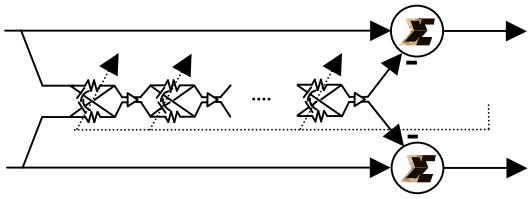
Madaline Model of Musical Pitch Perception

Jon Dattorro



In fulfillment of the requirements
for EE373**B**
Neural Networks
Stanford University
Spring 2000

Prof. Bernard Widrow
Aaron Flores, TA

Revision date: December 18, 2000
This document originated in Microsoft Word 2000,
and is available from the author's website.

Introduction: The Existence of Templates for Pitch Perception

No universal algorithm, model, or formula yet exists that can predict the perceived pitch of a tonal complex[1] to the degree of precision we would like (0.1 cent or better)[2]. In my opinion, the *central processor theory of pitch perception*, invented by professor Julius Goldstein[3] in 1973, [2] [3] [16] comes closest to providing an accurate formula. Goldstein's theory is based on the assumption that neural information is not sharply defined at spatial positions of resonance maxima along the basilar membrane, but instead fluctuates statistically about average values. Goldstein proposes that the mature auditory system's response is to minimize the difference between the input stimulus and some stored *template*. These templates are presumed formed by parental speech patterns during childhood. [7]

In this project, we develop an analog signal processing model for pitch perception. Because pitch templates will likely vary from person to person, there can be no universal model of pitch perception. Hence, any good model that we may develop must account for these natural variances.[4] Therefore, the hope of a high precision model may only be likely for a single individual.[5]

Our most important insight is that the human pitch processor behaves, from an engineering perspective, like an imperfect autocorrelator; equivalently, an imperfect periodicity detector. What is novel, and presented herein, is that we illustrate a plausible neural network that performs periodicity detection and inherently possesses a source of imperfection that creates errors in the periodicity estimate. We show that the errors made by the model are much like the human errors when compared to an autocorrelator. Consequently, we adopt the view that there is a strong bond between periodicity detection and human pitch perception.

---

[1] In the auditory analysis literature, a **complex** is defined as an audible tone that is not sinusoidal; i.e., a tone having more than one partial. A **partial** is a component sinusoid that is not necessarily harmonically related to the other partials in the complex. The first partial in an inharmonic complex shares the same ordinal position as the second harmonic in a harmonic complex. When partials are evenly spaced in frequency at multiples of the fundamental frequency f_o, as in $n f_o$, then they are called **harmonics**. The first harmonic (n = 1) is also called the **fundamental**.

[2] $\#cents = 1200 \ \log_2(f_2 / f_1)$

[3] Goldstein, Julius L., professor of electrical engineering, Washington University, St. Louis, MO 63130, (314) 935-9821, jlg@ee.wustl.edu

[4] I remember that as a young child, I was unable to discern the pitch of the very lowest notes on the piano. When I became older, I was finally able to ascribe pitches there.

[5] There is the ancillary phenomenon called *perfect pitch* that we only mention here. It is often called *absolute pitch*; the ability to determine the absolute frequency of a pitch percept, without an external reference. Not only do some individuals possess templates for pitch perception, as we all do, but they also possess an absolute frequency reference. Perfect pitch is not present in the majority of people, however, and there are fewer instances of famous musicians having this gift than not. Ludwig van Beethoven is a noteworthy example of an eminent musician possessing perfect pitch. Ironically, he became quite deaf later in life. Still, he managed to compose some of his greatest work in spite of his deafness; e.g., his 9th symphony, popularly referred to as "Ode to Joy". Seji Ozawa, renowned conductor of the Boston and San Francisco symphony orchestras, is well known for his demand for a positive deviation from the tuning standard 440 Hz (the note A_4 on the piano). All the orchestras of the world tune their instruments to that standard, except when conducted by Ozawa who insists upon the higher reference 442 Hz.

Some Pitch Model Requirements

A great scientist and professor of psychoacoustics, Ernst Terhardt, [6] agrees in essence with Goldstein, and explains the learning process of pitch perception in his 1978 paper "Psychoacoustic Evaluation of Musical Sounds". We provide a synopsis of his elucidation for which there exists ample evidence: [1]

The Bullets

- 1**)** The auditory system is able to **resolve** several partials of a complex tone; to perceive them separately.[6]

- 2**)** It is assumed that the auditory system is able to recognize and memorize the frequency intervals existing between those partials which are normally resolved. In typical human life, certain periodic signals having various fundamental frequencies are of highest biological relevance and occur often; the voiced sounds of speech. Hence, the chance is high that the auditory system will memorize that specific harmonic arrangement of partials as a matter of course; a template.

- 3**)** More generally, when a higher sensory system has memorized a specific stimulus arrangement, this particular arrangement, if presented again, will be perceived as a distinct entity rather than as a meaningless complex.

- 4**)** When a higher sensory system has developed the ability to perceive a pattern as an entity, the identification of that pattern will not break down when some of its elements are absent, provided that the remaining elements are objectively sufficient to specify the pattern.[7]

- 5**)** The higher sensory system may develop the habit to identify even a complete pattern preferably by means of only a restricted set of all the elements. Thus it is assumed that in the specific case of pitch perception, the hearing system deduces the pitch of a complex tone from those partials which lie in a frequency region where the system itself is working optimally; the speech band, approximately 300 to 1500 Hz.[8]

[6] It is well known that trained musicians have highly superior resolving power when compared to laymen.
[7] A good example of this is a very low piano tone, whose fundamental is always physically nearly absent **(**even on the best concert grands**)**. Another example is the loudspeaker in cheap transistor radios, that are incapable of producing frequencies below about 100 Hz, the fundamental frequency region of the male speaking voice.
[8] Ritsma [4] corroborates Terhardt's assumption.

Autocorrelation

A model of pitch perception based upon the process called *autocorrelation* could be made to satisfy the bulleted requirements. Bracewell [10,pg.41] defines the autocorrelation of a real signal x like so:

$$r(\tau) \triangleq \int x(t)\ x(t-\tau)\ dt \qquad (1.1)$$

Autocorrelation can be regarded as a good measure of periodicity or, equivalently, of *self similarity* as a function of delay τ. Other researchers have made much progress with autocorrelation-based models. [11] [12] Models based upon autocorrelation will not account for the *stretched template* phenomenon, however.

Stretched Templates

Terhardt professes that the templates formed through learning the harmonic patterns of speech are themselves stretched along the frequency axis during the natural course of human development. [1] [5] The aural impact of template stretching on perceived pitch is striking and musically significant; from an engineering perspective, it is second order. One simple analytical description of template stretching is,

$$f_n = n^{\sigma} f_o \qquad ; n = 1,2,3\ldots \qquad (1.2)$$

where f_n is the frequency of the n^{th} stretched harmonic, σ is the *stretch factor* nominally equal to 1, and f_o is the perceived fundamental frequency. When $\sigma = 1$ exactly, then there is no stretching. In that case, Equation (1.2) describes a harmonic series; Fourier series frequencies. When $\sigma > 1$, then the frequencies become stretched.

Any description of template stretching must satisfy the defining condition:

$$f_{n+2} - f_{n+1} > f_{n+1} - f_n \qquad (1.3)$$

There are many mathematical descriptions of stretching that will satisfy Equation (1.3). If we momentarily adopt the description in Equation (1.2), then Terhardt's stretch data tells us that values of σ close to 1 correspond well to human pitch templates;

$$\sigma = 1 + \iota \quad ; \iota \geq 0 \qquad (1.4)$$

where ι (iota) is on the order of 0.01.

In addition to the requirements specified in the bullets above, any accurate pitch model we develop will need to account for stretching of the template. Later we will show a simple analog neural network model that, at once, accounts for this stretching phenomenon, and subsumes autocorrelation. The hypothetical Equation (1.2) is now left behind.

Examples of Pitch Percept

We adopt the view that, in humans, pitch processing is almost synonymous with periodicity processing; i.e., autocorrelation and pitch perception share some characteristics that lead us to believe there must be some relationship between the two processes. [11] [12] Before proceeding with the development of our analog model, it is prudent to be apprised of the expected human response to some common test stimuli. We learn, from the following examples, that human periodicity processing is inherently flawed; that is to say, the periodicity estimate produced by autocorrelation is not equivalent to the human pitch percept to a high degree of precision.

Example 1: One Sinusoid.

One thing that we agree upon [8] by convention is the sensation of pitch produced by a lone sinusoid; the pitch frequency is the sinusoid frequency in this case. A lone sinusoid is therefore utilized as a pitch reference against which any other tonal complex can be aurally compared.[9]

The periodicity and pitch frequency are identical in this case.

Example 2: Harmonic Complex.

Consider a tone A_3 having the Fourier series of fundamental frequency 220 Hz, shown in Figure 1(a). The phase is not important. The perceived pitch is about 218 Hz. [1,pg.488] Ignorant of Terhardt's stretched template theory, one may have guessed that the pitch would be closer to the actual fundamental because of the perfect waveform periodicity.

