please return to Dick Lyon.
(415) 494-4325

XEROX
Palo Alto Research Center

System Sciences Lab
LSI Systems Area

A Signal-Processing Model of Hearing

Richard F. Lyon -- July 11, 1978

Abstract -- **The fields of physiology, physics, psychoacoustics, neurology, signal processing, etc., all contribute to the design of a model which emulates the hearing process. Results from these fields are surveyed, and combined to guide the design of a realizable model, as described in this paper. The resulting digital signal processing structure may be designed into custom VLSI silicon chips, and will be useful in several real-time sound processing applications, such as speech recognition systems. The model is preliminary, without determination of all the parameters.**

The first version of this report is being written as a term paper for Stanford University course EE208, Biological Information Processing, and is not intended for further distribution at this time.

INTRODUCTION

The intent of this research paper is to collect information describing how all the processing layers of hearing work (to the current state of knowledge and some speculation), to extract the relevant information processing functions from the various types of descriptions, and to assemble those functions into a coherent signal processing model. The model should be especially useful in speech processing systems (such as recognition), but also useful for music, noise, and other kinds of sound processing. Models of the intermediate mechanisms (such as neurons and the central nervous system) are considered for what they tell us about the information processing being done, but are not in themselves important to the model being developed.

This paper is organized around five major parts of the model, plus introductory sections on history, the scope of features considered, and guiding principles, and followed by concluding sections on implementation and summarizing discussion.

HISTORICAL OVERVIEW OF HEARING MODELS

> "Even in our era of technological wonders. the performances of our most amazing machines are still put in the shade by the sense organs of the human body. Consider the accomplishments of the ear. It is so sensitive that it can almost hear the random rain of air molecules bouncing against the eardrum. Yet in spite of its extraordinary sensitivity the ear can withstand the pounding of sound waves strong enough to set the body vibrating. The ear is equipped, moreover, with a truly impressive selectivity. In a room crowded with people talking, it can suppress most of the noise and concentrate on one speaker. From the blended sounds of a symphony orchestra the ear of the conductor can single out the instrument that is not performing to his satisfaction." (Georg von Bekesy, 1957)

The ability to implement computing machines has come a long way in the past few decades, and we can now confidently say that it is possible to build a machine that can do all the computations necessary to duplicate the performance of the ear, if only we knew what computations those were. We are ready to do the work necessary to merge the VLSI silicon implementation technology with the knowledge and speculations of how hearing works, to arrive at an electronic ear -- a forerunner of a new age of truly amazing machines.

How do we hear things? Obviously with our ears, but what happens inside to convert minute pressure waves into recognizable sounds? The early work on sound reflection and conduction by Kircher (1650) and Schelhammer (1684) was important, but did not help answer the hard part of the question, perception of pitch and timbre. The first reasonable contributions to the answer came when surgeons were able to find and understand the importance of the cochlea (from the Greek for a snail with a spiral shell). Bekesy reports that the gross structure of the ear was known to pioneering anatomists of the sixteenth and seventeenth centuries, but the details did not start to come out till late in the eighteenth century. A good early picture showing the coiled tube structure of the cochlea was produced in 1837 by Lincke. His drawings of the cochlear partition, however, contain little important detail. The structure of the organ of Corti (the active part of the cochlear partition) was first crudely drawn by Corti, in 1851, with the aid of improved microscopes and tissue fixation techniques. Further refinements were contributed by Politzer (1873), Retzius (1881), Ranvier (1875), Renaut (1899), Kolmer (1911), and Krause (1927). See von Bekesy (1960) for details and references. See Figures 1a and 1b for good modern pictures of the hearing mechanism in man. For more history (including Lucretius, Tartini, Ohm, Seeback, Helmholtz) and a general review of hearing models see Schroeder (1975).

During this time theories on hearing began to focus on the action of the hair cells, which were viewed at sensor cells with cilia that reacted to signals at various places along the cochlea. The question then was: what are the signals that these sensors receive, and how do they respond? The

simple and nearly correct early view was that different sounds stimulated different places on the organ of Corti, and that the distinguishing feature of these different sounds was their frequency. The perceptual concept of pitch was thought to relate directly to the physical *place*. This worked quite well for signals that were pure sine waves, but not well for speech and other complicated sounds. An alternative theory that developed was that pitch was perceived from the *volleys* of nerve impulses that resulted from stimulation with a bursty stimulus like voiced speech or a complex musical tone. Both of these theories were useful in understanding hearing at a beginning level, but the hard part was in combining them into a physicaly sound theory that would work correctly for all sorts of sounds. This required a better understanding of the mechanical behaviour of the cochlea.

The idea that the cochlea converts frequencies to places of maximum vibration on the cochlear partition had to be translated into a model of waves in the fluid filled spiral cavity. The main question there concerned the propagation properties of this type of "waveguide". Particularly, whether waves should be considered standing waves or travelling waves (travelling waves won, and are well described in von Bekesy, 1960). It took many years of continued experiments, and mathematical exploration, to home in on a fairly accurate characterization of the spatiotemporal response to general stimuli. But as we will see, there is still considerable controversy on several points.

The basic model is a fluid medium which is long, not very wide or tall, and has a stretchable membrane dividing it into two sections, with stiffness changing from very stiff near the input end (where the fluid in one side is driven through the oval wondow) to very stretchable near the terminal end (where finally there is an opening, the helocotrema, through to the other side of the fluid-filled cavity). The characteristic frequency (or CF, the frequency that causes the most displacement of a point on the membrane) changes geometrically by a factor of about 1000 (about 10 octaves). This model is represented by an electrical analog, using inductance for mass, capacitance for stretchable membranes, voltage for force, current for velocity, etc. The result is a tapered transmission line, with a discrete approximation as shown in Figure 2. The response can be studied by using this discrete-section lumped RLC approximation, or the transmission line can be solved for wave propagation characterisics directly by assuming no reflections; this is true if stretchiness changes slowly enough with distance, as in "the cochlear compromise" of Zweig (1975). The approximate correctness of this one-dimensional model requires the assumption that the wavelengths of waves moving through the cochlea are long compared to the height of the fluid channel, which is not a correct assumption if examined closely near the place of resonance for the wave.

To answer this objection, two-dimensional fluid models were examined by computer simulation techniques. This excellent work showed that good models can fit experimental data very well, while simpler models just come fairly close (Allen 1977). Figure 3 shows an idealized three-dimensional model of the cochlea, in which the width dimension is ignored to arrive at the two-dimensional

model.

A useful simple characterization of any of these models is the tuning curve, a graph of frequency response for a particular place on the basilar membrane (refered to as BM, the main part of the cochlear partition, holding the organ of Corti). There has not been much agreement on what the real tuning curves are, since different experiments give different results. The differences are mainly between the two classes mechanical tuning response and neural tuning response; the neural response, or internal representation of sounds, always seems to be much sharper than the measured or calculated mechanical resonances. The theories to explain the difference are based on two ideas: a second filter that sharpens the response by neural interactions (e.g. lateral inhibition), and sharpening due to the micromechanics of the organ of Corti (Hall, Allen, etc). This controversy is being actively pursued in recent literature. Another view has arisen recently, based on nonlinear modelling and experimental data; it is that the small-signal mechanical and neural responses are really the same, but that the experimental techniques used to measure mechanical response drive the cochlea into the nonlinear region, where selectivity is reduced (Kim 1975). If this is so, and if we can get by with modelling just the small-signal response, then the second filter *vs.* cochlear micromechanics question goes away.

If the acousto-mechanical model works, it describes the vibrations sensed by the hair cells; how do they respond to those vibrations, and what happens to their output signals? A main problem here is that the energy that is to be detected varies over 12 orders of magnitude (120 dB) from threshold of hearing to pain. The detector (or transducer) mechanism must compress this into something less than 2 orders of magnitude for representation as neural firing rates (about 10 to 1000 pulses per second). To accomodate this requirement, sensory organs are generally thought to have an overall "pseudo-log" response, which works very nicely for compressing the dynamic range of inputs. Even our telephone systems employ compression networks that are nearly logarithmic over about a 40 dB range of inputs.

The problem of describing in detail how the mechanical signals are converted to neural signals is known as the encoding problem. New analytical techniques have been used (Rose 1971, de Boer 1977) to show that the short-time response is very nearly like an ideal half-wave rectifier (linear relation of output signal amplitude to input signal amplitude), and that the longer-term compression characteristic is due to a fast adaptation process that changes the linear gain; the gain is modelled by a depletion mechanism (Schroeder 1975). The question of exactly which mechanical signal is being encoded is also very important; the tuning curves for position, velocity, pressure, and their spatial derivatives, are quite different in sharpness. Von Bekesy (1960) proposes that shear forces due to membrane bending are being detected, based on impedance matching considerations; Hall (1976) proposes that the first or second spatial derivative of displacement might be used; and Allen (1977) proposes a micromechanical mechanism for detecting a linear combination of pressure and displacement, giving tuning curves that match de Boer's data quite well.

When sounds have been converted to neural pulses by the fantastic mechanisms discussed above, the work has really just begun. As signals travel along nerves out of the cochlea and toward the brain, more processing is being done at various places along the way. The sites of processing are known as ganglia and nuclei, which are clusters of nerve cells outside of the central nervous system (think of them as micro-brains). The first, or most peripheral, such site is the spiral ganglion, which is on the end of the auditory nerve inside the cochlear spiral. Here the structure is analogous to the ganglia in the retina of the eye, where each ganglion cell gets inputs from several nearby sensor cells, and implements some kind of processing (second filter, or lateral inhibition?). The structure of such a network is shown in Figure 4, from observations of the retina; the ear may not have the same processing structure, but the general diagram should be similar, as in Figure 5 (Molnar 1968). In von Bekesy's collection of papers on sensory inhibition (von Bekesy 1967), he models such networks as simple linear filters in the space domain, whether for hearing, sight, feeling, or other senses.

More processing is done later, in places such as the AVCN and PVCN (antero- and postero-ventral cochlear nucleus). See Figure 6 for an overall picture of the neural processing in the auditory pathway. Only recently has progress been made in deciphering what goes on in the various nuclei and higher brain centers. The details are not yet available for this paper, but an important generalization is that many processing functions can be modelled as correlation computations. Early experiments on the visual-motor response of a beetle proved the effectiveness of autocorrelation as a model of low-level visual processing (Reichardt 1961, via Wooldridge 1963). We will use a similar model for processing the detector outputs to arrive at a representation of pitch and timbre.

Evolution has several clues to the relative importance of peripheral and central processing. Even low animals without evidence of "higher functions" can learn to recognize and respond to sensory patterns. The parrot evidently has all the intelligence it needs to remember and even repeat complicated sounds (much more versatile than just human-sounding speech). This is the reason we are confident that we can stop our modelling before getting into the complications of the cerebral cortex, the center of intelligence. But that still leaves us needing to model elementary memory and recognition processes. It is clear that the mechanism must be some kind of associative store, which responds to patterns by indicating whether the pattern has been stored before, and if so returning more information stored with the pattern (such as its name or meaning). A recent article (Menzel 1978) explores the organization of learning and memory in the tiny brain of a bee, which has less than a million neurons. See Sagan (1977) for a good discussion (including some very insightful speculation) on the structure of the brain and its relation to evolution.

FEATURES AND EFFECTS TO BE CONSIDERED

We cannot attempt to model all the intricacies of the hearing process, since we don't know what they all are; we should not attempt to make our model duplicate all experimentally observed phenomena, since some are subject to controversy, and many may be unimportant or detrimental to something like speech recognition. In this section we discuss some features of the hearing mechanism, and effects of processing, that we feel are worth considering for inclusion in the model.

Clearly, we can build a model which responds reasonably to sine-wave stimuli; the next step in trying to model response to more complicated stimuli is to study and model two-tone interactions. There have been many experiments and theories over the years to study the perception of pairs of tones with various amplitudes and frequencies. The results are diverse and fascinating; one might hear both tones separately, only one of the tones, a rough tone mixture, or the original tones plus one or more combination tones, depending on the various parameters. If we use a model which in a simple way accounts for most or all of these phenomena, then we can be confident that the model will respond reasonably to more complicated combinations of sounds. Some of the more subtle masking effects depend not only on the relative amplitudes of the tones, but also on their amplitude relative to the threshold of nonlinearities in the cochlea (one tone needs to be about 80 dB SPL to show clear nonlinear effects); we will avoid modelling these effects under the hope that the small signal response is more important to normal perception. For a good discussion of perception of complex musical sounds, including the concepts of harmony, consonance, and dissonance, see Roederer (1973) and Geldard (1972). For more on nonlinear two-tone interactions see Hall (1976), Kim (1975), and Schroeder (1975).

Nonlinearities may be needed in other places in the model, such as in detectors and limiters, compression curves, etc., but hopefully in simpler forms. The demodulation of complex waveforms into representations of timbre cannot be done by linear processes.

The question of phase sensitivity has bothered acousticians since they started trying to verify, quantify, or modify Ohm's Acoustic Law (that the ear is sensitive only to the amplitude spectrum). Schroeder (1975) has a good discussion of the issues and problems, but fails to note that by simply combining his models of cochlear mechanics and neural transduction, slightly phase-sensitive effects will be observed just as in psychoacoustic experiments. Thus, we expect to be able to model fairly accurately the slight phase sensitivity of the ear as a by-product of the other parts of the model.

