

6

Pitch perception models

Alain de Cheveigné

1 Introduction

This chapter discusses models of pitch, old and recent. The aim is to chart their common points – many are variations on a theme – and differences, and build a catalog of ideas for use in understanding pitch perception. The busy reader might read just the next section, a crash course in pitch theory that explains why some obvious ideas don't work and what are currently the best answers. The brave reader will read on as we delve more deeply into the origin of concepts, and the intricate and ingenious ideas behind the models and metaphors that we use to make progress in understanding pitch.

2 Pitch Theory in a Nutshell

Pitch-evoking stimuli usually are periodic, and the pitch usually is related to the period. Accordingly, a pitch perception mechanism must estimate the period T (or its inverse, the fundamental frequency F_0) of the stimulus. There are two approaches to do so. One involves the *spectrum* and the other the *waveform*. The two are illustrated with examples of stimuli that evoke pitch, such as pure and complex tones.

2.1 Spectrum

The spectral approach is based upon Fourier analysis. The spectrum of a pure tone is illustrated in Figure 1A. An algorithm to measure its period (inverse of its frequency) is to look for the spectral peak and use its position as a cue to pitch. This works for a pure tone, but consider now the sound illustrated in Figure 1B, that evokes the same pitch. There are several peaks in the spectrum, but the previous algorithm was designed to expect only one. A reasonable modification is to take the *largest* peak, but consider now the sound illustrated in Figure 1C. The largest spectral peak is at a higher harmonic, yet the pitch is still the same. A reasonable modification is to replace the largest peak by the peak of *lowest frequency*, but consider now the sound illustrated in Figure 1D. The lowest peak is at a higher harmonic, yet the pitch is still the same. A reasonable modification is to use the *spacing* between partials as a measure of period. That is all the more

reasonable as it often determines the frequency of the *temporal envelope* of the sound, as well as the frequency of possible *difference tones* (distortion products) due to nonlinear interaction between adjacent partials. However, consider now the sound illustrated in Figure 1E. None of the inter-partial intervals corresponds to its pitch, which (for some listeners) is the same as that of the other tones.

This brings us to a final algorithm. Build a histogram in the following way: for each partial, find its subharmonics by dividing the frequency of the partial by successive small integers. For each subharmonic, increment the corresponding histogram bin. Applied to the spectrum in Figure 1E, this produces the histogram illustrated in Figure 1F. Among the bins, some are larger than the rest. The *rightmost* of the (infinite) set of largest bins is the cue to pitch. This algorithm works for all the spectra shown. It illustrates the principle of *pattern matching* models of pitch perception.

2.2 Waveform

The waveform approach operates directly on the stimulus waveform. Consider again our pure tone, illustrated in the time domain in Figure 2A. Its periodic nature is obvious as a regular repetition of the waveform. A way to measure its period is to find *landmarks* such as peaks (shown as arrows) and measure the interval between them. This works for a pure tone, but consider now the sound in Figure 2B that evokes the same pitch. It has two peaks within each period, whereas our algorithm expects only one. A trivial modification is to use the *most prominent* peak of each period, but consider now the sound in Figure 2C. Two peaks are equally prominent. A tentative modification is to use *zero-crossings* (e.g. negative-to-positive) rather than peaks, but then consider the sound in Figure 2D, which has the same pitch but several zero-crossings per period. Landmarks are an awkward basis for period estimation: it is hard to find a marking rule that works in every case. The waveform in Figure 2D has a clearly defined *temporal envelope* with a period that matches its pitch, but consider now the sound illustrated in Figure 2E. Its pitch does not match the period of its envelope (as long as the ratio of carrier to modulation frequencies is less than about 10, see Plack and Oxenham, Chapter 2).

This brings us to a final algorithm that uses, as it were, every sample as a “landmark”. Each sample is compared to every other in turn, and a count is kept of the inter-sample intervals for which the match is good. Comparison is done by taking the *product*, which tends to be large if samples $x(t)$ and $x(t-\tau)$ are similar, as when τ is equal to the period T . Mathematically:

$$r(\tau) = \int x(t)x(t-\tau)dt \quad (1)$$

defines the *autocorrelation function*, illustrated in Figure 2F. For a periodic sound, the function is maximum at $\tau=0$, at the period, and at all its multiples. The *first* of these maxima with a *strictly positive* abscissa can be used as a cue to the period. This algorithm is the basis of what is known as the autocorrelation (AC) model of pitch. Autocorrelation and pattern matching are both adequate to measure periods as required by a pitch model, and they form the basis of modern theories of pitch perception.

We reviewed a number of principles, of which some worked and others not. All have been used in one pitch model or another. Those that use a flawed principle can (once the flaw is recognized) be ruled out. It is harder to know what to do with the models that remain. The rest of this chapter tries to chart out their similarities and differences. The approach is in part historical, but the focus is on the future more than on the past: in what direction should we take our next step to improve our understanding of pitch?

2.3 What is a model?

An important source of disagreement between pitch models, often not explicit, is what to expect of a *model*. The word is used with various meanings. A very broad definition is: *a thing that represents another thing in some way that is useful*. This definition also fits other words such as *theory, map, analogue, metaphor, law*, etc., all of which have a place in this review. “Useful” implies that the model represents its object faithfully, and yet is somehow easier to handle and thus *distinct* from its object. Norbert Wiener is quoted as saying: “The best material model of a cat is another, or preferably the same, cat.” I disagree: a cat is no easier to handle than itself, and thus not a useful model. Model and world must differ. Faithfulness is not sufficient. Figure 3 gives an example of a model that is obviously “wrong” and yet useful.

There are several corollaries. Every model is “false” in that it cannot match reality in all respects (Hebb 1959). Mismatch being allowed, multiple models may usefully serve a common reality. One pitch model may predict behavioral data quantitatively, while another is easier to explain, and a third fits physiology more closely. Criteria of quality are not one-dimensional, so models cannot always be ordered from best to worst. Rather than pit them one against another until just one (or none) remains, it is fruitful to see models as *tools* of which a craftsman might want several. Taking a metaphor from biology, we might argue for the “biodiversity” of models, which excludes neither competition nor the concept of “survival of the fittest”. Licklider (1959) put it this way:

The idea is simply to carry around in your head as many formulations as you can that are self-consistent and consistent with the empirical facts you know. Then, when you make an observation or read a paper, you find yourself saying, for example, “Well that certainly makes it look bad for the idea that sharpening occurs in the cochlear excitation process”.

Beginners in the field of pitch, reading of an experiment that contradicts a theory, are puzzled to find the disqualified theory live on until a new experiment contradicts its competitors. De Boer (1976) used the metaphor of the swing of a pendulum to describe such a phenomenon. An evolutionary metaphor is also fitting: as one theory reaches dominance, the others retreat to a sheltered ecological niche (where they may actually mutate at a faster pace). This review attempts yet another metaphor, that of “genetic manipulation”, in which pieces of models (“model DNA”) are isolated so that they may be recombined, hopefully speeding the evolution of our understanding of pitch. We shall use a historical perspective to help

isolate these significant strands. Before that, we need to discuss two more subjects of discord: the physical dimensions of stimuli and the psychological dimensions of pitch.

2.4 Stimulus descriptions

A second source of discord is *stimulus descriptions*. There are several ways to describe and parameterize stimuli that evoke a pitch. Some fit a wide range of stimuli, others a narrower range but with some other advantage. The “best choice” depends on the problem at hand. Whatever the choice, it is important to realize that the stimulus usually differs more or less from its idealized description (one could speak of a “model” of the stimulus). We use this opportunity to introduce some notations that will be useful later on.

A first description is the *periodic* signal (Fig. 4A). A signal $x(t)$ is periodic if there exists a number $T \neq 0$ such that $x(t) = x(t - T)$ for all time t . If there is one such number, there is an infinite set of them, and the *period* is defined as the smallest strictly positive member of that set (others are integer multiples). This representation is *parameterized* by the period T and by the shape of the waveform during a period: $x(t)$, $0 < t \leq T$. Stimuli differ from this description in various ways: they may be of finite duration, inharmonic, modulated in frequency or amplitude, or mixed with noise, etc. The description is nevertheless useful: stimuli that fit it well tend to have a clear pitch that depends on T .

A second description is the *sinusoid*, defined as $x(t) = A \cos(ft + \phi)$ where A is amplitude, f frequency and ϕ the starting phase (Fig. 4B). A sinusoid is periodic with period $T = 1/f$, so this description is a special case of the previous one. Sinusoids have an additional useful property: feeding one to a *linear time-invariant* system produces a sinusoid at the output. Its amplitude is multiplied by a fixed factor and its phase is shifted by a fixed amount, but it remains a sinusoid and its frequency is still f . Many acoustic processes are linear and time invariant. This makes the sinusoid an extremely useful description.

Supposing our stimulus is almost, but not quite, sinusoidal, should we use the better-fitting periodic description, or the more tractable sinusoidal description? The advantages of the latter might make us tolerate a less good fit. Disagreement between pitch perception models can be traced, in part, to a different answer to this question.

A third way of describing a pitch-evoking stimulus is as a *sum of sinusoids*. Fourier’s theorem says that any time-limited signal may be expressed as a sum of sinusoids:

$$x(t) = \sum_k A_k \cos(2\pi f_k t - \phi_k) \quad (2)$$

The number of terms in the sum is possibly infinite, but a nice property is that one can always select a finite subset (a “model of the model”) that fits the signal as closely as one wishes. The parameters are the set (f_k, A_k, ϕ_k) . The appeal of this description is that the effect of passing the stimulus through a linear time-invariant system may be predicted from its effect on each sinusoid in the sum. It thus combines useful features of the previous two

descriptions, but adds a new difficulty: each of the frequencies (f_k) could plausibly map to pitch.

A special case is the *harmonic* complex, for which all (f_k) are integer multiples of a common frequency F_0 . Parameters then reduce to F_0 and (A_k, ϕ_k) . Fourier's theorem tells us that the description is now *equivalent* to that of a periodic signal. It fits exactly the same stimuli, and the theorem allows us to translate between parameters $x(t)$, $0 < t \leq T$ and (A_k, ϕ_k) . This description fits many pitch-evoking stimuli and is very commonly used.

A fourth description is sometimes useful. The *formant* is a special case of a sum-of-sinusoids in which amplitudes A_k are largest near some frequency f_{LOCUS} (Fig. 4E). Its relevance is that a stimulus that fits this model may have a pitch related to f_{LOCUS} , and if the signal is also periodic with period $T=1/F_0$, pitches related to F_0 and f_{LOCUS} may both be heard (some people tend to hear one more easily than the other).

These various parameterizations appear repeatedly within the history of pitch. None is “good” or “bad”: they are all tools. However, multiple stimulus parameterizations pose a problem, as parameters are the “physical” dimensions that *psychophysics* deals with.

2.5 What is pitch?

A third possible source of discord is the definition of pitch itself (Plack and Oxenham, Chapter 1). The American National Standard Institute defines pitch as *that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high* (ANSI 1973). It doesn't mention the physical characteristics of the sounds. The French standards organization adds that pitch is *associated with frequency* and is low or high *according to whether this frequency is smaller or greater* (AFNOR 1977). The former definition is psychological, the latter psychophysical.

Both definitions assume a single perceptual dimension. For *pure tones* this makes sense, as the relevant stimulus parameter (f) is one-dimensional. Other perceptual dimensions such as brightness might exist, but they necessarily co-vary with pitch (Plomp 1976). For other pitch-evoking stimuli the situation is more complex. Depending on the stimulus representation (see Section 2.4), there might be several frequency parameters. Extrapolating from the definitions, one cannot exclude the possibility of *multiple* pitch-like dimensions. Indeed, a stimulus that fits the “formant” signal model may evoke a pitch related to f_{LOCUS} instead of, or in addition to, the pitch related to F_0 . Listeners may attend more to one or the other, and the outcome of experiments may be task- and listener-dependent (Smoorenburg 1970). For such stimuli, pitch has at least two dimensions, as illustrated in Figure 5. The pitch related to F_0 is called *periodicity pitch*, and that related to f_{LOCUS} is called *spectral pitch*¹. A pure tone also fits the formant model, but its periodicity and spectral pitches are not distinct (diagonal in Fig. 5). For other formant-like sounds they are distinct. As illustrated in Figure 5, periodicity pitch exists only within a limited region

¹ The term *spectral pitch* is used by Terhardt (1974) to refer to a pitch related to a resolved partial (Sec. 4.1, 7.2). We call that pitch a *partial pitch*.

of the parameter space. Spectral pitch is sometimes said to be mediated by place cues, and periodicity pitch by temporal cues (see below). However spectrum and time are closely linked, so it is wise to reserve judgment on this point.

Periodicity pitch varies according to a linear stimulus dimension (ordinate in Fig 5) but it has been proposed that the *perceptual structure* of periodicity pitch is helical, with pitches distributed circularly according to *chroma* and linearly according to *tone height*. Chroma accounts for the similarity (and ease of confusion) of tones separated by an octave, and tone height for the difference between the same chroma at different octaves (Bigand and Tillmann, Chapter 9). Tone height is sometimes assumed to depend on f_{LOCUS} . However, we saw that f_{LOCUS} is a distinct stimulus dimension (abscissa in Fig 5) and correlate of a perceptual quantity that we called *spectral pitch*, probably related to the dimension of brightness in timbre. Tone height and spectral pitch can be manipulated independently (Warren et al. 2003).

The pitch attribute is thus more complex than suggested by the standards, and further complexities arise as one investigates *intonation* in speech, or *interval*, *melody* and *harmony* in music (see Bigand and Tillmann, Chapter 9). We may usefully speak of models of the pitch attribute of varying complexity. The rest of this chapter assumes the simplest model: a one-dimensional attribute related to stimulus period.

3 Early Roots of Place Theory

Pythagoras (6th century BC) is credited for relating *musical intervals* to ratios of string length on a monochord (Hunt 1992). The monochord is a device comprising a board with two bridges between which a string is stretched (Fig. 6). A third and movable bridge divides the string in two parts with equal tension but free to vibrate separately. Consonant intervals of unison, octave, fifth and fourth arise for length ratios of 1:1, 1:2, 2:3, 3:4, respectively. This is an early example of *psychophysics*, in that a perceptual property (musical interval) is related to a ratio of physical quantities. It is also an early example of a model.

Aristoxenos (4th century BC) gives a clear, authoritative description of both interval and *pitch* (Macran 1902). A definition of a musical *note* that parallels our modern definition of pitch (ANSI 1973) was given by the Arab music theorist Safi al-Din (13th century): “a sound for which one can measure the excess of gravity or acuity with respect to another sound” (Hunt 1992). The qualitative dependency of pitch on *frequency* of vibration was understood by the Greeks (Lindsay 1966) but the quantitative relation was established much later by Marin Mersenne (1636) and Galileo Galilei (1638). Mersenne proceeded in two steps. First he confirmed experimentally the laws of strings, according to which frequency varies inversely with the length of a string, proportionally to the root of its tension, and inversely with the square root of its weight per unit length. This done, he stretched strings long enough to count the vibrations and, halving their lengths repeatedly, he derived the frequencies of every note of the scale.

Du Verney (1693) offered the first *resonance theory* of pitch perception (although the idea of resonance within the ear has earlier roots):

...[the spiral lamina,] being wider at the start of the first turn than the end of the last ... the wider parts can be caused to vibrate while the others do not ... they are capable of slower vibrations and consequently respond to deeper tones, whereas if the narrower parts are hit, their vibrations are faster and consequently respond to sharper tones...

Du Verney thought that the bony spiral lamina, wide at the base and narrow at the apex, served as a resonator. Note the concept of *selective response*.

...in the same way as the wider parts of a steel spring vibrate slowly and respond to low tones, and the narrower parts make more frequent and faster vibrations and respond to sharp tones...

Du Verney used a *technological metaphor* to convince himself, and others, that his ideas were reasonable.

...according to the various motions of the spiral lamina, the spirits of the nerve which impregnate its substance [that of the lamina] receive different impressions that represent within the brain the various aspects of tones

Thus was born the concept of *tonotopic projection* to the brain. This short paragraph condenses many of the concepts behind place models of pitch. The progress of anatomical knowledge up to (and beyond) Du Verney is recounted by von Békésy and Rosenblith (1948).

Mersenne was puzzled to hear, within the sound of a string or of a voice, pitches corresponding to the first five harmonics. He couldn't understand how a string vibrating at its fundamental could at the same time vibrate at several times that rate. He did however observe that a string could vibrate sympathetically to a string tuned to a multiple of its frequency, implying that it could also vibrate at that higher frequency.

Sauveur (1701) observed that a string could indeed vibrate *simultaneously* at several harmonics (he coined the words *fundamental* and *harmonic*). The laws of strings were derived theoretically in the 18th century (in varying degrees of generality) by Taylor, Daniel Bernoulli, Lagrange, d'Alembert, and Euler (Lindsay 1966). A sophisticated theory to explain superimposed vibrations was built by Daniel Bernoulli, but Euler leap-frogged it by simply invoking the concept of *linearity*. Linearity implies the *principle of superposition*, and that is what Mersenne lacked to make sense of the several pitches he heard when he plucked a string².

Mersenne missed the fact that the vibration he saw could reflect a sum of vibrations, with periods at integer submultiples of the fundamental period. Any such sum has the same period as the fundamental, but not necessarily the same shape. Indeed, adding *sinusoidal* partials produces variegated shapes depending on their amplitudes and phases (A_k, ϕ_k). That *any* periodic

² Mersenne pestered Descartes with this question but was not satisfied with his answers. Descartes finally came up with a qualitative explanation based on the idea of superposition in 1634 (Tannery and de Waard, 1970). Superposition can be traced earlier to Leonardo da Vinci and Francis Bacon (Hunt 1992).

wave can be thus obtained, and with a *unique* set of (A_k, ϕ_k) , was proved by Fourier (1822). The property had been used earlier, as many problems are solved more easily for sinusoidal movement. For example, the first derivation of the speed of sound by Newton in 1687 assumed “pendular” motion of particles (Lindsay 1966). Euler’s principle of superposition generalizes such results to *any sum of sinusoids*, and Fourier’s theorem adds merely that this means *any waveform*. This result had a tremendous impact.

4 Helmholtz

The mapping between pitch and period established by Mersenne and Galileo leaves a question open. An infinite number of waves have the same period: do they *all* map to the same pitch? Fourier’s theorem brings an additional twist by showing that a wave can be decomposed into elementary sinusoids. Each has its own period so, if the theorem is invoked, the period-to-pitch mapping is no longer one-to-one.

“Vibration” was commonly understood as a regular series of excursions in one direction separated by excursions in the other, but some waves have exotic shapes with several such excursion pairs per period. Do they too map to the same pitch? Seebeck (1841, Boring 1942) found that stimuli with two or three irregularly-spaced pulses per period had a pitch that matched the period. Spacing them evenly made the pitch jump to the octave (or octave plus fifth for three pulses). In all cases the pitch was consistent with the stimulus period, regardless of shape.

Ohm (1843) objected. In his words, he had “always previously assumed that the components of a tone, whose frequency is said to be f , must retain the form $a \cdot \sin 2\pi f t$ ”. To rescue this assumption from the results of Seebeck and others, he formulated a law saying that a tone evokes a pitch corresponding to a frequency f if and only if it “carries in itself the form $a \cdot \sin 2\pi(f t + p)$ ”³. In other words, every sinusoidal partial evokes a pitch, and no pitch exists without a corresponding partial. In particular, periodicity pitch depends on the presence of a *fundamental partial* of non-zero amplitude. This is more restrictive than Seebeck’s condition that a stimulus merely be periodic.

Ohm’s law was attractive for two reasons. First, it drew on Fourier’s theorem, seemingly tapping its power for the benefit of hearing theory. Second, it explained the higher pitches reported by Mersenne. Paraphrasing the law, Helmholtz (1877) stated that the sensation evoked by a pure tone is “simple” in that it does not support the perception of such higher pitches. From this he concluded that the sensation evoked by a complex tone is *composed* of the sensations evoked by the pure tones it contains. A corollary

³ Presence of the “form” was ascertained by applying Fourier’s theorem to consecutive waveform segments of size $1/f$. Ohm required that p and the *sign* of a (but not its *magnitude*) be the same for each segment. He said: “the necessary impulses must follow each other in time intervals of the length $1/f$ ”. This could imply that he was referring to the pitch of the *fundamental* partial and not (as was later assumed) other partials. Authors quoting Ohm usually reformulate his law, not always with equal results.

is that sensation cannot depend on the relative *phases* of partials. This he verified experimentally for the first eight partials or so, while expressing some doubt about higher partials.

To summarize, the Ohm/Helmholtz psychoacoustic model of pitch refines the simpler law of Mersenne: (a) Among the many periodic vibrations with a given period, only those containing a nonzero fundamental partial evoke a pitch related to that period; (b) Other partials might also evoke additional pitches; (c) Relative partial amplitudes affect the quality (timbre) of the vibration, but not its pitch, as long as the amplitude of the fundamental is not zero; (d) Relative phases of partials (up to a certain rank) affect neither quality nor pitch.

The theory also included a physiological part. Sound is analyzed within the cochlea by the basilar membrane (BM) considered as a bank of radially taut *strings*, each loosely coupled to its neighbors. Resonant frequencies are distributed from high (base) to low (apex), and thus a sound undergoes a spectral analysis, each locus responding to partials that match its characteristic frequency. From constraints on time resolution (see Section 10.2) Helmholtz concluded that selectivity must be limited. Thus he viewed the cochlea as an *approximation* of the Fourier transformer needed by the psychoacoustic part of the model. Limited frequency resolution was actually welcome, as it helped him account for *roughness* and *consonance*, bringing together mathematics, physics, elementary sensation, harmony, and aesthetics into an elegant unitary theory.

