Accepting the aforementioned restrictions and caveats, we see the LP-SACF model comes much closer to capturing human pitch judgements of the ACT stimuli — in particular their dependency on timbre — than our simplistic first implementation. The 0 dB broadband stimuli are consistently classified as high-pitched up until an amplitude ratio of $r = 2.25$, beyond which the percept drops to the low pitch (Figure 5.14A, blue curve). In comparison, low-pass filtered click trains (red curve) are judged as low-pitched more often in comparison, matching human behaviour well. The predicted psychometric curve for noisy broadband stimuli (green) falls in between, even though the effect of this manipulation is smaller than observed psychophysically and the only notable difference occurs at $r = 2.89$. However, this relative success in modelling the pitch of alternating click trains is accompanied by a significant, previously unseen failure of the model to capture the independence of pitch and timbre of the harmonic complex tones (Figure 5.14A). While human pitch judgements are virtually unaffected by our timbral manipulation in this case, the model judges 62% of the 250 Hz tones with broadband timbre to be high in pitch (500 Hz). Thus, it seems that the model (in this already unrealistically favourable regime) is now too unspecifically biased towards high pitches.

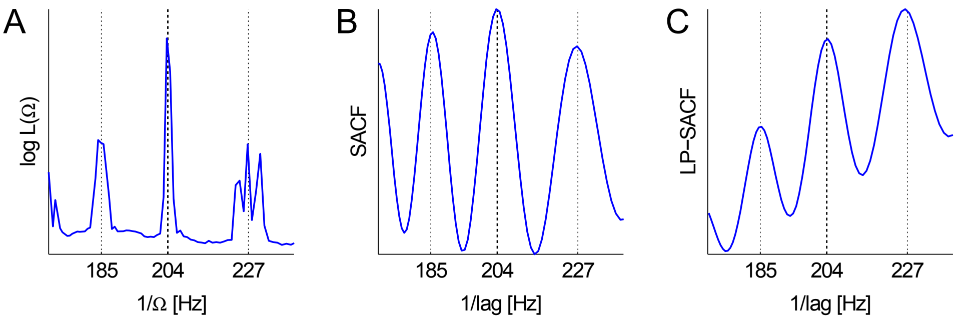


Figure 5.15: Comparing pitch estimates for a 200 Hz sinusoidally amplitude-modulated tone across three models. Dotted lines indicated typical human pitch matches. A: Bayesian model with prior pitch-timbre coupling. B: Simple summary autocorrelation; peripheral front-end as in the Bayesian model, unweighted summation across all channels for all time lags (SACF). C: Summary autocorrelation with physiologically plausible front-end, channel- and lag-dependent summation weights and low-pass filtering of the summary autocorrelation function over time (LP-SACF).

We confirmed this diagnosis with a further test of the model. As was previously discussed in sections 2.1.2 and 4.4, the pitch of an 200 Hz sinusoidally amplitude-modulated tone with carrier frequencies of 2040 Hz is typically reported by listeners as 204 Hz, while additional pitches around approximately 185 Hz and 227 Hz can be heard out when listeners are appropriately cued or encouraged to do so (Schouten et al., 1962). We compared the predictions of the Bayesian model *with* pitch-timbre dependence in the prior, and of our two implementations of the summary autocorrelation model (SACF and LP-SACF; LP-SACF in the same intensity range around 45 dB SPL) for this particular stimulus. While the dominant pitch of 204 Hz, as well as the two minor modes, are well-predicted by the Bayesian model and the simple SACF model, LP-SACF is again unduly biased towards the higher of these two alternative pitches and predicts instead a pitch of 227 Hz (this was robust to changes in the acoustic noise level). Hence, only the extended Bayesian model is able to capture the timbre-dependence of the pitch of alternating click trains without simultaneously incurring an unspecific high-pitch bias affecting its judgements of two types of control stimuli: amplitude-modulated and harmonic complex tones.

5.5 Harmonic complex tones revisited: the strength of missing-f_0 pitch

In chapter 4, we investigated the behaviour of our simple, uncoupled model across a range of pitch-evoking stimuli. While the model succeeded in predicting the pitch *height* of many periodic and aperiodic sounds, we found that the model lacked signs of the qualitative perceptual difference in pitch *strength* between HCTs with and without low-rank harmonics present in the spectrum. Pitch strength and f_0-discriminability of a HCT declines markedly, if its spectrum contains only high-rank harmonics above approximately the tenth (Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2003). In our uncoupled model, we found no sign of such a transition: the likelihood ratio between the true periodicity and a near-by value (e.g. a quarter tone above or below), which we treat as an indicator for the subjective certainty of the model about its own estimate on a single trial, was unaffected by the rank of the lowest harmonic.

Following our extension of the model by a prior that couples periodicity and spectral envelope in the generative process, we decided to revisit the issue of pitch strength of the missing fundamental. Based on the overall nature of this coupling, whereby low-pitched sounds are expected to have less high-frequency content than high-pitched sounds (cf. section 5.2.2), we might now expect to find a dependence of pitch strength on the rank of the lowest harmonic. When the model encounters evidence for high-frequency spectral content in the AN activity pattern while evaluating $\mathcal{L}(\Omega)$ for long periods Ω, it should now be inclined to attribute these high-frequency components to the acoustic background noise, rather than to the periodic signal embedded in it. By attributing high-frequency spectral components to noise, however, they are effectively discarded as evidence in favour of Ω or against it. Since the estimation of Ω (for low Ω and spectrally bright sounds) is based on less evidence, we would reasonably expect the certainty of the estimate to be reduced.

In order to test this intuition, we replicated a simple version of the experiments by Houtsma and Smurzynski (1990) in the model. We generated a series of harmonic complex tones with a fundamental of 250 Hz and 11 successive harmonics each, varying the harmonic number (or rank) n of the lowest component from 4 to 25 in steps of three. We computed the log-likelihood function $\ln \mathcal{L}(\Omega)$ for a narrow range of periodicities around 250 Hz and determined the height of its peak at 250 Hz compared to values of $\ln \mathcal{L}(\Omega)$ one quarter-tone above and below.

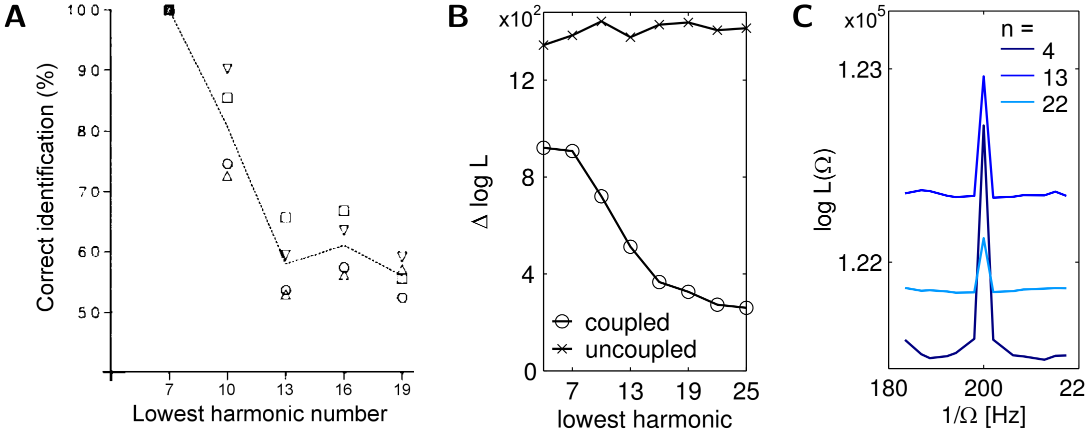


Figure 5.16: Effect of lowest harmonic number n on the pitch strength of 11-component HCTs. A: Human interval identification performance as a function of n; chance level was 14% (from Houtsma and Smurzynski, 1990). B: Local peak height of the log-likelihood function as a function of n in the model. C: Log-likelihood functions for three different stimuli with $n \in \{4, 13, 22\}$.

Using the same kind of stimuli, Houtsma and Smurzynski (1990) had found that both musical interval identification and f_0-discriminability declined rather sharply as n increased above 7 with no significant further deterioration at and above 16 (see Figure 5.16A). We found a very similar transition in the coupled model. As shown in 5.16B, the height of the local peak in $\ln \mathcal{L}(\Omega)$ around 250 Hz drops steeply at first as $n > 7$, but the decline tails off as $n \geq 16$ (though not entirely flat). The uncoupled model, in comparison, shows no deterioration of peak height across the entire range of values tested (as previously observed; cf. section 4.2).

A similar effect was demonstrated by Shackleton and Carlyon (1994). Complex tones with fundamental frequencies of either 88 or 250 Hz were filtered into either a "LOW", "MID" or "HIGH" frequency region (125 – 625 Hz, 1375 – 1875 Hz and 3900 – 5400 Hz respectively). In the LOW region, both HCTs contained resolved, low-rank partials, while both complexes contained only unresolved, high-rank partials in the HIGH region. In the MID region however, the low-pitched sound contained only peripherally-unresolved harmonics, while the high-pitched sound was still resolved. Measuring f_0 difference limens (F0DLs) for all six stimuli, it was found that F0DLs were low for the two LOW-filtered stimuli, and high for the two HIGH-filtered stimuli. In the MID region, there was a discrepancy between the low-pitched stimulus which was poorly discriminated, and the high-pitched stimulus which was well-discriminated by listeners (Figure 5.17A). Shackleton and Carlyon (1994) explained their results as a consequence of *resolvability* and argued that two different pitch mechanisms might be at work for resolved and unresolved complex sounds — an argument that Carlyon (1998) later used to refute the autocorrelation model of Meddis and O'Mard (1997) which showed no such dependence on resolvability.

We tested the behaviour of our coupled model on stimuli much like the ones used by Shackleton and Carlyon (1994). We used a low f_0 of 100 Hz (instead of 88 Hz), in order to ensure that both low and high stimulus periodicities matched our sampling rate of 20 kHz. The frequency range of the MID region was adjusted accordingly. We measured the local peak height of the log-likelihood function around the true fundamental frequencies as a proxy for pitch strength in the model. The results are shown in Figure 5.17B. Despite minor discrepancies, our model captures the main psychophysical effect: pitch strength *appears* to be determined by resolvability. In line with our previous results for the stimuli of Houtsma and Smurzynski (1990), no such effect was found in the uncoupled model (results not shown). This means, however, that the discrepancy in the MID frequency region cannot be due to resolvability *per se* in the model, as our change in the model prior does not affect the peripheral transduction stage at all. Instead, the effect can be explained by the changing spectral centre of mass in relation to the fundamental frequency, resulting in a tendency to regard the high-frequency content of low-pitched sounds as noise, and to thereby disregard it as evidence for the presence of the fundamental as discussed above.