Binaural effects are very important to normal mammalian hearing, since sound source localization has obviously important survival value. We do not intend to incorporate binaural effects into the model at this time, but we should examine models of binaural processing to get ideas about the kinds of processing mechanisms likely to be found in the nervous system. The important idea that emerges is cross-correlation processing (Roederer 1973 and Licklider 1959, or coincidence detection, which is similar, in Colburn 1977).

An intriguing mystery in hearing is the function of the outer hair cells. They receive efferent signals from the brain, and may influence the electrical, chemical or micromechanical properties of the organ of Corti. Is an understaning of this mechanism necessary to a successful emulation of the hearing process? We better hope not, but we can speculate that this efferent path is important to the focus of attention on one of several competing sound sources. That is, the ear may really be an adaptive filter, inside a giant feedback loop that directs the selection of a certain signal out of a field of interference. It would be nice to include this in our model, but we don't know enough yet, so we won't. Besides, in the lower animals there is no system of outer hair cells, and they are still able to remember and recognize patterns of sound.

It is important to implement other kinds of adaptive behaviour that we can understand. The overall model should be like a living organism in the sense of having enough feedback loops to be completely self-stabilizing.

GUIDING PRINCIPLES FOR THE MODEL

Converting our ideas about what the right processing steps are into a realizable model will require that we first resolve some fundamental questions about representation and methodology. In this section we develop such guiding principles, for application in the following sections on the model.

The representation problem is central to the modelling problem. The model will be built of digital logic elements, which will carry encoded representations of the signals of interest in the ear and brain. In the ear, signals are represented as small variations of pressure and velocity over time and space. In the brain, signals are represented by complicated patterns of discrete electrical pulses in time and space. Rather than modelling these representations directly, we should think about the underlying signal being conveyed, and select a representation to fit. Given the digital, technology, the natural representation for signals of all sorts is numerical; finite discrete numbers, at finite discrete places and times, can be used to represent arbitrarily complicated signals over time and multiple space dimensions, if certain sampling criteria are met.

Neural pulses represent signals in several ways, not just to confuse us, but because they are cleverly designed to make the most of the neural technology. Information is conveyed in the baseband component of the neural pulses (average rate of pulses represents signal intensity, roughly), but more information is conveyed in the detailed timing of the pulses (with a resolution of a few *micro*seconds in some cases, such as binaural localization). Does that mean we need to do similar tricks to capture all the information? No, if we sample fast enough in the first place, the various processing functions that need the information will be able to extract what they need and convert it to other forms. This means that the sample rate needed to encode neural information is considerably higher than the maximum pulse rate of a neuron. While this high-rate uniformly-sampled-data representation is not very efficient, it is easy to work with. Accordingly, we will adopt the sampled waveform approach to signal encoding. Another implication of this is that all signals

are analog in nature, and thus each sample must be represented by a numerical encoding with sufficient precision (a multiplicity of bits); that is, a hard (or single-bit) decision should not be used at any place within the processing to represent a signal.

The goal of the processing done by the ear and brain, and by our model, is to extract the interesting featural information from sound waves, for storage and later recognition. A clearly important principle is that we should attempt to extract as much as possible of the useful information, using the ear as our guide to what is useful, and to represent it in as few bits as possible without compromising quality. The signal output of the neural processing stage, which is the input to the memory, will consist of a low rate sampled feature vector (still consistent with discrete representation of intrinsically analog signals).

How fast do we perceive changes in sounds? Where is the boundary between hearing quickly changing sounds and hearing coarsely textured sounds (for example between beats and roughness in two-tone experiments)? The answer to these questions is vital to the model, since we must attempt to extract the roughness as a characterization of a steady sound, while representing the beats as slowly changing values of the feature vectors. For voiced speech, we clearly want pitch to be represented as a roughness, not as discrete bumps in the feature vectors. Thus the fastest signal to be represented in the feature vector will be somewhere below 60 Hz. All faster effects will have to be demodulated into baseband signals. We know that in vision we can accept a time sampling rate of 24 or 30 samples per second with little loss. We propose to use exactly a 60 Hz sample rate for this model of hearing (feature vectors limited to components below 30 Hz), in order to make it easy to view intermediate results on a television screen.

We should try to arrive at an ouput of feature vectors that is not only easy to handle, but also has a nicely structured space. Consider each element of the vector to be the coordinate along one dimension in a feature space. We want a space with good discriminability, or with dimensions not highly correlated. Literature on pattern classification provides some good guidelines, but not any effective algorithms for optimizing the choice of feature space. If the structure of the feature space is not too bad, the comparison problem in the associative memory is greatly simplified.

We must be careful to avoid using processing techniques that are badly behaved. We require robust algorithms, which will tolerate abuse and confusion of all kinds without giving spurious results. Thus many techniques used in speech recognition and other speech processing fields are immediately discarded--they fall apart under the influence of background noises like air conditioners and typewriters because they rely too much on the assumed structure of speech, which is supposed to be the only input. An important special case of a confusing input is perfect silence. We must be careful not to have something like a logarithm or reciprocal which will try to output an infinite value.

Most of us are familiar with the unpleasant reaction to a sudden loud sound (or any sudden sensory input which is much larger than the previous background stimulus). This is evidence that our

adaptation mechanisms can not cope perfectly with quick increases of signals over many orders of magnitude; but it happens rarely enough that we must admit that adaptation works amazingly well, and when we are caught by surprise the shock value may be important to our survival. Our sensory adaptation mechanisms seem to be optimally adjusted to help us notice similarities and steady state values, while simultaneously emphasizing differences and rates of change. We must strive to duplicate this performance in the model. This will require adaptation in all phases. However, to make the design and analysis problem tractable, we should cleanly separate the main processing functions from the adaptation functions. Thus, adaptation mechanisms are treated as a separate part of the model, even they are logically interspersed with the other parts.

Neural processing is full of nonlinearities. We could spend forever optimizing the nonlinearities for detector and adaptation curves, but we won't. We restrict ourselves to easily realizable nonlinearities of two classes. The first is a simple instaneous monadic operator (a memoryless function of one variable). The other is a bilinear gain-controlled amplifier (a multiplier); it is linear in either input with the other held constant. The instantaneous functions to be considered will be restricted to those implementable with simple arithmetic operations (not series), and possibly log and antilog. We will use simple approximations to whatever operations we want; if simple approximations won't work, the algorithm is too sensitive to be acceptable.

In the following sections, we propose digital signal processing structures to implement the various stages of processing that we have discussed.

MODEL PART 1--ACOUSTO-MECHANICAL FILTERING

If we start at the beginning, and follow sound information through all the processing operations that it encounters, we should study the properties of sound waves in air, in confined spaces as in the middle ear, in mechanical linkages, etc. But the important aspect of all this processing is that waves in air are effeciently transformed into waves in the cochlea; it is just a conversion of representations, with the usual impedance matching problem. This class of problems we will relegate to designers of microphones and preamplifiers. We will also let them have the problem of the overall frequency response between the air and the cochlear fluid, by incorporating a simple tone control in the preamplifier. The conversion of analog signals to digital discrete-time representation is the entrance to our model; for this we propose a wide dynamic range analog-to-digital converter (such as the 16-bit unit developed at CMU for speech research), running at a sample rate that depends on the desired fidelity and on cost (probably about 20 kHz). Thus the cochlea model will simply be an arithmetic processor that operates on a single fixed-rate stream of numbers (in real time, of course).

The cochlea takes a single input signal, and responds with different signals at a continuum of places along its length. We will simplify this to provide only a finite number of discrete outputs, each corresponding to a point on the cochlea. Between the input and any output there is a transfer function (or a more complicated input/output relation if nonlinearities are involved), and thus we

can look at this structure as a filterbank, a collection of different filters with a common input.

Actually, each of these filters is quite complicated, and they are highly interrelated; it makes sense, then, to actually make this filterbank from some shared structure that models the wave propagation in the cochlea. We are faced with the decision of what model, at what level of complexity, to adopt and implement. We will start with the two-dimensional fluid model, draw the equivalent RLC circuit for the long-wave approximation, and assume linearity (small signal model) for simplicity. It would be interesting to actually implement a more complicated version at a later time, to see if it is any more useful than the simplified model.

The RLC transmission line model has a discrete number of outputs (say five per octave over six octaves in a low-budget version), but it uses continuous-time signals and components. To convert this to a digital structure, we need to study the literature on wave-digital filters, structures which can directly model waves, impedances, etc. These structures map voltage/current combinations that represent incident and reflected waves into two-way digital wave ports. But are reflected waves needed in a cochlear model? Zweig (1975) and others say no; the response of the cochlea can be matched almost exactly by assuming no reflected waves at all; the cochlear compromise is inherited by the cochlear model. Thus we can use a simpler structure, a simple cascade of filters (this is like buffering the stages in the RLC model to prevent any backward signal flow).

The buffered RLC model and its digital filter equivalent are shown in Figure 7, along with typical transmission pole and zero locations. To determine the coefficients for the digital filters, we need to calculate the transmission poles and zeroes of the analog structure in the s-plane, map them by a simple transformation into the z-plane, and convert these to coefficients by standard methods. For reasonable pole and zero position accuracy, we should use coefficients of about 16 bits, and data words of about 22 bits. These word length requirements can be reduced some if we use another second-order form such as coupled first-order sections (Gold 1969), instead of the canonic form.

The detector output of this filter section (analogous to the displacement or velocity of the BM at a point) is not the same as the signal that is fed forward to the next stage (analogous to pressure in the cochlear fluid); rather, the detector output is the signal that comes from applying the poles due to the series resonance, which are close to the zeroes of the transmission gain. The result is that the filter response to the detector is highly tuned, and that the response to the next section is low-pass, with each section lowering the cutoff. Thus waves propagate through filter sections to the point of maximum resonance, then are quickly attenuated, just as in the cochlea (von Bekesy 1960).

The approximate tuning curves for this model can be found in Zweig (1975), since he did a similar transformation of the model by assuming no reflected waves in his solution approximation technique. See Figure 8 for Zweig's tuning curves, compared to measured data. Of course, the parameters can be varied to make the tuning more or less sharp. If we only use five sections per octave, the tuning sharpness should be reduced to avoid spurious results due to our limited frequency resolution. The resonant frequencies of the sections will be chosen to closely

approximate the relative frequency selectivity of the ear; the result is a geometric progression from about 500 Hz to 4000 Hz, and somewhat sparser above and below that.

The shape of the skirts on the tuning curve will depend on the density of sections per octave, among other things. If not enough sections are used, the first thing to suffer will be the steepness of these skirts; this is especially true of the high side skirt, which counts on the cumulative attenuation of all the previous sections.

The acousto-mechanical part of the model is now complete, and as can be seen, there are only two components required to implement it: memory and multiplier/adders. Both of these types of components can be built as regular arrays, suitable for silicon VLSI embodiment. In fact, such components already exist in MSI and LSI; digital filters are routinely built with boards full of them. In summary, the acousto-mechanical filter part of the model consists of a simple cascade of about 30 or so second-order pole-zero filter sections, with output taps from within each section.

MODEL PART 2--TRANSDUCERS OR DETECTORS

In part 1 of the model we mentioned that each section of the transmission line model has a separate ouptut to go to the detectors. Does the detector look at just one such output, or does it look at the difference or second difference of two or three adjacent outputs (the spatial derivative approach), or does it look at a linear combination of an output and a delayed or phase shifted version of the same (similar to the Allen approach)? We might postpone the decision and preserve the generality of all these approaches by letting each detector see a linear combination of three adjacent outputs and one or more delayed replicas of each. Then we would have to decide what linear combination of six or so inputs we really want. That's too much generality to handle, so first observe that the output of a section is to first approximation a delayed version of the previous output, but filtered somewhat due to the difference in resonant frequencies. So the question of whether to look at signals adjacent in time *vs.* in space is probably not important. Since adjacent signals in space are already available, use them; for simplicity just use two (call them X_i and X_{i+1}), but for a little remaining generality leave the linear combination unspecified, as in

$$Y_i = X_i + k \cdot X_{i+1}.$$

Different values of k will give different amounts of sharpening or broadening in the response of the signal Y_i.

Next we need the actual detector nonlinearity. A good power detector is a square-law nonlinearity, as used in photocells, resistor-thermistor power meters, etc. One might speculate that the square-law heatup of the lossy elements in the RLC model could be detected; but the amounts of power involved in the ear are so low that the temperature increase would not be perceivable. Power or full-wave envelope detection is not really the right thing to do. As mentioned before, the nonlinearity is known to be nearly an ideal half-wave rectifier, based on analyses of signals on single nerve fibers. Anyone who has played with a model of the cochlea should appreciate the reason for the half-wave nature. If we look at a smoothed version of a detector output, when the stimulus is a pulse train or voiced speech, we see mostly the envelope detected as "lumps" at the pitch or pulse frequency (see the wideband sonogram in Figure 9, for example); if we look at the output when the stimulus is a pure tone with the same pitch, we will see similar "lumps" (but coming mostly from the low frequency places on the cochlear model). If, on the other hand, we had a square-law or absolute value (full-wave) detector, the apparent pitch of the pure tone would have doubled, making it sound (or look) very unlike the pulse train with the same pitch. Since the ear uses half-wave detection, we can see that pitch is simply translated into periodic patterns after the detectors, relatively independent of place (for low enough pitch). An interesting effect is to have complex tones which are in separate frequency (place) regions, so they don't interfere much, but with different periodicity pitches. The model will clearly show different places responding at different pitches, which are completely unrelated to those places. We have observed this effect in speech (see Figure 10 for an example with two apparent pitches), which may explain why unique pitch determination is such a hard problem.