Helmholtz linked the decomposition of the stimulus to a decomposition of sensation, extending the principle of superposition to the sensory domain, and to the psychoacoustic mapping between stimulus and sensation. In doing so, he assumed compositional properties of sensation and perception for which his arguments were eloquent but not quite watertight. True, his theory implies the phase-insensitivity that he observed, but to be conclusive the argument should show that it is the *only* theory that can do so. It explains Mersenne's upper pitches (each suggestive of an elementary sensation) but begs the question of why they are so rarely perceived. More seriously, it predicts something already known to be false at the time. The pitch of a periodic vibration does *not* depend on the physical presence of a fundamental partial. That was known from Seebeck's experiments, from earlier observations on beats (see Section 10.1), and from observations of contemporaries of Helmholtz cited by his translator Ellis (traduttore traditore!).

Helmholtz was aware of the problem, but argued that theory and observation could be reconciled by supposing *nonlinear interaction* within the ear (or within other people's sound apparatus). Distortion within the ear was accepted as an adequate explanation by later authors (von Békésy, Fletcher) but, as Wever (1949) remarks, it does not save the psychoacoustic law. The *coup de grâce* was given by Schouten (1938) who showed that complete cancellation of the fundamental partial within the ear leaves the pitch unchanged. Licklider confirmed that that partial was dispensable by masking it, rather than removing it. The weight of evidence against the theory as the sole explanation for pitch perception is today overwhelming (Plack and Oxenham, Chapter 2).

Nevertheless the place theory of Helmholtz is still used in at least four areas: (1) to explain pitch of *pure tones* (for which objections are weaker),

(2) to explain the extraction of frequencies of *partials* (required by pattern matching theories as explained below), (3) to explain *spectral pitch* (associated with a spectral locus of power concentration), (4) in textbook accounts (as a result of which the “missing fundamental” is rediscovered by each new generation). Place theory is simmering on a back burner in many of our minds.

It is tempting to try to “fix” Helmholtz’s theory retrospectively. The Fourier transform represents the stimulus according to the “sum of sinusoids” description (Section 2.4), but among the parameters f_k of that description none is obviously related to pitch. We’d need rather an operation that fits the “periodic” or “harmonic complex” signal description. Interestingly, a string does just that. As Helmholtz (1857) himself explained, a string tuned to F_0 responds to *all harmonics* kF_0 . By superposition it responds to every sum of harmonics and therefore to *any periodic sound of period $1/F_0$* (Fig. 7). Helmholtz used the metaphor of a piano with dampers removed (or a harpsichord as suggested by Le Cat 1758) to explain how the ear works, and his physiological model invoked a bank of “strings” within the cochlea. However he preferred to treat cochlear resonators as spherical resonators (which respond each essentially to a single sinusoidal component). Had he treated them as strings there would have been no need for the later introduction of pattern matching models. The “missing fundamental” would never be missed. Period-tuned cochlear resonators were actually suggested by Weinland in 1894 (Bonnier 1901). Of course, such a “fixed” theory holds only as long as one sees the ear as a bank of strings.

Helmholtz invoked for his theory the principle of “specific energies” of his teacher Johannes Müller, according to which each nerve represents a different *quality* (in this case a different pitch). To illustrate it, he drew upon a technological metaphor: the telegraph, in which each wire transmits a *single* message. Alexander Graham Bell, who was trying to develop a *multiplexing* telegraph to overcome precisely that limitation (Hounshell 1976), read Helmholtz and, getting sidetracked, invented the *telephone* that later inspired to Rutherford (1886) a theory that he opposed to that of Helmholtz...

The next section shows how the missing fundamental problem was addressed by modern pitch theory.

5 Pattern Matching

The partials of a periodic sound form a *pattern* of frequencies. We are good at recognizing patterns. If they are incomplete, we tend to perceptually “reconstruct” what is missing. A pattern matching model assumes that pitch emerges in this way. Two parts are involved: one produces the *pattern* and the other looks for a match within a set of *templates*. Templates are indexed by pitch, and the one that gives the best match indicates the pitch. The best-known theories are those of Goldstein (1973), Wightman (1973) and Terhardt (1974).

5.1 Goldstein, Wightman and Terhardt

For Goldstein (1973) the pattern consists of a series f_k of partial frequency estimates. Each estimate is degraded by a *noise*, modeled as a Gaussian process with mean f_k , and a variance that is function of f_k . Only *resolved* partials (those that differ from their neighbors by more than a resolution limit) are included, and neither amplitudes nor phases are represented. A “central processor” attempts to account for the series as *consecutive* multiples of a common fundamental (the consecutiveness constraint was later lifted by Gerson and Goldstein 1978). Goldstein suggested that the f_k were possibly, but not necessarily, produced in the cochlea according to a place model such as that of Helmholtz. Srulovicz and Goldstein (1983) showed that they can also be derived from temporal patterns of auditory nerve firing. Interestingly, Goldstein mentions that estimates do not need to be ordered, and thus *tonotopy* need not be preserved once the estimates are known.

For Wightman (1973) the pattern consists of a tonotopic “peripheral activity pattern” produced by the cochlea, similar to a smeared power spectrum. This pattern undergoes Fourier transformation within the auditory system to produce a second pattern similar to the autocorrelation function (the Fourier transform of the power spectrum). Pitch is derived from a peak in this second pattern.

For Terhardt (1974) the pattern consists of a “specific loudness pattern” originating in the cochlea, from which is derived a pattern of *partial pitches*, analogous to the elementary sensations posited by Helmholtz⁴. From the pattern of partial pitches is derived a “gestalt” *virtual pitch* (periodicity pitch) via a pattern matching mechanism. Perception operates in either of two modes, analytic or synthetic, according to whether the listener accesses partial or virtual pitch, respectively. Analytic mode adheres strictly to Ohm’s law: there is a one-to-one mapping between resolved partials and partial pitches. Partial pitch is presumably innate, whereas virtual pitch is *learned* by exposure to speech. Listening is normally synthetic (virtual pitch).

The three models are formally similar despite differences in detail (de Boer 1977). The idea of pattern matching has roots deeper in time. It is implicit in Helmholtz’s notion of “unconscious inference” (Helmholtz 1857; Turner 1977). According to the “multicue mediation theory” of Thurlow (1963), listeners use their voice as a template (pitch then equates to the motor command that best matches an incoming sound). De Boer (1956) describes pattern matching in his thesis. Finally, pattern matching fits the behavior of the oldest metaphor in pitch theory: the *string* (compare figures 1F and 7C).

⁴ Terhardt called them *spectral pitches*, a term we reserve to designate the pitch associated with a concentration of power along the spectral axis.

5.2 Relation to signal processing methods

Signal processing methods are a source of inspiration for auditory models. Pattern matching is used in several methods of speech F0 estimation (Hess 1983). The “period histogram” of Schroeder (1968) accumulates all possible subharmonics of each partial (as in Terhardt’s model), while the “harmonic sieve” model of Duifhuis et al. (1982) tries to find a sieve that best fits the spectrum (as in Goldstein’s model). Subharmonic summation (Hermes 1988) or SPINET (Cohen et al. 1995) work similarly, and there are many variants. One is to cross correlate the spectrum with a set of “combs”, each having “teeth” at multiples of a fundamental. Rather than combs with sharp teeth, other regular patterns may be used, for example sinusoids. Cross correlating with sinusoids implements the Fourier transform. The Fourier transform applied to a *power* spectrum gives the *autocorrelation* function (as in Wightman’s model). Applied to a *logarithmic* spectrum it gives the *cepstrum*, commonly used in speech processing (Noll, 1967). There is a close connection between pattern matching and these representations.

Cochlear filters are narrow at low frequencies and wide at high. Wightman took this into account by applying *nonuniform smoothing* to the spectrum. Smoother parts of the spectrum require a smaller density of channels, so the spectrum can be *resampled* non-uniformly. This is the idea behind the so-called “mel spectrum” and MFCC (mel-frequency cepstrum coefficients), popular in speech processing. These are analogous to the logarithmic spectra of Versnel and Shamma (1998). Non-uniform sampling causes the regular structure of a harmonic spectrum to be lost and thus is not very useful for pitch.

A final point is worth mentioning. We usually think of frequency as positive, but the mathematical operation that relates power spectrum to ACF (or log power spectrum to cepstrum) applies to spectra that extend over *positive and negative* frequencies. The negative part is obtained by reflecting the positive part over 0 Hz. Spectra are then symmetric and their Fourier spectra contain only *cosines*, which always have a peak at 0 Hz. A similar constraint in a harmonic comb model is to anchor a tooth at 0 Hz, and it turns out that this is important to account for the pitch of inharmonic complexes. We know that the pitch of a set of harmonics spaced by Δf shifts if they are all mistuned by an equal amount. Pitch varies in proportion to the central partial in a first approximation (so-called “first effect”). In a second approximation it follows a lower frequency, sometimes even lower than the lowest partial (“second effect”). Without the constraint, the best fitting comb has teeth spaced by Δf regardless of the mistuning, implying no pitch shift. This led Jenkins (1961) and Schouten et al. (1962) to rule out spectrum-based pattern matching models. With the constraint, the best fit is a slightly stretched comb and this allows the pitch shift to be accounted for.

5.3 The learning hypothesis

Pattern matching requires a set of harmonic templates. Terhardt (1978, 1979) suggested that they are *learned* through exposure to harmonic-rich sounds such as speech. To explain how, Roederer (1975) proposed that

spectral patterns from the cochlea are fed to a neural net. At the intersection between a channel tuned to the fundamental, and channels tuned to its harmonics, synapses are reinforced through Hebbian learning (Hebb 1949). Licklider (1959) had earlier invoked Hebbian learning to link together the period and spectrum axes of his “duplex” model. Learning was also suggested by de Boer (1956) and Thurlow (1963), and is implicit in Helmholtz’s dogma of unconscious inference (Warren and Warren, 1968).

The harmonic patterns needed for learning may be found in the harmonics of a complex tone such as speech. They exist also in the series of its “superperiods” (subharmonics). This suggests that one could do away with Terhardt’s requirement of early exposure to *harmonically rich* sounds, since a pure tone too has superperiods. Readers in need of a metaphor to accept this idea should consider Figure 7. Panel A illustrates the template (made irregular by the logarithmic axis) formed by the partials of a harmonic complex tone. Panel B illustrates a similar template formed by the superperiods of a pure tone. Harmonically rich stimuli are not essential for the learning hypothesis.

Shamma and Klein (2000) went a step further and showed that template learning *does not require exposure to periodic sounds*, whether pure or complex. Their model is a significant step in the development of pattern matching models. Ingredients are: (1) an input pattern of phase locked activity, spectrally sharp or sharpened by some neural mechanism based on synchrony, (2) a nonlinear transformation such as half-wave rectification, and (3) a matrix sensitive to spike coincidence between each channel and every other channel. In response to noise or random clicks, each channel rings at its characteristic frequency (CF). The nonlinearity creates a series of harmonics of the ringing that correlate with channels tuned to those harmonics, resulting in Hebbian reinforcement (reinforcement of a synapse by correlated activity of pre- and postsynaptic neurons) at the intersection between channels. The loci of reinforcement form diagonals across the matrix, and together these diagonals form a harmonic template. Shamma and Klein made a fourth assumption: (4) sharp phase transitions along the BM near the locus tuned to each frequency. This seems to be needed only to ensure that learning occurs also with nonrandom sounds. Shamma and Klein note that the resulting “template” is not a perfect comb. Instead it resembles somewhat Figure 7C.

Exposure to speech or other periodic sounds is thus unnecessary to learn a template. One can go a step further and ask whether *learning* itself is necessary. We noted that the string responds equally to its fundamental and to all harmonics, and thus behaves as a pattern-matcher. That behavior was certainly not learned. We’ll see later that other mechanisms (such as autocorrelation) have similar properties. Taking yet another step, we note that the string operates directly on the waveform and not on a spectral pattern. So it would seem that *pattern matching* itself is unnecessary, at least in terms of function. It may nevertheless be the way the auditory system works.

6 Pure Tones and Patterns

Pattern matching allows the response to a complex tone to be treated (in the pattern stage) as the sum of sensory responses to *pure* tones. This is fortunate, as much effort has gone into the psychophysics of pure tones. Pattern matching is not particular about how the pattern is obtained, whether by a cochlear place mechanism or centrally from temporal fine structure. It *is* particular about its quality: the number and accuracy of partial frequency estimates it can operate on.

6.1 Sharpening

Helmholtz's estimate of cochlear resolution (about one semitone) implied that the response to a pure tone is spread over several sensory cells. Strict application of Müller's principle would predict a "cluster" of pitches (one per cell) rather than one. Gray (1900) answered this objection by proposing that a single pitch arises at the place of *maximum stimulation*. Besides reducing the sensation to one pitch, the principle allows accuracy to be independent of peak width: narrow or wide, its locus can be determined exactly (in the absence of noise), for example by competition within a "winner-take-all" neural network (Haykin 1999). However, if noise is present before the peak is selected, accuracy obviously *does* depend on peak width. Furthermore, if two tones are present at the same time their patterns may interfere. One peak may vanish, being reduced to a "hump" on the flank of the other, or its locus may be shifted as a result of riding on the slope of the other. These problems are more severe if peaks are wide, so sharpness of the initial tonotopic pattern is important.

Recordings from the auditory nerve or the cochlea (Ruggero 1992) show tuning to be narrower than the wide patterns observed by von Békésy, which worried early theorists. Narrow cochlear tuning is explained by *active* mechanisms that produce negative damping. The occasional observation of spontaneous oto-acoustic emissions suggests that tuning might in some cases be *arbitrarily* narrow (e.g. Camalet et al. 2000), such as to sometimes cross into instability. However, these active mechanisms being nonlinear, one cannot extrapolate tuning observed with a pure tone to a combination of partials. Sharp tuning goes together with a boost of gain at the resonant frequency. The phenomenon of *suppression*, by which the response to a pure tone is suppressed by a neighboring tone, suggests that the boost (and thus the tuning) is lost if the tone is not alone. If hyper-sharp tuning requires that there be only one partial, it is of little use to sharpen the responses to partials a complex tone. Similar remarks apply to measures of selectivity in conditions that minimize suppression (Shera et al. 2002).

Indeed, at medium-to-high amplitudes, profiles of auditory-nerve fiber response to complex tones lack evidence of harmonic structure in cats (Sachs and Young 1979). However, profiles are better represented in the subpopulation of *low-spontaneous rate* fibers (see Winter, Chapter 4). Furthermore, Delgutte (1996; Cedolin and Delgutte 2004) argues that filters might be narrower in humans. Psychophysical forward masking patterns indeed show some harmonic structure (Plomp, 1964). Schofner (Chapter 3)

discusses the issues that arise when comparing measures between humans and animal models.

A “second filter” after the BM was a popular hypothesis before modern measurements showed sharply tuned mechanical responses. A variety of mechanisms have been put forward: mechanical sharpening (e.g. sharp tuning of the cilia or tectorial membrane, or differential tuning between tectorial and basilar membranes), sharpening in the transduction process, or sharpening by neural interaction. Huggins and Licklider (1951) list a number of schemes. They are of interest in that the question of a sharper-than-observed tuning arises repeatedly (e.g. in the template-learning model of Shamma and Klein). Some of these mechanisms might be of use also to sharpen ACF peaks (see Section 9).

Sharpening can operate on the cross-frequency profile of amplitudes, on the pattern of phases, or on both. A simple sharpening operation is an *expansive nonlinearity*, e.g. implemented by coincidence of several neural inputs from the same point of the cochlea (on the assumption that probability of coincidence is the product of input firing probabilities). Another is *spatial differentiation* (more generally *spatial filtering*) of the amplitude pattern, e.g. by summation of excitatory and inhibitory inputs of different tuning. Sharp patterns can also be obtained using phase, for example by transduction of the differential motion of neighboring parts within the cochlea, or by neural interaction between phase-locked responses. The *Lateral Inhibitory Network* (LIN) of Shamma (1985) uses both amplitude and phase. Partials of low frequency (<2 kHz) are emphasized by phase transitions along the BM, and those of high frequency by spatial differentiation of the amplitude pattern. The hypothesis is made attractive by a recent model that uses a different form of phase-dependent interaction to account for loudness (Carney et al. 2002). In the *Average Localized Synchrony Rate* (ALSR) or *Measure* (ALSM) of Young and Sachs (1979) and Delgutte (1984), a narrowband filter tuned to the characteristic frequency of each fiber measures synchrony to that frequency. The result is a pattern where partials stand out clearly. The *matched filters* of Srulovicz and Goldstein (1983) operate similarly. These are examples from a range of ingenious schemes to sharpen peaks of response patterns.

Alternatives to peak sharpening are to assume that a pure tone is coded by the *edge* of a tonotopic excitation pattern (Zwicker 1970), or that that partials of a complex tone are coded using the location of *gaps* between fibers responding to neighboring partials (Whitfield 1970).

6.2 Labeling by synchrony

In place theory, the frequency of a partial is signaled by its position along the tonotopic axis. LIN and ALSR use phase locking merely to measure the position more finely. Troland (1930) argued that position is unreliable, and that it is better to label a channel by phase locking at the partial’s frequency, an idea already put forward by Hensen in 1863 (Boring, 1942). Peripheral filtering would serve merely to *resolve* partials, so that frequency can be measured and each channel labeled clearly. A nice feature of this idea is that all channels that respond to a partial contribute to characterize it (rather than just some predetermined set). Tonotopy is not required, as noted by

Goldstein (1973), but the “labels” still need to be decoded to whatever dimension underlies the harmonic templates to which the pattern is to be matched.

A possible decoder is some form of central filterbank. In the *dominant component* scheme of Delgutte (1984), each channel of the neural response is analyzed over a central filterbank, and the resulting spectral profiles combined over channels. A related principle underlies the *modulation filterbank* (e.g. Dau et al. 1996), discussed later on in the context of temporal models. An objection is that the hypothesis requires several filterbanks, one peripheral and one (or more) central. What is gained over a single filterbank? A possible answer is that transduction nonlinearity recreates the “missing fundamental” component for stimuli that lack one. However, one wonders why this is better (in terms of function) than Helmholtz's assumption of a mechanical nonlinearity preceding the cochlear filter.

From this discussion, it appears that the frequency of a pure tone (or partial) might be derived from either place *or* time cues. To decide between them, Siebert (1968, 1970) used a simple model assuming triangle-shaped filters, nerve spike production according to a Poisson process, and optimal processing of spike trains. Calculations showed that place alone was sufficient to account for human performance. Time allowed *better* performance, and Siebert tentatively concluded that the auditory system does *not* use time. However, a reasonable form of suboptimal processing (filters matched to interspike interval histograms) gives predictions closer to behavior (Goldstein and Srulovicz 1977). In a recent computational implementation of Siebert's approach, Heinz et al. (2001) found, as Siebert did, that place cues are sufficient and time cues more than sufficient to predict behavioral thresholds. However, predicted and observed thresholds were parallel for time but not for place (Fig. 8), and Heinz et al. tentatively concluded that the auditory system *does* use time. Interestingly, despite the severe degradation of time cues beyond 5 kHz (Johnson, 1980), useful information could be exploited up to 10 kHz at least, and predicted and observed thresholds remained parallel up to the highest frequency measured, 8 kHz. Extrapolating from these results, the entire partial frequency pattern of a complex might be derived from temporal information.

To summarize, a wide range of schemes produce spectral patterns adequate for pattern matching. Some rely entirely on BM selectivity, while others ignore it. No wonder it is hard to draw the line between “place” and “time” theories! We now move on to the second major approach to pitch: time.

7 Early Roots of Time Theory

Boethius (Bower, 1989) quotes the Greek mathematician Nicomachus (2nd century), of the Pythagorean school:

...it is not, he says, only one pulsation which emits a simple measure of sound; rather a string, struck only one time, makes many sounds, striking the air again and again. But since its velocity of percussion is such that

one sound encompasses the other, no interval of silence is perceived, and it comes to the ears as if one pitch.

We note the idea, rooted in the Pythagorean obsession with number, that a sound is *composed* of several elementary sounds. Ohm and Helmholtz thought the same, but their “elements” were sinusoids. The notion of overlap between successive elementary sounds prefigures the concept of impulse response and convolution. Boethius continues:

If, therefore, the percussions of the low sounds are commensurable with the percussions of the high sounds, as in the ratios which we discussed above, then there is no doubt that this very commensuration blends together and makes one consonance of pitches.

Ratios of pulse counts play here the role later played by ratios of frequency in spectral theories. The origin of the relation between pitch and pulse counts is unclear, partly because the vocabulary of early thinkers (or translators, or secondary sources) did not clearly distinguish between rate of vibration, speed of propagation, amplitude of vibration, and the speed (or rate) at which one object struck another to make sound (Hunt, 1992). Mersenne and Descartes clarified the roles of vibration rate and speed of propagation, finding that the former determines, while the latter is independent of, pitch. It is interesting to observe Mersenne (1636) struggle to explain this distinction using the same word (“fast”) for both.

The rate-pitch relation being established, a pitch perception model must explain how rate is measured within the listener. Mersenne and Galileo both measured vibrations by *counting* them, but they met with two practical difficulties: the lack of accurate time standards (Mersenne initially used his heartbeat, and in another context the time needed to say “Benedicam dominum”) and the impossibility of counting fast enough the vibrations that evoke pitch. These difficulties can be circumvented by the use of calibrated *resonators* that we mentioned earlier on, with their own set of problems due to instability of tuning. Here is possibly the fundamental contrast between time and place: is it more reasonable to assume that the ear counts vibrations, or contains calibrated resonators?

This question overlaps that of *where* measurement occurs within the listener, as the ear seems devoid of counters but possibly equipped with resonators. Counting, if it occurs, occurs in the brain. The disagreement about where things happen can be traced back to Anaxagoras (5th century BC) for whom hearing depended simply on *penetration of sound to the brain*, and Alcmaeon of Crotona (5th century BC) for whom *hearing is by means of the ears, because within them is an empty space, and this empty space resounds* (Hunt, 1992). The latter sentence seems to “explain” more than the first: the question is also how much “explanation” we expect of a model.