Terhardt found that his subjects, who were able to resolve individual harmonics, dependably reported higher than actual frequency for each component sinusoid. [5] The first harmonic, however, was reported to be about 218 Hz; 2 Hz lower in frequency than the actual fundamental. The departure is also illustrated in Figure 1(a).

The periodicity of the composite waveform, in this case, is exactly 220 Hz. The autocorrelation $r(\tau)$ has a distinct peak at $1/\tau = 220$ Hz as shown in Figure 2. This means that an autocorrelation model would not reflect the inherent inaccuracy of periodicity perception in the normal human hearing system.

[9] There remains a dependency of the pitch percept upon absolute amplitude. Terhardt also points out that there may even be a difference in pitch percept between the two ears of the same individual. [8] Nonetheless, we agree to an ANSI standard, demanding that the sinusoid have about the same perceived amplitude as the tonal complex under scrutiny. Long before the ANSI standard, musicians used a metal tuning fork that, when held against the mastoid bone behind the pinna, produces a nearly perfect sinusoid.

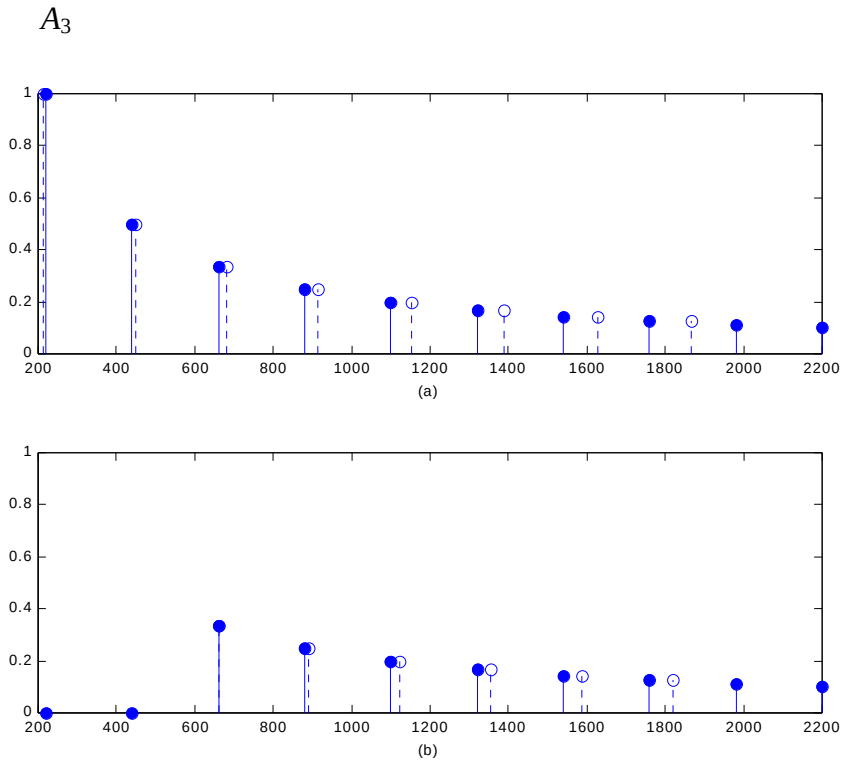


Figure 1. Solid is actual input stimulus, dotted is perceived.
(a**)** Example 2. Amplitude spectrum (1/*n*) of harmonic complex A_3 . Actual f_o = 220 Hz .
(b**)** Example 3. Same input as before except first two harmonic amplitudes zeroed.

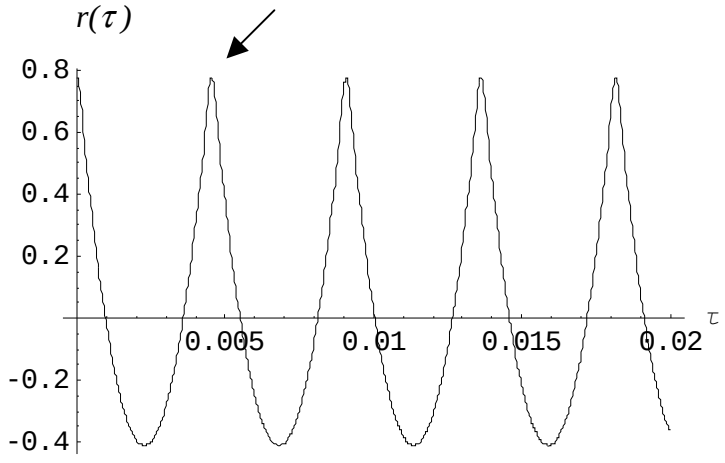


Figure 2. Autocorrelation of the harmonic complex in Example 2. The periodicity estimate (at the arrow) is not equal to the pitch percept.

Example 3: Harmonic Complex Absent Fundamental and Second Harmonic.
Now consider again the same tone A_3 having a Fourier series of fundamental frequency 220 Hz, but having both the fundamental and second harmonic amplitudes set to 0. The perceived pitch is now about 215 Hz. [1,pg.489] The pitch is lower than it is when all the harmonics are physically present as in Example 2!

Terhardt's subjects reported higher than actual frequency for each component sinusoid, except for the third harmonic (660 Hz) which was reported at its nominal sinusoid frequency. [5] This departure is illustrated in Figure 1(b).

The periodicity of the composite waveform is, once again, exactly 220 Hz as shown in Figure 3. Autocorrelation again fails to predict the human pitch percept accurately in this case.

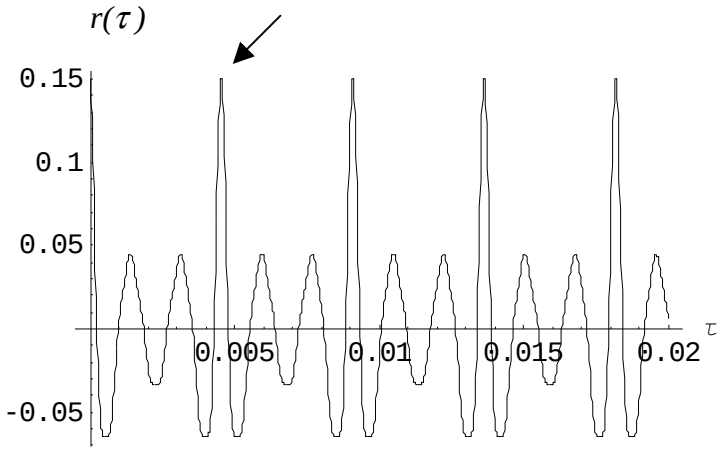


Figure 3. Autocorrelation of the signal in Example 3. The periodicity estimate (indicated by the arrow) is not the same as the perceived pitch.

Summary of the Examples

Pitch perception, like autocorrelation, is largely insensitive to the phase of individual partials. It is not difficult to conceive an autocorrelation model that could be made to account for the phenomena listed in the bulleted requirements. These facts motivated the use of autocorrelation as a starting point for the development of an accurate pitch model. From these few common examples we learn, however, that autocorrelation can only approximate the inherently inaccurate periodicity processing in the human hearing system. For an autocorrelation model to more accurately reflect human periodicity processing, the stretched template phenomenon must be incorporated.

Harmonic Template Model of Pitch Perception

From an engineering perspective, the human hearing system may be said to possess an inaccurate periodicity detector that is used for determining pitch. We learned from the examples, that if an autocorrelation process exists in the human hearing system for periodicity detection, then that process must not be perfectly implemented there. The question we hope to shed some light upon pertains to how the human pitch processor is neurologically constructed. In this section, we look at an alternative but equivalent implementation of autocorrelation called "self similarity". Based upon this alternative implementation, we posit a plausibly inaccurate neurological autocorrelator in the next section.

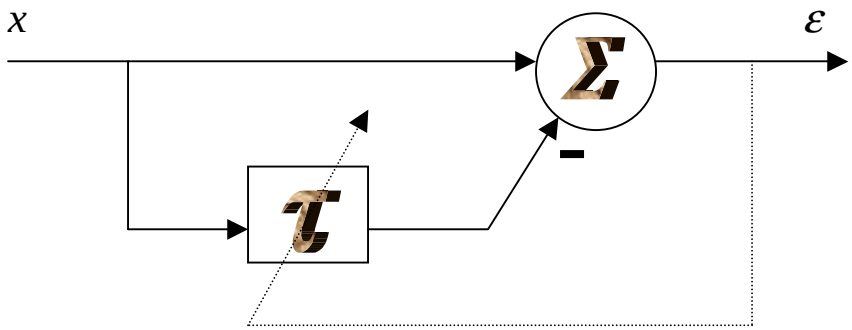


Figure 4. Self similarity (or harmonic template) model of pitch perception, having variable delay τ. This model does not give accurate estimates of the pitch percept.

