1995-01

A Spectral Network Model of Pitch Perception

A SPECTRAL NETWORK MODEL OF PITCH PERCEPTION

Michael A. Cohen, Stephen Grossberg, and Lonce L. Wyse

August 1992
Revised: February 1994
Revised: September 1994
Revised: January 1995

Technical Report CAS/CNS-92-024

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
111 Cummington Street
Boston, MA 02215

A spectral network model of pitch perception

by

Michael A. Cohen, Stephen Grossberg, and Lonce L. Wyse

Center for Adaptive Systems
and
Department of Cognitive and Neural Systems

Boston University
111 Cummington St.
Boston MA 02215

Technical Report # CAS/CNS-TR-92-024
Send reprint requests to:
Professor Stephen Grossberg

Received: August 1992
Revised: February 1994
Revised: September 1994

Revised: January 1995

Running Title: Cohen, Grossberg, and Wyse: Spatial Pitch Network

Abstract

A model of pitch perception, called the Spatial Pitch Network or SPINET model, is developed and analyzed. The model neurally instantiates ideas from the spectral pitch modeling literature and joins them to basic neural network signal processing designs to simulate a broader range of perceptual pitch data than previous spectral models. The components of the model are interpreted as peripheral mechanical and neural processing stages, which are capable of being incorporated into a larger network architecture for separating multiple sound sources in the environment.

The core of the new model transforms a spectral representation of an acoustic source into a spatial distribution of pitch strengths. The SPINET model uses a weighted "harmonic sieve" whereby the strength of activation of a given pitch depends upon a weighted sum of narrow regions around the harmonics of the nominal pitch value, and higher harmonics contribute less to a pitch than lower ones. Suitably chosen harmonic weighting functions enable computer simulations of pitch perception data involving mistuned components, shifted harmonics, and various types of continuous spectra including rippled noise. It is shown how the weighting functions produce the dominance region, how they lead to octave shifts of pitch in response to ambiguous stimuli, and how they lead to a pitch region in response to the octave-spaced Shepard tone complexes and Deutsch tritones without the use of attentional mechanisms to limit pitch choices. An on-center off-surround network in the model helps to produce noise suppression, partial masking and edge pitch. Finally, it is shown how peripheral filtering and short term energy measurements produce a model pitch estimate that is sensitive to certain component phase relationships.

PACS numbers: 43.66.Hg 43.66.Ba

I Introduction and Overview

A fundamental problem of auditory perception is the identification and separation of multiple acoustic sources. Such a process enables human listeners to perceive and recognize the contents of discriminable auditory streams, in a process called auditory scene analysis by Bregman (1990). The process utilizes a variety of cues including synchrony, harmonicity, and binaural timing and intensity information to assign acoustic components to the appropriate auditory stream. This article describes a model for generating a spatial representation for the pitch of an acoustic source that can be naturally embedded in an architecture for source separation.

The Spatial Pitch Net, or SPINET, is a type of spectral "pattern matching" model, briefly reported in Cohen, Grossberg and Wyse (1992a,b). The input to the pitch detecting module is a spectral representation, and the output is a function across pitch. Other models that transform a spectral representation of the signal to a pitch representation include the pitch models of Goldstein (1973), Wightman (1973), and Terhardt (1972). The SPINET model properties simulate many significant pitch perception data for reasons similar to those of the spectral models mentioned above, whose formal kinship has been demonstrated by de Boer (1976) (see Appendix A for a summary of data addressed by different models). Despite the formal similarities, each of the spectral models suggest a different mechanism for implementing what turns out to be similar functions of pitch. Wightman (1973) computes the peak in a cosine Fourier transform of a smeared spectrum. The process analogous to smearing the spectrum is accomplished in the Goldstein (1973) model by perturbing the signal frequency components with noise. A harmonic template matching process then produces the most likely pitch. In the Terhardt (1972) model, input components have "virtual pitches" at subharmonics. When different components have virtual pitches that coincide, the strength of the virtual pitch is increased. This process is similar to increasing the pitch strength when multiple harmonic components fall through holes in the sieve of a

harmonic template.

A key component in each model is a set of filters with bandwidths that scale with the filter center frequency and which spread or randomize the ultimate effect of a component across frequency. One difference between Goldstein's Optimal Processor model and the other models is that the Goldstein model is not deterministic. The frequency scaling function is the variance of a normally distributed noise process given the input frequency. The model then produces a maximum likelihood estimation of the pitch using an idealized harmonic template. Wightman's bandwidth-scaling filters model the peripheral auditory filters, and are intended to approximate the resolving powers of the basilar membrane place coding. A cosine Fourier transform measures the periodicity in the spectral representation to produce a deterministic pitch function. The spreading function in Terhardt's model is the "coincidence interval" parameter which determines the contributions to a pitch made by nearby subharmonics of different input components (Terhardt, Stoll and Seewann, 1982a). To sum up the functional relationships between the maximum likelihood estimator and the deterministic pitch strength models, the smearing of the effect of spectral components (whether by a noise process or by the spread of activation) determine a pitch function (whether a probability density or an activation level) with various modes that (explicitly or implicitly) are the result of different harmonic number assignments to the peaks in the spectral representation. These components will discussed in more detail in the context of the Spatial Pitch Network.

II SPINET Structure

The stages of the SPINET model are summarized in Figure 1. The input to the model is computer generated sound sampled at a rate of 16 kHz. All sounds were 25 ms in duration including a 5 ms raised cosine onset and offset ramp.

[Figure 1 about here.]

A. Model Equations

The pressure variation at the oval window of the cochlea initiates a traveling wave along the basilar membrane (von Békésy, 1928) and produces a maximal displacement at a position along the basilar membrane as a function of frequency. High frequencies produce their maximum displacement near the basal end of the cochlea, low frequencies near the apex. Each point along the membrane can thus be considered as a bandpass mechanical frequency filter.

The processing stage modeling the mechanical filtering of the basilar membrane (Figure 1, stage 2) consists of a bank of bandpass filters, each with a frequency response approximating a fourth-order Gammatone filter (Holdsworth *et al.*, 1988; Patterson *et al.*, 1988):

$$GT(f) = [1 + j(f - f_i)/b(f_i)]^4, \tag{1}$$

and implemented as a cascade of four first-order digital filters where f_i is the center frequency of the ith filter, and $b(f_i)$ controls the bandwidth of the filter as a function of center frequency as described in Eq. 3. The complete set consists of 512 filters with center frequencies spaced evenly in Equivalent Rectangular Bandwidth (ERB) units (Moore and Glasberg, 1983) from 50 Hz to 5 kHz to cover the extent of the "existence region" for residue pitch (Ritsma, 1962).

Following Moore and Glasberg (1983), the equivalent rectangular bandwidth (ERB) of the filter centered at a frequency f, is a function of the filter center frequency:

$$ERB(f_i) = 6.2310^{-6} f_i^2 + 93.3910^{-3} f_i + 28.52. \tag{2}$$

Holdsworth *et al.* (1988) showed that if the power passed through the fourth-order Gammatone filter is set equal that passed by a rectangular filter with gain one, then the bandwidth parameter $b(f_i)$ is related to the ERB by

$$b(f_i) = ERB(f_i)/.982. \tag{3}$$

Equation 2 implies that such filters above 1 kHz have bandwidths that are approximately a constant percentage of their center frequency, and become relatively wider as the center frequency becomes lower.

The output of the filter bank is measured to derive a spectral representation of the signal using the equation

$$Y(f_i,n) = \frac{BB(f_i)(1-e^{-\beta})^{1/2}}{N} \sum_{i=1}^{N} \sqrt{\left(\sum_{j=1}^{n} x^2(f_i, n-\Delta(i+j))e^{-\beta j}\right)} \tag{4}$$

where $x(f_i,n)$ is the signal passing through the Gammatone filter with center frequency f_i at time n, $\beta = 8.637 \times 10^{-3}$, Δ is the sampling period fixed throughout at $1/16000sec$, and $N = 80 = 5ms/\Delta$ is the averaging window length. Input sound levels were chosen so that $\max_i Y(f_i,n)$ is the same for all sounds. By Eq 4., $Y(f_i,n)$ is a measure of the square root of the power passed through the filter centered at that frequency multiplied by an exponential time window which decays to half its maximum over approximately 5 ms. This measure is averaged over a 5 ms window in each filter to yield the spectral input (Figure 1, Stage 3) to the next processing layer. The function $BB(f_i)$ is a lumped model of processes contributing to a broad bandpass effect on the contribution of frequency regions to pitch which is assumed to include the outer and middle ear transfer function (Dadson and King, 1952) as well as the phase locking capabilities of 8th nerve neurons. Unlumping these properties would add to the complexity of the model without having a substantial effect on the simulated results. This stage (Figure 1, Stage 4) is thus modeled in the frequency domain by the gamma function:

$$BB(f_i) = sf_i \exp(-sf_i) \tag{5}$$

where $s = .001$ producing a peak gain at 1 kHz and a region between 500 Hz and 2 kHz that is flat within a 3 dB range.

The next stage (Figure 1, Stage 5) models cooperative interactions across nearby frequencies and competitive interactions across a broader frequency band of the averaged power spectrum $Y(f_i,n)$. Interactions fall off with distance as the Gammatone function of Eq. 1.

The inhibitory region is larger than the width of the excitatory region, and both scale with the ERB of the channel. The power spectrum of the Gammatone function with center frequency f_i and a bandwidth proportionality factor of κ is

$$|H(f_i, f, \kappa)|^2 = [1 + ((f - f_i)/(\kappa b(f_i))^2]^{-4} \tag{6}$$

where $b(f_i)$ is as defined in Eq. 3. The result of the cooperative-competitive interactions is

$$S(f_i, n) = \sum_{j=1}^{512} Y(f_i, n)[\frac{|H(f_i, f_j, \kappa_{ex})|^2}{A_{ex}(f_i)} - \frac{|H(f_i, f_j, \kappa_{in})|^2}{A_{in}(f_i)}] \tag{7}$$

where f_i is the center frequency of the channel, and $\kappa_{ex} = .4$ and $\kappa_{in} = .6$ define the excitatory and inhibitory regions as a constant proportion of the ERB of the frequency channel. The area of of the excitatory region ($A_{ex}(f_i)$) and that of the inhibitory region ($A_{in}(f_i)$) is defined to be the sum of the Eq. 6 function values taken over the center frequencies of the filter bank. Although each inhibitory region is wider than that of the excitatory region centered at the same frequency, the two regions are normalized in Eq. 7 to be equal. Thus, if the power spectrum measured from the peripheral filter bank is flat, then the output from this layer is zero across the frequency spectrum. Equation 7 models the equilibrium response of neurons organized in an on-center off-surround anatomy. It is assumed that the neurons track the inputs fast enough to remain in approximate equilibrium with them.

The next two stages carry out a weighted (Figure 1, Stage 6) harmonic summation (Figure 1, Stage 7). The pitch strength P is a sum of non-negative spectral strengths S, weighted by the distance between the nominal pitch p and the frequency of the harmonic mp, as in

$$P(p, n) = \sum_m [S(mp, n)]^+ h(m), \tag{8}$$

where

$$[x]^+ = \begin{cases} x & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

and

$$h(m) = \begin{cases} 1 - M\ log_2(m) & \text{for } M\ log_2(m) < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Parameter M in Eq. 9 determines the slope of the falloff with harmonic number m that a harmonic makes to the strength P of pitch p, and has the value $M = .15$ in our simulations.

The output of the network is taken to be the pitch that has the strongest activation level; that is, the "best fitting fundamental" is taken to be that pitch p which maximizes the output function $P(p, n)$. When modeling experiments where the pitch responses were restricted to a given region, the pitch is taken to be the maximal pitch in that region. A winner-take-all operation, which can be implemented by an on-center off-surround feedback network (Grossberg 1973, 1988) or another contrast enhancing competitive neural network, can be used to select the maximally activated pitch. The combination of filter (8) followed by a contrast-enhancing operation is a specialized case of a competitive learning, or self-organizing feature map, neural network (Carpenter and Grossberg, 1991; Grossberg, 1976, 1982; Kohonen, 1989), again solved at equilibrium with respect to the current inputs. Some further assumptions will also be suggested below as a way to interpret the information across the entire pitch function.