The doctrine of internal air, “aer internus”, had a deep influence up to the 18th century, when it merged gradually into the concepts of resonance and “animal spirits” (nerve activity) that eventually culminated in Helmholtz’s theory. The *telephone theory* of Rutherford (1886) was possibly a reaction against the authority of that theory (and its network of mutually supporting assumptions, some untenable such as Ohm’s law). In the minimalist spirit of Anaxagoras, Rutherford proposed that the ear merely transmits vibrations to the brain like a telephone receiver. The contrast

between his modest theory (2 pages), and the monumental opus of Helmholtz that it opposed, is striking. To its credit, Rutherford's two-page theory was parsimonious, to its discredit it just shoved the problem one stage up.

An objection to the telephone theory was that nerves do not fire fast enough to follow the higher pitches. Rutherford observed transmission in a frog motor nerve up to relatively high rates (352 times per second). He did not doubt that the auditory nerve might respond faster. The need for high rates was circumvented by the *volley theory* of Wever and Bray (1930), according to which several fibers fire in turn such as to produce, together, a rate several times that of each fiber. Later measurements within fibers of the auditory nerve proved the theory wrong, in that firing is *stochastic* rather than regular (Galambos and Davis 1943, Tasaki 1954), but right in that fibers can indeed represent frequencies higher than their discharge rate. Steady-state discharge rates in the auditory nerve are limited to about 300 spikes per second, but the pattern of *instantaneous probability* can carry time structure that can be measured up to 3-5 kHz in the cat (Johnson, 1980). The limit is lower in the guinea pig, higher in the barn owl (9 kHz, Köppl 1997), and unknown in humans.

A pure tone produces a BM motion waveform with a single peak per period, a simple pattern to which to apply the volley principle (in its probabilistic form). However, Section 2.2 showed the limits of peak-based schemes for more complex stimuli. The idea that pitch follows their temporal *envelope* (Fig. 2E), via some demodulation mechanism, was proposed by Jenkins (1961) among others. It was ruled out by the experiments of de Boer (1956) and Schouten et al. (1962) in which the partials of a modulated-carrier stimulus were mistuned by equal amounts, producing a pitch shift (as mentioned earlier). The envelope stays the same, and this rules out not only the envelope as a cue to pitch (except for stimuli with unresolved partials, Plack and Oxenham, Chapter 2), but also *inter-partial spacing* or *difference tones*. De Boer (1956) suggested that the effective cue is the spacing between *peaks of the waveform fine structure* closest to peaks of the envelope, and Schouten et al. (1962) pointed out that zero-crossings or other "landmarks" would work as well.

The waveform fine structure theory was criticized on several accounts, the most serious being that it predicts greater *phase-sensitivity* than is observed (Wightman 1973). The solution to this problem was brought by the autocorrelation (AC) model. Before moving on to that, I'll describe an influential but confusing concept: the residue.

8 Schouten and the Residue

In the tradition of Boethius, Ohm and Helmholtz thought that a stimulus is composed of elements. They believed that the sensation it evokes is composed of elementary sensations, and that a one-to-one mapping exists between stimulus elements and sensory elements. The *fundamental* partial mapped to periodicity pitch, and higher partials to higher pitches that some people sometimes hear. Schouten (1940a) agreed to all these points but one:

periodicity pitch should be mapped to a different part of the stimulus, called the *residue*. He reformulated Ohm's law accordingly.

Schouten (1938) had confirmed Seebeck's observation that the fundamental partial is dispensable. Manipulating individual partials of a complex with his optical siren, he trained his ear to hear them out (as Helmholtz had done before using resonators). He noted that the fundamental partial too could be heard out. The stimulus then seemed to contain *two* components with the same pitch. Introspection told him that their qualities were identical, respectively, to those of a pure tone at the fundamental and of a complex tone without a fundamental. The latter carried a salient low pitch. From his new law, Schouten reasoned that the missing-fundamental complex must either *contain* or *be* the residue. He noticed that removing additional low partials left the sharp quality intact. Low partials can be heard out, and each carries its own pitch, so Schouten reasoned that they are *not* part of the residue, whereas removing higher partials reduces the sharp quality that Schouten associated with the residue. Thus he concluded that the residue must consist of these *higher partials perceived collectively*. It somehow escaped him that periodicity pitch remains salient when the higher partials are absent.

Exclusion of resolvable partials from the residue put Schouten's theory into trouble when it was found that they actually dominate periodicity pitch (Ritsma 1967; Plomp 1967a). Strangely enough, Schouten gave as an example a bell with characteristic tones fitting the highly resolvable series 2:3:4 (Schouten 1940b,c). Its strike note fits the missing fundamental, yet all of its partials are resolvable. De Boer (1976) amended Schouten's definition of residue to include all partials, which is tantamount to saying that the residue *is* the sound, rather than part of it. Schouten (1940a) had mentioned that possibility, but he rejected it as causing "a great many difficulties" without further explanation. Possibly, he believed that interaction *in the cochlea* between partials, strong if they are unresolved, is necessary to measure the period. The AC model (next Section) shows that it is not.

The residue concept is no longer useful and the term "residue pitch" should be avoided. The concept survives in discussions of stimuli with "unresolved" components, commonly used in pitch experiments to ensure a complete absence of spectral cues (Section 10.4). Their pitch is relatively weak, which confirms that the residue (in Schouten's narrow definition) is *not* a major determinant of the periodicity pitch of most stimuli.

9 Autocorrelation

Autocorrelation, like pattern matching, is the basis of several modern models of pitch perception. It is easiest to understand as a measure of *self-similarity*.

9.1 Self-similarity

A simple way to detect periodicity is to take the *squared difference* of pairs of samples $x(t)$, $x(t-\tau)$ and smooth this measure over time to obtain a

temporally stable measure of self-similarity:

$$d(\tau) = (1/2) \int [x(t) - x(t - \tau)]^2 dt \quad (3)$$

This is simply half the Euclidean distance of the signal from its time-shifted self. If the signal is periodic, the distance should be zero for a shift of one period. A relation with the *autocorrelation function* or ACF (Eq. 1) may be found by expanding the squared difference in Equation 3. This gives the relation:

$$d(\tau) = e - r(\tau) \quad (4)$$

where e represents signal energy and r the autocorrelation function. Thus, $r(\tau)$ increases where $d(\tau)$ decreases, and peaks of one match the valleys of the other. Peaks of the ACF (or valleys of the difference function) can be used as cues to measure the period. The variable τ is referred to as the *lag* or *delay*. The difference function d and ACF r are illustrated in Figs. 9B and C, for the stimulus illustrated in A.

9.2 Licklider

Licklider (1951, 1959) proposed that autocorrelation could explain pitch. Processing occurs within the auditory nervous system, after cochlear filtering and hair-cell transduction. It can be modeled as operating on the half-wave rectified basilar-membrane displacement. The result is a 2-dimensional pattern with dimensions characteristic frequency (CF) and lag (Figure 9D). If the stimulus is periodic, a ridge spans the CF dimension at a lag equal to the period. Pitch may be derived from the position of this ridge, but Licklider didn't actually give a procedure for doing so.

Meddis and Hewitt (1991a,b) repaired this oversight by simply summing the 2D pattern across frequency to produce a "summary ACF" (SACF) from which the period may be derived (Fig. 9E). They also included relatively realistic filter and transduction models in their implementation, and showed that the model could account for many important pitch phenomena. "AC model" in this chapter designates a class of models in the spirit of Licklider, and Meddis and Hewitt. The SACF is visually similar to the ACF of the stimulus waveform (Fig. 9C), which has been used as a simpler predictive model (de Boer 1956; Yost 1996).

Licklider imagined an elementary network made of neural *delay* elements and *coincidence counters*. A coincidence counter is a neuron with two excitatory synapses, that fires if spikes arrive within some short time window at both synapses. Its firing probability is the *product* of firing probabilities at its inputs, and this implements the product within the formula of the ACF. Licklider supposed that this elementary network was reproduced within each channel from the periphery. It is similar to the network proposed by Jeffress (1948) to explain localization on the basis of interaural time differences.

Figure 9 illustrates the fact that the AC model works well with stimuli with resolved partials. Individual channels do not show fundamental periodicity (D), and yet the pattern that they form collectively is periodic at the fundamental. The period is obvious in the SACF (E). Thus, it is not necessary that partials interact on the BM to derive the period, a fact that escaped Schouten (and perhaps even Licklider himself). In the absence of

half-wave rectification, the SACF would be *equal* to the ACF of the waveform (granted mild assumptions on the filterbank). Differences between ACF and SACF (Figs. 9C and E) reflect the effects of nonlinear transduction and amplitude normalization.

9.3 Phase Sensitivity

Excessive phase sensitivity was a major argument against temporal models (Wightman, 1973). Phase refers to the parameter ϕ of the sinusoid model, or ϕ_k of the sum-of-sinusoids model (Section 2.4). Changing ϕ is equivalent to shifting the time origin, which doesn't affect the sound. Likewise, a change of ϕ_k by an amount *proportional to the frequency* f_k is equivalent to shifting the time origin. For a steady-state stimulus, manipulations that obey this property are imperceptible. This is de Boer's (1976) phase rule. However, phase changes that do *not* obey de Boer's rule may also be imperceptible. This is Helmholtz's rule, corollary of Ohm's law (if perception is composed from sensations, each related to a partial, there is no place for interaction *between* partials, and thus no place for phase effects). Helmholtz limited its validity to resolved partials. For stimuli with non-resolved partials, phase changes may be audible and may affect pitch, primarily the distribution of matches for ambiguous stimuli (such as illustrated in Fig. 2 E). For example, a complex with unresolved partials in alternating sine/cosine (ALT) phase may have a pitch at the *octave* of its true period (Plack and Oxenham, Chapter 2).

How does the AC model fare in this respect? Autocorrelation discards phase, but it is preceded by transduction nonlinearities that *are* phase-sensitive, themselves preceded by narrow-band filters that tend on the contrary to *limit* phase-sensitive interaction. These filters are however non-linear, and they produce *combination tones* (see Section 10.1) that behave as extra partials with phase-dependent amplitudes.

Concretely: ACFs from channels that respond to *one* partial do not depend on phase (unless that partial is a phase-dependent combination tone). Channels that respond to *two* partials are only slightly phase-dependent if the partials are of high rank. Channels responding to *three* harmonics or more are more strongly phase-dependent, but phase affects mainly the shape of the ACF and usually not the position of the period cue. Its *salience* may however change relative to competing cues at other lags. For example, within channels responding to several partials, the ACF is sensitive to the envelope of the waveform of their sum. For complexes in ALT phase (Plack and Oxenham, Chapter 2), the envelope period is half the fundamental period, which may explain why their pitch is at the octave.

Other forms of phase sensitivity, such as to time reversal, may be accounted for by invoking a particular implementation of the AC model (de Cheveigné 1998) or related models (Patterson 1994a,b, see Section 9.5). Pressnitzer et al. (2002, 2003) describe an interesting quasi-periodic stimulus for which both the pitch and the AC model period cue *are* phase-dependent. To summarize, the limited phase (in)sensitivity of the AC model accounts in large part for the limited phase (in)sensitivity of pitch (Meddis and Hewitt, 1992b). See also Carlyon and Shamma (2003).

9.4 Histograms

Licklider’s “neural autocorrelation” operation is equivalent to an *all-order interspike interval* (ISI) histogram, one of several formats used by physiologists to represent spike statistics of single-electrode recordings (Ruggero 1973; Evans 1986). Other common formats are *first-order ISI*, *peristimulus time* (PST), and *period* histograms. ISI histograms count intervals between spikes. First-order ISIs span consecutive spikes, and all-order ISIs span spikes both consecutive or not. The PST histogram counts spikes relative to the stimulus onset, and the period histogram counts them as a function of phase within the period.

Cariani and Delgutte (1996a,b) used all-order ISI histograms to quantify auditory nerve responses in the cat to a wide range of pitch-evoking stimuli. Results were consistent with the AC model. However, first-order ISI histograms are more common in the literature (e.g. Rose et al. 1967) and models similar to Licklider’s have been proposed that use them (Moore 1977; van Noorden 1982). In those models, a histogram is calculated for each peripheral channel, and histograms are then summed to produce a summary histogram. The “period mode” (first large mode at non-zero lag) of the summary histogram is the cue to pitch.

Recently there has been some debate as to whether first- or all-order statistics determine pitch (Kaernbach and Demany 1998; Pressnitzer et al. 2002, 2003). Without entering the debate, we note that all-order statistics may usefully be applied to the aggregate activity of a *population* of N fibers. There are several reasons why one should wish to do so. One is that refractory effects prevent single fiber ISIs from being shorter than about 0.7 ms, meaning that frequencies above 800 Hz don’t evoke a period mode in the histogram of a single fiber. Another is that aggregate statistics make more efficient use of available information, because the number of intervals increases with the *square* of N . Aggregate statistics may be simulated from a single-fiber recording by pooling post-onset spike times recorded to N presentations of the same stimulus. Intervals between spikes from the same fiber or stimulus presentation are either included (de Cheveigné 1993) or preferably excluded (Joris 2001).

In contrast, first-order statistics cannot usefully be applied to a population because, as the aggregate rate increases, most intervals join the zero-order mode (mode near zero lag, due to multiple spikes within the same period). The period mode becomes depleted, an effect accompanied by a shift of that mode towards shorter intervals (this phenomenon has actually been invoked to explain certain pitch shifts, Ohgushi 1978, Hartmann 1993). The all-order histogram does not have this problem and is thus a better representation.

It is important to realize that any statistic *discards* information. Different histograms are not equivalent, and the wrong choice of histogram may lead to misleading results. For example, the ISI histogram applied to the response to certain inharmonic stimuli reveals, as expected, the “first effect of pitch shift” whereas a period histogram locked to the envelope does not (Evans 1978). Care must be exercised in the choice and interpretation of statistics.

9.5 Related models

The *schematic model* of Moore (1977, 2003) embodies the essence of the AC model. Its description includes features (such as an upper limit on delays) that allow it to account for most important aspects of pitch (Moore 2003).

The *cancellation model* (de Cheveigné 1998) is based on the *difference function* of Eq. 3 instead of the ACF of Eq. 1. Equation 4 relates the two functions, and cancellation and AC models are therefore formally similar. Peaks of the ACF (Fig. 10A) correspond to valleys of the difference function (Fig. 10B). The appeal of cancellation is that it may account also for *segregation* of harmonic sources (de Cheveigné 1993, 1997a), which makes it useful in the context of multiple pitches (see Section 10.6). A “neural” implementation, on the lines of Licklider’s, is obtained by replacing an excitatory synapse of the coincidence neuron by an *inhibitory* synapse, and assuming that every excitatory spike is transmitted unless it coincides with an inhibitory spike. Roots of this model are to be found in the Equalization-Cancellation model of binaural interaction of Durlach (1963), and the Average Magnitude Difference Function (AMDF) method of speech F0 estimation of Ross et al. (1974) (see Hess 1983 for similar earlier methods).

The *Strobed Temporal Integration* (STI) model of Patterson et al. (1992) replaces autocorrelation by cross-correlation with a train of “strobe” pulses:

$$STI(\tau) = \int s(t)x(t - \tau)dt \quad (5)$$

where $s(t)$ is a train of pulses derived by some process such as peak picking. Processing occurs within each filter channel, and produces a 2D pattern similar to Licklider’s. In contrast to autocorrelation, the STI operation itself is phase-sensitive. It thus predicts perceptual sensitivity to time reversal of some stimuli (Patterson 1994a), although it is not clear that it also predicts the *insensitivity* observed for others. A possible advantage of STI over the ACF is that the *strobe* can be delayed instead of the signal:

$$STI(\tau) = \int s(t - \tau)x(t)dt \quad (6)$$

in which case the implementation of the delay might be less costly (a pulse is cheaper to delay than an arbitrary waveform). Within the brainstem, octopus cells have strobe-like properties, and their projections are well represented in man (Adams 1997). A possible weakness of STI is that it depends, as do early temporal models, on the assignment of a marker (strobe) to each period.

The term *Auditory Image Model* (AIM) refers, according to context, either to STI or to a wider class including autocorrelation. Thanks to strobed integration, the fleeting patterns of transduced activity are “stabilized” to form an *image*. As in similar displays based on the ACF (e.g. Lyon 1984; Weintraub 1985; Slaney 1990), we can hope that *visually* prominent features of this image might be easily accessible to a central processor. An earlier incarnation of the image idea is the “camera acustica” model of Ewald (1898; Wever 1949) in which the cochlea behaved as a resonant membrane. The pattern of standing waves was supposed to be characteristic of each stimulus. STI and AIM evolved from earlier *pulse ribbon* and *spiral* detection models (Patterson 1986, 1987).

The *dominant component* representation of Delgutte (1984) and the *modulation filterbank* model (e.g. Dau et al. 1996) were mentioned earlier. After transduction in the cochlea, the temporal pattern within each cochlear channel is Fourier-transformed, or split over a bank of internal filters, each tuned to its own “best modulation frequency” (BMF). The result is a 2D pattern (cochlear CF vs. modulation Fourier frequency or BMF). To the degree that this pattern resembles a power spectrum, modulation filterbank and AC models are related. The modulation filterbank was designed to explain sensitivity to slow modulations in the infrapitch range, but it has also been proposed for pitch (Wiegrebe et al. 2004).

Interestingly, the *string* can be seen as belonging to the AC model family. Autocorrelation involves two steps: delay and multiplication followed by temporal integration, as illustrated in Figure 10A. Cancellation involves delay, *subtraction* and squaring as illustrated in Figure 10B. Delgutte (1984) described a comb-filter consisting of delay, *addition* and (presumably) squaring as in Figure 10C. This last circuit can be modified as illustrated in Figure 10D. The frequency characteristics of both circuits have peaks at all multiples of $f=1/\tau$, but the peaks of the latter are sharper. A *string* is, in essence, a delay line that feeds back onto itself as in Figure 10D. Cariani (2003) recently proposed that neural patterns might circulate within *recurrent timing nets*, producing a build-up of activity within loops that match the period of the pattern. This too fits the description of a string.

These examples show that autocorrelation and the string (and thus pattern matching) are closely related. They differ in the important respect of *temporal* resolution. At each instant, the ACF reflects a relatively short interval of its input (sum of the delay τ and the duration of temporal smoothing). The string reflects the past waveform over a much longer interval, as information is recycled within the delay line. In effect, this allows comparisons across *multiples* of τ , which improves frequency resolution at the expense of time resolution. Another way to capture regularity over longer intervals is the *narrowed AC function* (NAC) of Brown and Puckette (1989) in which high-order modes of the ACF are scaled and added to sharpen the period mode. The NAC was invoked by de Cheveigné (1989) and Slaney (1990) to explain acuity of pure tone discrimination. Another twist is to fit the AC histogram to exponentially-tapered “periodic templates” (Cedolin and Delgutte 2004), the best-fitting template indicating the pitch. NAC and periodic template can be seen as “subharmonic” counterparts of “harmonic” pattern-matching schemes. Once again we find strong connections between different models.

To conclude on a historical note, a precursor of autocorrelation was proposed by Hurst (1895), who suggested that sound propagates up the tympanic duct, through the helicotrema, and back down the vestibular duct. Where an ascending pulse meets a descending pulse, the BM is pressed from both sides. That position characterizes the period. More recently, Loeb et al. (1983) and Shamma et al. (1989) invoked the BM as an alternative to neural delays. The BM is dispersive and behaves as a delay line only for a narrow-band stimulus. Delay can then be equated to *phase*, which brings us very close to some of the spectral sharpening schemes evoked earlier.

9.6 Selecting the period mode

The description of the AC model is not quite complete. The ACF or SACF of a periodic stimulus has *several* modes, one at each multiple of the period, including zero (Fig. 11A). The cue to pitch is the *leftmost* of the modes at *positive* multiples (dark arrow). To be complete a model should specify the mechanism by which that mode is selected. A pattern-matching model is confronted with the similar problem of choosing among candidate subharmonics (Fig. 1F). This seemingly trivial step is one of the major difficulties in period estimation, rarely addressed in pitch models. There are several approaches.

The easiest is to set limits for the period range (Fig. 11B). To avoid more than one mode within the range (in which case the cue would still be ambiguous), the range must be at most one octave, a serious limitation given that musical pitch extends over about 7 octaves. A second approach is to set a lower period limit and use some form of bias to favor modes at shorter lags (Fig. 11C). Pressnitzer et al. (2001) used such a bias (which occurs naturally when the ACF is calculated from a short-term Fourier transform, as in some implementations) to deemphasize pitch cues beyond the lower limit of musical pitch. A difficulty is that the period mode is sometimes less salient than the zero-order mode (or a spurious mode near it) (Fig. 11D). The difficulty can be circumvented by various heuristics, but they tend to be messy and to lack generality. A solution recently proposed in the context of F0 estimation (de Cheveigné and Kawahara 2002) is based on the difference function (Eq. 3, Fig 9B). A normalization operation removes the dip at zero lag, after which the period lag may be selected reliably.

Once the mode (or dip) has been chosen, its *position* must be accurately measured. Supposing there is internal noise, it is not clear how the relatively wide modes obtained for a pure tone (Fig. 11) can be located with accuracy consistent with discrimination thresholds (about 0.2% at 1 kHz, Moore 1973). One solution is to suppose that higher-order modes contribute to the period estimate (e.g. de Cheveigné 1989, 2000). Another is to suppose that histograms are fed to matched filters (Goldstein and Srulovicz 1977). If the task is pitch *discrimination*, it may not be necessary to actually choose or locate a mode. For example Meddis and O'Mard (1997) used Euclidean distance between SACF patterns to predict discrimination thresholds. However, it is not easy to explain on that basis how a subject decides that one of two stimuli is *higher* in pitch, or how a manifold of stimuli (with same period but diverse timbres) maps to a common pitch.

To summarize, the AC model characterizes periodicity by measuring *self-similarity across time*, either of the acoustic waveform or of the internal patterns it gives rise to. At an abstract level, autocorrelation and pattern matching are linked via an important mathematical theorem, the Wiener-Khintchine theorem, which says that the ACF is the Fourier transform of the power spectrum. At a detailed level, they differ considerably in how they might be implemented in the auditory system. There are also important conceptual differences. For pattern matching, pure tones have the status of *elementary* stimuli. For the AC model they are like any other periodic stimulus, special only in that they affect a limited set of peripheral channels.