The self similarity model in Figure 4 is a classical adaptive system. [9] Since we assume that there is some connection between waveform periodicity and pitch perception, [11] the model above is designed to estimate periodicity. Periodicity is detected by minimizing the difference between the input signal x and a delayed replica of itself. Hence, this pitch model is called a self similarity model. Next, we derive the equivalence between autocorrelation and self similarity.

For the self similarity model, the delay at the minimum difference, $\tau_{\min}$, is the periodicity estimate. This minimizing delay is identical to what would be found by maximizing the autocorrelation of the input signal; viz.,

$$\tau_{\min} = \arg\min_{\tau} \int \big(x(t) - x(t-\tau)\big)^2 \, dt = \arg\left(k - \max_{\tau} \int x(t)x(t-\tau)dt \right) \qquad (1.5)$$

where the value of k is constant for a particular x and is independent of delay τ. This Equation (1.5) says that the τ found by maximizing the autocorrelation is identical to that found by minimizing a self similarity function.[10] Hence the equivalence of autocorrelation and self similarity.

In other words, the periodicity estimate of autocorrelation can be produced by minimizing the signal power with respect to τ at the output ε of the self similarity model in Figure 4. This point of view has a nice frequency domain interpretation that leads to the harmonic template. Employing Rayleigh's theorem [10,pg.112] we may write

$$\tau_{\min} = \arg\min_{\tau} \int \big(x(t) - x(t-\tau)\big)^2 \, dt = \arg\min_{\tau} \int \big|X(f)\big|^2 \, \sin^2(\pi f \tau) \, df \qquad (1.6)$$

The template is now revealed as the $\sin^2(\pi f \tau)$ term in the frequency domain from Equation (1.6); i.e., the normalized magnitude square Fourier transfer function of the model in Figure 4,

$$\begin{aligned} H_{\mathrm{f}}(f) &= \frac{E(f)}{X(f)} = 1 - e^{-j2\pi f \tau} \qquad ; \boxed{E(f) \Leftrightarrow \varepsilon(t)} \\ \big|H_{\mathrm{f}}(f)\big|^2 &= 4\sin^2(\pi f \tau) \end{aligned} \qquad (1.7)$$

Plotted for various fixed τ, the template appears as in Figure 5.

[10] The equivalence follows simply by expanding the square. More specifically, from Equation (1.5) we learn that maximizing the autocorrelation is equivalent to minimizing the particular self similarity function whose metric is the 2-norm. It is more than likely that the human hearing system uses a different metric. Use instead of the 1-norm, for example, yields different $\tau_{\min}$ for aperiodic x. We do not consider other metrics any further.

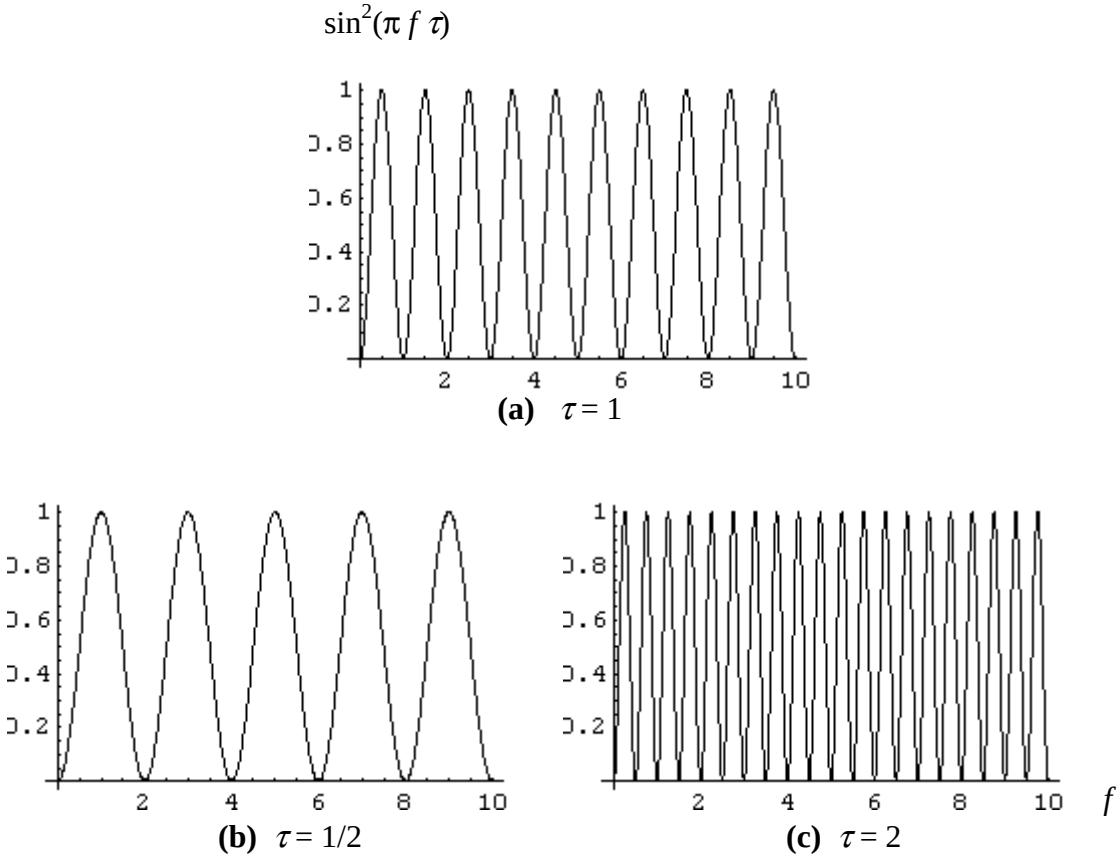


Figure 5. The harmonic template $|H_f(f)|^2/4$ evaluated for various τ.

From Figure 5 we see that the template is simply a comb filter having zeroes at multiples of the inverse delay $1/\tau$. The template is harmonic because the zeroes are spaced at integer multiples of $1/\tau$. From Equation (1.6) we see that as τ is adapted, the comb attempts to cancel the input signal spectrum. When the input spectrum is itself harmonic, as are the voiced sounds of speech, then the comb could annihilate it if τ were adjusted properly. Hence, the self similarity model in Figure 4 is also called the harmonic template model.

More generally, for any model transfer function $H(f, \tau)$ we may say that

$$\tau_{\min} = \arg\min_{\tau} \int |X(f)|^2 \, |H(f,\tau)|^2 \, df \qquad (1.8)$$

where the template $|H(f,\tau)|^2/4$ is also a function of τ. The estimate of pitch percept is $1/\tau_{\min}$. Since the function to be minimized is typically multi-modal,[11] the assumed sense of minimization is local.

[11] We will examine the multi-modal behavior later in Equation (1.15) and the associated figures.

Madaline Model[12] of Pitch Perception

Recall that the harmonic template model in Figure 4 is equivalent to an autocorrelator. We now present an analog network that can imitate the behavior of the harmonic template model. We show that this imitation becomes exact only in the limit of a multiplicity of component stages. In this analog network, we will for the first time account for stretching of the templates as proposed by Terhardt. [1] [5]

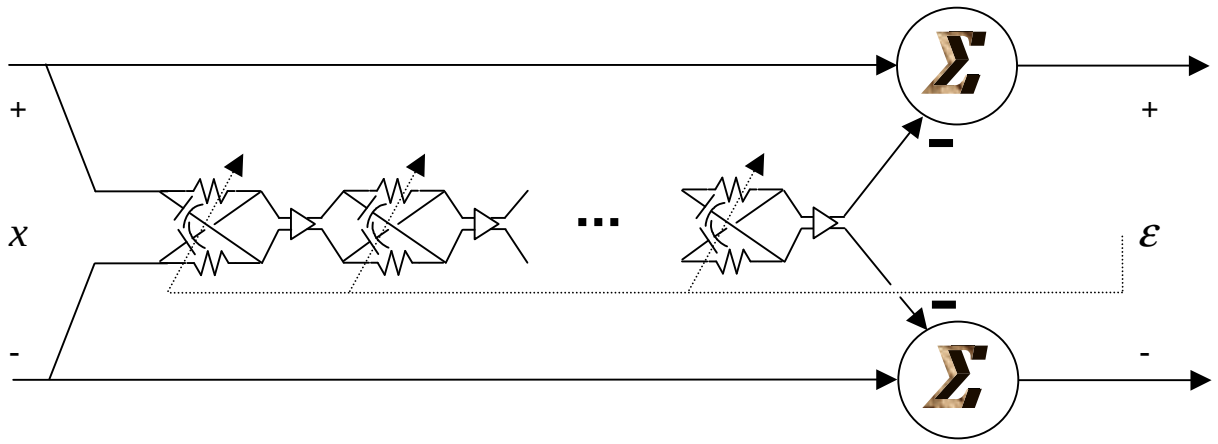


Figure 6. Neural network model of pitch perception. High impedance differential buffers (the triangles) appear at each *axon* pair output.[13] This model can produce more accurate estimates of pitch percept.

