Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification

Ray Meddis and Michael J. Hewitt
Department of Human Sciences, University of Technology, Loughborough LE11 3TU, United Kingdom

(Received 18 September 1990; revised 6 November 1990; accepted 24 January 1991)

Licklider [Experientia 7, 128–133 (1951)] presented a theory of pitch highlighting the role of auditory-nerve interspike-interval timing information in the process of pitch extraction. His theory is simplified and amended and presented here as a computer implementation. This implementation has been successfully tested using simulations of a wide range of classical demonstrations of pitch phenomena including the missing fundamental, ambiguous pitch, pitch shift of equally spaced, inharmonic components, musical chords, repetition pitch, the pitch of interrupted noise, the existence region, and the dominance region for pitch. The theory is compared with a number of alternative theories and the physiological plausibility of a temporal model is considered.

PACS numbers: 43.66.Nm, 43.66.Ba, 43.66.Hg [WAY]

INTRODUCTION

This article presents a detailed exploration of the properties of a model of pitch identification based upon the extraction of timing information from auditory nerve activity. Specifically, the model uses autocorrelation analysis in a manner similar to that proposed by Licklider (1951). He proposed that pitch could be extracted from eighth nerve firing patterns by a running autocorrelation function performed on the activity of individual fibers. We have extended his suggestion in a model that combines the results of such analyses across a range of fibers sensitive to different sound frequencies into a summary autocorrelation function. Decisions about pitch are based exclusively on this summary autocorrelation function. These decisions are then compared to psychophysical results based on classical studies using human listeners.

Licklider made his original suggestion in an attempt to explain the human ability to perceive the pitch of a complex tone even though that tone contained no spectral component corresponding to that pitch. He rejected the prevailing theory (Fletcher, 1924) that distortion products of nonlinear cochlear responses could wholly explain the phenomenon. He pointed to the fact that the waveform envelope of unresolved harmonic components could be used to extract pitch information if an autocorrelation analysis could be performed. He thought that this might be achieved by a delay line mechanism at a low level in the auditory nervous system.

His theory depended on the idea that the harmonic components were not fully resolved so that the pitch period would be represented in the periodicity of the firing of the individual auditory-nerve fibers. Subsequently, however, it became clear that the most salient pitch percepts were based on complex stimuli where the individual harmonic components were widely spaced and almost certainly resolved by auditory fibers (de Boer, 1956) such that the periodicity of individual fibers would prove a poor guide to the pitch of the stimulus. This gave rise to a generation of new theories that emphasised spectral cues to pitch (Goldstein, 1973; Terhardt, 1974; Wightman, 1973a; Yost, 1982) based largely upon the resolved harmonics of the tone complex.

Such theories were successful in explaining most results of psychophysical experiments in pitch perception but they sat uneasily alongside demonstrations of nonspectral pitch phenomena such as that arising from periodically interrupted or amplitude-modulated noise (Miller and Taylor, 1948; Burns and Viemeister, 1976) and experiments showing that pitch could be heard even when only demonstrably unresolved harmonics were present in the stimulus (Moore and Rosen, 1979).

Houtsma and Smurzynski (1990) have recently addressed this issue systematically and shown unequivocally the need for pitch extraction processing to occur with both resolved and unresolved harmonics of tone complexes. Their results show that unresolved harmonics can contribute to pitch identification in addition to resolved harmonics and that information from both sources is combined to generate the pitch percept. The model to be described combines periodicity information from both resolved and unresolved regions by a simple aggregation process and thus meets the requirements of recent psychophysical results.

A second issue concerns the relative merits of spectral and temporal analysis of the stimulus. Of course, it is generally accepted that the mechanical to neural transduction process at the cochlea effects a limited frequency analysis of the signal. The "place" theories of Goldstein, Terhardt, and Wightman use this resolution as the starting point for a pattern recognition process whereby a pitch is associated with a pattern of activity across a number of locations along the basilar membrane. Temporal theories, on the other hand, assume some form of analysis based on the time intervals among spikes in the individual AN fibers. Licklider's autocorrelation analysis is just one example of a range of possible temporal theories. Goldstein and Srulovicz (1977) and Srulovicz and Goldstein (1983) developed a hybrid model that exploited temporal patterns of firing to extract enhanced spectral information which could then be used in a template matching system for identifying pitch.

 0001-4966/91/062866-17$00.80

Temporal analysis has been encouraged by physiological studies demonstrating phase locking of auditory-nerve (AN) fiber activity to tone period (Kiang *et al.*, 1965) and using autocorrelation techniques to isolate pitch effects (e.g., Evans, 1986; Horst *et al.*, 1986). The insensitivity of place methods of analysis to formant structure in high-amplitude speech sounds (Young and Sachs, 1979) has also encouraged the development of ALSR (averaged localized synchronous rate) temporal representations of AN activity.

These ideas have been reflected in modeling activity which has concentrated on the temporal information in the signal (e.g., Broadbent, 1975; Moore, 1982; van Noorden, 1982; Loeb *et al.*, 1983; Lyon, 1984; de Cheveigné, 1986; Patterson, 1987; Lazzaro and Mead, 1989; Slaney and Lyon, 1990). All authors have emphasised the advantages of temporal analysis in extracting pitch. Some voice-separation algorithms seek to use pitch difference between voices in order to select and enhance one voice at the expense of the other. In various ways these models use autocorrelation to identify the pitches of the voices (Weintraub, 1985; Gardner, 1989; Assmann and Summerfield, 1990; Meddis and Hewitt, 1990).

A systematic examination of the properties of temporal models of auditory perception is clearly required at this stage. Unfortunately, the variety of models makes this difficult. However, we have developed one such model which is sufficiently close to the main stream of recent developments to make reasonable claim to be representative of the genre. The model performs autocorrelations on the activity of groups of simulated nerve fibers and aggregates these autocorrelation functions (ACFs) to produce a *summary* autocorrelation function. It will be shown that this summary ACF contains the necessary information for the purpose of simulating human listener performance in a wide range of psychophysical studies including ambiguous pitch, pitch shift of equally spaced harmonics, pitch of stimuli with inharmonic components, pitch of musical chords, repetition pitch, and the existence and dominance regions for pitch perception. In a companion paper (Meddis and Hewitt, 1991b), we explore the ability of the same model to simulate the human listener's sensitivity to stimulus component phase effects.

I. MODEL DESCRIPTION

A. Introduction

The model outlined in Fig. 1 consists of a number of stages: (1) outer-ear frequency bandpass function; (2) middle-ear low- and high-frequency attenuation; (3) mechanical filtering of the basilar membrane; (4) mechanical to neural transduction at the hair cell; (5) refractory inhibition of firing of auditory-nerve fibers; (6) estimation of the distribution of intervals among all spikes originating from fibers within the same channel; (7) summation of interval estimates across channels; and (8) pitch extraction by inspection of the summary ACF.

Each of these stages will be described individually. However, the following characteristics are valid throughout. The signal was sampled and the model updated 20 000 times per second. From stage 4 onward, we describe the process in terms of spikes and intervals among spikes. However, the computation was carried out exclusively in terms of the probability of a spike's occurring; we did not generate and monitor individual spikes. Our measure of the time intervals among spikes was also based on the time interval between each spike and all other spikes occurring within the same channel.

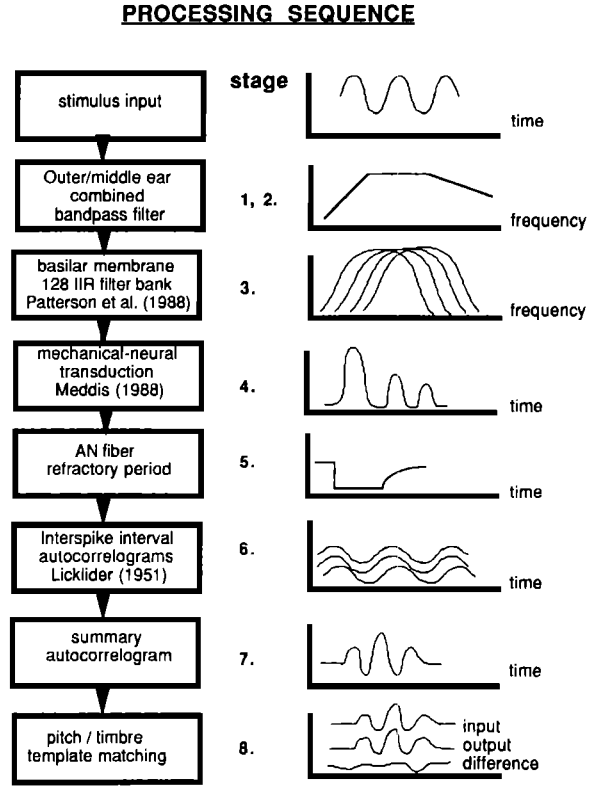


FIG. 1. Processing sequence of the model.

Our stimuli are purely number sequences and have no physical dimensions but we use the convention that a signal rms of 1 is treated as 0 dB1 (decibel *re*: rms = 1). (This is a departure from previous publications, e.g., Meddis, 1986, 1988, where a signal rms of 1 was treated as 30 dB.) Since the scale is arbitrary, we have chosen values that show a fairly close parallel with SPL ratings in psychophysical studies. On this scale, our standard auditory-nerve fiber to be described below has a threshold of 15 dB1.

Except where explicitly stated, signals are 100 ms in duration and the results represent the state of the model at the end of this time. For periodic stimuli, the stimulus length may be adjusted slightly to ensure that the stimulus ends at the major amplitude peak of the cycle. Signal waveforms in figures always represent the last 7.5 ms of the stimulus, which is three times the time constant of the model (see stage 6). This time window represents the section of the stimulus that contributed to the final running autocorrelations.

B. Stages 1 and 2: Outer- and middle-ear effects

Sound entering the outer ear is subject to a pressure gain at the tympanic membrane relative to the entrance to the ear canal; this pressure gain is maximal in the region between 2

and 5 kHz (Wiener and Ross, 1946; Djupesland and Zwislocki, 1972; Shaw, 1974). The middle ear, which couples sound energy from the external auditory meatus to the cochlea also has a bandpass pressure-transfer function but, on this occasion, the peak is nearer 1 kHz and has a much steeper slope at the low frequencies (as shown by Nedzelnitsky, 1980, for the cat). These two functions have been combined to obtain an approximately flat-topped bandpass function. The combined outer- and middle-ear functions are shown in Fig. 2. We implemented this combined function using a digital bandpass filter on the input.[1]

C. Stage 3: Basilar membrane filtering

The simulation of the bandpass filtering effect of the basilar membrane was achieved using a set of 128 digital critical-band (gammatone) filters supplied to us by Roy Patterson and John Holdsworth (Patterson *et al.*, 1988).[2] The center frequencies of the overlapping filters are equally spaced, approximately 0.25 equivalent rectangular bandwidths (ERBs) apart (Fletcher, 1940), along a scale between 80 Hz and 8 kHz. The theory behind these rounded-exponential (roex) filter shapes is given in Patterson and Moore (1986). The gammatone function provides a close approximation to the impulse response of the roex filters.