Pattern matching solves the missing fundamental problem; for the AC model that problem does not occur. Pattern matching and autocorrelation, through their many variants, are the main contenders today for explaining pitch perception.

10 Advanced Topics

Modern pitch models account for major phenomena equally well. To decide between models, one must look at more arcane phenomena, second-order effects and implementation constraints. A model should ideally be able to fit them all; should it fail we may look to alternate models. In a sense, here is the cutting edge of pitch theory. The casual reader should skip to Section 11 and come back on a rainy day. Brave reader, read on.

10.1 Combination Tones

When two pure tones are added, their sum fluctuates (*beats*) at a rate equal to the difference of their frequencies. Young (1800) suggested that beats of the appropriate frequency could give rise to a pitch, and thus explain the “Tartini” tones sometimes observed in music (Boring 1942). By construction, the stimulus contains no partial at the beat frequency. The pitch that it evokes is therefore a counter-example to Ohm’s law.

If the medium is *nonlinear*, distortion products (harmonics and combination tones) may arise at the beat frequency and various other frequencies. If such were the case every time a pitch is heard, then Ohm’s law could be saved. Perhaps for that reason, there seems to have been a strong tendency to believe this hypothesis, and to assign any pitch not accounted for by a partial to a distortion product.

If the stimulus is a pure tone of frequency f , distortion products are harmonics nf . If the stimulus contains two partials at f and g , they also include terms of the form $\pm nf \pm mg$ (where m and n are integers). Their amplitudes depend on the amplitudes of the primaries and the shape of the nonlinearity. *If the nonlinearity can be expanded as a Taylor series* around zero, these amplitudes can be calculated relatively easily (Helmholtz 1877; Hartmann 1997). The first term (linear) determines the primaries f and g . The second term (quadratic) determines the even harmonics and the difference tone $g-f$. The third (cubic) determines the odd harmonics and the “cubic difference tone” $2f-g$. Higher-order terms introduce other products. Amplitudes increase at a rate of 2 dB per dB for the difference tone, and 3 dB per dB for the cubic difference tone, as a function of the amplitude of the primaries. However all this holds only if the nonlinearity can be expanded as a Taylor series. There is no reason why that should always be the case. As a counter-example, distortion products of a half-wave rectifier vary in direct proportion to the amplitude of primaries.

The *difference tone* $g-f$ played an important role in the early history of pitch theory. Its frequency is the same as that of *beats*, so it could account for the pitches that they evoke (“Tartini tones”), and also for the pitch of a “missing-fundamental” stimulus. Helmholtz argued that distortion might

arise (a) within equipment used to produce “missing-fundamental” stimuli, (b) within the ear. The first argument faded with progress in instrumentation. It was already weak because periodicity pitch is salient at low amplitudes, and apparently unrelated to measurements or calculations of the difference tone.

We already noted that the second argument does not save Ohm’s law, as that law claims to relate *stimulus* components (as opposed to internally produced) to pitches. Not only that, it is possible to *cancel* (and at the same time estimate) any difference tone produced by the ear, by adding an external pure tone of equal frequency, opposite phase, and appropriate amplitude (Rayleigh 1896). Adding a second low-amplitude pure tone at a slightly different frequency, and checking for the absence of beats, makes the measurement very accurate (Schouten 1938, 1970). After this very weak distortion product is canceled the pitch remains the same, so the difference tone $g-f$ cannot account for periodicity pitch.

The *harmonics nf* played a confusing role. Being higher in frequency than the primaries they are expected to be more susceptible to masking than difference tones. Indeed, they are not normally perceived except at very high amplitudes. Yet Wegel and Lane (1924) found beats between a primary and a probe tone near its octave. This, they thought, indicated the presence of a relatively strong second harmonic. They estimated its amplitude by adjusting the amplitude of the probe tone to maximize the salience of beats. This method of *best beats* was widely used to estimate distortion products. Eventually, the method was found to be flawed: beats can arise from the slow variation in phase between nearly harmonically related partials (Plomp 1967b). Beats do not require closely spaced components, and thus do *not* indicate the presence of a harmonic.

This realization came after many such measurements had been published. As “proof” of non-linearity, aural harmonics bolstered the hypothesis that the difference-tone accounts for the missing fundamental. Thus they added to confusion (on the role of difference products, see Pressnitzer and Patterson 2001). Similarly confusing were measurements of distortion products in cochlear microphonics (Newman et al. 1937), or auditory nerve-fiber responses. They arise because of nonlinear mechanical-to-nervous or electrical transduction, and do *not* reflect BM distortion components equivalent to stimulus partials, and thus are not of significance in the debate (Plomp 1965).

In contrast to other products, the *cubic difference tone 2f-g* is genuinely important for pitch theory. Its amplitude varies roughly in proportion with the primaries (and not as their cube as expected from a Taylor-series nonlinearity). It increases as f and g become closer, but it is only measurable (by Rayleigh’s cancellation method) for g/f ratios above 1.1, at which point it is about 14 dB below the primaries (Goldstein 1970). Amplitude decreases rapidly as the frequency spacing increases. A combination tone, even if weak, can strongly affect pitch if it falls within the *dominance region* (Plack and Oxenham, Chapter 2). Difference tones of higher order ($f-n(g-f)$) can also contribute (Smooenburg 1970).

Combination tones are important for pitch theory. They are necessary to explain the “second effect” of pitch shift of frequency-shifted complexes (Smooenburg 1970; de Boer 1976). As their amplitudes are phase sensitive, they allow spectral theories to account for aspects of phase sensitivity. Their

effect can be conveniently “modeled” as additional stimulus components, with parameters that can be calculated or measured by the cancellation method (e.g. Pressnitzer and Patterson 2001). To avoid having to do so, most pitch experimenters now add *lowpass noise* (e.g. pink noise) to mask distortion products.

10.2 Temporal integration and resolution

A question has puzzled thinkers on and off: waves (or pulses, or particles) follow each other in time, how is it that we hear a *continuous* sound? Bonnier (1901), for example, argued that unipolar excitation of cochlear sensory cells would evoke an intermittent sensation if the BM did not act as a *delay line* (of 30-50 ms): at every instant, at least one cell along the delay line is excited by the excitatory phase of the waveform, allowing sensation to be continuous at least for F0s above about 20-30 Hz. Here we have the notion that patterns must be *integrated* over time to ensure smoothness (or stability of estimates over time). All models need temporal integration. It may be explicit as here, or implicit via build-up and decay of resonance.

On the other hand, Helmholtz argued that smoothing must not be excessive, because the ear needs to follow “shakes” of up to 8 notes per second that occur in music. Using 1/8 s as an upper limit on the response time of the resonators in his model, he derived a lower limit on their bandwidth, anticipating the *time-frequency tradeoff* of Gábor (1947) (analogous to Heisenberg’s principle of uncertainty in quantum mechanics). The tradeoff is expressed as:

$$\Delta f \Delta t \geq k \quad (7)$$

where Δf and Δt are frequency and time uncertainties respectively, and k is a constant that depends on how they are measured. Fine *spectral* resolution thus requires a long *temporal* analysis window. Moore (1973) calculated the resolution Δf with which pure tones of duration d could be discriminated on the basis of excitation pattern amplitude changes of at least 1 dB. He found the relation $\Delta f \cdot d \geq 0.24$, analogous to Equation 7. He also found that psychophysical frequency difference limens were about *10 times better* than the relation implies. As Gábor’s relation is so very fundamental, this is puzzling.

The puzzle was explained by Nordmark (1968, 1970). The word “frequency” commonly carries two different meanings. One is the reciprocal of the interval between two events of equal phase, called *phase frequency* by Kneser (1948; Nordmark 1968, 1970). The other is *group frequency* as measured by Fourier analysis:

For a time function of limited duration, [Fourier] analysis will yield a series of sine and cosine waves grouped around the phase frequency. No exact value can be given [to] the group frequency, which is thus subject to the uncertainty relation (Nordmark, 1970).

In contrast to group frequency, phase frequency can be determined with *arbitrary accuracy* by measuring time between two “events”. This strong claim seems to imply the superiority of event-based (temporal) over spectral models, but we argued earlier that *events* themselves are hard to extract

reliably (Section 2.2). Could a similar claim be made for a model that does not use events, say, for autocorrelation?

Take an ongoing signal $x(t)$ that is known to be periodic with some period T . Given a signal chunk of duration D , suppose that we find $T \leq D/2$ such that $x(t) = x(t+T)$ for every t such that both t and $t+T$ fall within the chunk. T *might* be the period, but can we rule out other candidates $T' \neq T$? Shorter periods can be ruled out by trying every $T' \leq T$ and checking if we have $x(t) = x(t+T')$ for every t such that both t and $t+T'$ fall within the chunk. If this fails we can rule out a shorter period. However, we cannot rule out that the true period is *longer than* $D-T$, because our chunk might be part of a larger pattern. To rule this out we must know the *longest expected period* T_{MAX} , and we must have $D \geq T + T_{MAX}$. If this condition is satisfied, then there is no limit to the *resolution* with which T is determined. These conditions can be transposed to the short-term running ACF:

$$r(\tau) = \int_{t=0}^W x(t)x(t-\tau)dt \quad (8)$$

Two time constants are involved: the window size W , and the maximum lag τ_{MAX} for which the function is calculated. They map to T_{MAX} and T respectively in the previous discussion. The required duration is their sum, and depends thus on the *lower limit* of the expected F0 range. A rule of thumb is to allow at least $2T_{MAX}$.

As an example, the lower limit of melodic pitch is near 30 Hz (period \approx 33 ms) (Pressnitzer et al. 2001). To estimate arbitrary pitches requires about 66 ms. If the F0 is 100 Hz (period = 1/10 ms) the time can be shortened to $33+10=43$ ms. If we *know* that the F0 is no lower than 100 Hz, the duration may be further shortened to $10+10=20$ ms. These estimates apply in the absence of noise. With noise present, internal or external, more time may be needed to counter its effects.

We might speculate that *pattern matching* allows even better temporal resolution, because periods of harmonics are shorter and require (according to the above reasoning) less time to estimate than the fundamental. Unfortunately, harmonics must be resolved, and for that the signal must be stable over the duration of the impulse response of the filterbank that resolves them.

Suppose now that the stimulus is *longer* than the required minimum. The extra time can be used according to at least three strategies. The first is to increase integration time to reduce noise. The second is to test for self-similarity across period multiples, so as to refine the period estimate. The third (so-called “multiple looks” strategy) is to cut the stimulus into intervals, derive an estimate from each, and average the estimates. The benefit of each can be quantified. Denoting as E the extra duration, the first strategy increases integration time by a factor $n_1=(E+W)/W$, and thus reduces variability of the *pattern* (e.g. ACF) by a factor of $\sqrt{n_1}$. The second reduces variability of the *estimate* by a factor of at least $n_2=(E+T)/T$, by estimating the period multiple n_2T and then dividing. It could probably do even better by including also estimates of smaller multiples of the period. The third allows $n_3=(E+D)/D$ multiple looks (where $D \geq T+W$ is interval duration), and thus reduces variability of the *estimate* by a factor of $\sqrt{n_3}$. The benefit of the first strategy is hard to judge without knowledge of the relation between pattern variability and estimate variability. The second strategy seems better than the third (if n_2 and n_3 are comparable). Studies

that invoke the third strategy often treat intervals as if they were *surrounded by silence* and thus discard structure across interval boundaries. This is certainly suboptimal. A priori, the auditory system could use any of these strategies, or some combination. The second strategy suggests a roughly inverse dependency of discrimination thresholds on duration (as observed by Moore 1973 for pure tones up to 1-2 kHz), while the other two imply a shallower dependency.

What parameters should be used in models? Licklider (1951) tentatively chose 2.5 ms for the size of his exponentially shaped integration windows (roughly corresponding to W). Based on the analysis above, this size is sufficient only for periods shorter than 2.5 ms (frequencies above 250 Hz). A larger value, 10 ms, was used by Meddis and Hewitt (1992). From experimental data, Wiegrebe et al. (1998) argued for *two* stages of integration separated by a nonlinearity. The first had a 1.5 ms window and the second some larger value. Wiegrebe (2001) later found evidence for a *period-dependent* window size of about twice the stimulus period, with a minimum of 2.5 ms. These values reflect the *minimum* duration needed.

In Moore's (1973) study, pure tone thresholds varied inversely with duration up to a frequency-dependent limit (100 ms at 500 Hz), beyond which improvement was more gradual. In a task where isolated harmonics were presented one after the other in noise, Grose et al. (2002) found that they merged to evoke a fundamental pitch only if they spanned less than 210 ms. Both results suggest also a *maximum* integration time.

Obviously, an organism does not want to integrate for longer than is useful, especially if a longer window would include garbage. Plack and White (2000a,b) found that integration may be *reset* by transient events. Resetting is required by sampling models of frequency modulation (FM) or glide perception. Resetting is also required to compare intervals across time in discrimination tasks. Those tasks also require *memory* for the result of sampling, and it is conceivable that integration and sensory memory have a common substrate.

10.3 Dynamic pitch

Aristoxenos distinguished the stationarity of a musical note, with a pitch from deep to high, from the continuity of the spoken voice or transitions between notes, with qualities of tension or relaxation. The exact terms chosen by the translator (Macran 1902) are of less interest than the fact that the concepts of *static* and *dynamic* pitch were so carefully distinguished. It is indeed conceivable that dynamic pitch is perceived differently from static pitch. For example FM might be transformed to amplitude modulation (AM) and perceived by an AM-sensitive mechanism (Moore and Sek 1994), or frequency glides might be decoded by a mechanism directly sensitive to the derivative of frequency (Sek and Moore 1999). The alternative is that frequency is *sampled* by the mechanism used for static pitch, and the samples compared across time (Hartmann and Klein 1980; Dooley and Moore 1988). For this to work, the estimation mechanism must be tolerant to frequency change.

Estimation is not instantaneous (Section 10.2), so frequency "sampling" makes sense only in a limited way. Frequency change impairs periodicity,

and this makes estimation more difficult. Integration over time of unequal frequencies "blurs" the estimate of *the* frequency at any instant. A shorter window reduces the blur, but at the expense of the accuracy of the estimation process (Section 10.2).

Discrimination of frequency-modulated patterns is thus expected to be poor. Strangely, Demany and Clément (1997) observed what they called "hyperacute" discrimination of *peaks* of frequency modulation. Thresholds were smaller than expected given the lack of stable intervals long enough to support a sampling model. A possible explanation is that periods shrink during the up-going ramp, and expand during the down-going ramp. Cross-period measurements that *span* the modulation peak are therefore relatively stable, leading to relatively good discrimination (de Cheveigné 2001).

The case might be made for the opposite proposition, that tasks involving *static* pitch (such as frequency discrimination) actually involve detectors sensitive to *frequency change* (Okada and Kashino 2003; Demany and Ramos 2004). It is often noted that weak pitches become more salient when they change (Davis 1951), so change may play a fundamental role in pitch. In the extreme one could propose that pitch is *not* a linear perceptual dimension, but rather some combination of sensitivities to pitch-change and to musical interval. Whether or not this is the case, we still need to explain the extraction of the quantity that changes.

If listeners are asked to judge the *overall* pitch of a frequency-modulated stimulus, the result can usually be predicted from the *average instantaneous frequency*. If amplitude changes together with frequency, overall pitch is well predicted by the intensity- or envelope-weighted average instantaneous frequency (IWAIF or EWAIF) models (Anantharama et al. 1993; Dai et al. 1996). Even better predictions are obtained if frequency is weighted inversely with *rate of change* (Gockel et al. 2001).

10.4 Unresolved partials

For Helmholtz, Ohm's law applied only to *resolved* partials. Schouten later extended the law by assigning the remaining *unresolved* partials to a new sensory component, the residue. The resolved vs. unresolved distinction is crucial for pattern matching because resolved partials alone can offer a useful pattern. It was once crucial also for temporal models, because unresolved partials alone can produce, on the BM, the fundamental periodicity that was thought necessary for a "residue pitch".

The distinction is still made today. Many modern studies use only stimuli with unresolved partials (to rule out "spectral cues"). Others contrast them with stimuli for which at least some partials are resolved. "Unresolved stimuli" are produced by a combination of high-pass filtering, to remove any resolved partials, and addition of low-pass noise to mask the possibly resolvable combination tones. Reasons for this interest are of two sorts. Empirically, pitch-related phenomena are surprisingly different between the two conditions (Plack and Oxenham, Chapter 2). Theoretically, pattern matching is viable only for resolved partials, so phenomena observed with unresolved partials cannot be explained by pattern matching. Autocorrelation is viable for both, but the experiments are nevertheless used to test it too. The argument is: "Autocorrelation being equally capable of

handling both conditions, large differences between conditions imply that autocorrelation is *not* used for both”. It applies to any unitary model. I find the argument not altogether convincing for two reasons: other accounts might fit the premises, and the premises themselves are not clear-cut.

Auditory filters have roughly constant Q , and thus unresolved partials are necessarily of high rank. *Rank*, rather than resolvability, might limit performance. Indeed, Moore (2003) suggested a maximum delay of $15/CF$ in each channel, implying a maximum rank of 15. Other possible accounts are: (a) Spectral region staying the same, unresolved stimuli must have *longer periods*, and longer periods may be penalized. (b) Period staying the same, unresolved stimuli must occupy *higher spectral regions*, and high-frequency channels might represent periodicity less well. (c) Lowpass noise added to lower spectral regions (that normally dominate pitch) in unresolved conditions may have a *deleterious* effect that penalizes those conditions. (d) The auditory system may *learn* to ignore channels where partials are unresolved, for example because they are phase sensitive (and thus more affected by reverberation), etc. These accounts need to be ruled out before effects are assigned to resolvability.

A clear behavioral difference between resolved and unresolved conditions is the order-of-magnitude step in F_0 discrimination thresholds between complex tones that include lower harmonics and those that don't. The limit occurs near the 10th harmonic and is quite sharp (Houtsma and Smurzynski 1990; Shackleton and Carlyon 1994; Bernstein and Oxenham 2003). Higher thresholds are attributed to the poor resolvability of higher harmonics.

If such is the case, we expect direct measures of partial resolvability to show a breakpoint near this limit. A resolvable partial must be capable of evoking its own pitch (at least according to Terhardt's model). An *isolated* partial certainly does, but two are individually perceptible only if their frequencies differ by at least 8% at 500 Hz, and somewhat more at higher or lower frequencies (Plomp 1964). Closer spacing yields a single pitch, function of the *centroid* of the power spectrum (Dai et al. 1996) (this justifies the assertion made in Sect. 2.5 that *spectral pitch* depends on the locus of a spectral concentration of power). The 10th harmonic is about 9% from its closest neighbor, so this measure is roughly consistent with the breakpoint in complex F_0 discrimination.

However, with neighbors on both sides, a partial is less well resolved. Harmonics in a complex are resolved only up to rank 5-8 (Plomp 1964). This does not agree with a breakpoint at rank 10. By pulsating the partial within the complex, Bernstein and Oxenham (2003) found a higher resolvability limit (10-11) that fit well with F_0 discrimination thresholds in the same subjects. However, when *even* and *odd* partials were sent to different ears (thus doubling their spacing within each cochlea), partials were resolvable to about the 20th, and yet the breakpoint in F_0 discrimination limens still occurred at a low rank. The two measures of resolvability do not fit.

Various other phenomena show differences between resolved and unresolved conditions: *frequency modulation detection* (Plack and Carlyon 1995; Carlyon et al. 2000), *streaming* (Grimault et al. 2000), *temporal integration* (Plack and Carlyon 1995; Micheyl and Carlyon 1998), *pitch of concurrent harmonic sounds* (Carlyon 1996), *F_0 discrimination between*

resolved and unresolved stimuli (Carlyon and Shackleton 1994; see also Oxenham et al. 2004), etc. If breakpoints always occurred at the same point along the resolved-unresolved continua, the resolvability hypothesis would be strengthened. However the parameter space is often sampled too sparsely to tell. A popular stimulus set (FOs of 88 and 250Hz and frequency regions of 125-625, 1375-1875, and 3900-5400 Hz) offers several resolved-unresolved continua but each is sampled only at its well-separated endpoints. Inter-partial distances are drastically reduced if complex tones are added; yet “resolvability” (as defined for an *isolated* tone) seems to govern the salience of pitch within a *mixture* (Carlyon 1996). The *lower limit of musical pitch* increases in higher spectral regions, as expected if it was governed by resolvability, but the boundary follows a different trend, and extends well within the unresolvable zone (Pressnitzer et al. 2001). Some data do not fit the resolvable/unresolvable dichotomy.

To summarize, many modern studies focus on stimuli with unresolved partials. Aims are: (a) to test the hypothesis of distinct pitch mechanisms for resolved and unresolved complexes (next Section), (b) to get more proof (if needed) that pitch can be derived from purely temporal cues, or (c) to obtain an analogue of the impoverished stimuli available to cochlear implantees (Moore and Carlyon, Chapter 7). This comes at a cost, as it focuses efforts on a region of the parameter space where pitch is weak, quite remote from the musical sounds that we usually take as pleasant. It is justified by the theoretical importance of resolvability.

10.5 The two-mechanism hypothesis

Pattern matching and autocorrelation each has its strengths and followers. It is tempting to adopt both and assign to each a different region of parameter space: pattern matching to stimuli with resolved harmonics, and autocorrelation to stimuli with no resolved harmonics. The advantages are a better fit to data, and better relations between tenants of each approach. The disadvantages are that two mechanisms are involved, plus a third to integrate the two.

The temptation of multiple explanations is not new. Vibrations were once thought to take two paths through the middle ear: via ossicles to the oval window, and via air to the round window. Müller’s experiment reduced them to one (Fig. 3). Du Verney (1683) believed that the trumpet-shaped semicircular canals were tuned like the cochlea, while Helmholtz thought the ampullae handled noise-like sounds until he realized that cochlear spectral analysis could take care of them too. Bonnier (1896-98) assigned the sacculus to sound localization (as a sort of “auditory retina”) and the cochlea to frequency analysis. Bachem (1937) postulated two independent pitch mechanisms, one devoted to tone height, the other to chroma, the latter better developed in possessors of absolute pitch. Wever (1949) suggested that low frequencies are handled by a temporal mechanism (volley theory) and high frequencies by a place mechanism, and Licklider’s duplex model implemented both (with a learned neural network to connect them together). The motivation is to obtain a better fit with phenomena, and perhaps sometimes also to find a use for a component that a simpler model would ignore.

