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, to build a catalog of ideas of use for understanding of pitch perception. The busy reader might read just the next section. It explains in a nutshell the problem, why some obvious ideas don't work, and what are currently the best answers. The brave reader will read on as we delve deeper into the origin of concepts, and the intricate and ingenious ideas behind the models and metaphors upon which our understanding of pitch makes progress.

2 Pitch theory in a nutshell

Pitch-evoking stimuli are usually periodic, and the pitch usually follows the period. Accordingly, a pitch perception mechanism must estimate the period $\boldsymbol{T}$ (or its inverse, the fundamental frequency $\boldsymbol{f_0}$) of the stimulus. There are two approaches to do so. One involves the *spectrum* and the other the *waveform*.

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 is to look for a 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 eventual *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 is the same as that of the other tones.

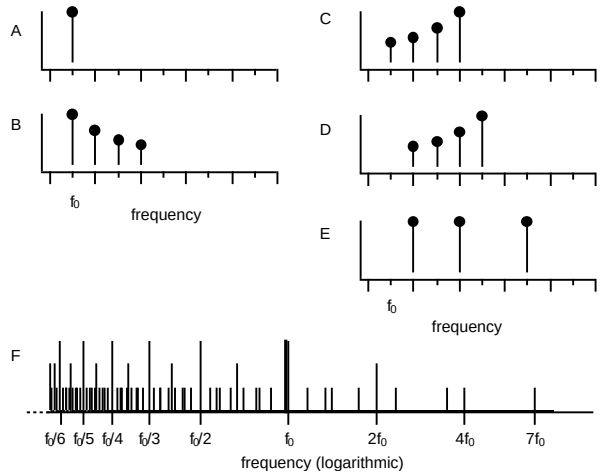

Figure 1. Spectral approach. (A-E): Schematized spectra. (F): Subharmonic histogram of spectrum in (E). Choosing the *peak* in the spectrum reveals the pitch in (A) but not in (B) where there are several peaks. Choosing the *largest* peak works in (B) but fails in (C). Choosing the peak with *lowest 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 is the rightmost among the largest bins (bold line).

This brings us to a final algorithm. Build a histogram in the following way: for each partial, find its subharmonics by dividing its frequency 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 criterion 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.

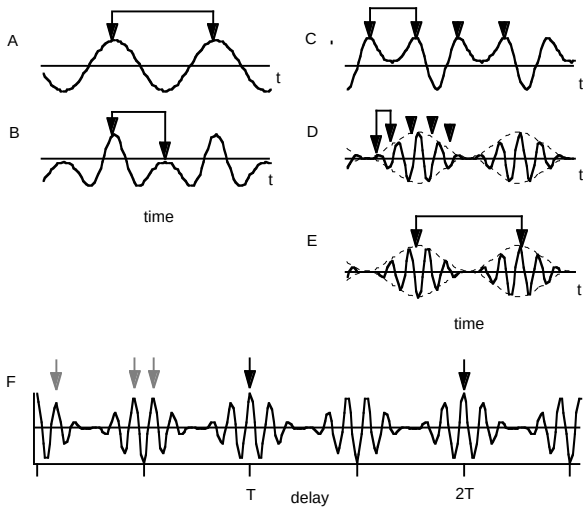

Figure 2. Temporal approach. (A-E): Waveforms of pitch-evoking stimuli. (F): Autocorrelation function of waveform in (E). Taking the interval between *successive* peaks works in (A) but fails in (B). The interval between *highest* peaks works in (B) but fails in (C). The interval between positive-going *zero-crossings* works in (C) but fails in (D). 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. Such stimuli may evoke several pitches corresponding to the gray arrows instead of (or in addition to) the pitch corresponding to the period.

This brings us to a final algorithm. Every sample is compared to all others in turn (as if all samples were "landmarks"), 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-τ)*** are similar, as when ***τ*** is equal to the period ***T***. Mathematically:

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

defines the *autocorrelation function*, illustrated in Figure 2F. For a periodic sound, the function is maximum at ***τ=0*** and all multiples of the period. 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. Models that use a flawed principle can (once the flaw is recognized) be ruled out. It is harder to know what to do with those 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 *model* 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 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 desirable but not sufficient.

There are several corollaries. Every model is "false" in that it cannot match reality in all respects (Hebb 1959). Multiple models may serve a common reality. One pitch model may predict 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 strive to falsify models until just one (or none) remains, it is fruitful to see them 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 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 evolve 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 ideas. Before that, we need to discuss two more subjects of discord: the physical dimensions of stimuli and the psychological dimensions of pitch.

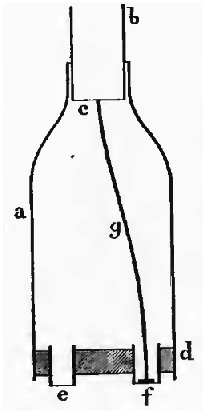

Figure 3. Johannes Müller built this model of the middle ear to convince himself that sound is transmitted from the ear drum (c) via the ossicular chain (g) to the oval window (f), rather than by air to the round window (e) as was previously thought. The model is obviously "false" (the ossicular chain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

2. 4 Stimulus models

A second source of discord is stimulus description. Many different stimuli evoke a pitch, and there are several ways to describe and parametrize them. Some ways fit a wide range of stimuli. Others fit a narrower range but with some other advantage. The "best choice" depends on which criteria are privileged, but in every case, the real stimulus differs more or less from its "idealized" description, and thus one can speak of a "stimulus model." We use the opportunity to introduce some notation that will be useful later on.

A first model for describing a pitch-evoking stimulus 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 exists one such number, there exist many: the *period* is the smallest strictly positive member of this set, and the other members are its multiples. This model is *parametrized* by the period T, and also 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 model is nevertheless useful: stimuli that fit it well tend to have a clear pitch that depends on T.

A second model 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 model is a particularization 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. Supposing our stimulus is almost, but not quite, sinusoidal, should we use the better-fitting periodic model, or the more tractable sinusoidal model? Given the advantages of the latter, it might seem reasonable to tolerate a less good fit. Part of the disagreement between pitch perception models can be traced to a different answer to this question.

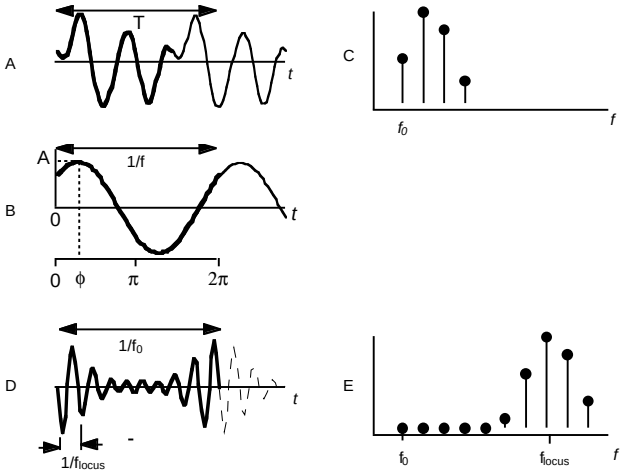

Figure 4. Stimulus models. (A): Periodic stimulus. The parameters ***o***f the model are T and the values of the stimulus during one period: s(t), $0<t\leq T$. (B): Pure tone. The parameters are f, A and ϕ. (C): Amplitude spectrum of the signal in (A). Together with phase (not shown) this provides an alternative parametrization of the signal in (A). (D,E) Waveform and spectrum of a formant-like periodic stimulus. This stimulus may evoke a pitch related to f_0 or f_{LOCUS} or both.

A third stimulus model is the *sum of sinusoids* . Fourier's theorem says that any time-limited signal may be expressed as a sum of sinusoids:

$$x(t) = \sum_k a_k \cos(f_k t + \phi_k) \qquad (2)$$

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 effect of a linear time-invariant system on the stimulus may be predicted from its effect on each sinusoid in the sum. This model thus combines useful features of the previous two, but it adds a new difficulty: it involves *several* frequencies that could plausibly map to pitch. A special case is if they are all 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 model is now *equivalent* to the periodic signal model (i.e. it describes exactly the same set of stimuli) and translates between their parameters $x(t)$, $0<t\leq T$ and the set (a_k, ϕ_k).

A fourth model, the *formant*, is a special case of the sum-of-sinusoids model 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}. If the signal is also periodic with period $T=1/f_0$, pitches related to f_0. and f_{LOCUS} may compete with each other.

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

2. 5 What is pitch?

[pointer to chapter 1]
A third possible source of discord is the definition of pitch itself. 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 - check). 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.

For *pure tones* the unidimensional nature of pitch assumed by the standards seems reasonable, as the stimulus parameter space is essentially one-dimensional. Other "perceptual dimensions" such as brightness might exist, but for pure tones they co-vary with pitch as frequency varies (Plomp 1976). For other pitch-evoking stimuli the situation is more complex. Stimuli that fit 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 to one or the other, and the outcome of experiments using these stimuli tends to be task- and listener-dependent. 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]. Pure tones also fit the formant model, and for them periodicity and spectral pitches co-vary. For other periodic sounds they are distinct. As illustrated in Figure 5, periodicity pitch has a limited *region of existence* that a pitch model should hopefully explain. Spectral pitch is often assumed to be mediated by place cues and periodicity pitch by temporal cues. However spectrum and time are closely linked, so it is wise to reserve judgment on this point.

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

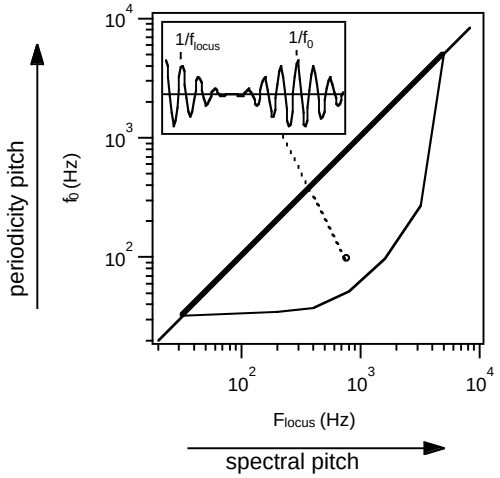

Figure 5. Formant-like stimuli may evoke two pitches, periodicity and spectral, that map to the f_0 and f_{LOCUS} stimulus dimensions respectively. The parameter space is limited by construction to the region below the diagonal. Stimuli that fall outside the closed region do not evoke a periodicity pitch with a musical nature (Semal and Demany - check; Pressnitzer et al. 2001). For pure tones (diagonal) periodicity and spectral pitch co-vary. Insert: autocorrelation function of a formant-like stimulus.