The filters have been based on psychophysical studies but there has been a qualitative convergence of physiological results and psychophysical theory (Moore, 1986) to such an extent that the use of these filters in a physiological model can be justified. In this respect an ERB [which closely resembles a critical band (Zwicker *et al.*, 1957)] represents a distance of approximately 0.9 mm on the basilar membrane. The Patterson and Holdsworth digital filters are "straight-sided" and therefore differ from physiological tuning curves that often display a tail at extreme low frequencies. However, this divergence only occurs at 30 dB below the central tip of the filter. The filters are also linear and have the same bandpass parameters at all signal amplitudes, a feature that, it is becoming clear, is also at variance with the physiological performance of auditory-nerve fibers (e.g., Deng and Geisler, 1987). Despite these reservations, we believe that the filters work well as a first approximation.[3]

Figure 3(a) shows the output of the filters[4] given a

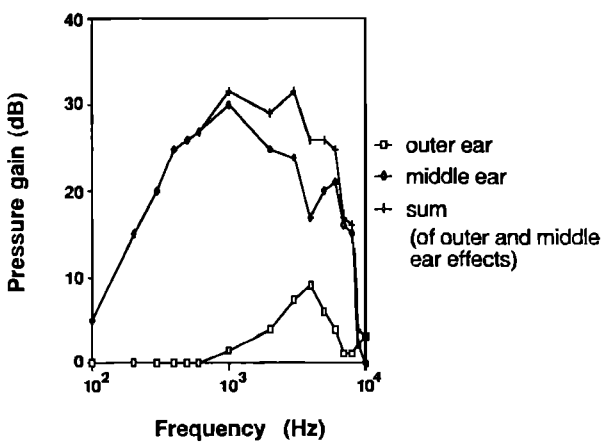


FIG. 2. Pressure gains at the outer and middle ear (see text).

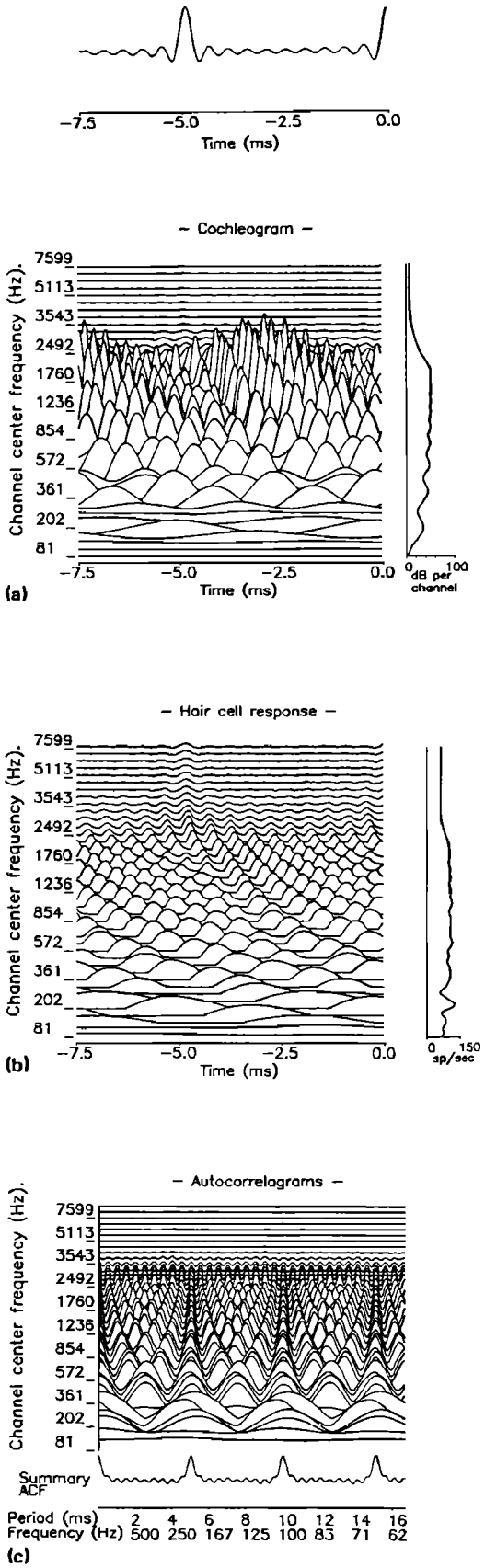


FIG. 3. Model output at three stages of processing in response to a stimulus composed of the first ten harmonics of 200 Hz at 50 dB1 per component. Only the last 7.5 ms of the model operation is shown: (a) cochleogram, basilar membrane filter output. The graph on the right represents the power output of the filters (dB1/channel); (b) hair-cell response [the graph on the right represents the spike-rate profile (spikes/s)]; and (c) the individual spike-interval histograms (ACFs) and the summary ACF (see text).

pulsed stimulus composed of ten equal amplitude (50 dB1) harmonics of 200 Hz. The vertical line graph to the right of the filter outputs in Fig. 3(a) shows the power of the output from each filter expressed in dB per channel. Because of the wide range of stimulus intensities used, the cochleogram scale is changed from display to display and the vertical line graph is necessary for a proper interpretation of the figure. The stimulus in Fig. 3(a) has been chosen to allow a comparison with the calculated excitation pattern given in Moore and Glasberg (1987, Fig. 6); there is little difference between the two.

D. Stage 4: Mechanical to neural transduction

The output of each filter was passed to a hair-cell simulator (Fig. 4) which converted the mechanical motion of the BM at that point to a probability of spike occurrence in the post-synaptic auditory nerve. The model and its properties are given more fully in Meddis (1986, 1988) and a computer program is given in Meddis *et al.* (1990).

The model assumes that the probability of a spike occurrence is linearly related to the amount of transmitter substance in the synaptic cleft between the inner hair cell and its corresponding auditory nerve

$$p(t) = h\,c(t)\,dt, \tag{1}$$

where $p(t)$ is the probability of a spike's occurring during the period t to $t + dt$, $c(t)$ is the amount of transmitter in the cleft, h is a constant of the model and dt is 1/sample rate (normally, 0.00005 s).

Transmitter substance is released into the cleft from the hair cell in amounts that depend upon the permeability of the hair cell membrane. This permeability is modulated by the signal amplitude:

$$k(t) = g\,dt\,[x(t) + A\,]/[x(t) + A + B)\,],$$

$$\text{for } [x(t) + A\,] > 0,$$

and

$$k(t) = 0, \quad \text{for } [x(t) + A\,] < 0, \tag{2}$$

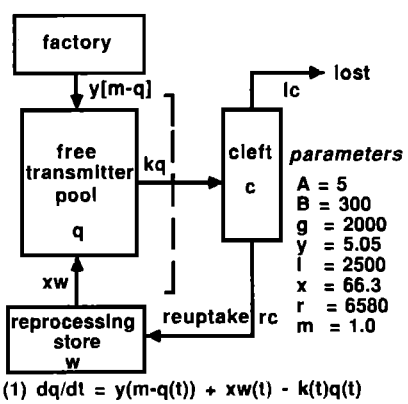


FIG. 4. Flow diagram, difference equations and parameters of the hair cell model after Meddis (1988).

where $k(t)$ is the permeability of the membrane, $x(t)$ is the instantaneous amplitude of the signal after filtering, and g, A, and B are parameters of the model.

The amount of transmitter released into the cleft at time t is $k(t)\,q(t)\,dt$, where $q(t)$ is the amount of transmitter in a free store lying close to the membrane.

Transmitter is recovered from the cleft by reuptake into the hair cell at a rate $r\,c(t)$. Some of it is, however, lost from the cleft and from the system altogether at a rate $l\,c(t)$. This loss of transmitter from the system causes adaptation of the spike rate. Recovered transmitter is held briefly in a reprocessing store [contents, $w(t)$], whence it is returned to the free transmitter store at a rate $x\,w(t)$, where it is again available for release from the cell; r, l, and x are parameters of the model.

Most of the transmitter simply cycles through the free transmitter, cleft and reprocessing stores, but deficiencies in the amount of free transmitter (caused by the losses from the cleft) are slowly restored by manufacture of new transmitter from the "factory" at a rate $y\,[m - q(t)]$. Amounts of transmitter held in the free transmitter pool are expressed here as fractions of an arbitrary maximum value m, which has been set to unity in this implementation.

The operation of the model is simulated using the permeability equation [Eq. (2)] and the three differential equations that are given in Fig. 4. The parameters of the model are also given in Fig. 4 and these have been chosen to yield a high spontaneous-rate fiber (60 spikes/s). This fiber has a saturated rate of 95 spikes/s, a rate-intensity threshold of 15 dB1 and a dynamic range of 25 dB1.

This model simulates a number of the crucial properties of auditory-nerve firing (Fig. 5). The steady-state rate-intensity function in response to pure tones is sigmoidal with a limited dynamic range. In response to a short tone burst, the firing rate shows adaptation which can be characterized by the sum of two exponentials having time constants of 75 and 7.7 ms (at sound intensities 20 dB above rate threshold) and a recovery rate (following stimulus offset) of 46 ms. The model also shows phase locking to stimulus fine structure up to 5 kHz. This latter property is crucial to the phenomena discussed later in this article. Further details of the model's performance can also be found in Meddis (1986, 1988) and Meddis *et al.* (1990). The properties of the hair-cell/AN fiber are not based on observations of any one neuron but are typical of high-spontaneous fibers which make up the majority of fibers in the auditory nerve.

Figure 3(b) shows the activity of a bank of hair cells expressed in terms of the probability of spike occurrence within the population of fibers innervating the channel. Note that for channels carrying a high-amplitude filter output, the hair cell response is approximately a compressed, half-wave rectified version of the filtered input. For low-amplitude filter outputs, however, the output follows the input with little obvious rectification or compression (see, for example, channels with center frequencies in the region of 2.5 kHz). This is also characteristic of recordings from the auditory nerve itself. The vertical line graph to the right of Fig. 3(b) shows the rate profile (spikes/s) across channels for the preceding 7.5 ms.

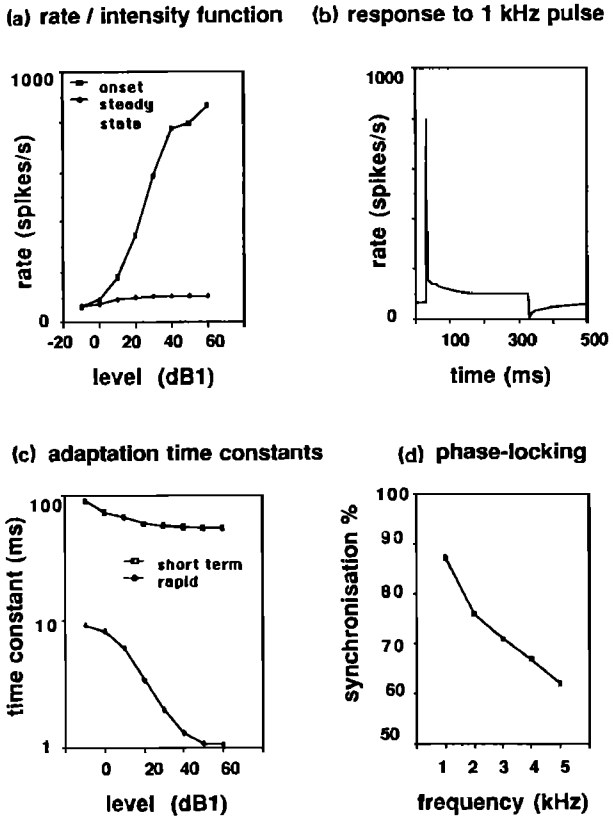


FIG. 5. Some properties of the hair cell model: (a) rate-intensity function at onset (1 ms) and during steady state; (b) rate response to a 1-kHz, 300-ms tone pulse; (c) time constants of rapid and short-term adaptation; and (d) phase locking to stimuli in the range 1–5 kHz.

Only one hair cell is implemented per channel. However, the model assumes that a large number of identical hair cells are present within each channel and that the probability function for one cell is the same for all. This is an assumption made in the interest of computational speed. However, real fibers do show a variety of different properties, particularly in spontaneous rate and dynamic range. That variety has not been modeled in this implementation. Greenberg (1986) has argued that medium and low spontaneous-rate fibers may offer an even better basis for the kind of analysis reported below.

E. Stage 5: Refractory effects

The probability of spike generation in a fiber depends on the recent history of firing of that fiber. This effect was approximated by the following formula

$$p'(t) = p(t)\left(1 - \sum_{i=1}^{\infty} p'(t-T)\, W(t-T)\right), \qquad (3)$$

where $T = i\,dt$, $p'(t)$ is the probability of spike occurrence after adjustment for refractory effects, $p(t)$ is the unadjusted probability (based on the amount of transmitter in the cleft), and $W(t)$ is an empirically derived weighting function. Note that the values of $p'(t-T)$ must be calculated before $p'(t)$ can be evaluated.