The interpretation of the results by Shackleton and Carlyon (1994) as an effect of harmonic number, rather than resolvability, is strengthened by an experiment by Bernstein and Oxenham (2003). Here, a monaurally unresolved complex sound was played to subjects *dichotically*, with even and odd harmonics presented to opposite ears. The harmonics were perceptually fused across ears, and pitch height of the sound remained unaffected by this manipulation. However, as the spacing of the components in each ear doubled, the sound became peripherally resolved. If resolvability was the major deter-

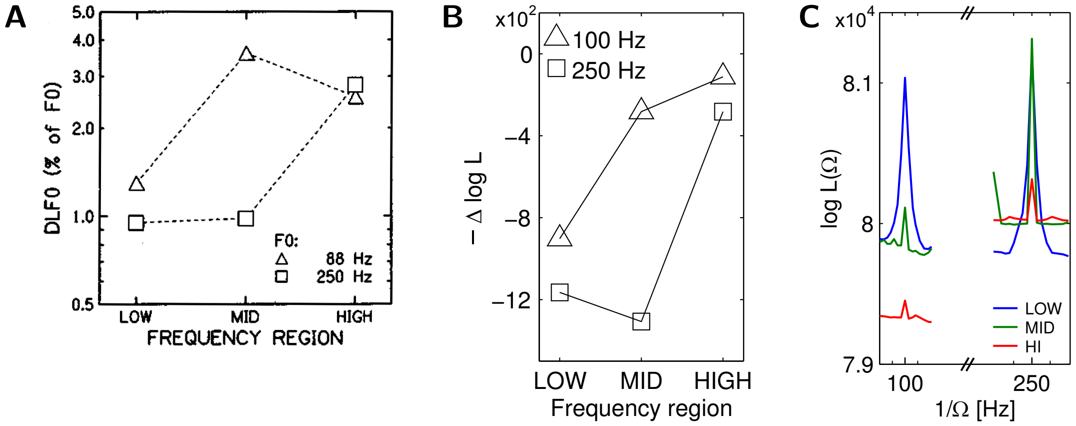


Figure 5.17: Pitch strength of two complex tones filtered into different frequency bands. A: F0DLs for two sounds with f_0s of 88 and 250 Hz, filtered into LOW, MID and HIGH frequency regions. B: *Negative* peak height of $\ln \mathcal{L}(\Omega)$ in the model (f_0s 100 and 250 Hz; high negative values indicate high (local) certainty about Ω). C: Log-likelihood functions for the six stimuli.

minant of pitch strength, then one might expect an improvement in f_0-discriminability for the dichotic stimulus compared to the diotic one. No such effect on discriminability was found, arguing for harmonic number, rather than resolvability as the key determining factor (see also Moore and Gockel, 2011 for a recent review and discussion).

Chapter 6

Conclusions

6.1 Summary

In this thesis, we set out to develop a Bayesian probabilistic model of human pitch perception, based on Helmholtz' notion of perception as unconscious inference and harnessing the power of modern statistical estimation techniques. In chapter 3, we described a generative model of naturalistic pitch-evoking sounds and evoked responses in the peripheral auditory system. We discussed how the model can be inverted to perform optimal perceptual inference about the periodicity of an arbitrary waveform, indirectly observed by the central auditory system through time-varying neural firing rates in the auditory nerve. Due to the inherent non-linearity of the sensorineural transduction process and the high dimensionality of the latent variable space, exact inference in the model is intractable. We adapted two established approximate inference techniques for use with our model.

Two variants of the model were presented, which differ in their prior assumptions regarding the relationship between the periodicity of a sound and the characteristics of its spectral envelope. In the uncoupled model, these two acoustic properties are treated as independent. Our evaluation of the uncoupled model in chapter 4 revealed that we can account for the pitch frequency of a large variety of periodic and non-periodic pitch-evoking sounds under this assumption. However, we also saw indications of sensitivity to spectral characteristics of sounds other than their harmonicity in human listeners, which the uncoupled model is unable to explain.

In chapter 5, we introduced a coupling between periodicity and spectral envelope into the assumed generative process. The coupling was chosen such that the low-pass characteristic of the spectral envelope, and concomitantly its spectral centre of mass, varies with periodicity: high-pitched sounds are expected to contain more high-frequency content than low-pitched sounds. This coupling was motivated in part by qualitative, psychophysical evidence for a perceptual dependency of pitch on timbre. The coupling parameters, however, were qualitatively fit to a database of natural pitch-evoking sounds. As a result of this dependency, the spectral envelope of natural sounds becomes to some degree informative about their periodicity. In an ideal observer, we therefore expect to observe a biasing effect of timbre on pitch that is most pronounced when the uncertainty regarding the sound periodicity is high.

We designed a psychophysical experiment to test these predictions, using stimuli that allow for precise control over the degree of octave-uncertainty in human listeners. Our results indicate a timbre-dependent bias of human pitch perception that is quantitatively well-described by our coupled Bayesian model. The uncoupled Bayesian model fails to account for these effects alongside a variety of other, non-probabilistic models lacking the power to express this kind of dependency. Thus, human psychophysical behaviour is non-trivially explained by our coupled model, indicating that human listeners exploit statistical regularities of natural sounds during pitch perception in order to improve the accuracy of their periodicity estimates.

Finally, we investigated the effect of coupling pitch and timbre in the model on periodicity estimation for harmonic complex tones with missing fundamental. Whereas pitch strength in the uncoupled model was insensitive to the rank of the lowest harmonic present in the stimulus, pitch strength in the coupled model decreases as this rank is increased, similar to the percept in human listeners. We therefore suggest that decreased pitch strength of HCTs with high-rank harmonics does not primarily constitute a potentially disadvantageous performance limit. Instead, it may reflect an optimal or near-optimal adaptation of the auditory system to natural listening conditions.

As envisioned by Helmholtz, human pitch perception bears specific hallmarks of an optimal inferential process, in which ambiguities inherent in the immediate, incoming sensory evidence are resolved by recourse to prior knowledge (learnt or innate) regarding the occurrence of their physical causes in the external world and the processes that map causes onto sensations. The act of inference itself is highly non-trivial in the case of our model. We have derived approximate inference schemes without consideration of the computational and mechanistic constraints of their potential physiological substrate. We justify this in two ways. Firstly, in order to establish the behaviour of a true ideal observer with minimal bias, we should prioritise accuracy of the approximation over implementational constraints regarding time and memory demand or its likely mechanistic building-blocks (within the limits of overall tractability). Secondly, there is very little hard evidence for the exact locus, size or structure of the neuronal network involved in the determination and representation of our percept. Whatever evidence is available seems insufficient to delineate even a rough boundary between viable and inviable estimation algorithms.

Nevertheless, implementational considerations are important. With our current arsenal of inference techniques, we are severely limited technically in terms of the maximal stimulus duration, number of trials and density with which we can feasibly evaluate the likelihood function. We have seen that in many instances, heuristic models such as summary autocorrelation and its close relatives make predictions similar to our Bayesian model at a fraction of the computational cost. Rather than further improving the validity of these heuristics purely by phenomenological adjustments, the *functional* insights gained from our ideal observer approach may serve as valuable guidance in the development of more accurate heuristics. Chiefly amongst these are the consideration of natural sounds and listening conditions, which are expected to shape the behaviour of an ideal observer, as well as the efficient combination of information from multiple sources where they may be available, weighted according to their reliability.

6.2 Outlook

In this thesis we have taken the first steps towards a systematic Bayesian characterisation of human pitch perception. Needless to say, many possible routes have been left unexplored along the way. Some easily-attainable extensions that would not require structural changes to the model have already been discussed in chapter 3. We could include a more realistic prior distributions over periodicities, that reflects, for example, the prevalence of speech amongst the pitched sounds likely to be encountered in our environment, as well as the rarity of extremely high- or low-pitched sounds. In our current model, the fundamental sensitivity of AN fibres to acoustic stimulation is uniform across all CFs. Human sensitivity, in contrast, varies non-uniformly with frequency, owing to diverse factors such as the outer- and middle-ear transfer functions or a varying density of hair cells along the length of the basilar membrane. We have preformed preliminary studies with filter gains adjusted according to absolute threshold hearing levels, and it stands to reason that a modification such as this may be required to capture subtle phenomena regarding the dominance of some spectral regions over others in determining the pitch of a sound. At the level of the auditory nerve, we could include static non-linear compression in the demodulation stage, requiring only minor changes to the inference algorithms to account for the different derivatives of the rectification function.

By assumption, we have restricted ourselves to short, monaural stimuli of constant periodicity, presented in an otherwise unstructured background of noise. This immediately suggests possible next steps. Amongst these, an extension of the model towards binaural processing of pitch is perhaps the least interesting by itself. On the one hand, binaural pitch phenomena are largely limited to artificial listening conditions and will hardly ever be encountered in natural situations. On the other hand, those phenomena encountered under natural conditions persist even for monaural stimulation.

Extending the generative model to mixtures of periodic sounds seems conceptually straightforward. We can simply imagine the waveform to be a sum of component sounds, each of which is distributed according to our acoustic model. However, the dimensionality of the latent variable space that we need to integrate over in order to evaluate the likelihood $\mathcal{L}(\Omega_1, \ldots, \Omega_K)$ grows linearly with the number of components, in addition to the exponential growth in the number of combinations of periodicities $(\Omega_1, \ldots, \Omega_K)$ itself. Needless to say, the computational demand of joint estimation along the lines of our *current* inference algorithms seems absolutely prohibitive. Interestingly, binaural processing may play a more important role in the perception of mixtures of pitched sounds: interaural time- and level differences provide salient cues for the grouping of frequency components in the stimulus according to their spatial origin, which would be expected to interact with harmonicity-based grouping cues.

More immediately attainable, perhaps, is an extension towards time-varying pitch. We could, for example, consider a Hidden Markov model in which short observation segments, each some tens of milliseconds long and distributed according to our generative model, are linked by a prior over consecutive periodicities but independent otherwise.

Even though the number of likelihood evaluations grows linearly with the number of segments, the dynamic programming algorithm involved in computing the trajectory of periodicities through time is easily parallelised, and the use of shorter individual time windows could in fact benefit performance. With a model of this kind (crude as the Markov assumption may be) we could start to investigate sequential pitch effects such as the tritone paradox (Deutsch, 1986) in the perception of Shepard tones and its dependence on stimulus context and priming (discussed in section 2.1).

Pitch perception is no end in itself. If, as we have argued, pitch is used to identify sound sources or even to infer the semantic content of a speech utterance, then pitch perception should ideally be closely intermeshed with these "higher-level" auditory perceptual tasks. This view is supported by the limited available evidence for the involvement of higher auditory cortical areas in the processing and representation of pitch. Assuming for example, that identifying the gender (or alternatively, size or age) of a speaker from their voice is behaviourally useful, we could express the joint distribution over pitch and timbre as mixture distribution over male and female speakers, where each gender has its own characteristic range of pitches and timbres. Knowing the pitch of a speaker's voice (in addition, for example, to the content of the speech utterance) is obviously useful in establishing his or her gender. Knowing the speaker's gender conversely restricts the pitch range we can expect to hear. This may in turn help us in segregating the speaker's voice from a noisy background, which is another common perceptual task the auditory system faces. It is one of the great appeals of the Bayesian probabilistic framework that it allows for the consistent and optimal integration of evidence, knowledge, expectations and inferences across multiple levels of a hierarchically structured task model as we have just sketched. Thus, by expressing pitch perception as Bayesian inference, we pave the way for its seamless integration with probabilistic models of auditory scene analysis in general.