We propose that the electrical network in Figure 6 is a first cut in the design of a plausible neurological model of the human pitch processor. We discuss the plausibility of this simple model, but we leave it to Haykin to justify the circuit components. [14,ch1.2] Each RC lattice and differential buffer[14] together form a *neuron* pair. All neuron pairs are replications; i.e., all resistors and capacitors have the same respective value (each neuron pair makes one allpass filter stage, and all the stages are identical). Those equal values are presumably a result of the childhood learning process and template formation that Terhardt and Goldstein postulate. [1] [2] [3] [7] [16] We further presume that the RC product (resistance times capacitance) remains variable and adaptable throughout life after learning has ended. This cascade of identical adaptable neuron pairs constitute a special case of Widrow's analog madaline.

---

[12] The terms adaline and **madaline** were coined by Bernard Widrow, madaline meaning **m**ultiple **adaline** [13,ch.2.2.3]; i.e., **m**ultiple **ada**ptive **li**near **ne**uron.

[13] An **axon** is a biological component of a neuron, and may be modeled as an RC transmission line. [14]

[14] The buffer may also emulate the so called **activation function** of the neuron, but we do not investigate that possibility here. The activation function is just a nonlinearity; [14] usually a threshold function **(**the step or Heaviside function [10]**)** or a sigmoidal function. The presence of a nonlinearity would make our model more realistic, but also makes analysis difficult.

With $H_p(f)$ for the network in Figure 6 defined like we did in Equation (1.7),[15] we have

$$H_p(f) = \frac{E(f)}{X(f)} = 1 - \left(\frac{\frac{\mathrm{N}}{\tau} - j\pi f}{\frac{\mathrm{N}}{\tau} + j\pi f} \right)^{\mathrm{N}} = 1 - e^{-j2\mathrm{N}\arctan(\pi f \tau / \mathrm{N})} \tag{1.9}$$

$$\left|H_p(f)\right|^2 = 4\sin^2\left(\mathrm{N}\arctan\left(\frac{\pi f \tau}{\mathrm{N}}\right)\right)$$

where N is the total number of neuron pairs in Figure 6, and where each neuron pair contributes only a portion of the delay τ according to the formula,

$$2\ \mathrm{R}\ \mathrm{C} = \frac{\tau}{\mathrm{N}} \tag{1.10}$$

Notice the similarity between Equation (1.7) and Equation (1.9). The network in Figure 6 reduces to the model in Figure 4 in the limit of an infinite number of neuron pairs; [15]

$$\lim_{\mathrm{N}\to\infty} H_p(f) = H_{\mathrm{f}}(f) = 1 - e^{-j2\pi f \tau} \tag{1.11}$$

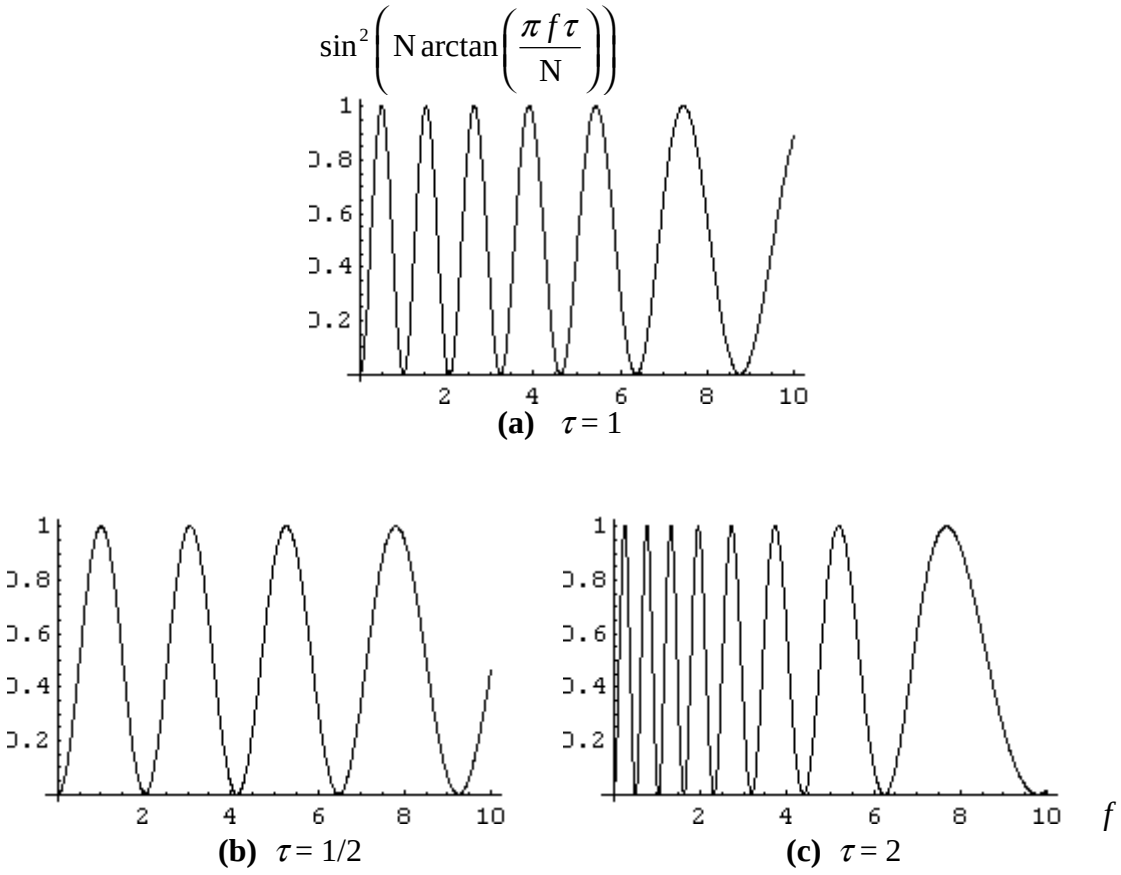
Hence Equation (1.9) becomes identical to Equation (1.7) in the limit. Thus for an infinite number of identical adaptable neuron pairs, we have succeeded in constructing a variable delay τ out of resistors and capacitors. Therefore, the neural network model in Figure 6 can, in the limit, be made to perform autocorrelation as proved by Equation (1.5).

It is more plausible that any physical network would employ only a finite number of neuron pairs. The neural network behavior, then, only approximates an autocorrelator. From Example 2 and Example 3, we learned that the human pitch processor can be viewed as an inaccurate periodicity detector. Hopefully, this analog approximation in Figure 6 is the de facto approximation that the human hearing system makes. If that is the case, then we have discovered the main source of inaccuracy in the human pitch processor.

[15] The subscript p stands for *phaser*, while the subscript f stands for *flanger*. [15]

In contrast to the harmonic template for the self similarity model in Figure 4, the template for our neural network model in Figure 6 is stretched as illustrated in Figure 7. Compare the stretched template for our neural network, to the harmonic template in Figure 5. A stretched template, whose zeroes satisfy Equation (1.3), is an essential characteristic of *any* model in light of Terhardt's theory. [1]

The particular stretching characteristic of our neural network model is an artifact induced by the cascade of a *finite* number of identical neuron pairs attempting to emulate a delay. The exact mathematical description of the stretching will depend upon the number of pairs N utilized by the model. Regardless of N, Equation (1.3) remains satisfied.[16]



(a) $\tau = 1$

(b) $\tau = 1/2$

(c) $\tau = 2$

Figure 7. The stretched template $|H_p(f)|^2 / 4$ evaluated for N = 20 and various τ.

[16] We speculate that the nominal number of neuron pairs in a neurological cascade specialized for pitch perception, is a variable and characteristic of each individual human. The number is most likely dependent upon the extent (or lack) of musical training, while a severe deficit may explain what is commonly known as "tone deafness". Tone deafness might also be explained by a cascade of neuron pairs that were never trained properly; i.e., each neuron pair in the cascade is not identical to the next.

Accounting for the Bullets

Bullet 1**)** The ability to resolve partials is achieved via the selectivity of the comb shape.

Bullet 2**)** The template is the stretched comb.

Bullet 3**)** The fact that the pitch percept is perceived as one entity is related to the replication of identical neuron pairs. Recall that the cascade of replicated neuron pairs constitute a single delay, in the limit of replication number. The network adapts that delay to a single value in response to a particular harmonic stimulus.

Bullet 4**)** Because a template is constrained to have nearly harmonically spaced troughs, the absence of one or a few harmonics in a stimulus does not radically change predicted pitch. Output ε signal power minimization demands that a template be aligned over the remaining harmonics. The spacing between template trough frequencies, then, becomes most important.

Bullet 5**)** Stretched templates and the fact that this network only approximates an autocorrelator, make our model plausible. But to make the network more realistic, we insert a frequency dependent transmission loss into the delay path. If the transmission loss is modeled properly, then a restricted set of partials will predominate the template.

Transmission Loss

Recall that all the neuron pairs are presumed identical in Figure 6 as a result of template development during childhood. We hypothesize the existence of delay path transmission loss that is caused by respective mismatches in the resistive and capacitive elements. Because those elements are variable during adaptation in response to a stimulus, then the transmission loss must itself be variable and must somehow track τ in Equation (1.10).