We set

$$W(t-T) = 1, \quad \text{for } t - T < 0.001 \text{ s} \qquad (4)$$

and

$$W(t-T) = 0, \quad \text{for } t - T > 0.001 \text{ s},$$

to create an absolute 1-ms refractory period. More elaborate functions for $W(t)$ are available (Gaumond *et al.*, 1982).

Note that Fig. 3(b) represents the "drive" of the hair cell and is prior to any refractory effects in the auditory-nerve fibers. For the stimuli used in this study (medium intensity, continuous stimuli), refractory effects make very little difference to the probability of firing. As a consequence, no figure is offered to show the result of including the refractory effects.

F. Stage 6: Distribution of time intervals

A histogram of the time intervals among spikes was generated separately for each filtered channel. Intervals between each and every spike were used (not just successive spikes).[5] This is convenient computationally because the probability of observing a time interval of length δt between two spikes, one of which begins at time t is

$$H(t, \delta t) = p'(t)\, p'(t - \delta t)\, dt. \qquad (5)$$

We have n fibers operating within the channel:

$$h(t, \delta t) = n\, p'(t)\, p'(t - \delta t)\, dt. \qquad (6)$$

However, the value n is a constant; it has no significant influence on what follows and it will be omitted. We assume that n is large enough to justify a smooth approximation to the spike-occurrence function.

In principle, we could study interspike intervals of any length but in the examples below we shall arbitrarily restrict δt to values between $\delta t = 0.00005$ s (the smallest model epoch length) and $\delta t = 0.01667$ s; the latter period corresponds to a pitch of 60 Hz.

Licklider specified that the summation over time should be limited by a time constant Ω, of approximately 2.5 ms. The histogram entry for the interval δt at time t for channel k can be found as follows:

$$h(t, \delta t, k) = \sum_{i=1}^{\infty} p(t-T)\, p(t - T - \delta t)\, e^{-T/\Omega}\, dt, \qquad (7)$$

where $T = i\,dt$.

The computation of the interspike interval histogram is in the form of a running autocorrelation calculation. Licklider (1959) formally demonstrated the autocorrelation nature of his system. We shall, therefore, call the histograms autocorrelation functions (ACFs) to emphasize the point. In this way, we shall also avoid confusion with the more regular use of the term "interval histogram," which is used almost exclusively, by physiologists, to refer to intervals between successive spikes. Licklider does not explain why the time constant should be 2.5 ms but more recent work by Viemeister (1979) on the temporal modulation transfer function suggests a similar value (3 ms).

For $T > 3\Omega$, expression (7) returns only very small values. Accordingly, the histogram was, in practice, computed only over a 7.5-ms period (i.e., from t to $t - 0.0075$ s). Examples of the ACFs can be seen in the body of Fig. 3(c).

G. Stage 7: Summation across channels

Channel ACFs are not used individually but are averaged across channels to generate a summary correlogram:

$$s(t,\ \delta t) = \sum_{k=1}^{128} \frac{h(t,\ \delta t,\ k)}{128}, \tag{8}$$

where $h(t, \delta t, k)$ is the running ACF $h(t, \delta t)$ for the kth channel sampled at time t.

Figure 3(c) shows the summary ACF at the bottom of the figure. Because the the scale is given in terms of time interval, frequency must be read from right to left with low frequencies on the right-hand side of the figure.

H. Stage 8a: Pitch extraction

The highest point of the summary ACF is used as a simple indicator of the perceived pitch. For many purposes, this gives a satisfactory result. For example, in Fig. 3(c), the summary ACF gives a clear peak at the pulse repetition frequency of 200 Hz (period, 5 ms) and this matches our perception of a clear pitch quality at that frequency, when a number of equally high peaks are present, the peak with the shortest lag (highest frequency) is used.

II. PITCH STUDIES

A. Missing fundamental

Figure 3(c) illustrates the pitch extraction ability of the model for a pulse train. The stimulus is composed of a fundamental and a set of successive and equal-amplitude harmonics in cosine phase. However, listeners perceive the same pitch even when the fundamental component is missing. Figure 6 illustrates the response of the model to a stimulus composed of only the 3rd, 4th, and 5th harmonics of a 200-Hz fundamental at 50 dB1 per component. The summary ACF shows a clear peak at a period of 5 ms which corresponds to a frequency of 200 Hz even though there is no component present at that frequency in the stimulus.

The mechanism of this effect is easy to see. The ACFs for channels in the region responding to the 600-Hz component show peaks at 1.66, 3.33, 5.00, 6.67, 8.33,..., ms, while the ACFs for the channels close to the 4th harmonic have peaks at 1.25, 2.5, 3.75, 5.0, 6.25,..., ms and those close to the 5th harmonic have peaks at 1, 2, 3, 4, 5, 6,..., ms. When the ACFs are summed vertically, the presence of a peak at 5 ms in all active channels creates a major peak in the summary ACF.

In this example, the three stimulus components are resolved, i.e., separated by the filter bank. However, the model produces appropriate pitch estimates even when the components are not resolved. Figure 7 shows the model's response to a three-component stimulus harmonic complex where the component tones are the 14th, 15th, and 16th harmonics of 200 Hz. Here, the individual components are not resolved by the filters. The model produces an ACF peak at 200 Hz.

Virtual pitch is clearer for resolved harmonics; unresolved harmonics give rise to a more diffuse pitch sensation. This is reflected in the multiplicity of peaks in the vicinity of the highest peak at 200 Hz. We shall return to this issue below in connection with the "existence region" of virtual

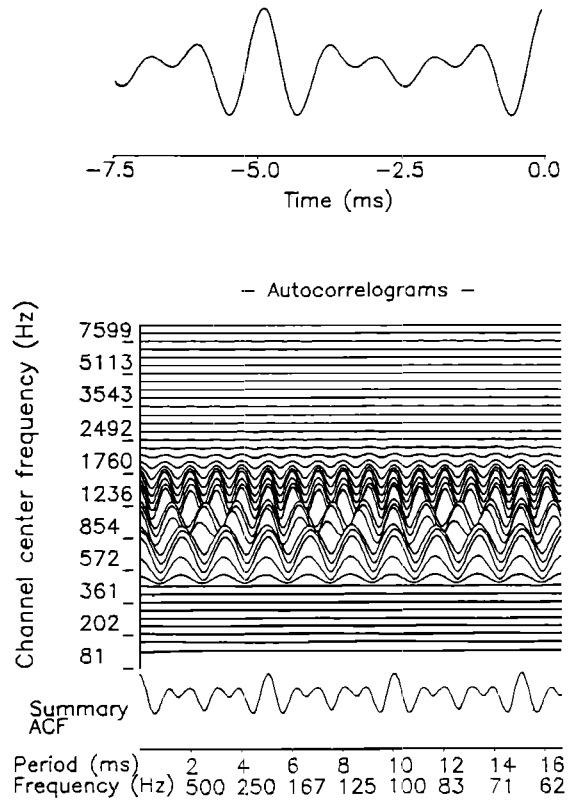


FIG. 6. Model output in response to resolved harmonics. The stimulus consists of the 3rd, 4th, and 5th harmonics of a 200-Hz fundamental.

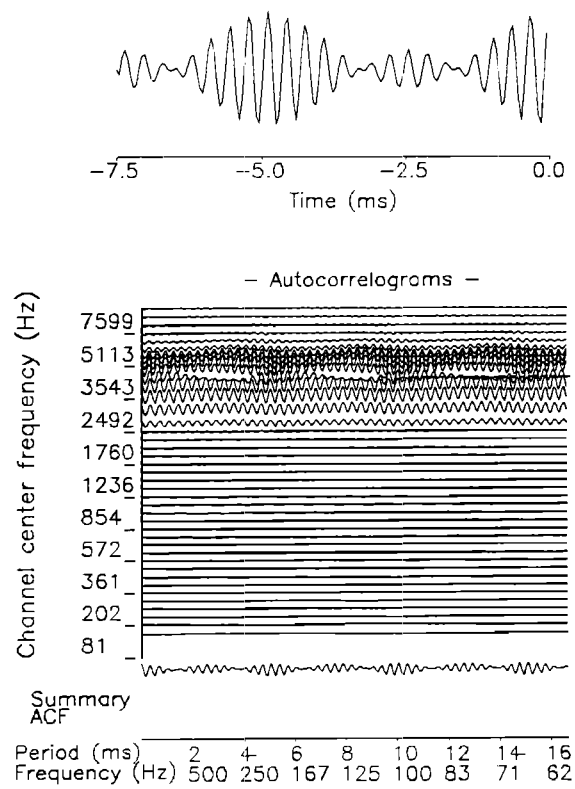


FIG. 7. Model output in response to unresolved harmonics. The stimulus consists of the 14th, 15th, and 16th harmonics of a 200-Hz fundamental.

pitch. The important point here is that the model identifies a pitch in both circumstances and it does so within the same explanatory framework. Both kinds of pitch percept arise from the same mechanism—the aggregate of a set of interspike interval histograms.

B. Ambiguous pitch

Many stimuli with a pitch quality yield a number of discrete pitch matches in matching experiments. Schouten *et al.* (1962) presented to their listeners a stimulus consisting of a 100% amplitude modulation of a 1990-Hz carrier by a 199-Hz sine function (i.e., a stimulus composed of the 9th, 10th, and 11th harmonics of a 199-Hz fundamental). Their subjects were asked to match the pitch of this stimulus with a second harmonic stimulus of variable pitch. Pitch matches were observed in the discrete regions of 145, 160, 178, 198, and 220 Hz. Between these values, there were regions where pitch matches were absent.

Figure 8 shows the summary ACF for this stimulus; it has peaks at 152.5, 166, 181, 199, and 221 Hz. The valleys between these peaks represent regions in which pitch matches would not be expected. We shall discuss the discrepancies in pitch estimates below. The first point to note is the model's property of generating a number of *discrete* candidate pitch percepts.

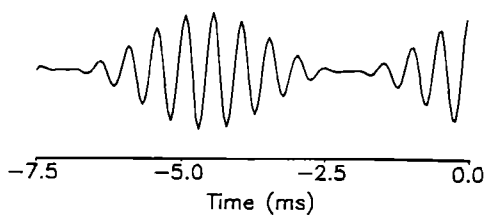
The pitch matches of Schouten *et al.* are slightly different from those predicted by the model. The two lowest matches (145 and 160 Hz) are substantially lower than predicted by the model (152.5 and 166 Hz). This effect is believed to be caused by nonlinear effects in the cochlea which generate combination tones lower in frequency (but harmonically related) to the component tones. These additional components appear to blend with the original stimulus and alter the frequency of the ambiguous virtual pitches on either side of the predominant (fundamental) pitch.

As a first approximation to this nonlinear effect, we introduced an additional component at 1592 Hz corresponding to the eighth harmonic of the series (in phase with the other harmonics and at the same amplitude), a component likely to be present as à nonlinear distortion product. It corresponds to the combination tone $2f_1 - f_2$ for the 9th and 10th harmonic as well as the combination tone $3f_1 - 2f_2$ for the 10th and 11th harmonics. The resulting summary ACF looks very similar to that in Fig. 8 but the peaks have shifted slightly to the new values of 144.5, 157.5, 179, 199, and 223 Hz. The biggest discrepancy between the data and the model's response is now less than 3 Hz. We have not pursued this line of investigation further, in this article, because the appropriate method for modeling the nonlinearities introduced by the cochlea remain controversial.

C. Pitch shift of equally spaced components

Inharmonic equally spaced tone complexes can also give rise to virtual pitch percepts (Walliser, 1969; de Boer, 1956; Schouten *et al.*, 1962; Plomp, 1976). We can think of these complexes as harmonic complexes whose components have all been shifted in frequency by the same amount. Inharmonic stimuli such as these are important theoretically because the envelope of the stimulus is not affected by the frequency shift.