B. Implementation

The computations were performed by three separate programs; one for the Gammatone filter bank, another for the energy and the last for the pitch computation including the on-center off-surround convolution. All programs were written in C and run on a time-shared Sun Sparc-10 workstation. For 25ms sounds sampled at 16 kHz, 512 frequency channels and 200 pitch channels, the computation times were Filterbank: 5.5 seconds, Energy: 1.2 seconds, Pitch: .7 seconds, each including input/output file read/write time. The model uses only local feedforward network interactions that will run in real time when implemented as a chip.

III Model Components and Other Models

A. Peripheral Filters

A comparison of the peripheral frequency spreading functions followed by the template match in both the SPINET and the Optimal Processor models clarifies the differences between a mechanistic and an information theoretic theory. The interpretation of the peripheral frequency spreading effect represents uncertainty about the precise frequency of a stimulus component in the Optimum Processor theory, and a deterministic spatial weighting function in the network model. In the SPINET model, the peripheral spreading functions as part of a pitch matching template. In the statistical model, the Gaussians do not function as part of the template, but instead represent the uncertainty of the frequency of an input component. The template is matched to the perturbed signal wherein the contribution of a tone to a pitch falls off with the Euclidean distance between the noise perturbed tone and nearest template component location. The Optimum Processor produces a pitch with the maximum likelihood given the uncertainty of the input, or more generally, produces a probability density function across pitch. The SPINET model produces a continuous, spatially organized, "strength of activation" function of pitch.

The Optimal Processor and the SPINET model predict a similar multimodal distribution of possible pitch matches. In both models, the different modes correspond to different estimates of the harmonic numbers assigned to the components. In order to compare the deterministic models to statistical models or statistical performance data, the activation vs. pitch function must be related to a probability density function of a random decision variable. Houtsma (1979) used one such technique to compare the models of Terhardt (1974) and of Wightman (1973) with Goldstein's model. He used the relative heights of the main modes in a region of the deterministic pitch activation functions as their relative likelihoods, with non-peak regions considered to have zero likelihood. The pitch function was then normalized

so that the sum of the heights was one, to yield something analogous to a discrete probability density function.

Many models of pitch use a broad bandpass function near the periphery that has been variously interpreted as representing the spectral-dominance phenomenon (Terhardt *et al.*, 1982a), or as representing the transfer function of the outer and middle ear together (Meddis and Hewitt, 1991), or the inverse of the minimum audible pressure (MAP) threshold function (eg., Yost and Hill, 1979). The shape of this weighting function bears an inverse relationship to the tone frequency JND function, which is used as partial justification for the shape of the function relating the variance of the noise distribution to frequency in Goldstein's pitch model. Despite the different mechanisms suggested, the shape of the function plays a similar role and is used to address the same data, particularly dominance region data, by the different models.

B. On-center Off-surround Interactions

Yost and Hill (1979) introduced the use of lateral inhibition into the pitch modeling literature in their Peripheral Weighting Model. They were primarily concerned with modeling the pitch of anharmonic rippled noise which is produced by subtracting Gaussian white noise from itself delayed by an interval τ. The spectrum has peaks spaced by $1/\tau$, which in a log frequency representation become closer together at higher frequencies.

They showed that a "dominance region" emphasizing the spectrum in the neighborhood of $4/\tau$ produced the best pitch predictions for the this kind of noise. They used the center-surround mechanism with bandwidths proportional to their center frequencies as a means of inhibiting frequencies above $4/\tau$ without the model having to know *a priori* the value of τ. In terms of the center frequency f of the filters, the lateral interactions used to fit the pitch data were about $1/6f$ for the excitatory region, and an inhibitory region extending another $1/6f$ beyond the excitatory region. The sizes of the center-surround interactions agree with

those found in physiological studies of the cat cochlear nucleus (Bilsen *et al.*, 1975) and psychoacoustically in humans (Houtgast, 1977; Shannon, 1976).

The on-center off-surround lateral interactions (Eq. 7) play several important roles in the SPINET model, one of which is to attenuate the DC level of activation in the spectral layer. Consider pink noise with constant power per octave. Since the excitatory and inhibitory regions are of equal area, the spectral representation used to derive pitch has a constant zero level of activation. A pitch model that sums spectral regions near harmonics that does not control for noise in some analogous fashion would be biased toward lower fundamentals, since their harmonics become more closely spaced in the regions of high noise density. The effect of incorporating surround inhibition is to flatten the pitch response to noise.

The center-surround mechanism also serves to increases the effective resolution of the spectral representation, making pattern matching pitch determinations possible where they would not otherwise be. The excitation pattern (the output from Stage 4, Figure 1) for 6 harmonics of 100 Hz between 1000 and 1500 Hz is shown in Figure 2(a). The output of the center-surround processing is shown in Figure 2(b), where all but the component at 1400 Hz are represented by a distinct peak in the representation.

Also visible in Figure 2(b) is the increased weight afforded the extreme frequency components of this stimulus relative to the middle components. This a "partial masking" effect which is explicitly incorporated by another mechanism in the model of Terhardt, Stoll and Seewann (1982a). The "dominance" of outer components in frequency discrimination for both individual components and the pitch of the complex when low harmonics are missing has also been suggested by Moore, Glasberg and Shailer (1984). This edge enhancement of the spectral contour is also responsible in the model for the "edge pitch" associated with the filter cutoff frequencies of narrow bandpass noise (Bilsen, 1977; Fastl, 1971).

[Figure 2 about here.]

C. Harmonic Summation

The "harmonic sieve" (Duifhuis *et al.* 1982; Scheffers, 1983) is a kind of template matching where the "holes" in the sieve have a rectangular shape around each harmonic of a pitch. That is, an input component either contributes to the pitch or it does not, depending upon whether or not it is close enough to a harmonic of the pitch to fall through the sieve. Moore *et al.* (1985) measured the influence of a harmonic on pitch by mistuning the components one at a time, and observing the effect on the shift in the pitch of the complex. As a single component in a harmonic complex is mistuned, the perceived pitch of the complex begins to shift at first in the same direction as the component. As the component is mistuned beyond 3% of its original frequency, its effect on the pitch begins to diminish and the pitch shifts back toward its original f_0 (Figure 3). When the component is mistuned by roughly 8% of its original frequency, its effect on the pitch is negligible. Moore *et al.* (1985) suggested that if a harmonic sieve is operating, one possible explanation of these data is that a component does not fall through the sieve in an all-or-none fashion. In the SPINET model, the frequency spreading due to the energy measure of the Gammatone filters (Eq. 1) followed by the punctate template (Eq. 8) is equivalent to using spectral peaks and a sieve with gradual skirts around the harmonics and is responsible for the gradual effect on the pitch as a harmonic is shifted.

[Figure 3 about here.]

A problem arises in models that give equal weight to all harmonics of a fundamental because they predict equal pitch strengths (or likelihoods) for all subharmonics of that fundamental. Additional mechanisms are needed to explain how even the pitch of a single tone is unambiguously perceived. This problem occurs in the Optimum Processor theory where the mean squared error used to evaluate the fit between a harmonic template and a stimulus gives the same result for a nominal pitch value and all its submultiples because the components of a template with a given spacing are a subset of the components of all templates

with submultiple spacings. The SPINET model without the harmonic weighting function (so that $h(n) = 1$ in Eq. 8) exhibits such a response. This can be seen in Figure 4a, which shows the pitch activation due to a single input tone at 1 kHz with peaks equally prominent at all subharmonics. In Goldstein *et al.* (1978), two mechanisms are considered which would help prevent the subharmonic match. One is to restrict the number of harmonics that could contribute to pitch so that lower pitches would not be predicted by high components. The other is to restrict the range of pitches included in the template matching process.

[Figure 4 about here.]

Gerson and Goldstein (1978) elaborated this second method by introducing an *a priori* expectation into the Optimum Processor theory. This prior expectation, due to experimental conditions and subject biases, is presumed to correspond to a rectangular distribution determining the upper and lower bounds of pitch perception. Given this rectangular *a priori* expectation, the model computes a maximum likelihood estimate evaluated over the region within the bounds of the expectation.

In the SPINET model, as in Terhardt, Stoll and Seewann (1982a), it is assumed that the greater the ratio of a component frequency to a nominal pitch value, the less the contribution the component makes to that pitch. The SPINET model uses a decreasing function, linear in log frequency. To the extent that frequency is represented neurally as a tonotopic map, this model property represents a decreasing effect of cells on each other with distance across the map.

In response to a harmonic complex or single tone stimulus, the weighted network model produces a unique maximum in the pitch activation function at the pitch corresponding to the periodicity of the stimulus. Figure 4b shows the output of the model using the decreasing weighting functions, in response to a single tone at 1000 Hz. No additional attentional mechanism is required for the model to respond unambiguously with the pitch at the frequency of a single tone or harmonic complex. For single tones, this mechanism

is consistent with the the fact that the range of possible pitch percepts is *a priori* much wider than a single octave. This is not to deny that frequency-specific attentional effects are sometimes operative, for example in detecting signals in noise (Dai, Scharf and Buus, 1991), or in hearing pitches that correspond to the minor modes of the activation function, but in the absence of such active attentional focusing, the default "expectation" is assumed to be essentially unbiased.

D. The Dominance Region

The dominance region is that part of the spectrum where components have the strongest influence on pitch. It is a function of both the frequency of the input components and fundamental frequency (Plomp 1967; Ritsma, 1967). In terms of the Optimum Processor theory, the dominance region is the spectral region where two complexes differing slightly in fundamental frequency are most discriminable (Goldstein, 1973).

There are two different kinds of errors predicted by the shape of the σ/f function (Figure 5) in the model contributing to its account of the dominance region. One kind of error is due to pitches in the secondary modes of the probability density function which are the result of assigning the wrong harmonic numbers to the noise perturbed signal. This type of error becomes more likely as harmonics become more closely spaced in log frequency (as harmonic numbers increase). The other kind or error is caused by pitches in the main mode of the probability density function, but where the variance is high due to the low precision of the component frequency estimates at low and high frequencies.

[Figure 5 about here.]

For fundamentals below 300 Hz, resolution improves as harmonic numbers increase until their frequencies reach the peak in the resolution curve of Figure 5, thereby partially offsetting the degradation in performance due to the closer component spacing for the low fundamentals. For high fundamentals, as harmonic numbers increase, the wide component

spacings imply that their frequencies quickly surpass the peak in the resolution curve, thereby causing a faster deteriation in performance.

The SPINET model shows a similar pattern relating harmonic number to pitch strength. For low fundamentals, the function first increases with harmonic number because the bandpass function (Equation 5) increases with component frequency faster than the distance-dependent harmonic weighting function attenuates the contribution to pitch. For high fundamentals, even low-order harmonic contributions to pitch are attenuated by both the bandpass function and the harmonic weighting function (Equation 9). The effect of the unimodal σ/f function in the Optimal Processor theory is thus analogous to the effect of the bandpass function in the SPINET model.

Since the shape of the pitch function resembles the shape of the probability density function produced by the Optimum Processor, it is interesting to consider interpreting the deterministic model statistically for comparison. Houtsma (1979) did this with the models of Terhardt (1972, 1973) and Wightman (1973) by taking the percent correct in performance as the ratio of the height of the main pitch mode to the sum of the heights of all the modes within a roughly half-octave "attentional" band around the main mode.

When the SPINET model is analyzed in this way, it does not produce a fundamental-frequency dependent variation in percentage correct as is seen in the data and predicted by the Optimum Processor. This is because the entire pitch strength function, for a given pair of stimulus components with fixed harmonic numbers, scales across fundamental frequency while leaving the shape (that is, the *relative* heights of the modes) invariant. One method we are exploring to preserve the f_0 dependence of the strength function, discussed above in the context of the dominance region, is to add a constant level of noise across frequency to the pitch function before taking the maximally activated pitch as the model output.

IV SPINET Simulations

A. Pitch Shifts with Component Shifts

When harmonic components ($f_n = nf_0, n = 1, ...$) are all shifted by a constant amount, Δ, in frequency so that they maintain their spacing of f_0, ($f_n = nf_0 + \Delta, n = 1, ...$), the pitch shift in linear frequency is slower than that of the components (Patterson and Wightman, 1976; Schouten *et al.* 1962). Typical data show an ambiguous pitch region at shift values of $\Delta = lf_0$, $l = .5, 1.5, 2.5, ...$ where the most commonly perceived pitch jumps down to below the value of f_0. Figure 6a shows the pitch of components spaced by $f_0 = 100Hz$ as a function of the lowest component's harmonic number, l. When the shift value Δ is near a harmonic of f_0 ($\Delta = lf_0$, $l = 0, 1, 2, ...$), then the pitch is unambiguous and near 100 Hz.