For simplicity, we lump the delay path transmission loss all together into $\alpha(f)$ as shown in Figure 8.[17] We do not know the exact form of $\alpha(f)$ but its presence in some form is necessary to corroborate the earlier examples, Terhardt's pitch "trend" data [1,Fig.8], and Goldstein's pitch "striation" data [16,Fig.1].

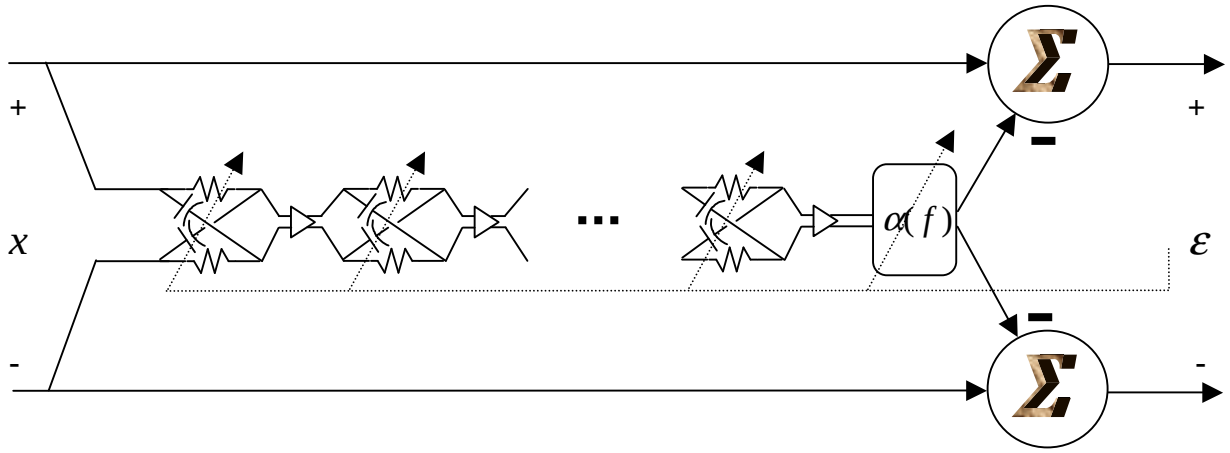


Figure 8. Neural network model of pitch perception showing transmission loss $\alpha(f)$. This lossy model produces the most accurate pitch estimates thus far.

[17] The network in Figure 8 simply represents transmission loss inserted into the delay path of the network in Figure 6. We recognize that transmission loss may in fact be distributed among the differential buffers.

Terhardt's Pitch Trend Data

Terhardt's trend data, [1,Fig.8] the solid curves sketched in Figure 9, shows that the stretched template phenomenon causes larger departures in perceived pitch of harmonic complexes from their fundamental frequency, as fundamental frequency goes down. The lower solid curve indicates that a more pronounced departure occurs when the fundamental and second harmonic are physically absent from the complex. Terhardt's data, shown sub-sampled and linearly interpolated between six points in Figure 9, should be regarded only as indicative of a trend; the data should not be considered exact because it will vary from person to person. If our model is to be realistic, it must predict this effect.

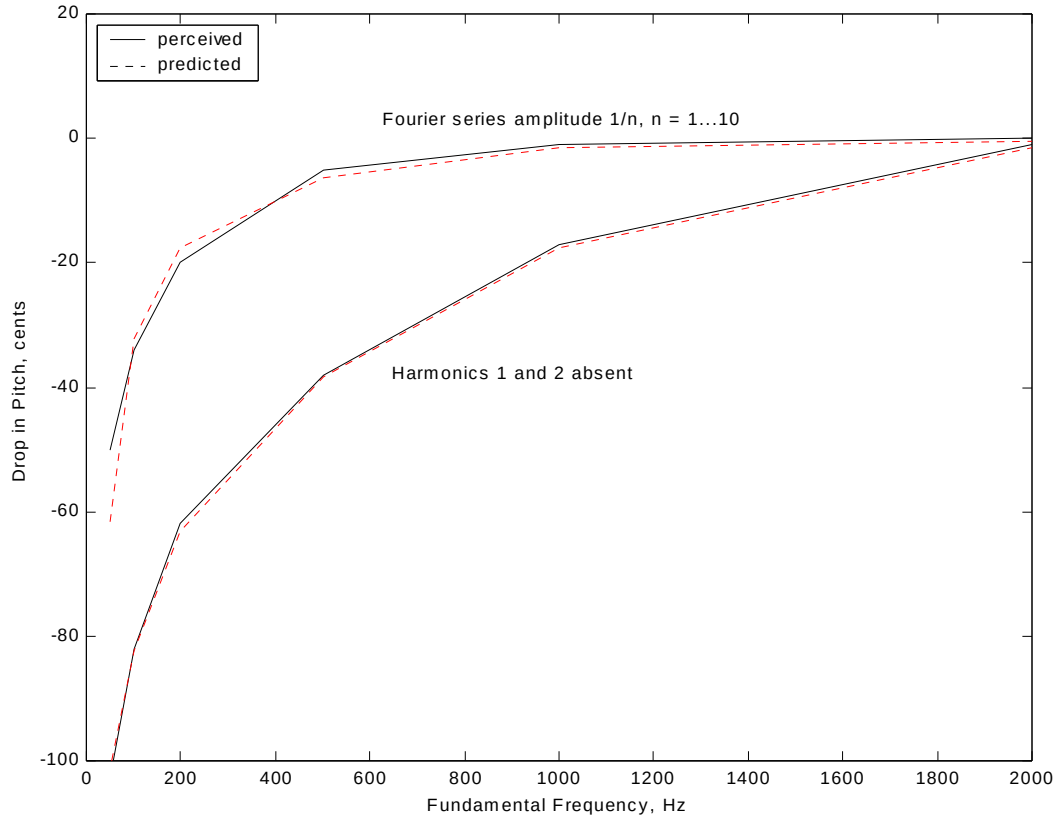


Figure 9. Solid curves show Terhardt's trend data; [1,Fig.8] the departure of perceived pitch from the fundamental of a harmonic complex. Plots are for six input fundamentals, at the breakpoints f_o = {50, 100, 200, 500, 1000, 2000} Hz, and for two different input types described in terms of their Fourier series. Dotted curves show the departure predicted by the model in Figure 8, N = 26.

We suppose that an all-pole second order analog filter can sufficiently model the transmission loss. Then, by using Terhardt's trend data in Figure 9, we can determine how that filter tracks τ. The Fourier transfer function of the analog filter we choose is,

$$\alpha(f) = \frac{\left(\frac{\Omega}{Q}\right)^2}{(j2\pi f)^2 + \frac{\Omega}{Q} j2\pi f + \Omega^2} \tag{1.12}$$

where Ω is the approximate peak center frequency in radians along the $j\omega$ axis, and where Q is the approximate[18] filter selectivity defined as the peak center frequency divided by the peak bandwidth.

By cut and try, we determined that for N = 26 the loss-filter tracking of τ wants to see

$$\left.\begin{aligned} Q &\approx 2 + 8\left(\frac{f_o}{50}\right)^{\frac{-1}{\ln(2)}} \\ &\\ &\textit{When amplitude } \frac{1}{n},\ n = 1...10, \\ \Omega &\approx (16.1395 + 3.96073\ f_o - 0.000130299\ f_o^2 + 4.65894\times10^{-8}\ f_o^3)2\pi \\ &\\ &\textit{When harmonics } 1 \textit{ and } 2 \textit{ absent}, \\ \Omega &\approx (24.3234 + 4.46174\ f_o - 0.000764190\ f_o^2 + 1.91367\times10^{-7}\ f_o^3)2\pi \end{aligned}\right\} \tag{1.13}$$

The delay τ tracks fundamental frequency f_o for harmonic stimuli. Hence, these ad hoc heuristic tracking equations are written in terms of f_o.

[18] We say "approximate" because for these two parameters to be exact, the filter transfer function would have a zero at DC.

The dotted curves in Figure 9 represent the pitch estimates made using the model in Figure 8 and the transmission loss filter in Equation (1.12) parameterized by Equation (1.13). The pitch estimates are calculated using Equation (1.8), where the input signal spectrum $X(f)$ is presumed to be composed of *spectral lines*,[19] and where the lossy model transfer function is

$$H(f) = H_{\alpha}(f) = \frac{E(f)}{X(f)} = 1 - e^{-j2\mathrm{N}\arctan(\pi f \tau / \mathrm{N})}\alpha(f) \qquad (1.14)$$

Figure 10 shows the transmission loss filters $\alpha(f)$ required to make the pitch estimates in Figure 9; Equation (1.12) parameterized by Equation (1.13). Each loss filter corresponds to one particular input signal x. Each dotted curve in Figure 10 represents the loss filter required when the first two harmonics are absent from the corresponding Fourier series constituting the input signal.