A reasonable alternative for the detection nonlinearity is a "soft" half-wave rectifier, since in nature it is unusual to find a sharp break as in the ideal half-wave rectifier. Schroeder uses the relation:

$$Z = Y + (Y^2+1)^{.5},$$

which approaches $2 \cdot Y^+$ (twice the positive part of Y) for large absolute values of Y. Near Y = 0, however, this function makes a smooth transition from no response to positive response, without a sharp corner. See Figure 11 for the characteristics of this nonlinearity. The offset value of Z = 1 at Y = 0 corresponds to the spontaneous firing rate of the detector neuron; when the hair cell is stimulated, the firing rate increases during positive half cycles, and decreases to near zero during negative half cycles. The result is that signals with amplitude less than about unity contribute much less average firing rate than they would with a sharp rectifer characteristic. This is qualitatively consistent with the knee in the subjective loudness curve (sones *vs.* phons) at about 20 dB SPL (van Bergeijk 1960). There are problems in scaling unity in this nonlinearity to the signal amplitude in the model, and in deciding whether such scaling should be affected by adaptation; these problems don't exist with the sharp nonlinearity. Therefore, without excluding further improvements, we will stay with the ideal half-wave rectifier for now.

The output of the detectors is a good place to do the first step of sample rate reduction. This is ideally a null information processing operation (a no-op), but actually requires some loss of information due to bandwidth reduction. We will apply a simple smoothing (lowpass) filter to the detector outputs, and resample at about 2000 Hz (an order of magnitude reduction is possible here because most of the information in the fast sampled waveform has been demodulated to baseband or envelope information by the detectors). This new sample rate is high enough to encode most of the information that can be conveyed by a pulse train limited to about 1000 pulses per second, so it should adequately handle the representation of neural signals.

We must be very careful about sample rates and bandwidths. Even if we have sampled at 20 kHz to capture sound waves up to 10 kHz, we can get into trouble with aliasing at the detector. An instantaneous nonlinearity will change a band-limited signal into a much wider-band signal in the continuous-time domain, but will cause strange results due to aliasing in the discrete-time domain. Particularly, observe that any detector will have a strong second-order component, which will produce both DC (representing envelope) and a double frequency term from the sinusoidal or bandpass input waveform. Therefore, a signal frequency of 5000+f Hz will produce a spurious output at 10000-2f Hz. For example, a 9000 Hz channel will have to cope with spurious signals around 2000 Hz coming out of the detector, which is especially a problem if such signals are not filtered out before resampling at a lower rate; resampling at 2000 Hz aliases signals in this area right into the baseband, for maximum confusion.

A good smoothing filter is a cascade of a few first-order (RC type) lowpass sections, plus a structure that puts transmission zeroes at all places that would alias to DC. A simple sum of N adjacent samples gives the required N-1 zeroes, where N is the sample rate reduction factor (N = 10 in our example). The first-order filters are similarly simple, but require multipliers. The cutoff at 1000 Hz

should not be sharp, but should be a gradual rolloff from about 100 Hz, so that as pitch increases, the volley representation decreases gradually toward zero. Higher pitches than 1000 Hz will be represented only as places of maximum envelope response. The human ear may actually have some volley timing information at considerably higher pitches, with some neurons sensitive to timing differences on the order of a few microseconds, but the effects are subtle, and not relevant to speech. If we want to accomodate these effects, we might need faster sampling and broader lowpass filters. Nature's neural pulse train representation is used very efficiently in this respect, to carry not only envelope information (in the firing rate), but also phase and frequency information (in the detailed timing of individual pulse positions).

Adaptation is covered later, but we should say a few words here about the depletion model of transducer adaptation (Schroeder 1975). Hair cells cause primary auditory neurons to fire by manufacturing and delivering to them a chemical transmitter; the rate of stimulation is limited by a relative shortage of this commodity. In addition, there is a finite storage area for this commodity, and a decay rate. Figure 12 is an electrical analog of this mechanism, where the conductance g(t) is the rectified signal from the transducer (like the membrane conductance caused by bending the cilia), the battery and resistor are the production and decay mechanism, and the capacitor is the storage (this model is like Schroeder's except that he used a Norton equivalent, with a current source instead of a battery, which was better for the production-depletion explanation but did not generalize the way we wanted it to for the discussion below). The output is the current through the conductance g(t); this is just the detected signal multiplied by a gain proportional to the voltage on the storage capacitor. The product RC is a recovery time constant: the time to fill to within 1/e of the distance to the steady-state capacity, where production matches decay. The attack time constant, the time required to deplete the storage enough that a very large stimulation will deliver only a reasonable level of transmitter, is short but variable. Essentially, with a very large stimulation, the whole store will be transmitted at once, causing a barrage of nerve firings which quickly decays to the saturation level, that which is just supported by the production of transmitter. The response of the adaptive transducer (with Schroeder's soft nonlinearity) to tone bursts of various levels is shown in Figure 13.

This mechanism has to be examined in several ways. It has short-term (waveform distortion), medium-term (adaptation to attack), and long-term (recovery) properties. It has small signal (linear) and large signal (limiting) behaviour, and a crossover region. But more importantly, it has spatial characteristics, relative to the other detector channels. If each saturates independently, there will be a tremendous flattening of response with increasing signal level. Some kind of inhibitory action is needed to reduce this effect, or to reverse it to give sharpening instead of flattening. One way to do this is to couple the production-depletion mechanisms so that adjacent channels compete for the transmitter commodity. Such competition is modelled by replacing the batteries in the model with coupled RC storage mechanisms much like the original (with higher capacities, lower average saturation levels, and slower recovery times). This distributed RC structure might represent the input of some primary energy commodity (such as ATP) by diffusion from the nearby fluids or

blood vessels. It might be preceded by yet another similar stage, even more tightly coupled or completely shared, to represent the electrical potential in the scala media, which also affects the gain of the transducers. For more description of the relevant structures, see the section on adaptation.

The effect of this adaptation mechanism removes the need for a "pseudo-log" nonlinearity, since the response is linear for very low signals and stabilizes with very high signals. Still, it is important to define where the crossover region from small to large signal is for each layer of the adaptation mechanism. This will probably have to be set by some simulation and experimentation, but should not be critical.

In summary, the model for each channel of the transducer array is a cascade of four simple steps, as follows:

1. A linear combination of two adjacent filterbank outputs.
2. An ideal half-wave rectifer.
3. A controlled-gain amplifier, controlled by the output of the adaptation model.
4. A smoothing filter and sample rate reduction.

The outputs should only need a dynamic range of about 2 to 3 orders of magnitude (40 to 60 dB), which can be accomodated in seven to ten bits. Since very large transients may occur at the onset of a sound, some overflow protection limiting should be included. Notice that all signal values after the rectifiers are strictly positive, potentially simplifying the arithmetic units required.

MODEL PART 3--PERIPHERAL NEURAL PROCESSING

Think of this part of the model as an array processor. It receives a set of several dozen inputs, several thousand times per second, and has to extract and output a smaller set of numbers, at a lower rate. It decomposes into two main sections, the first operating on the inputs as separate streams, expanding them into even more streams but reducing the rates, and the second linearly combining these streams into a smaller number of streams, with a better structured space. The actual neural processing in mammalian hearing mechanisms is probably not so cleanly divided, but we need to restrict ourselves to something we know how to do.

The problem is similar to the original waveform analysis problem; the fine structure and periodicities in time of the input waveforms have to be converted to patterns in space, and represented at a much lower rate (say 60 Hz, more than an order of magnitude reduction). Should we apply another filterbank analysis and extract the spectra of these signals? There is no known mechanism that would support such an analysis in a neural network; but we can arrive at another form of spectral information by computing an autocorrelation estimate of the signal. This is something neurons can do.

How can neurons compute autocorrelations? Figure 14 shows a signal processing structure that computes smoothed products of a signal times various delayed versions of itself; these are autocorrelation estimates parameterized by the delay τ. Neurons can easily implement the required delays by path length differences. The multipliers are also naturals for neurons, if they don't have to be particularly accurate. Suppose that a neuron fires only if it is stimulated nearly simultaneously from two different inputs; then it forms a coincidence detector, or AND-gate, which is a fair approximation to a multiplier (or exactly a one-bit digital multiplier). The real response will be smoother than that, so a multiplier-like function is not unlikely. Even a crude approximation is adequate to give roughly the same results as a good autocorrelation analysis if the signals are dithered by noise.

Consider the response of this structure to the periodic pulses that come from the detectors when the stimulus is a low-pitch tone or pulse train. As the detected pulses travel down the delay line, they come to places where they line up with new pulses, resulting in a maximum product. Thus the autocorrelation function has a peak that identifies the pitch period with the delay τ. In general, the shape of the autocorrelation function is related to the timbre of the sound.

We need to decide what values of τ to implement; both the resolution and the maximum range are important. Suppose the total delay (maximum τ) is 16 msec, or 32 samples at 2 kHz; this corresponds to a full cycle of a 60 Hz pitch, or a half cycle of a 30 Hz pitch, so the pitch resolving ability will roll off between 30 and 60 Hz, which is about what we want. If we implement taps after every .5 msec delay, we will expand each channel into 32 channels, which is too many, especially considering the relative invalue of pitch to speech perception. Instead, use taps that are geometrically spaced, with a factor of two separation. A convenient set of values would be .5, 1, 2, . . ., 16 msec, for a total of only six taps (and perhaps more at 0 and 32 msec to make eight). These cover full cycles of pitches from 60 to 1000 Hz.

Just picking off the desired taps from a full delay line has two big problems. First, it uses too much memory (32 taps times 32 frequency channels is 1024 words of several bits each, which might fill up a whole chip). Second, and more important, is the fact that the resulting autocorrelation estimates do not include much information about the intermediate taps, making the estimates for large values of delay depend too much on the short-time structure of the signal. Both of these problems can be alleviated by using a lowpass filter cascade instead of a simple delay cascade. The filters can be fairly trivial; they simply add two successive values and cut the sample rate by two. Thus the delay of each stage doubles as its sample rate decreases, and the bandwidth is halved at the same time. Such a structure, with exponentially decreasing sample rate, can be built with a word of memory per stage and a single adder, for any number of stages. The outputs will be multiplied by the input at the full 2000 Hz sample rate, and the result will be averages of autocorrelation coefficients, over various output taps of the original structure.

The smoothing lowpass filters in the autocorrelation estimator will reduce the signal bandwidth to

less than 30 Hz, for resampling at 60 samples per second. The result is a slowly changing plane of values, with axes of frequency and delay time. A display of these values as blackness (or brightness) on the two-dimensional surface of a CRT display in real time would be an impressive dynamic representation of sounds. This is quite unlike a sonogram, which parameterizes in terms of frequency only, and is plotted against time on a fairly fine scale along the other dimension; it changes too fast to present the dynamics in real time. Our representation is much more closely analogous to vision, in that it produces patterns in a plane, changing only slowly in time. Thus we can look at it easily.

The array of point streams thus generated may be further filtered to enhance edges or smooth over noises, in time and two space dimensions. In light of the transformation operation in the second part of the neural processing model (below), spatial filtering becomes irrelevant (it is incorporated in the transformation matrix, implicitly). Time filtering should not be needed if the adaptation mechanism has contributed the desired edge enhancement effects. We leave open for now the option to put in some kind of filter at this stage.

In the second part of the neural processing model, the several hundred values generated as discussed above are combined into a smaller number of feature vectors by a memoryless linear transformation. The motivation for this operation is to extract a few uncorrelated features from a "picture" that has high correlation between adjacent elements. Each such feature is a weighted sum (and difference) of any or all of the elements of the picture. Neurons could implement such a function by integrating excitory and inhibitory inputs from many different places in the "auditory plane".

Early speech recognition systems (and more recent ones such as marketed by Threshold Technology Inc.) used such a model of neural processing to extract features from a bandpass filterbank model of the ear. Unfortunately, they have typically used only a single bit output; the resulting analog-to-digital feature extractors are called "analog threshold logic."

How can we determine a good transformation matrix to be used to map from the input space to the output space? If we have a collection of sounds that we want to be able to discriminate among, we have a classical problem in pattern classification; there are methods of computing the transformation to maximize discriminability in the output space (given enough data and computing time), along with other useful techniques and theorems. Since we do not know what sounds we want to discriminate, we need another approach.

Suppose that for each output component we choose coefficients at random, independently for each element of the input space. Maybe that is how our neural networks get built. The resulting outputs will likely have low correlation, but may not capture much of the useful information unless there are many of them. We still need another approach.

The approach we recommend is to pick a set of smooth orthogonal functions, and use them. A natural set is sines and cosines of the two dimensions. Disregard the delay time dimension for a moment, and consider the inputs as just a graph of compressed intensity *vs.* frequency; then transforming with sinusoids is like performing an inverse Fourier transform, similar to that used in cepstral analysis, but with a distorted frequency scale. Such a mel-based cepstral analysis method of feature extraction was recently used in an experimental speech recognition system (Mermelstein 1978).

Another justification for the use of sinusoids in the log(f) domain comes from a paper entitled "Correlation and Dimensionality of Speech Spectra" (Li 1969), in which techniques of pattern classification are applied to sampled vowel spectra to arrive at estimates of the eigenfunctions, or principle components. Plots of these eigenfunction estimates are somewhat noisy and irregular, but strongly resemble cosines of log(f).