There is evidence for both temporal *and* place mechanisms (e.g. Gockel et al. 2001; Moore 2003). The assumption of independent mechanisms for resolved and unresolved harmonics is also becoming popular (Houtsma and Smurzynski 1990; Carlyon and Shackleton 1994). It has also been proposed that a unitary model might suffice (Houtsma and Smurzynski 1990; Meddis and O'Mard 1997). The issue is hard to decide. Unitary models may have serious problems (e.g. Carlyon 1998a,b) that a two-mechanism model can fix. On the other hand, assuming two mechanisms is akin to adding free parameters to a model: it automatically allows a better fit. The assumption should thus be made with reluctance (which does not mean that it is not correct). A two-mechanism model compounds vulnerabilities of both, such as lack of physiological evidence for delay lines or harmonic templates.

10.6 Multiple pitches

Pitch models usually account for a single pitch, but some stimuli evoke more than one: (a) stimuli with an ambiguous periodicity pitch, (b) narrow-band stimuli that evoke both a periodicity pitch and a spectral pitch, (c) concurrent voices or instruments, (d) complex tones in analytic listening mode.

Early experiments with stimuli containing few harmonics sometimes found multimodal distributions of pitch matches (de Boer 1956; Schouten et al. 1962). Pitch models usually produce multiple or ambiguous cues for such stimuli (e.g. Fig. 2F), and with appropriate weighting they should account for “multiple” pitches of this kind.

A formant-like stimulus may produce a *spectral pitch* related to the formant frequency (Section 2.5). The spectral pitch may coexist with a lower periodicity pitch if the stimulus is a periodic complex. For pure tones the two pitches are confounded. In so-called diphonic singing styles of Mongolia or Tibet, spectral pitch carries the melody while periodicity pitch serves as a drone. Some listeners may be more sensitive to one or the other (Smooenburg 1970). It is common to attribute periodicity pitch to temporal analysis, and spectral pitch to cochlear analysis, reflecting two different mechanisms. However one cannot exclude a common mechanism. A sharp spectral locus implies quasi-periodicity in the time domain, and this shows up as modes at short lags in the ACF (insert in Figure 5).

In music, instruments often play together, each with its own pitch, and appropriately gifted or trained people may perceive their *multiple pitches* (see Darwin, Chapter 8). Reverberation may transform a monodic melody into polyphony of two parts or more (the echo of a note accompanies the next). Sabine (1907) suggested that this is why scales appropriate for harmony emerged before polyphonic style. Models described so far address only the *single* pitch of an *isolated* tone, and cannot account for more without modification. A simple idea is to take the pattern that produced a pitch cue for an isolated tone, and scan it for *several* such cues. As an example, Assmann and Summerfield (1990) estimated the F0s of two concurrent vowels from the *largest* and *second-largest* peaks of the SACF. Unfortunately, distinct peaks do not always exist (simulations based on this procedure gave comparatively poor results; de Cheveigné 1993).

A better procedure is to estimate pitches *iteratively* (de Cheveigné and Kawahara 1999), by estimating first one period and then removing it. In the context of *pattern matching*, this is known as the “harmonic sieve” (Duifhuis et al. 1982; Parsons 1976). An initial F0 estimate is derived from the pattern of partials. Partial that fit its harmonic series (within some tolerance) are removed, and a second F0 is estimated from the remainder. The process may be iterated, each F0 controlling the sieve in turn. Scheffers (1983) tested the idea using spectral analysis similar to that of the ear, but found that F0s were rarely both estimated correctly. The reason given was lack of spectral resolution. As discussed in Section 10.4, partials within 8-10% of another partial are not readily resolved (they tend to merge and give rise to a single, intermediate pitch). Since many partials of a mixture have closer spacing, the applicability of a “harmonic sieve” is limited.

Iterative estimation works also with the *AC model*. A first period is estimated from the SACF, channels dominated by that period are discarded, and a second period is estimated from the remainder. Weintraub (1985) and Meddis and Hewitt (1992) used this procedure to segregate speech sounds. *Cancellation* (Section 9.5) can be used in place of autocorrelation, but it offers additional options. A period may be suppressed *within a channel*, for example to estimate a tone too weak to dominate any channel. The steps of suppression and estimation may also be merged into a *joint estimation* procedure (de Cheveigné and Kawahara 1999).

The harmonic sieve requires that partials be spaced wide enough to be resolved. Meddis and Hewitt's scheme requires spectral envelopes, with features (e.g. formants) broad enough to be resolved. Cancellation (if implemented perfectly) does not depend on peripheral resolution. Carlyon (1996) found that subjects could *not* perceive two pitches within pairs of “unresolved” complexes (see Section 10.4) so the effectiveness of cancellation, if used by the auditory system, must have limits.

As noted by Mersenne (1636), careful listening to a complex reveals higher pitches in addition to the fundamental. Helmholtz (1857, 1877) attributed each partial pitch to an elementary *sensation* produced by a sinusoidal partial⁵. Partial pitches are not commonly heard, but for Helmholtz they nevertheless underlie all musical perception. We access the lowest partial pitch to perceive the *note*, the next partial pitches to hear *overtones*, and the ensemble of partial pitches to hear *timbre* (Watt 1917 used the word “pitch-blend”). Schouten instead mapped the note to the residue, and Terhardt mapped it to the pattern of partial pitches (his “spectral pitches”), but neither disagreed with Helmholtz’s compositional model of auditory perception.

To account for partial pitches, a pattern-matching model must access the *inputs* of the pattern-matching stage in addition to its *output* (e.g. Terhardt et al. 1982; see also Martens 1982). The AC model instead accounts for them by restricting its processing to particular channels from the periphery.

⁵ Helmholtz's translator Ellis remarked that a partial pitch might correspond instead to a series of harmonically related partials. For example, the partial pitch at the octave might correspond to the series (2, 4, 6, etc.) rather than to the 2nd harmonic, and might even exist in the absence of harmonic 2...

Helmholtz (1857) noted that partials are easier to hear out if *mistuned*. Mistuning also produces a systematic *shift* of the partial pitch (Hartmann and Doty 1996) for which an explanation, based on a time-domain process akin to the harmonic sieve, was proposed by de Cheveigné (1997b, 1999).

To summarize, there are several ways to allow pitch models to handle more than one pitch. Pattern matching models split patterns according to a “harmonic sieve” before matching. AC models divide cochlear channels among sources before periodicity estimation. Cancellation models allow joint estimation of multiple periods. For pattern matching, a partial pitch is a preexisting sensory element, perceptible if it manages to escape fusion. For AC models, it results from a segregation mechanism that involves peripheral (and possibly central) filtering. There are close relations between pitch and segregation (Hartmann 1996; Darwin, Chapter 8). More behavioral data are needed to understand multiple pitch perception.

10.7 Harmony, melody and timbre

Music science was central to science up to the 17th century. The work of Beeckmann, Descartes, Mersenne, the Galilei, and others, were largely aimed at questions such as musical consonance and musical scales (Cohen 1985). Later progress required isolating pitch from the musical context, but it obviously remains relevant and a pitch model should account for its effects. Chroma, intervals, harmony, tonality, effects of context, or the relation between pitch and timbre (Bigand and Tillmann, Chapter 9) are a challenge for pitch models.

Chroma designates a set of equivalence classes based on the *octave* relation. In some cases chroma seems the dominant mode of pitch perception. For example, absolute pitch appears to involve mainly chroma (Bachem 1937; Miyazaki 1990; Ward 1999). Demany and Armand (1984) found that infants treated octave-spaced pure tones as equivalent. A spectral account of octave equivalence is that all partials of the upper tone belong to the harmonic series of the lower tone. A temporal account is that the period of the lower tone is a superperiod of the higher. In both cases the relation is not reflexive (the lower tone contains the upper tone but not vice-versa) and is thus not a true equivalence. Furthermore, similar (if less close) relations exist also for ratios of 3, 5, 6, etc., for which equivalence is not usually invoked. Octave *equivalence* is not an obvious emergent property of pitch models.

Absolute pitch is rare. BM tuning and neural delays being relatively stable, it should be the rule rather than the exception. Relative pitch involves the potentially harder task of abstracting interval relations between period cues along a periodotopic dimension. Some intervals involve simple numerical ratios for which coincidence between partials or subharmonics might be invoked, but accurate interval perception appears to be possible for nonsimple ratios too. Interval perception is not an obvious emergent property of pitch models.

Some aspects of harmony may be “explained” on the basis of simple ratios between period counts or partial frequencies (Rameau 1750; Helmholtz 1877; Cohen 1985). Terhardt et al. (1982, 1991) and Parncutt (1988) explain chord roots on the basis of Terhardt’s pattern-matching

model. To the extent that pattern-matching models are equivalent to each other and to autocorrelation, similar accounts might be built upon other pitch perception models (e.g. Meddis and Hewitt 1991a), but it is not clear how they account for the strong effects of tonal *context* described by Bigand and Tillmann in Chapter 9. Dependency of pitch on context or *set* was emphasized by de Boer (1976).

Section 2.5 pointed out that certain stimuli may evoke two pitches, one dependent on periodicity, and another on the spectral locus of a concentration of power. The latter quantity also maps to a major dimension of *timbre* (brightness) revealed by multidimensional scaling (MDS) experiments (e.g. Marozeau et al. 2003). Historically there has been some overlap in the vocabulary and concepts used to describe pitch (e.g. “low” vs. “high”) and timbre (e.g. “sharp” vs. “dull”) (Boring 1942). In an MDS experiment Plomp (1970) showed that periodicity and spectral locus map to independent subjective dimensions. Tong et al. (1983) similarly found independent dimensions for place and rate of stimulation in a subject implanted with a multielectrode cochlear implant, while McKay and Carlyon (1999) found independent dimensions for carrier and modulator with a *single* electrode (see Moore and Carlyon, Chapter 7). As stressed by Bigand and Tillmann (Chapter 9), the musical properties of pitch must be taken into account by pitch models.

10. 8 Binaural Effects

Binaural hearing has more than once played a key role in pitch theory. The proposal that sounds are localized on the basis of binaural time of arrival (Thompson 1882) implied that *time* (and not just spectrum) is represented internally. Once that is granted, a temporal account of pitch such as Rutherford’s telephone theory becomes plausible. Binaural release from masking (Licklider 1948; Hirsh 1948) later had the same implication. In the “Huggins’ pitch” phenomenon (Cramer and Huggins 1958), a pitch is evoked by white noise, identical at both ears apart from a narrow *phase* transition at a certain frequency. As there is no spectral structure at either ear, this was seen as evidence for a temporal account of pitch.

Huggins’ pitch had prompted Licklider (1959) to formulate the triplex model, in which his own autocorrelation network was preceded by a network of binaural delays and coincidence counters, similar to the well-known localization model of Jeffress (1948). A favorable interaural delay was selected using Jeffress’s model, and pitch was then derived using Licklider’s model. The triplex model used the temporal structure at the output of the binaural coincidence network.

Jeffress’s model involves multiplicative interaction of delayed patterns from both ears. Another model, the Equalization-Cancellation (EC) model of Durlach (1963) invoked *addition* or *subtraction* of patterns from both ears. These could also have been used to produce temporal patterns to feed the triplex model. However Durlach chose instead to use the profile of activity across CFs as a static *tonotopic* pattern. It turns out that many binaural phenomena, including Huggins’ pitch, can be interpreted in terms of a “central spectrum”, analogous to that produced monaurally by a stimulus with a structured (rather than flat) spectrum (Bilsen and Goldstein

1974; Bilsen 1977; Raatgever and Bilsen 1986). Phenomena seen earlier as evidence of a temporal mechanism were now evidence of a *place* mechanism situated at a central level.

In a task involving pitch perception of two-partial complexes, Houtsma and Goldstein (1972) found essentially the same performance if partials went to the same or different ears. In the latter case there is no fundamental periodicity at the periphery. They concluded that pitch cannot be mediated by a temporal mechanism and must be derived centrally from the pattern of resolved partials. These data were a major motivation for pattern matching. However, we noted earlier that Licklider's model does *not* require fundamental periodicity within a peripheral channel. It can derive the period from resolved partials, and it is but a small step to admit that they can come from both ears. Houtsma and Goldstein found that performance was no better with binaural presentation, despite the better resolution of the partials, favorable to pattern matching. Thus, their data could equally be construed as going *against* pattern matching.

An improved version of the EC model gives a good account of most binaural pitches (Culling et al. 1998a,b; Culling 2000). As in the earlier models of Durlach, or Bilsen and colleagues, it produces a tonotopic profile from which pitch cues are derived, but Akeroyd and Summerfield (2000) showed that the *temporal* structure at the output of the EC stage could also be used to derive a pitch (as in the triplex model). A possible objection to that idea is that it requires two stages of time domain processing, which might be costly in terms of anatomy. However, de Cheveigné (2001) showed that the same processing may be performed as one stage. The many interactions between pitch and binaural phenomena (e.g. Carlyon et al. 2001) suggest that periodicity and binaural processing may be partly common.

10. 9 Physiological models

Models reviewed so far proceed by working out an account of how pitch *might* be extracted. The hope is that physiology will eventually provide support, but so far it has not obliged (Winter, Chapter 4). A strong objection to the AC model is the lack of evidence of autocorrelation patterns, or delays of the duration required (at least 30 ms). There is likewise little evidence in favor of pattern matching. A different approach is to start from known anatomy and physiology, and work towards a functional model. This seems a sound approach, as it only allows ingredients known to exist in the auditory system. Weaknesses are: (a) sparse sampling or technical difficulties may prevent the observation of an indispensable ingredient, (b) experiment design and reporting are model-driven, and in particular (c) the wrong choice of stimuli or descriptive statistics might bias model building in an unhelpful way.

The model of Langner (1981, 1998) tries to explain pitch and at the same time account for physiological responses to amplitude-modulated sinusoidal carriers. The basic circuit has two inputs. One is a pulse train phase-locked to the stimulus *carrier* (period $\tau_c = 1/f_c$). The other is a strobe pulse locked to the *modulation envelope* (period $\tau_m = 1/f_m$). The strobe triggers two parallel delay circuits that converge upon a coincidence neuron

that activates if the *delay difference* between pathways equals the modulation period (or an integer multiple $n_m \tau_m$ of that period). An array of such circuits covers periods in the pitch range.

The model has elements reminiscent of those of Licklider and Patterson (Section 9). A distinctive feature is the use of *two* delay circuits rather than one. One (called an “integrator” or “reductor”), accumulates carrier pulses up to some threshold and thus produces a delay (relative to the strobe) equal to an integer multiple of the *carrier period* ($n_c \tau_c$). The other is an oscillator circuit that produces a burst of spikes triggered by the strobe, with a particular “intrinsic oscillation” period τ_i , (a small integer multiple of a synaptic delay of 0.4 ms). The circuit thus actually outputs *several* delayed spikes, all integer multiples of the *oscillator period* ($n_o \tau_o$). Putting things together, coincidence can only occur if the “periodicity equation” is true:

$$n_m \tau_m = n_c \tau_c - n_o \tau_o$$

Since the required integers might not always exist, certain periods might be missing. From this one might predict a step-like trend of psychophysical pitch matches, that Langner (1981) did indeed observe but that Burns (1982) failed to replicate. On the other hand, the equation allows many possible combinations of the six quantities that it involves. As a consequence, the behavior of the model is hard to analyze and compare with other models.

This example illustrates a difficulty of the physiology-driven approach. The physiological data were gathered in response to *amplitude-modulated sinusoids*, which don’t quite fit the stimulus models of Section 2.4. Pitch varies with (f_c, f_m), but the parameter space is non-uniform: regions of true and approximate periodicity alternate, evoking either clear or weak and ambiguous pitch. The choice of parameters leads naturally to posit a model that extracts them in order to get at the pitch, but in this case the task is hard. In contrast, a study *starting from pitch theory* might have used stimuli with parameters easier to relate to pitch, and produced data conducive to a simpler model.

In a different approach, Hewitt and Meddis (1994), and more recently Wiegbe and Meddis (2004) suggested that *chopper cells* in the cochlear nucleus (CN) converge on coincidence cells in the central nucleus of the inferior colliculus (ICC). Choppers tend to fire with spikes regularly spaced at their characteristic interval. Firing tends to align to stimulus transients and, if the period is close to the characteristic interval, the cell is *entrained*. Cells with similar properties may align to similar features and thus fire precisely at the *same instant* within each cycle, leading to the activation of the ICC coincidence cell. A different stimulus period would give a less orderly entrainment, and a smaller ICC output, and in this way the model is tuned. It might seem that periodicity is encoded in the highly regular *interspike intervals*. Actually, it is the temporal alignment of spikes across chopper cells, rather than ISI intervals within cells, that codes the pitch. A feature of this approach is the use of *computational* models of the auditory periphery and brainstem (Meddis 1988; Hewitt et al. 1992) to embody relevant physiological knowledge. Winter (Chapter 4) discusses physiologically based models more deeply.

10.10 Computer models

Material models were once common (e.g. Fig. 3), but nowadays the substrate of choice is software. The many available software packages will not be reviewed, because progress is rapid and information quickly outdated, and because up-to-date tools can easily be found using search tools (or by asking practitioners in the field). The computer allows models of such a complexity that they are not easily understood (a situation that may arise also with mathematical models). The scientist is then in the uncomfortable position of requiring a second model (or metaphor) to understand the first. This is probably unavoidable, as the gap is wide between the complexity of the auditory nervous system and our limited cognitive abilities. We should nevertheless perhaps worry when a researcher treats a model as if it were as opaque as the auditory system. Special mention should be made of the *sharing of software* and source code. In addition to making model production much easier, it allows models to be *communicated*, including those that are not easily described.

10.11 Other modeling approaches

The ideas outlined in this subsection were chosen for their rather unusual view of neural processing of auditory patterns, and thus pitch.

Many theories invoke a *spatial* internal representation, for example tonotopic or periodotopic. A spatial map of pitch fits the high vs. low spatial metaphor that we use for pitch, and thus gives us the feeling of "explaining" pitch. However that metaphor may be recent (Duchez 1989): the Greeks instead used words that fit their experience with stringed instruments, such as "tense" or "lax". A different argument is that distinct pitches must map to (spatially) distinct motor neurons to allow distinct behavioral responses (Whitfield 1970). Licklider (1959) accepted the idea of a map, but questioned the need for it to be spatially *ordered*. The need for the map itself may also be questioned. Cariani (2001) reviews a number of alternate processing and representation schemes based on time.

Maps are usually understood as *rate* vs. place representations, but *time* (of neural discharge relative to an appropriate reference) has been proposed as an alternative to rate (Thorpe et al. 1996). Maass (1998) gave formal proofs that so-called "spiking neural networks" are as powerful, and in some cases more powerful (in terms of network size for a given function), than networks based on rate. Time is a natural dimension of acoustic patterns, and its use within the auditory system makes sense. Within the auditory cortex, transient responses have been found with latencies reproducible to within a millisecond (Elhilali et al. 2004), consistent with a code in terms of spike time relative to a reference spike, itself triggered by a stimulus feature. Maass also pointed out that spiking networks allow arbitrary impulse responses to be synthesized by combining appropriately delayed excitatory and inhibitory post-synaptic potentials (EPSPs and IPSPs). Time-domain filters can thus be implemented within dendritic trees.

Barlow (1961) argued that a likely role of sensory relays is to recode incoming patterns so as to minimize the average number of spikes needed to represent them. For example, supposing the relay has M outputs, the most

common input pattern would map to *no* spike, the M next-most common patterns to *one* spike on one output neuron, etc. Rare patterns would map to patterns with more spikes. The advantages are at least threefold. First, neural activity (and metabolic cost) is minimized, all the more so as M is large. Second, the relay extracts *regularities* in incoming patterns, and thus serves to characterize them. Third, reduced response to common patterns may increase sensitivity to less common events. Early relays would handle simple stimulus-related structure, and the later ones more abstract regularities. *Periodicity* is a candidate for early recoding, and the cancellation model (Section 9.5) actually implements it in some sense.

If Barlow's principle is valid, stimulus-related structure should give way to neural patterns that are *sparse*, as common patterns are coded by few spikes, and *labile*, as the system adjusts to the changing statistics of incoming patterns (Nelken et al. 2004). If so, stable maps of stimulus structure (tonotopy, etc.) at levels beyond brainstem and midbrain might reflect mainly irrelevant leftover structure. Barlow's principle fits well with Bayesian models of information processing (Barlow 2001).

Maass (2003) recently proposed a model of neural processing in two stages. The first performs a large number of non-linear transformations on incoming patterns (he calls it a "liquid state machine"). The only requirement on transforms is that they be sufficiently diverse. The second stage *learns* linear combinations of these transforms. Theoretical analysis and simulations show that this model can efficiently learn arbitrary patterns. Transforms are, as it were, *selected* according to their usefulness. Networks such as Shamma and Klein's harmonic template, Licklider's autocorrelation, or cancellation, if they occurred, would be likely candidates for selection. This is an alternative form of the "learning hypothesis" (Section 5.3).

Licklider's (1951) pitch model is closely related to Jeffress's (1948) binaural model, and success of the latter (Joris et al 1998) has bolstered the former. Recently the Jeffress model has been questioned (McAlpine et al. 2001). It assumes an *array* of spatially-tuned channels within each cochlear frequency band, the channel with maximal activation indicating azimuth. McAlpine and colleagues instead found evidence in the guinea pig for a mechanism analogous to that which encodes color within the visual system. Azimuth affects the balance of activation of *two* channels within each frequency band, one encoding "leftness" and the other "rightness". In other words, within each cochlear frequency band, delay can be assimilated to phase and synthesized as the weighted sum of two quadrature signals. It is logical to ask if a similar mechanism could work for pitch, for example to synthesize delays required by the AC model.