[Figure 6 about here.]

The model's correspondence with these data (Figure 6b) is due to the gradual reduction in the contribution a component makes to a pitch as it is mistuned, combined with the effect of filters whose widths are approximately constant in log coordinates for high frequencies. As the components shift together in linear frequency away from harmonicity, the higher components move into the shallow skirts of the filters centered at harmonics of the original nominal pitch frequency much more slowly than do the lower components, thereby slowing the shift away from the original pitch. For the same reason, as the lowest stimulus component increases in harmonic number, all components are moving through broader filters, so the slopes of the pitch shift become less steep, as can be seen in both the data and the model output in Figure 6.

B. Pitch Shift Slopes with Component Shifts

One of the main findings of Patterson and Wightman (1976) was the difference in the slope of the pitch shift between low and high fundamentals as the components shift while maintaining

their spacing in linear frequency (Figure 7a). The slopes converge as lower components are removed. Figure 7b shows a plot of the slopes found in the model measured at the point where components are harmonic. The difference in slopes for the two fundamental frequencies is due to the region of dominance induced by the combined effect of three weighting functions: the BB (Eq. 5) broad bandpass function, the harmonic falloff (Eq. 9) giving more weight to low-order harmonics, and the inhibitory interactions (Eq. 7) which, being of roughly constant width in log frequency, inhibit the higher frequency components more than the low.

[Figure 7 about here.]

These weighting functions are insufficient to explain the entire data set. When the slope of the least-mean-squares best fitting straight line through all the pitches is measured, including those in the ambiguous regions, then the model produces too little difference in slopes between the different f_0's (Figure 7c). There are several possible explanations for the disparity between the model measurements when the ambiguous region is included, and the data of Patterson and Wightman (1976).

1) **Combination tones.** The "second effect of pitch" is that when the shifting stimulus consists of lower frequency components, the shift of the pitch is steeper than when it consists of higher frequency components. The addition of the $f_i - n(f_{i+1} - f_i)$ combination tones (Goldstein, 1967) arising from the peripheral interaction of two successive components, would be exactly at the frequencies in the equal spacing pattern of the Patterson and Wightman (1976) paradigm, albeit at lower levels. Their effect is thus easy to predict and, as noted by many authors (*e.g.* Smoorenburg, 1970), would indeed be to make the slopes greater. By the SPINET mechanism discussed for the first effect of the pitch shift, the addition of lower components would increase the slope for the shift in the model as well.

2) **Secondary Modes.** The slopes measured by Patterson and Wightman (1976) were the slopes of mean pitch matches made by the subjects. In the ambiguous region, there are more modes in the pitch function whose strength rivals that of the main mode. Above the

fundamental, the secondary modes are higher than the main mode; below the fundamental, the secondary modes are lower. If subjects matched to these secondary modes in the ambiguous region, the mean pitch would be further from the main mode and thus the slopes would be steeper in the ambiguous region.

3) **Grouping effects.** In his doctoral dissertation, de Boer (1956) suggested that the "second effect of the pitch shift", the systematic decrease in the slope of the shift as lower components are eliminated from the signal, could be due to a preferential weighting given to the lower components. Without considering component groupings, it does not seem logical that the components that are shifting the fastest out of their harmonic relationship with the rest of the components (measured as a percentage shift from their harmonic frequency) should be the ones to be accorded the most weight. Furthermore, the current model explains the rate of the shift using the fact that the higher components move through the frequency-scaled Gammatone filters more slowly than do the lower components, thereby maintaining their contribution to pitches near the spacing frequency at higher shift values than do the lower components.

However, complexes in the ambiguous region often sound like multiple sources. If there exists a separate grouping process with the capacity to separate the influence on pitch of different frequency regions of the peripheral (in this case spectral) representation of the signal, then the pitch being primarily influenced by the lower tones would move faster than it does when it is forced to take into account all frequency regions of the peripheral representation. The addition of such a grouping mechanism to a larger architecture containing the SPINET model would thus produce better estimates of the shift slopes in the ambiguous region, while leaving the good performance of the model near harmonic regions intact.

C. The Tritone Paradox

Deutsch (1992a, 1992b), has investigated a phenomenon called the tritone paradox. Stimuli are composed of sinusoids spaced by octaves with a raised cosine amplitude envelope across the entire range of hearing. In musical notation, notes spaced by an octave have the same pitch class (the same name, *e.g.* C#), which is suggestive of their perceptual similarity. Thus, perceptually fused complexes of octave spaced components have a clear pitch class, but an ambiguous octave designation. Shepard (1964) found that, when presented with two successive stimuli of different pitch classes, the interval that subjects identified was that corresponding to the shortest distance between the two pitch classes. Thus, the interval C-G was heard as descending 5 semitones rather than as ascending 7 semitones. Indeed, Shepard found that when a sequence of these octave-component complexes is presented which repeatedly traversed the semitone scale, pitch appears to ascend endlessly in a kind of barbershop pole illusion, despite the octave equivalence of notes spaced by 12 semitones.

When the interval between two such complexes is exactly half an octave (a "tritone" in musical terminology), proximity obviously cannot be used to judge the direction of the interval. In fact, Deutsch found strong intra-subject consistency of the judgments depending upon the pitch class of the tones. For tritones based on half the pitch classes, the intervals were heard as ascending, while intervals based on the other half were heard as descending.

These data are consistent with the explanation that pitch judgments are all taken to be within a single octave, which is the behavior exhibited by the SPINET model, as well as the Virtual Pitch model (Terhardt, Stoll and Seewann, 1982b), in response to such stimuli. Figure 8 shows the SPINET model's circularity of the judgments with pitch class. The effect is due to a combination of the broad bandpass function (Eq. 5) and the falling harmonic weighting function (Eq. 9). If, for example, only the broadband filter were operative and all harmonics were weighted equally, then the lowest possible submultiple of the components would always be the chosen pitch. The combination of the two mechanisms results

in all pitches occurring within an octave that is centered below the peak in the amplitude envelope of the stimulus, and well above the lowest possible pitch (Figure 8). A tritone interval that spans the discontinuity in the pitch function of Figure 8 produces nominal pitch values that descend, while the same interval comprised of pitch classes that do not span the discontinuity produces nominal pitch values that ascend. The variability that Deutsch found between subjects can be explained in model terms by the manipulation of the BB and harmonic weighting functions (Eqs. 5, 9). Small changes in the parameters governing these functions shift the octave region of maximal pitch responses without substantially affecting the response to other pitch stimuli.

[Figure 8 about here.]

D. Rippled Noise Spectra

Noise with a rippled spectrum is also capable of producing a pitch sensation. One such spectrum is produced by summing Gaussian white noise with itself delayed by an interval τ. The average spectral power density is

$$\phi(f, \tau, g) = 1 + g\cos(2\pi f\tau), \tag{10}$$

where g is the gain parameter applied to the delayed signal (Bilsen and Ritsma, 1970) . The result is often referred to as $Cos+$ noise, and has peaks separated by $1/\tau$. For $Cos+$ noise, the peaks are at harmonics of the frequency corresponding to the reciprocal of the delay τ, and a pitch is induced at this frequency. The SPINET response is shown in Figure 9a.

If a delayed white noise signal is subtracted from itself, the result is $Cos-$ noise which has an average power spectrum density of

$$\phi(f, \tau, g) = 1 - g\cos(2\pi f\tau). \tag{11}$$

The $Cos-$ spectrum is thus seen to be a shifted version of the $Cos+$ spectrum with a shift value equal to $1/(2\tau)$. These rippled noise stimuli produce a pitch sensation similar to the

residue pitches induced with tones at the locations of the noise peaks. Specifically, the *Cos*–spectrum produces an ambiguous pitch that is generally matched to $.9/\tau$ and $1.1/\tau$ (see Yost, Hill and Perez-Falcon (1978) for a review). This should not be surprising, as the peaks are in the same locations as the tones in the "ambiguous region" discussed in Section A. Figure 9 shows the SPINET model response to rippled noise which shows peaks near $.9/\tau$ and $1.1/\tau$, the location of the most frequently matched pitches.

[Figure 9 about here.]

E. Pitch of Narrow Bands of Noise

von Békésy (1963) reported that pitches could be observed corresponding to the upper and lower edges of an octave band of noise between 400 and 800 Hz, and made the analogy to Mach bands at luminance edges in vision (Mach, 1865). Small and Daniloff (1967) used noise for matching with cutoff frequencies in a region an octave above or below the test stimulus. They found that low and highpass filtered noise could invoke a pitch sensation corresponding to the noise edges when the cutoff frequencies were as high as 10 kHz for both low and highpass noise, and as low as 80 Hz for high-pass noise and 600 Hz for lowpass noise. When the bandwidth of the noise is less than approximately 1/5 octave, the pitch is heard to be near the center of the band of noise (Fastl, 1971), and only at larger bandwidths do pitches begin to show at the edges of the noise. Figure 10 shows the response of the SPINET model to bands of noise created by summing randomly spaced sinusoids (spaced by an average of 2 Hz) with random phase in bands centered at 500 Hz, with bandwidths of 1/10, 1/5 and 2/5 of an octave. The pitch functions are averaged over ten trials. The model chooses the location of the maximum as the pitch on each trial, and individual trials tend to have one dominant peak even when the average function shows a peak at both noise band edges.

[Figure 10 about here.]

F. The Dominance Region

It has long been known that certain harmonics have more influence on pitch perception than others. Ritsma (1967) and Plomp (1967), using a similar experimental procedure, showed that the region of the 3rd, 4th and 5th harmonics is dominant in determining the pitch of a harmonic complex. Plomp presented subjects with two stimuli A and B in succession, where

$$A = \sum_{n=1}^{12} \cos(2\pi n f t) \tag{12}$$

and

$$B = \sum_{n=1}^{m} \cos[2\pi n(0.9f)t] + \sum_{m+1}^{12} \cos[2\pi n(1.1f)t]. \tag{13}$$

Plomp asked subjects whether the pitch of B was higher or lower than of A. Responses were plotted as a function of m, the cut-off number for harmonics above which harmonics of B were mistuned up, and below which they were mistuned down. For fundamental frequencies above 1400 Hz, subjects reported that the pitch of B was lower than A for all m; that is, even when only one component was lower, the pitch was perceived as moving down. For lower fundamentals, m could be as high as 5, and the pitch of B was still identified as being higher. Since for lower fundamentals, the direction that the 3rd, 4th and 5th harmonics were tuned determined which way the pitch was heard as moving, these harmonics became known as constituting the "dominance region".

The SPINET model predictions for the dominance region can be seen in Figure 11 for fundamental frequencies of 100 and 1400 Hz. The plot shows the pitch strength function in response to the Plomp (1967) stimulus B for 5 different values of m between 1 and 5. The two peaks are centered around the fundamental frequency of stimulus A. For the 100 Hz fundamental, the peak on the lower side of A does not approach the value of the peak on the high side until $m > 4$, while for 1400 Hz, the peak on the lower side is maximal for $m > 1$. The contribution that a component makes to a pitch falls off more quickly with harmonic number for high fundamentals even though the harmonic weighting function has the same

slope (M in Eq. 9) for all pitches because of the steep falloff in the BB function (Eq. 5) at high frequencies. This causes the dominance region to move significantly toward the lowest harmonic as as the fundamental increases.

[Figure 11 about here.]

G. Distant Modes and Octave Drops for Ambiguous Stimuli

Much of the pitch shift data has been gathered by focusing the attention of experimental subjects on a narrow pitch region centered at f_0, and has thus neglected the true extent of the ambiguity of the pitch sensation in the ambiguous region (Patterson and Wightman, 1976; Schouten, Ritsma and Cardozo, 1962). As Schouten *et al.* (1962) showed, the distribution of pitch matches is multi-modal with the various modes being clearly separated by a region where no matches occur. Several of the modes are near f_0, but some modes are further away. In the data on pitch as a function of the shift in equally spaced components, the *ambiguous region* is characterized by the components being near the frequencies $f_n = f_0(1/2+n)$, which can be written as $Nf_0/2$ for odd integer N. This ambiguous region, where pitch identification jumps discontinuously from one side of f_0 to the other when matching is constrained to a narrow band about f_0, is the region where all the components are near the odd harmonics of $f_0/2$. Gerson and Goldstein (1978) showed that, in fact, when the lowest frequency component in the stimulus was an odd multiple of $f_0/2$, the lower pitch, $f_0/2$, could be heard when pitch matches were not restricted to be in a narrow band around f_0. Some of their data for a four-component stimulus are summarized in Figure 12a. The model's maximum pitch as a function of the lowest harmonic number, without the restrictions of an attentional window, predicts this octave drop, as shown in Figure 12b.