Figure 9 shows the response of the model when we apply



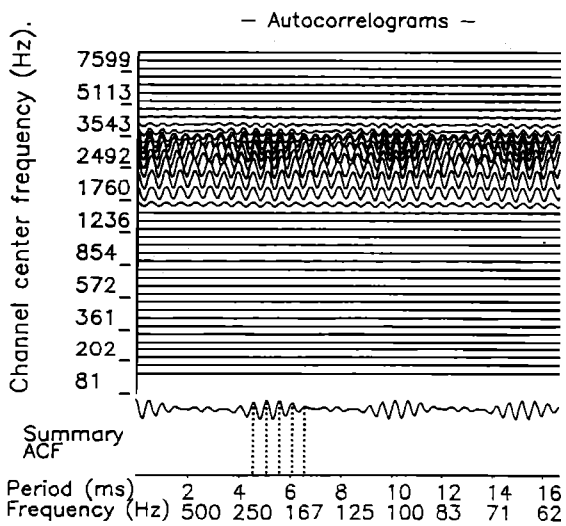


FIG. 8. Discrete pitch matches (Schouten *et al*, 1962). The stimulus is the 9th, 10th, and 11th harmonics of a 199-Hz fundamental. The marked peaks in the summary ACF suggest that pitch matches could be made at 152.5, 166, 181, 199, and 221 Hz.

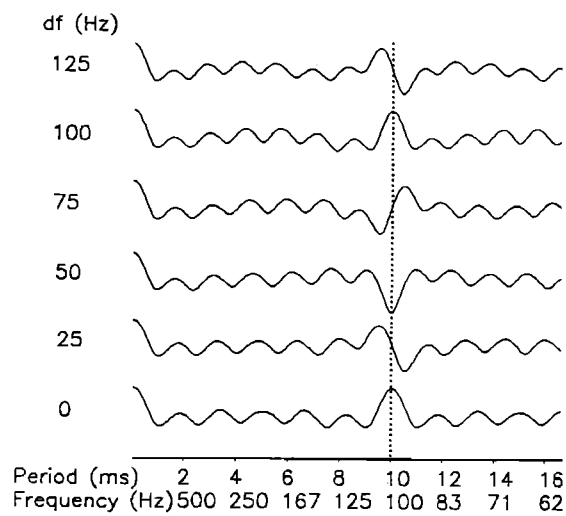


FIG. 9. Pitch of stimuli with (inharmonic) equally spaced components. The stimuli consist of a six equal-amplitude harmonics of 100 Hz with a common shift added to each component. The major peaks of the summary ACFs in the region of 100 Hz indicate possible pitch matches.

25-Hz shifts to all six tone components of a complex initially consisting of the first six harmonics of a 100-Hz fundamental. For the initial condition, a clear peak can be seen in the region of 100 Hz. The same is true when the components are all shifted by 25 Hz but the major peak is shifted to a shorter period, i.e., the complex is expected to have a slightly higher pitch. When a 50-Hz shift is applied, the summary ACF shows two peaks of equal amplitude on either side of the 100-Hz point. The 75-Hz shift condition yields a dominant peak slightly lower than 100 Hz but a 100-Hz shift again brings the components back into a harmonic relationship (2nd–7th harmonics of 100 Hz) and the dominant peak is again at 100 Hz.

Figure 10(b) shows the dominant pitch matches arising from a more complete exploration of the effect of shifts. The stimulus is again composed of six tones separated by 100 Hz. Successive stimuli have a 25-Hz shift applied to all six components. The sequence begins with a stimulus whose lowest component is 100 Hz (harmonic number is 1). After the first four 25-Hz shifts, the stimulus becomes harmonic again (harmonic number is 2). The x axis refers to the harmonic number of the lowest harmonic in the complex. The pitch estimates are restricted to a region between 90 and 100 Hz and oscillate between these extremes, always returning to 100 Hz when the stimulus is harmonic. These results are essentially the same as those reported by Patterson and Wightman (1976) on whose study these explorations are based.

Figure 11 plots the pitch matches for a similar inharmonic series based on a 400-Hz fundamental. In this case, however, we have used both six and twelve harmonics. Here, the pitch matches oscillate within a region bounded by 340 to 460 Hz but the pattern is basically similar to the 100-Hz series. In both cases the slope of the lines joining points in each segment gets less from left to right. This is a feature that is widely reported.

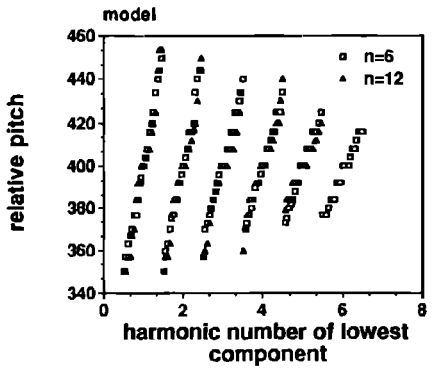


FIG. 11. Same as Fig. 10. The stimuli are equally spaced, equal-amplitude components with a 400-Hz spacing. The points plotted indicate the major peaks of the summary ACFs in the region of 400 Hz. Both 6 and 12 harmonics were used.

Patterson and Wightman (1976) also claim to have found a consistent if small tendency for the slope for 6-component stimuli to be steeper than for 12-component stimuli. This effect is present in our results but is so slight as to be hardly measurable. We take this to be consistent with the small size of the empirical result.

They observed a clear tendency for the slope of the lines in the 400-Hz condition to begin (at low harmonic numbers) higher than those for the 100-Hz condition but to become similar at higher harmonic numbers. Figure 12 illustrates that this also is true of the pitch-match predictions of the model.

When frequency shifts are applied to equally spaced

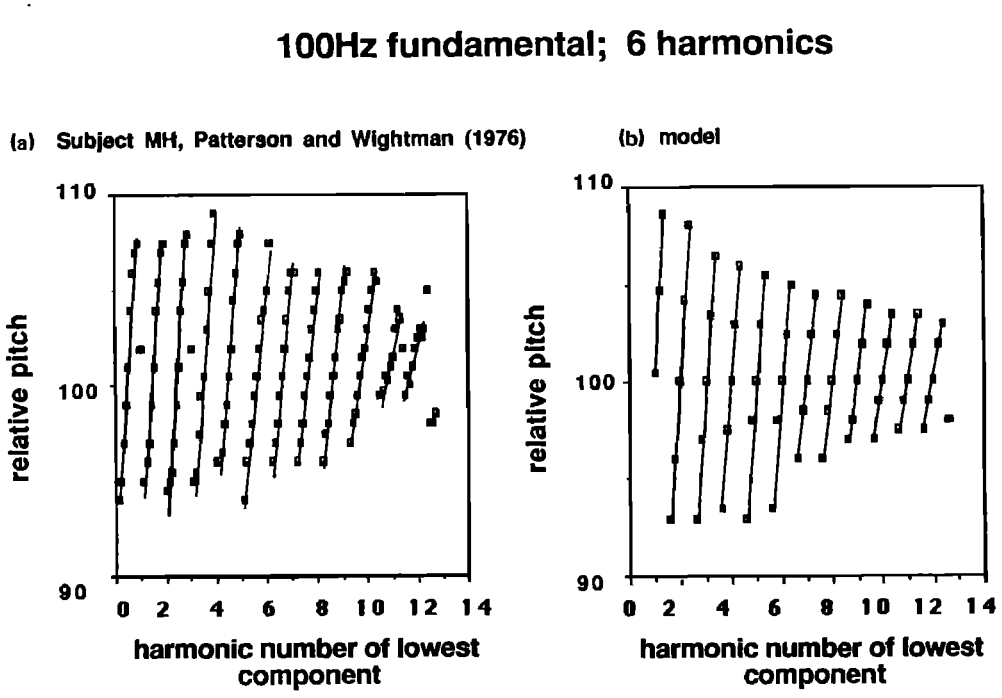


FIG. 10. Pitch of stimuli with (inharmonic) equally spaced components. Pitch-match results for stimuli consisting of six equal-amplitude harmonics of 100 Hz with a common shift added to each component. The points plotted [Fig. 10(b)] indicate the major peaks of the summary ACFs in the region of 100 Hz. The results are the same as the empirical results [Fig. 10(a)] of Patterson and Wightman (1976). The x axis shows the frequency of the lowest component in the tone complex divided by the frequency spacing. When the stimulus is harmonic, this corresponds to the harmonic number.

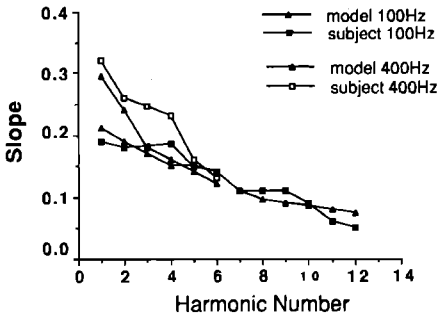


FIG. 12. Slopes of the predicted pitches in Figs. 10 and 11 (100- and 400-Hz spacing) given as a function of harmonic number of the lowest component of the complex. The results are similar to those of one of the two subjects studied by Patterson and Wightman (1976).

components, the envelope of the signal is unaffected. Accordingly, the success of the model in predicting human responses to such stimuli demonstrates that the model is not responding directly to the signal envelope but to the stimulus fine structure within the envelope.

D. Musical chords

Terhardt (1979) showed that his model could simulate an important aspect of the perception of a musical chord. He used a C-major triad produced on an electronic organ. The triad is composed of three separate tone complexes with fundamental frequencies 392, 523.2, and 659.2 Hz corresponding to G^4, C^5, and E^5, respectively. His complex consisted of the first four harmonics of 392 Hz and the first three harmonics of 523.2 and 659.2 Hz.

Figure 13 shows the summary interval ACF that results when this complex is presented to the model. The most notable feature is the prominent peak at 130 Hz. This peak corresponds to the musical note C^3 which is the root note of the chord, a note that is not present in the chord but that defines the identity of the chord and explains its harmonic quality. This demonstrates a congruence between our model and Terhardt's model on this point.

E. Repetition pitch

When a broadband noise stimulus is delayed and added to itself, the resulting stimulus has a pitch that falls as the delay increases (Bilsen, 1966, 1970; Bilsen and Ritsma, 1970; Fourcin, 1965; Yost, *et al.*, 1978). If the delayed noise is added to itself with positive sign (cos + stimulus), the pitch of the stimulus is normally judged to be $1/t$ Hz, where t is the delay in seconds. However, if the noise is added back with negative sign (cos − stimulus), the pitch is weaker and also ambiguous; it may be reported as either slightly higher or slightly lower than $1/t$ Hz. Figures 14 and 15 show the output of the model for cos + and cos − stimuli with a delay of 0.002 s. The weakness of the effect is entirely consistent with observations using human listeners. The time constant of integration (Ω) of the autocorrelation is only 2.5 ms in this implementation. As a consequence, a single summary autocorrelogram reflects the activity of the model over a period of less than 7.5 ms. Because the effect is weak and variable from stimulus to stimulus, we have repeated the analysis many times to establish the reliable features. We assume that listeners average this weak percept over a long period which involves many samples of the stimulus. Each figure shows six example summary ACFs in response to six different noise stimuli.