Appendix A

Gradient and Hessian of $\ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$

We want to find expressions for the gradient and Hessian with respect to $\boldsymbol{x}$ of

$$\begin{aligned} \ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega) &= \ln \mathrm{P}(A \mid \boldsymbol{x}) + \ln \mathrm{P}(\boldsymbol{x} \mid \Omega) \\ &= \ln \mathrm{P}(A \mid \boldsymbol{x}) + \ln \sum_s \mathrm{P}(\boldsymbol{x}, s \mid \Omega) \\ &= \sum_i \ln \mathrm{P}(\boldsymbol{a}_i \mid \boldsymbol{x}) + \ln \sum_s \mathrm{P}(\boldsymbol{x}, s \mid \Omega) \end{aligned} \tag{A.1}$$

where $s = 1 \dots S$ is the component indicators in our (Gaussian) mixture model of $\boldsymbol{x}$.

A.1 Gradient $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \mid \Omega)$

We will treat the two gradients, $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(A \mid \boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} \ln \sum_s \mathrm{P}(\boldsymbol{x}, s \mid \Omega)$, separately in the following two section.

A.1.1 $\nabla_{\boldsymbol{x}} \ln \sum_s \mathrm{P}(\boldsymbol{x}, s \mid \Omega)$

$\mathrm{P}(\boldsymbol{x} \mid \Omega)$ is a mixture distribution: $\mathrm{P}(\boldsymbol{x} \mid \Omega) = \sum_s \mathrm{P}(\boldsymbol{x}, s \mid \Omega)$. As we will proof below, we can relate the gradient of $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(\boldsymbol{x} \mid \Omega)$ (note the logarithm) to the gradients of the log-component distributions $\ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega)$ by a simple weighted average as follows:

$$\nabla_{\boldsymbol{x}} \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) = \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \nabla_{\boldsymbol{x}} \ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega) \tag{A.2}$$

In words: $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega)$ is the sum over the individual log-component gradients, weighted by their linear, *posterior* responsibilities $\mathrm{P}(s \,|\, \boldsymbol{x}, \Omega)$, given via Bayes rules by

$$\mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) = \frac{\mathrm{P}(\boldsymbol{x} \,|\, s, \Omega)\,\mathrm{P}(s \,|\, \Omega)}{\sum_{s'} \mathrm{P}(\boldsymbol{x} \,|\, s', \Omega)\,\mathrm{P}(s' \,|\, \Omega)} \tag{A.3}$$

Note that, typically for our purposes, the prior mixture weights are uniform and independent of Ω: $P(s|\Omega) = \frac{1}{S}$, simplifying above expression further.

With this convenient result at hand, we can compute $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$ easily, as the component distributions are all Gaussian in $\boldsymbol{x}$ and their individual gradients are:

$$\nabla_{\boldsymbol{x}} \ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega) = -\boldsymbol{x}(\Sigma^s_\Omega)^{-1} \quad , \tag{A.4}$$

where Σ^s_Ω is the generative covariance matrix as defined in equation (3.17) (in case of the less general, uncoupled model, we can simply drop the Ω-index; cf equation (3.10)).

Proof of equation (A.2):

Since $\sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) = 1$, we can trivially rewrite $\ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega)$

$$\ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) = \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) \tag{A.5}$$

Exploiting the identities

$$\mathrm{P}(\boldsymbol{x} \,|\, \Omega) = \frac{\mathrm{P}(\boldsymbol{x}, s \,|\, \Omega)}{\mathrm{P}(s \,|\, \boldsymbol{x}, \Omega)} \tag{A.6}$$

and

$$\sum_s \nabla_x \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) = \nabla_x \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) = \nabla_x 1 = 0 \tag{A.7}$$

we obtain the gradient $\nabla_x \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega)$ as follows:

$$\begin{aligned}
\nabla_x \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) &= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \nabla_x \ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega) \\
&\quad - \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \nabla_x \ln \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \qquad \text{(A.8a)} \\
&\quad + \sum_s \nabla_x \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) \\
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \nabla_x \ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega) \\
&\quad - \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \frac{\nabla_x \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega)}{\mathrm{P}(s \,|\, \boldsymbol{x}, \Omega)} \qquad \text{(A.8b)} \\
&\quad + \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) \cdot \nabla_x \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \\
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \nabla_x \ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega) \qquad \text{(q.e.d)} \qquad \text{(A.8c)}
\end{aligned}$$

A.1.2 $\nabla_{\boldsymbol{x}} \ln \mathrm{P}(A \,|\, \boldsymbol{x})$

We recall from section (3.23), that

$$\ln \mathrm{P}(A \,|\, \boldsymbol{x}) = -\frac{1}{2\sigma_A^2} \sum_i (\boldsymbol{a}_i - \, \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})^2 + \mathcal{Z} \quad , \qquad \text{(A.9)}$$

where $(\cdot)^2$ denotes the inner product of a vector with itself. For a single channel i, we obtain the gradient $\nabla \ln \mathrm{P}(\boldsymbol{a}_i \,|\, \boldsymbol{x})$:

$$\nabla \ln \mathrm{P}(\boldsymbol{a}_i \,|\, \boldsymbol{x}) = \frac{1}{\sigma_A^2} \nabla \boldsymbol{a}_i (\, \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})^\intercal - \frac{1}{2\sigma_A^2} \nabla (\, \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{b})^2 \qquad \text{(A.10)}$$

We will consider the two terms on the RHS of equation (A.27) separately, starting with $\nabla \boldsymbol{a}_i (\, \mathbf{r}(\boldsymbol{b}_i * \boldsymbol{x}) * \boldsymbol{l})^\intercal$. Writing out both convolutions explicitly, we get

$$\boldsymbol{a}_i (\, \mathbf{r}(\boldsymbol{b}_i * \boldsymbol{x}) * \boldsymbol{l})^\intercal = \sum_t a_i(t) \sum_{t'} l(t - t' + 1) \, \mathrm{r} \left(\sum_{t''} b_i(t' - t'' + 1) x(t'') \right) \qquad \text{(A.11)}$$

with indices

$$t' = \max(t - N + 1, 1), \ldots, t - 1 \text{ and } t'' = \max(t' - M + 1, 1), \ldots, t' - 1 \quad , \qquad \text{(A.12)}$$

where N and M are the lengths of $\boldsymbol{l}$ and $\boldsymbol{b}$, respectively. Hence,

$$\frac{\partial}{\partial x_k} \boldsymbol{a}_i^{\intercal}(\,\mathbf{r}(\boldsymbol{b}_i * \boldsymbol{x}) * \boldsymbol{l})$$

$$= \sum_t a_i(t) \sum_{t'} l(t-t')\ \frac{\partial}{\partial x_i}\,\mathrm{r}\left(\sum_{t''} b_i(t'-t''+1)x(t'')\right) \tag{A.13}$$

$$= \sum_t a_i(t) \sum_{t'} l(t-t')\,\mathrm{r}'((\boldsymbol{x} * \boldsymbol{b}_i)(t')) \begin{cases} b_i(t'-k+1) & \text{if } k \le t' \le i+M+1 \\ 0 & \text{otherwise} \end{cases} \tag{A.14}$$

$$= \sum_{t'} \left(\sum_t a_i(t) l(t-t') \right) \mathrm{r}'((\boldsymbol{x} * \boldsymbol{b}_i)(t')) b_i(t'-k+1) \tag{A.15}$$

with $t' = k \ldots \min(k+M-1, T)$ and $t = t' \ldots \min(t'+N-1, T)$

$$= \sum_{t'} B_i(t')\,\mathrm{r}'((\boldsymbol{x} * \boldsymbol{b}_i)(t')) b_i(t'-k+1) \tag{A.16}$$

$$= (\boldsymbol{B}_i \circ \mathbf{r}'(\boldsymbol{x} * \boldsymbol{b}_i)) * \overleftarrow{\boldsymbol{b}_i}(k) \quad , \tag{A.17}$$

where

$$B_i(t) = \sum_t a_i(t) l(t-t') = (\boldsymbol{a}_i * \overleftarrow{\boldsymbol{l}}\,)(t') \tag{A.18}$$

and $\overleftarrow{\boldsymbol{b}_i}$ and $\overleftarrow{\boldsymbol{l}}$ denote the time-reversed filter kernels $\boldsymbol{b}_i$ and $\boldsymbol{l}$.

Similarly, for $\nabla(\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{b})^2$ we get

$$\frac{\partial}{\partial x_k}(\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{b})(\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})^{\intercal}$$

$$= \ \frac{\partial}{\partial x_k} \sum_t \left(\sum_{t'} l(t-t')\,\mathrm{r}\left((\boldsymbol{x} * \boldsymbol{b}_i)(t')\right) \right)^2 \tag{A.19}$$

$$= 2 \sum_t (\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})(t)\ \frac{\partial}{\partial x_i} \left(\sum_{t'} l(t-t')\,\mathrm{r}\left((\boldsymbol{x} * b_i)(t')\right) \right) \tag{A.20}$$

$$= 2 \sum_t (\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})(t)$$

$$\cdot \sum_{t'} l(t-t')\,\mathrm{r}'((\boldsymbol{x} * \boldsymbol{b})(t')) \begin{cases} b_i(t'-k) & \text{if } k \le t' \le k+M-1 \\ 0 & \text{otherwise} \end{cases} \tag{A.21}$$

$$= 2\,((\boldsymbol{C}_i \circ \mathbf{r}'(\boldsymbol{x} * \boldsymbol{b}_i)) * \overleftarrow{\boldsymbol{b}_i})(k) \tag{A.22}$$

where

$$C_i(t) = ((\,\mathbf{r}'(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l}) * \overleftarrow{\boldsymbol{l}}\,)(t') \quad . \tag{A.23}$$

Thus,

$$\nabla \ln \mathrm{P}(A \,|\, x) = \sum_i ((\boldsymbol{B}_i - \boldsymbol{C}_i) \circ \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i)) * \overleftarrow{\boldsymbol{b}_i} \tag{A.24}$$

A.2 Hessian $\nabla^2_{\boldsymbol{x}} \ln \mathrm{P}(A, \boldsymbol{x} \,|\, \Omega)$

As with the gradient, we will consider $\nabla^2_{\boldsymbol{x}} \ln \mathrm{P}(A \,|\, \boldsymbol{x})$ and $\nabla^2_{\boldsymbol{x}} \ln \sum_s \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega)$, the two constituent parts of the Hessian, separately.

A.2.1 $\nabla^2_{\boldsymbol{x}} \ln \sum_s \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega)$

In section A.1.1, showed that the *gradient* of $\ln \sum_s \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega)$ is simply related to the gradients of its log-component distributions $\ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega)$. As will will proof below, a similar relationships also holds for its Hessian matrix:

$$\begin{aligned}\nabla^2_x \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega) = &\sum_s \mathrm{P}(s \,|\, \boldsymbol{x}, \Omega) \cdot \left(\nabla^2_x \ln \mathrm{P}(\boldsymbol{x}, s \,|\, \Omega) + [\nabla_x \ln \mathrm{P}(x, s \,|\, \Omega)]^2\right) \\ &- [\nabla_x \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega)]^2 \quad ,\end{aligned} \tag{A.25}$$

where $[\boldsymbol{v}]^2$ denotes the *outer* product matrix of a vector $\boldsymbol{v}$ with itself, and the mixture responsibilities $\mathrm{P}(s \,|\, \boldsymbol{x}, \Omega)$ can be computed as in section A.1.1.