The inclusion of the autocorrelation delay time axis complicates the picture, but it is still reasonable to think that most of the important information will be captured by transforming with a set of smooth functions of the two dimensions. See Figure 15 for an example of the receptive field of a neuron with this kind of two-dimensional excitory/inhibitory input network (from Swigert 1971). How many such "eigenfunctions" should we use? We propose to use sixteen, to give sixteen nearly independent featural dimensions.

In summary, the peripheral neural processing part of the model consists of these steps:

1. An autocorrelation processor on every frequency input.
2. Low-pass smoothing and sample-rate reduction.
3. An *optional* filter in time and two space dimensions.
4. A linear transformation to a length-sixteen feature vector.

MODEL PART 4--LEARNING, MEMORY, AND RECOGNITION

Suppose some sequences of the feature vectors generated above represent important sounds that we want to remember, and that we have therefore stored these sequences in a memory. The bit number, coordinate number, and time index all map into the address space of the memory in some way, but we will treat each vector as a word of storage, and simply say that a time index is mapped into a word address. Addresses are really unimportant, except to tie feature vectors together in an ordered sequence, from beginning to end, for each sound unit (template) stored.

When the "ear", or the analysis part of our hearing model, delivers something interesting to the "brain", or memory, it is recognized if a similar sequence of vectors has been previously stored. It should not be necessary to identify the beginning and end of a potentially interesting sequence, and search through the memory for it; each pattern in memory should match itself against what is coming in, continuously, in real time. Thus the memory is intrinsically associative; that is, stored values are not referenced by address, but by content matching. The memory sends back responses like "That sounds like Bach," not just "I've heard that before." See Menzel (1978) for more discussion of associative memory, in the context of visual pattern recognition in bees.

There are several problems in implementing this associative memory concept. First, what do we mean by similar? This requires the definiton of a distance measure between vectors, and a way of composing distances over sequences of vectors. Second, how are interesting sequences selected for inclusion in the memory, and how is the memory updated? Both of these issues are discussed briefly in this section.

The transformation to a sixteen-dimensional feature space in the neural process model above was designed to result in low correlation between dimensions. As a result, the Euclidean distance (square-root of sum of squares of differences of components) is a reasonable way to measure the similarity of feature vectors. Other distance measures will not be considered.

To compose distances over sequences, let us use the *sum-of-squares* of the individual distances between corresponding vectors in those sequences. Thus, the distance between sequences is the sum over corresponding time indices, of the sum over sixteen dimensions, of the squared difference of components. (An intermediate square root and square cancel each other.) Other ways of composing distances will not be considered, either.

Now we really have a problem. How do we know what time index of a stored sequence corresponds to what time index of the current unknown input? We don't. The comparison mechanism associated with each stored pattern continually checks to see if a recent segment of the unknown input, up to the current time, matches the stored sequence up to its end. That would be easy if sequences always came at the same rate, with a one-to-one correspondence of time indices between unknown and stored template. Figure 16 shows how it would be done (since things

happen at such a low rate, much of the logic would be shared).

The problem of matching sequences is much more complicated if the sequences are generated by processes that can change their rate, such as the human speech production process. The composition of distances should be redefined to be the lowest distance obtainable by trying all possibilities in an allowable class of time-warping index transformations. The result is the distance to the most similar interpretation of the unknown that could be found, and maybe information about how it was warped to get that match.

Checking all allowable correspondences of time indices leads to a combinatorial explosion of computation; but the principle of optimality can be applied to reduce the problem to a dynamic programming problem, in which computation grows only as the length of the sequence being matched. The computations required to implement a dynamic programming solution to this problem are vector distance calculation, selection of minimum, and addition, all at the rate of M per sample interval, where M is the length of the template. The algorithm and some variations are discussed in detail in Sakoe (1978). The data flow and memory requirements are also simple, but are not presently included here due to time limitations.

The question of how to decide what patterns to put in the memory is also difficult. Suppose we simplify the problem by considering a simple isolated word recognition system for a single user. The templates would be just the words in the vocabulary to be recognized, as spoken by the intended user. The problem then is to find a good training strategy to enter representative templates and to update them as the user's pronounciation changes. This and other high-level strategy-related functions are beyond the scope of this paper.

To use the same associative memory in unrestricted vocabulary, multi-user, continuous speech recognition, we might try loading it with many variations of short phonetic units, such as syllables, phonemes, or transemes. The decision of what units to use is probably of central importance, but is also beyond the scope of this paper. From the resulting distances to phonetic units, one could compute likely words and sentences by searching through a finite-state model of the language, as has been done in the IBM continuous speech understanding project (Bahl 1978).

In summary, the learning, memory, and recognition part of the model is an associative store, with a method of entering new templates from the analyzed input, and a method of concurrent pattern matching, with the time-warp search complexity limited by the use a dynamic programming algorithm. The details are not presented, but digital signal processing concepts should be applied to assure reasonable range and resolution of the numbers involved. The memory requirements for each template of length M (which is variable) are M "big words" for the M stored vectors, and M "small words" for the intermediate results of the dynamic programming (a typical half-second word has M = 30).

MODEL PART 5--ADAPTATION MECHANISMS

The goal of adaptation is make the featural outputs robust with respect to variations is loudness, rate, room acoustics, microphone orientation, etc., but not to completely remove all dependence on these conditions. For example, is we decide to "normalize" the data at some point in the model to remove all dependence on loudness, we will boost the significance of the noise floor and not be able to distinguish it from a good signal; this must be avoided.

Ideally, an intelligent device with a good idea of how the model works would monitor various signals and continuously readjust the model parameters to keep it tuned to the conditions or to the signal of interest. This is a much too difficult and general approach. Instead, we will look at methods of locally stabilizing the signals at various places in the model, without much idea of what features are of interest. The function of the outer hair cells in the organ of Corti may be of the former type, and therefore outside the scope of this model.

The general philosophy of local adaptation should be applied in all the representation spaces of the model, from waveform input to feature vector output. At each stage, the range of adaptation should be fairly small compared to the total; the rates of adaptation at the various stages should be chosen relative to the characteristics of the local signals.

The adaptation machanisms that we currently intend to include are the following:

1. An overall AGC (automatic gain control), possibly before the analog-to-digital converter, to reduce the dynamic range of the signal by about 20 dB (factor of ten gain change). The adaptation should be gradual (less than two-to-one compression on a log scale) and its rate should be slow compared to the characteristic time (or propagation delay) of the cochlear model, so that it doesn't distort waveforms at all, or envelopes very much.
2. A gain-controlled amplifier at the transducers, controlled by a resource-depletion model, which itself has two or three layers (it doesn't matter if it is before or after the transducer nonlinearity, as long as that nonlinearity is an ideal half-wave rectifier). This can be partitioned as follows:

2a. An overall gain control will model the effect of the electrical potential in the scala media. This can adapt more quickly than the input AGC, since gain changes after the frequency analyzer will not cause distortion products. The model for the potential is a first order filter with inputs from a reference constant and from all the detected signals summed together.

2b. A per-channel "battery", a row of coupled first-order filters that model the diffusion of an enery source (ATP) into the hair cells, where it is used to produce the chemical transmitter. The coupling is a resistive type, which accurately models diffusion into the storage (capacitors).

2c. The "battery" recharges the store of transmitter, which is itself stored in a first-order

linear depletion model (another leaky capacitor) per channel, but without coupling--each hair cell saves all the transmitter it makes for itself.

3. Deemphasis of the slowly changing part of the feature vectors by time filtering them may be included later. This is considered to be an adaptation, so it was not mentioned previously under neural processing, where it could fit.

See Figure 17 for the method of interconnection of the various adaptation mechanisms. Each mechanism has its own way of limiting the peaks of large signals, and of changing their transient shape. For example, the transducers limit by emptying their stores of transmitter, very quickly; they recharge fairly quickly, too (about 100 msec). But how far they recharge is limited by the surrounding activity, within a few diffusion lengths, by depletion of the energy source, which recovers more slowly. It is important that signals are passed through the fast limting adaptation as soon as possible after the frequency analyzer, to transform between the huge dynamic range of mechanical signals to the limited dynamic range of neural signals. Hopefully, we will succeed in doing this while treating the main signal path as a well behaved signal, and adding the adaptation on the side to control the gains, to make the signals behave.

There are an incredible variety of possible automatic gain control types of adaptation designs, distinguished by the type of filter and nonlinearity (if any) used to estimate the signal strength, the type of compression nonlinearity used (if any), the form and parameters of the feedback filter, and the nonlinearity of the controlled gain element. The depletion model is particularly simple, because the signal to be detected is unfiltered, it needs no nonlinearity since it is already a strength, the feedback filter is a leaky integrator (first-order), and the controlled gain element is bilinear. Other variations on the design will certainly have their own advantages, but their appreciation will have to wait for some analysis and experimentation.

IMPLEMENTATION CONSIDERATIONS

Where is the meat of the hearing model? It is in the detailed implementation design, and the selected parameters, which have not yet been done. From the system design presented in this paper, it should be possible to design the details of a hardware or software implementation of the entire model, and then to experiment with the parameters. All the parts of the model are easy to design and build; an important goal is to design them in a consistent design methodology so that they will all fit together into a coherent digital system. A consistent method of interfacing the pieces at the levels of bits, words, frames, etc. is necessary before the detailed design can go very far. We are working on this aspect of the design now.

Currently, the amount of circuitry we can put on a single large silicon chip is enough to do perhaps the entire first part of the model. But we must design the system in smaller pieces, always keeping in mind that they should be *inward compatible*, so that by the time we have finished the design, and tested the pieces, we can take advantage of the newest technology and pull the whole system inward onto a single chip.

Obviously, there is some good sense is building a simplified version of the model during the development stages. Just building the first three parts, without adaptation, might be enough to discover powerful new applications for real-time sound display.

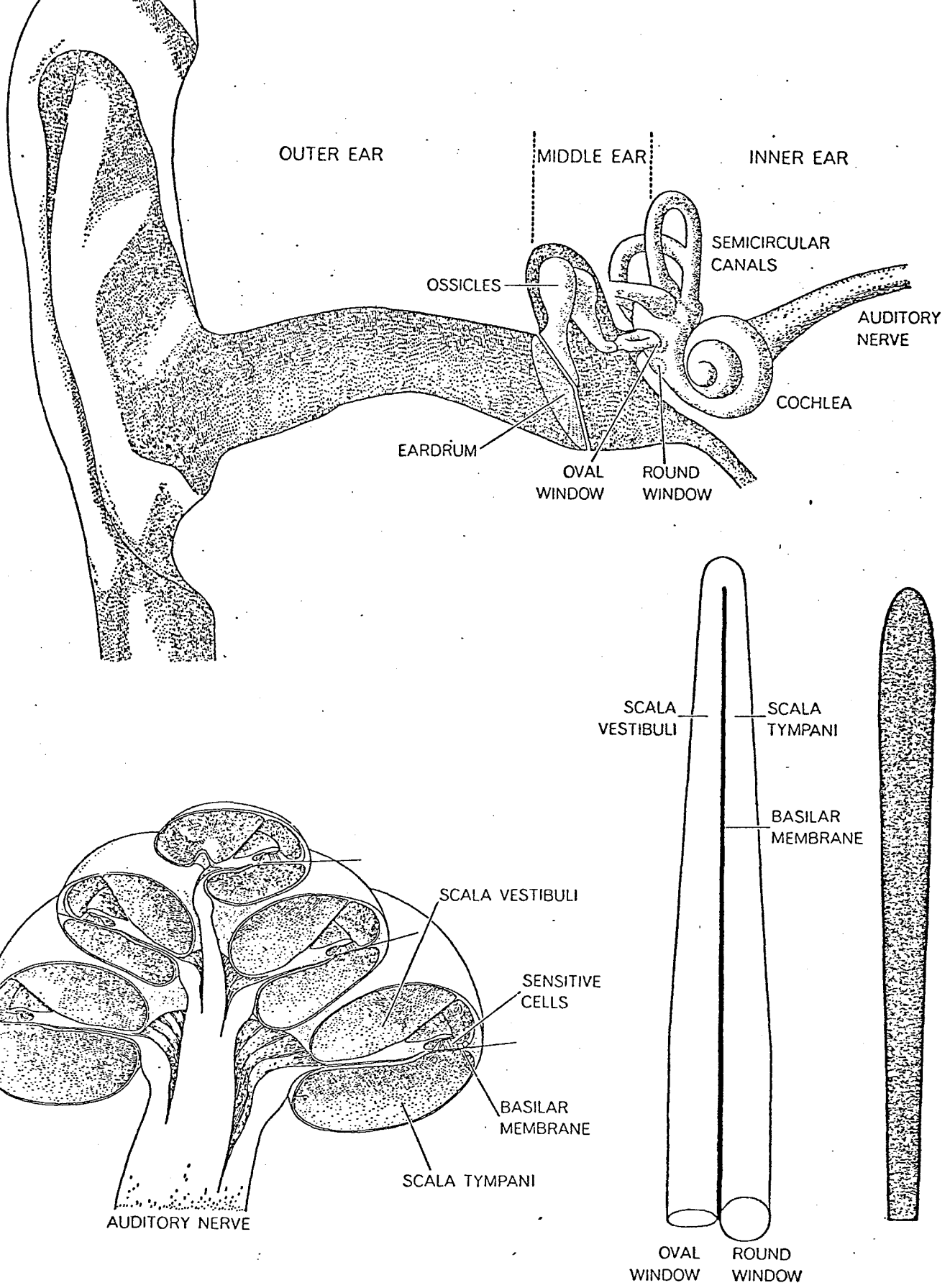
DISCUSSION, SUMMARY AND CONCLUSIONS

We already know most of what we need to know about the ear to build a much better signal-processing model of it than has ever been built. Proven techniques from the fields of pattern classification and word recognition can be applied to this new representation to get performance which is probably much better than that attainable in the past on recognition tasks. As new information on how the nervous system processes neural auditory information becomes available, that information can be used to refine the model. Many parameters of the model will be soft, and can be improved as we learn what is right, without any redesign.