Because the parameter equation Ω depends upon the form of the input signal x, it is apparent that the loss filters are signal dependent. We speculate that this signal dependency might be reduced somewhat by taking into account the impact of the cochlear response curves illustrated in Figure 11. [17] [18] Estimating the pitch percept using Equation (1.8) would then require convolution of the input spectral lines with the corresponding cochlear frequency response.[20]

In Figure 12 we show the stretched template of the lossy neural network model in Figure 8 for various τ, and for the loss filter $\alpha(f)$ corresponding to an input x having Fourier series amplitude $1/n$, $n = 1\ldots10$.[21] The effect of the loss filter upon the template is to make benign the very high harmonics in the input signal, and to selectively weight the remaining harmonics.

[19] A **spectral line** is a Dirac delta function in the frequency domain; $\delta(f)$. It is a characteristic of the spectrum of periodic functions. [10]

[20] Because each cochlear response occupies such a broad spectral region, its domain crosses into neighboring harmonics. The sensitivity of the pitch estimate to the individual harmonic is thereby reduced. Because each cochlear response leaks more below its peak frequency than above, the input harmonics which are higher in frequency achieve more spectral dominance. Sensitivity of the pitch estimate would then be less for loss of low frequency harmonics than for loss of high frequency harmonics. A reduction in sensitivity translates to a reduction in signal dependency for the loss filters; but again, this is all speculative.

[21] Recall that the phase is not important. Compare these templates to those in Figure 5 and Figure 7.

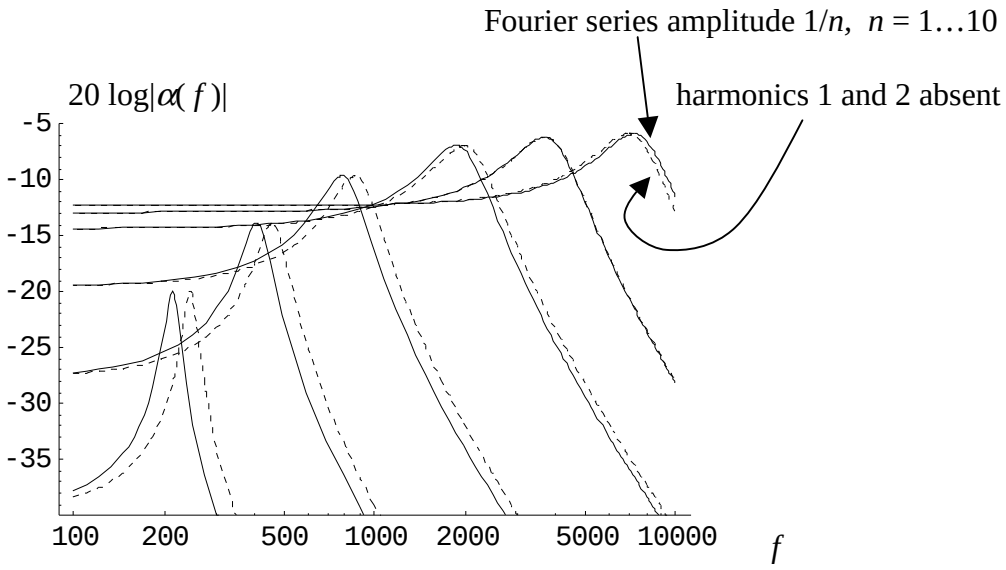


Figure 10. Transmission loss filters for N = 26 and f_o = {50, 100, 200, 500, 1000, 2000} Hz. Dotted curves show filter required when corresponding input x has a Fourier series missing its first two harmonics. The loss filter tracks τ, hence it tracks fundamental frequency f_o. Notice that the loss filter required for a fundamental frequency of 2000 Hz is centered at about 7.5 kHz. Pitch estimates are unaffected by scaling absolute filter amplitude.

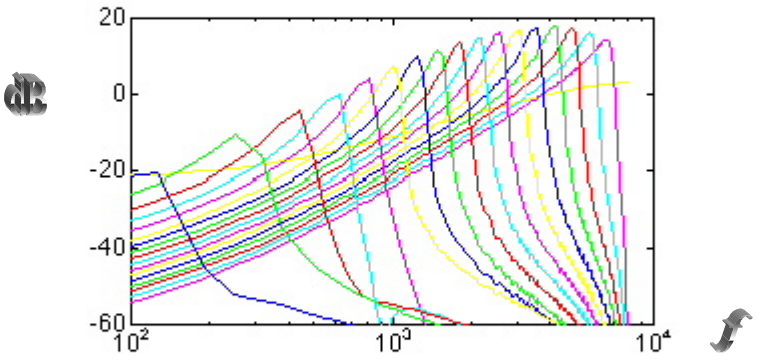


Figure 11. Lyon's cochlear model. [17]

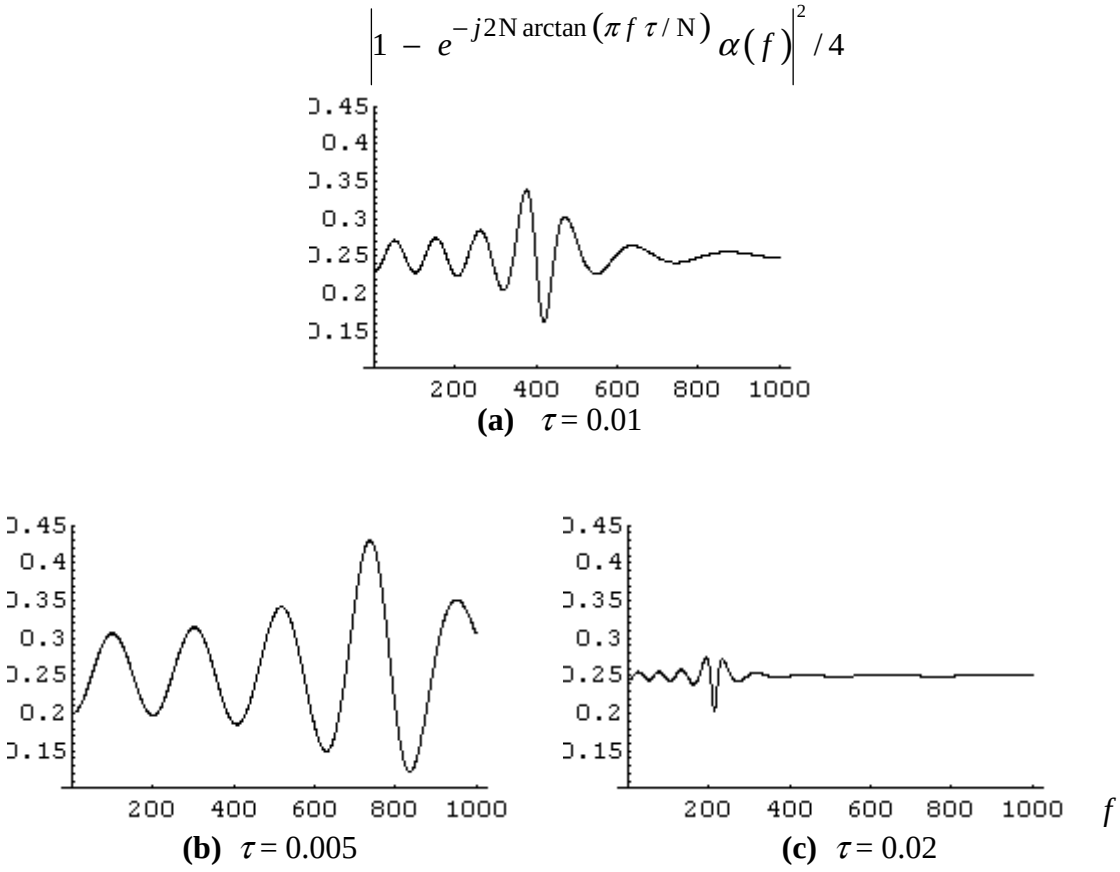


Figure 12. The stretched lossy template $|H_\alpha(f)|^2 / 4$ evaluated for $\mathrm{N} = 20$, various τ. The loss filter $\alpha(f)$ in Equation (1.12) tracks τ in these plots; i.e., $f_o \leftarrow 1/\tau$ in Equation (1.13). Each template asymptotically approaches a value of 0.25.

Goldstein's Pitch Striation Data

Now we leave behind harmonic stimuli; the input stimuli are no longer described well by a Fourier series. The stimulus we consider instead consists of four equal-amplitude sinusoids equi-spaced in frequency by $\Delta = 200$ Hz. The fundamental f_L in this inharmonic complex is randomly set between 20 and 800 Hz while the 200 Hz Δ is maintained. Because Δ remains constant while f_L varies, this stimulus is referred to as the *pitch-shift* stimulus.[22] Goldstein's striation data [16,Fig.1][23] is represented in Figure 13 illustrating the perceived pitch of the pitch-shift stimulus.