For the Hessian, let us consider the partial derivatives $\frac{\partial^2}{\partial x_i \partial x_j} \ln \mathrm{P}(\boldsymbol{x} \,|\, \Omega)$ (for ease of notation, we will drop any dependency on Ω in the following derivation):

$$\frac{\partial^2 \ln \mathrm{P}(\boldsymbol{x})}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_j} \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial}{\partial x_i} \ln \mathrm{P}(\boldsymbol{x}, s) \tag{A.26a}$$

$$\begin{aligned} &= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial^2}{\partial x_i \partial x_j} \ln \mathrm{P}(\boldsymbol{x}, s) \\ &\quad + \sum_s \frac{\partial}{\partial x_j} \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial}{\partial x_i} \ln \mathrm{P}(\boldsymbol{x}, s)\end{aligned} \tag{A.26b}$$

$$\begin{aligned}
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial^2}{\partial x_i \partial x_j} \ln \mathrm{P}(\boldsymbol{x}, s) \\
&\quad + \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial}{\partial x_j} \ln \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial}{\partial x_i} \ln \mathrm{P}(\boldsymbol{x}, s)
\end{aligned} \tag{A.26c}$$

$$\begin{aligned}
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial^2}{\partial x_i \partial x_j} \ln \mathrm{P}(\boldsymbol{x}, s) \\
&\quad + \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial}{\partial x_j} \ln \mathrm{P}(\boldsymbol{x}, s) \cdot \frac{\partial}{\partial x_i} \ln \mathrm{P}(\boldsymbol{x}, s) \\
&\quad - \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \frac{\partial}{\partial x_j} \ln \mathrm{P}(\boldsymbol{x}) \cdot \frac{\partial}{\partial x_i} \ln \mathrm{P}(\boldsymbol{x}, s)
\end{aligned} \tag{A.26d}$$

$$\begin{aligned}
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \left[\frac{\partial^2 \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i \partial x_j} + \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_j} \cdot \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i} \right] \\
&\quad - \frac{\partial \ln \mathrm{P}(\boldsymbol{x})}{\partial x_j} \cdot \frac{1}{\mathrm{P}(\boldsymbol{x})} \sum_s \mathrm{P}(\boldsymbol{x}, s) \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i}
\end{aligned} \tag{A.26e}$$

$$\begin{aligned}
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \left[\frac{\partial^2 \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i \partial x_j} + \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_j} \cdot \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i} \right] \\
&\quad - \frac{\partial \ln \mathrm{P}(\boldsymbol{x})}{\partial x_j} \cdot \frac{1}{\mathrm{P}(\boldsymbol{x})} \frac{\partial}{\partial x_i} \sum_s \mathrm{P}(\boldsymbol{x}, s)
\end{aligned} \tag{A.26f}$$

$$\begin{aligned}
&= \sum_s \mathrm{P}(s \,|\, \boldsymbol{x}) \cdot \left[\frac{\partial^2 \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i \partial x_j} + \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_j} \cdot \frac{\partial \ln \mathrm{P}(\boldsymbol{x}, s)}{\partial x_i} \right] \\
&\quad - \frac{\partial \ln \mathrm{P}(\boldsymbol{x})}{\partial x_j} \cdot \frac{\partial \ln \mathrm{P}(\boldsymbol{x})}{\partial x_i}
\end{aligned} \tag{A.26g}$$

$$\text{q.e.d.} \tag{A.26h}$$

A.2.2 $\nabla^2_{\boldsymbol{x}} \ln \mathrm{P}(A \,|\, \boldsymbol{x})$

$$\nabla^2 \ln \mathrm{P}(\boldsymbol{a}_i \,|\, \boldsymbol{x}) = \frac{1}{\sigma_A^2} \nabla^2 \boldsymbol{a}_i (\, \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})^{\top} - \frac{1}{2\sigma_A^2} \nabla^2 (\, \mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{b})^2 \tag{A.27}$$

$$= \frac{1}{\sigma_A^2} \left[\mathrm{diag}(\boldsymbol{B}_i \circ \boldsymbol{r}_i'') - \mathrm{diag}(\boldsymbol{C}_i \circ \boldsymbol{r}_i'') - G \circ [\boldsymbol{r}_i]^2 \right] \underset{\text{2d}}{*} [\boldsymbol{b}_i]^2_{180^\circ} \tag{A.28}$$

Here, $\mathrm{diag}(\boldsymbol{v})$ denotes a diagonal matrix with diagonal elements $\boldsymbol{v}$, $\boldsymbol{r}_i$ is shorthand notation for $\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i)$ (similarly for the derivatives of $\boldsymbol{r}$), $[\cdot]^2$ denotes the outer product of a vector with itself, and $[\cdot]_{180^\circ}$ stands for the 180 degree rotation of a matrix. Note that, as $\boldsymbol{b}_i$ is a filter kernel, the convolution of a matrix with $[\boldsymbol{b}_i]^2_{180^\circ}$ results in the sequential row- and column-wise filtering of that matrix with the reverse kernel, $\overleftarrow{\boldsymbol{b}_i}$.

From section A.1.2, we already know the gradients $\nabla \boldsymbol{a}_i(\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})^\intercal$ and $\nabla(\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{b})^2$. Using those, we obtain (from similarly elementary algebra) the second partial derivatives for the first term (cf. equations (A.17) and (A.18)):

$$\frac{\partial}{\partial x_k \partial x_l} \boldsymbol{a}_i(\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{l})^\intercal \tag{A.29}$$

$$= \sum_{t'=1}^{T} B_i(t) r_i''(t') \begin{cases} b_i(t'-k+1) b_i(t'-l+1) & \text{if } 1 \leq t'-k+1 \leq M \\ & \wedge\ 1 \leq t'-l+1 \leq M \\ 0 & \text{otherwise} \end{cases} \tag{A.30}$$

And for the second term (cf. equations (A.22) and (A.23)):

$$\frac{\partial}{\partial x_k \partial x_l} (\,\mathbf{r}(\boldsymbol{x} * \boldsymbol{b}_i) * \boldsymbol{b})^2 \tag{A.31}$$

$$= 2 \sum_{t'=1}^{T} C_i(t') r_i''(t') \begin{cases} b_i(t'-k+1) b_i(t'-l+1) & \text{if } \ldots \\ 0 & \text{otherwise} \end{cases} \tag{A.32}$$

$$+ 2 \sum_{t'=1}^{T} \sum_{s} G(t', s) r_i'(t') r_i'(s) \begin{cases} b_i(t'-k+1) b_i(s-l+1) & \text{if } \ldots \\ 0 & \text{otherwise} \end{cases} \tag{A.33}$$

$$\tag{A.34}$$

with

$$G(t', s) = \sum_{t=\max(t'+n+1,1)}^{\min(t'+N-1,T)} l(t-t'+1)\, l(t-s+1) \tag{A.35}$$

Appendix B

Method for estimating f_0 and f_c of musical instrument and vocal sounds

For each recording in the collection of instrumental and vocal sounds used in chapter 5 to fit the dependency between pitch and timbre in the coupled model (cf. section 5.2.2), the fundamental frequency (f_0) and spectral envelope width (f_c) of all notes were estimated as follows:

As each recording is a scale segment containing several different musical notes, individual notes are first identified based on the minima and maxima of the amplitude envelope. Each note is then further analysed in overlapping windows of 1 s duration (75% overlap). The Discrete Fourier Transform (DFT) of each windowed sound segment is computed and peaks in the amplitude spectrum are identified, with a minimum required difference of 40 Hz between neighbouring peaks and a minimum amplitude of 1% of the maximum in that particular segment.

In order to determine the fundamental frequency of each segment, the waveform is demodulated at 4 kHz and its autocorrelation function (ACF) $R(\tau)$ is computed. We then determine the shortest lag $\tau^* > 0.2\,\text{ms}$ at which $R(\tau)$ reaches a peak within 5% of its maximum. In chosing τ^* this way, rather than simply picking $\operatorname{argmax} R(\tau)$, errors can be avoided that would otherwise occur due the effect of noise on harmonically related peaks in $R(\tau)$ of near-equal height. For the same reason, the initial estimate

τ^* of the sound periodicity is iteratively lowered further, one octave at a time, if the amplitude spectrum contains a significant peak one octave below $1/\tau^*$. f_0 is estimated as the inverse of the final periodicity estimate τ^*.

Following the estimation of f_0, the width of the spectral envelope is determined. All spectral peaks $\{f_1, \ldots f_K\}$ that fall within 10% of a harmonic of the estimated fundamental are selected. The amplitude A and width f_σ of a scaled, half-normal function

$$e(f) = A \exp(-\frac{f^2}{f_\sigma^2}) \tag{B.1}$$

is fitted to the amplitudes $\{a_1, \ldots a_K\}$ of the selected spectral peaks by minimising the summed squared error,

$$E = \sum_{k=1}^{K} (e(f_k) - a_k)^2 \quad . \tag{B.2}$$

Thus, f_σ is the width of a half-normal approximation to the spectral envelope of the sound segment. The mean of a half-normal distribution of width f_σ equals $f_c = \sqrt{2/\pi} f_\sigma$. This value f_c is returned as a measure of the spectral envelope width. We will, somewhat imprecisely, call f_c the "spectral centroid" of a sound on occasion, when in fact it is the centroid of an approximation to its spectral envelope. For sounds with an approximately half-normal spectral envelope, f_c will closely approximate the *harmonic spectral centroid* as defined by the MPEG-7 ISO standard, which is simply the amplitude-weighted mean of the harmonic spectral peaks on linear scales of both frequency and amplitude (Kim et al., 2005). Figure B.1A shows the outcome of the f_0 and f_c determination algorithm for one example note from the music instrument database. Figure B.1B shows the estimates of f_0 and f_c across the entire collection of sounds, and demonstrates a good correspondence between the instrumental and vocal samples concerning the dependency of f_c on f_0. There is, however, also considerable inter-instrumental variability, which is not taken into account by the coupled generative model as it only attempts to fit the overall dependency across all instruments (cf. section 5.2.2).