Periodicity pitch depends on a *linear* stimulus dimension (ordinate in Fig 5), but a *helical* perceptual structure has been proposed in which pitches are 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 accounts for their difference on a high-low psychological dimension. Tone height is sometimes assumed to depend on f_{LOCUS}. However, that is a distinct stimulus dimension (abscissa in Fig 5), correlate of a perceptual quantity that we call *spectral pitch* (Sec 2.5, [glossary]), probably related to the dimension of brightness in timbre.

This description of pitch is more complex than implied by the standard, and yet it does not cover concepts such as *intonation* (in speech) or *interval*, *melody* and *harmony* (in music) [pointer to Emmanuel's chapter]. 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. 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.

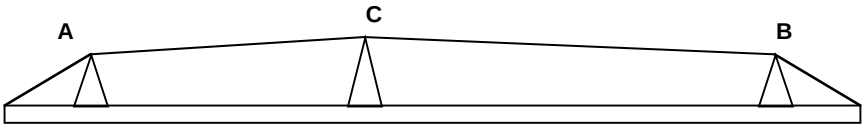

Figure 6. Monochord. A string is stretched between two fixed bridges A,B on a sounding board. A movable bridge C is placed at an intermediate position in such a way that the tension on both sides is equal. The pitches form a consonant interval if the lengths of segments AC and CB are in a simple ratio.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

Aristoxenos (4th century BC) gives a clear, authoritative description of both interval and *pitch* (Macran 1902). A definition of 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 established 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 root of its linear density. 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 figure 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 vibrate at that higher frequency). Sauveur (1701) noted that this vibration could involve *simultaneously* 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 theoretical explanation of superimposed vibrations was built by Daniel Bernoulli, but Euler leap-jumped it [check this expression] 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 string[2].

Mersenne missed the fact that the vibration he saw could reflect a combination of vibrations, with periods at fractions of the fundamental period. Any such sum has the period of the fundamental. Adding in this way *sinusoidal* partials produces variegated shapes depending on their amplitudes and phases ($\boldsymbol{a_k}$,$\boldsymbol{\phi_k}$). That *any* periodic wave can be thus obtained, and with a *unique* set of ($\boldsymbol{a_k}$,$\boldsymbol{\phi_k}$), was proved by Fourier (1822). The property had been used earlier, as many problems can be 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). The principle of superposition

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

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 one 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 suggesting that a wave can be decomposed into elementary sinusoids, each with its own period. If the theorem is invoked, the pitch-to-period mapping is no longer one-to-one.

"Vibration" was commonly understood as regularly spaced condensation separated by rarefaction, but some periodic waves have exotic shapes with several condensation/rarefaction phases 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 fit the period. Making them evenly spaced made the pitch jump to the octave (or octave plus a perfect fifth for three pulses). In every case 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 $\boldsymbol{m}$, must retain the form $\boldsymbol{a.sin2\pi mt}$". 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 $\boldsymbol{m}$ if and only if it "carries in itself the form $\boldsymbol{a.sin2\pi(mt+p)}$"[3]. In other words, every sinusoidal partial evokes a pitch, and no pitch exists without a corresponding partial of nonzero amplitude. In particular, periodicity pitch depends on the presence of the *fundamental partial* rather than on periodicity per se.

Ohm's law was attractive for two reasons. First, it drew on Fourier's theorem, 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

[3] Ascertained by applying the theorem of Fourier to consecutive waveform segments of size $\boldsymbol{1/m}$. Ohm required that the phase $\boldsymbol{p}$ of the fundamental component be the same for each segment, and that its amplitude $\boldsymbol{a}$ have the same sign but not necessarily the same magnitude. The statement "the necessary impulses must follow each other in time intervals of the length $\boldsymbol{1/m}$" could imply that the stimulus must actually be *periodic*, in which case Ohm's law would really govern only the pitch of the *fundamental* partial and not (as is usually assumed) other partials. Authors quoting Ohm usually reformulate his law, not always with equal results.

concluded that the sensation evoked by a complex tone is *composed* of the sensations evoked by the pure tones it contains. That granted, it follows that a sensation cannot depend on the relative *phases* of partials. He verified phase invariance 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 that related pitch to period: (a) among the many periodic vibrations with a given period, only those containing a nonzero fundamental partial have a pitch related to that period, (b) other partials may sometimes also evoke a pitch determined by their frequency, (c) relative amplitudes of partials affect the quality (timbre) of the vibration but not its pitch as long as the amplitude of the fundamental is not zero, (d) the 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 considered as a bank of radially taught *strings*, each loosely coupled to its neighbors. Resonant frequencies are distributed from high (basis) 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 below) 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 accounted for *roughness* and *consonance*, and thus helped bridge together mathematics, physics, elementary sensation, harmony, and aesthetics into an elegant unitary theory.

Helmholtz linked a decomposition of the stimulus to a decomposition of sensation, extending the principle of superposition to the sensory domain, and its psychoacoustic relation with the stimulus. 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 indicative 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, as evident from Seebeck's experiments, from earlier observations on beats (see below), 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 is today overwhelming.

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 of 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" model, but among the parameters f_k of that model none is obviously related to periodicity pitch. We'd need rather an operation that implements the "periodic" signal model. Interestingly, a string does just that. As Helmholtz (1857) himself explains, the string responds to its fundamental *and all of its harmonics* (Fig. 7). He used the metaphor of a bank of strings (a piano with dampers removed) to explain how the ear works, and his physiological model invoked a bank of "strings" within the cochlea. However he preferred to describe them as behaving as spherical resonators (each of which responds essentially only to a pure tone). Had he treated them as strings, each tuned to its own periodicity, there would have been no need for the later introduction of pattern-matching models. The "missing fundamental" would never have been missed.

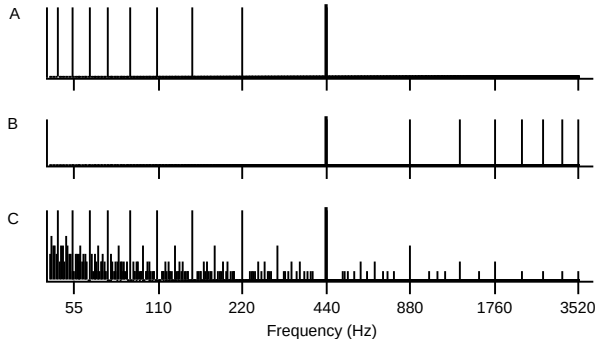

Figure 7. (A): Strings that respond to a 440 Hz *pure tone* (the abscissa of each pulse represents the tuning of a string). (B): Partials of a 440 Hz *complex tone.* (C): Strings that respond to a 440 Hz complex tone. Pulses are scaled in proportion to the power of the response. The string resonator is selective to periodicity rather than Fourier frequency.

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. Searching to overcome precisely that limitation, Alexander Graham

Bell read Helmholtz and used his ideas to try to develop a *multiplexing* telegraph (Hounsell 1976). Getting sidetracked, he invented instead the telephone, a metaphor later used by Rutherford (1886) in a theory opposed to that of Helmholtz...

The next section shows how the missing fundamental problem was fixed 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 in such a model: one produces the *pattern* and the other matches it to 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,* and thus modeled as a gaussian process with mean f_k and variance function of f_k. Only *resolved* partials (those that differ from their neighbors by more than a resolution limit) are included, and neither amplitude nor phase are represented. A "central processor" attempts to account for the series as *consecutive* multiples of a common fundamental (the constraint that partials are consecutive was later raised by Gerson and Goldstein 1978). Goldstein suggested that estimates were possibly, but not necessarily, produced in the cochlea according to Helmholtz's model. 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 Helmholtz[4]. From this is derived a "gestalt" *virtual pitch* (periodicity pitch). Perception operates in either of two modes, analytic or synthetic, according to whether the listener consciously tries to access spectral or partial pitch. Partial pitch is presumably innate, but virtual pitch perception is *learned* by exposure to speech. Normal listening is synthetic. Analytic mode adheres strictly to Ohm's law: there is a one-to-one mapping between partial pitches and resolved partials.

Despite differences of detail, these three models are formally similar (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). The "multicue mediation theory" of Thurlow (1963) suggests that listeners use their own 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)...

5. 2 Relation to signal processing methods

Pattern-matching is used in signal processing methods to estimate the period of signals such as speech. 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, Willems and Sluyter (1983) 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. Applied to a *power* spectrum, this 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's model took that into account by applying nonuniform smoothing to the input pattern. High-frequency channels being more smoothed, they may be sampled more sparsely. Doing so produces the *MFCC* (mel-frequency cepstrum coefficients), popular in speech processing and analogous to the logarithmic spectra of Versnel and Shamma (1998). Resampling the frequency axis scrambles harmonicity information, so these are not very useful for pitch.

---

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

A detail is worth mentioning. We often think of frequency as extending along an axis from 0 Hz to some upper limit (or infinity). However the spectra that relate to the ACF and cepstrum are defined over an axis that includes *positive and negative* frequencies. Those spectra are symmetrical around 0 Hz, so only cosines are involved in their Fourier decomposition. Transposing from the cosine to other shapes of "harmonic comb", this amounts to anchoring a tooth of every comb at 0 Hz. Without that constraint, and considering only the positive frequency axis, any set of partials spaced by Δf is best matched by a comb with teeth spaced by Δf. In that case, one does not predict the pitch shifts observed for inharmonic complexes (first effect). Jenkins (1961) and Schouten et al. 1962) raised this as an objection against spectral models. With the constraint, however, the best match is a slightly stretched comb labeled by the *pseudofundamental* of the partials, that implies a first effect, and thus the objection falls.