[Figure 12 about here.]

As can be seen in the Gerson and Goldstein data, the relationship between the lowest harmonic number and subjects' pitch matches is one-to-many (Smoorenburg, 1970) . Since

the lower octave pitch match implies the assignment of non-successive harmonic numbers to the stimulus components, these data motivated the least-mean-squares template matching extension to the Optimal Processor theory so that it no longer presumed that the stimulus was comprised of successive harmonics of some fundamental frequency (Gerson and Goldstein, 1978; Goldstein *et al.*, 1978). In addition, as the lower components are removed, the octave drop becomes less likely. Under these stimulus conditions, model behavior is best understood by examining the entire pitch function rather than just the maximal pitch.

Raatgever and Bilsen (1991) provide further data for comparison. They presented "anharmonic" noise stimuli that were produced by passing white noise through a delay line with delay T and feeding a fraction g of the delayed version back to the input with a sign inversion. This is different from the rippled noise stimuli discussed earlier where only a delay, but no feedback, is used. These comb-filtered noise signals have peaks and valleys at the same spectral locations as rippled noise, but the peaks are sharper (Raatgever and Bakkum, 1986), having power spectra of the form

$$P(f) = \frac{1}{1 - 2g\cos(2\pi fT) + g^2}. \tag{14}$$

The anharmonic noise was passed through a highpass filter with a variable cutoff frequency. As the lower peaks in the anharmonic spectrum are removed, the perception of the lower octave percept disappears, giving way to matches on either side of the pitch with the nominal frequency of the spectral peak spacing (Figure 13). The SPINET model behavior can be seen by looking at the whole pitch function, where the lower octave peak moves from having the highest level of activation to a relatively lower level as the filter cutoff frequency increases (Figure 14a-d).

[Figure 13 about here.]

[Figure 14 about here.]

H. The Phase of a Mistuned Component

The SPINET model is sensitive to aspects of the fine temporal structure of the input signal because the spectral representation on which the pitch decision is based derives from finite time measurements of the signal. Thus the model can be tested on multi-tone complex stimuli with varying phase relationships. As in the autocorrelation model of Meddis and Hewitt (1991), SPINET pitch measurements are sensitive to relative phases of components by virtue of within-channel cancellation or reinforcing interactions. If components are completely resolved, no phase effects appear in the pitch output.

Hartmann (1988) performed a discrimination experiment using a harmonic signal composed of the the first seven harmonics of 800 Hz in one interval. In the other interval, the same stimulus was used except that the fourth harmonic was mistuned by 2.5%. Hartmann manipulated the duration of the signal and found as an overall trend that the subjects did better the longer the stimuli. The improvement was not monotonic however, but had dips and troughs as a function of duration. Discounting the long term improvement trend, the dips and troughs were cyclic with the period of the stimulus.

Meddis and Hewitt (1991b) showed that their autocorrelation model produces the same pattern of dips and troughs, but since they used only one time constant for the running autocorrelation functions, the gradual improvement was not superimposed. Indeed, duration *per se* is not the the critical variable; rather, it is the phase of the signal over the time window in which the pitch function is measured.

Meddis and Hewitt (1991b) plotted this effect by computing the Euclidean distance between the model summary autocorrelation function for the non-mistuned component stimulus and the stimulus with the mistuned component at different "durations". They assumed that the percent correct (which Hartmann measured) would have the same trend as this distance metric. The SPINET spectral model produces the same phase sensitivity when interpreted in this fashion (Figure 15) but, like the autocorrelation model, shows no long term trend

due to the absence of any integration mechanism that spans time intervals on the scale of 100 *ms.*

[Figure 15 about here.]

I. A spectral explanation of a classical phase experiment

To test human sensitivity to phase, Ritsma and Engel (1964) used a quasi-frequency modulated (QFM) signal with the center frequency component shifted in relation to the flanking tones by 90 degrees:

$$x(t) = 0.5m \sin[2\pi(n-1)ft] + \sin[2\pi nft + \pi/2] + 0.5m \sin[2\pi(n+1)ft]. \qquad (15)$$

When n, the harmonic number of the middle component, was equal to 11 and 13, Ritsma and Engel (1964) found that subjects matched pitches to both the fundamental frequency f and to $2f$. When $n = 12$ however, they found pitch matches above and below f and $2f$, but rarely in between. The results are consistent with a fine temporal structure "peak-picker" algorithm which they advocated. Wightman (1973b) was unable to duplicate the results of the experiment however, finding pitches at f (the region about $2f$ was not tested) for each $n = 10, 11, 12, 13$, thereby refuting the idea of phase sensitivity to such stimuli. Wightman did not test for pitches near $2f$.

Meddis and Hewitt (1991b) showed that their model predictions agreed with Wightman's (1973b) findings that for each $n = 10, 11, 12, 13$, pitches are found at f but not nearby. In the region of $2f$, however, they found pitches slightly above or slightly below $2f$, but not at $2f$ when n is even, and at exactly $2f$ for n odd, which agrees with the Ritsma and Engel's (1964) observations.

A possible explanation of these data is in terms of the fine temporal structure of the signal (Ritsma and Engel,1964; Moore, 1977). In both phase cases, the envelope has a major peak at the fundamental period $1/f_0$, and a secondary peak at $1/2f_0$. For the zero-phase condition, the secondary envelope peak is much weaker than the major peak, while

for the phase-shifted condition, the two envelope peaks are almost equal in magnitude. Now consider how the fine temporal structure is superimposed on the envelope structure. When the harmonic number n of the center component is odd, the fine structure has peaks that correspond with both envelope peaks. When n is even, however, the fine structure has peaks that flank the $1/2f_0$ envelope peak and line up exactly only with the $1/f_0$ envelope peak. If pitch is determined by measuring the period between fine structure peaks that occur near envelope peaks, the system could make pitch matches near $2f_0$ when n is even.

The pitch output of the SPINET model also agrees with Ritsma and Engel's (1964) split-peak findings around $2f_0$ for n even, but the difference between the shape of the pitch functions near $2f_0$ for n even versus n odd can be explained without reference to temporal fine structure, and is, in fact, independent of the phase shift of the middle stimulus component. Figure 16 shows the SPINET pitch functions for $n = 12$, and Figure 17 shows the same for $n = 11$.

[Figure 16 about here.]

[Figure 17 about here.]

The explanation for the behavior parallels the explanation for the ambiguous region in the paradigm of equally spaced shifted components discussed in Section G. When n is even, two of the three components are odd multiples of f_0, and therefore are shifted to frequency values that are exactly half way between harmonics of $2f_0$. The presence of these two components make the $2f_0$ pitch match unlikely. In the SPINET model, these two components contribute to a dip in the pitch function, working against the middle component that contributes to the strength of the $2f_0$ pitch. When n is odd, two of the three components are even harmonics of f_0 and are therefore (successive) harmonics of $2f_0$. A pitch peak a $2f_0$, regardless of the component phase relationships, is thus not surprising.

Further contributing to the absence of any pitch match at $2f_0$ when n is even, is that the two anharmonic peaks are on the "edge" of the signal spectrum, while the only harmonic of

$2f_0$ is interpolated between them and is subject to the "partial masking" effect discussed at the end of Sec. III B. In terms of the model, this edge effect occurs because the competitive interactions (Eq. 7) between frequency locations in the spectral representation, enhance the edges of the excitation pattern coming from the bank of peripheral filters.

V Conclusion

The Spatial Pitch Net model generates a spatial representation of pitch from a spectral representation of the auditory stimulus. A key feature of the model is a set of weighting functions for harmonics that decrease with harmonic number. The weighting functions obviate the need for an *a priori* attentional window to prevent all subharmonics of a given pitch from assuming an equal pitch strength. The forms of the weighting functions are capable of explaining the "dominance region" data for harmonic contributions to pitch. The model can handle continuous spectra such as rippled noise as well as the more standard spectra of discrete tones.

The SPINET model is constructed using components similar to those found in several different spectral and neural network models. The synthesis has enabled the model to be successfully tested on a breadth of data not attempted by any single spectral model previously. Using one model to explore such a range of data brings a coherency of explanation to, for example, the utility of center-surround mechanisms for modeling rippled noise data, psychophysically and physiologically measured inhibitory interactions, and phenomena such as partial masking and edge pitch. Due to the frequency component interactions in the peripheral filters used to derive the spectral representation, some temporal effects such as component phase relationships can be simulated which are not typically explored with formal spectral models.

The SPINET model produces as output a strength value across a spatial representation of pitch, rather than the frequency of the most likely pitch. It is based on the idea of a

"central spectrum" representation of the auditory signal, rather than the fine structure of a temporal waveform (Licklider, 1951; Meddis and Hewitt, 1991). A spatial representation of activation across pitch in response to each stimulus is important in part because it can provide an explanation for data on responses to ambiguous stimuli. More importantly, such a spatial representation can be naturally integrated into the dynamics of a larger architecture for auditory and speech perception (cf., Boardman, Cohen, and Grossberg, 1993; Boardman, Grossberg, and Cohen, 1994; Cohen and Grossberg, 1986; Cohen, Grossberg and Stork, 1988; Govindarajan, Grossberg, Wyse, and Cohen, 1994; Grossberg, Boardman, and Cohen, 1994; Grossberg and Stone, 1986). For example, if attentional factors are used to prime a particular frequency region, then the spatial pitch representation plays an important role in understanding how attentional focusing can alter the ensuing pitch percept. This kind of model can also use pitches as cues to group together the components of the same sound source and to separate different sources from one another in the auditory scene. Govindarajan, Grossberg, Wyse, and Cohen (1994) have embedded the SPINET model into a larger neural architecture for auditory scene analysis and source separation in which both pitch and spatial location cues can be used to separate harmonically overlapping sound sources, as in a cocktail party situation.

Acknowledgements

Michael A. Cohen and Stephen Grossberg were supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225).

Lonce Wyse was supported in part by the American Society for Engineering Education, and by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0225).

We wish to express our gratitude to Adrian Houtsma, Roy Patterson and an anonymous reviewer for their careful reading and comments on earlier drafts of this paper.

Appendix A: Summary of Data addressed by Various Models

Tables I and II summarize the pitch data that have been addressed by various models, either in the original modeling work, or in modified versions or discussions by the original others or others in the literature. A "$\surd$" means that the model produces a reasonable fit to the data, "NT" (not tried) means there has been no published discussion, and "X" means that the model has been shown not to work for the particular data. This table is intended only for a quick comparison, and it should be understood that many of the models have several different incarnations that might change an entry in the table.

[Table 1 about here.]

[Table 2 about here.]

References

Bilsen, F. (1977). "Pitch of noise signals: Evidence for a "central spectrum"," J. Acoust. Soc. Am. **61**, 150–159.

Bilsen, F. and Goldstein, J. (1974). "Pitch of dichotically delayed noise and its possible spectral basis," J. Acoust. Soc. Am. **55**, 292–297.

Bilsen, F. and Ritsma, R. (1970). "Some parameters influencing the perceptibility of pitch," J. Acoust. Soc. Am. **47**, 469–475.

Bilsen, F., ten Kate, J., Buunen, T., and Raatgever, J. (1975). "Response of single units in the cochlear nucleus of the cat to cosine noise," J. Acoust. Soc. Am. **58**, 858–866.

Boardman, I., Cohen, M.A., and Grossberg, S. (1993). "Variable rate working memories for phonetic categorization and invariant speech perception," in *Proceedings of the world congress on neural networks, Portland, Oregon* (Erlbaum Associates, Hillsdale, NJ), Vol. 3, pp. 2–5.

Boardman, I., Grossberg, S., and Cohen, M. (1994). "Neural dynamics of phonetic trading relations for variable-rate CV syllables." *Technical Report CAS/CNS-TR-94-037* (Boston University, Boston, MA).