Mach (1884, Boring 1942) actually proposed a two-channel "color scheme" to code pitch height as a combination of "brightness" and "dullness", while a third channel coded "richness of timbre". Köhler (1913, Boring 1942) used a similar idea to represent "vocality" (a quality assimilated to chroma), and Schouten (1940c) mentioned a "color" scheme to represent *periodicity* at each point of the basilar membrane. Helmholtz (1877) had suggested combining adjacent sensory cells to represent intermediate values of pitch, in an effort to preempt the objection that their numbers were too few to code the finer grades of pitch.

Applying a scheme analogous to McAlpine's to pitch involves difficulties of two kinds. First, except in the case of pure tones close in

frequency (Dai et al. 1996), adding sounds of different pitch does not produce a sound of intermediate pitch, as when colors are mixed. Second, the requirements of pitch are harder to satisfy than localization. For a narrow band signal (such as in a cochlear channel), delay can be assimilated to phase and synthesized as the weighted sum of two signals in quadrature phase ($\pm 90^\circ$). Up to 1.7 kHz (most of the range of frequencies studied by McAlpine et al. 2001), delays of up to $\pm 150 \mu\text{s}$ (largest guinea pig ITD) can be synthesized in this way, and if negative weights are allowed, the range can be doubled. Beyond that, the phase-delay mapping is ambiguous. The entire existence region of pitch (Fig. 5) involves delays longer than the period of any partial.

True, for a sufficiently narrow band signal, a large delay can be equated to phase and implemented as a delay shorter than the period (or as the weighted sum of quadrature signals). However this mapping is ambiguous and is hard to see how a pitch model can be built in this way. Nevertheless there may be some way to formulate a model along these lines that works. Certainly the need for a high-resolution array of pitch-sensitive channels might be alleviated, as originally suggested by Helmholtz.

Du Verney (1638) proposed that the eardrum is *actively* tuned by muscles of the middle ear to match the pitch of incoming tones (he did not say how the tunable eardrum and fixed cochlear resonators might share roles). Most pitch models are of the “fixed” sort, but tuning is possibly an option. Perception often involves some form of action, for example moving one’s head to resolve localization ambiguity. *Efferent* pathways are as ubiquitous within the auditory system as their role is little known (Sahey et al. 1997), and it is conceivable that pitch is extracted according to a tunable version of, say, the AC model. It might be cheaper, in terms of neural circuitry, to have one or more tunable delay/coincidence elements rather than the full array posited by the standard AC model. Tuning might explain the common lack of absolute pitch (absolute pitch would then be explained by the uncommon presence of fixed tuned elements).

To summarize Section 10, specialized issues give insight as to which model of pitch is correct, as simpler phenomena are explained equally well by most models. Special phenomena may sometimes require specialized models, but it should be understood that they all address facets of the same object, the auditory system. Hopefully some day they will merge into a unitary model worthy of Helmholtz.

11 Of Models and Men

This book is about pitch, but the hero of the chapter is the model. Model-making itself is a metaphor of perception. Like the shadows on the back of Plato’s cave, models reflect the world outside (or in our case: inside the ear) in the same way as the pattern of activity on the retina reflects the structure of a scene. Perception guides *action*, and effective action leads to survival of the organism. Reversing the metaphor, a criterion for judging our models is what we *do* with them. For society, the bottom line is to adequately address technical, economical, medical, etc. issues. For the researcher it is to

“publish or perish”. Ultimately, here is the meaning of the word “useful” in our definition.

Over the past, pitch theory has progressed unevenly. Various factors appear to have hastened or slowed the pace. Models are made by people, who are driven by whims and animosities and the need to “survive” scientifically. Ego-involvement (to use Licklider’s words) drives the model-maker to move forward, and also to thwart competition. At times, progress is fueled by the intellectual power of one person, such as Helmholtz. At others, it seems hampered by the authority of that same power. Controversy is stimulating, but it tends to lock opponents into sterile positions that slow their progress (Boring 1929, 1940).

Certain desirable features make a model fragile. A model that is *specific* about its implementation is more likely to be proven false than one that is vague. A model that is unitary or simple is more likely to fail than one that is narrow in scope or rich in parameters. These forces should be compensated, and at times it may be necessary to *protect* a model from criticism. It is my speculation that Helmholtz knew the weakness of his theory in respect to the missing fundamental, but felt it necessary to resist criticism that might have led to its demise. The value and beauty of his monumental bridge across mathematics, physiology and music were such that its flaws were better ignored. To that one must agree. Yet Helmholtz’s theory has cast a long shadow across time, still felt today and not entirely beneficial.

This chapter was built on the assumption that a healthy menagerie of models is desirable. Otherwise, writing sympathetically about them would have been much harder. There are those who believe that theories are not entirely a good thing. Von Békésy and Rosenblith (1948) expressed scorn for them, and stressed instead anatomical investigation (and technical progress in *instrumentation* for that purpose) as a motor of progress. Wever (1949), translator of the model-maker von Békésy, distrusted material and mathematical models. Boring (1926) called out for “fewer theories and more theorizing”. Good theories are falsifiable, and some put their best efforts into falsifying them. If, as Hebb (1959) suggests, every theory is already false by essence, such efforts are guaranteed to succeed. The falsifiability criterion is perhaps less useful than it seems.

On the other hand, progress in science has been largely a process of weeding out theories. The appropriate attitude may be a question of balance, or of a judicious alternation between the two attitudes, as in de Boer’s metaphor of the pendulum. This chapter swings in a model-sympathetic direction, future chapters may more usefully swing the other way.

Inadequate *terminology* is an obstacle to progress. The lack of a word, or worse, the sharing of a word between concepts that should be distinct is a source of fruitless argument. Mersenne was hindered by the need to apply the same word (“fast”) to both vibration rate and propagation speed. Today, “frequency” is associated with spectrum (and thus place theory) in some contexts, and rate (and thus temporal theory) in others. “Spectral pitch” and “residue” are used differently by different authors. We must recognize these obstacles.

Metaphors are useful. Our experience of resonating objects (Du Verney’s steel spring, or Le Cat’s harpsichord) makes the idea of resonance within the ear easy to grasp and convey to others. In this review the

metaphor of the *string* has served to bridge time (from Pythagoras to Helmholtz to today) and theory (from place to autocorrelation). Helmholtz used the *telegraph* to convince himself of the adequacy of his version of Müller's principle, but, had it been invented earlier, the *telephone* might have convinced him otherwise.

A final point has to do with the collective dimension of theory making. Mersenne was known to be impatient with his opponents. In 1634, Nicolas-Claude Fabri de Pieresc warned him: "... you must refrain from putting criticism on others... without urgent necessity, to induce no one to try to bite you in revenge." Mersenne changed radically, became affable and developed an intense correspondence with the best minds of the time. In an age without scientific journals, that did possibly more for the advancement of knowledge than his own discoveries and inventions (Tannery and de Waard 1970).

12 Summary

Historically, theories of pitch were often theories of *hearing*. It is good to keep in mind this wider scope. Pitch determines the survival of a professional musician today, but the ears of our ancestors were shaped for a wider range of tasks. It is conceivable that pitch grew out of a mechanism that evolved for other purposes, for example to segregate sources, or to factor redundancy within an acoustic scene (Hartmann 1996). The "wetware" used for pitch certainly serves other functions, and thus advances in understanding pitch benefit our knowledge of hearing in general.

Ideally, understanding pitch should involve choosing, from a number of plausible mechanisms, the one used by the auditory system, on the basis of available anatomical, physiological or behavioral data. Actually, many schemes reviewed in Sections 2.1 and 2.2 were *functionally* weak. Understanding pitch also involves weeding out those schemes that "do not work", which is all the more difficult as they may seem to work perfectly for certain classes of stimuli. Two schemes (or families of schemes) are functionally adequate: pattern matching and autocorrelation. They are closely related, which is hardly surprising as they both perform the same function: period estimation. For that reason it is hard to choose between them.

My preference goes to the autocorrelation family, and more precisely to cancellation (that uses minima rather than maxima as cues to pitch, Section 9.5). This has little to do with pitch, and more with the fact that cancellation is useful for segregation and fits the ideas on redundancy-reduction of Barlow (1961). I am also, as Licklider put it, "ego involved". Cancellation could be used to measure periods of resolved partials in a pattern-matching model, but the pattern-matching part would still need accounting for. A period-sized delay seems an easy way to implement a harmonic template or sieve. Although the existence of adequate delays is controversial, they are a reasonable requirement compared to other schemes. If a better scheme were found to enforce harmonic relations, I'd readily switch from autocorrelation/cancellation to pattern matching. For now, I try to keep both in my mind as recommended by Licklider.

It is conceivable that the auditory system uses neither. A reason to believe so is that they don't seem to fit with every feature described by the physiologist, the psychoacoustician or the musician. Another is that both models were designed to be simple and easily understood. Obviously the auditory nervous system has no such constraint, so the actual mechanism might be far more complex than we can easily apprehend. Our current models may still be useful as tools to *understand* such a complex mechanism. Judging from yesterday's progress, however, it is wise to assume that yet better tools are to come.

This chapter reviewed models, present and past. Not to write a history, nor to select the best of today's models, but rather to help with the development of *future* models. To quote Flourens (Boring, 1963): 'Science *is* not. It becomes.'

13 Sources

Delightful introductions to pitch theory (unfortunately hard to find) are Schouten (1970) and de Boer (1976). Plomp gives historical reviews on resolvability (Plomp 1964), beats and combination tones (Plomp 1965, 1967b), consonance (Plomp and Levelt 1965), and pitch theory (Plomp 1967a). The early history of acoustics is recounted by Hunt (1990), Lindsay (1966) and Schubert (1978). Important early sources are reproduced in Lindsay (1973), and Schubert (1979). The review of von Békésy and Rosenblith (1948) is oriented towards physiology. Wever (1949) reviews the many early theories of cochlear function, earlier reviewed by Watt (1917), and yet earlier by Bonnier (1896-1898, 1901). Boring (1942) provides an erudite and in-depth review of the history of ideas in hearing and the other senses. Cohen (1984) reviews the progress in musical science in the critical period around 1600. Turner (1977) is a source on the Seebeck/Ohm/Helmholtz dispute. Original sources were consulted whenever possible, otherwise the secondary source is cited. For lack of linguistic competence, sources in German (and Latin for early sources) are missing. This constitutes an important gap.

14 Acknowledgements

I thank the many people who offered ideas, comments or criticism on earlier drafts, in particular Yves Cazals, Laurent Demany, Richard Fay, Bill Hartmann, Stephen McAdams, Ray Meddis, Brian Moore, Andrew Oxenham, Chris Plack, Daniel Pressnitzer, François Raveau. Michael Heinz kindly provided data for Figure 8.

15 References

- Adams JC (1997) Projections from octopus cells of the posteroventral cochlear nucleus to the ventral nucleus of the lateral lemniscus in cat and human. *Aud Neurosci* 3:335-350.
- AFNOR (1977) Recueil des normes françaises de l'acoustique. Tome 1 (vocalulaire), NF S 30-107, Paris: Association Française de Normalisation.
- Akeroyd MA, and Summerfield AQ (2000) A fully-temporal account of the perception of dichotic pitches. *Br J Audiol* 33(2):106-107.
- Anantharaman JN, Krishnamurti AK, and Feth LL (1993) Intensity weighting of average instantaneous frequency as a model of frequency discrimination. *J Acoust Soc Am* 94:723-729.
- ANSI (1973) American national psychoacoustical terminology - S3.20. New York.
- Assmann PF, and Summerfield Q (1990) Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J Acoust Soc Am* 88:680-697.
- Bachem A (1937) Various kinds of absolute pitch. *J Acoust Soc Am* 9:145-151.
- Barlow HB (1961) Possible principles underlying the transformations of sensory messages. In Rosenblith WA (ed) *Sensory Communication*. Cambridge Mass: MIT Press, 217-234.
- Barlow HB (2001) Redundancy reduction revisited. *Network: Comput. Neural Syst.* 12: 241-253.
- von Békésy G, and Rosenblith WA (1948) The early history of hearing - observations and theories. *J Acoust Soc Am* 20:727-748.
- Bilsen FA (1977) Pitch of noise signals: evidence for a "central spectrum". *J Acoust Soc Am* 61:150-161.
- Bilsen FA, and Goldstein JL (1974) Pitch of dichotically delayed noise and its possible spectral basis. *J Acoust Soc Am* 55:292-296.
- de Boer E (1956) On the "residue" in hearing. PhD Thesis
- de Boer E (1976) On the "residue" and auditory pitch perception. In Keidel WD and Neff WD (eds) *Handbook of sensory physiology*, vol V-3. Berlin: Springer-Verlag, 479-583.
- de Boer E (1977) Pitch theories unified. In Evans EF and Wilson JP (eds) *Psychophysics and physiology of hearing*. London: Academic, 323-334.
- Bonnier P (1896-1898) *L'oreille - Physiologie - Les fonctions*. Paris: Masson et fils Gauthier-Villars et fils.
- Bonnier P (1901) *L'audition*. Paris: Octave Doin.
- Boring EG (1926) Auditory theory with special reference to intensity, volume and localization. *Am J Psych* 37:157-188.
- Boring EG (1929) The psychology of controversy. *Psychological Review* 36:97-121 (reproduced in Boring, 1963).
- Boring EG (1942) *Sensation and perception in the history of experimental psychology*. New York: Appleton-Century.
- Boring EG (1963) *History, Psychology and Science* (Edited by R.I. Watson and D.T. Campbell). New York: John Wiley and sons.
- Bower CM (1989) *Fundamentals of Music* (translation of *De Institutione Musica*, Anicius Manlius Severinus Boethius, d524). New Haven: Yale University Press.
- Brown JC, and Puckette MS (1989) Calculation of a "narrowed" autocorrelation function. *J Acoust Soc Am* 85:1595-1601.
- Burns E (1982) A quantal effect of pitch shift? *J Acoust Soc Am* 72:S43.

- Camalet S, Duke T, Jülicher F, and Prost J (2000) Auditory sensitivity provided by self-tuned critical oscillations of hair cells. *Proc Natl Acad Sci* 97:3183–3188.
- Cariani PA (2001) Neural timing nets. *Neural Networks* 14:737-753.
- Cariani PA (2003) Recurrent timing nets for auditory scene analysis. *Proc IEEE IJCNN*, 1575-1580.
- Cariani PA, and Delgutte B (1996a) Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J Neurophysiol* 76:1698-1716.
- Cariani PA, and Delgutte B (1996b) 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. *J Neurophysiol* 76:1717-1734.
- Carlyon RP (1996) Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker. *J Acoust Soc Am* 99
- Carlyon RP (1998a) The effects of resolvability on the encoding of fundamental frequency by the auditory system. In Palmer A, Rees A, Summerfield AQ and Meddis R (eds) *Psychophysical and physiological advances in hearing*. London: Whurr, 246-254.
- Carlyon RP (1998b) Comments on "A unitary model of pitch perception" [*J Acoust Soc Am* 102, 1811-1820 (1997)]. *J Acoust Soc Am* 104:1118-1121.
- Carlyon RP, and Shamma S (2003) An account of monaural phase sensitivity. *J. Acoust. Soc. Am.* 114:333-348.
- Carlyon RP, and Shackleton TM (1994) Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms? *J Acoust Soc Am* 95:3541-3554.
- Carlyon RP, Moore BCJ, and Micheyl C (2000) The effect of modulation rate on the detection of frequency modulation and mistuning of complex tones. *J Acoust Soc Am* 108:304-315.
- Carlyon RP, Demany L, and Deeks J (2001) Temporal pitch perception and the binaural system. *J Acoust Soc Am* 109:686-700.
- Carney LH, Heinz MG, Evilsizer ME, Gilkey RH, and Colburn HS (2002) Auditory phase opponency: a temporal model for masked detection at low frequencies. *Acta Acustica United with Acustica* 88:334-347.
- Cedolin L, and Delgutte B (2004) Representations of the pitch of complex tones in the auditory nerve. In Pressnitzer D, de Cheveigné A, McAdams S and Collet L (eds) *Auditory signal processing: psychophysics, physiology and modeling*. New York: Springer, in press.
- de Cheveigné A (1989) Pitch and the narrowed autocoincidence histogram. *Proc ICMPC, Kyoto*, 67-70.
- de Cheveigné A (1993) Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J Acoust Soc Am* 93:3271-3290.
- de Cheveigné A (1997a) Concurrent vowel identification III: A neural model of harmonic interference cancellation. *J. Acoust. Soc. Am.* 101:2857-2865.
- de Cheveigné A (1997b) Harmonic fusion and pitch shifts of inharmonic partials. *J. Acoust. Soc. Am.* 102:1083-1087.
- de Cheveigné A (1998) Cancellation model of pitch perception. *J Acoust Soc Am* 103:1261-1271.
- de Cheveigné A (1999) Pitch shifts of mistuned partials: a time-domain model. *J Acoust Soc Am* 106:887-897.
- de Cheveigné A (2000) A model of the perceptual asymmetry between peaks and troughs of frequency modulation. *J Acoust Soc Am* 107:2645-2656.
- de Cheveigné A (2001) Correlation Network model of auditory processing. *Proc Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis, Aalborg (Denmark)*.
- de Cheveigné A, and Kawahara H (1999) Multiple period estimation and pitch perception model. *Speech Communication* 27:175-185.

- de Cheveigné A, and Kawahara H (2002) YIN, a fundamental frequency estimator for speech and music. *J Acoust Soc Am* 111:1917-1930.
- Cohen HF (1984) *Quantifying music*. Dordrecht: D. Reidel (Kluwer).
- Cohen MA, Grossberg S, and Wyse LL (1995) A spectral network model of pitch perception. *J Acoust Soc Am* 98:862-879.
- Cramer EM, and Huggins WH (1958) Creation of pitch through binaural interaction. *J Acoust Soc Am* 30:413-417.
- Culling JF (2000) Dichotic pitches as illusions of binaural unmasking. III. The existence region of the Fourcin pitch. *J Acoust Soc Am* 107:2201-2208.
- Culling JF, Marshall D, and Summerfield Q (1998a) Dichotic pitches as illusions of binaural unmasking II: the Fourcin pitch and the Dichotic Repetition Pitch. *J Acoust Soc Am* 103:3525-3539.
- Culling JF, Summerfield Q, and Marshall DH (1998b) Dichotic pitches as illusions of binaural unmasking I: Huggin's pitch and the "Binaural Edge Pitch". *J Acoust Soc Am* 103:3509-3526.
- Dai H, Nguyen Q, Kidd GJ, Feth LL, and Green DM (1996) Phase independence of pitch produced by narrow-band signals. *J Acoust Soc Am* 100:2349-2351.
- Dau T, Püschel D, and Kohlrausch A (1996) A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J Acoust Soc Am* 99:3615-3622.
- Davis H, Silverman SR, and McAuliffe DR (1951) Some observations on pitch and frequency. *J Acoust Soc Am* 23:40-42.
- Delgutte B (1984) Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds. *J Acoust Soc Am* 75:879-886.
- Delgutte B (1996) Physiological models for basic auditory percepts. In Hawkins HL, McMullen TA, Popper AN and Fay RR (eds) *Auditory computation*. New York: Springer-Verlag, 157-220.
- Demany L, and Armand F (1984) The perceptual reality of tone chroma in early infancy. *J Acoust Soc Am* 76:57-66.
- Demany L, and Clément S (1997) The perception of frequency peaks and troughs in wide frequency modulations. IV. Effects of modulation waveform. *J Acoust Soc Am* 102:2935-2944.
- Demany L, and Ramos C (2004) Informational masking and pitch memory: Perceiving a change in a non-perceived tone. *Proc CFA/DAGA*.
- Dooley GJ, and Moore BCJ (1988) Detection of linear frequency glides as a function of frequency and duration. *J Acoust Soc Am* 84:2045-2057.
- Duchez M-E (1989) La notion musicale d'élément <<porteur de forme>>. Approche épistémologique et historique. In McAdams S and Deliège I (eds) *La musique et les sciences cognitives*. Liège: Pierre Mardaga, 285-303.
- Duifhuis H, Willems LF, and Sluyter RJ (1982) Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *J Acoust Soc Am*:1568-1580.
- Durlach NI (1963) Equalization and cancellation theory of binaural masking-level differences. *J Acoust Soc Am* 35:1206-1218.
- Elhilali M, Klein DJ, Fritz JB, Simon JZ, and Shamma SA (2004) The enigma of cortical responses: Slow yet precise. In Pressnitzer D, de Cheveigné A, McAdams S and Collet L (eds) *Auditory signal processing: psychophysics, physiology and modeling*. New York: Springer-Verlag, in press.
- Evans EF (1978) Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons. *Audiology* 17:369-420.
- Evans EF (1986) Cochlear nerve fibre temporal discharge patterns, cochlear frequency selectivity and the dominant region for pitch. In Moore BCJ and Patterson RD (eds) *Auditory frequency selectivity*. Plenum Press, 253-264.
- Fletcher H (1924) The physical criterion for determining the pitch of a musical tone. *Phys Rev* (reprinted in Shubert, 1979, 135-145) 23:427-437.
- Fourier JBJ (1820) *Traité analytique de la chaleur*. Paris: Didot.