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. Synapses at the intersection between channels tuned to the fundamental and harmonics would be reinforced through Hebbian learning (Hebb 1949; Roederer 1975). Licklider (1959) had 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 *subharmonics* (or "superperiods"). This suggests that one could do away with Terhardt's requirement of early exposure to *harmonically rich* sounds: a pure tone also has subharmonics, so it could serve as well. Readers in need of a metaphor to accept this idea should consider Figure 7. Panel B illustrates the template (made irregular by the logarithmic axis) formed by the partials of a harmonic complex tone. Panel A illustrates a similar template formed by the superperiods of a *pure* tone.

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, (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). Through nonlinearity there arises within the channel a series of harmonic components that correlate with other channels tuned to those frequencies,

resulting in Hebbian reinforcement [define] at their intersection. The loci of reinforcement form diagonals across the matrix. Together these diagonals form a harmonic template. Shamma and Klein made a fourth assumption: (4) sharp phase transitions along the basilar membrane near the locus tuned to each frequency. Apparently this is 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 that 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 place mechanism or on the basis of temporal fine structure. It *is* particular about its quality: the number and accuracy of partial frequency estimates.

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 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 from peak width: narrow or wide its locus can be determined exactly, for example by competition within a "winner-take-all" neural network (Haykin 1999). This works for a single tone, and in the absence of internal noise. 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 the 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 tuning is explained by *active* cochlear 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), to the extent that it sometimes produces spontaneous oscillations. However active mechanisms are nonlinear, so one cannot extrapolate behavior observed with a pure tone to a combination of partials. The phenomenon of *suppression* (by which the auditory nerve response to a pure tone is suppressed by a neighboring tone) suggests that the gain boost at resonance that contributes to sharp tuning is lost if the tone is not alone. If so, such hyper-sharp tuning is of little use to sharpen the response pattern of each partial of a complex tone. At medium-to-high amplitudes, profiles of response to complex tones lack any evidence of harmonic structure (Sachs and Young 1979) [check Delgutte].

A "second filter" after the basilar membrane was a popular hypothesis before measurements from the cochlea 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 membrane), 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, or in autocorrelation models to derive accurate estimates from blunt peaks).

Sharpening can operate purely on the amplitude pattern, for example spatial differentiation or filtering (e.g. summation of excitatory and inhibitory inputs of different tuning), or an expansive nonlinearity (e.g. coincidence of similarly-tuned inputs). It may also use *phase*, for example the differential motion of neighboring parts within the cochlea, or neural interaction of phase-locked responses from neighboring parts of the basilar membrane. The Lateral Inhibitory Network (LIN) of Shamma (1985) uses both amplitude and phase. Partials of low frequency (<2 kHz) are emphasized by rapid phase transitions along the basilar membrane, and those of high frequency by spatial differentiation of the amplitude pattern. Neural interaction of a slightly different form may account for loudness (Carney et al. 2002). In the Average Localized Synchrony Rate (ALSR) of Young and Sachs (1979), a narrowband filter tuned to the characteristic frequency of a 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 that exist to improve sharpening.

An alternative to sharpening is to assume that a pure tone is coded by the *edge* of a tonotopic excitation pattern rather than its peak (Zwicker 1970). Likewise, Whitfield (1970) proposed that a complex tone might produce a pattern of responses to each partial separated by *gaps*.

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 refine this position. 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 in 1863 by Hensen (Boring, 1942). Peripheral filtering would serve merely to *resolve* partials so that each channel is labeled clearly by a single frequency. A nice feature of this idea is that all channels responding to a partial may be used to characterize it (rather than just some predetermined set). Tonotopy is not needed to carry the information, as noted by Goldstein (1973), but 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 filter bank. In the *dominant component* scheme of Delgutte (1984), each channel is subjected to a Fourier transform. Transforms are summed over channels. The same principle underlies the *modulation filterbank* (e.g. Dau, Püschel and Kohlrausch 1996) discussed below in the context of temporal models. An objection to these hypotheses is that they reproduce centrally an operation (filtering) that is available peripherally, and thus lose some of the appeal of Helmholtz's original idea.

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 optimum processing of spike trains. Calculations showed that place alone was sufficient to account for human performance, whereas time allowed better performance. Siebert tentatively concluded that the auditory system probably does *not* use time. However, *suboptimal* processing of temporal cues (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, Colburn and Carney (2001) found, as Siebert, that place cues are sufficient, and time cues more than sufficient, to predict behavioral thresholds. However temporal cues predicted dependencies on frequency and level that were parallel to those observed. Place cues did not. Heinz et al. tentatively concluded that the auditory system probably *does* use time. Interestingly, despite a severe degradation of time cues beyond 5 kHz (Johnson, 1980), useful information could be exploited up to 10 kHz at least. Behavioral and predicted thresholds were parallel up to the highest frequency measured, 8 kHz.

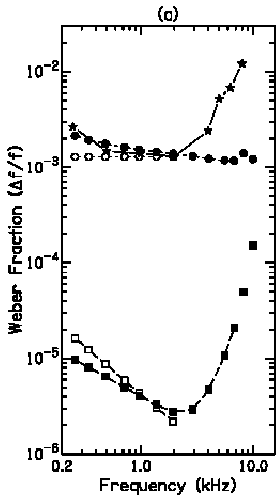

Figure 8. Pure tone frequency discrimination by humans and models. Stars: threshold for a 200 ms pure tone with equal loudness as a function of frequency (Moore, 1973). Circles: predictions for a place-only model. Squares: predictions of a time-only model. Open symbols are Siebert's (1970) analytical model, closed symbols are Heinz et al's (2001) computational model. [a high-quality epsf exists for this figure, but it needs cropping]

To summarize, a wide range of schemes produce spectral patterns adequate for pattern matching. Some rely entirely on basilar membrane 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. In the notion of overlap between

successive elementary sounds, one might read a prefiguration of 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 own heartbeat, or 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 evoked 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 explanation seems to explain more: 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). According to von Békésy and Rosenblith (1948), resonance was alternately supported and opposed (for example by Johannes Müller) until it was firmly established by Helmholtz. The *telephone theory* of Rutherford (1886) was possibly a reaction against the authority of Helmholtz's theory and its network of mutually supporting assumptions, some untenable such as Ohm's law. In the minimalist spirit of Anaxagoras, Rutherford

suggested that the ear merely transmits vibrations to the brain like a telephone receiver. It is striking to note the contrast between this modest theory (2 pages) and the monumental opus of Helmholtz that it opposed. To its credit, Rutherford's two-page theory was parsimonious, to its discredit it just shoved the problem one stage up. The telephone, recently invented and vigorously promoted towards scientific circles (Hounshell, 1976) was bound to be seized as a metaphor.

An objection 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), and did not doubt that the auditory nerve might respond faster. A more satisfying answer was 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. Measurements in auditory nerve fibers proved the theory wrong in that firing is *stochastic* rather than regular (Galambos and Davis 1943, Tasaki 1954), and right in that several 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 measurable up to 3-5 kHz in the cat (Johnson, 1980). The limit is lower in the guinea pig, higher (9 kHz, Köppl 1997) in the barn owl, and unknown in humans.

Pure tone waveforms have a single peak per period, and should evoke a simple firing pattern to which to apply the volley principle (in its probabilistic form). However we saw earlier (Sect. 2.2) the limits of this and other simple schemes as a basis for period estimation as required by a pitch model. The idea that pitch follows the *envelope*, presumably 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) with inharmonic stimuli. As the partials of a modulated-carrier stimuli are mistuned, the pitch shifts, but the envelope stays the same, ruling out not only the envelope as a cue to pitch, 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 AC model. Before moving on to that, we'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 and sensory elements. The *fundamental* partial maps to periodicity pitch, and higher partials to higher pitches that some people sometimes hear. Schouten (1940a) agreed to all 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 observations. 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. In that case the stimulus appeared 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. Reasoning from his new law, Schouten assumed that the missing-fundamental complex must contain (or be) the residue. Removing additional low partials left the sharp quality intact. Furthermore, 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.

The exclusion of resolvable partials from the residue put Schouten's theory into trouble when it was found that they actually dominate periodicity pitch (Ritsma 1962, 1963). Strangely enough, Schouten gave as an example of residue a bell with characteristic tones fitting the (highly resolvable) series 2:3:4 (Schouten 1940b,c), that produces a strike note that fits the missing fundamental. De Boer (1976) amended Schouten's definition to include all partials, which is tantamount to saying that the residue *is* the sound, rather than part of it. Schouten (1940a) mentioned this possibility but rejected it as causing "a great many difficulties" without further explanation. Possibly, he believed that *interaction between partials* in the cochlea, 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. Their pitch is relatively weak, which confirms that the residue (in Schouten's narrow definition) is *not* a major determinant of periodicity pitch.

9 Autocorrelation

Autocorrelation, like pattern-matching, is the basis of 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-τ)*** and smooth this measure over time to obtain a temporally stable measure of self-similarity:

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

This is simply 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. The relation with the *autocorrelation function* or ACF (Eq. 1) is found by expanding the squared difference in Equation 3. This gives the relation:

$$d(\tau) = p - r(\tau) \qquad (4)$$

where ***p*** represents signal power. Power ***p*** can be considered constant if integration occurs over a long window. Thus, if ***r(τ)*** increases ***d(τ)*** 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 andC, for a the stimulus illustrated in A.

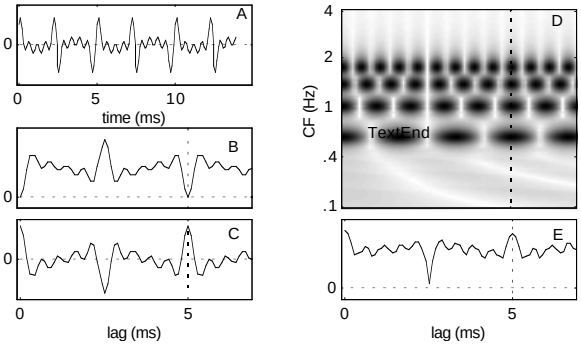

Figure 9. (A): Stimulus consisting of odd harmonics 3, 5, 7, 9. (B): Difference function ***d(τ)***. (C): AC function ***r(τ)***. (D): Array of ACFs as in Licklider's model. (E): Summary ACF as in Meddis and Hewitt's model. Vertical dotted lines indicate the position of the period cue. Note that the 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.