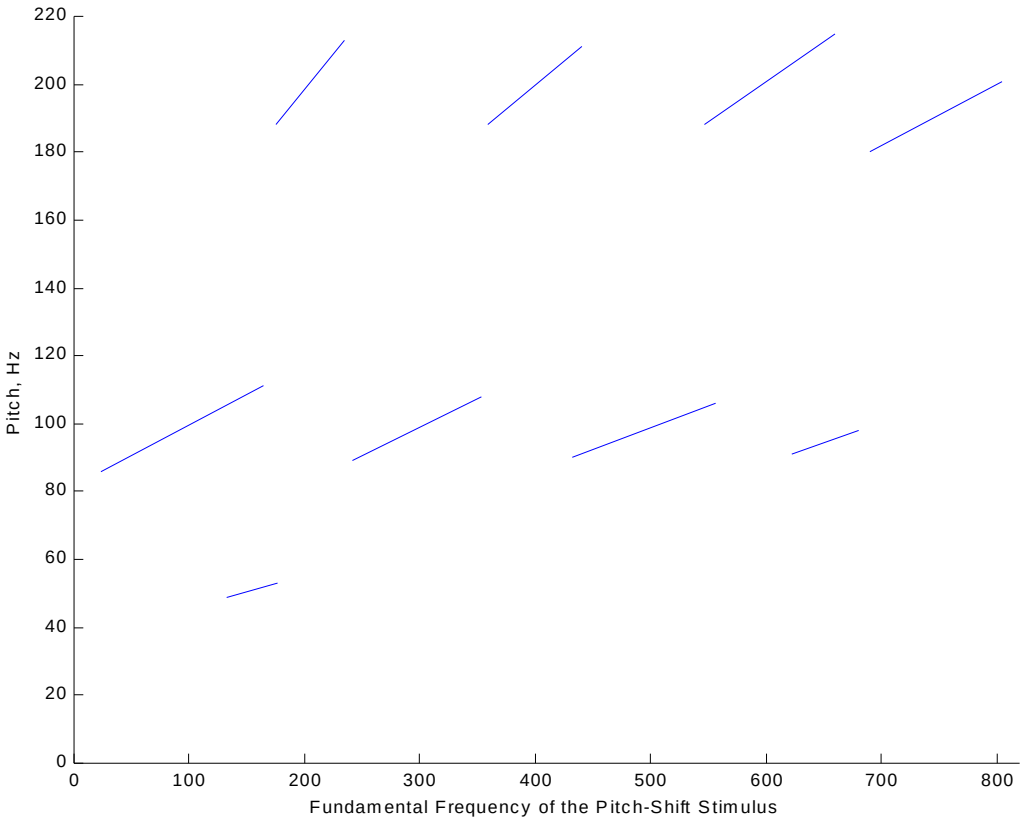


Figure 13. Striation curves show Goldstein's pitch data [16,Fig.1] averaged from one particular subject responding to the pitch-shift stimulus.

[22] For the special cases occurring when the greatest common divisor of this complex lies on the abscissa of the graph in Figure 13, only then is the pitch-shift stimulus considered harmonic.

[23] Goldstein's pitch reference is not a pure sinusoid as per the ANSI standard. Instead he uses the fourth through seventh equal-amplitude harmonics of a Fourier series as the pitch reference. As we are interested only in corroborating the striation characteristic that Goldstein discovered, the offset in perceived pitch consequent to the use of such a reference will be safely ignored.

By **striation** we mean the near-parallel slashes constituting each graphical row of data in Figure 13. The striations are interesting for three reasons: 1**)** The slope in a particular row decreases as f_L increases, 2**)** there are distinct rows of striations, and 3**)** the pitch is ambiguous. This particular subject exhibits distinct rows centered at pitches of about 50, 100, and 200 Hz. The striation data in Figure 13 is averaged data and valid only for one particular subject. The data, as we present it, should be considered indicative only of a trend. Indeed, data from other subjects can look different, but the gross striation characteristics we noted are similar. [16] If our pitch model is to be realistic, then it must reflect striation.

The most easily observable characteristic that this subject shares with all the others is ambiguity of the perceived pitch; i.e., perceived pitch for a given complex is not unique per individual as evident from Figure 13. The manifestation of pitch ambiguity is well known;[24] the perceived pitch of church bells, for example. [19] The ambiguity implies that we must look for a more localized minimum in Equation (1.8) that we repeat here for convenience.

$$\tau_{\min} = \arg\min_{\tau} \int |X(f)|^2 \, |H(f,\tau)|^2 \, df \tag{1.8}$$

We do not know the form of the transmission loss $\alpha(f)$ for this particular subject of Goldstein. That will make it difficult to predict the pitches in Figure 13. What we will do instead is to use the pitch model we developed for Terhardt's subject **(**Equation (1.12) and Equation (1.13), absent-harmonics case**)** to see if that model exhibits the striation characteristics. We reveal the striations by examining the function to be minimized in Equation (1.8),[25]

$$\int \triangleq -\int |X(f)|^2 \, |H_\alpha(f,\tau)|^2 \, df \tag{1.15}$$

Since we presume the pitch-shift stimulus $X(f)$ to be composed of spectral lines, this equation simplifies to

$$\int = -\left(|H_\alpha(f_L,\tau)|^2 + |H_\alpha(f_L+\Delta,\tau)|^2 + |H_\alpha(f_L+2\Delta,\tau)|^2 + |H_\alpha(f_L+3\Delta,\tau)|^2 \right) \tag{1.16}$$

Figure 14 shows Equation (1.16) plotted in terms of $1/\tau$ and f_L. The striations of Figure 13 might be imagined to ride along the peaks in Figure 14. But since the data in Figure 13 and Figure 14 are for two different subjects, we cannot expect a match.

We generated a contour map of Figure 14 to show the peaks more plainly in Figure 15. The lightest shades in the contour map represent striations; hence, they represent all the possible pitches that Terhardt's subject might ascribe to the pitch-shift stimulus. Striation characteristics revealed by Goldstein's data are also evident in Terhardt's subject. So, at the very least, we have revealed some unity in the characteristics of human pitch perception from the data of two eminent scholars.

[24] Ambiguity is also inherent to any periodicity-based model; that is one more reason we claim a bond between periodicity-based models and pitch perception.
[25] We flip the function to make the troughs appear as peaks.

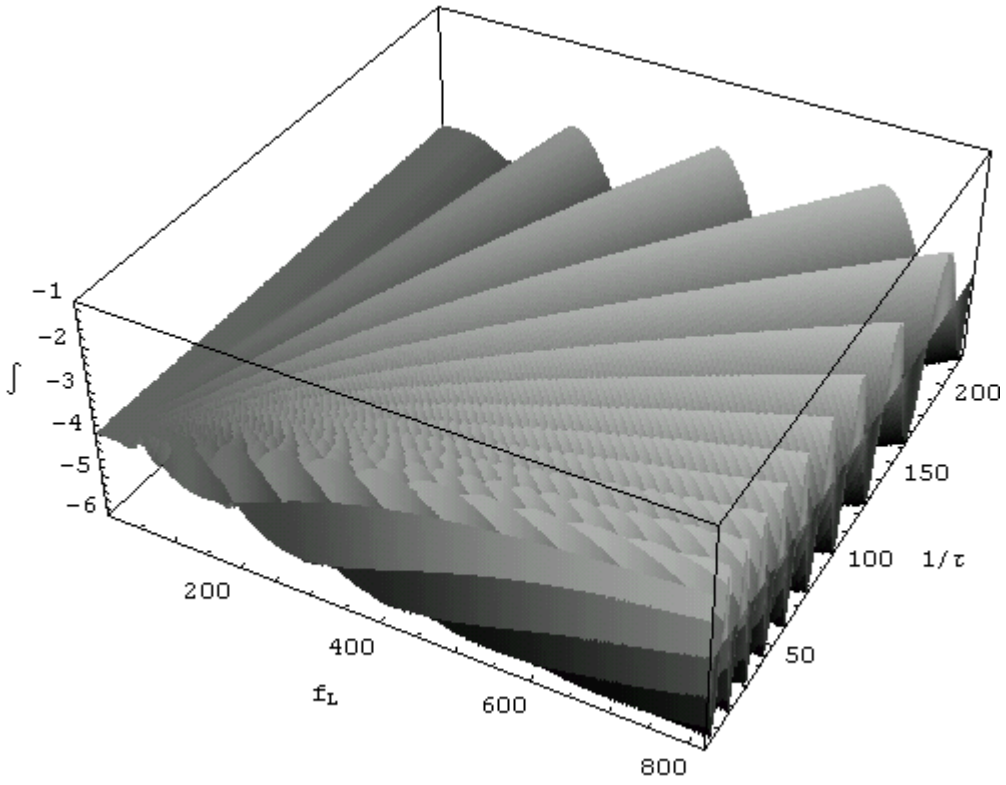


Figure 14. Plausible explanation for the origin of Goldstein's striation data; Equation (1.15).