- Gábor D (1947) Acoustical quanta and the theory of hearing. *Nature* 159:591-594.
- Galambos R, and Davis H (1943) The response of single auditory-nerve fibers to acoustic stimulation. *J Neurophysiol* 6:39-57.
- Galilei G (1638) *Mathematical discourses concerning two new sciences relating to mechanics and local motion, in four dialogues*. Translated by Weston, London: Hooke (reprinted in Lindsay, 1973, pp 40-61).
- Gerson A, and Goldstein JL (1978) Evidence for a general template in central optimal processing for pitch of complex tones. *J Acoust Soc Am* 63:498-510.
- Gockel H, Moore BCJ, and Carlyon RP (2001) Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *J Acoust Soc Am* 109:701-712.
- Goldstein JL (1970) Aural combination tones. In Plomp R and Smoorenburg GF (eds) *Frequency analysis and periodicity detection in hearing*. Leiden: Sijthoff, 230-247.
- Goldstein JL (1973) An optimum processor theory for the central formation of the pitch of complex tones. *J Acoust Soc Am* 54:1496-1516.
- Goldstein JL, and Srulovicz P (1977) Auditory-nerve spike intervals as an adequate basis for aural frequency measurement. In Evans EF and Wilson JP (eds) *Psychophysics and physiology of hearing*. London: Academic Press, 337-347.
- Goldstein JG, and Oxenham A (2003) Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? *J. Acoust. Soc. Am.* 113:3323-3334.
- Gray AA (1900) On a modification of the Helmholtz theory of hearing. *J Anat Physiol* 34:324-350.
- Grimault N, Micheyl C, Carlyon RP, Arthaud P, and Collet L (2000) Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency. *J Acoust Soc Am* 108:263-271.
- Grose JH, Hall JW, III, and Buss E (2002) Virtual pitch integration for asynchronous harmonics. *J Acoust Soc Am* 112:2956-2961.
- Hall JW, III, and Peters RW (1981) Change in the pitch of a complex tone following its association with a second complex tone. *J Acoust Soc Am* 71:142-146.
- Hartmann WM (1996) Pitch, periodicity, and auditory organization. *J. Acoust. Soc. Am.* 100:3491-3502.
- Hartmann WM (1997) *Signals, sound and sensation*. Woodbury, N.Y.: AIP.
- Hartmann WM, and Klein MA (1980) Theory of frequency modulation detection for low modulation frequencies. *J Acoust Soc Am* 67:935-946.
- Hartmann WM (1993) On the origin of the enlarged melodic octave. *J Acoust Soc Am* 93:3400-3409.
- Hartmann WM, and Doty SL (1996) On the pitches of the components of a complex tone. *J Acoust Soc Am* 99:567-578.
- Haykin S (1999) *Neural networks, a comprehensive foundation*. Upper Saddle River, New Jersey: Prentice Hall.
- Hebb DO (1949) *The organization of behavior*. New York: Wiley.
- Hebb, DO (1959) A neuropsychological theory. In S. Koch (ed) *Psychology, a study of a science*. New York: McGraw-Hill, vol I, pp. 622-643.
- Heinz MG, Colburn HS, and Carney LH (2001) Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve. *Neural Computation* 13:2273-2316.
- Hess W (1983) *Pitch determination of speech signals*. Berlin: Springer-Verlag.
- von Helmholtz H (1857, translated by A.J. Ellis, reprinted in Warren & Warren 1968) On the physiological causes of harmony in music. pp 25-60.
- von Helmholtz H (1877) *On the sensations of tone* (English translation A.J. Ellis, 1885, 1954). New York: Dover.
- Hermes DJ (1988) Measurement of pitch by subharmonic summation. *J Acoust Soc Am* 83:257-264.

- Hewitt MJ, Meddis R, and Shackleton TM (1992) A computer model of a cochlear nucleus stellate cell. Responses to amplitude-modulated and pure-tone stimuli. *J Acoust Soc Am* 91:2096-2109.
- Hewitt MJ, and Meddis R (1994) A computer model of amplitude-modulation sensitivity of single units in the inferior colliculus. *J Acoust Soc Am* 95:2145-2159.
- Hirsh I (1948) The influence of interaural phase on interaural summation and inhibition. *J Acoust Soc Am* 20:536-544.
- Hounshell DA (1976) Bell and Gray: contrasts in style, politics and etiquette. *Proc IEEE* 64:1305-1314.
- Houtsma AJM, and Goldstein JL (1972) The central origin of the pitch of complex tones. Evidence from musical interval recognition. *J Acoust Soc Am* 51:520-529.
- Houtsma AJM, and Smurzynski J (1990) Pitch identification and discrimination for complex tones with many harmonics. *J Acoust Soc Am* 87:304-310.
- Huggins WH, and Licklider JCR (1951) Place mechanisms of auditory frequency analysis. *J Acoust Soc Am* 23:290-299.
- Hunt FV (1992, original: 1978) *Origins in acoustics*. Woodbury, New York: Acoustical Society of America.
- Hurst CH (1895) A new theory of hearing. *Proc Trans Liverpool Biol Soc* 9:321-353 (and plate XX).
- Jeffress LA (1948) A place theory of sound localization. *J Comp Physiol Psychol* 41:35-39.
- Jenkins RA (1961) Perception of pitch, timbre and loudness. *J Acoust Soc Am* 33:1550-1557.
- Johnson DH (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *J Acoust Soc Am* 68:1115-1122.
- Joris PX, Smith PH, and Yin TCT (1998) Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron* 21:1235-1238.
- Joris PX (2001) Sensitivity of inferior colliculus neurons to interaural time differences of broadband signals: comparison with auditory nerve firing. In Breebaart DJ, Houtsma AJM, Kohlrausch A, Prijs VF and Schoonhoven R (eds) *Physiological and psychophysical bases of auditory function*. Maastricht: Shaker BV, 177-183.
- Kaernbach C, and Demany L (1998) Psychophysical evidence against the autocorrelation theory of pitch perception. *J Acoust Soc Am* 104:2298-2306.
- Köppl C (1997) Phase locking to high frequencies in the auditory nerve and cochlear nucleus magnocellularis of the barn owl *Tyto alba*. *J Neurosci* 17:3312-3321.
- Langner G (1981) Neuronal mechanisms for pitch analysis in the time domain. *Exp Brain Res* 44:450-454.
- Langner G, and Schreiner CE (1988) Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J Neurophysiol.* 60:1799-1822.
- Le Cat C-N (1758) *La Théorie de l'ouïe: supplément à cet article du traité des sens*. Paris: Vallat-la-Chapelle.
- Licklider JCR (1948) The influence of interaural phase relations upon the masking of speech by white noise. *J Acoust Soc Am* 20:150-159.
- Licklider JCR (1951) A duplex theory of pitch perception (reproduced in Schubert 1979, 155-160). *Experientia* 7:128-134.
- Licklider, JCR (1959) Three auditory theories. In S. Koch (ed) *Psychology, a study of a science*. New York: McGraw-Hill, I, pp. 41-144.
- Lindsay RB (1966) The story of acoustics. *J Acoust Soc Am* 39:629-644.
- Lindsay RB (1973) *Acoustics: historical and philosophical development*. Stroudsburg: Dowden, Hutchinson and Ross.
- Loeb GE, White MW, and Merzenich MM (1983) Spatial cross-correlation - A proposed mechanism for acoustic pitch perception. *Biol Cybern* 47:149-163.

- Lyon R (1984) Computational models of neural auditory processing. Proc IEEE ICASSP, 36.1.(1-4).
- Maass W (1998) On the role of time and space in neural computation. Lecture notes in computer science 1450:72-83.
- Maass W, Natschläger T, and Markram H (2003) Computation models for generic cortical microcircuits. In J. Feng (ed) Computational Neuroscience: A Comprehensive Approach. CRC-Press, to appear.
- Macran HS (1902) The harmonics of Aristoxenus. Oxford: The Clarendon Press (reprinted 1990, Georg Olms Verlag, Hildesheim)
- Marozeau J, de Cheveigné A, McAdams S, and Winsberg S (2003) The dependency of timbre on fundamental frequency. J. Acoust. Soc. Am. 114:2946-2957.
- Martens JP (1981) Comment on "Algorithm for extraction of pitch and pitch salience from complex tonal signals" [J. Acoust. Soc. Am. 71, 679-688 (1982)]. J. Acoust. Soc. Am. 75:626-628.
- McAlpine D, Jiang D, and Palmer A (2001) A neural code for low-frequency sound localization in mammals. Nature Neuroscience 4:396-401.
- McKay CM, and Carlyon RP (1999) Dual temporal pitch percepts from acoustic and electric amplitude-modulated pulse trains. 105:347-357.
- Meddis R (1988) Simulation of auditory-neural transduction: further studies. J Acoust Soc Am 83:1056-1063.
- Meddis R, and Hewitt MJ (1991a) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. J Acoust Soc Am 89:2866-2882.
- Meddis R, and Hewitt MJ (1991b) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: phase sensitivity. J. Acoust. Soc. Am. 89:2883-2894.
- Meddis R, and Hewitt MJ (1992) Modeling the identification of concurrent vowels with different fundamental frequencies. J Acoust Soc Am 91:233-245.
- Meddis R, and O'Mard L (1997) A unitary model of pitch perception. J Acoust Soc Am 102:1811-1820.
- Mersenne M (1636) Harmonie Universelle. Paris: Cramoisy (reprinted 1975, Paris: Editions du CNRS).
- Micheyl C, and Carlyon RP (1998) Effects of temporal fringes on fundamental-frequency discrimination. J Acoust Soc Am 104:3006-3018.
- Miyazaki K (1990) The speed of musical pitch identification by absolute-pitch possessors. Music Perception 8:177-188.
- Moore BCJ (1973) Frequency difference limens for short-duration tones. J Acoust Soc Am 54:610-619.
- Moore BCJ (1977) An introduction to the psychology of hearing. London: Academic Press (first edition).
- Moore BCJ (2003) An introduction to the psychology of hearing. London: Academic Press (fifth edition).
- Moore BCJ, and Sek A (1994) Effects of carrier frequency and background noise on the detection of mixed modulation. J. Acoust. Soc. Am. 96:741-751.
- Nelken I, Ulanovsky N, Las L, Bar-Yosef O, Anderson M, Chechik G, Tishby N, and Young E (2004) Transformation of stimulus representations in the ascending auditory system. In Pressnitzer D, de Cheveigné A, McAdams S and Collet L (eds) Auditory signal processing: psychophysics, physiology and modeling. New York: Springer, in press.
- Newman EB, Stevens SS, and Davis H (1937) Factors in the production of aural harmonics and combination tones. J Acoust Soc Am 9:107-118.
- Noll AM (1967) Cepstrum pitch determination. J Acoust Soc Am 41:293-309.
- van Noorden L (1982) Two channel pitch perception. In Clynes M (ed) Music, mind, and brain. London: Plenum press, 251-269.
- Nordmark J (1963) Some analogies between pitch and lateralization phenomena. J Acoust Soc Am 35:1544-1547.

- Nordmark JO (1968) Mechanisms of frequency discrimination. *J Acoust Soc Am* 44:1533-1540.
- Nordmark JO (1970) Time and frequency analysis. In Tobias JV (ed) *Foundations of modern auditory theory*. New York: Academic Press, 55-83.
- Ohgushi K (1978) On the role of spatial and temporal cues in the perception of the pitch of complex tones. *J Acoust Soc Am* 64:764-771.
- Ohm GS (1843) On the definition of a tone with the associated theory of the siren and similar sound producing devices. *Poggendorf's Annalen der Physik und Chemie* 59:497ff (translated and reprinted in Lindsay, 1973, pp 242-247).
- Okada M, and Kashino M (2003) The role of spectral change detectors in temporal order judgment of tones. *Neuroreport* 14
- Oxenham A, Bernstein LR, and Micheyl C (2004) Pitch perception of complex tones within and across ears and frequency regions. In Pressnitzer D, de Cheveigné A, McAdams S and Collet L (eds) *Auditory signal processing: physiology, psychophysics and modeling*. Springer-Verlag, in press.
- Parncutt R (1988) Revision of Terhardt's psychoacoustical model of the roots of a musical chord. *Music Perception* 6:65-94.
- Parsons TW (1976) Separation of speech from interfering speech by means of harmonic selection. *J Acoust Soc Am* 60:911-918.
- Patterson RD (1987) A pulse ribbon model of monaural phase perception. *J Acoust Soc Am* 82:1560-1586.
- Patterson RD (1994a) The sound of a sinusoid: time-domain models. *J Acoust Soc Am* 96:1419-1428.
- Patterson RD (1994b) The sound of a sinusoid: spectral models. *J Acoust Soc Am* 96:1409-1418.
- Patterson RD, and Nimmo-Smith I (1986) Thinning periodicity detectors for modulated pulse streams. In Moore BCJ and Patterson RD (eds) *Auditory frequency selectivity*. New York: Plenum Press, 299-307.
- Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, and Allerhand M (1992) Complex sounds and auditory images. In Cazals Y, Horner K and Demany L (eds) *Auditory physiology and perception*. Oxford: Pergamon Press, 429-446.
- Plack CJ, and Carlyon RP (1995) Differences in frequency detection and fundamental frequency discrimination between complex tones consisting of resolved and unresolved harmonics. *J Acoust Soc Am* 98:1355-1364.
- Plack CJ, and White LJ (2000a) Perceived continuity and pitch perception. *J Acoust Soc Am* 108:1162-1169.
- Plack CJ, and White LJ (2000b) Pitch matches between unresolved complex tones differing by a single interpulse interval. *J Acoust Soc Am* 108:696-705.
- Plomp R (1964) The ear as a frequency analyzer. *J Acoust Soc Am* 36:1628-1636.
- Plomp R (1965) Detectability threshold for combination tones. *J Acoust Soc Am* 37:1110-1123.
- Plomp R (1970) Timbre as a multidimensional attribute of complex tones. In Plomp R and Smoorenburg GF (eds) *Frequency analysis and periodicity detection in hearing*. Leiden: Sijthoff, 397-414.
- Plomp R (1967a) Pitch of complex tones. *J Acoust Soc Am* 41:1526-1533.
- Plomp R (1967b) Beats of mistuned consonances. *J Acoust Soc Am* 42:462-474.
- Plomp R (1976) *Aspects of tone sensation*. London: Academic Press.
- Plomp R, and Levelt WJM (1965) Tonal consonance and critical bandwidth. *J Acoust Soc Am* 38:545-560.
- Pressnitzer D, and Patterson RD (2001) Distortion products and the pitch of harmonic complex tones. In Breebaart DJ, Houtsma AJM, Kohlrausch A, Prijs VF and Schoonhoven R (eds) *Physiological and psychophysical bases of auditory function*. Maastricht: Shaker, 97-104.
- Pressnitzer D, Patterson RD, and Krumbholz K (2001) The lower limit of melodic pitch. *J Acoust Soc Am* 109:2074-2084.

- Pressnitzer D, Winter IM, and de Cheveigné A (2002) Perceptual pitch shift for sounds with similar waveform autocorrelation. *Acoustic Research Letters Online* 3:1-6.
- Pressnitzer D, de Cheveigné A, and Winter IM (2003) Physiological correlates of the perceptual pitch shift of sounds with similar waveform autocorrelation. *Acoustic Research Letters Online*, 5:1-6.
- Raatgever J, and Bilsen FA (1986) A central spectrum model of binaural processing. Evidence from dichotic pitch. *J Acoust Soc Am* 80:429-441.
- Rameau J-P (1750) *Démonstration du principe de l'harmonie*, Paris: Durand (reproduced in "E.R. Jacobi (1968) Jean-Philippe Rameau, Complete theoretical writings, V3, American Institute of Musicology, 154-254).
- Lord Rayleigh (1896) *The theory of sound* (second edition, 1945 re-issue, New York: Dover).
- Ritsma RJ (1967) Frequencies dominant in the perception of the pitch of complex tones. *J Acoust Soc Am* 42:191-198.
- Roederer JG (1975) *Introduction to the physics and psychophysics of music*. New York: Springer Verlag.
- Rose JE, Brugge JF, Anderson DJ, and Hind JE (1967) Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *J Neurophysiol* 30:769-793.
- Ross MJ, Shaffer HL, Cohen A, Freudberg R, and Manley HJ (1974) Average magnitude difference function pitch extractor. *IEEE Trans. ASSP* 22:353-362.
- Ruggero MA (1973) Response to noise of auditory nerve fibers in the squirrel monkey. *J Neurophysiol* 36:569-587.
- Ruggero MA (1992) Physiology of the auditory nerve. In Popper AN and Fay RR (eds) *The mammalian auditory pathway: neurophysiology*. New York: Springer Verlag, 34-93.
- Rutherford E (1886) A new theory of hearing. *J Anat Physiol* 21:166-168.
- Sabine WC (1907) Melody and the origin of the musical scale. In Hunt FV (ed) *Collected papers on acoustics by Wallace Clement Sabine* (1964). New York: Dover, 107-116.
- Sachs MB, and Young ED (1979) Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. *J Acoust Soc Am* 66:470-479.
- Sahey TL, Nodar RH, and Musiek FE (1997) *Efferent auditory system*. San Diego: Singular.
- Sauveur J (1701) *Système général des intervalles du son*, *Mémoires de l'Académie Royale des Sciences* 279-300:347-354. (translated and reprinted in Lindsay, 1973, pp 88-94).
- Scheffers MTM (1983) *Sifting vowels*. PhD Thesis Groningen.
- Schouten JF (1938) The perception of subjective tones. *Proc Kon Acad Wetensch (Neth.)* 41:1086-1094 (reprinted in Schubert 1979, 146-154).
- Schouten JF (1940a) The residue, a new component in subjective sound analysis. *Proc Kon Acad Wetensch (Neth.)* 43:356-356.
- Schouten JF (1940b) The residue and the mechanism of hearing. *Proc Kon Acad Wetensch (Neth.)* 43:991-999.
- Schouten JF (1940c) The perception of pitch. *Philips technical review* 5:286-294.
- Schouten JF (1970) The residue revisited. In Plomp R and Smoorenburg GF (eds) *Frequency analysis and periodicity detection in hearing*. London: Sijthoff, 41-58.
- Schouten JF, Ritsma RJ, and Cardozo BL (1962) Pitch of the residue. *J Acoust Soc Am* 34:1418-1424.
- Schroeder MR (1968) Period histogram and product spectrum: new methods for fundamental-frequency measurement. *J Acoust Soc Am* 43:829-834.
- Schubert ED (1978) History of research on hearing. In Carterette EC and Friedman MP (eds) *Handbook of perception*. New York: Academic Press, IV, pp. 41-80.

- Schubert ED (1979) *Psychological acoustics (Benchmark papers in Acoustics, v 13)*. Stroudsburg, Pennsylvania: Dowden, Hutchinson & Ross, Inc.
- Sek A, and Moore BCJ (1999) Discrimination of frequency steps linked by glides of various durations. *J Acoust Soc Am* 106:351-359.
- Semal C, and Demany L (1990) The upper limit of musical pitch. *Music Perception* 8:165-176.
- Shackleton TM, and Carlyon RP (1994) The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *J Acoust Soc Am* 95:3529-3540.
- Shamma SA (1985) Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *J Acoust Soc Am* 78:1622-1632.
- Shamma SA, Shen N, and Gopalaswamy P (1989) Stereausis: binaural processing without neural delays. *J Acoust Soc Am* 86:989-1006.
- Shamma S, and Klein D (2000) The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. *J Acoust Soc Am* 107:2631-2644.
- Shera CA, Guinan JJ, and Oxenham AJ (2002) Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proc Natl Acad Sci USA* 99:3318-3323.
- Siebert WM (1968) Stimulus transformations in the auditory system. In Kolars PA and Eden M (eds) *Recognizing patterns*. Cambridge Mass: MIT Press, 104-133.
- Siebert WM (1970) Frequency discrimination in the auditory system: place or periodicity mechanisms. *Proc IEEE* 58:723-730.
- Slaney M (1990) A perceptual pitch detector. *Proc ICASSP*, 357-360.
- Smooenburg GF (1970) Pitch perception of two-frequency stimuli. *J Acoust Soc Am* 48:924-942.
- Srulovicz P, and Goldstein JL (1983) A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *J Acoust Soc Am* 73:1266-1276.
- Tannery M-P, and de Waard C (1970) *Correspondance du P. Marin Mersenne*, vol. XI (1642). Paris: Editions du CNRS.
- Tasaki I (1954) Nerve impulses in individual auditory nerve fibers of guinea pig. *J Neurophysiol* 17:97-122.
- Terhardt E (1974) Pitch, consonance and harmony. *J Acoust Soc Am* 55:1061-1069.
- Terhardt E (1978) Psychoacoustic evaluation of musical sounds. *Percept & Psychophys* 23:483-492.
- Terhardt E (1979) Calculating virtual pitch. *Hearing Research* 1:155-182.
- Terhardt E (1991) Music perception and sensory information acquisition: relationships and low-level analogies. *Music Perception* 8:217-240.
- Terhardt E, Stoll G, and Seewann M (1982) Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J Acoust Soc Am* 71:679-688.
- Thompson SP (1882) On the function of the two ears in the perception of space. *Phil Mag (S5)* 13:406-416.
- Thorpe S, Fize F, and Marlot C (1996) Speed of processing in the human visual system. *Nature* 381:520-522.
- Thurlow WR (1963) Perception of low auditory pitch: a multicue mediation theory. *Psychol Rev* 70:461-470.
- Tong YC, Blamey PJ, Dowell RC, and Clark GM (1983) Psychophysical studies evaluating the feasibility of speech processing strategy for a multichannel cochlear implant. *J Acoust Soc Am* 74:73-80.
- Troland LT (1930) Psychophysiological considerations related to the theory of hearing. *J Acoust Soc Am* 1:301-310.
- Turner RS (1977) The Ohm-Seebeck dispute, Hermann von Helmholtz, and the origins of physiological acoustics. *Brit J Hist Sci* 10:1-24.