9. 2 Licklider

Licklider (1951, 1959) suggested that autocorrelation could explain pitch. Processing must occur 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. Licklider didn't actually give a procedure for doing so. Meddis and Hewitt (1991) simply summed the 2D pattern across the frequency dimension to produce a "summary AC function" (SACF) from which the period may be derived (Fig. 9E). They showed that the model could account for many important pitch phenomena. 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 inputs. Its firing probability is the *product* of firing probabilities at its inputs. 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), but 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 basilar membrane to derive the period, a fact that escaped Schouten (and possibly 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 filter bank). The differences between Figs. 9C and E reflect nonlinear transduction.

9. 3 Phase Sensitivity

Excessive phase sensitivity was a major argument against temporal models (Wightman, 1973). It is important to see how autocorrelation fares in this respect. Phase refers to the parameter ϕ of the of the sinusoid model, or ϕ_k of the sum-of-sinusoids model (Sect. 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[5]. Helmholtz limited its validity to resolved partials.

Three factors contribute to make the AC model phase-insensitive. First, ACFs from channels that respond to a *single* partial do not depend on phase. Second, ACFs in channels that respond to *two* partials of sufficiently high rank are only slightly phase-dependent. Third, ACFs in channels responding to *three channels or more* are phase-dependent, but this dependency does not affect the position of the period cue[6]. Phase may affect the *relative salience* of that cue relative to cues at competing lags. This explains changes in the distribution of pitch matches for stimuli with ambiguous pitch.

Since some phase changes *are* perceptible, one would like a model of hearing (if not pitch) to predict them (Patterson 1994a,b). In the AC model, phase sensitivity may arise from several sources: nonlinear basilar membrane mechanics (in particular via the production of combination tones), hair-cell adaptation, nonlinearity of hair-cell transduction, and also properties of hypothetical physiological implementations of the ACF (de Cheveigné 1998).

9. 4 Histograms

Licklider's "autocorrelation-like" 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. Intervals are between consecutive spikes for the first-order histogram, and between all spikes, consecutive or not, for the all-order histogram. 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 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 1982, 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 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). Without entering the debate, we note that all-order statistics may usefully be

[5] 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.

[6] This is true for periodic stimuli. For certain quasi-periodic stimuli the pitch cue may vary with phase, as does the pitch itself (Pressnitzer et al. 2001).

applied to a *population* of fibers. This has several desirable effects. There is no longer a lower limit on interval duration (~0.7 ms) due to refractory effects, so frequencies above 800 Hz may be represented by their period mode. In single-fiber statistics the period mode tends to be distorted or suppressed. Better use is made of available information, as the number of intervals increases with the *square* of the aggregate number of spikes [add figure]. First-order statistics cannot usefully be applied to a population because, as aggregate rate increases, more intervals belong to the zero-order mode and the period mode is depleted. The effect is accompanied by a shift of the period mode towards shorter intervals, a phenomenon that has actually been invoked to explain certain pitch shifts (Ohgushi 1978, Hartmann 1993). The all-order histogram has a functional advantage.

It is important to realize that histogram statistics *discard* information, each in its own way. 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 the expected "first effect of pitch shift", whereas a period histogram applied to the same data does not (Evans 1978). Care must be exercised in the choice of statistics and their interpretation.

9. 5 Related models

The *cancellation* model uses the difference function of Eq. 3 instead of the ACF (de Cheveigné 1998). It is formally quite similar to the AC model, its major advantages being its relations with *harmonic segregation* and the ease with which it can be extended to account for multiple pitches (see below). A "neural" implementation on the lines of Licklider's is obtained by replacing an excitatory synapse by an *inhibitory* synapse, and assuming that the coincidence neuron transmits all spikes that arrive at the excitatory synapse unless they coincide with an inhibitory spike. The 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 f_0 estimation of Ross et al. (1974).

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

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

where $s(t)$ is a train of pulses ("strobes"), derived by some process such as peak picking. Processing occurs within each filter channel to produce a 2D pattern similar to Licklider's. STI preserves phase to some degree. Thus it can explain sensitivity to temporal asymmetry observed for some stimuli, 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:

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

in which case the implementation of the delay is less costly (a pulse is cheaper to delay than an arbitrary waveform). Within the brain stem, 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 early temporal models, on the assignment of a marker (strobe) to each period. The *Auditory Image Model* (AIM) designates, according to context, either STI or a wider class including autocorrelation. Thanks to strobed integration, the fleeting patterns of transduced activity are "stabilized" to form an image. The idea, as in similar simulations based on the ACF (e.g. Lyon 1984, Weintraub 1985, Slaney 1990), is that *visually* prominent features of this image should be easily accessible to a central processor. An earlier avatar of this idea might be the "camera acustica" model of Ewald (1898, Wever 1949) in which the cochlea behaved as a resonant string, responding to different frequencies by patterns of resonance characteristic of each sound. STI and AIM evolved from earlier *pulse ribbon* and *spiral* detection models (Patterson 1986, 1987).

The *dominant component* scheme of Delgutte (1984) and 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). Power spectrum and ACF form a Fourier transform pair, and in this sense the 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. In this role it is prone to the same "missing fundamental" problem that plagues place theory. In some sense it seems strange to posit a central Fourier transformer given that a peripheral one (of sorts) exists in the cochlea.

Interestingly, the *string* can be seen as belonging to the same 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 this circuit and the previous one have peaks at all multiples of $f=1/\tau$, but the peaks are sharper for the latter. A *string* is, in essence, a delay line that feeds back onto itself as in Figure 10D.

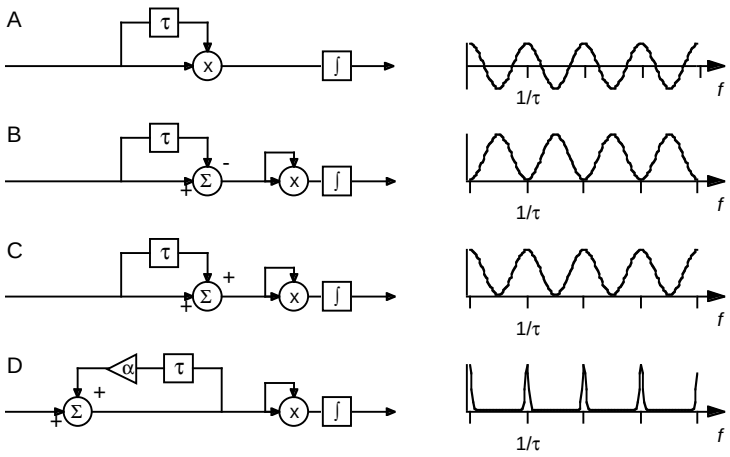

Figure 10. Processing involved in various pitch models. (A) Autocorrelation involves *multiplication*. (B) Cancellation involves *subtraction*. (C) The feed-forward comb-filter (Delgutte 1984) involves *addition*. (D) In the feed-back comb-filter, the *delayed sum* is added to the input (after attenuation),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

These examples emphasize the deep relations between autocorrelation and pattern-matching. It is worth pointing out an important difference between ACF and string: *temporal* resolution. At each instant, the output of the ACF reflects a relatively short interval of its input (determined by the delay $\boldsymbol{\tau}$ and temporal smoothing). That of the string reflects the past over many multiples of $\boldsymbol{\tau}$, as information is recycled within the delay line. In effect, comparisons are made across *multiple* periods, thus improving frequency resolution at the expense of time resolution. This is reminiscent of the *narrowed AC function* (Brown and Puckette 1989) that uses high-order modes of the ACF to sharpen the period mode. The idea was used by de Cheveigné (1989) and Slaney (1990) to explain acuity of pure tone discrimination. Once again we find strong connections between different models.

To conclude on a historical note, an early version of autocorrelation can be found in model of Hurst (1885, Wever 1949), who suggested that sound propagates up the tympanic duct, through the helicotrema and back down the vestibular duct. The basilar membrane is pressed from both sides wherever an ascending pulse meets the previous descending pulse, and the position of this point characterizes the period. More recently, Loeb et al. (1983) and Shamma et al. (1989) invoked the basilar membrane as an alternative to neural delays. The basilar membrane is dispersive and behaves as a delay line only for a narrow-

band stimulus. In that case delay can be assimilated 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 a mode at zero lag, and modes at positive multiples of the period (Fig. 11A). The cue to pitch is the *leftmost* of the latter (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.

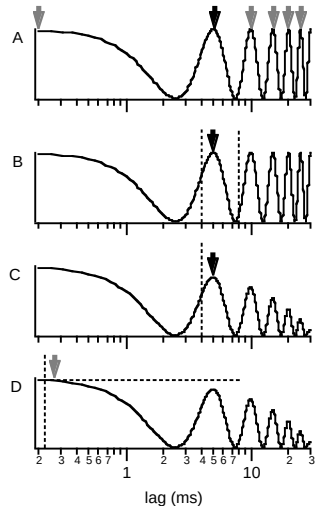

Figure 11. SACFs in response to a 200 Hz pure tone. The abscissa is logarithmic and covers roughly the range of periods that evoke a musical pitch (0.2 to 30 ms). The pitch mechanism must choose the mode that indicates the period (dark arrow in A) and reject the others (gray arrows). This may be done by setting lower and higher 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 within the search range (D).

The first and 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 9 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). Such a bias occurs naturally when the ACF is calculated from a short-term Fourier transform, and Pressnitzer et al. (2001) used it to deemphasize pitch cues at longer lags. The difficulty with such a bias is that the period mode (or a minor mode near it) can happen to be less salient than the zero-order mode (Fig. 11D). The difficulty can be circumvented by various heuristics, but the result tends to be messy and of limited generality. A good solution was proposed recently by de Cheveigné and Kawahara (2002).

Once the mode has been chosen, its *position* must be determined precisely. Supposing internal noise, it is not clear how the relatively wide modes obtained for a pure tone (Fig. 11) can be located with accuracy commensurable with discrimination thresholds (about 0.2% at 1kHz, 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 *discrimination*, it may not be necessary to actually choose or locate a mode. For example Meddis and O'Mard (1997) use the Euclidean distance between SACF patterns to predict discrimination thresholds. This avoids choosing a mode, but it does not account for generalization of pitch across different SACF shapes (i.e. different timbres), nor does it allow predicting which of two stimuli is *higher* in pitch.