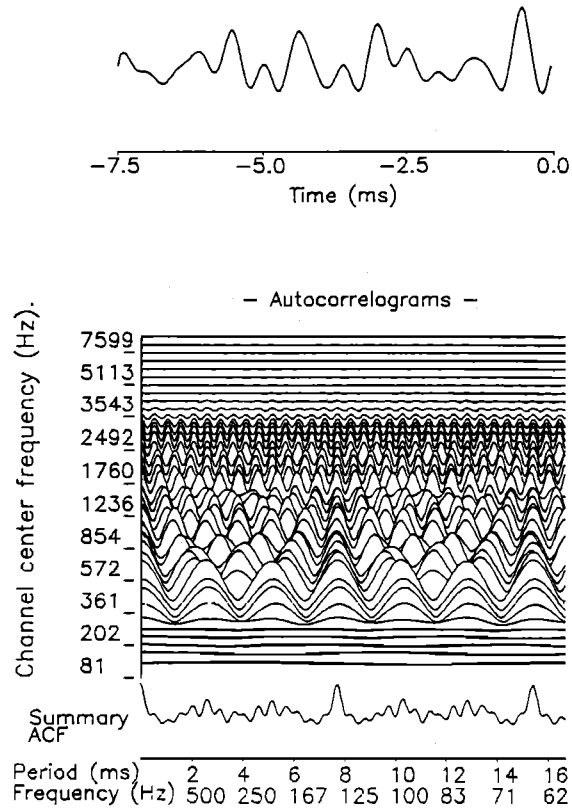


FIG. 13. C-major triad. The stimulus consists of the first four harmonics of 392 Hz (G^4) and the first three harmonics of 523.2 Hz (C^5) and 659.2 Hz (E^5). The summary ACF has a peak at 130 Hz corresponding to the note C^3, which is the root-note of the chord.

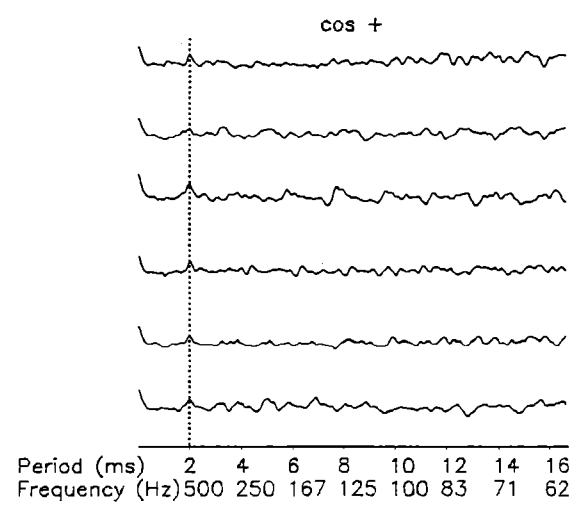


FIG. 14. Repetition pitch. Six example summary ACFs produced by the model in response to stimuli consisting of Gaussian noise that has been delayed by 2 ms and then *added* to itself. Note the peaks in the summary ACF at 2 ms (500 Hz). This peak is consistently present for all stimuli, while other peaks are inconsistent.

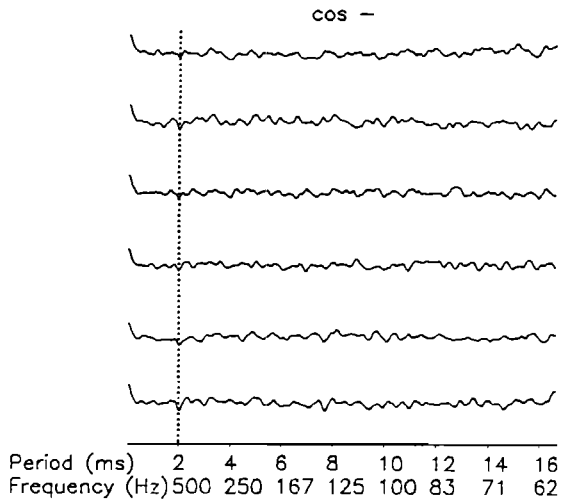


FIG. 15. Repetition pitch. Six example summary ACFs produced by the model in response to stimuli consisting of Gaussian noise that has been delayed by 2 ms and then *subtracted* from itself. Note the *dip* in the summary ACF at 2 ms (500 Hz). This dip is consistently present for all stimuli, while other peaks are inconsistent. The weak peaks on either side of the dip can be identified as approximately 550 and 450 Hz by averaging the ACFs for a number of stimuli.

1. cos+ stimuli

For the cos + stimulus, the summary ACFs all show a peak at a period of 2 ms (Fig. 14), corresponding to a pitch of 500 Hz. This is not the only peak in the summary ACF but it is the largest and is consistent over every sample of cos + noise that we have tested while the other peaks vary in height and location from sample to sample. Figure 16(a) shows how the location of this pitch peak in the average of 30 summary ACFs varies as a function of the delay t. It is a straight line function giving a pitch equal to $1/t$ in agreement with the empirical results of Yost *et al.* (1978).

This result is hardly surprising. Cos + noise has a long-term ripple spectrum with spectral peaks at frequencies $1/t$, $2/t$, $3/t$, etc. In the long run, therefore, this stimulus should behave like a harmonic stimulus with a fundamental frequency of $1/t$. In the short run, the spectrum is highly variable and, as a consequence, the pitch will be weaker. While a single summary ACF shows a number of competing candidate peaks, in the long run the peak at $1/t$ emerges as the consistently highest peak.

2. cos− stimuli

The picture is even less clear for the cos − stimulus (Fig. 15). The main feature here is a consistently observed *dip* at the lag of $1/t$. On either side of the dip are poorly defined peaks. In order to compare the model results with those of Yost *et al.* (1978), 30 different ACFs were computed for each delay and averaged. The pitch predictions given in Fig. 16(b) represent the two highest peaks in the averaged summary ACF within $\pm$ 20% of the location of the main dip at $1/t$. The straight lines show the fit suggested by Yost *et al.* (1978) for their empirical data. They suggest slopes of $1.1/t$ and $0.9/t$. The model results fit these values well.

Like cos + noise, cos − noise also has a long-term ripple spectrum with spectral peaks at $k/t + 1/2t$, i.e., it has much in common with an *inharmonic* complex of tones equally spaced midway between the corresponding harmonics. We have already seen (Fig. 9) that such inharmonic complexes give rise to an ambiguous pitch percept that offers a choice of two weak pitches on either side of the fundamental of the corresponding harmonic series. Accordingly, our autocorrelation approach would be expected to produce a prediction of an ambiguous pitch percept for both cases. While these considerations are true for the long-term spectrum of cos − noise, the random fluctuations in the stimulus over time do mean that short-term sampling gives rise to an insecure pitch estimate.

The exact value for the pitch percept for cos − noise and its strength depends on the bandpass characteristics of the noise (Yost *et al.*, 1978). The reason for this, in terms of the autocorrelation method, is that each spectral peak produces an autocorrelation peak in the region of $1/t \pm 1/(2kt)$. While the dip is constant across components at a period of $1/t$, the location of the adjacent autocorrelation peak varies with the harmonic number k. For high harmonic numbers this peak is close to $1/t$ but at low harmonic numbers it is well separated from $1/t$. A wide-band signal, therefore, generates a pitch that is a compromise over a number of values. A narrow-band signal will reduce the range of the locations of peaks on either side of the dip at $1/t$. The results in Fig. 16(b) were collected using broadband noise which would be expected to produce the least clearly defined percept.

F. Sinusoidally amplitude-modulated noise

Miller and Taylor (1948), using interrupted broadband noise, reported accurate pitch matches at the rate of interruption of the noise. Unlike cos + or cos − noise, the spectra of interrupted-noise stimuli contain no spectral peaks at either the perceived pitch frequency or its harmonics but are invariant with respect to modulation frequency. The pitch of interrupted noise offers a serious problem, therefore, to pitch theories that extract patterns from spectral representations. The existence of this pitch effect was one of the phenomena encouraging the development of temporal accounts of pitch perception. However, the percept is weak and the interpretation of empirical results can be problematic for technical reasons (see Burns and Viemeister, 1976, for a review). Many of these technical problems can be dealt with by using amplitude-modulated (AM) noise, and Burns and Viemeister (1976) showed that the pitch percept is weak but at least strong enough for the recognition of melodies and musical intervals.

The model was therefore tested using broadband noise 100% amplitude modulated at a range of frequencies between 100 and 400 Hz. Initial testing showed that the peaks in the summary ACF occurred in roughly the expected places but were small and inconsistent. The same method as described above for cos + and cos − noise was used to get a clearer picture; for each AM frequency, 30 stimuli were presented to the model and the resulting summary ACFs were averaged. Figure 16(c) shows the values of the highest peaks in the averaged summary ACFs between 80 and 400 Hz.

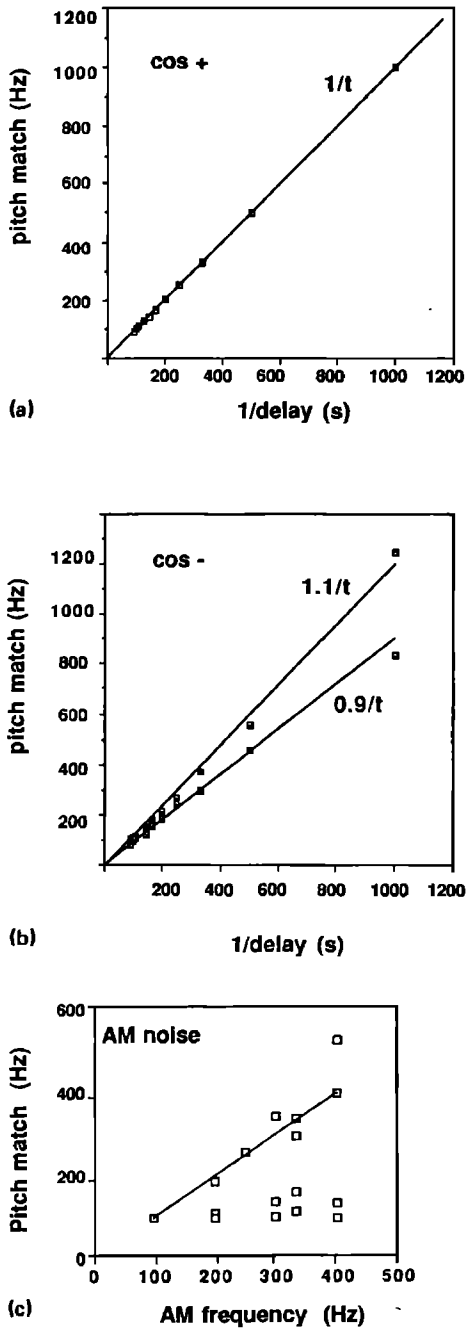


FIG. 16. Pitches for noise stimuli: (a) best pitch matches for cos + signals as a function of delay using the average of the summary ACFs for 30 cos + noise stimuli (see text); (b) best pitch matches for cos − signals as a function of delay using the average of the summary ACFs for 30 cos − noise stimuli (see text); the lines represent the best fit functions for the psychophysical data collected by Yost *et al.* (1978); and (c) possible pitch matches for 100% AM noise using the average of the summary ACFs for 30 AM noise stimuli at each AM frequency (see text).

First, the highest peak was identified and then any other peaks within 5% of the highest peak were noted. The picture is obviously confused, but it can be seen that peaks were always observed that correspond to the AM frequency. It is important to remember that pitch matching is extremely difficult for these stimuli. Burns and Viemeister were forced to use the recognition of musical intervals in their study for this very reason. It is possible that the weak results obtained with the model correspond to the observation that pitch matching is difficult with AM noise.

G. Existence region for virtual pitch

Ritsma (1962) explored the existence region for pitch using amplitude-modulated signals. He found that a clear pitch percept was limited to carrier frequencies below 5 kHz and that pitch values were limited to a region between 60 and 800 Hz. Using the ratio of the carrier frequency to modulation rate as the harmonic number of the stimulus, he also showed that the pitch percept weakened as the harmonic number increased.

We presented 100% amplitude-modulated carriers to the model using harmonic numbers 5,10,15,20,25 applied to carrier frequencies (f_c) of 1,2,3,4, and 5 kHz. The stimuli used were computed as follows:

$$x(t) = 0.5\sin[2\pi(n-1)ft] + \sin[2\pi nft] + 0.5\sin[2\pi(n+1)ft], \quad (9)$$

where n is the harmonic number and $f\,(=f_c/n)$ is the fundamental of the set of harmonics. The resulting summary ACFs are shown in Fig. 17.