In summary, the model, shown in block diagram form in Figure 18, consists of digital signal processing structures to perform the following functions:

1. Model the distributed resonance of the cochlea as a filter cascade.
2. Model the action of the hair-cells in detecting the vibrations of the cochlea.
3. Model peripheral neural processing as autocorrelation and feature extraction.
4. Model storage and associative matching of sounds using dynamic programming.
5. Model the adaptive mechanisms that affect all stages.

The value of most hearing models in the past has been in explaining and exploring the workings of the ear and brain. The value of the signal-processing model is in emulating these workings, so we can make a machine that hears the way we do. Such a machine has the potential to be very useful if implemented in real-time hardware.



PERCEPTION OF SPEECH begins in the ear, shown at top in simplified cross section. The eardrum transmits sound vibrations to the three small bones called ossicles, which cause waves in fluid in the cochlea. The cochlea, seen in cross section at bottom left, contains the basilar membrane (*color*), on which rest the sensitive cells that excite auditory-nerve fibers. At bottom center cochlea is rolled out, with basilar membrane in side view. Front view of the basilar membrane (*bottom right*) shows that it is wider at one end than the other. The wide region vibrates in response to low frequencies, whereas the narrow region responds to high frequencies.

Figure 1a. Drawings of the human hearing apparatus, from Broadbent, 1962.

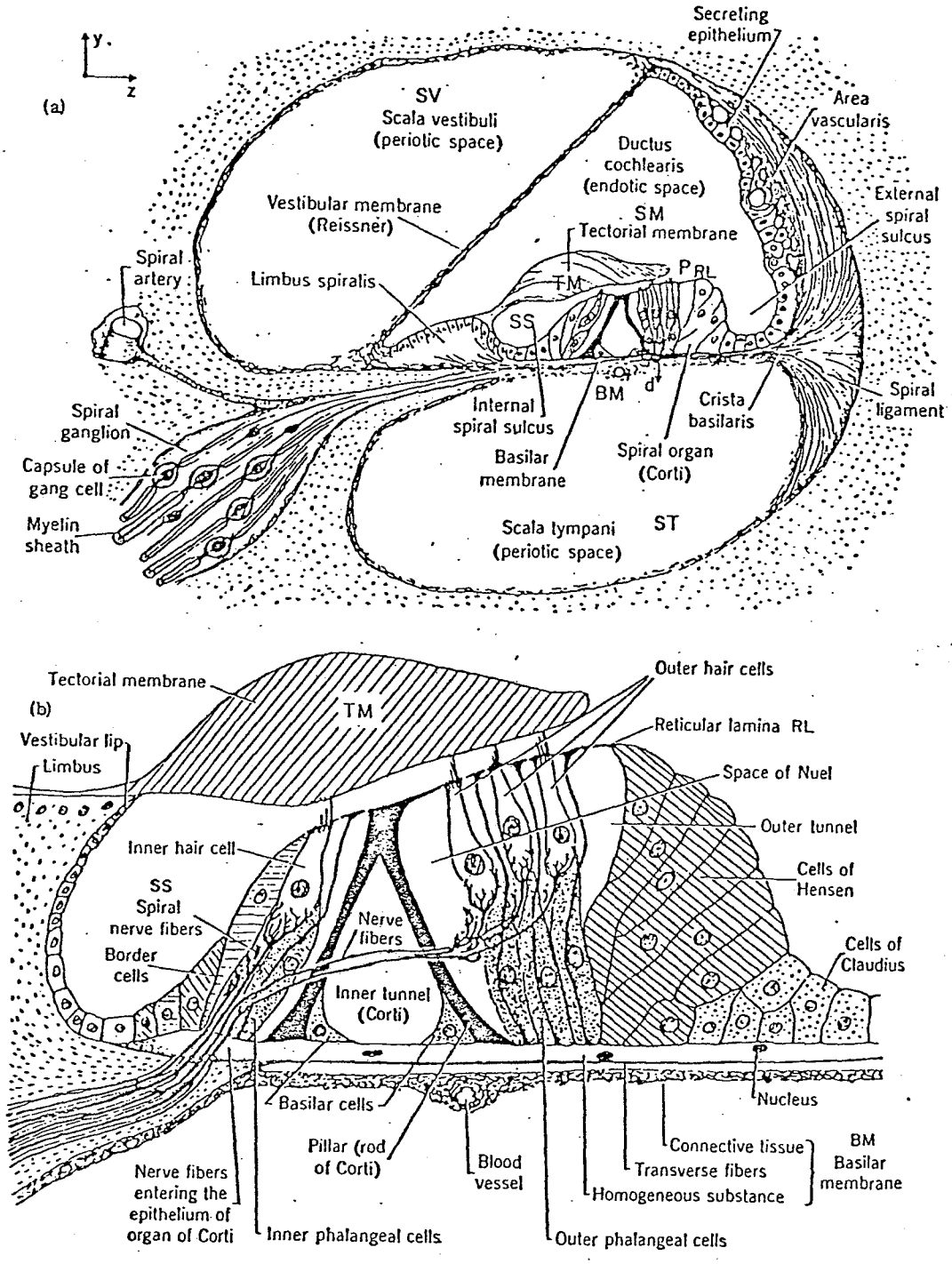


FIG. 5. , (a) Drawing of a cross section of the cochlear canal due to Rasmussen. p_{RL} is the pressure at the reticular lamina while d represents the displacement of the basilar membrane. (b) A detailed view of the subtectorial region.

Figure 1b. More drawings of the human hearing apparatus, from Allen, 1977.

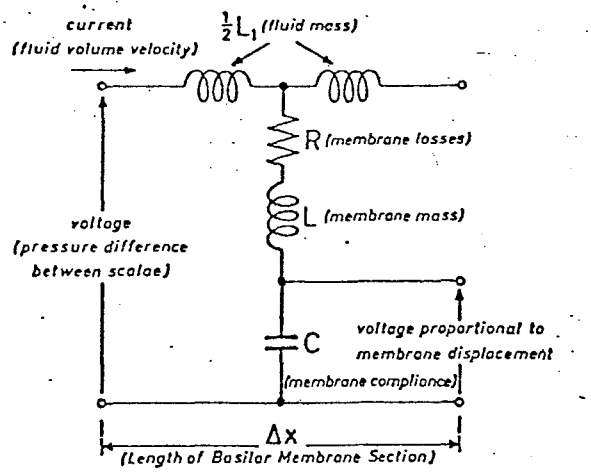


Fig. 9. Electrical analog circuit for short section of basilar membrane ("long-wave model"). For low frequencies ($\omega << 1/\sqrt{LC}$), a given section acts essentially as delay line. For high frequencies ($\omega >> 1/\sqrt{LC}$), circuit represents (inductive) attenuator. For $\omega = 1/\sqrt{LC}$, transverse current (corresponding to basilar membrane velocity) has resonance peak. Different frequencies peak at different places along membrane.

Figure 2. An RLC approximation to the cochlear transmission line model, from Schroeder, 1975.

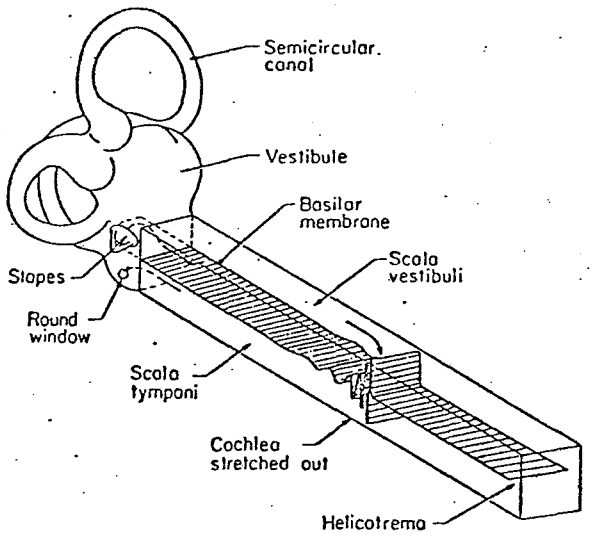
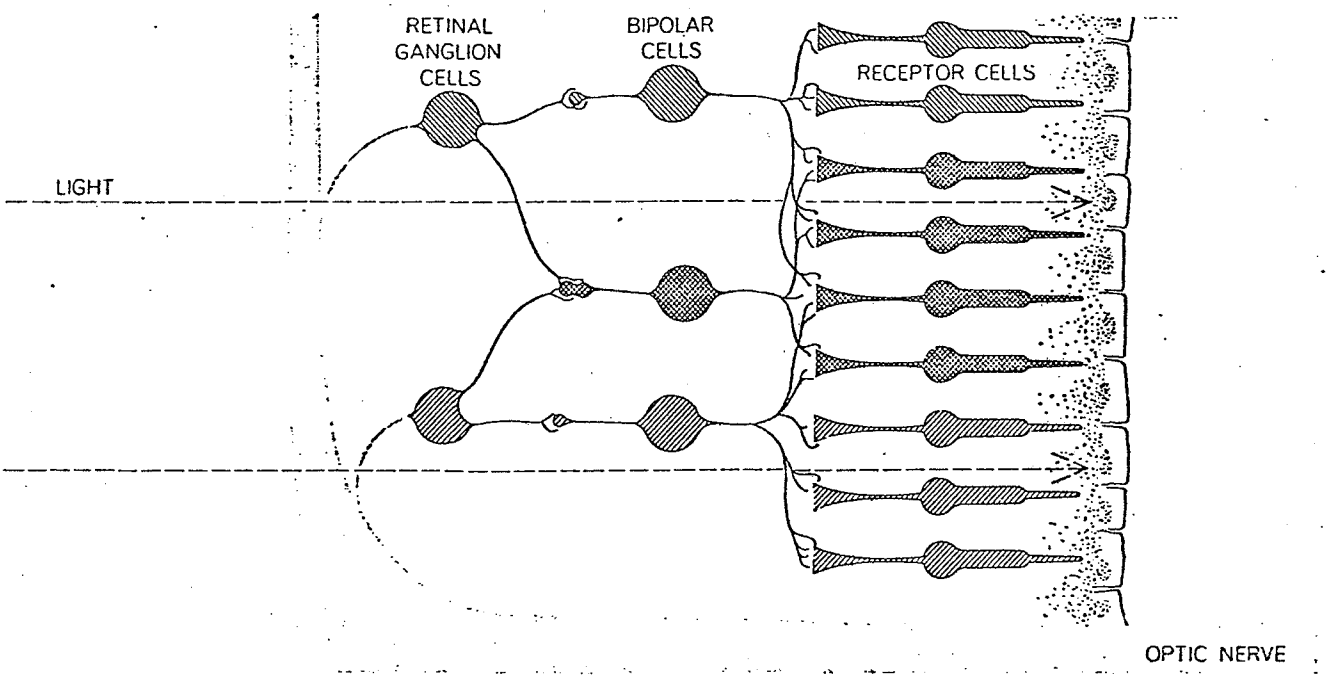


FIG. 1. Simplified physical model of the cochlea. The cochlea is uncoiled and approximated by two fluid-filled rigid-walled compartments (scalae vestibuli and tympani) separated by an elastic sheet (basilar membrane). The stapes, which is connected to the eardrum, is vibrated by sound, setting up a fluctuating pressure difference across the basilar membrane, which drives its motion. The response of the basilar membrane at an instant of time to a pure tone is schematically indicated. The vertical size of the scalae and the displacement of the basilar membrane are greatly exaggerated. The arrows in the cochlea show fluid flow. When the wavelength of the wave on the basilar membrane is larger than the height of the compartments, the region between the stapes and the helicotrema acts like a transmission line, a section of which is shown in Fig. 2.

Figure 3. Idealization of the fluid cavities of the cochlea, from Zweig, 1975.



STRUCTURE OF RETINA is depicted schematically. Images fall on the receptor cells, of which there are about 130 million in each retina. Some analysis of an image occurs as the receptors transmit messages to the retinal ganglion cells via the bipolar cells. A group of receptors funnels into a particular ganglion cell, as indicated by the shading; that group forms the ganglion cell's receptive field. Inasmuch as the fields of several ganglion cells overlap, one receptor may send messages to several ganglion cells.

Figure 4. The network of neurons innervating the retina, from Hubel, 1963.

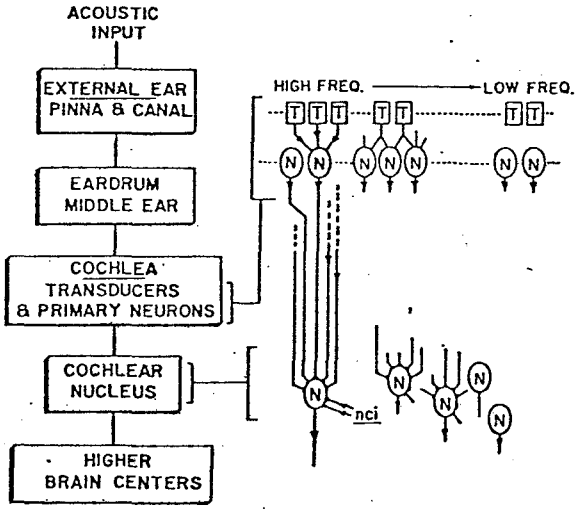


Fig. 1. Block diagram showing the path of the stimulating signal through the afferent auditory system. The transfer function of the external and middle ears of the cat has been studied extensively.[19],[53] Properties of the cochlea have been experimentally measured.[1] Studies of the primary neurons and neurons in the cochlear nucleus are cited throughout the text.