10 Pros and Cons of Pattern-matching vs Autocorrelation

Pattern-matching and autocorrelation are the two major explanations of pitch today. As we have seen they are deeply related. They are also largely equivalent in terms of behavior, so it is hard to choose between them on that basis.

Pattern-matching operates on sensory correlates that are supposed equivalent to those of pure tones. Much effort has been invested in the psychophysics of pure tones, and it is satisfying to assume that the results are relevant to the perception of complex sounds. The benefit is analogous, potentially, to that of the superposition principle in physics. A drawback is lack of parsimony, in that pattern-matching requires two stages (spectral analysis and pattern-matching) where a time-domain model requires one. On the other hand, the first of these stages *is* known to occur, at least in a rough form. Autocorrelation measures directly the feature relevant for pitch: periodicity. In a sense it embodies pattern extraction *and* pattern matching in a single step. The drawback then is that it does not seem to really need peripheral filtering, a prominent feature of auditory processing. However that feature has other useful functions so this is perhaps not a serious objection. The following paragraphs examine the debate in more detail.

10. 1 Anatomy and physiology

The strongest argument for pattern-matching is the presence of peripheral filtering. Tonotopy is observed at many stages in the auditory system (although one might expect relevant patterns to be *processed* rather than replicated). Rate/frequency tuning is sharp for pure tones, but less so for partials embedded within a complex, particularly at high amplitudes. This may rule out a rate/place representation for patterns used for pitch. Partials may still be resolved in temporal patterns: certain fibers or cells respond exclusively to a single partial of a complex tone. However there is little evidence for neural circuits decoding these patterns (e.g. ALSR-type matched filters; Sec 5.1), or of a resulting central spectrum pattern. There is no strong evidence for the existence of harmonic templates, of the pattern-matcher that uses them, or of a representation of its output (which might show up as periodotopy, or at least periodicity-tuning). Shamma and Klein (2000) showed plausibly how templates could emerge, but evidence is lacking for certain of the ingredients of the model (e.g. spectral sharpening), or its results.

Accurate phase-locking in the auditory nerve and the first relays of the brain stem favors the AC model, but the gradual loss of synchrony as one progresses within the system makes it unlikely that the model is implemented at a level beyond IC. For the AC model is the simplicity and plausibility of the processing required, but against it is the lack of evidence of that processing actually occurring, or of a neural representation of its result (which might show up as periodotopy, or at least periodicity-tuning). There is also no clear evidence that the neural *delays* that are required can actually be dealt with by the auditory system..

It is hard to weigh the lack of clear evidence. Measurements are sparse and relevant data may have passed unnoticed, so one cannot rule out either model. Yet the lack of evidence might also mean that processing follows an entirely different principle. In that case the models may still be of use as first-order accounts (they serve at least to highlight the functions required of a pitch mechanism). Licklider (1959) pointed out that there is no strong reason to expect relevant patterns to be *spatially ordered*. Furthermore, recoding of the sort suggested by Barlow (1968) (Sec. 11.10) would radically scramble patterns. If Barlow's principles are used by the auditory system, a fully functional implementation of either model might produce no easily-interpreted pattern.

10. 2 The issue of resolved harmonics

Stimuli with resolved harmonics tend to evoke a strong pitch, while unresolved harmonics (Schouten's "residue") produce a weaker pitch. The AC model is expected to handle both equally well, and this is sometimes held as evidence *against* it. The argument is not decisive, as factors may intervene to degrade the performance of an otherwise effective mechanism (the converse is not true:

insufficient performance would decisively rule out the model). It may be argued that the evidence better fits pattern-matching models that require partials to be resolved. Those models *fail* for unresolved partials, but the objection may be answered by supposing two mechanisms (Sec. 10.3).

A confusing factor is that experiments that manipulate resolution also manipulate other factors. For example, to reduce resolution while keeping the period constant, the stimulus must be restricted to high frequencies. To reduce resolution while keeping the frequency range constant requires long periods. It is conceivable that frequency and/or the cost of longer delays are the determinant factor in each case. Furthermore, there are some elements that are inconsistent with the hypothesis that resolution per se is the factor that determines the superiority of resolved harmonics for pitch. [review]

Pattern-matching models require the frequencies of each partial to be independently measurable, in which case it should be capable of evoking its own pitch (at least according to Terhardt's model). An *isolated* partial certainly evokes a pitch, but two concurrent partials are individually perceptible only if their frequencies differ by about 8% at 500Hz, and somewhat more above and below that frequency (Plomp 1964). Partials closer in frequency evoke a single pitch, approximately dependent on the centroid of the power spectrum (Dai et al. 1996) (this justifies the assertion made in Section 2.5 that spectral pitch depends on the locus of a spectral concentration of power). A partial flanked by neighbors on both sides is even harder to resolve: partials within a complex tone may be resolved individually only up to rank 5-8 (Plomp 1964).

10. 3 The two-mechanism hypothesis

Pattern-matching and autocorrelation each have their own features and following. It is tempting to adopt both, assigning each to 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 and an easier consensus. The disadvantages are that two mechanisms are involved, plus a third to integrate the two.

The situation 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 illustrated in Figure 3 reduced the two paths to one. Du Verney (1683) believed that the trumpet-shaped semicircular canals were tuned just like the cochlea, and Helmholtz initially thought that they handled noise[7] before he realized that cochlear spectral analysis could handle it too. The labyrinth had been shown to be nonauditory by Flourens in the early 19th century (Boring 1942). Bachem (1937) postulated two independent pitch mechanisms, one devoted to tone height, the other to chroma, the latter more

[7] Todd and Cody (2000) recently suggested that the sacculus plays a role in the perception of loud dance music [Cazals, Science 198x].

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. Licklider's duplex model implemented both, and suggested how tonotopic and periodotopic dimensions could map to a common pitch dimension, thanks to Hebbian learning within a neural network.

The existence of temporal *and* place mechanisms is a common assumption today (e.g. Gockel et al. 2001). That of independent mechanisms for resolved and unresolved harmonics (Houtsma and Smurzynski 1990) is becoming popular (e.g. 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. The assumption of multiple mechanisms is an easy fix to problems of a unitary model. Like adding free parameters to a model, it is guaranteed to give a better fit. Precisely for that reason, it should be resisted (though not necessarily refused). In any case, it does not address objections such as lack of physiological evidence for delay lines or harmonic templates.

11 Advanced Topics

Modern pitch models account for major phenomena equally well. To decide between them, one must look at implementation constraints, second-order effects, and more complex aspects of pitch. If we suppose that the same mechanism underlies them all, these effects and constraints may reveal its nature. In a sense, this is the cutting edge of pitch theory. The casual reader should skip to Section 10 and come back on a rainy day. Brave reader, read on.

11. 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 premise, and assigning any pitch not accounted for by a partial to a distortion product.

If the stimulus is a pure tone of frequency $\boldsymbol{f}$, distortion products are harmonics $\boldsymbol{nf}$. If the stimulus contains two partials at $\boldsymbol{f}$ and $\boldsymbol{g}$, they also include

terms of the form ***±nf ± 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 only 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. Their amplitudes increase at a rate of 2dB per dB for the difference tone and 3dB 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 be always possible. 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 such 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 salience is large 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 far as that law claims to relate *stimulus* components (as opposed to internally-produced components) to pitches. In addition, it is possible to simultaneously estimate and *cancel* the 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 more accurate (Schouten 1938, 1970). Canceling this very weak distortion product does not affect the pitch. 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. 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 harmonically-related partials (Plomp 1967b), and thus they do *not* indicate the presence of a component. This realization came after many such measurements had been published. As proof of nonlinearity, aural harmonics reinforced the difference-tone hypothesis, otherwise weak for lack of direct evidence. Thus they added to confusion. Similarly confusing were measurements of distortion products in cochlear microphonics (Newman et al.

1937) or auditory nerve-fiber responses. These arise because of nonlinear mechanical-to-nervous or electrical transduction. They do *not* signal the presence of 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 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 measurable (by Rayleigh's cancellation method) only for ***g/f*** ratios above 1.1 (Goldstein 1970), at which point it is about 14 dB below the primaries. Amplitude decreases rapidly as the frequency spacing becomes larger. Even weak, a combination tone can strongly affect the pitch if it falls within the *dominance region* (see below). Difference tones of higher order (***f-n(g-f)***) can also contribute (Smoorenburg 1970).

Combination tones are important for pitch theory. They are necessary to explain the "second effect" of pitch shift of harmonic complexes (de Boer 1976; Smoorenburg 1970), and they allow spectral theories to account for 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 experiments now add *lowpass noise* (e.g. pink noise) to mask distortion products.

11. 2 Temporal integration and resolution

Helmholtz reasoned that the ear must follow "shakes" of up to 8 notes per second that occur in music. From this he derived a lower limit on the bandwidth of resonators in his model. This prefigured the *time-frequency tradeoff* of Gabor (1947) (analogous to Heisenberg's principle of uncertainty in quantum mechanics) that can be expressed as:

$$\Delta f \Delta t \geq 1 \qquad (7)$$

given appropriate definitions of frequency and time uncertainties ***Δf*** and ***Δt***. Fine spectral resolution requires a long temporal analysis window, and fine temporal resolution likewise implies coarse spectral features. As Gabor's relation is so very fundamental, it is puzzling to learn that *arbitrarilily accurate* estimates can be derived from a short segment of waveform (on the order of two periods). The solution to this puzzle was pointed out by Nordmark (1968a,b). The word "frequency" carries two different meanings in this context. One is *group frequency* as measured by Fourier analysis:

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

The other is the reciprocal of the interval between two events of equal phase, called *phase frequency* by Kneser (1948; Nordmark 1968a,b). It can be determined with arbitrary accuracy from a short chunk of signal.