Bregman, A. (1990). *Auditory Scene Analysis* (M.I.T. Press, Cambridge).

Carpenter, G.A. and Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks* (M.I.T. Press, Cambridge).

Cohen, M. and Grossberg, S. (1986). "Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory," Human Neurobiology **5**, 1–22.

Cohen, M., Grossberg, S., and Stork, D. (1988). "Speech perception and production by a self-organizing neural network," in *Evolution, learning, cognition and advanced architectures*, edited by Y. Lee (World Scientific Publishers, Hong Kong).

Cohen, M., Grossberg, S., and Wyse, L. (1992a). "Harmonic weighting functions in a neural network model of pitch detection and representation," in *Proceedings of the International Joint Conference on Neural Networks, Beijing, P.R.China* (Institute of Electrical and Electronic Engineers, Piscataway NJ), Vol. 2, pp. 149–154.

Cohen, M., Grossberg, S., and Wyse, L. (1992b). "A neural network for synthesizing the pitch of an acoustic source," in *Proceedings of the International Joint Conference on Neural Networks. Baltimore, Maryland* (Institute of Electrical and Electronic Engineers, Piscataway, NJ), Vol. 4, pp. 649–654.

Dadson, R. and King, J. (1952). "A determination of the normal threshold of hearing and its relation to the standardization of audiometers," J. Laryngol. Otol. **66**, 366–378.

Dai, H., Scharf, B., and Buus, S. (1991). "Effective attenuation of signals in noise under focused attention," J. Acoust. Soc. Am. **89**, 2837–2842.

de Boer, E. (1976). "Pitch theories unified," in *Psychophysics and Physiology of Hearing*, edited by E. Evans and J. Wilson (Academic Press, London).

Deutsch, D. (1992a). "Paradoxes of musical pitch," Scientific American **264**, 88–95.

Deutsch, D. (1992b). "Some new pitch paradoxes and their implications," Philosophical transactions of the royal society of London **336**, 391–397.

Duifhuis, H., L.F.Willems, and Sluyter, R. (1982). "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," J. Acoust. Soc. Am. **71**, 1568–1580.

Fastl, H. (1971). "Ueber Tonhöhenempfindungen bei Rauschen," Acustica **25**, 350–354.

Gerson, A. and Goldstein, J. (1978). "Evidence for a general template in central optimal processing for pitch of complex tones," J. Acoust. Soc. Am. **63**, 498–510.

Goldstein, J. (1967). "Auditory nonlinearity," J. Acoust. Soc. Am. **41**, 676–689.

Goldstein, J. (1973). "An optimum processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. **54**, 1496–1515.

Goldstein, J., Gerson, A., Srulovicz, P., and Furst, M. (1978). "Verification of the optimal probabilistic basis of aural processing in pitch of complex tones," J. Acoust. Soc. Am. **63**, 486–497.

Govindarajan, K.K., Grossberg, S., Wyse, L.L., and Cohen, M.A. (1994). "A neural network model of auditory scene analysis and source segregation." *Technical Report CAS/CNS-TR-94-039* (Boston University, Boston, MA).

Grossberg, S. (1973). "Contour enhancement, short-term-memory, and constancies in reverberating neural networks," Studies in Applied Mathematics **52**, 217–257.

Grossberg, S. (1976). "Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors," Biol. Cyb. **23**, 121–134.

Grossberg, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control* (Kluwer Academic Publishers, Dordrecht).

Grossberg, S. (1988). "Nonlinear neural networks: Principles, mechanisms, and architectures," Neural Networks **1**, 17–61.

Grossberg, S., Boardman, I., Cohen, M. (1994). "Neural dynamics of variable-rate speech categorization." *Technical Report CAS/CNS-TR-94-038* (Boston University, Boston, MA).

Grossberg, S. and Stone, G. (1986). "Neural dynamics of word recognition and recall: Attentional priming, learning and resonance," Psychological Review **93**, 46–74.

Hartmann, W. (1988). "Pitch perception and the segregation and integration of auditory entities," in *Auditory Function: Neurobiological Bases of Hearing*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (John Wiley and Sons, New York), pp. 623–645.

Hill, R. and Yost, W. (1978). "Strength of the pitches assiciated with ripple noise," J. Acoust. Soc. Am. **64**, 485–492.

Holdsworth, J., Nimmo-Smith, I., Patterson, R., and Rice, P. (1988), "Implementing a gammatone filter bank," Annex C of the SVos Final Report: The auditory filter bank. APU Report 2341.

Houtgast, T. (1977). "Auditory-filter characteristics derived from direct-masking data and pulsation-threshold data with a rippled-noise masker," J. Acoust. Soc. Am. **62**, 409–415.

Houtsma, A. (1979). "Musical pitch of two-tone complexes and predictions by modern pitch theories," J. Acoust. Soc. Am. **66**, 87–99.

Houtsma, A. and Goldstein, J. (1972). "The central origin of the pitch of complex tones: Evidence from musical interval recognition," J. Acoust. Soc. Am. **51**, 520–529.

Kohonen, T. (1989). *Self-organization and associative memory* (Springer-Verlag, Berlin), 3rd edition.

Licklider, J. (1951). "A duplex theory of pitch perception," Experientia **7**, 128–133.

Mach, E. (1865). "Über die wirkung der räumlichen vertheilung des lichtreizes auf die netzhaut, I," Sitzber. Math.-Naturw. Kl. Kaiser. Akad. Wiss. **52**, 303–322.

Meddis, R. and Hewitt, M. (1991a). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I:Pitch identification," J. Acoust. Soc. Am. **89**, 2866–2882.

Meddis, R. and Hewitt, M. (1991b). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery II:Phase sensitivity," J. Acoust. Soc. Am. **89**, 2883–2893.

Moore, B. and Glasberg, B. (1983). "Suggested formulae for calculating auditory filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

Moore, B., Glasberg, B., and Peters, R. (1985). "Relative dominance of individual partials in determining the pitch of complex tones," J. Acoust. Soc. Am. **77**, 1853–1860.

Moore, B. (1977). "Effects of relative phase of the components on the pitch of three-component complex tones," in *Psychophysics and Physiology of Hearing*, edited by E. Evans and J. Wilson (Academic, New York).

Patterson, R., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (1988). "An efficient auditory filterbank based on the gammatone function," in Annex B of the SVos Final Report: The auditory filter bank. APU Report 2341.

Patterson, R. and Wightman, F. (1976). "Residue pitch as a function of component spacing," J. Acoust. Soc. Am. **59**, 1450–1459.

Plomp, R. (1967). "Pitch of complex tones," J. Acoust. Soc. Am. **41**, 1526–1533.

Raatgever, J. and Bakkum, M. (1986). "Spectral domainance for noise signals with monaural and dichotic comb spectra," in *Proc. 12th Int Congr. Acoust.* (, Toronto), Vol. B2(4).

Raatgever, J. and Bilsen, F. (1991). "The pitch of anharmonic comb filtered noise reconsidered," in *Auditory Physiology and Perception*, edited by Y. Cazals, C. Demany, and K. Homer (Pergamon Press, Oxford), Vol. 83 of *Advances in the Biosciences*, pp. 215–222.

Ritsma, R. (1962). "Existence region of the tonal residue I," J. Acoust. Soc. Am. **34**, 1224–1229.

Ritsma, R. (1967). "Frequencies dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am. **42**, 191–198.

Ritsma, R. and Engel, F. (1964). "Pitch of frequency-modulated signals," J. Acoust. Soc. Am. **36**, 1637–1644.

Scheffers, M. (1983). "Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter," J. Acoust. Soc. Am. **74**, 1716–1725.

Schouten, J., Ritsma, R., and Cardozo, B. (1962). "Pitch of the residue," J. Acoust. Soc. Am. **34**, 1418–1424.

Shannon, R. (1976). "Two-tone unmasking and suppression in a forward-masking situation," J. Acoust. Soc. Am. **59**, 1460–1471.

Shepard, R. (1964). "Circularity in judgments of relative pitch," J. Acoust. Soc. Am. **36**, 2346–2353.

Smoorenburg, G. (1970). "Pitch perception of two-frequency stimuli," J. Acoust. Soc. Am. **48**, 924–942.

Terhardt, E. (1972). "Zur Tonhöhenwahrnehmung von Klängen," Acustica **26**, 173–199.

Terhardt, E. (1974). "Pitch, consonance, and harmony," J. Acoust. Soc. Am. **55**, 1061–1069.

Terhardt, E., Stoll, G., and Seewann, M. (1982a). "Algorithm for extraction of pitch and pitch salience from complex tonal signals," J. Acoust. Soc. Am. **71**, 679–688.

Terhardt, E., Stoll, G., and Seewann, M. (1982b). "Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions," J. Acoust. Soc. Am. **71**, 671–678.

von Békèsy, G. (1928). "Zur Theorie des Hörens; die Schwingungsform der Basilar Membran," Phys Z. **29**, 793–810.

von Békésy, G. (1963). "Hearing theories and complex sound," J. Acoust. Soc. Am. **35**, 588–601.

Wightman, F. (1973a). "The pattern-transformation model of pitch," J. Acoust. Soc. Am. **54**, 407–416.

Wightman, F. (1973b). "Pitch and stimulus fine structure," J. Acoust. Soc. Am. **54**, 397–406.

Yost, W., Hill, R., and Perez-Falcon, T. (1978). "Pitch and pitch discrimination of broadband signals with rippled power spectra," J. Acoust. Soc. Am. **63**, 1166–1173.

Yost, W. and Hill, R. (1979). "Models of the pitch and pitch strength of ripple noise," J. Acoust. Soc. Am. **66**, 400–410.

List of Figures

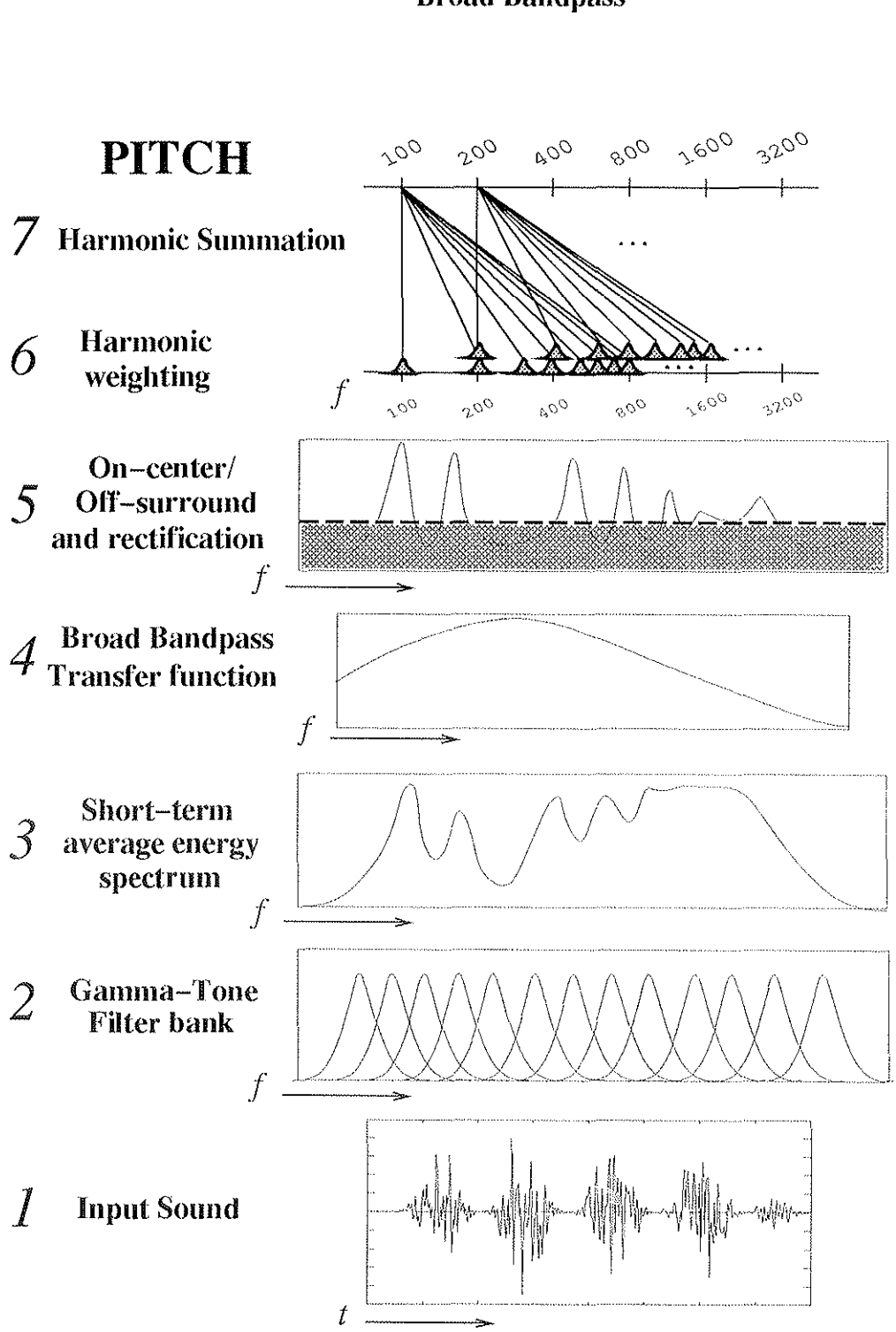
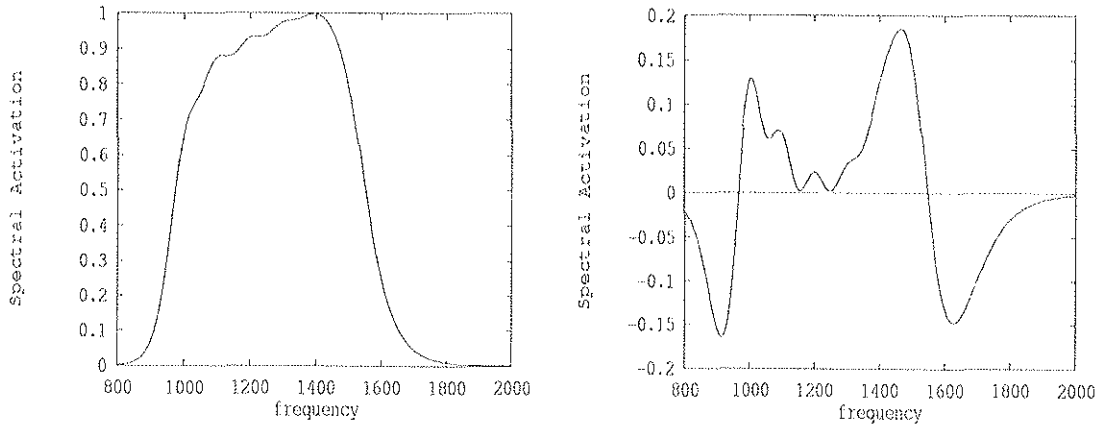