Right: Schematic diagram of the connection between the transducers in the inner ear and the neurons in the cochlear nucleus. T, transducer; N, neuron; and nci, noncochlear input.

Figure 5. Block and schematic diagrams of the hearing process, from Molnar, 1968.

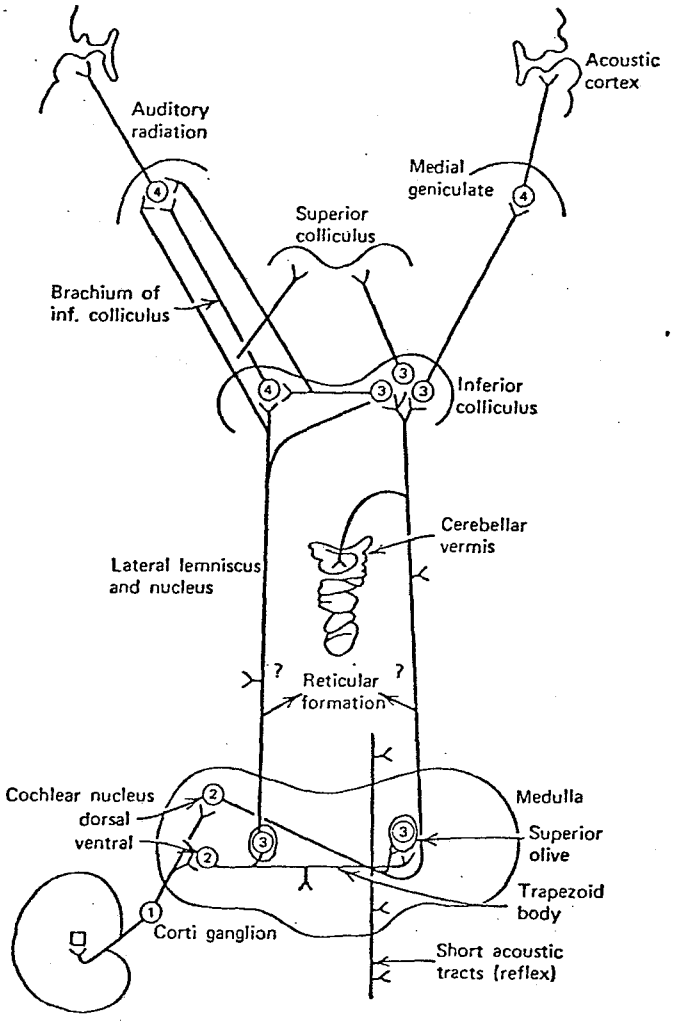
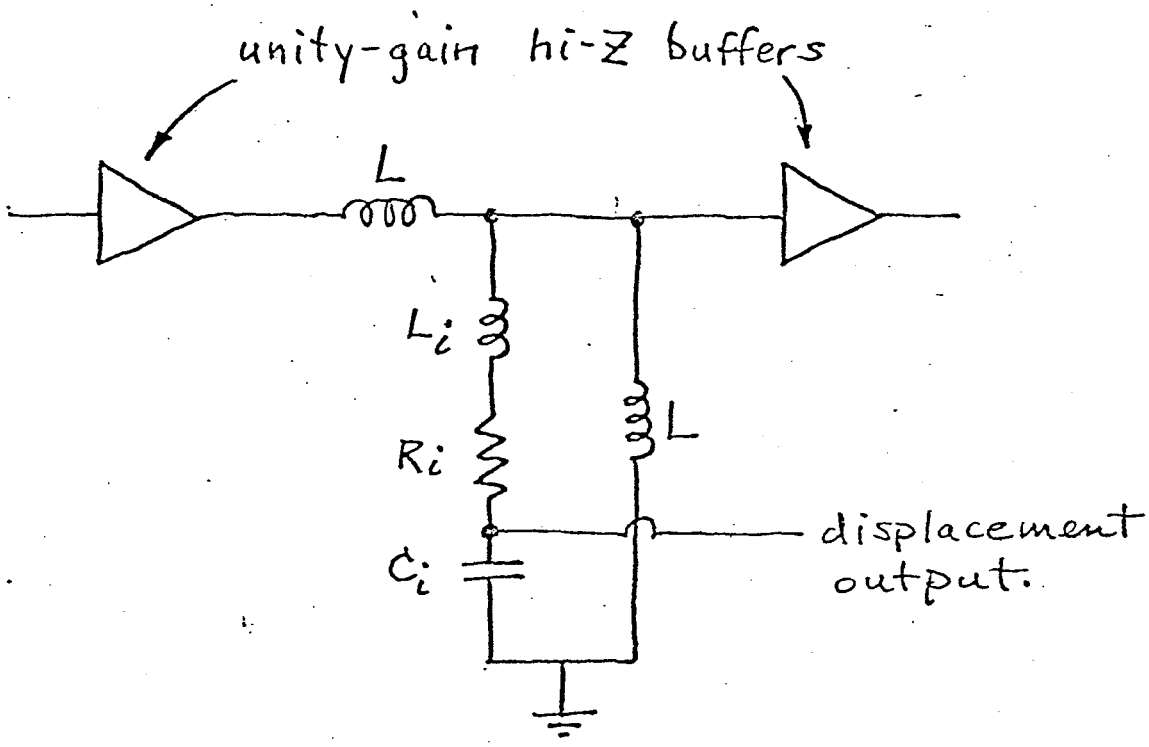
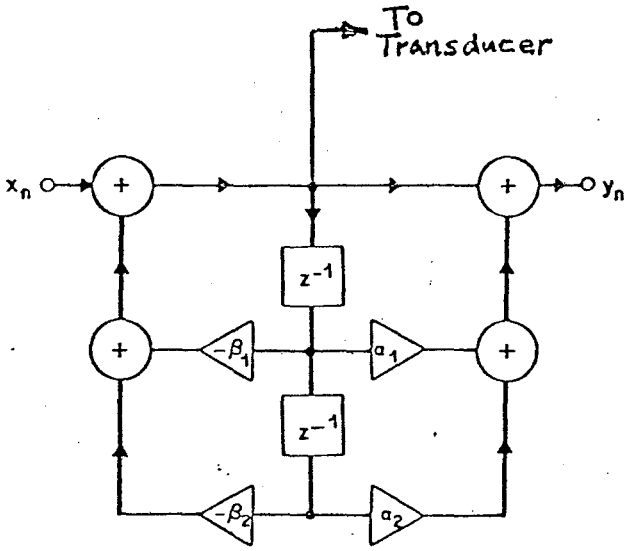


Fig. 6–14. Neurons of the auditory pathway (highly schematic and simplified). From Thompson (*572*) after Davis.

Figure 6. A schematic diagram of the auditory pathways, from Geldard, 1972.





Recursive second-order section.

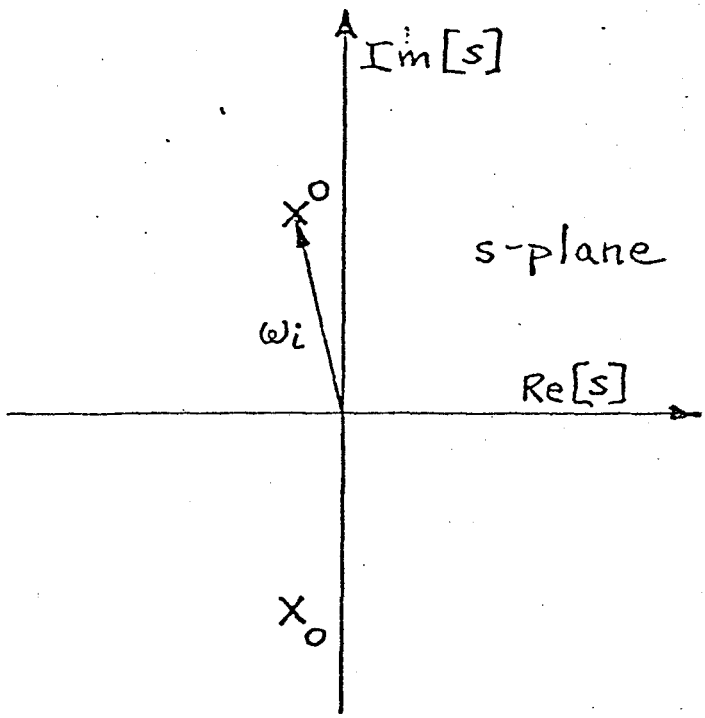


Figure 7. The buffered RLC cochlear model, its digital equivalent, and the pole-zero diagram of a section, by the author.

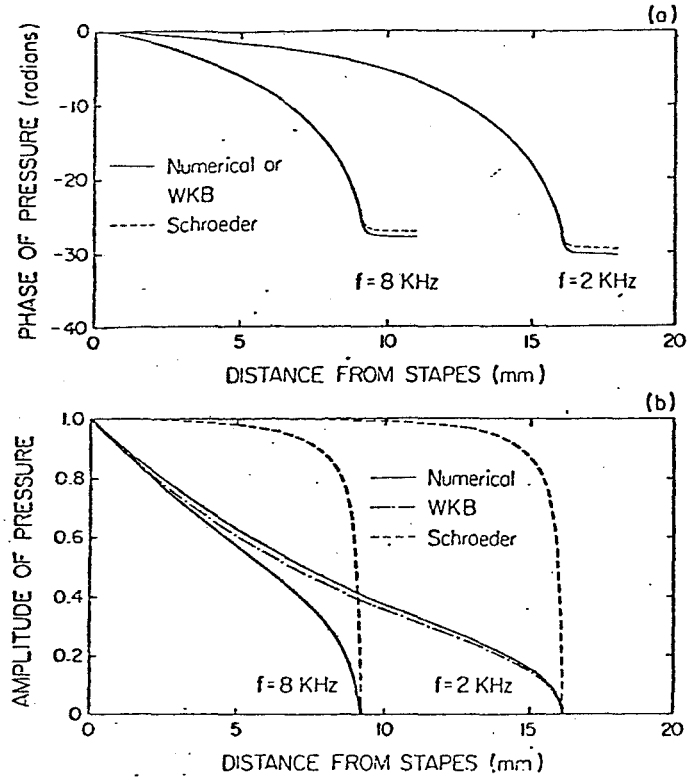


FIG. 3. Comparison of the approximate WKB solution with an accurate numerical solution for the pressure for frequencies (f) of 2 and 8 kHz. Parameter values are $N=5$, $\delta=0.02$, $\omega_m=2\pi$ (50 kHz), and $d=\frac{1}{2}$ cm. These values were chosen to be typical of those used in fitting experimental data. Schroeder's (1973) solution is also shown. (a) Phase as a function of basilar membrane position. The numerical and WKB solution are identical on the scale of the graph. Schroeder's solution for the phase is essentially identical. (b) Amplitude (normalized to unity at the stapes) as a function of basilar membrane position. The numerical and WKB solution at high frequency (8 kHz) are virtually indistinguishable. Schroeder's solution at high and low (2 Hz) frequency is significantly different and violates energy conservation.[9]

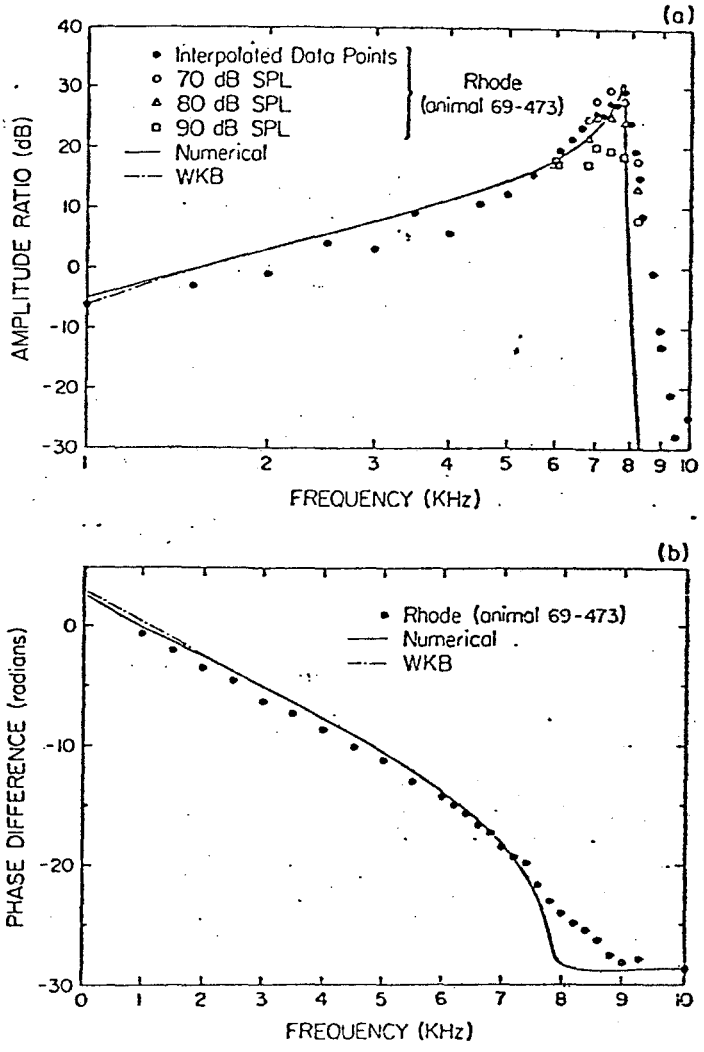


FIG. 4. Comparison of predictions from the model with Mössbauer measurements of the ratio of basilar membrane displacement at a fixed location on the basilar membrane to malleus displacement. Both the WKB approximate solution and the more accurate numerical solution are shown. The corresponding parameters are $\omega_r=2\pi$ (7.8 kHz), $N=5$, $\delta=0.02$, and $d=\frac{1}{2}$ cm, $\omega_m \gg \omega_r$. (a) Amplitude ratio as a function of frequency (logarithmic scale). (b) Phase difference measured in radians as a function of frequency (linear scale). Although there is qualitative agreement between theory and experiment, a detailed comparison in the region of large displacements and "breaking" phase is confounded by the nonlinearities in the data found at the very high pure-tone sound intensities used for stimulation. These nonlinearities are evident in (a), where the amplitude ratio is given for different sound-pressure levels.