As this is an important issue, the claim needs clarification. Suppose the period estimation algorithm is based on autocorrelation:

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

Two time constants are involved: the window size $\boldsymbol{W}$, and the maximum lag $\boldsymbol{\tau_{MAX}}$ for which the function is calculated. The duration of the chunk of signal needed is their sum. What are the minimum values needed to estimate the period of a signal? The answer depends on the period $\boldsymbol{T}$, of course, and also on the largest *possible* period $\boldsymbol{T_{MAX}}$. The reason why the latter intervenes is that estimation involves verifying both that $\boldsymbol{T}$ is a satisfactory candidate for the observed chunk, and that the chunk is not part of some larger periodic pattern. Concretely, $\boldsymbol{\tau_{MAX}}$ should be at least equal to $\boldsymbol{T}$, and $\boldsymbol{W}$ at least equal to $\boldsymbol{T_{MAX}}$ (de Cheveigné and Kawahara 2002). Thus, the duration required depends on the lower limit of the $\boldsymbol{f_0}$ range.

As an example, given that the lower limit of melodic pitch is 30Hz ≈ 1/30ms (Pressnitzer et al. 2001), it takes at least 40 ms to ensure that a tone within the melodic pitch range has an $\boldsymbol{f_0}$ of 100Hz = 1/10ms. If it is known that the $\boldsymbol{f_0}$ is no lower than 100 Hz, the duration can be shortened to 20 ms. In special cases, if more is known about the stimulus (e.g. the waveform shape) the duration might even be shortened almost to 10 ms, but no shorter[8]. These estimates apply in the absence of noise, internal or external. With noise present, obviously more time will be needed to counteract its effects.

Suppose now that the stimulus is longer than the required minimum. The extra time can be used for two purposes. One is to increase integration time to reduce noise. The other is to measure super-periods (extending the search for self similarity up to $\boldsymbol{nT}$ to refine the estimate of $\boldsymbol{T}$). A third option is to divide the stimulus into intervals and process each as if it were a shorter stimulus (surrounded by silence), and then average the estimates. This option is wasteful, as it ignores structure across interval boundaries. A priori, the auditory system could use any of these strategies or some combination of them.

Licklider (1951) tentatively chose 2.5 ms for the size of his exponentially-shaped integration windows (roughly corresponding to $\boldsymbol{W}$). Based on the analysis above, this size is sufficient only for frequencies beyond 250 Hz. A larger value of 10 ms was used by Meddis and Hewitt (1992). Wiegrebe et al. (1998) argued for two stages of integration separated by a nonlinearity, with a

[8] One might hope that a pattern-matching model could provide better temporal resolution, as each partial has a shorter period and thus requires less time to estimate than the fundamental period. Unfortunately, each partial must first be *resolved*, and this requires a filter with an impulse response at least as long as twice the fundamental period.

1.5 ms window for the first stage and some larger value for the second. Wiegrebe (2001) later found a task-dependent window size of about twice the stimulus period used, with a minimum of 2.5 ms. These values were determined experimentally from the minimum duration needed for a task, but discrimination is known to improve with duration (e.g. Moore 1973), so the auditory system is obviously capable of integrating information over longer periods. Plack and White (2000a,b) found that such integration could be *reset* by transient events. Resetting is obviously necessary to obtain samples for comparison across time, as required in discrimination tasks, or by sampling models of FM or glide perception. The latter also require *memory*, and it is conceivable that integration and sensory memory have a common substrate. Grose et al. (2002) estimated the *maximum* integration time for pitch to be about 210 ms.

11. 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. Of interest is less the exact terms chosen by the translator (Macran 1902), than the fact that these concepts were so carefully distinguished. It is indeed conceivable that dynamic pitch is perceived differently from static pitch. For example FM might be transformed to AM and perceived by an AM-sensitive mechanism. Frequency glides might be decoded by a mechanism sensitive to the derivative of frequency [ref]. The alternative is that the modulation is simply sampled by the standard pitch mechanism, and the samples compared across time (Hartmann and Klein 1980; Dooley and Moore 1988). This hypothesis entails several requirements: temporal resolution must be sufficient to follow the modulation, the mechanism must be tolerant to the residual frequency variation over each analysis window, and the system must include memory to allow comparison over time.

Estimation is not instantaneous, so "sampling" of a frequency-modulated stimulus makes sense only in a limited way[9]. Due to the frequency change, periodicity is approximate and estimation may be compromised. Furthermore, integration of unequal frequency estimates makes the choice of *the* frequency at any instant uncertain, so discrimination is expected to be poor. A possible exception occurs at extrema of frequency modulation. In that situation (where Demany and Clément 1997 observed "hyperacute" frequency discrimination), opposite frequency changes on either side of the extremum might compensate each other (de Cheveigné 2000).

[9] "Instantaneous frequency" seems to allow perfect temporal sampling of a changing frequency, but it should be recalled that spectral analysis is required to estimate it. This involves integration over time.

The case might be made for the opposite proposition, that tasks involving static pitch (e.g. discrimination between successive stimuli) are performed on the basis of detectors sensitive to *change*. It has been noted that certain stimuli acquire a pitch quality only when that pitch is made to vary (Davis 1951). In the extreme one could propose that pitch is not a linear perceptual dimension, but rather a combination of pitch-change sensitivity and sensitivity to musical interval.

If listeners are asked to judge the *overall* pitch of a modulated stimulus, the result usually falls within the range of modulation. The match can be predicted by taking the *average* of instantaneous frequency. If instantaneous amplitude changes together with instantaneous frequency, the latter is best weighted by the former, according to the intensity- or envelope-weighted average instantaneous frequency (IWAIF or EWAIF) models (Anantharama et al. 1993; Dai et al. 1996). An even better prediction is obtained by weighting inversely with *rate of change* (Gockel et al. 2001).

11. 4 Pitch salience

[More here]

11. 5 Multiple pitches

Most pitch models assume that a stimulus evokes one 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) complex tones in analytic listening mode, (d) concurrent voices or instruments.

Some stimuli used in early experiments were reported to have an ambiguous periodicity pitch (de Boer 1956; Schouten et al. 1962). Most pitch models produce multimodal cues for those stimuli, that fit the multimodal histograms of pitch matches. A standard pitch model is adequate in this case.

A narrow-band formant-like stimulus may produce a *spectral pitch* related to the formant frequency (Sect. 2.5). If the stimulus is periodic, the spectral pitch competes with the lower periodicity pitch. For example in the so-called diphonic singing styles of Mongolia, spectral pitch carries the melody while the low pitch serves as a drone. Some listeners seem more sensitive to one pitch and others to the other (e.g. Smoorenburg 1970). The two pitches are confounded for pure tones and distinct for complex tones within a restricted region of parameters (Fig. 5). It is common to attribute one to temporal cues, and the other to place cues derived from cochlear analysis, so that periodicity and spectral pitch would emerge from different mechanisms. Licklider (1959) suggested that these two mechanisms could be mapped together via Hebbian learning within a neural network. However one cannot exclude a common mechanism. The sharp

spectral locus implies quasi-periodicity that shows up as modes at short lags in the ACF (insert in Figure 5).

As noted by Mersenne, careful listening to a complex tone reveals higher pitches in addition to the fundamental. Helmholtz (1857, 1877) attributed each to the elementary *sensation* produced by a sinusoidal *partial*[10]. Percepts rarely emerge from these sensations, yet an argument for their existence is that they may be heard. They are, as it were, "latent" pitches. However, the more salient period pitch emerges despite the lack of a fundamental and thus, logically, in the absence of a corresponding sensation. Such partial-related sensations are neither necessary nor sufficient to produce a pitch. Their nature is problematic.

Pattern-matching models may account for partial pitches by supposing that a pitch may be associated with an *input* as well as the *output* of the pattern-matching stage (e.g. Terhardt 1991). Temporal models account for them by restricting processing to a particular channel from the periphery (e.g. Hartmann and Doty 1996). Partials are then audible (with attention) if they can be *isolated* in the temporal pattern of some channel. As Helmholtz (1857) noted, the salience of a partial increases if the partial is mistuned. Spectral theories might explain this by the action of a *harmonic sieve* (Duifhuis et al. 1982). Temporal theories might explain it as a local disruption of periodicity patterns. The systematic pitch shifts observed when partials are mistuned (Hartmann and Doty 1996) were accounted for by a model of de Cheveigné (1997, 1999).

In music, instruments often play together, each with its own pitch, and appropriately gifted people can perceive *several pitches* within a stimulus. Reverberation may transform a monodic melody into a polyphony of two parts or more (the echo of one note accompanying the next), and Sabine (1907) suggested that this is why scales appropriate for harmony emerged before polyphonic style. Methods and models for multiple period estimation are reviewed by de Cheveigné (1993) and de Cheveigné and Kawahara (1999). They may be classified into *spectral*, *spectro-temporal*, and *temporal*.

The *spectral* approach is based on analysis of the compound sound into partials. Partials belonging to each sound are then reassembled, and this allows a pitch to be derived from each. Parsons (1976) built a system for separating concurrent voices, based on a relatively high-resolution spectrum. A harmonic comb was fit to the spectrum (as in pattern-matching models) to determine one f_0. Spectral peaks that fit the comb were removed, and the remaining peaks were used to determine a second f_0. Scheffers (1983) tested the idea using spectral analysis similar to that occurring in the ear, but found that it rarely estimated both f_0s correctly.

---

[10] In a footnote, Helmholtz's translator Ellis points out that each "harmonic" may consist of a *harmonic series* with a fundamental frequency multiple of the fundamental of the tone, rather than a single partial at that frequency. A "partial pitch" could be assigned to the corresponding series rather than to the partial.

The *spectrotemporal* approach is exemplified by Weintraub (1985) or Meddis and Hewitt (1992). ACFs are calculated within each channel of a cochlear model, and on this basis channels dominated by one voice or the other voice are grouped together. Channels for each group may be added to estimate the corresponding f_0 (the models cited do not actually go that far).

A *temporal* approach is described by de Cheveigné (1993) and de Cheveigné and Kawahara (1999). Estimation is performed by applying as many time-domain comb filters as voices present. Adjusting their parameters for complete cancellation produces a set of period estimates.