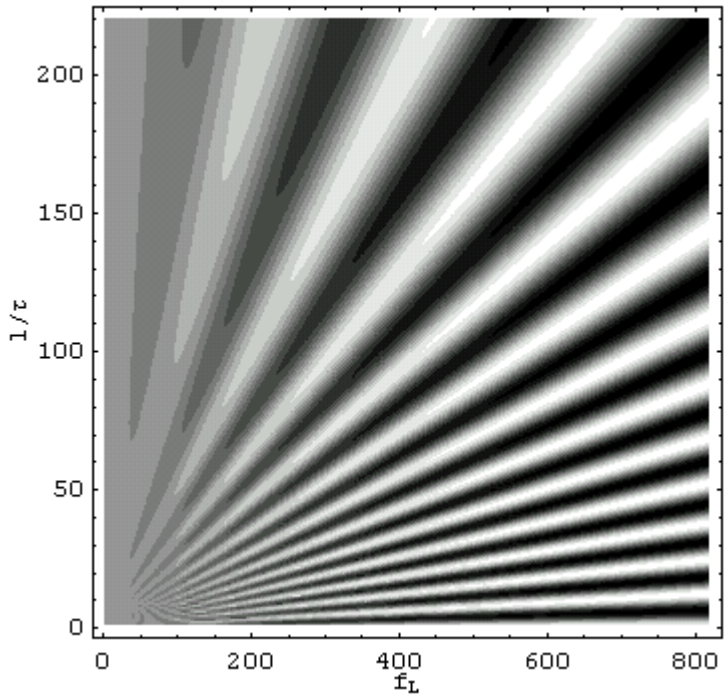


Figure 15. Contour map of previous figure. Lighter shades represent greater values.

Origin of the Striations

Aside from fitting our model to a particular subject, it is perhaps more important to realize that its exhibition of striation characteristics, in response to the pitch-shift stimulus, is somewhat independent of transmission loss $\alpha(f)$. More to the point, if we ignore the transmission loss by setting $\alpha(f) = 1$ (no loss), then the striation characteristic of the contour map, generated by exercising the model, remains intact. That observation is in keeping with our implicit assumption that all human responses exhibit the gross striation phenomenon regardless of any particular individual's transmission loss.

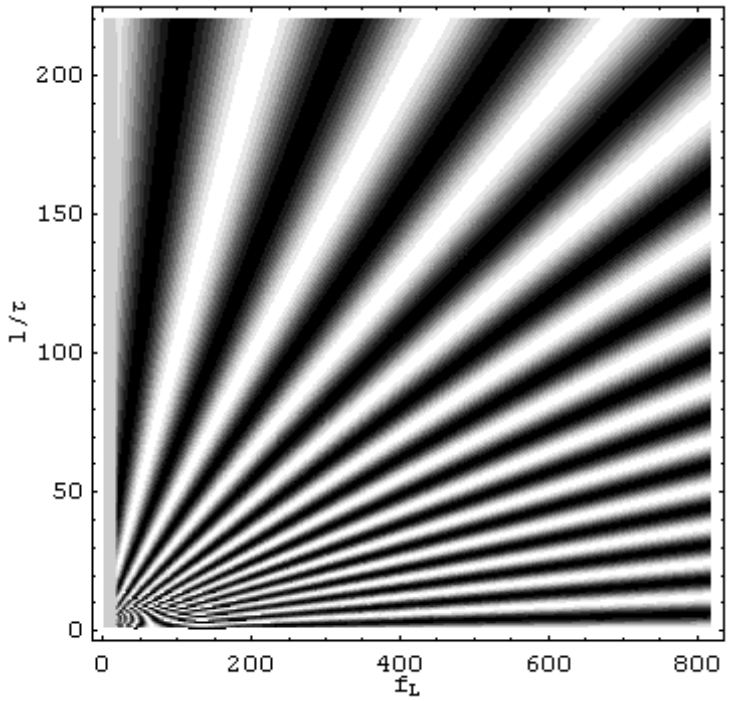


Figure 16. Same model as before except transmission loss $\alpha(f)$ set to 1.

Having the transmission loss absent from our model, as in Figure 6, the predominant circuit component becomes the cascade of many identical adaptable neuron pairs (allpass filter stages). Recall from Equation (1.11) that in the limit of an infinite number of identical adaptable neuron pairs, the cascade becomes equivalent to a variable delay element as in Figure 4. The existence of the striation characteristic seems, then, to originate from the existence of signal delay in the neurological model.

Conclusions

We adopted and justified the view that human pitch perception is based upon some sort of periodicity detection scheme. When compared to an autocorrelator, however, the human periodicity estimates are different. We said that from an engineering perspective, the human periodicity detector is flawed. The predominant flaw exhibited is Terhardt's stretched template phenomenon. We showed how to design a neural network that intrinsically incorporates stretched templates while emulating an autocorrelator.

We began our discussion of pitch perception by considering harmonic stimuli. Our model of human pitch processor response instead to inharmonic stimuli or noise, requires no modification; regardless of the particular stimulus characteristics, the pitch processor always adapts itself so as to align a template with the input signal spectrum as best it can. When the input perfectly matches a template, then there is complete annihilation at the processor signal output. Most of the time, annihilation does not occur; then, the adaptation produces a local minimization. We presumed that the spectrum of pitched sounds, presented to the model input, consists of the spectral lines of Fourier series. We ignored the impact of the cochlear response on any and all stimuli. This aspect certainly requires reconsideration.

We ignored the possibility of using metrics other than the 2-norm to adapt our pitch models. This area is certainly fertile for further research as it is unlikely that the human system squares signals to perform comparisons.

Another area that we left untouched is the replacement of the differential buffers in the neural networks in Figure 6 and Figure 8 by (nonlinear) activation functions. [14] This possibility warrants further consideration, as this replacement may be more realistic in terms of what is physically plausible. In any case, inter-neuron loading should be kept minimal, otherwise transmission loss will become excessive.

The exact form of the dynamic transmission loss $\alpha(f)$ in the delay path requires determination. The model of transmission loss must be distributed. For this project, we modeled the transmission loss as simply as possible, but the lumped second order section we used is not necessarily the correct model. A more plausible model would be derived from element mismatch in the delay path.

This project culminated with some unification of human response data from two distinct researchers, Ernst Terhardt and Julius Goldstein. Using a neural network model for pitch perception that we developed by emulating Terhardt's pitch trend data, [1] we were able to explain the origin of the striation phenomenon discovered by Goldstein. [16]

REFERENCES

[1] Ernst Terhardt, “Psychoacoustic Evaluation of Musical Sounds”, Perception and Psychophysics, vol.23(6), pp.483-492, 1978.

[2] Julius L. Goldstein, “An optimum processor theory for the central formation of the pitch of complex tones”, Journal of the Acoustical Society of America 54, 1496-1516, 1973.

[3] Julius L. Goldstein, “Verification of the optimal probabilistic basis of aural processing in pitch of complex tones”, Journal of the Acoustical Society of America 63, 486-497, 1978.

[4] R. J. Ritsma, “Frequencies dominant in the perception of the pitch of complex sounds”, Journal of the Acoustical Society of America 42, 191-198, 1967.

[5] Ernst Terhardt, “Pitch shifts of harmonics, an explanation of the octave enlargement phenomenon”, Proceedings of the 7th International Congress on Acoustics, Budapest 3, 621-624, 1971

[6] Ernst Terhardt’s home page;
http://www.mmk.e-technik.tu-muenchen.de/persons/ter.html

[7] Shihab Shamma, David Klein, “The Case of the Missing Pitch Templates: How Harmonic Templates Emerge in the Early Auditory System”,
http://www.isr.umd.edu/TechReports/ISR/1999/TR_99-27/TR_99-27.phtml

[8] Ernst Terhardt discusses the definition of pitch,
http://www.mmk.e-technik.tu-muenchen.de/persons/ter/top/defpitch.html

[9] Bernard Widrow, Samuel D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985

[10] Ronald N. Bracewell, *The Fourier Transform and Its Applications*, second edition, revised, McGraw-Hill, 1986

[11] Malcolm Slaney, Richard F. Lyon, “A Perceptual Pitch Detector”, Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, vol.1, pp.357-360, April 1990

[12] Andrea Megela Simmons, autocorrelation processing in bullfrogs,
http://neuroscience.brown.edu/Faculty/MSimmons.html

[13] A. Cichocki, R. Unbehauen, *Neural Networks for Optimization and Signal Processing*, Wiley, 1993

[14] Simon Haykin, *Neural Networks, a Comprehensive Foundation*, second edition, Prentice-Hall, 1999

[15] W. M. Hartmann, “Flanging and Phasers”, Journal of the Audio Engineering Society, vol.26, no.6, pp.439-443, June 1978

[16] A. Gerson, J. L. Goldstein, “Evidence for a general template in central optimal processing for pitch of complex tones”, Journal of the Acoustical Society of America 63, no.2, pp.498-510, February 1978

[17] Malcolm Slaney, MATLAB Auditory Toolbox,
http://web.interval.com/papers/1998-010/

[18] Martin McKinney's Homepage, convolution of spectral lines with cochlear response, the octave enlargement effect,
http://mount.ne.mediaone.net/

[19] Arthur H. Benade, *Fundamentals of Musical Acoustics*, second revised edition, Dover, 1990