As the *carrier frequency* of the signal increases, there is a reduction in the prominence of the individual peaks in the ACF; i.e., the peak/trough ratio declines; clear peaks are visible at 1-kHz carrier frequency but the ACF descends to little more than a ripple at 5 kHz. This occurs in the model because, at high frequencies, phase locking to the fine structure of the waveform is dramatically reduced.

As the *harmonic number* of the stimulus is increased, the number of candidate pitch peaks increases in the region of the pitch value of the stimulus. The percept, therefore, would be expected to become weaker because of the greater difficulty in selecting between adjacent peaks. This effect was anticipated by Moore (1977) in his analysis of the problem.

These two results are therefore consistent with Ritsma's results because they predict a weakening of the pitch percept as the harmonic number increases and as the carrier frequency increases. However, the model predicts no attenuation of the low pitch percept for very low fundamental frequencies. As long as the interval ACF is extended to cover longer periods, the model will accommodate lower pitches.

To accommodate this limitation, the model would need to be elaborated with some *ad hoc* principle. However, it has been suggested (Moore, personal communication) that low pitch can be heard as low as 30 Hz. Accordingly, it may be appropriate to delay any change to the model until a consensus emerges on this point.

H. Dominance region

The pitch of harmonic complexes is generally clearer when lower harmonics are present. This effect has been systematically studied by Ritsma (1967), Plomp (1967), Bilsen (1973) and, more recently, Houstma and Smurzynski (1990). For fundamental frequencies below 400 Hz, harmonics lying between the 3rd and the 5th contribute most to the strength of the pitch percept. Plomp (1967) suggested

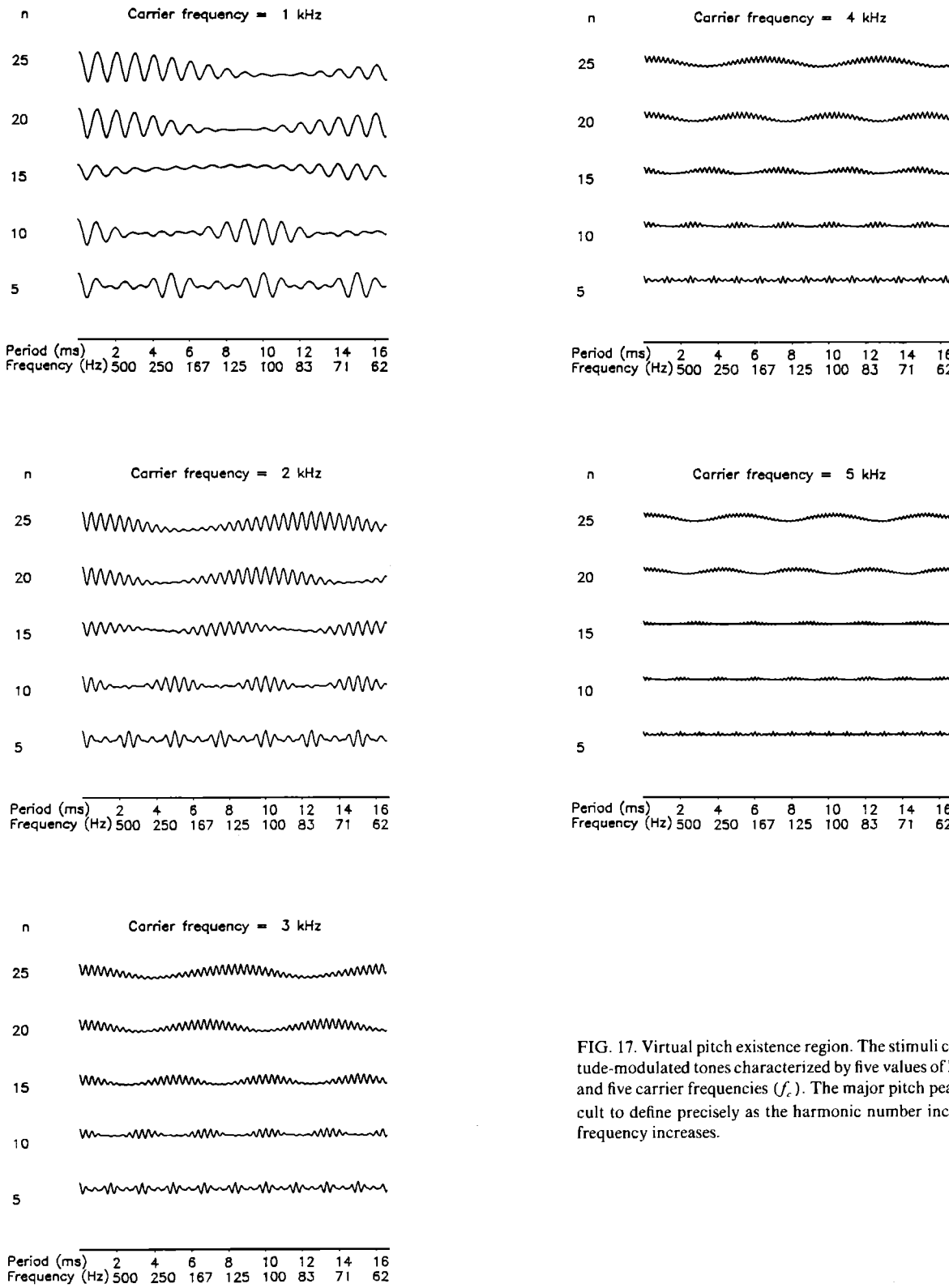


FIG. 17. Virtual pitch existence region. The stimuli consist of 100% amplitude-modulated tones characterized by five values of harmonic number (n) and five carrier frequencies (f_c). The major pitch peak becomes more difficult to define precisely as the harmonic number increases and the carrier frequency increases.

that center of the dominance region falls as the pitch of the complex rises.

It was decided to compare the model's performance with the results of Plomp's study because he covers the widest pitch range. Where the pitch range overlaps, his results can be shown to be consistent with Ritsma's and Bilsen's studies (Plomp, 1976, p. 117). Plomp asked subjects to compare two stimuli and say which had the higher pitch. One stimulus was a harmonic complex

$$x(t) = a \sum_{n=1}^{12} \cos(2\pi nft). \tag{10}$$

For the second stimulus, he shifted harmonics 1 to m down by 10% in frequency and harmonics $m + 1$ to 12 up in frequency by 10%:

$$x(t) = a \sum_{n=1}^{m} \cos[2\pi n(0.9)ft] + a \sum_{n=m+1}^{12} \cos[2\pi n(1.1)ft]. \quad (11)$$

The issue was whether the upward-shifted harmonics would dominate the downward-shifted harmonics in determining the perceived pitch. Plomp found that a small number of low harmonics shifted downward would outweigh the upward shift of the rest of the higher harmonics; that is, stimuli where the low and high harmonics were equally balanced would typically be characterized by a low value of m.

When testing the model we used the second stimulus only [Eq. (12)] and noted the height of the two peaks corresponding to the shifted pitches. These two peaks were always on either side of and close to the single peak generated by the standard stimulus. If the higher pitch peak was bigger than the lower pitch peak, we treated the predicted pitch percept as higher in pitch than the standard stimulus and vice versa.

Figure 18 illustrates the procedure with a 300-Hz fundamental. The first stimulus ($m = 1$) consists of 1 lowered component (270 Hz) and 11 raised components (660, 990, 1320,..., Hz). The summary ACF shows a single prominent peak in the region of 330 Hz. The stimulus for the top row ($m = 4$) consists of four lowered components (270, 540, 810, 1080 Hz) and eight raised components (1650, 1980, 2310,..., Hz). Its summary ACF has its highest peak at 270 Hz. The dominance crossover point occurs just above $m = 2$. On interpolating the numerical representation of the figures, we can specify a crossover point at $m = 2.4$.

Figure 19 shows the results for a range of fundamental frequencies (200, 300, 450, and 675 Hz); the points on the graph indicate the point at which the lower harmonics came to dominate the higher harmonics. Plomp's (1976, Fig. 45, p. 116) results and the model data follow a parallel course and demonstrate a clear low harmonic dominance in the working of the model.

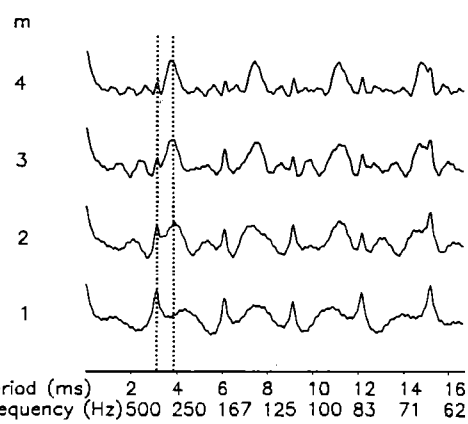


FIG. 18. Summary ACFs for four stimuli consisting of the m lowest harmonics of 300 Hz shifted 10% down in frequency and $(12 - m)$ highest harmonics shifted 10% up. As m increases, the major pitch-peak shifts from 330 to 270 Hz The transition point occurs between $m = 2$ and $m = 3$ and is numerically estimated as $m = 2.4$ for the purpose of Fig. 19.

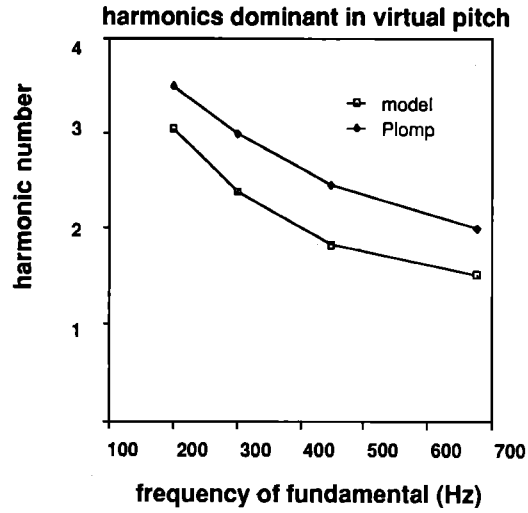


FIG. 19. Transition points in Plomp's dominance experiment (see Fig. 18) for 200-, 300-, 450-, and 675-Hz fundamentals. Plomp's data are interpolated from his results (Plomp, 1976, p. 116, Fig. 45).

The effect, in the model at least, appears to result from two underlying causes. In a competitive situation, such as that devised by Plomp, the lower frequencies dominate the summary ACF because the density of the channels is much greater at lower frequencies. As a result, more of these channels contribute to the summary ACF. On the other hand, low-frequency components are subject to greater relative attenuation at the outer and middle ear. For low-frequency components, around 100 Hz, this relative attenuation may outweigh the advantage gained by the increased number of channels in this region. As a result, it is harmonics above the first and second that will dominate for a fundamental near 100 Hz.

As the fundamental frequency increases, however, there is less attenuation of low-frequency components due to middle-ear effects, while the relative number of contributing channels remains roughly constant. This gives a clear advantage to the lower frequency components which are more richly represented. As a consequence, the region of dominance shifts toward the lower harmonic numbers. In this way the model offers an explanation for Plomp's result which shows the center of dominance declining as fundamental frequency rises.

Evans (1986) and Greenberg (in the discussion to Evan's paper), review this matter in the context of actual nerve-fiber recordings. Both emphasize the role of fibers midway between two harmonics which often show a periodicity at the fundamental due to the interaction of the two adjacent harmonics (even though the harmonics are resolved in the activity of fibers at their own center frequencies). This offers an alternative explanation of the dominance region which emphasizes the need to have adjacent harmonics high enough to permit interaction. The phenomenon of "central pitch" (Houtsma and Goldstein, 1972) to be discussed below suggests that harmonics which cannot interact with each other at the level of the AN (because they are presented to different ears) may still contribute to the pitch of the stim-

ulus. Harmonic interaction cannot, therefore, be a complete explanation of the dominance region. Nevertheless, harmonic interaction does have a role to play and this should be captured by the present model if the filter functions are realistic.