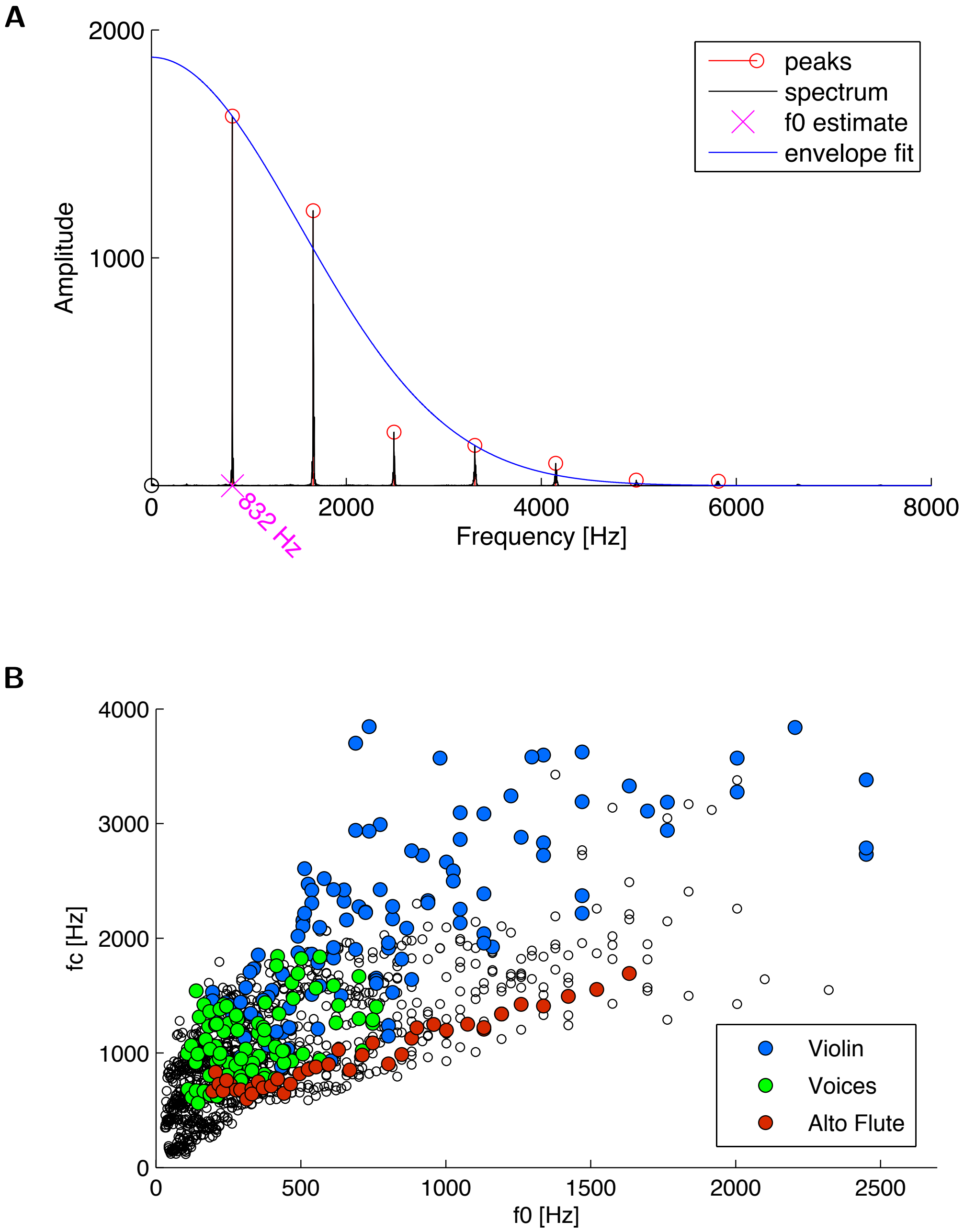


Figure B.1: A: Estimated f_0 (pink) and best-fit half-normal approximation (blue) to the spectral envelope of a sample note from the music instrument data base (Viola A$\flat$5). The linear amplitude spectrum is shown in black, significant peaks as determined by the algorithm in red. B: Estimates f_0 and f_c for all sounds from the collection of samples (instrumental and vocal).

Bibliography

A. M. H. J. Aertsen and P. I. M. Johannesma. Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, 38:223–234, 1980.

M. B. Ahrens and M. Sahani. Observers exploit stochastic models of sensory change to help judge the passage of time. *Current Biology*, 21(3):200–206, 2011.

M. A. Akeroyd, B. C. J. Moore, and G. A. Moore. Melody recognition using three types of dichotic-pitch stimulus. *Journal of the Acoustical Society of America*, 110 (3):1498–1504, 2001.

L. A. Anderson, M. N. Wallace, and A. R. Palmer. Identification of subdivisions in the medial geniculate body of the guinea pig. *Hearing Research*, 228(1-2):156–167, 2007.

ANSI. *American National Standard Acoustical Terminology.* ANSI S1.1 1994. American National Standards Association, New York, 1994.

ASA. *Acoustical Terminology SI, 1-1960.* American Standards Association, New York, 1960.

R. Ashley. Musical pitch space across modalities: spatial and other mappings through language and culture. In R. G. P. W. S. Lipscomb, R. Ashley, editor, *Proceedings of the International Conference on Music Perception and Cognition*, pages 64–71. Causal Productions, 2004.

F. Attneave and R. K. Olson. Pitch as a medium: A new approach to psychophysical scaling. *The American Journal of Psychology*, 84(2):147–166, 1971.

A. Bachem. Various types of absolute pitch. *Journal of the Acoustical Society of America*, 9(2):146–151, 1937.

A. Bachem. Chroma fixation at the ends of the musical frequency scale. *Journal of the Acoustical Society of America*, 20(5):704–705, 1948.

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

E. Balaguer-Ballester, S. L. Denham, and R. Meddis. A cascade autocorrelation model of pitch perception. *Journal of the Acoustical Society of America*, 124(4):2186–2195, 2008.

E. Balaguer-Ballester, N. R. Clark, M. Coath, K. Krumbholz, and S. L. Denham. Understanding pitch perception as a hierarchical process with top-down modulation. *PLoS Computational Biology*, 5(3):e1000301, 2009.

E. L. Bartlett and X. Wang. Neural representations of temporally modulated signals in the auditory thalamus of awake primates. *Journal of Neurophysiology*, 97(2): 1005–1017, 2007.

J. Beck and W. A. Shaw. The scaling of pitch by the method of magnitude-estimation. *The American Journal of Psychology*, 74(2):242–251, 1961.

D. Bendor and X. Wang. The neuronal representation of pitch in primate auditory cortex. *Nature*, 436(7054):1161–1165, 2005.

P. Berkes, R. Turner, and M. Sahani. On sparsity and overcompleteness in image models. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. MIT Press, 2008.

J. G. Bernstein and A. J. Oxenham. Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? *Journal of the Acoustical Society of America*, 113(6):3323–3334, 2003.

J. G. W. Bernstein and A. J. Oxenham. An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination. *Journal of the Acoustical Society of America*, 117(6):3816–3831, 2005.

U. W. Biebel and G. Langner. Evidence for interactions across frequency channels in the inferior colliculus of awake chinchilla. *Hearing Research*, 169(1-2):151–168, 2002.

F. A. Bilsen. Repetition pitch - monaural interaction of a sound with erpetition of same but phase shifted sound. *Acoustica*, 17(5):295–300, 1966.

F. A. Bilsen. Pitch of noise signals: Evidence for a ”central spectrum”. *Journal of the Acoustical Society of America*, 61(1):150–161, 1977.

J. K. Bizley, K. M. M. Walker, B. W. Silverman, A. J. King, and J. W. H. Schnupp. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of Neuroscience*, 29(7):2064–2075, 2009.

J. K. Bizley, K. M. M. Walker, A. J. King, and J. W. H. Schnupp. Neural ensemble codes for stimulus periodicity in auditory cortex. *Journal of Neuroscience*, 30(14): 5078–5091, 2010.

C. C. Blackburn and M. B. Sachs. The representations of the steady-state vowel sound /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology*, 63(5):1191–1212, 1990.

D. Bolinger. Intonation across languages. In *Universals of human language, Vol.2: Phonology*, pages 471–524. Stanford University Press, Stanford, 1978.

D. H. Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10:433–436, 1997.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* The MIT Press, Cambridge, MA, 1990.

J. P. Brokx and S. G. Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10(1):23–36, 1982.

W. E. Brownell, C. R. Bader, D. Bertrand, and Y. de Ribaupierre. Evoked mechanical responses of isolated cochlear outer hair cells. *Science*, 227(4683):194–196, 1985.

E. M. Burns. Pure-tone pitch anomalies. I. Pitch-intensity effects and diplacusis in normal ears. *Journal of the Acoustical Society of America*, 72(5):1394–1402, 1982.

E. M. Burns. Intervals, scales and tuning. In D. Deutsch, editor, *The Psychology of Music*, pages 215–264. Academic Press, New York, 2nd edition, 1998.

E. M. Burns and N. F. Viemeister. Nonspectral pitch. *Journal of the Acoustical Society of America*, 60(4):863–869, 1976.

E. M. Burns and N. F. Viemeister. Played-again sam: Further observations on the pitch of amplitude-modulated noise. *Journal of the Acoustical Society of America*, 70(6):1655–1660, 1981.

P. A. Cariani and B. Delgutte. Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology*, 76(3):1698–1716, 1996a.

P. A. Cariani and B. Delgutte. Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. *Journal of Neurophysiology*, 76(3):1717–1734, 1996b.

R. P. Carlyon. Comments on “A unitary model of pitch perception”. *Journal of the Acoustical Society of America*, 104(2):1118–1121, 1998.

R. P. Carlyon and T. M. Shackleton. Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms? *Journal of the Acoustical Society of America*, 95(6):3541–3554, 1994.

L. Cedolin and B. Delgutte. Spatiotemporal representation of the pitch of harmonic complex tones in the auditory nerve. *Journal of Neuroscience*, 30(38):12712–12724, 2010.

C. Chambers and D. Pressnitzer. The effect of context in the perception of an ambiguous pitch stimulus. *Association for Research in Otolaryngology Abstract*, pages 346–347, 2011.

M. Clynes, editor. *Music, Mind and Brain.* Plenum Press, New York, 1982.

M. A. Cohen, S. Grossberg, and L. L. Wyse. A spectral network model of pitch perception. *Journal of the Acoustical Society of America*, 98(2):862–879, 1995.

N. P. Cooper and W. S. Rhode. Mechanical responses to two-tone distortion products in the apical and basal turns of the mammalian cochlea. *Journal of Neurophysiology*, 78(1):261–270, 1997.

E. M. Cramer and W. H. Huggins. Creation of pitch through binaural interaction. *Journal of the Acoustical Society of America*, 30(5):413–417, 1958.

J. F. Culling, A. Q. Summerfield, and D. H. Marshall. Dichotic pitches as illusions of binaural unmasking. I. Huggins' pitch and the “binaural edge pitch”. *Journal of the Acoustical Society of America*, 103(6):3509–3526, 1998.

H. Dai. On the relative influence of individual harmonics on pitch judgment. *Journal of the Acoustical Society of America*, 107(2):953–959, 2000.

L. Dawe, J. Platt, and E. Welsh. Spectral-motion aftereffects and the tritone paradox among canadian subjects. *Attention, Perception & Psychophysics*, 60:209–220, 1998.

E. de Boer. Pitch of inharmonic signals. *Nature*, 178(4532):535–536, 1956a.

E. de Boer. *On the "Residue" in Hearing.* PhD thesis, University of Amsterdam, Faculty of Mathematics and Physics, 1956b.

E. de Boer. Pitch theories unified. In E. F. Evans and J. P. Wilson, editors, *Psychophysics and Physiology of Hearing: An International Symposium.* Academic Press, London, 1977.

E. de Boer and H. R. de Jongh. On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America*, 63(1):115–135, 1978.

E. de Boer and C. Kruidenier. On ringing limits of the auditory periphery. *Biological Cybernetics*, 63:433–442, 1990.

E. de Boer and P. Kuyper. Triggered correlation. *IEEE Transactions on Biomedical Engineering*, BME-15(3):169 –179, 1968.

A. de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 93(6):3271–3290, 1993.

A. de Cheveigné. Cancellation model of pitch perception. *Journal of the Acoustical Society of America*, 103(3):1261–1271, 1998.

A. de Cheveigné. Pitch perception models. In C. Plack, R. Fay, A. Oxenham, and A. Popper, editors, *Pitch: Neural Coding and Perception*, volume 24 of *Springer Handbook of Auditory Research*, pages 169–233. Springer New York, 2005.