Broad Bandpass
PITCH
7 Harmonic Summation
6 Harmonic weighting
5 On–center/ Off–surround and rectification
4 Broad Bandpass Transfer function
3 Short–term average energy spectrum
2 Gamma–Tone Filter bank
1 Input Sound
100
200
400
800
1600
3200
f
t

FIG. 1.


Spectral Activation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
800
1000
1200
1400
1600
1800
2000
frequency
Spectral Activation
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
800
1000
1200
1400
1600
1800
2000
frequency

(a) *(b)*

FIG. 2.

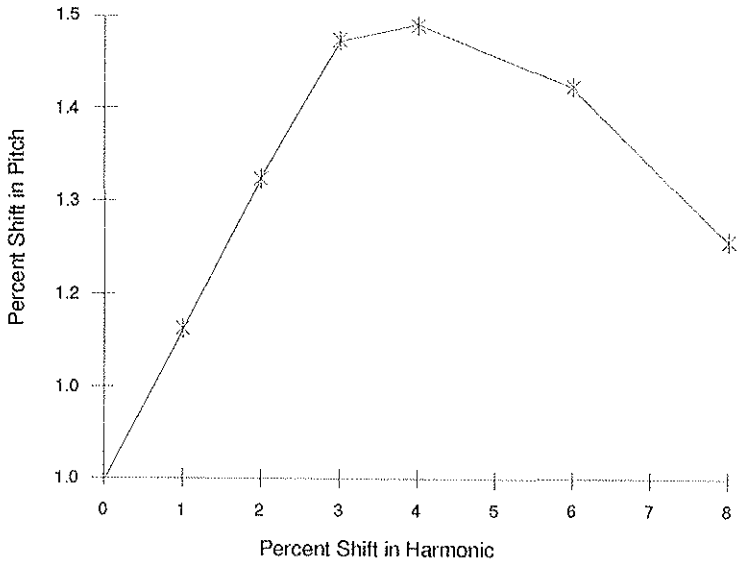
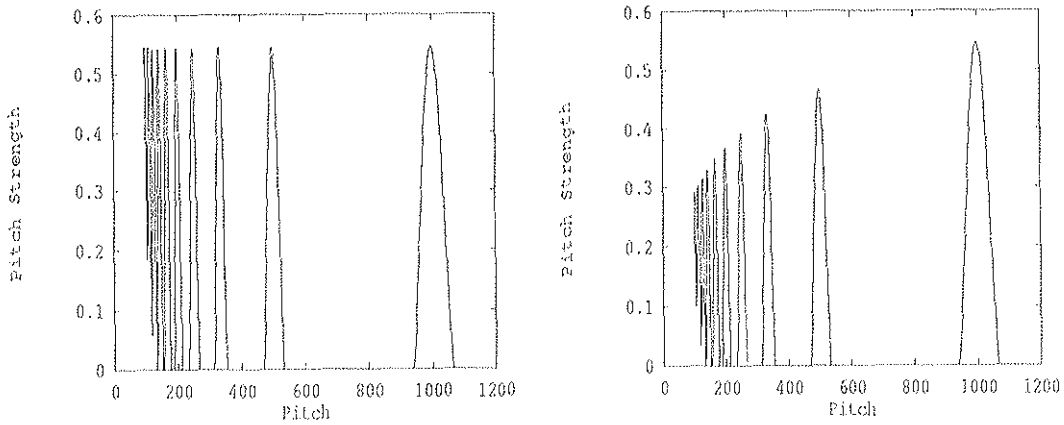

1.5
1.4
1.3
1.2
1.0
1.0
Percent Shift in Pitch
0
1
2
3
4
5
6
7
8
Percent Shift in Harmonic

FIG. 3.


Pitch Strength
0.6
0.5
0.4
0.3
0.2
0.1
0
0 200 400 600 800 1000 1200
Pitch
Pitch Strength
0.6
0.5
0.4
0.3
0.2
0.1
0
0 200 400 600 800 1000 1200
Pitch

(a) *(b)*

FIG. 4.

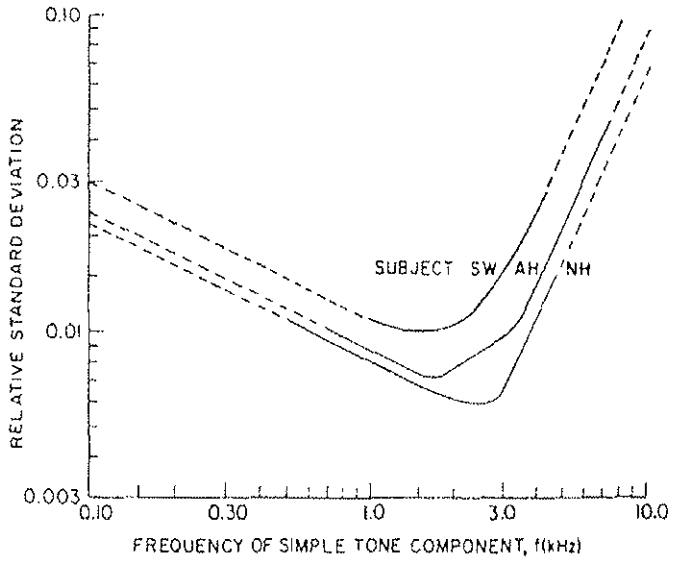
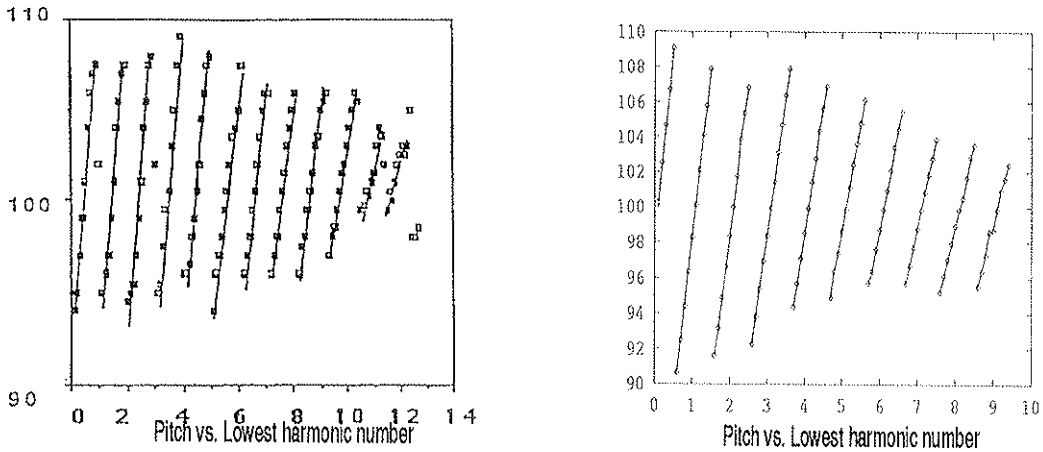
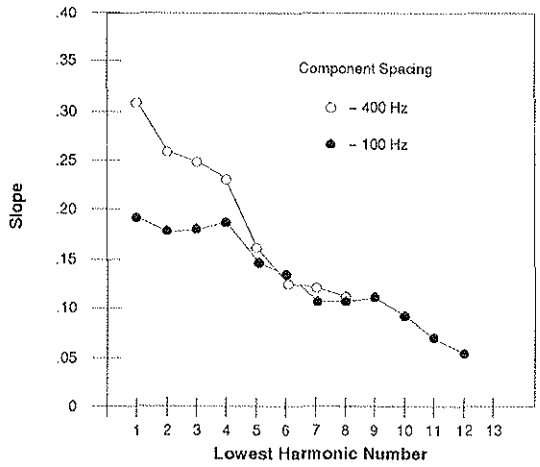

0.10
0.03
0.01
0.003
RELATIVE STANDARD DEVIATION
SUBJECT SW AH NH
0.10
0.30
1.0
3.0
10.0
FREQUENCY OF SIMPLE TONE COMPONENT, f(kHz)

FIG. 5.

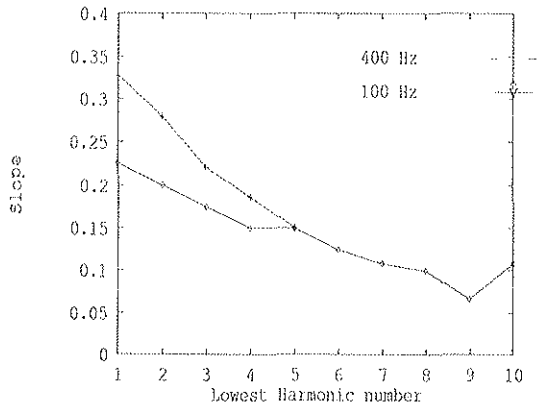

110
100
90
0 2 4 6 8 10 12 14
Pitch vs. Lowest harmonic number
110
108
106
104
102
100
98
96
94
92
90
0 1 2 3 4 5 6 7 8 9 10
Pitch vs. Lowest harmonic number

(a) *(b)*

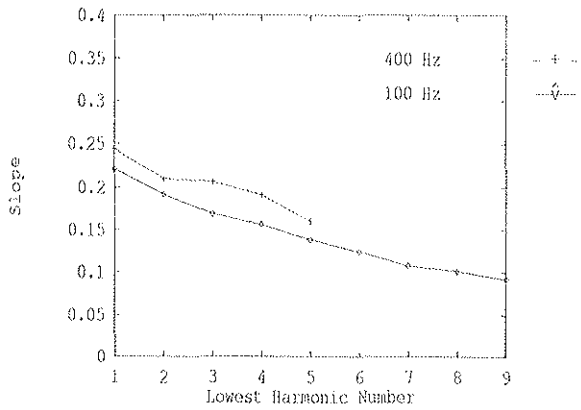
FIG. 6.