III. DISCUSSION

A. Modeling pitch

We have supplemented a model of the auditory periphery with a modified version of Licklider's (1951) suggestion that auditory-nerve fiber interspike intervals be aggregated in a manner that is mathematically similar to autocorrelation functions. The only novel feature in the model was the principle of aggregating information across frequency-selective channels. With this amendment, his approach has proven satisfactory as a model of human listener's pitch perception.

Below we shall explore the similarities and differences between this model and established theories of pitch. The similarities will be much more prominent than the differences. The model has the merit of being able to deal with pitch percepts arising from both resolved and unresolved signal components while offering an explanation for the dominance of low-frequency harmonics in giving rise to the sensation of "low pitch." In so doing, it combines insights from two rival schools of thought, championing, respectively, place and temporal theory. We may note in this context that Licklider's original concept of a "duplex" theory was aimed at this very goal.

The major departure from Licklider's theory concerns the cross-channel combination of information. While he was content to pass some of the more difficult aspects of pitch extraction up to higher "learning centers," we have shown that the simple expedient of adding his running ACFs together to form a summary ACF provides us with an effective way of predicting a wide range of pitch phenomena. This means that a periodicity which is common to a number of channels will be clearly represented in the summary, whether or not that periodicity is most prominent in any individual channel.

The use of a short time constant (2.5 ms) for the running ACF followed Licklider's original suggestion. It has relatively few implications for the extraction of pitch from continuous and unchanging tone complexes. Such a short time constant does imply that only very short exposures to stimuli are necessary to perceive pitch. While this may be the case for some stimuli, repetition pitch for noiselike stimuli may require considerably longer stimulation (Buunen, 1980). The model was not very successful in handling cosripple pitch and interrupted noise using this time constant; much longer integration periods were required to achieve satisfactory results. To simulate this longer integration period to help define the weak noise pitches, we averaged summary ACFs over 30 stimuli. This implies a second integration principle in addition to and following the 2.5-ms time constant of the ACFs.

B. Physiological parallels

The biggest conceptual difficulty with the time interval detection approach is knowing how it is achieved physiologically. There is a considerable gap between the knowledge that timing information is available and identifying a nucleus in the nervous system that does the job. We accept this and would prefer to present the model as a statement that the nervous system (a) extracts pitch using timing information and (b) pools this information across filtered channels.

It is known that the central nervous system can create and make use of delays. Goldberg and Brown (1969) have shown that cells in the medial superior olive in the brain stem are responsive to time disparities between signals presented to the left and right ear. Yin and Chan (1988) have demonstrated the viability of a cross-correlation hypothesis of sound source localization on the basis of a temporal analysis of single-cell activity in the inferior colliculus. Nevertheless, we do not know for certain of any location in the central nervous system where Licklider's scheme of auditory delay lines is manifestly active.

It is possible that the timing information, which is central to Licklider's model, is preserved but dealt with in a different way from his proposal. The ventral cochlear nucleus and the inferior colliculus are two nuclei containing cells that fire in response to the periodicity of the stimulus (e.g., Frisina, 1983; Rees and Palmer, 1989; Kim and Leonard, 1988). Frisina even reports cells showing a bandpass modulation transfer function for stimuli of moderate intensity. Schreiner and Langner (1988) report tonotopically arranged cells within the central nucleus of the inferior colliculus which have best modulation frequencies between 100 and 500 Hz. Within each frequency-selective lamina of the nucleus, the best modulation frequencies were arranged concentrically with higher modulation frequencies located more centrally. This mapping of modulation frequency to place is the best candidate for the physiological basis of the ACFs discussed above. Unfortunately, their research does not make clear the way in which the information is combined across channels.

The mechanism behind the sensitivity to amplitude modulation in cochlear nucleus cells remains unclear but it may involve inhibitory delays along the neural pathways (Manis and Brownell, 1983) or be caused by a nonlinear response to the AN rectification of the filtered waveform. Whether physiological delay lines can be found to support Licklider's simple scheme seems doubtful since no very long delays have ever been reported. Suga (1990), however, has presented suggestive results implying temporal coding of pitch in bats; while these need not apply to human listening, they do demonstrate related specializations of nervous systems.

The model, therefore, remains neutral on the exact mechanism whereby temporal information is extracted from the activity of the AN fibers although it clearly reflects the work of neurophysiologists such as Evans (1989) and Horst *et al.* (1986), who have shown for many stimuli that single-fiber recordings contain enough information in their temporal firing patterns to explain many psychoacoustic discriminations.

C. Existing pitch theories

While the model is exceedingly computationally intensive, its handling of the auditory-nerve spikes is very simple; it merely requires that time intervals among spikes within channels be aggregated across channels. All of the properties of the model demonstrated above are a consequence of this principle. It is of interest to see how this principle compares with other theories of pitch.

Goldstein's (1973) theory of pitch is concerned only with resolved components of a complex signal. At the heart of his theory is a system that takes a resolved tone component, say 600 Hz, and postulates a number of possible pitch values for the complete stimulus. These are submultiples of the harmonic, *viz.*, 300, 200, 150, 120 Hz,..., etc. Another harmonic, say 750 Hz, suggests pitch values of 375, 250, 187.5, 150, 125,..., etc. The only value they have in common is 150 Hz and the system estimates that this is the most likely value for the pitch.

Goldstein's theory is considerably more subtle than this, however. It takes into account the likely error in estimating the frequency of the signal components and the consequences this will have for the pitch estimates. However, the essence of the system is the discovery of a common submultiple or an approximation thereof. For resolved harmonics, our physiological model is reasonably faithful to this prescription. The individual ACFs for the two channels featuring the 600- and the 750-Hz components will have peaks at periods corresponding to the candidate pitches indicated above. When the two ACFs are added together, the coincidence of peaks at a period corresponding to 150 Hz will result in a prominent peak at this point in the summary ACF.

We differ, however, in our explanation of the dominance effect whereby low harmonics dominate the pitch percept. In Goldstein's theory, this is explained in terms of the accuracy of estimation of the frequency of individual components, while in our model it emerges as a trade-off between the amplitude attenuation of low-frequency components and the greater salience of lower harmonics caused by the higher density of the low-frequency channels.

Goldstein's theory is a "central" theory in that the pitch effects require an explanation beyond the cochlea. We intend our version of Licklider's ideas to be a central theory too because the autocorrelation calculations are neural. While the model has been presented as monaural, we anticipate a binaural version where the spikes from two ears are combined centrally. Such a model could embrace the demonstration by Houstma and Goldstein (1972) that two harmonics presented dichotically can produce a pitch sensation. It is a feature of the model that it combines information from different channels within the same ear and a simple extension to a binaural version would allow it to combine information from two ears and generate a "central pitch."

Terhardt's theory of virtual pitch (Terhardt, 1974, 1979, 1980), like Goldstein's, uses submultiples of isolated signal components but, unlike Goldstein, he assumes that the pitch is identified using a pattern recognition system based on extensive learning. This is in sharp contrast to our model which has an explicit mechanism for extracting pitch information owing little to learning or experience. Licklider's theory was itself dependent on a learning system and therefore can be considered a precursor of Terhardt's view. However, in our version of Licklider's model, we have abandoned the learning component in favor of a more direct computational mechanism.

The theories of Wightman, (Wightman, 1973a, 1973b; Wightman and Green, 1974) and Yost (Yost and Hill, 1978, 1979; Yost *et al.*, 1978; Yost, 1982) involve performing an autocorrelation on the spectral profile representing the energy output of the filter bank. They are successful in predicting many of the same phenomena as the other theories. Indeed, de Boer (1977) has shown a family similarity between Wightman's theory and those of Terhardt and Goldstein. They have not, however, been shown capable of explaining the weak pitch associated with interrupted noise. A critical feature of these models is their insensitivity to phase, which will be discussed more fully in a companion paper (Meddis and Hewitt, 1991b). Spectral profile theories are not sensitive to phase, while the modified Licklider model is sensitive to phase.

Temporal models in the tradition of Schouten, Licklider, Ritsma, Moore, and van Noorden are clearly much closer to our account. Early attempts to predict pitch on the basis of the filtered signal envelope or time intervals between signal function peaks have been abandoned—largely because of their inability to combine information across channels. Because virtual pitch can be heard using widely spaced low-frequency (and hence resolved) components, theories based on single channels must prove inadequate.

Moore (1977, 1982) saw that, in a physiological account, the time intervals would need to be based on individual spikes in the auditory nerve. In his "crude sketch" and elsewhere he anticipated many of the pitch effects described above. Our model is clearly closer to this view than any other model and his supporting arguments sustained us during the development of our system. There are numerous differences in detail and our account is more complete but the essential components are broadly similar.

Patterson's (1987) pulse ribbon model is also similar to our own, not least because we use his filter bank and we both work with neural events precisely located in time. Our use of spike probabilities is only a superficial distinction from his use of discrete spikes. We agree strongly with Patterson on the importance of the *pattern* of the nerve impulses. An important difference, however, is the introduction in our model of an explicit mechanism for specifying the pitch heard and for measuring the degree of perceptual similarity between stimuli with similar frequency components but phase differences. Patterson's pulse ribbon provides a clear visual insight into the phenomena while our model attempts a more precise numerical prediction. This stylistic difference does not, in itself, imply a fundamental difference of approach.

One point does merit attention, however. In Patterson's theory the traveling wave delays introduced by the cochlear filtering are troublesome. These delays impair the visual appeal of the pulse ribbon and make patterns difficult to identify. He aligns the ribbon so that the peak of each waveform occurs at the same time and clear patterns can be distin-

guished which provide better insights into the effect of phase changes. By contrast, our model makes no such adjustment and yet suffers no obvious penalty.

The propagation delay produces effects that are obvious only across channels. However, Licklider's model measures intervals only within channels; measurements that are, as a consequence, blind to the actual cochlear delay that applies to that channel. The cochlear delay is therefore discounted at the level of the interval ACFs *before* the cross-channel summation occurs. Accordingly, cochlear delay is not a problem that needs to be addressed directly. Of course, it may be that Patterson's suggestion that "the auditory system accommodates for the propagation delay in the cochlea" may be essentially the same as our view that the problem does not arise because of the nature of the processing arrangements. A full comparison with Patterson's model will have to wait until detailed numerical predictions based on his approach are forthcoming.

Despite these minor differences, our model and those of Moore, van Noorden, and Patterson sit comfortably together as a class which emphasize the explanatory power of temporal relationships among auditory-nerve fiber spikes. As such, they form a subset of temporal theories that feature directly the auditory-nerve response, a tradition explicitly adopted by Licklider (1951), whose original insights inspired the work reported above.

ACKNOWLEDGMENTS

We would particularly like to acknowledge our debt to Roy Patterson, Brian Moore, Quentin Summerfield, Malcolm Slaney, and Alain de Cheveigné for their comments and advice during the development of the model and the preparation of this article. We thank Trevor Shackleton for valuable comments on earlier versions of the paper. The article has also benefited considerably from the comments of a number of anonymous reviewers. This work was supported by a grant from the Science and Engineering Research Council Image Interpretation Initiative.

[1] The following digital filter coeficients were used:

$$y_i = 0.8878x_i - 0.8878x_{i-2} - 0.2243y_{i-1} + 0.7757y_{i-2}.$$

[2] Much of the early development of the model was based on a similar set of filters supplied by Martin Cooke, Department of Computer Science, Sheffield, UK.

[3] The filters work with 12-bit accuracy ($\pm$ 2047). To optimize accuracy in practice, the stimulus was rescaled to make the peak instantaneous amplitude equal to 2047 before filtering. Later, the output amplitude was restored to its original scale. As a result, the sensitivity of the filters was a joint function of the 12-bit resolution and the amplitude of the highest peak of the input signal.

[4] For clarity of presentation, only one in every three filter outputs is shown.