A. de Cheveigné. Pitch perception. In C. Plack, editor, *The Oxford Handbook of Auditory Science: Hearing*, volume 3, pages 71–104. Oxford University Press, 2010.

A. de Cheveigné and D. Pressnitzer. The case of the missing delay lines: Synthetic delays obtained by cross-channel phase interaction. *Journal of the Acoustical Society of America*, 119(6):3908–3918, 2006.

A. de Cheveigné, H. Kawahara, M. Tsuzaki, and K. Aikawa. Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. *Journal of the Acoustical Society of America*, 101(5):2839–2847, 1997.

B. Delgutte. Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. *Journal of the Acoustical Society of America*, 87(2):791–809, 1990.

D. Deutsch. A musical paradox. *Music Perception: An Interdisciplinary Journal*, 3(3): 275–280, 1986.

H. Duifhuis, L. F. Willems, and R. J. Sluyter. Measurement of pitch in speech: An implementation of goldstein's theory of pitch perception. *Journal of the Acoustical Society of America*, 71(6):1568–1580, 1982.

M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

G. Fant. *Acoustic Theory of Speech Production.* Mouton & Co, The Hague, 1960.

H. Fastl. Ueber Tonhöhenempfindungen bei Rauschen. *Acustica*, pages 350–354, 1971.

H. Fastl and E. Zwicker. *Psychoacoustics: Facts and Models.* Springer series in information sciences. Springer, 3rd edition, 2007.

Y. I. Fishman, D. H. Reser, J. C. Arezzo, and M. Steinschneider. Pitch vs. spectral encoding of harmonic complex tones in primary auditory cortex of the awake monkey. *Brain Research*, 786(1-2):18 – 30, 1998.

J. L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer, Berlin - New York, 2nd edition, 1972.

J. L. Flanagan and N. Guttman. On the pitch of periodic pulses. *Journal of the Acoustical Society of America*, 32(10):1308–1319, 1960.

J. L. Flanagan, N. Guttman, and B. J. Watson. Pitch of periodic pulses with nonuniform amplitudes. *Journal of the Acoustical Society of America*, 34(5):738–739, 1962.

G. Frank, W. Hemmert, and A. W. Gummer. Limiting dynamics of high-frequency electromechanical transduction of outer hair cells. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4420–4425, 1999.

R. D. Frisina, R. L. Smith, and S. C. Chamberlain. Encoding of amplitude modulation in the gerbil cochlear nucleus: I. A hierarchy of enhancement. *Hearing Research*, 44 (2-3):99–122, 1990.

J. Fritz, S. Shamma, M. Elhilali, and D. Klein. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience*, 6(11): 1216–1223, 2003.

J. B. Fritz, M. Elhilali, and S. A. Shamma. Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. *Journal of Neuroscience*, 25(33):7623–7635, 2005.

R. R. Gacek. *Neuroanatomy of the auditory system*, volume 2. Academic Press, New York, 1972.

W. R. Garner. *The Processing of Information and Structure.* Lawrence Erlbaum, Oxford, 1974.

J. Giangrand, B. Tuller, and J. A. S. Kelso. Perceptual dynamics of circular pitch. *Music Perception: An Interdisciplinary Journal*, 20(3):241–262, 2003.

B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103 – 138, 1990.

H. Gockel, B. C. J. Moore, C. J. Plack, and R. P. Carlyon. Effect of noise on the detectability and fundamental frequency discrimination of complex tones. *Journal of the Acoustical Society of America*, 120(2):957–965, 2006.

J. L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54(1):317–317, 1973.

C. Gough. Musical acoustics. In T. D. Rossing, editor, *Springer Handbook of Acoustics*, chapter 15, pages 531–668. Springer, New York, 2007.

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

S. Grossberg, K. K. Govindarajan, L. L. Wyse, and M. A. Cohen. ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Networks*, 17(4):511 – 536, 2004.

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

J. J. Guinan, Jr. Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans. *Ear and Hearing*, 27(6):589–607, 2006.

J. J. Guinan, Jr. and K. M. Stankovic. Medial efferent inhibition produces the largest equivalent attenuations at moderate to high sound levels in cat auditory-nerve fibers. *Journal of the Acoustical Society of America*, 100(3):1680–1690, 1996.

N. Guttman and J. L. Flanagan. Pitch of high-pass-filtered pulse trains. *Journal of the Acoustical Society of America*, 36(4):757–765, 1964.

N. Guttman and S. Pruzansky. Lower limits of pitch and musical pitch. *Journal of Speech and Hearing Research*, 5(3):207–214, 1962.

T. A. Hackett, T. M. Preuss, and J. H. Kaas. Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. *The Journal of Comparative Neurology*, 441(3):197–222, 2001.

D. A. Hall and C. J. Plack. Pitch processing sites in the human auditory brain. *Cerebral Cortex*, 19(3):576–585, 2009.

W. M. Hartmann. *Signals, Sound, and Sensation.* AIP series in modern acoustics and signal processing. American Institute of Physics, 1997.

W. M. Hartmann and C. D. McMillon. Binaural coherence edge pitch. *Journal of the Acoustical Society of America*, 109(1):294–305, 2001.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

H. Helmholtz. Ueber Combinationstöne. *Annalen der Physik*, 175(12):497–540, 1856. ISSN 1521-3889.

S. Hemilä, S. Nummela, and T. Reuter. What middle ear parameters tell about impedance matching and high frequency hearing. *Hearing Research*, 85(1-2):31–44, 1995.

D. J. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

H.-P. Hesse. The judgement of musical intervals. In Clynes (1982), pages 217–225.

M. J. Hewitt and R. Meddis. A computer model of amplitude-modulation sensitivity of single units in the inferior colliculus. *Journal of the Acoustical Society of America*, 95(4):2145–2159, 1994.

T. Holton and T. F. Weiss. Frequency selectivity of hair cells and nerve fibres in the alligator lizard cochlea. *Journal of Physiology*, 345(1):241–260, 1983.

T. Houtgast. Psychophysical evidence for lateral inhibition in hearing. *Journal of the Acoustical Society of America*, 51(6B):1885–1894, 1972.

A. J. M. Houtsma. Pitch and timbre: Definition, meaning and use. *Journal of New Music Research*, 26(2), 1997.

A. J. M. Houtsma and J. F. M. Fleuren. Analytic and synthetic pitch of two-tone complexes. *Journal of the Acoustical Society of America*, 90(3):1674–1676, 1991.

A. J. M. Houtsma and J. L. Goldstein. Perception of musical intervals : evidence for the central origin of the pitch of complex tones. Technical Report TK7855.M41 R43 no.484, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1971. (based on a Ph.D. thesis in the Dept. of Electrical Engineering, 1971, by A.J.M. Houtsma).

A. J. M. Houtsma and J. L. Goldstein. The central origin of the pitch of complex tones: Evidence from musical interval recognition. *Journal of the Acoustical Society of America*, 51(2B):520–529, 1972.

A. J. M. Houtsma and J. Smurzynski. Pitch identification and discrimination for complex tones with many harmonics. *Journal of the Acoustical Society of America*, 87 (1):304–310, 1990.

E. Jaynes. *Probability theory: the logic of science.* Cambridge University Press, 2003.

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

D. H. Johnson. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *Journal of the Acoustical Society of America*, 68(4):1115–1122, 1980.

B. M. Johnstone, R. Patuzzi, and G. K. Yates. Basilar membrane measurements and the travelling wave. *Hearing Research*, 22(1-3):147 – 153, 1986.

E. Joliveau, J. Smith, and J. Wolfe. Vocal tract resonances in singing: The soprano voice. *Journal of the Acoustical Society of America*, 116(4):2434–2439, 2004.

S. C. Kadia and X. Wang. Spectral integration in a1 of awake primates: Neurons with single- and multipeaked tuning characteristics. *Journal of Neurophysiology*, 89(3): 1603–1622, 2003.

D. Kersten, P. Mamassian, and A. Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, 55(1):271–304, 2004.

D. O. Kim, J. G. Sirianni, and S. O. Chang. Responses of DCN-PVCN neurons and auditory nerve fibers in unanesthetized decerebrate cats to AM and pure tones: Analysis with autocorrelation/power-spectrum. *Hearing Research*, 45(1-2):95–113, 1990.

H.-G. Kim, N. Moreau, and T. Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval.* John Wiley & Sons, 2005.

M. A. Klein and W. M. Hartmann. Binaural edge pitch. *Journal of the Acoustical Society of America*, 70(1):51–61, 1981.

M. Kleiner, D. H. Brainard, and D. G. Pelli. What's new in Psychtoolbox-3? *Perception*, 36, 2007. ECVP Abstract Supplement.

D. C. Knill and W. Richards, editors. *Perception as Bayesian inference.* Cambridge University Press, New York, 1996.

B. S. Krishna and M. N. Semple. Auditory temporal processing: Responses to sinusoidally Amplitude-Modulated tones in the inferior colliculus. *Journal of Neurophysiology*, 84(1):255–273, 2000.

K. Krumbholz, R. D. Patterson, and D. Pressnitzer. The lower limit of pitch as determined by rate discrimination. *Journal of the Acoustical Society of America*, 108(3): 1170–1180, 2000.

C. L. Krumhansl and P. Iverson. Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance*, 18(3):739–751, 1992.

G. Langner and C. E. Schreiner. Periodicity coding in the inferior colliculus of the cat. i. neuronal mechanisms. *Journal of Neurophysiology*, 60(6):1799–1822, 1988.

G. Langner, M. Sams, P. Heil, and H. Schulze. Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 181:665–676, 1997.

G. Langner, M. Albert, and T. Briede. Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (Chinchilla laniger). *Hearing Research*, 168(1-2):110–130, 2002.

D. J. Levitin and S. E. Rogers. Absolute pitch: perception, coding, and controversies. *Trends in Cognitive Sciences*, 9(1):26 – 33, 2005.

W. H. Lichte. Attributes of complex tones. *Journal of Experimental Psychology*, 28(6): 455–480, 1941.

J. C. R. Licklider. A duplex theory of pitch perception. *Experientia*, 7:128–134, 1951.

B. Lindblom and J. Sundberg. *The Human Voice in Speech and Singing*, pages 669–712. Springer, New York, 2007.

G. R. Lockhead and R. Byrd. Practically perfect pitch. *The Journal of the Acoustical Society of America*, 70(2):387–389, 1981.

N. K. Logothetis. The underpinnings of the bold functional magnetic resonance imaging signal. *The Journal of Neuroscience*, 23(10):3963–3971, 2003.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, 2003.

M. S. Malmierca and T. S. Hacket. Structural organization of the ascending auditory pathway. In A. Rees and A. R. Palmer, editors, *The Oxford Handbook of Auditory Science: The Auditory Brain*, volume 2. Oxford University Press, 2010.

P. Mamassian, M. S. Landy, and L. T. Maloney. Bayesian modelling of visual perception. In R. Rao, B. Olshausen, and M. Lewicki, editors, *Probabilistic Models of the Brain: Perception and Neural Function*, pages 13–36. MIT Press, Cambridge, MA, 2002.

J. Marozeau and A. de Cheveigné. The effect of fundamental frequency on the brightness dimension of timbre. *Journal of the Acoustical Society of America*, 121(1): 383–387, 2007.

J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg. The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, 114 (5):2946–2957, 2003.

J. P. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg. The perceptual interaction between the pitch and timbre of musical sound. *Journal of the Acoustical Society of America*, 109(5):2288–2288, 2001.