.40
.35
.30
.25
.20
.15
.10
.05
0
Slope
Component Spacing
○ - 400 Hz
● - 100 Hz
1 2 3 4 5 6 7 8 9 10 11 12 13
Lowest Harmonic Number

(a)


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
slope
400 Hz
100 Hz
1 2 3 4 5 6 7 8 9 10
Lowest Harmonic number

(b)


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
slope
400 Hz
100 Hz
1 2 3 4 5 6 7 8 9
Lowest Harmonic Number

(c)

FIG. 7.

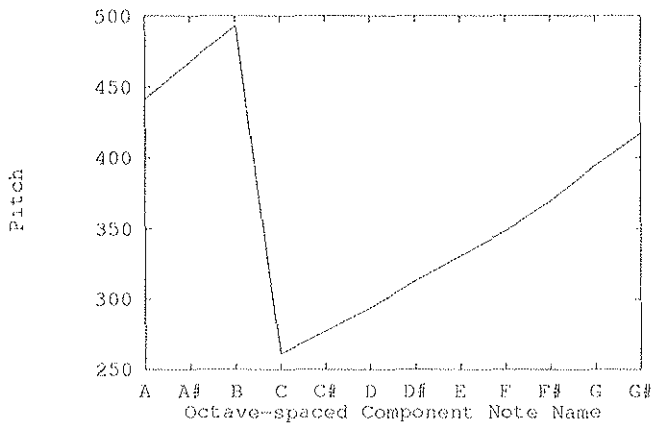
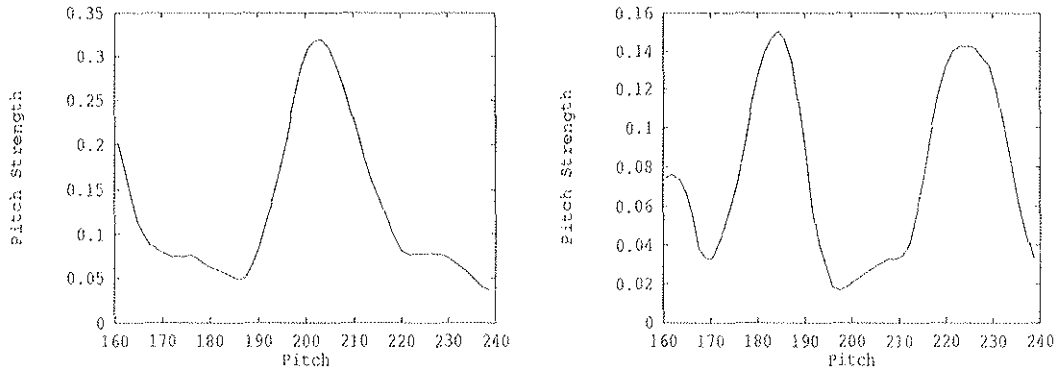
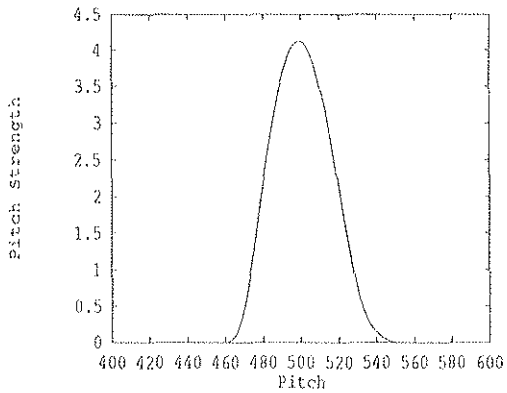

500
450
400
350
300
250
Pitch
A
A♯
B
C
C♯
D
D♯
E
F
F♯
G
G♯
Octave-spaced Component Note Name

FIG. 8.

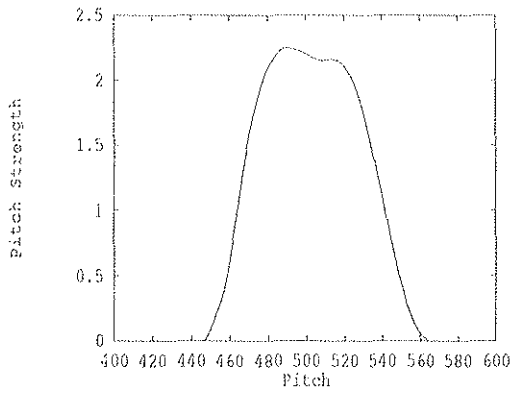

Pitch Strength
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
160 170 180 190 200 210 220 230 240
Pitch
Pitch Strength
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
160 170 180 190 200 210 220 230 240
Pitch

(a) *(b)*

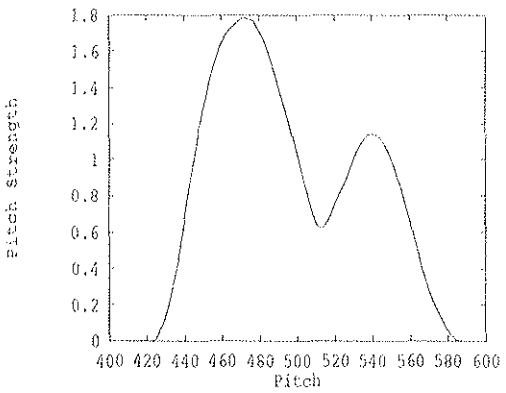
FIG. 9.


Pitch Strength
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
400 420 440 460 480 500 520 540 560 580 600
Pitch

(a)


Pitch Strength
2.5
2
1.5
1
0.5
0
400 420 440 460 480 500 520 540 560 580 600
Pitch

(b)


Pitch Strength
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
400 420 440 460 480 500 520 540 560 580 600
Pitch

(c)

FIG. 10.

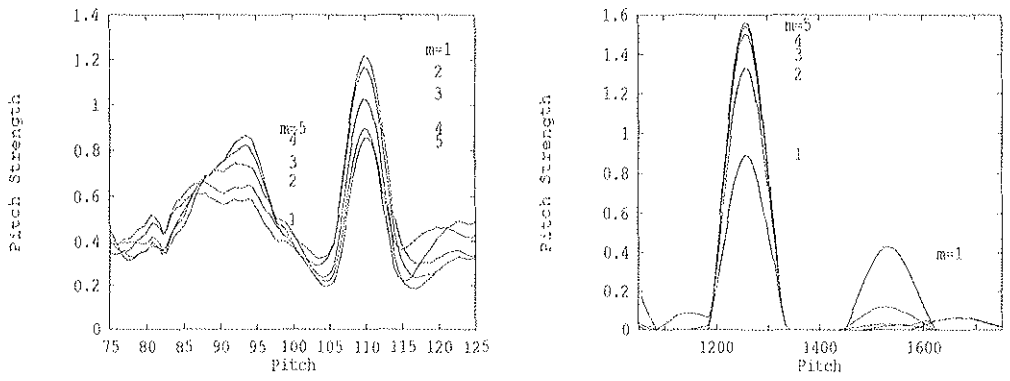
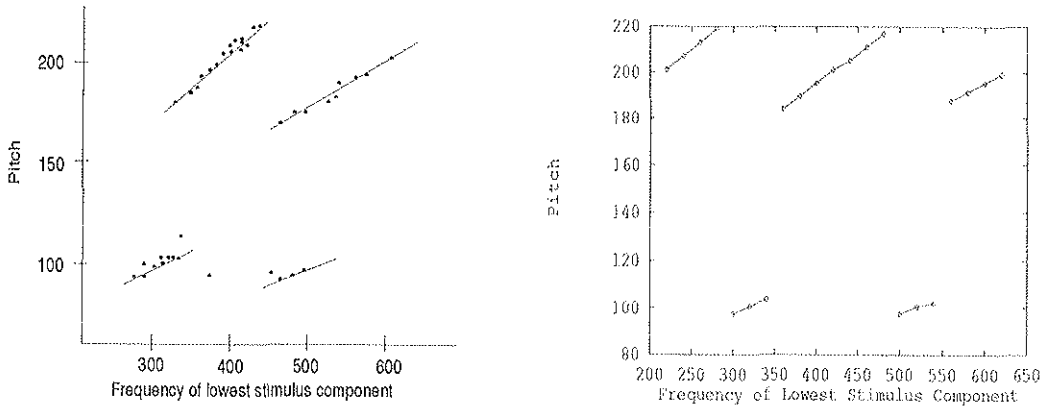

Pitch Strength
1.4
1.2
1
0.8
0.6
0.4
0.2
0
m=1
2
3
4
5
m=5
4
3
2
1
75 80 85 90 95 100 105 110 115 120 125
Pitch
(a)
Pitch Strength
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
m=5
4
3
2
1
m=1
1200
1400
1600
Pitch
(b)

FIG. 11.


200
150
100
Pitch
300
400
500
600
Frequency of lowest stimulus component
220
200
180
160
140
120
100
80
Pitch
200 250 300 350 400 450 500 550 600 650
Frequency of Lowest Stimulus Component

(a) *(b)*

FIG. 12.

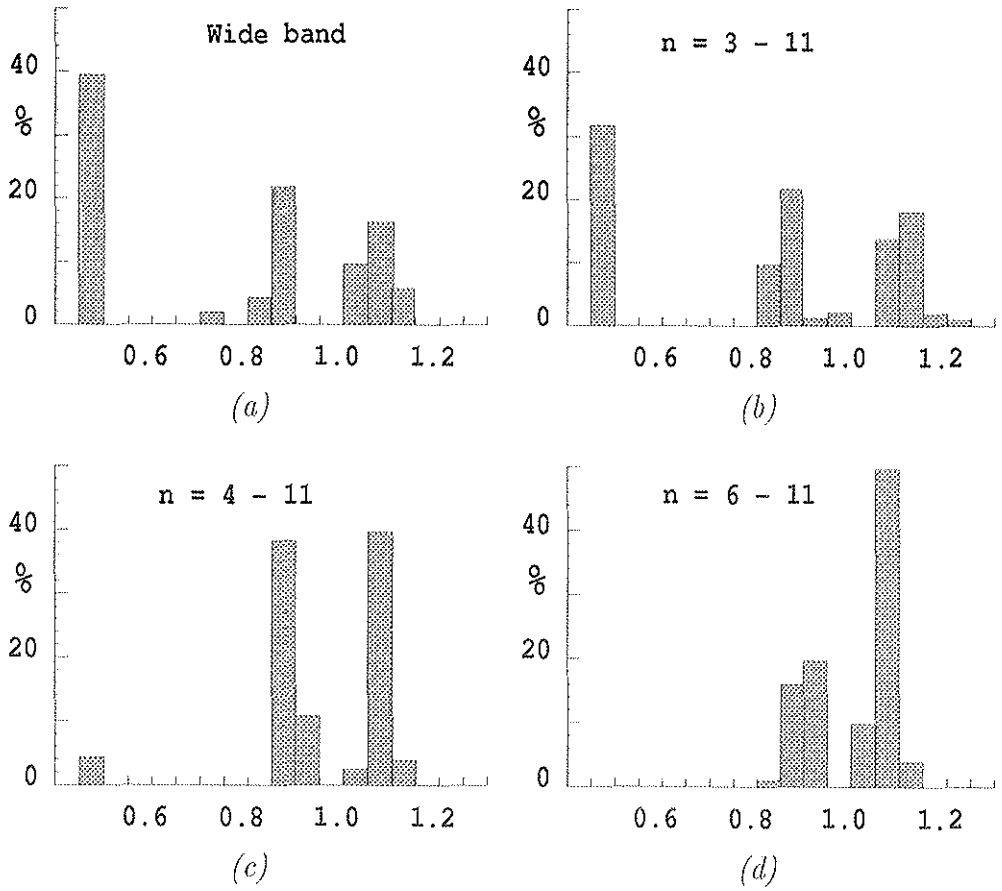


FIG. 13.