[5] This histogram is not the same as an interspike interval (ISI) histogram, which only measures time intervals between successive spikes.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–796.

Bilsen, F. A. (**1966**). "Repetition pitch: monaural interaction of a sound with the repetition of the same, but phase shifted, sound," Acustica **17**, 265–300.

Bilsen, F. A. (**1970**). "Repetition pitch: its implication for hearing theory and room acoustics," in *Frequency Analysis and Periodicity Detection in Hearing*," edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden).

Bilsen, F. A. (**1973**). "On the influence of the number and phase of harmonics on the perceptibility of the pitch of complex signals," Acustica **28**, 60–65.

Bilsen, F. A., and Ritsma, R. J. (**1970**). "Repetition pitch and its implication for hearing theory," Acustica **22**, 53–73.

Boer, E. de (**1956**). "On the residue in hearing," Doctoral dissertation, University of Leiden.

Boer, E. de (**1977**). "Pitch theories unified," in *Psychophysics and Physiology of Hearing*," edited by E. F. Evans and J. P. Wilson (Academic, London).

Broadbent, D. E. (**1975**). "The sixth annual Fairey lecture: waves in the eye and ear," J. Sound Vib. **41**, 113–125.

Burns, E. M., and Viemeister, N. F. (**1976**). "Nonspectral pitch," J. Acoust. Soc. Am. **60**, 863–869.

Buunen, T. J. F. (**1980**). "The effect of stimulus duration on the prominence of pitch," in *Psychological, Physiological and Behavioural Studies in Hearing*, edited by G. Van Den Brink and F. A. Bilsen (Delft U.P., The Netherlands).

de Cheveigné, A. (**1986**). "A pitch perception model," Proc. ICASSP-86, 897–900.

Deng, L., and Geisler, C. D. (**1987**). "A composite auditory model for processing speech sounds," J. Acoust. Soc. Am. **82**, 2001–2012.

Djupesland, G. and Zwislocki, J. J. (**1972**). "Sound pressure distribution in the outer ear," Scand. Audiol. **1**, 197–203.

Evans, E. F. (**1986**). "Cochlear nerve fibre temporal discharge patterns, cochlear frequency selectivity and the dominant region for pitch," in *Auditory Frequency Selectivity*, edited by B. C. J. Moore and R. D. Patterson (Plenum, New York).

Evans, E. F. (**1989**). "Representation of complex sounds in the peripheral auditory system with particular reference to pitch perception," in *Structure and Perception of Electroacoustic Sound and Music*, edited by S. Nielzen, and O. Olsson, Exerpta Medica (Elsevier, Amsterdam).

Fletcher, H. (**1924**). "The physical criterion for determining the pitch of musical tone," Phys. Rev. **23**, 427–437

Fletcher, H. (**1940**). "Auditory patterns," Rev. Mod. Phys. **12**, 47–65.

Fourcin, A. J. (**1965**). "The pitch of noise with periodic spectral peaks," in Rapports 5^{e} Congres International d'Acoustique, Lige, Vol. Ia, B. 42.

Frisina, R. D. (**1983**). "Enhancement of responses to amplitude modulation in the gerbil cochlea nucleus," Ph.D. thesis, University of Syracuse, New York.

Gardner, R. B. (**1989**). "An algorithm for separating simultaneous vowels," Br. J. Audiol. **23**, 170–171.

Gaumond, R. P., Molnar, C. E., and Kim, D. O. (**1982**). "Stimulus and recovery dependency of cat cochlear nerve spike discharge probability," J. Neurophysiol. **48**, 856–873.

Goldberg, J. M., and Brown, P. B. (**1969**). "Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localisation," J. Neurophysiol. **32**, 613–636.

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

Goldstein, J. L., and Srulovicz, P. (**1977**). "Auditory-nerve spike interval as an adequate basis for aural frequency measurement," in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson (Academic, London).

Greenberg, S. (**1986**) "Possible role of low and medium spontaneous rate cochlear nerve fibers in the encoding of waveform periodicity," in *Auditory Frequency Selectivity*, edited by B. C. J. Moore and R. D. Patterson (Plenum, New York).

Horst, J. W., Javel, E., and Farley, G. F. (**1986**). "Coding of spectral fine structure in the auditory nerve. I. Fourier analysis of period and interspike interval histograms," J. Acoust. Soc. Am. **79**, 398–416.

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

Houtsma, A. J. M., and Smurzynski, J. (**1990**). "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. **87**, 304–310.

Kiang, N. Y. -S., Watanabe, T., Thomas, E. C., and Clark, L. F. (**1965**). "Discharge of auditory fibers in the cat's auditory nerve," Res. Mon. **35** (MIT, Cambridge, MA).

Kim, D. O., and Leonard, G. (**1988**). "Pitch-period following response of cat cochlear nucleus neurons to speech sounds," in *Basic Issues in Hearing*, edited by H. Duifhuis, J. W. Horst, and H. P. Wit (Academic, London).

Lazzaro, J., and Mead, C. (1989). "Silicon modeling of pitch perception," Proc. Natl. Acad. Sci. U.S.A. 86, 9597–9601.
Licklider, J. C. R. (1951). "A duplex theory of pitch perception," Experientia 7, 128–133.
Licklider, J. C. R. (1959). "Three auditory theories," in *Psychology: A Study of a Science*, edited by S. Koch, (McGraw-Hill, New York).
Loeb, G. E., White, M. W., and Merzenich, M. M. (1983). "Spatial cross-correlation," Biol. Cybern. 47, 149–163.
Lyon, R. F. (1984). "Computational models of neural auditory processing," IEEE ICASSP 84, 3.
Manis, P. B., and Brownell, W. E. (1983). "Synaptic organisation of eight nerve afferents to cats dorsal cochlear nucleus," J. Neurophysiol. 50, 1156–1181.
Meddis, R. (1986). "Simulation of mechanical to neural transduction in the auditory receptor," J. Acoust. Soc. Am. 79, 702–711.
Meddis, R. (1988). "Simulation of auditory-neural transduction: Further studies," J. Acoust. Soc. Am. 83, 1056–1063.
Meddis, R. and Hewitt, M. J. (1991a). "Modeling the identification of concurrent vowels with different fundamental frequencies," submitted to J. Acoust. Soc. Am.
Meddis, R., and Hewitt, M. J. (1991b). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: Phase sensitivity," J. Acoust. Soc. Am. 89, 2883–2894.
Meddis, R., Hewitt, M. J., and Shackleton, T. M. (1990). "Implementation details of a computational model of the inner hair-cell/auditory-nerve synapse," J. Acoust. Soc. Am. 87, 1813–1818.
Miller, G. A., and Taylor, W. G. (1948). "The perception of repeated bursts of noise," J. Acoust. Soc. Am. 20, 171–182.
Moore B. C. J. (1986). "Parallels between frequency selectivity measured psychophysically and in cochlear mechanics," Scand. Audiol. Suppl. 25, 139–152.
Moore, B. C. J. (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. F. Evans, and J. P. Wilson (Academic, New York).
Moore, B. C. J. (1982). *An Introduction to the Psychology of Hearing* (Academic, London), 2nd ed.
Moore, B. C. J., and Glasberg, B. R. (1987). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Hear. Res. 28, 209–225.
Moore, B. C. J., and Rosen, S. M. (1979). "Tune recognition with reduced pitch and interval information," J. Exp. Psychol. 31, 229–240.
Nedzelnitsky (1980). "Sound pressures in the basal turn of the cat cochlea," J. Acoust. Soc. Am. 68, 1676–1689.
Noorden, L. van (1982). "Two channel pitch perception," in *Music, Mind and Brain*, edited by M. Clynes (Plenum, London).
Patterson, R. D. (1987). "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am. 82, 1560–1586.
Patterson, R. D., and Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity*, edited by B. C. J. Moore (Academic, London).
Patterson, R. D., and Wightman, F. L. (1976). "Residue pitch as a function of component spacing," J. Acoust. Soc. Am. 59, 1450–1459.
Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). "Spiral vos final report, Part A: The auditory filterbank," Cambridge Electronic Design, Contract Rep. (Apu 2341).
Plomp, R. (1967). "Pitch of complex tones," J. Acoust. Soc. Am. 41, 1526–1533.
Plomp, R. (1976). *Aspects of Tone Sensation* (Academic, London).
Rees, A., and Palmer, A. R. (1989). "Neuronal responses to amplitude modulation and pure-tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise," J. Acoust. Soc. Am. 85, 1978–1994.
Ritsma, R. J. (1962). "Existence region of the tonal residue. I," J. Acoust. Soc. Am. 34, 1224–1229.
Ritsma, R. J. (1967). "Frequencies dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am. 42, 191–198.
Schouten, J. F., Ritsma, R. J., and Cardozo, B. J. (1962). "Pitch of the residue," J. Acoust. Soc. Am. 34, 1418–1424.
Schreiner, C. E., and Langner, C. E. (1988). "Coding of temporal patterns in the central auditory nervous system," in *Auditory Function*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York).
Scrulovicz, P., and Goldstein, J. L. (1983). "A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum," J. Acoust. Soc. Am. 73, 1266–1276.
Shaw, E. A. G. (1974). "The external ear," in *Handbook of Auditory Physiology*, edited by W. D. Keidel and W. D. Neff (Springer, Berlin).
Slaney, M., and Lyon, R. F. (1990). "A perceptual pitch detector," Proc. 1990 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Albuquerque, NM, 357–360.
Suga, N. (1990). "Biosonar and neural computation in bats," Sci. Am., June 1990, 34–41.
Terhardt, E. (1974). "Pitch, consonance and harmony," J. Acoust. Soc. Am. 55, 1061–1069.
Terhardt, E. (1979). "Calculating virtual pitch," Hear. Res. 1, 155–182.
Terhardt, E. (1980). "Towards understanding pitch perception: problems, concepts and solutions," in *Psychophysical, Physiological and Behavioural Studies of Hearing*, edited by G. van der Brink, and F. A. Bilsen (Delft University, Delft).
Viemeister, N. F. (1979). "Temporal modulation transfer functions based on modulation thresholds," J. Acoust. Soc. Am. 66, 1364–1379.
Walliser, K. von (1969). "Zusammenhänge zwischen dem Schallreiz und der Periodentonhöhe," Acoust. 21, 319–329.
Weintraub, M. (1985). "A theory and computational model of monaural auditory sound separation," doctoral dissertation, Stanford University, Stanford, CA.
Wiener, F. M., and Ross, D. A. (1946). "The pressure distribution in the auditory canal in a progressive sound field," J. Acoust. Soc. Am. 18, 401–408
Wightman, F. L. (1973a). "The pattern transformation model of pitch," J. Acoust. Soc. Am. 54, 407–416.
Wightman, F. L. (1973b). "Pitch and stimulus fine structure," J. Acoust. Soc. Am. 54, 397–406.
Wightman, F. L., and Green, D. M. (1974). "The perception of pitch," Am. Sci. 62, 208–215.
Yin, T. C. T., and Chan, J. C. K. (1988). "Neural mechanisms underlying interaural time sensitivity to tones and noise," in *Auditory Function*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York).
Yost, W. A. (1982). "The dominance region and ripple noise pitch: A test of the peripheral weighting model," J. Acoust. Soc. Am. 72, 416–425.
Yost, W. A. and Hill, R. (1979). "Models of the pitch and pitch strength of ripple noise," J. Acoust. Soc. Am. 66, 400–410.
Yost, W. A. and Hill, W. (1978). "Strength of pitches associated with ripple noise," J. Acoust. Soc. Am. 64, 485–492.
Yost, W. A., 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.
Young, E. D., and Sachs, M. B. (1979). "Representation of steady-state vowels in the temporal aspects of discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. 66, 1381–1403.
Zwicker, E., Flotorp, G., and Stevens, S. S. (1957). "Critical bandwidth in loudness summation," J. Acoust. Soc. Am. 29, 548–557.