- Du Verney JG (1683) *Traité de l'organe de l'ouïe, contenant la structure, les usages et les maladies de toutes les parties de l'oreille*. Paris.
- Versnel H, and Shamma S (1998) Spectral-ripple representation of steady-state vowels. *J Acoust Soc Am* 103:5502-2514.
- Ward WD (1999) Absolute pitch. In Deutsch D (ed) *The psychology of music*. Orlando: Academic press, 265-298.
- Warren RM, and Warren RP (1968) *Helmholtz on perception: its physiology and development*. New York: Wiley.
- Warren JD, Uppenkamp S, Patterson RD, and Griffith TD (2003) Separating pitch chroma and pitch height in the human brain. *Proc Natl Acad Sci US* 100:10038-19942.
- Watt HJ (1917) *The psychology of sound*. Cambridge: The University Press.
- Wegel RL, and Lane CE (1924) The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear. *Phys Rev* 23:266-285 (reproduced in Schubert 1979, 201-211).
- Weintraub M (1985) A theory and computational model of auditory monaural sound separation. PhD Thesis, Stanford.
- Wever EG, and Bray CW (1930) The nature of acoustic response: the relation between sound frequency and frequency of impulses in the auditory nerve. *Journal of experimental psychology* 13:373-387.
- Wever EG (1949) *Theory of hearing*. New York: Dover.
- Whitfield IC (1970) Central nervous processing in relation to spatio-temporal discrimination of auditory patterns. Plomp R and Smoorenburg GF (eds) *Frequency analysis and periodicity detection in hearing*. Leiden: Sijthoff, 136-152.
- Wiegrebe L, Patterson RD, Demany L, and Carlyon RP (1998) Temporal dynamics of pitch strength in regular interval noises. *J Acoust Soc Am* 104:2307-2313.
- Wiegrebe L (2001) Searching for the time constant of neural pitch integration. *J Acoust Soc Am* 109:1082-1091.
- Wiegrebe L, Stein A, and Meddis R (2004) Coding of pitch and amplitude modulation in the auditory brainstem: One common mechanism? In Pressnitzer D, de Cheveigné A, McAdams S and Collet L (eds) *Auditory signal processing: psychophysics, physiology and modeling*. New York: Springer, in press.
- Wiegrebe L, and Meddis R (2004) The representation of periodic sounds in simulated sustained chopper units of the ventral cochlear nucleus. *J. Acoust. Soc. Am.*, in press
- Wightman FL (1973) The pattern-transformation model of pitch. *J Acoust Soc Am* 54:407-416.
- Yost WA (1996) Pitch strength of iterated rippled noise. *J Acoust Soc Am* 100:3329-3335.
- Young T (1800) Outlines of experiments and inquiries respecting sound and light. *Phil Trans of the Royal Society of London* 90:106-150 (and plates).
- Young ED, and Sachs MB (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *J Acoust Soc Am* 66:1381-1403.
- Zwicker E (1970) Masking and psychoacoustical excitation as consequences of the ear's frequency analysis. In Plomp R and Smoorenburg GF (eds) *Frequency analysis and periodicity detection in hearing*. Leiden: Sijthoff, 376-396.

Figure 1. Spectral approach. (A) to (E) are schematized spectra of pitch-evoking stimuli; (F) is the subharmonic histogram of the spectrum in (E). Choosing the *peak* in the spectrum reveals the pitch in (A) but not in (B) where there are several peaks. Choosing the *largest* peak works in (B) but fails in (C). Choosing the peak with *lowest frequency* works in (C) but fails in (D). Choosing the *spacing* between peaks works in (D) but fails in (E). A *pattern-matching* scheme (F) works with all stimuli. The cue to pitch here is the rightmost among the largest bins (bold line).

Figure 2. Temporal approach. (A) to (E) are waveform samples of pitch-evoking stimuli. (F) is the autocorrelation function of the waveform in (E). Taking the interval between *successive* peaks (arrows) works in (A) but fails in (B). The interval between *highest* peaks works in (B) but fails in (C). The interval between positive-going *zero-crossings* works in (C) but fails in (D) where there are several zero-crossings per period. The *envelope* works in (D), but fails in (E). A scheme based on the *autocorrelation* function (F) works for all stimuli. The leftmost of the (infinite) series of main peaks (dark arrows) indicates the period. Stimuli such as (E) tend to be ambiguous and may evoke pitches corresponding to the gray arrows instead of (or in addition to) the pitch corresponding to the period.

Figure 3. Johannes Müller built this model of the middle ear to convince himself that sound is transmitted from the ear drum (c) via the ossicular chain (g) to the oval window (f), rather than by air to the round window (e) as was previously thought. The model is obviously “false” (the ossicular chain is not a piece of wire) but it allowed an important advance in understanding hearing mechanisms (Müller 1838; von Békésy and Rosenblith 1948).

Figure 4. Descriptions of pitch-evoking stimuli. (A): Periodic waveform. The parameters of the description are T and the values of the stimulus during one period: $s(t)$, $0 < t \leq T$. (B): Sinusoidal waveform. The parameterization (f , A and ϕ) is simpler, but the description fits a smaller class of stimuli (pure tones). (C): Amplitude spectrum of the signal in (A). Together with phase (not shown) this provides an alternative parameterization of the stimulus in (A). (D) Waveform of a formant-like periodic stimulus. (E): Spectrum of the same stimulus. This stimulus may evoke a pitch related to F_0 , or to f_{LOCUS} , or both.

Figure 5. Formant-like stimuli may evoke two pitches, periodicity and spectral, that map to F_0 and f_{LOCUS} stimulus dimensions respectively. The parameter space includes only the region below the diagonal, and stimuli that fall outside the closed region do not evoke a periodicity pitch with a musical nature (Semal and Demany 1990; Pressnitzer et al. 2001). For pure tones (diagonal) periodicity and spectral pitch co-vary. Insert: autocorrelation function of a formant-like stimulus.

Figure 6. Monochord. A string is stretched between two fixed bridges A,B on a sounding board. A movable bridge C is placed at an intermediate position in such a way that the tension on both sides is equal. The pitches form a consonant interval if the lengths of segments AC and CB are in a simple ratio. The *string* plays an important role as model and metaphor in the history of pitch.

Figure 7. (A): Partials that excite a string tuned to 440 Hz. (B): Strings that respond to a 440 Hz pure tone (the abscissa of each pulse represents the frequency of the lowest mode of the string). (C): Strings that respond to a 440 Hz complex tone. Pulses are scaled in proportion to the power of the response. The *rightmost* string with a full response indicates the period. The string is selective to periodicity rather than Fourier frequency.

Figure 8. Pure tone frequency discrimination by humans and models, replotted from Heinz et al (2001). Open triangles: threshold for a 200 ms pure tone with equal loudness as a function of frequency (Moore, 1973). Circles: predictions of place-only models. Squares: predictions of time-only models. Open circles and squares are for Siebert's (1970) analytical model, closed circles and squares are for Heinz et al's (2001) computational model.

Figure 9. (A): Stimulus consisting of odd harmonics 3, 5, 7, and 9. (B): Difference function $d(\tau)$. (C): AC function $r(\tau)$. (D): Array of ACFs as in Licklider's model. (E): Summary ACF as in Meddis and Hewitt's model. Vertical dotted lines indicate the position of the period cue. Note that the partials are resolved and form well-separated horizontal bands in (D). Each band shows the period of a *partial*, yet their sum (E) shows the fundamental period.

Figure 10. Processing involved in various pitch models. (A) Autocorrelation involves *multiplication*. (B) Cancellation involves *subtraction*. (C) The feed-forward comb-filter (Delgutte 1984) involves *addition*. (D) In the feedback comb-filter, the *delayed output* is added to the input (after attenuation), rather than the delayed input. This circuit behaves like a string. Plots on the right show, as a function of frequency, the value measured at the output for a pure-tone input. For a frequency inverse of the delay, and all of its harmonics, the product (A) is maximum, the difference (B) is minimum, the sum (C) is maximum. Tuning is sharper for the feedback comb-filter (D).

Figure 11. SACFs in response to a 200 Hz pure tone. The abscissa is logarithmic and covers roughly the range of periods that evoke a musical pitch (0.2 to 30 ms). The pitch mechanism must choose the mode that indicates the period (dark arrow in A) and reject the others (gray arrows). This may be done by setting lower and upper limits on the period range (B), or a lower limit and a bias to favor shorter lags. (C). The latter solution may fail if the period mode is less salient than the portion of the zero-lag mode that falls within the search range (D).

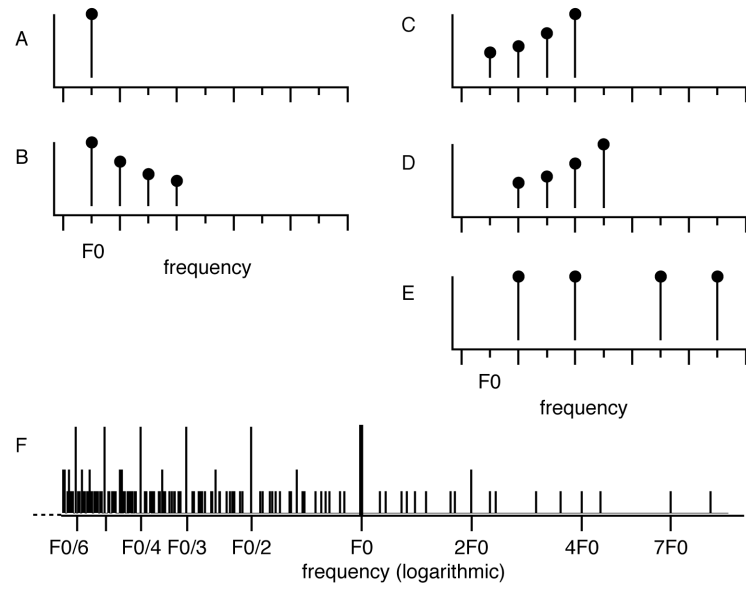


Figure 1.

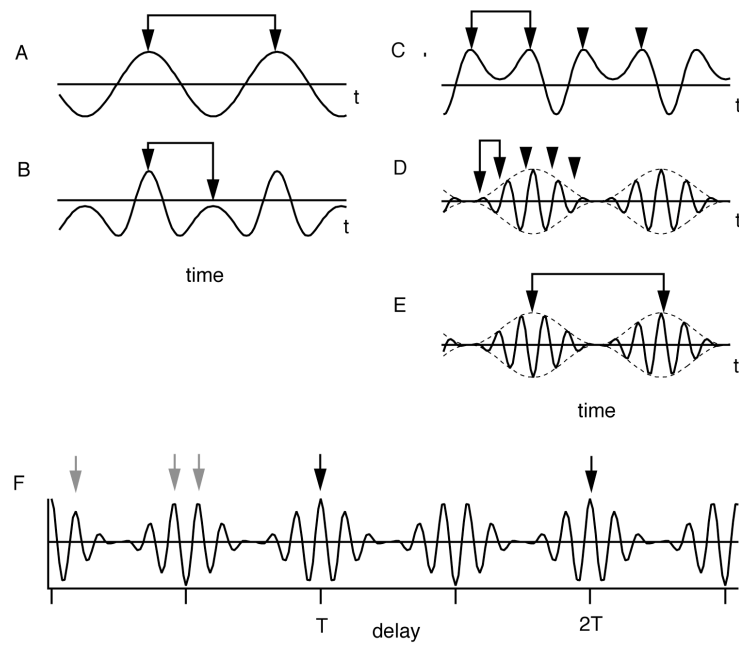


Figure 2.

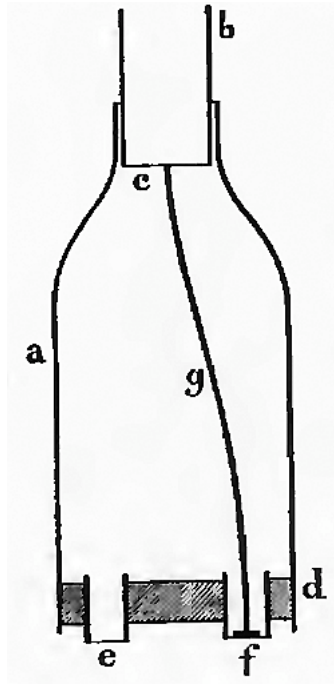


Figure 3.

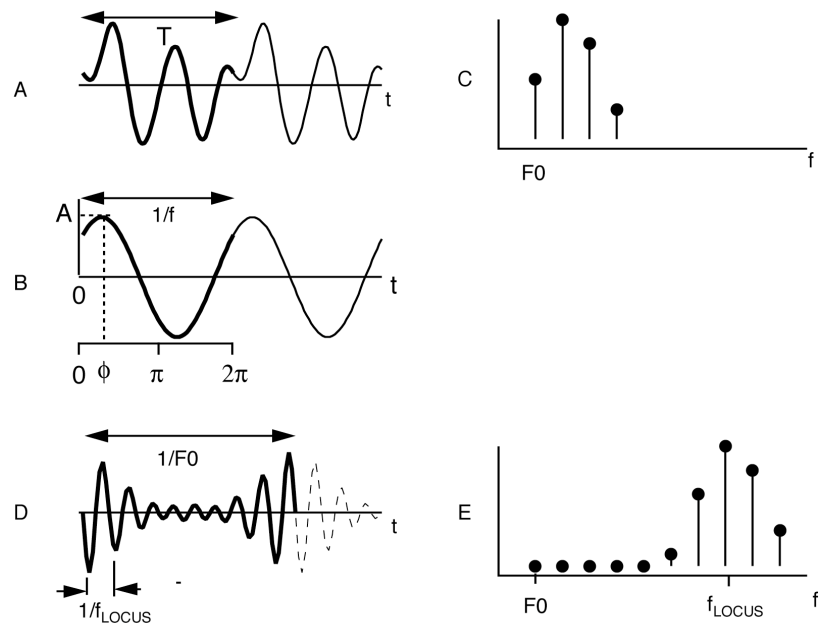


Figure 4.

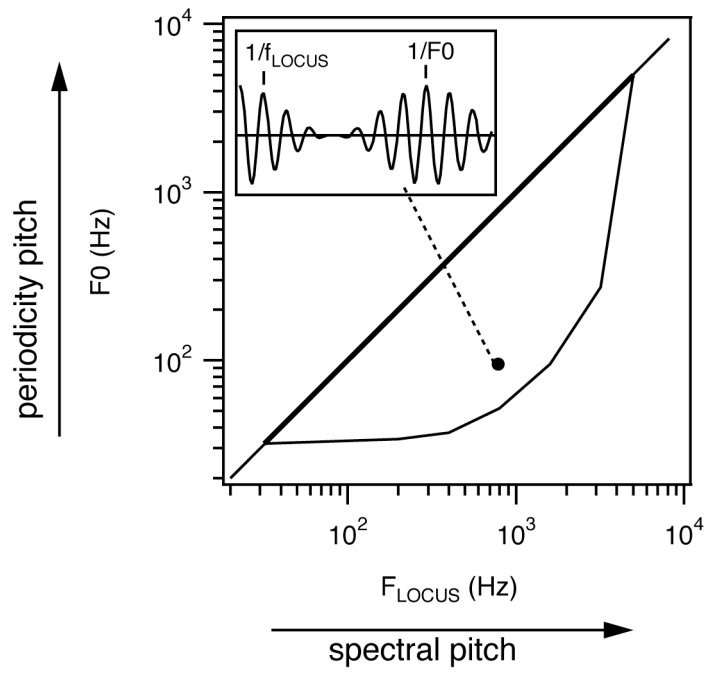


Figure 5.

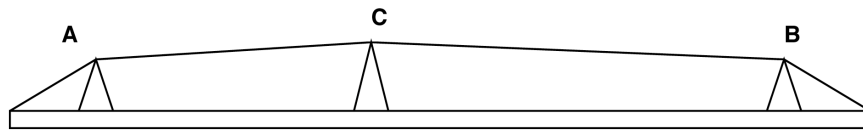


Figure 6.

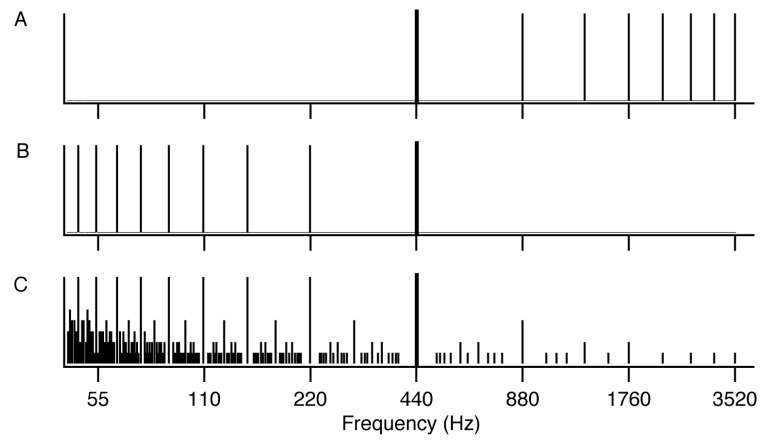


Figure 7.

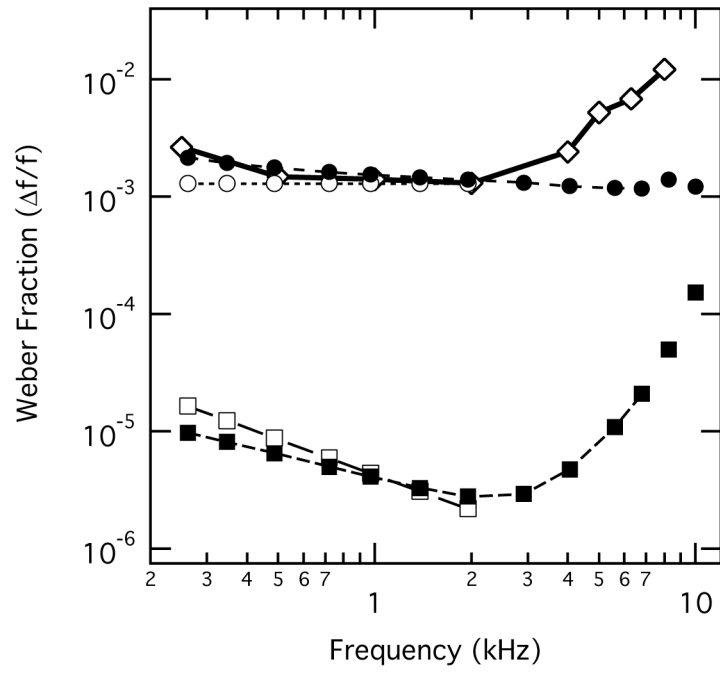


Figure 8.

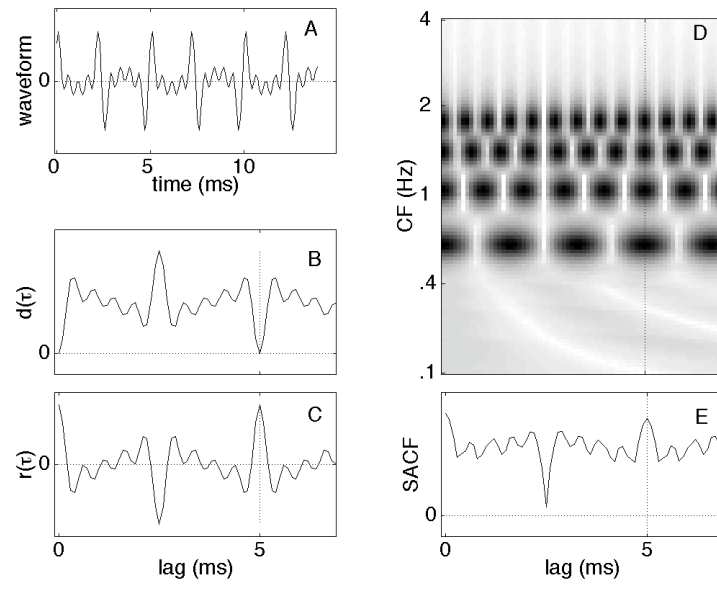


Figure 9.

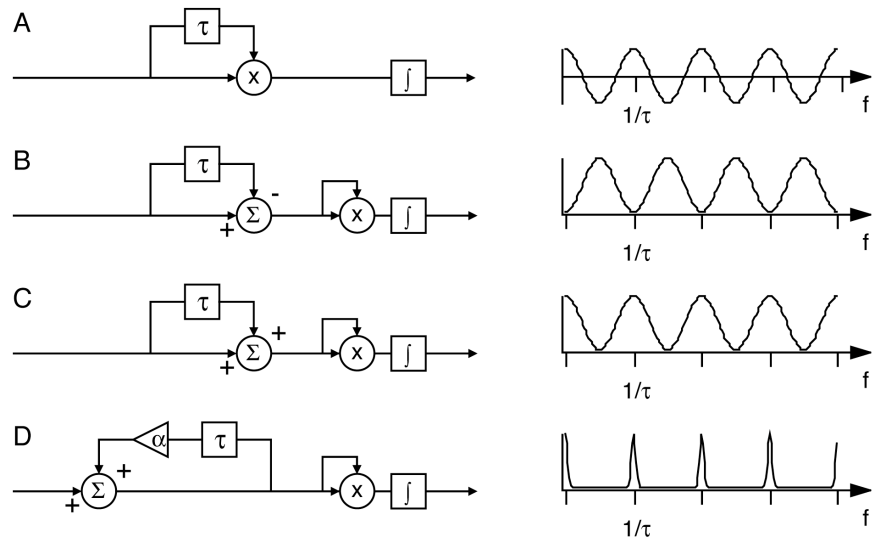


Figure 10.

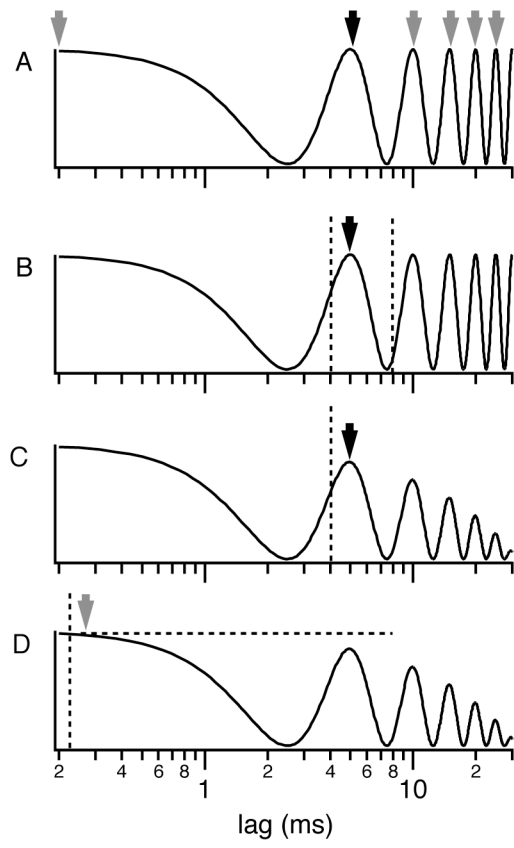


Figure 11.