11. 6 Binaural Effects

Binaural effects have played several important roles in pitch theory. The first was to show that hearing cannot use only the spectral cues allowed by Helmholtz's doctrine of phase-blind frequency analysis. Localization on the basis of binaural time-of-arrival cues (Thompson 1882) requires that they (and not just the spectrum) be represented internally. If time is represented internally, an account of pitch such as Rutherford's telephone theory becomes more plausible. Binaural release of masking (Licklider 1948; Hirsch 1948) had the same implication. In Huggins' pitch (Cramer and Huggins 1958), white noise presented to either ear of a listener, identical at both ears apart from a narrow *phase* transition at a certain frequency, evokes a pitch in the complete absence of spectral structure at either ear. This was strong evidence in favor of a temporal account of pitch.

The next role was on the contrary in favor of a spectral account. Huggins' pitch prompted Licklider to formulate the *triplex* model, an extension of his AC model including an initial stage of binaural processing. This involved a network of binaural delays and coincidence counters similar to the well-known localization model of Jeffress (1948). The output of each coincidence counter was fed to the input of an AC network. Some favorable interaural delay was selected, and pitch was then derived from the AC network attached to it. The triplex model thus used the *temporal structure* of the coincidence counter output to feed its pitch-extracting stage (in Jeffress's model it was simply smoothed over time). Binaural interaction in the Jeffress model is *multiplicative*, whereas the Equalization-Cancellation (EC) model of Durlach (1963) allows also *addition* or *subtraction*, that could also have been used to feed the triplex model. However Durlach chose instead to use the profile of activity across CFs as a *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.

Houtsma and Goldstein (1972) found essentially the same performance when partials of two-partial complexes went to the same or different ears. In the latter case there is no fundamental periodicity at the periphery, so they concluded that pitch cannot be mediated by a temporal mechanism such as Licklider's, 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 any peripheral channel. It can derive the period by combining measurements of individual partials, and it is but a small step to assume this can work from inputs of both ears. Furthermore, Houtsma and Goldstein report that performance was no better with binaural presentation, as one would have expected from better resolution of the partials in that case. Thus, their data could equally be construed as going *against* pattern-matching.

An improved version of the EC model was recently proposed by Culling, Summerfield, and Marshall (1998a,b; Culling 2000), that gives a good account of binaural pitches. As the earlier models of Durlach, or Bilsen and colleagues, it produces a tonotopic profile from which pitch cues are derived. However Akeroyd and Summerfield (2000) showed that one can also use the *temporal* structure at the output to derive the pitch (as in the triplex model). A possible objection to that idea is that it requires two stages of time domain processing, possibly a costly assumption in terms of anatomy. However de Cheveigné (2001) showed that the same processing may be performed as one stage.

11. 7 Harmony, melody and timbre

Pitch has aspects that do not fit the model of a linear perceptual dimension function of period. These include chroma, intervals, harmony, melody and other features important in music, as well as the relation between pitch and timbre. Accounting for them is a challenge for pitch models.

Chroma designates a set of equivalence classes based on the *octave*. In some contexts chroma seems the dominant mode of pitch perception. For example, *absolute pitch* appears to involve mainly chroma (Bachem 1937; Myazaki 1990; Ward 1999). Demany and Armand (1984) found that infants treated octave-spaced pure tones as equivalent. A spectral model accounts for octave equivalence in that all partials of the upper tone belong to the harmonic series of the lower tone. A temporal model accounts for it in that an integer number of periods of the higher tone fits within a period of the lower tone. 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. True octave *equivalence* is not an obvious emergent property.

Basilar membrane tuning or neural delays should be relatively stable, so absolute pitch should be the rule rather than the exception. Relative pitch involves the potentially harder task of abstracting interval relations between cues

along the periodotopic dimension. Nevertheless the opposite prevails: absolute pitch is rare. Depending on the scale, some intervals may 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. Again, interval perception, as underlies relative pitch, is not an obvious emergent property of the models.

Some aspects of harmony may be explained on the basis of simple relations between period counts (e.g. Boethius) or partial frequencies (Rameau 1750; Helmholtz 1877). Terhardt et al. (1982) and Parncutt (1988) explain chord roots on the basis of Terhardt's pattern-matching model. To the degree that pattern-matching models are equivalent to each other and to autocorrelation, similar accounts may be built upon other pitch perception models. The question of contextual effects remains [pointer to Emmanuel's chapter].

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. Historically there has been some overlap in the vocabulary (and concepts) used to describe pitch (e.g. "low" vs "high") or brightness (e.g. "sharp" vs "dull") (Boring 1942). In an MDS experiment Plomp (1976 - check) 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.

11. 8 Physiological models

[Pointer to Ian's chapter]

Models reviewed so far start from a notion of how pitch is extracted, and then look to physiologists for confirmation. Another approach is to start from known anatomy and physiology and work upwards. On the basis of, say, the physiology of cells in the cochlear nucleus (choppers, onset cells, etc.), the pattern of their projections to the inferior colliculus, and the physiology of cells at that stage, can one explain pitch?

This seems a good approach, as the only ingredients allowed in the model are those that exist in the auditory system. Some weaknesses are: (a) sparse sampling or technical difficulties may prevent observing an indispensable ingredient, (b) physiological experimentation and description themselves are model-driven, and in particular (c) the wrong choice of stimuli (for example amplitude-modulated pure tones with arbitrary carrier/modulator frequency ratios) or descriptive statistics (for example period, or first-order ISI histograms) 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 response patterns in response to amplitude-modulated sinusoidal carriers. The basic circuit has two inputs. One is a

succession of pulses 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 (Sec 9). A distinctive feature is the use of *two* delays 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) 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 "periodidicity equation" is true:

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

For a given set (τ_m, τ_c, τ_o) the required integers may actually not exist, so circuits tuned to certain periods may be missing from the array. In agreement with this fact, Langner did indeed describe steplike trends of psychophysical pitch matches, but Burns 1982 failed to replicate them. On the other hand, the equation is satisfied by many possible combinations of the six quantities that it involves. As a consequence, the behavior of the model 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 Sec 2.4. In terms of *periodicity* the (f_c, f_m) parameter space is non-uniform, and contains alternating regions of variable pitch clarity or ambiguity. The functional dependency of pitch on two parameters makes it appear that both must be extracted and used to determine pitch, and this is a factor of complexity in Langner's model. In contrast, a study *starting from pitch theory* might have used stimuli with parameters easier to relate to pitch, and produced data conductive to a simpler model.

In a different approach, Hewitt and Meddis (1994) suggested that cochlear nucleus (CN) *chopper cells* converge on coincidence cells within the central nucleus of the inferior colliculus (ICC). Choppers tend to respond with spikes regularly spaced at their characteristic interval. The chopping pattern tends to align to transients or periodicity and, if the period is close to the characteristic interval, it is *entrained*. A population of similar cells that align to similar stimulus features will fire precisely at the *same instant* within each cycle, leading to the activation of the ICC coincidence cell. A different period gives 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. Instead, it is the temporal alignment of spikes across chopper cells, rather than ISI intevals 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.

11. 9 Computer models

Material models were once common, but nowadays the substrate of choice is software. The many available software packages will not be reviewed, because evolution is rapid and information quickly outdated, and because up-to-date tools can easily be found by a web search (or by asking practitioners of the field).

The computer allows models of a complexity such that their principle or behavior 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. It is probably unavoidable that the gap between the complexity of the auditory nervous system on one hand, and our limited cognitive abilities on the other, must be spanned by hard-to-understand models. One should nevertheless worry when a researcher treats a model as opaque and performs experiments to understand how it works. Special mention should be made of the *sharing of software*. In addition to making model production much easier, it allows models to be *communicated*, including those that are not easily described.

11. 10 Other modeling approaches

Certain models assume an internal *map*, a topographic pattern (e.g. tonotopic or periodotopic) of firing activity over space. Whitfield (1970) argued that distinct behavioral responses must involve different neurons, and thus that the final representation must be a place code. Licklider (1959) questioned the need for an *orderly* distribution of, for example, period-specific responses. An unordered distribution might be functionally as effective.

Time-of-arrival has recently been suggested as an alternative to rate to represent scalar quantity, for example in the visual system (Thorpe et al. 1996). Networks based on time are formally as powerful, and in some case more powerful (in terms of function for a given network size), than networks based on rate (Maass, 1998). Time is a natural dimension of patterns processed by the auditory system, and using it to represent information internally makes sense. It is perhaps of significance that transient responses with millisecond latency accuracy have been found in the cortex (Shamma - check).

Barlow (1968) argued that a likely role of a sensory relay is to recode its input so that a common event evokes few spikes (possibly none), whereas a rare

event is coded in a more verbose fashion. The effect is to save spikes and reduce the metabolic cost of processing. Information is not lost, but redundancy is reduced in the sense that parts of the pattern that are implied by others need not be represented. An additional benefit is that input regularities are characterized by the *coding rule*. Barlow did not mention pitch, but periodic patterns are obviously a good candidate for such recoding, and the cancellation model (Sec. 9.5) actually follows this principle.

Recoding might occur in multiple stages. It might also be an integral part of the processing that implements, say, pattern-matching. According to Barlow's calculations, the greater the fan-out (number of output channels), the smaller the cost in metabolic terms. If each significant pattern is recoded to a single spike on a specific neuron, meaningful activity is likely to be hard to find by recording.

Cariani (2001) reviews a number of processing schemes, in particular involving internal oscillators and reverberant circuits. Stimulus periodicitity such that evokes a pitch could cause locking of such oscillators, or be recoded to a non-sychronous internal activity.

To summarize this Section, modern pitch models account equally well for the simpler aspects of pitch perception, and therefore complex features and second-order effects are important to understand what is going on. Some of them possibly involve particular mechanisms, but more likely they are facets of the same basic process, much of which is shared by other hearing functions. Multiple specialized *models* might be needed, but they are all models of the same target: the auditory system.

12 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. As the shadows on the back of Plato's cavern, 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 sanction 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, 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 restrain competition. At times progress is driven by the intellectual power of one person, such as Helmholtz. At others it seems retarded

by the authority of that same power. Controversy is important in science (Boring 1929), but its most intense phases are perhaps not the most fruitful.

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 limitations of his theory, 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 cast a long shadow across time, that is still felt today, and that is not entirely beneficial.