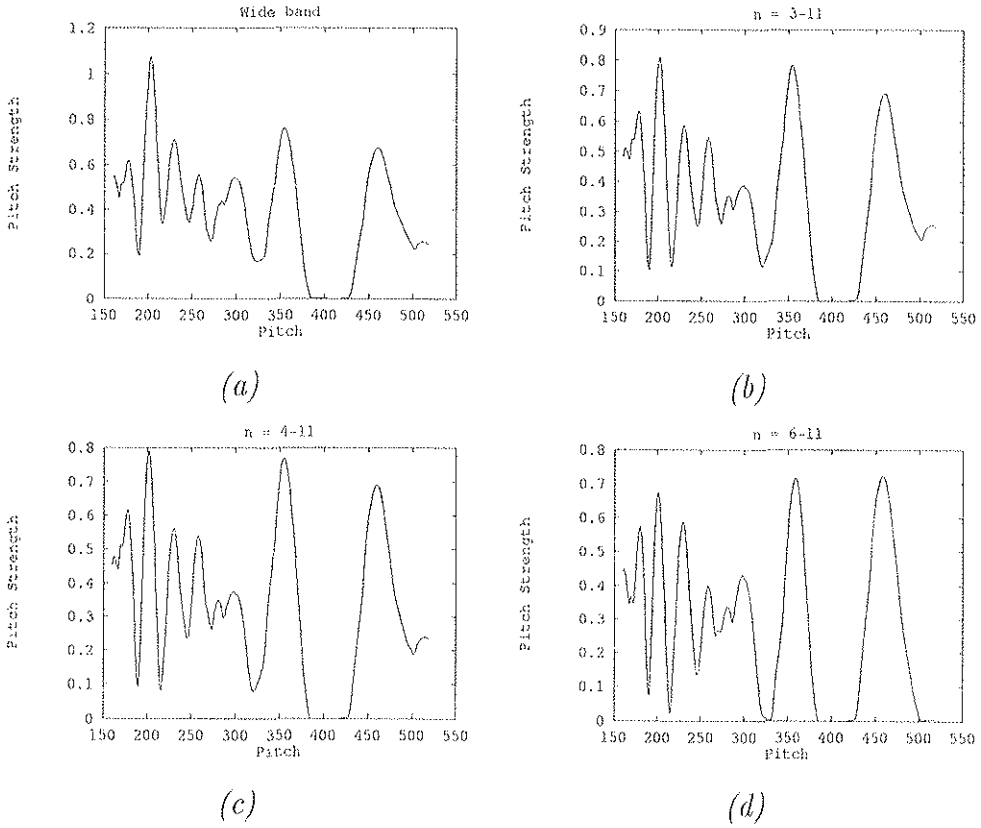
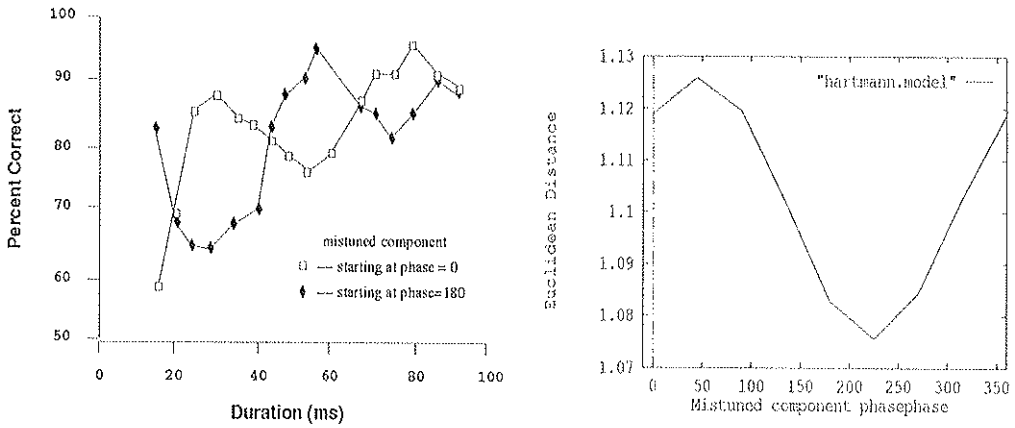


FIG. 14.


100
90
80
70
60
50
Percent Correct
0
20
40
60
80
100
Duration (ms)
mistuned component
— starting at phase = 0
— starting at phase=180
1.13
1.12
1.11
1.1
1.09
1.08
1.07
EuclGeom Distance
"hartmann.model"
0
50
100
150
200
250
300
350
Mistuned component phasephase

(a) *(b)*

FIG. 15.

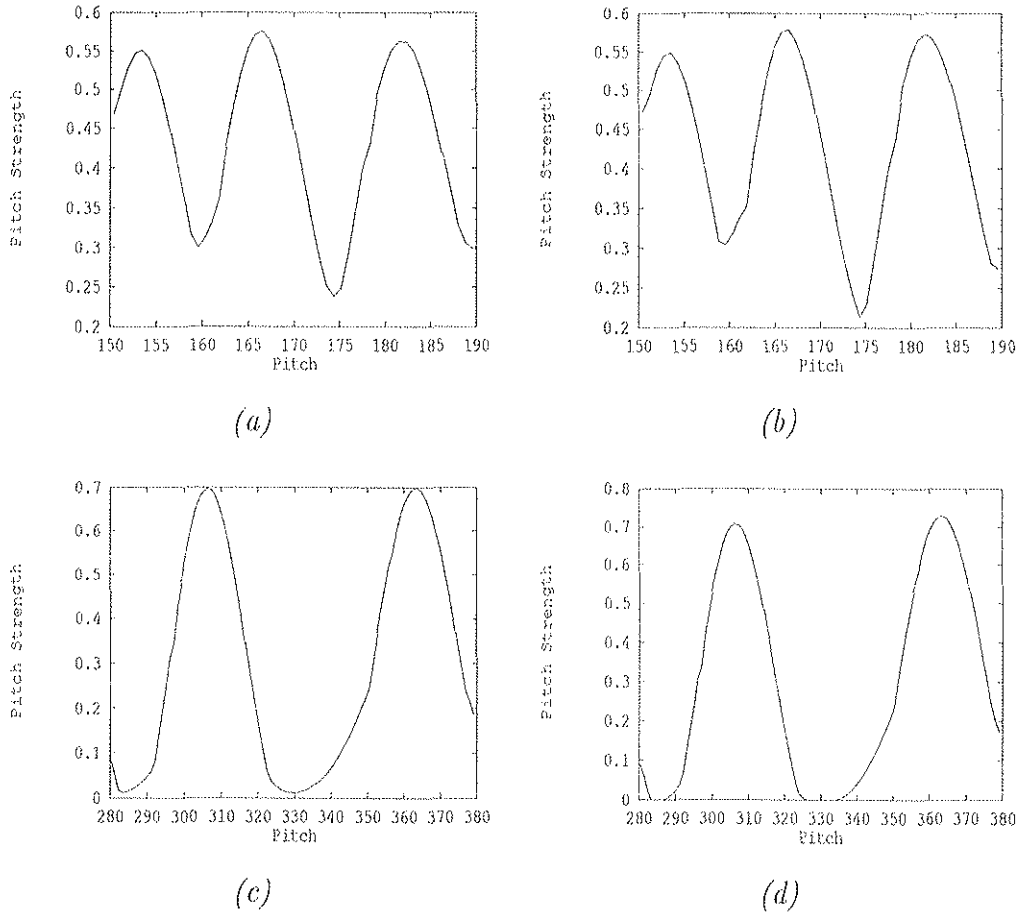


FIG. 16.

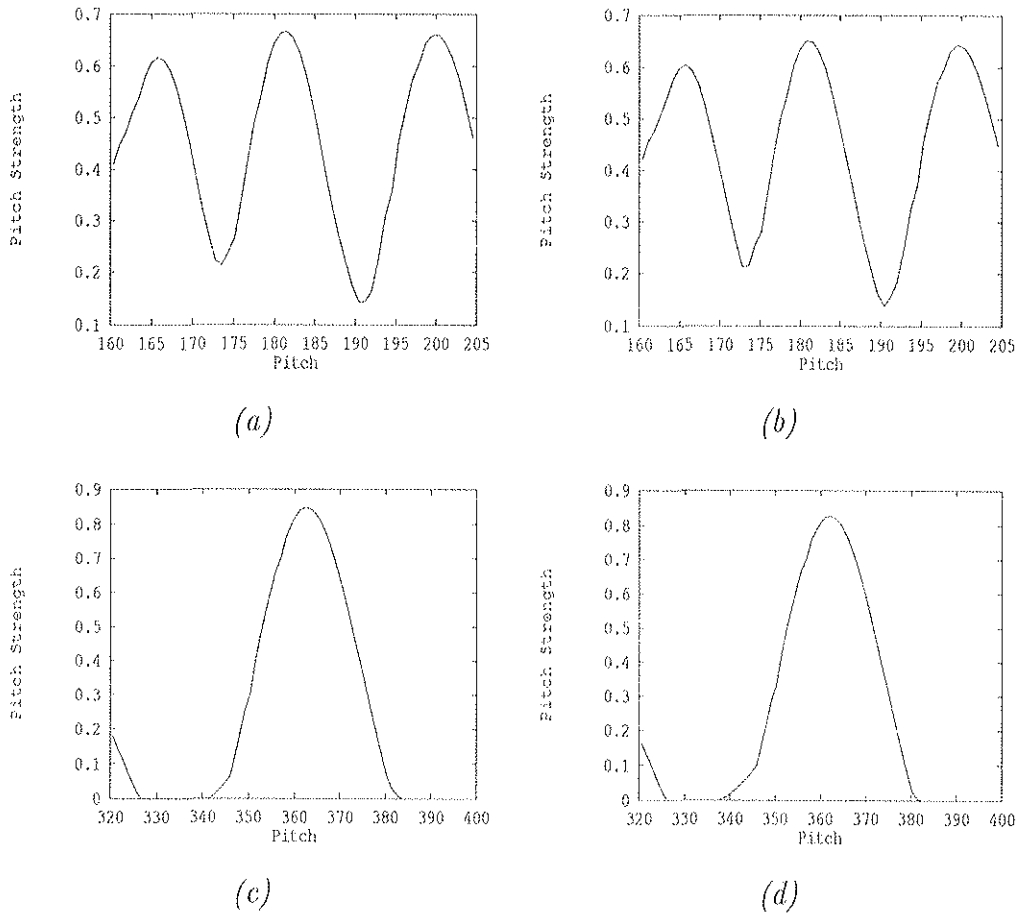

Pitch Strength
0.7
0.6
0.5
0.4
0.3
0.2
0.1
160 165 170 175 180 185 190 195 200 205
Pitch
(a)
Pitch Strength
0.7
0.6
0.5
0.4
0.3
0.2
0.1
160 165 170 175 180 185 190 195 200 205
Pitch
(b)
Pitch Strength
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
320 330 340 350 360 370 380 390 400
Pitch
(c)
Pitch Strength
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
320 330 340 350 360 370 380 390 400
Pitch
(d)

FIG. 17.

List of Tables

Model	multiple pitches	amb. region octave drop	AM noise	Cos+	Cos-	Tritone Paradox
SPINET	√	√	X	√	√	√
Goldstein	X[a]	√	X	√[b]	√[c]	NT
Wightman	√	NT[d]	X	√	X[e]	NT
Meddis & Hewitt	√	NT	√	√	√	NT
Yost	√	NT	X	√	√	NT
Terhardt	√	√	X	NT	NT	√
Duifhuis	√	NT	X	NT	NT	NT

TABLE A-II.

[a]But see Duifhuis' extension of Goldstein's model below

[b]Discussed by (Bilsen and Goldstein, 1974), they look for the optimal fit between the spectrum and a sinusoidal function

[c]Presumably the same as for Cos+ noise (Bilsen and Goldstein, 1974)

[d]But presumably would with a spectral attentional window

[e](Yost and Hill, 1979; Hill and Yost, 1978)

Model	mistune 1 component	existence region	dominance region	pitch shifts	pitch shift slopes
SPINET	√[a]	NT	√	√	√
Goldstein	NT	√	√[b]	√[c]	√
Wightman	NT	√	NT[d]	√	√
Meddis & Hewitt	√[e]	√	√	√	√
Yost	NT	NT	NT	√	√
Terhardt	NT	√	√	√[f]	√[g]
Duifhuis	NT	√	√	NT	NT

TABLE A-I.

[a]But not a great fit, as discussed in the text

[b]But with a different interpretation than the deterministic models

[c]Need to add combination tones

[d]They did not run the Plomp dominance region experimental paradigm, but presumably they could since they invoke the dominance region concept in explaining the performance of their model.

[e]But not a great fit, as discussed in the text

[f]Presumably only, discussed by (Patterson and Wightman, 1976)

[g]Presumably only, discussed by (Patterson and Wightman, 1976)