Figure 8. Approximate tuning curves for the cochlear filter model, from Zweig, 1975.

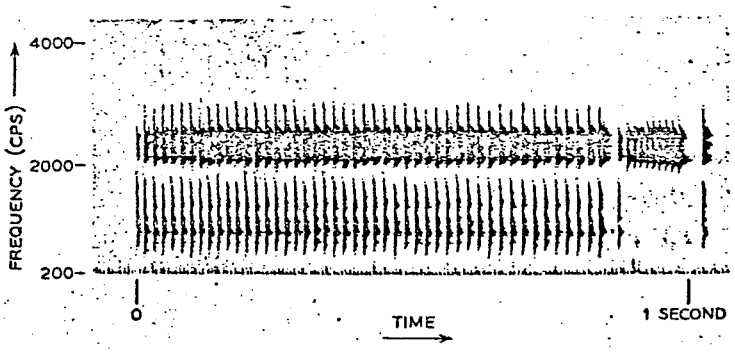


Fig. 7.3 A spectrogram of the sound of *Xenopus*, a South African toad.

Figure 9. A wideband sonogram showing pitch converted to time pulses, from van Bergeijk, 1960.

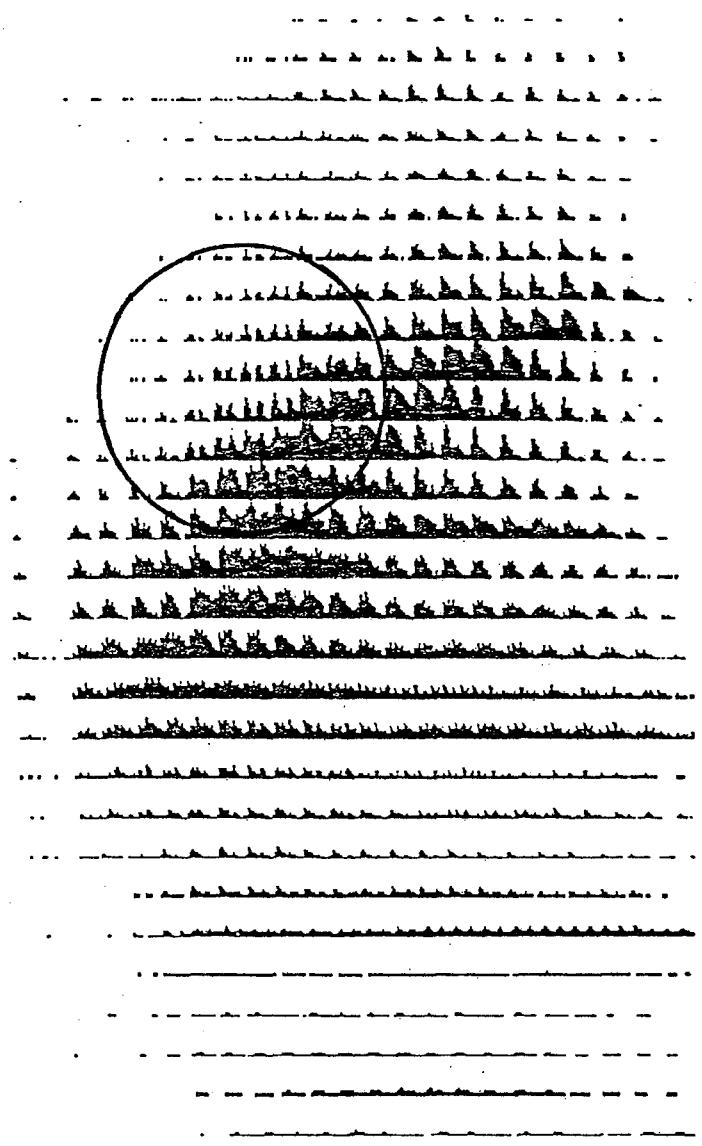


Figure 10. A sonogram showing pitch doubling in speech: the word "one" spoken and analyzed by the author.

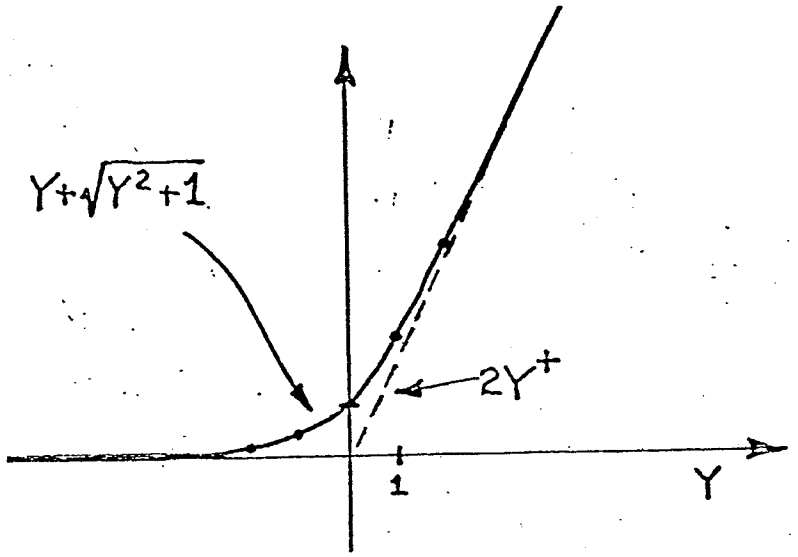


Figure 11. Characteristics of a soft half-wave rectifying transducer, from Schroeder, 1975.

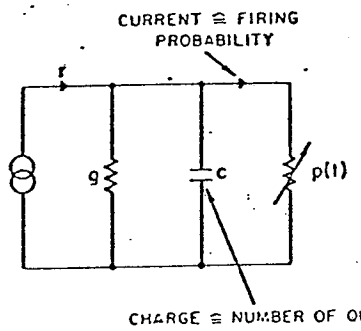


Fig. 18. Electrical analog circuit of model [39] for mechanical-to-neural transduction process effected by hair cell. Conductance $p(t)$ varies with acoustic simulation in half-wave rectifier fashion. High-average conductance discharges capacitor C, thereby limiting current representing firing probability of nerve attached to hair cell. This "depletion" process of charge on capacitor is thought to account for the neural "adaptation" seen in Fig. 15.

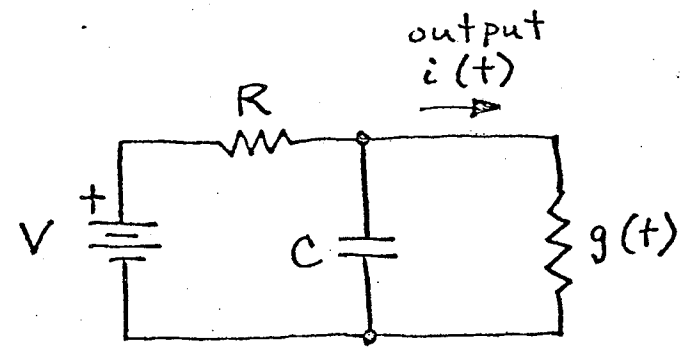


Figure 12. The depletion model of transducer adaptation, from Schroeder, 1975.

Figure 13. Response of the soft rectifier with adaptation, from Schroeder, 1975.

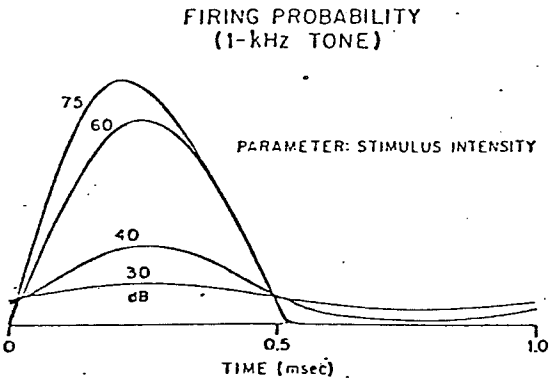


Fig. 19. Firing probabilities of model [39] for sinusoidal acoustic signal of 4 different input levels. Firing probability follows instantaneous signal amplitude linearly at 30 dB. Above 60 dB, firing probability corresponds to half-wave rectified signal and changes only slightly with further increases in signal level. This model result corresponds to period histograms shown in Fig. 17.

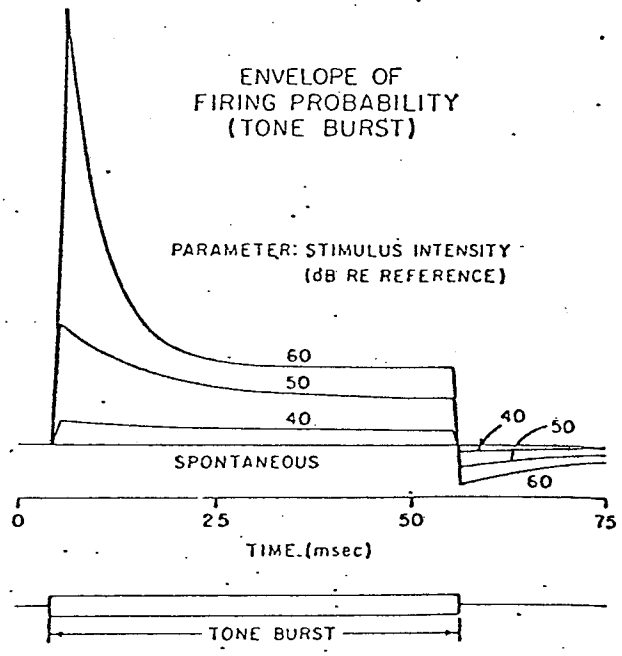


Fig. 20. Model result illustrating neural adaptation for stimulation with tone burst. At bottom, temporal extent (50 ms) of tone burst is shown. Compare with Fig. 15.

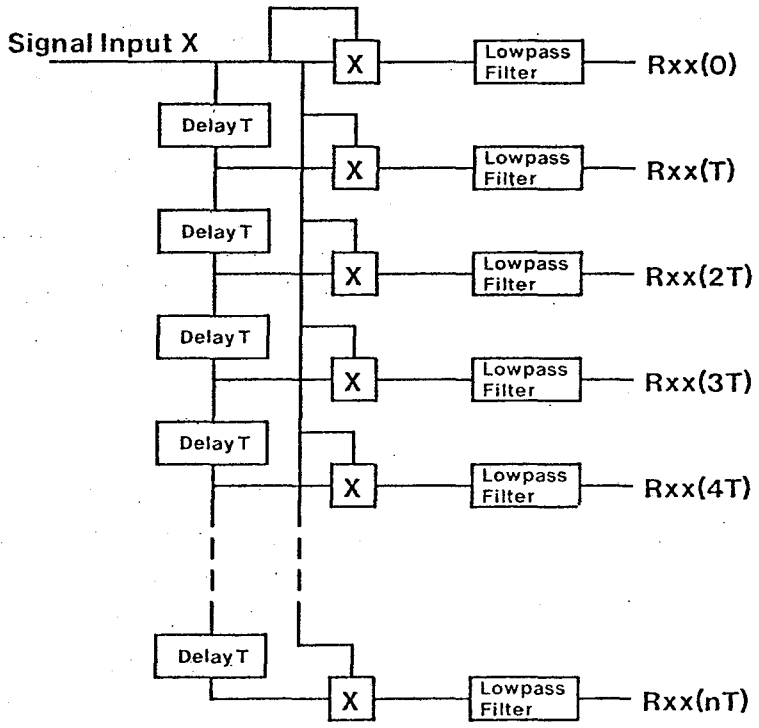


Figure 14. A network for estimating the autocorrelation function of a signal, by the author.