D. Marr. *Vision: A computational investigation into the human representation and processing of visual information.* W. H. Freeman, San Francisco, 1982.

S. McAdams and A. Bregman. Hearing musical streams. *Computer Music Journal*, 3 (4):26–60, 1979.

S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192, 1995.

D. McAlpine. Neural sensitivity to periodicity in the inferior colliculus: Evidence for the role of cochlear distortions. *Journal of Neurophysiology*, 92(3):1295–1311, 2004.

J. H. McDermott, A. J. Lehr, and A. J. Oxenham. Individual differences reveal the basis of consonance. *Current Biology*, 20(11):1035–1041, 2010.

R. Meddis. Auditory-nerve first-spike latency and auditory absolute threshold: A computer model. *Journal of the Acoustical Society of America*, 119(1):406–417, 2006.

R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America*, 89(6):679–688, 1991.

R. Meddis and L. O'Mard. A unitary model of pitch perception. *Journal of the Acoustical Society of America*, 102(3):1811–1820, 1997.

R. D. Melara and L. E. Marks. Interaction among auditory dimensions: timbre, pitch, and loudness. *Perception & Psychophysics*, 48(2):169–178, 1990.

T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005.

B. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, 5th edition, 2003.

B. C. Moore and H. E. Gockel. Resolvability of components in complex tones and implications for theories of pitch perception. *Hearing Research*, 276(1-2):88 – 97, 2011. Annual Reviews 2011.

B. C. J. Moore. Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America*, 54(3):610–619, 1973.

B. C. J. Moore, B. R. Glasberg, and R. W. Peters. Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, 77(5), 1985.

C. T. Morgan, W. R. Garner, and R. Galambos. Pitch and intensity. *Journal of the Acoustical Society of America*, 23(6):658–663, 1951.

T. Moser, A. Brandt, and A. Lysakowski. Hair cell ribbon synapses. *Cell and Tissue Research*, 326:347–359, 2006.

L. Mozart. *Versuch einer gründlichen Violinschule.* Augsburg, 1756. English translation by E. Knocker, *A Treatise on the Fundamental Principles of Violin Playing*, Oxford University Press, London, 1947.

I. Murray. *Advances in Markov chain Monte Carlo methods.* PhD thesis, Gatsby computational neuroscience unit, University College London, 2007.

R. M. Neal. Statistical inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Sience, University of Toronto, 1993.

R. M. Neal. Annealed importance sampling. Technical Report 9805, Department of Statistics, University of Toronto, 1998.

R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo: Methods and Applications.* CRC Press, 2010.

I. Nelken. Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology*, 14(4):474–480, 2004.

I. Nelken, J. K. Bizley, F. R. Nodal, B. Ahmed, A. J. King, and J. W. H. Schnupp. Responses of auditory cortex to complex stimuli: Functional organization revealed using intrinsic optical signals. *Journal of Neurophysiology*, 99(4):1928–1941, 2008.

M. A. Newton and A. E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48, 1994.

J. Nocedal and S. J. Wright. *Numerical optimization.* Springer, New York, 2nd edition, 1999.

D. Oertel, R. Bal, S. M. Gardner, P. H. Smith, and P. X. Joris. Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11773–11779, 2000.

G. S. Ohm. Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. *Annalen der Physik*, 135(8):513–565, 1843.

A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing*, chapter 7. Prentice Hall, 2nd edition, 1999.

A. J. Oxenham and C. A. Shera. Estimates of human cochlear tuning at low levels using forward and simultaneous masking. *Journal of the Association for Research in Otolaryngology*, 4:541–554, 2003.

A. J. Oxenham, J. G. W. Bernstein, and H. Penagos. Correct tonotopic representation is necessary for complex pitch perception. *Proceedings of the National Academy of Sciences of the United States of America*, 101(5):1421–1425, 2004.

A. J. Oxenham, C. Micheyl, M. V. Keebler, A. Loper, and S. Santurette. Pitch perception beyond the traditional existence region of pitch. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7629–7634, 2011.

C. Pantev, M. Hoke, B. Ltkenhner, and K. Lehnertz. Tonotopic organization of the auditory cortex: Pitch versus frequency representation. *Science*, 246(4929):486–488, 1989.

R. D. Patterson. The effects of relative phase and the number of components on residue pitch. *Journal of the Acoustical Society of America*, 53(6):1565–1572, 1973.

R. D. Patterson. Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 59(3):640–654, 1976.

R. D. Patterson. Auditory images: How complex sounds are represented in the auditory system. *Journal of the Acoustical Society of Japan E*, 21(4):183–190, 2000.

R. D. Patterson and B. C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. In B. C. J. Moore, editor, *Frequency Selectivity in Hearing*. Academic Press, London, 1986.

R. D. Patterson and F. L. Wightman. Resiude pitch as a function of component spacing. *Journal of the Acoustical Society of America*, 59(6):1450–1459, 1976.

R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In Y. Cazals, L. Demany, and K. Honer, editors, *Auditory Physiology and Perception*, pages 429–443. Pergamon, Oxford, 1992.

R. D. Patterson, M. H. Allerhand, and C. Giguère. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America*, 98(4):1890–1894, 1995.

R. D. Patterson, S. Uppenkamp, I. S. Johnsrude, and T. D. Griffiths. The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36(4):767–776, 2002.

R. D. Patterson, D. R. R. Smith, R. van Dinther, and T. C. Walters. Size information in the production and perception of communication sounds. In W. A. Yost, A. N. Popper, and R. R. Fay, editors, *Auditory Perception of Sound Sources*, volume 16. Springer Verlag, 2007.

R. D. Patterson, D. R. R. Smith, R. van Dinther, and T. C. Walter. Size information in the production and perception of communication sounds. In W. A. Yost, A. N. Popper, and R. R. Fay, editors, *Auditory perception of sound sources*, Springer Handbook of Auditory Research (vol. 29). Springer, 2008.

H. Penagos, J. R. Melcher, and A. J. Oxenham. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience*, 24(30):6810–6815, 2004.

G. E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184, 1952.

H. Pick, D. Warren, and J. Hay. Sensory conflict in judgments of spatial direction. *Attention, Perception & Psychophysics*, 6:203–205, 1969.

M. A. Pitt. Perception of pitch and timbre by musically trained and untrained listeners.

Journal of Experimental Psychology: Human Perception and Performance, 20(5): 976–986, 1994.

R. Plomp. Pitch of complex tones. *Journal of the Acoustical Society of America*, 41 (6), 1967.

R. Plomp. Timbre as a multidimensional attribute of complex tones. In R. Plomp and G. F. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*, pages 397–414. Sijthoff, Leiden, 1970.

R. Plomp and W. J. M. Levelt. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, 38(4):548–560, 1965.

R. Plomp and H. J. M. Steeneken. Pitch versus timbre. In *Proceedings of the 7th International Congress of Acoustics*, pages 377–380, Budapest, 1971.

I. Pollack. Amplitude and time jitter thresholds for rectangular-wave trains. *Journal of the Acoustical Society of America*, 50(4B):1133–1142, 1971.

D. Pressnitzer, R. D. Patterson, and K. Krumbholz. The lower limit of melodic pitch. *Journal of the Acoustical Society of America*, 109(5):2074–2084, 2001.

A. Preuss and P. Müller-Preuss. Processing of amplitude modulated sounds in the medial geniculate body of squirrel monkeys. *Experimental Brain Research*, 79(1), 1990.

S. Puschmann, S. Uppenkamp, B. Kollmeier, and C. M. Thiel. Dichotic pitch activates pitch processing centre in heschl's gyrus. *NeuroImage*, 49(2):1641–1649, 2010.

C. E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, Cambridge, MA, 2006.

R. Renken, J. E. C. Wiersinga-Post, S. Tomaskovic, and H. Duifhuis. Dominance of missing fundamental versus spectrally cued pitch: Individual differences for complex tones with unresolved harmonics. *Journal of the Acoustical Society of America*, 115 (5):2257–2263, 2004.

B. Repp and J. Thompson. Context sensitivity and invariance in perception of octave-ambiguous tones. *Psychological Research*, 74:437–456, 2010.

W. S. Rhode. Observations of the vibration of the basilar membrane in squirrel monkeys using the mössbauer technique. *Journal of the Acoustical Society of America*, 49(4B): 1218–1231, 1971.

W. S. Rhode. An investigation of postmortem cochlear mechanics using the Mössbauer effect. In A. Møller, editor, *Basic Mechanisms of Hearing*, pages 49–63. Academic Press, New York, 1973.

W. S. Rhode. Temporal coding of 200% amplitude modulated signals in the ventral cochlear nucleus of cat. *Hearing Research*, 77(1-2):43–68, 1994.

W. S. Rhode and S. Greenberg. Lateral suppression and inhibition in the cochlear nucleus of the cat. *Journal of Neurophysiology*, 71(2):493–514, 1994.

B. L. Riker. The ability to judge pitch. *Journal of Experimental Psychology*, 36(4): 331–346, 1946.

R. J. Ritsma. Existence region of the tonal residue I. *Journal of the Acoustical Society of America*, 34(9A):1224–1229, 1962.

R. J. Ritsma. Existence region of the tonal residue II. *The Journal of the Acoustical Society of America*, 35(8):1241–1245, 1963.

R. J. Ritsma. Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America*, 42(1):191–198, 1967.

D. W. Robinson and R. S. Dadson. A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5):166–181, 1956.

K. Robinson. Brightness and octave position: are changes in spectral envelope and in tone height perceptually equivalent? *Contemporary Music Review*, 9(1):83–95, 1993.

L. Robles, M. A. Ruggero, and N. C. Rich. Two-tone distortion in the basilar membrane of the cochlea. *Nature*, 349:413–414, Jan 1991.

L. Robles, M. A. Ruggero, and N. C. Rich. Two-tone distortion on the basilar membrane of the chinchilla cochlea. *Journal of Neurophysiology*, 77(5):2385–2399, 1997.

J. E. Rose, J. F. Brugge, D. J. Anderson, and J. E. Hind. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of Neurophysiology*, 30(4):769–793, 1967.

S. Rosen, R. J. Baker, and A. Darling. Auditory filter nonlinearity at 2 kHz in normal hearing listeners. *Journal of the Acoustical Society of America*, 103(5):2539–2550, 1998.

M. A. Ruggero and N. C. Rich. Furosemide alters organ of corti mechanics: evidence for feedback of outer hair cells upon the basilar membrane. *Journal of Neuroscience*, 11(4):1057–1067, 1991.

M. A. Ruggero and A. N. Temchin. Unexceptional sharpness of frequency tuning in the human cochlea. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18614–18619, 2005.

M. A. Ruggero, L. Robles, and N. C. Rich. Two-tone suppression in the basilar membrane of the cochlea: mechanical basis of auditory-nerve rate suppression. *Journal of Neurophysiology*, 68(4):1087–1099, 1992a.

M. A. Ruggero, L. Robles, N. C. Rich, A. Recio, A. M. Brown, and E. F. Evans. Basilar membrane responses to two-tone and broadband stimuli [and discussion]. *Philosophical Transactions: Biological Sciences*, 336(1278):307–315, 1992b.

M. A. Ruggero, N. C. Rich, and A. Recio. Alteration of basilar membrane responses to sound by acoustic overstimulation. In D. H., H. J.W., van Dijk P., and van Netten S. M., editors, *Biophysics of Hair Cell Sensory Systems*, pages 258–264. World Scientific, Singapore, 1993.