This chapter was built on the assumption that a healthy population of models is desirable. Otherwise writing sympathetically about them would have been much harder. There are those that 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". Some believe that the best quality of a theory is to be falsifiable, and put their efforts into falsifying them. If, as Hebb (1959) suggests, every theory is already false by essence, such efforts are guaranteed to succeed with every theory. Thus the criterion is less useful than it seems.

On the other hand, progress in science has been largely a process of weeding out theories. The appropriate attidude 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 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.

Conversely, *metaphors* are useful objects. Our experience of resonating objects (a piano, or du Verney's steel spring) makes the idea of resonance within the ear easy to grasp, and to 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, this did probably more for the advancement of knowledge than his own discoveries and inventions (Tannery and de Waard 1970).

13 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 is actually a spin-off of mechanisms that evolved for other purposes, for example to segregate sources, or to factor out redundancy within an acoustic scene. Mechanisms used for pitch are certainly used also for other functions, and thus advances in understanding pitch benefit our knowledge of hearing in general.

Ideally, understanding pitch should involve choosing, among a number of plausible mechanisms, the one used by the auditory system, in agreement with available anatomical, physiological or behavioral data. Actually, many schemes reviewed in Sect 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 deeply 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. This has little to do with pitch, and more with the fact that cancellation is also useful for segregation, and fits the ideas on redundancy-reduction of Barlow (1968). Cancellation could quite well be used by a pattern-matching model to measure periods of resolved partials. However the pattern-matching operation would still need accounting for. In the autocorrelation model (or its cancellation variant) it is embodied by a period-sized delay. Although the existence of adequate delays is controversial, this seems to me 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 autocorrelation nor pattern-matching. A reason to believe so is that many details described by the physiologist or psychoacoustician make little sense in their context. 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.'

14 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). Several 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. Boring (1942) provides an erudite and in-depth review of the history of ideas in hearing and other senses. 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.

15 Acknowledgements

I thank Ray Meddis, Daniel Pressnitzer, Laurent Demany, Stephen McAdams [others] for comments on this chapter.

16 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.

Akeroyd MA, and Summerfield AQ (2000) A fully-temporal account of the perception of dichotic pitches. Br. J. Audiol. 33(2):106-107.

AFNOR (1977) Recueil des normes françaises de l'acoustique. Tome 1 (vocalulaire), NF S 30-107, Paris: Association Française de Normalisation.

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: Author.

Bachem A (1937) Various kinds of absolute pitch. J. Acoust. Soc. Am. 9:145-151.

Barlow HB (1968) Possible principles underlying the transformations of sensory messages. In Kolers PA and Eden M (eds) Recognizing patterns. Cambridge Mass: MIT Press, 217-234.

von Bekesy G, and Rosenblith WA (1948) The early history of hearing - observations and theories. J. Acoust. Soc. Am. 20:727-748.

Bilsen FA, and Goldstein JL (1974) Pitch of dichotically delayed noise and its possible spectral basis. J. Acoust. Soc. Am. 55:292-296.

Bilsen FA (1977) Pitch of noise signals: evidence for a "central spectrum". J. Acoust. Soc. Am. 61:150-161.

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. 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.

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 (reproduced in Boring, 1963). Psychological Review 36:97-121.

Boring EG (1942) Sensation and perception in the history of experimental psychology. New York: Appleton-Century.

Boring EG (1953) The role of theory in experimental psychology (reprinted in Boring, 1963). Am. J. Psych. 66:169-184.

Boring EG (1963) History, Psychology and Science (Edited by R.I. Watson and D.T. Campbell). New York: John Wiley and sons.

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. P.N.A.S. 97:3183–3188.

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.

Cariani PA (2001) Neural timing nets. Neural Networks 14:737-753.

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.

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.
Cramer EM, and Huggins WH (1958) Creation of pitch through binaural interaction. J. Acoust. Soc. Am. 30:413-417.
de Cheveigné A (1989) Pitch and the narrowed autocoincidence histogram. Proc. 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, and Tsuzaki M (1997) A model of the pitch shifts of mistuned partials. Proc. Acoust. Soc. Japan spring meeting, 415-416.
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, and Kawahara H (1999) Multiple period estimation and pitch perception model. Speech Communication 27:175-185.
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 (Danmark).
de Cheveigné A, and Kawahara H (2002) YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111:1917-1930.
Cohen MA, Grossberg S, and Wyse LL (1995) A spectral network model of pitch perception. J. Acoust. Soc. Am. 98:862-879.
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.
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.
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.
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.

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.
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.
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.
Gabor 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 mechanicks and local motion, in four dialogues (translated by THO. WESTON, reprinted in Lindsay, 1973, pp 40-61). London: Hooke.
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.
Gray AA (1900) On a modification of the Helmholtz theory of hearing. J. Anat. Physiol. 34:324-350.
Grose JH, Hall JW, III, and Buss E (2002) Virtual pitch integration for asynchronous harmonics. J. Acoust. Soc. Am. 112:2956-2961.
Haykin S (1999) Neural networks, a comprehensive foundation. Upper Saddle River, New Jersey: Prentice Hall.
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.
Hartmann WM (1997) Signals, sound and sensation. Woodbury, N.Y.: AIP.
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, 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.
von Helmholtz H (1857 (translated by A.J. Ellis, reprinted in Warren & Warren 1968)) On the physiological causes of harmony in music. In 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.
Hirsch 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.
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.
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. Neuroscience 17:3312-3321.
Langner G (1981) Neuronal mechanisms for pitch analysis in the time domain. Exp. Brain Res. 44:450-454.
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.
Macran HS (1902) The harmonics of Aristoxenus (reprinted 1990, Georg Olms Verlag, Hildesheim). Oxford: The Clarendon Press.
Mersenne M (1636) Harmonie Universelle (reprinted 1975, Paris: Editions du CNRS). Paris: Cramoisy.
Meddis R (1988) Simulation of auditory-neural transduction: further studies. J. Acoust. Soc. Am. 83:1056-1063.
Meddis R, and Hewitt MJ (1991) 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 (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.
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 (1982) An introduction to the psychology of hearing (second edition). London: Academic 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.
Nordmark J (1963) Some analogies between pitch and lateralization phenomena. J. Acoust. Soc. Am. 35:1544-1547.
Nordmark JO (1968a) Mechanisms of frequency discrimination. J. Acoust. Soc. Am. 44:1533-1540.
Nordmark JO (1968b) 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 (translated by Lindsay, reprinted in Lindsay, 1973, pp 242-247. Poggendorf's Annalen der Physik und Chemie 59:497ff.
Parsons TW (1976) Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Am. 60:911-918.
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 (1987) A pulse ribbon model of monaural phase perception. J. Acoust. Soc. Am. 82:1560-1586.
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.

Patterson RD (1994a) The sound of a sinusoid: time-domain models. J. Acoust. Soc. Am. 94:1419-1428.
Patterson RD (1994b) The sound of a sinusoid: spectral models. J. Acoust. Soc. Am. 96:1409-1418.
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, and Levelt WJM (1965) Tonal consonance and critical bandwidth. J. Acoust. Soc. Am. 38:545-560.
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.
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. Journal of the Acoustical Society of America 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.
van Noorden L (1982) Two channel pitch perception. In Clynes M (ed Music, mind, and brain. London: Plenum press, 251-269.
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 (reproduced in "E.R. Jacobi (1968) Jean-Philippe Rameau, Complete theoretical writings, V3, American Institute of Musicology, 154-254), Paris: Durand.
[Rayleigh 1896]
Ritsma RJ (1962) Existence region of the tonal residue, I. J. Acoust. Soc. Am. 34:1224-1229.
Ritsma RJ (1963) Existence region of the tonal residue, II. J. Acoust. Soc. Am. 35:1241-1245.
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 mammlian 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.
Sauveur J (1701) Système général des intervales du son (translated by R.B. Lindsay as "General system of sound intervals and its application to sounds of all systems and all musical instruments", reprinted in Lindsay, 1973, pp 88-94). Mémoires de l'Académie Royale des Sciences 279-300:347-354.
Scheffers MTM (1983) Sifting vowels. PhD Thesis Gröningen.
Schouten JF (1938) The perception of subjective tones. Proc. Kon. Acad. Wetensch (Neth.) (reprinted in Schubert 1979, 146-154) 41:1086-1094.
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, Ritsma RJ, and Cardozo BL (1962) Pitch of the residue. J. Acoust. Soc. Am. 34:1418-1424.
Schouten JF (1970) The residue revisited. In Plomp R and Smoorenburg GF (eds) Frequency analysis and periodicity detection in hearing. Sijthoff, 41-58.
Schroeder MR (1968) Period histogram and product spectrum: new methods for fundamental-frequency measurement. J. Acoust. Soc. Am. 43:829-834.
Schubert E (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.
Semal C, and Demany L (1990) The upper limit of musical pitch. Music Perception 8:165-176.
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.
Siebert WM (1968) Stimulus transformations in the auditory system. In Kolers 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.
Slaney M (1991) Visualizing sound with auditory correlograms. J. Acoust. Soc. Am. (draft)
Smoorenburg 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, 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.
Terhardt E (1991) Music perception and sensory information acquisistion: relationships and low-level analogies. Music Perception 8:217-240.
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.
Todd NP, and Cody FW (2000) Vestibular responses to loud dance music: a physiological basis of the "rock and roll threshold"? J. Acoust. Soc. Am. 107:496-500.
Tong YC, Blamey PJ, Dowell RC, and Clark GM (1983) Psychophysical studies evaluating the feasability ofspeech 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. The Britsh Journal for the History of Science 10:1-24.
du Verney JG (1683) Traité de l'organe de l'ouie, 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: John Wiley and sons.
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 (reproduced in Schubert 1979, 201-211). Phys. Rev. 23:266-285.
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.
Weintraub M (1985) A theory and computational model of auditory monaural sound separation. PhD Thesis Stanford.
Wever EG (1949) Theory of hearing. New York: Dover.
Wever EG, and Lawrence M (1954) Physiological acoustics. Princeton: Princeton University Press.
Whitfield IC (1957) The physiology of hearing. Progr. in Biophys. and Biophys. Chem. 8:2-47.

Whitfield IC (1970) Central nervous processing in relation to spatio-temporal discrimination of auditory patterns. In R. P 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.

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 (plus 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.