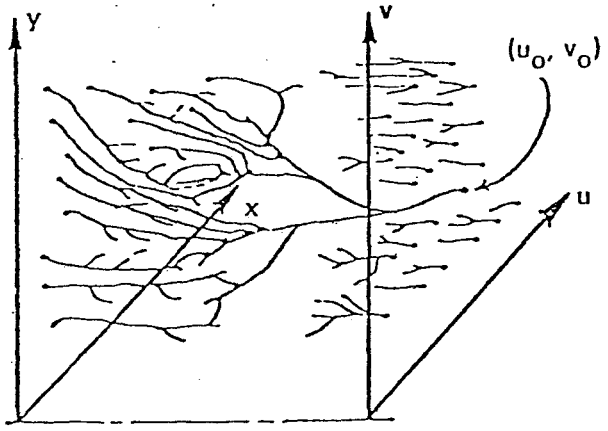


FIGURE 10 Schematic of effect of retinal field (x-y plane) upon cortical neuron at coordinates (u_0, v_0).

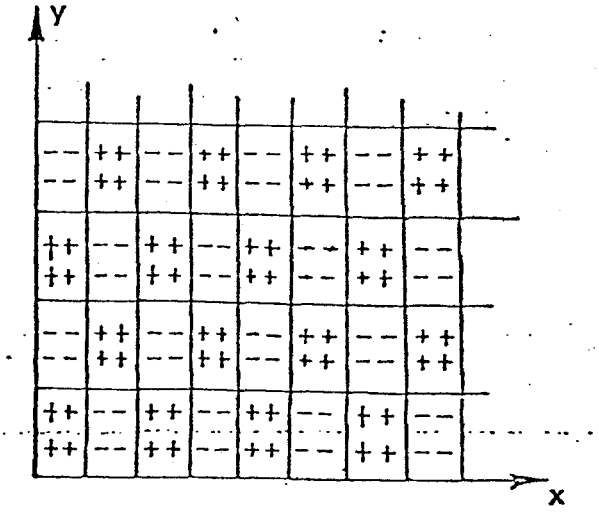


FIGURE 11 Receptive field of cortical neuron at (u_0, v_0) if the receptive field is of the form cos (u_0x) cos (v_0y). Positive is excitatory; Negative, inhibitory.

Figure 15. The receptive field of a neuron in a feature detector circuit, from Swigert, 1971.

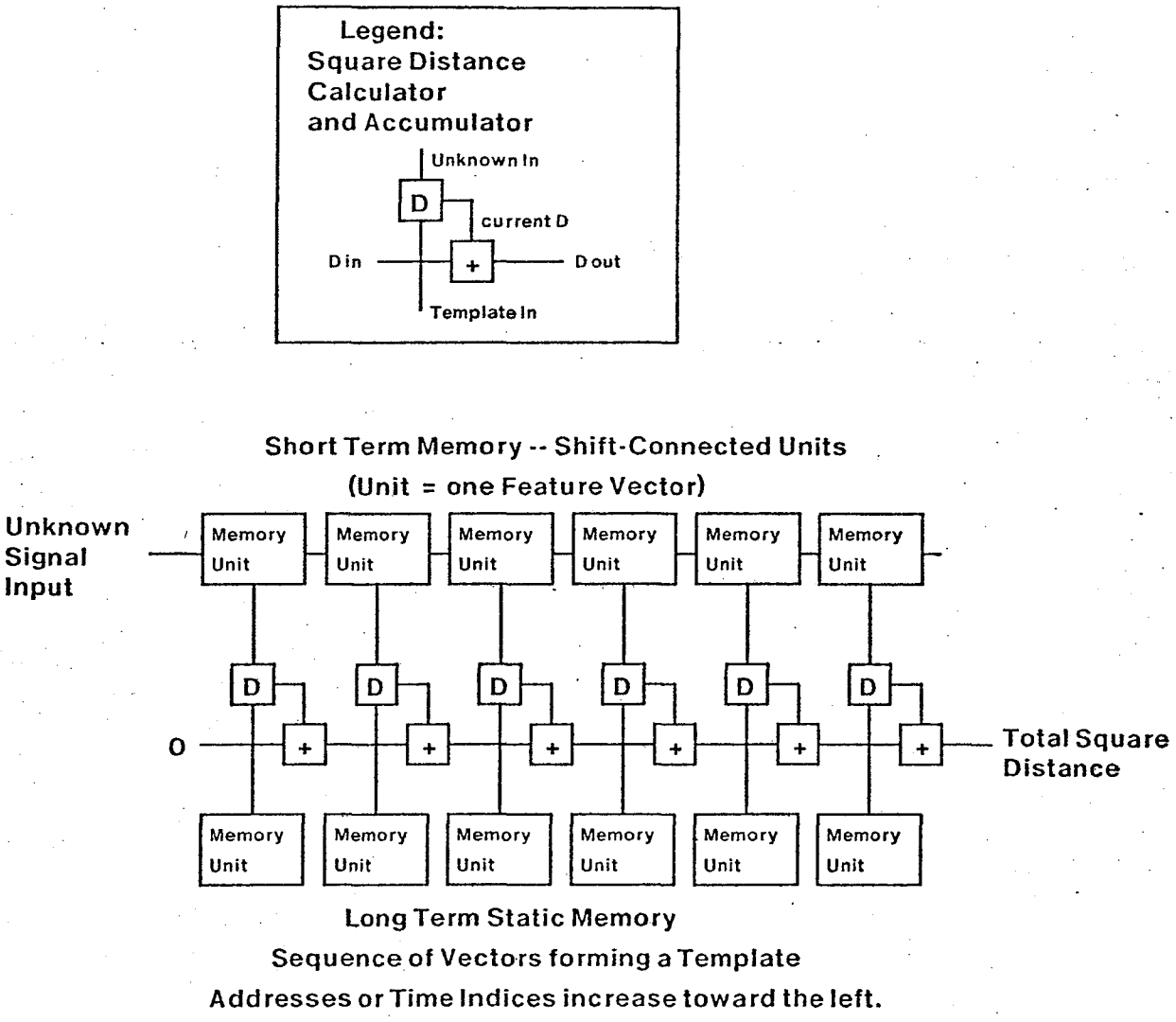


Figure 16. A pattern matching network for a simple associative memory, by the author.

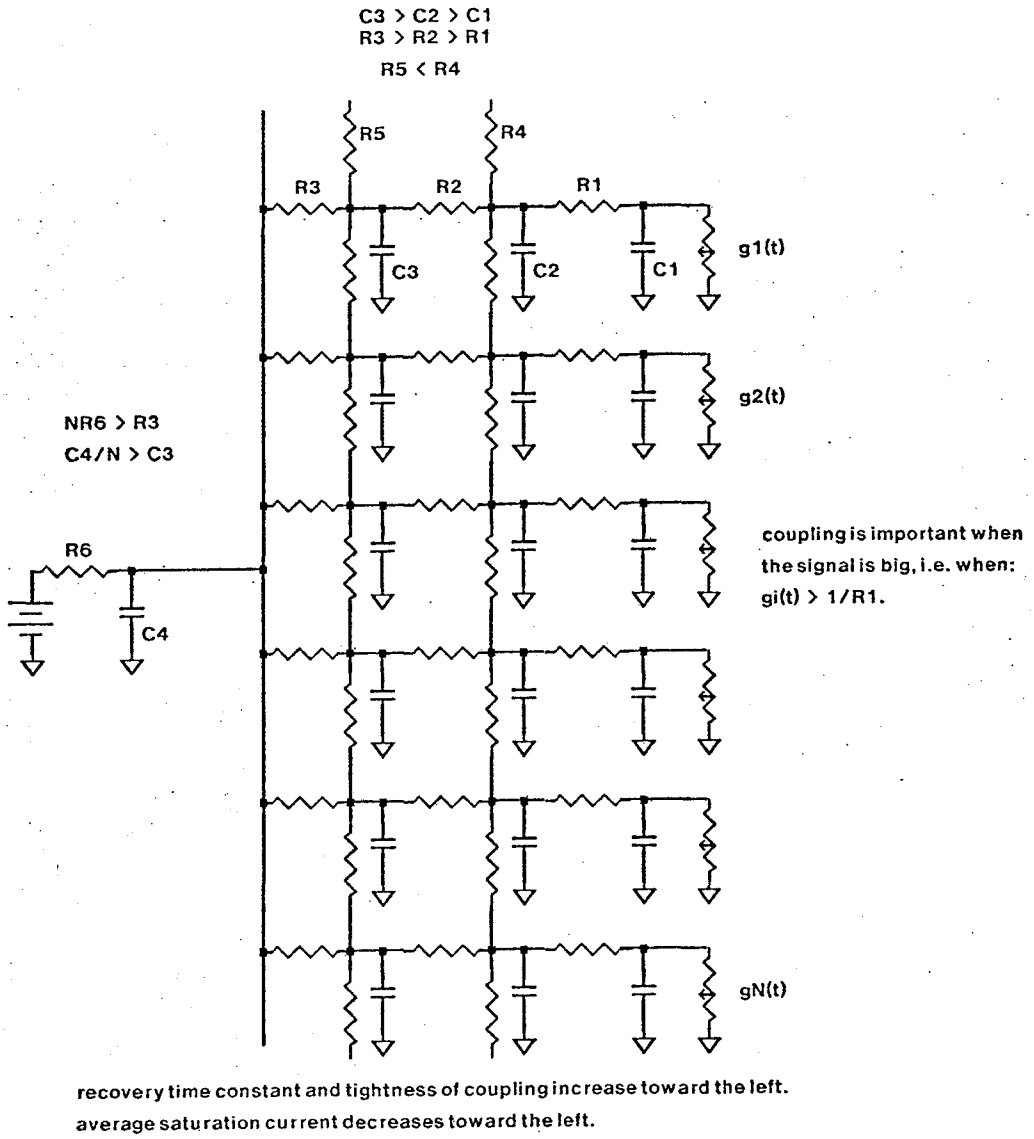


Figure 17. Interconnected Adaptation Mechanisms -- the Depletion/Diffusion Model by the author.

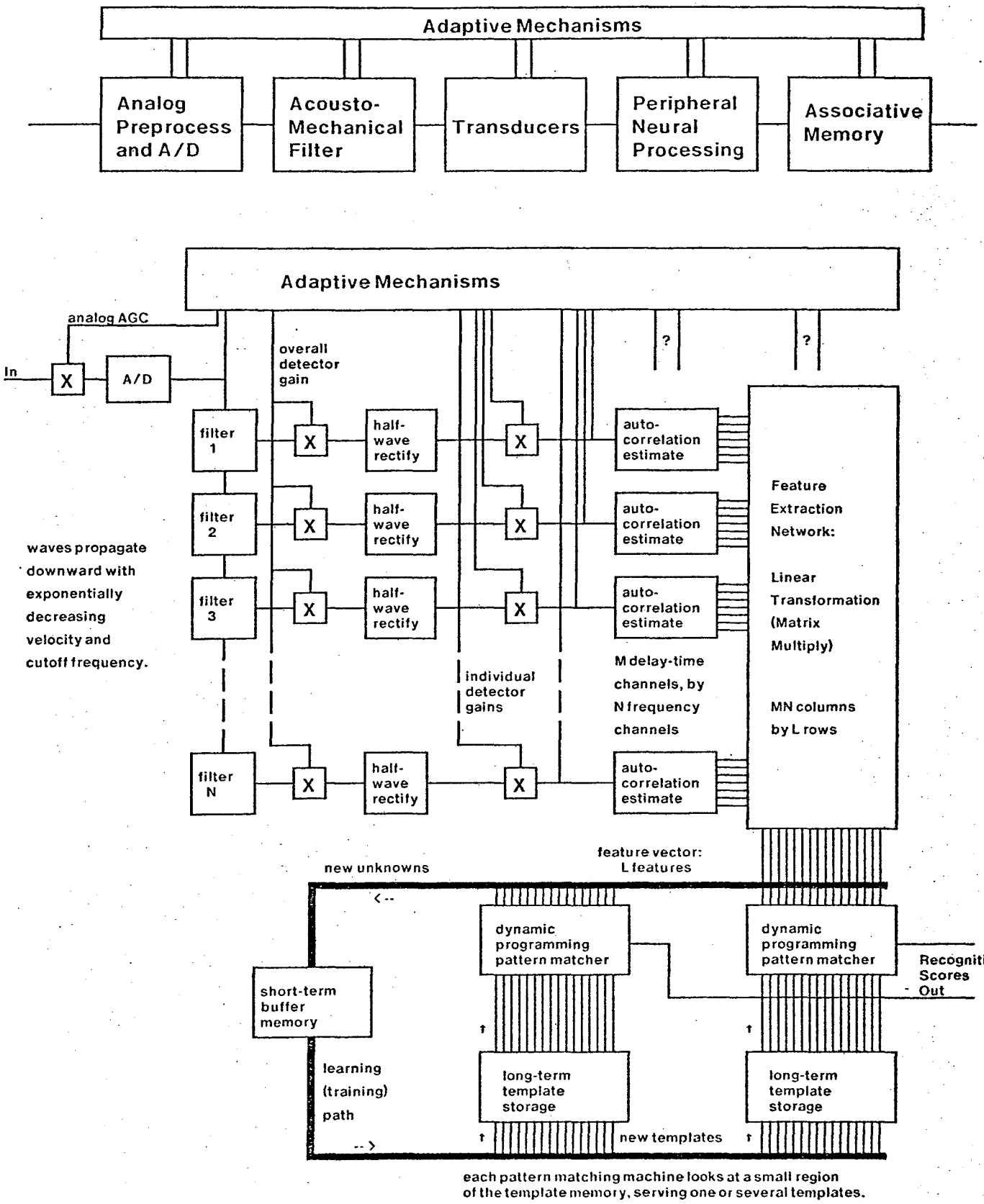


Figure 18. A block diagram of the Signal-Processing Model of Hearing, by the author.