M. A. Ruggero, N. C. Rich, L. Robles, and A. Recio. The effects of acoustic trauma, other cochlear injury, and death on basilar membrane responses to sound. In A. Axelsson, H. M. Borchgrevink, R. P. Hamernik, P.-A. Hellstrom, D. Henderson, and R. J. Salvi, editors, *Scientific Basis of Noise-Induced Hearing Loss*, pages 23 – 35. Thieme, Stuttgart, 1996.

M. A. Ruggero, N. C. Rich, A. Recio, S. S. Narayan, and L. Robles. Basilar-membrane responses to tones at the base of the chinchilla cochlea. *Journal of the Acoustical Society of America*, 101(4):2151–2163, 1997.

I. J. Russell and E. Murugasu. Medial efferent inhibition suppresses basilar membrane responses to near characteristic frequency tones of moderate to high intensities. *Journal of the Acoustical Society of America*, 102(3):1734–1738, 1997.

I. J. Russell and P. M. Sellick. Low-frequency characteristics of intracellularly recorded receptor potentials in guinea-pig cochlear hair cells. *Journal of Physiology*, 338(1): 179–206, 1983.

J. Santos-Sacchi. On the frequency limit and phase of outer hair cell motility: effects of the membrane filter. *Journal of Neuroscience*, 12(5):1906–1916, 1992.

J. F. Schouten. The perception of subjective tones. In *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, volume 41, pages 1086–1093, 1938.

J. F. Schouten. The perception of pitch. *Philips Technical Review*, 5(10), 1940.

J. F. Schouten, R. J. Ritsma, and B. L. Cardozo. Pitch of the residue. *Journal of the Acoustical Society of America*, 34(9B):1418–1424, 1962.

C. E. Schreiner and G. Langner. Periodicity coding in the inferior colliculus of the cat. II. topographical organization. *Journal of Neurophysiology*, 60(6):1823–1840, 1988.

M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

E. Schubert and J. Wolfe. Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica united with Acustica*, 92:820–825, 2006.

D. W. Schwarz and R. W. Tomlinson. Spectral response patterns of auditory cortex neurons to harmonic complex tones in alert monkey (macaca mulatta). *Journal of Neurophysiology*, 64(1):282–298, 1990.

A. Seebeck. Ueber die Sirene. *Annalen der Physik*, 136(12):449–481, 1843.

C. Semal and L. Demany. The upper limit of "musical" pitch. *Music Perception*, 8(2): 165–175, 1990. ISSN 0730-7829.

T. M. Shackleton and R. P. Carlyon. The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *Journal of the Acoustical Society of America*, 95(6):3529–3540, 1994.

S. Shamma and D. Klein. The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *Journal of the Acoustical Society of America*, 107(5):2631–2644, 2000.

L. Shams and U. R. Beierholm. Causal inference in perception. *Trends in Cognitive Sciences*, 14(9):425 – 432, 2010.

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

R. N. Shepard. Geometrical approximations to the structure of musical pitch. *Psychological Review*, 89(4):305 – 333, 1982.

C. A. Shera, J. J. Guinan, and A. J. Oxenham. Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5):3318–3323, 2002.

N. H. Silbert, J. T. Townsend, and J. J. Lentz. Independence and separability in the perception of complex nonspeech sounds. *Attention, Perception & Psychophysics*, 71 (8):1900–1915, 2009.

A. P. Simpson. Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2):621–640, 2009.

P. G. Singh and I. J. Hirsh. Influence of spectral locus and F0 changes on the pitch and timbre of complex tones. *Journal of the Acoustical Society of America*, 92(5): 2650–2661, 1992.

M. Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. Technical report, Apple Computer, Inc., 1993.

M. Slaney and R. Lyon. A perceptual pitch detector. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-90)*, volume 1, pages 357–360, 1990.

A. M. Small, Jr. and R. G. Daniloff. Pitch of noise bands. *Journal of the Acoustical Society of America*, 41(2):506–512, 1967.

E. C. Smith and M. S. Lewicki. Efficient coding of time-relative structure using spikes. *Neural Computation*, 17(1):19–45, 2005.

E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006.

G. F. Smoorenburg. Pitch perception of two-frequency stimuli. *Journal of the Acoustical Society of America*, 48(4B):924–942, 1970.

E. Snelson, C. E. Rasmussen, and Z. Ghahramani. Warped Gaussian processes. In *Advances in Neural Information Processing Systems 16 (NIPS 2003)*. MIT Press, 2004.

P. Srulovicz and J. L. Goldstein. A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *Journal of the Acoustical Society of America*, 73(4):1266–1276, 1983.

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

S. S. Stevens. Issues in psychophysical measurement. *Psychological Review*, 78(5):426 – 450, 1971.

S. S. Stevens and J. Volkmann. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, 53(3):329–353, 1940.

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3): 185–190, 1937.

E. A. Strickland and N. F. Viemeister. The effects of frequency region and bandwidth on the temporal modulation transfer function. *Journal of the Acoustical Society of America*, 102(3):1799–1810, 1997.

C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis. A revised model of the inner-hair cell and auditory-nerve complex. *Journal of the Acoustical Society of America*, 111(5):2178–2188, 2002.

E. Terhardt. Pitch, consonance, and harmony. *Journal of the Acoustical Society of America*, 55(5):1061–1069, 1974.

E. Terhardt. Response to E. de Boer: Pitch theories unified. In E. F. Evans and J. P. Wilson, editors, *Psychophysics and Physiology of Hearing: An International Symposium*. Academic Press, London, 1977.

E. Terhardt. Calculating virtual pitch. *Hearing Research*, 1(2):155 – 182, 1979.

E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustical Society of America*, 71(3):679–688, 1982.

R. E. Turner and M. Sahani. Demodulation as probabilistic inference. *IEEE Transactions on Audio, Speech, and Language Processing*, PP(99), 2011.

R. S. Turner. The Ohm-Seebeck dispute, Hermann von Helmholtz, and the origins of Physiological Acoustics. *The British Journal for the History of Science*, 10(1):1–24, 1977.

N. Ulanovsky, L. Las, and I. Nelken. Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, 6(4):39–398, 2003.

R. van Dinther and R. D. Patterson. Perception of acoustic scale and size in musical instrument sounds. *Journal of the Acoustical Society of America*, 120(4):2158–2176, 2006.

L. van Norden. Two channel pitch perception. In Clynes (1982), pages 251–269.

J. Vliegen and A. J. Oxenham. Sequential stream segregation in the absence of spectral cues. *Journal of the Acoustical Society of America*, 105(1):339–346, 1999.

G. von Békésy. *Experiments in Hearing.* McGraw-Hill, New York, 1960.

G. von Békésy. Hearing theories and complex sounds. *Journal of the Acoustical Society of America*, 35(4):588–601, 1963.

H. L. F. von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Vieweg, Brauschweig, 1863. English translation by A. J. Ellis, *On the Sensations of Tone*, Dover, New York, 1954.

H. L. F. von Helmholtz. *Handbuch der physiologischen Optik.* Leopold Voss, Leipzig, 1867. English translation by J. P. C. Southall, *Helmholtz's treatise on physiological optics*, Dover, New York, 1962.

A. Vurma, M. Raju, and A. Kuuda. Does timbre affect pitch? Estimations by musicians and non-musicians. *Psychology of Music*, 2010.

K. M. Walker, J. K. Bizley, A. J. King, and J. W. Schnupp. Cortical encoding of pitch: Recent results and open questions. *Hearing Research*, 271(1-2):74 – 87, 2011.

M. N. Wallace, T. M. Shackleton, and A. R. Palmer. Phase-locked responses to pure tones in the primary auditory cortex. *Hearing Research*, 172(1-2):160–171, 2002.

M. N. Wallace, L. A. Anderson, and A. R. Palmer. Phase-locked responses to pure tones in the auditory thalamus. *Journal of Neurophysiology*, 98(4):1941–1952, 2007.

X. Wang, T. Lu, D. Bendor, and E. Bartlett. Neural coding of temporal information in auditory thalamus and cortex. *Neuroscience*, 154(1):294–303, 2008.

W. D. Ward. Subjective musical pitch. *Journal of the Acoustical Society of America*, 26(3):369–380, 1954.

C. Warrier and R. Zatorre. Influence of tonal context and timbral variation on perception of pitch. *Attention, Perception & Psychophysics*, 64:198–207, 2002. ISSN 1943-3921.

Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604, 2002.

G. Westheimer. Was Helmholtz a Bayesian ? *Perception*, 37(5):642–650, 2008.

I. C. Whitfield. Auditory cortex and the pitch of complex tones. *Journal of the Acoustical Society of America*, 67(2):644–647, 1980.

L. Wiegrebe. Searching for the time constant of neural pitch extraction. *Journal of the Acoustical Society of America*, 109(3):1082–1091, 2001.

L. Wiegrebe and R. Meddis. The representation of periodic sounds in simulated sustained chopper units of the ventral cochlear nucleus. *Journal of the Acoustical Society of America*, 115(3):1207, 2004.

L. Wiegrebe and R. D. Patterson. Quantifying the distortion products generated by amplitude-modulated noise. *The Journal of the Acoustical Society of America*, 106 (5):2709–2718, 1999.

F. L. Wightman. The pattern-transformation model of pitch. *Journal of the Acoustical Society of America*, 54(2):407–416, 1973.

I. M. Winter. The neurophysiology of pitch. In C. Plack, R. Fay, A. Oxenham, and A. Popper, editors, *Pitch: Neural Coding and Perception*, volume 24 of *Springer Handbook of Auditory Research*, pages 99–146. Springer New York, 2005.

I. M. Winter, L. Wiegrebe, and R. D. Patterson. The temporal representation of the delay of iterated rippled noise in the ventral cochlear nucleus of the guinea-pig. *Journal of Physiology*, 537(Pt 2):553–566, Dec. 2001.

D. R. Wozny, U. R. Beierholm, and L. Shams. Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, 6(8):e1000871, 2010.

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

W. A. Yost and R. Hill. Models of the pitch and pitch strength of ripple noise. *Journal of the Acoustical Society of America*, 66(2):400–410, 1979.

W. A. Yost, R. Hill, and T. Perez-Falcon. Pitch and pitch discrimination of broadband signals with rippled power spectra. *Journal of the Acoustical Society of America*, 63 (4):1166–1173, 1978.

W. A. Yost, R. Patterson, and S. Sheft. A time domain description for the pitch strength of iterated rippled noise. *Journal of the Acoustical Society of America*, 99 (2):1066–1078, 1996.

J. E. Zakrisson. The effect of the stapedius reflex on attenuation and poststimulatory auditory fatigue at different frequencies. *Acta Otolaryngol Suppl*, 360:118–121, 1979.

R. J. Zatorre. Pitch perception of complex tones and human temporal-lobe function. *Journal of the Acoustical Society of America*, 84(2):566–572, 1988.

J. Zheng, W. Shen, D. Z. Z. He, K. B. Long, L. D. Madison, and P. Dallos. Prestin is the motor protein of cochlear outer hair cells. *Nature*, 405:149–155, May 2000.

G. Zweig. Basilar membrane motion. *Cold Spring Harbor Symposia on Quantitative Biology*, 40:619–633, 1976.