

## SPEECH PROCESSING BY REAL AND SILICON EARS

To demonstrate a new approach to sensory information processing, a model of the human auditory periphery is being implemented in silicon. The model uses analog processing methods to produce a real-time parallel output from speech that closely resembles the averaged auditory nerve outputs from the cochlea. In this article, the current understanding of the processing in the ear is reviewed, and a snapshot is given of the ongoing attempt to implement this processing in a very large scale integrated chip.

### INTRODUCTION

Why should the ear be of engineering interest? What characteristics of the ear would be valuable in a signal processor? How would such a processor be realized, and where would it be of use? The intent of this article is to address these questions. Specifically, it surveys an ongoing project in which the physiology of the ear is used in designing a hardware speech processor. The project, known as HEEAR (Hopkins Electronic Ear), involves engineers and researchers from the Electrical and Computer Engineering Department of the School of Engineering and the Applied Physics Laboratory of The Johns Hopkins University.

The acoustic signal of speech is the principal medium for human communication. The invention of the telephone by Alexander Graham Bell was an outgrowth of his intense interest in speech.<sup>1</sup> Today we see intense academic and industrial activity aimed at speech recognition by computer, which has the potential of making typewriters or keyboards "hands off" devices. An area of special interest to us is aiding the deaf in speech communication. Tactile aids and cochlear implants make speech accessible to deaf persons who get no benefit from hearing aids. These applications work best when the speech signal is not easily corrupted by noise and when strong amplitude compression occurs without distortion. (Dynamic range is the ratio, in decibels, between the weakest and strongest signals in the operating range of a system. Amplitude compression occurs when the output dynamic range is smaller than the input dynamic range.) The human ear has these characteristics, and a speech processor that emulates the ear could improve the performance of these devices.

In this article, the term "ear" signifies the pinna and ear canal (outer ear), the middle ear, the cochlea, and the bundle of nerve fibers called the auditory nerve connecting the cochlea to the brain. Within the cochlea, sound vibrations are converted to neurological signals carried to the brain by the auditory nerve fibers. Another term for this entire system is auditory periphery. We begin with a brief description of the structure and physiology of the mammalian ear, followed by an outline of the coding of some speech sounds by the ear and a model for

those processes. We then focus on the realization in silicon of a speech processor having some of the characteristics of the auditory periphery.

### THE AUDITORY PERIPHERY (EAR)

Figure 1 shows the gross anatomy of the human auditory periphery.<sup>2,3</sup> The main parts of the auditory periphery are the pinna, ear canal, middle ear, cochlea, and auditory nerve bundle. The ear canal allows the inner ear to be deep within the temporal bone, the hardest bone we have. It also has a broad resonance of about 2000 Hz, in the middle of the speech spectrum. The eardrum (tympanum) and three middle ear bones (malleus, incus, and stapes) allow efficient coupling of airborne sound at the tympanum to the fluids of the cochlea. The cochlea is a coiled, fluid-filled tube divided lengthwise by the basilar membrane into two large chambers, the scala vestibuli and the scala tympani. It also contains a small chamber called the scala media (see Fig. 2A). The basilar membrane moves in response to acoustic waves in the cochlear fluid excited by the eardrum.

Figure 2B shows the central portion of the cochlea in a cross-sectional view, including the nerve fibers leaving the receptor cells of the cochlea that are part of the auditory nerve shown in Figure 2A. The receptor cells of audition are called hair cells (so named because of the cilia at the apex of the cells; see Fig. 2C). (Analogously, the receptor cells of vision are called rods and cones.) The hair cells transduce the vibrational motion of the basilar membrane into nerve impulses in the fibers of the auditory nerve. The stiffness of the basilar membrane decreases by a factor of 100 from the base to the apex of the cochlea, and this is a major determinant in the tuning properties of the ear. The inside of the cochlea along the basilar membrane contains one row of about 3000 inner hair cells and three rows of outer hair cells totaling about 20,000. Nerve impulses induced by vibration of the basilar membrane travel through about 30,000 auditory nerve fibers into the brain. Over 95% of the auditory nerve fibers running from the cochlea to the brain are stimulated by chemical connections called synapses on the inner

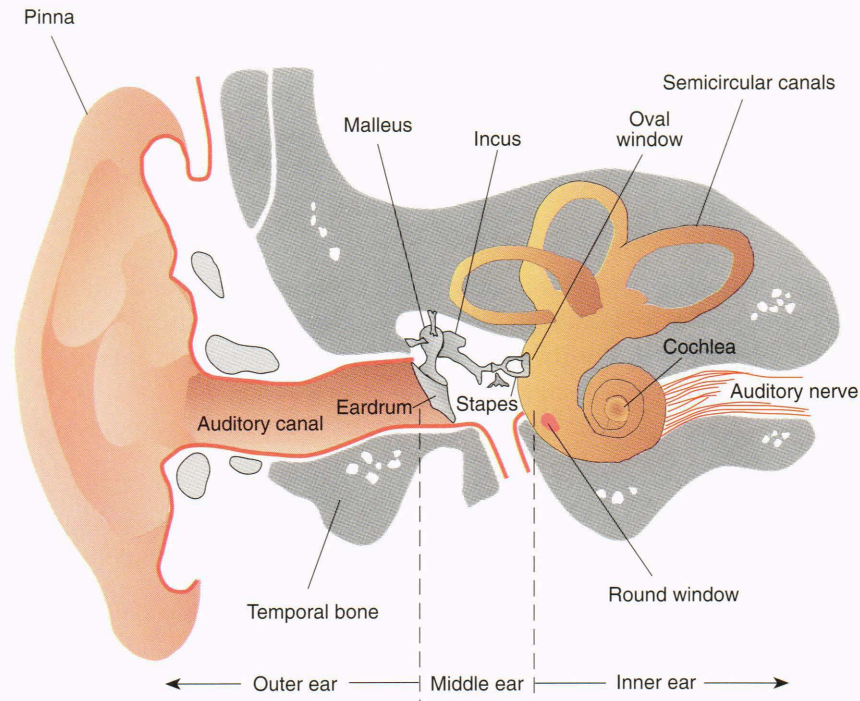


Figure 1. Gross anatomy of the auditory periphery.

hair cells (see Fig. 2C). Each inner hair cell has about ten synapses, and each synapse provides the sole excitation for one auditory nerve fiber.

In the 1950s, the functional properties of the cochlea were thought to be well understood, mainly because of the research of Georg von Békésy.<sup>4</sup> He was awarded the Nobel Prize in physiology in 1961 on the basis of his definitive studies of the mechanical properties of the ear. Figure 3 summarizes von Békésy's results on the vibration of the basilar membrane derived from experiments with preparations of excised temporal bone. He demonstrated that sinusoidal (or "pure tone") stimulation caused the traveling wave in the cochlea to move from base to apex (Fig. 3A). The maximum height of the wave appeared closer to the basal end of the cochlea for higher frequencies. In Figure 3B, the relative amplitude of vibration is plotted as a function of pure-tone frequency at several locations along the basilar membrane. The curves exhibit a broad resonance, and for locations closer to the stapes, the resonance occurs at higher frequencies.

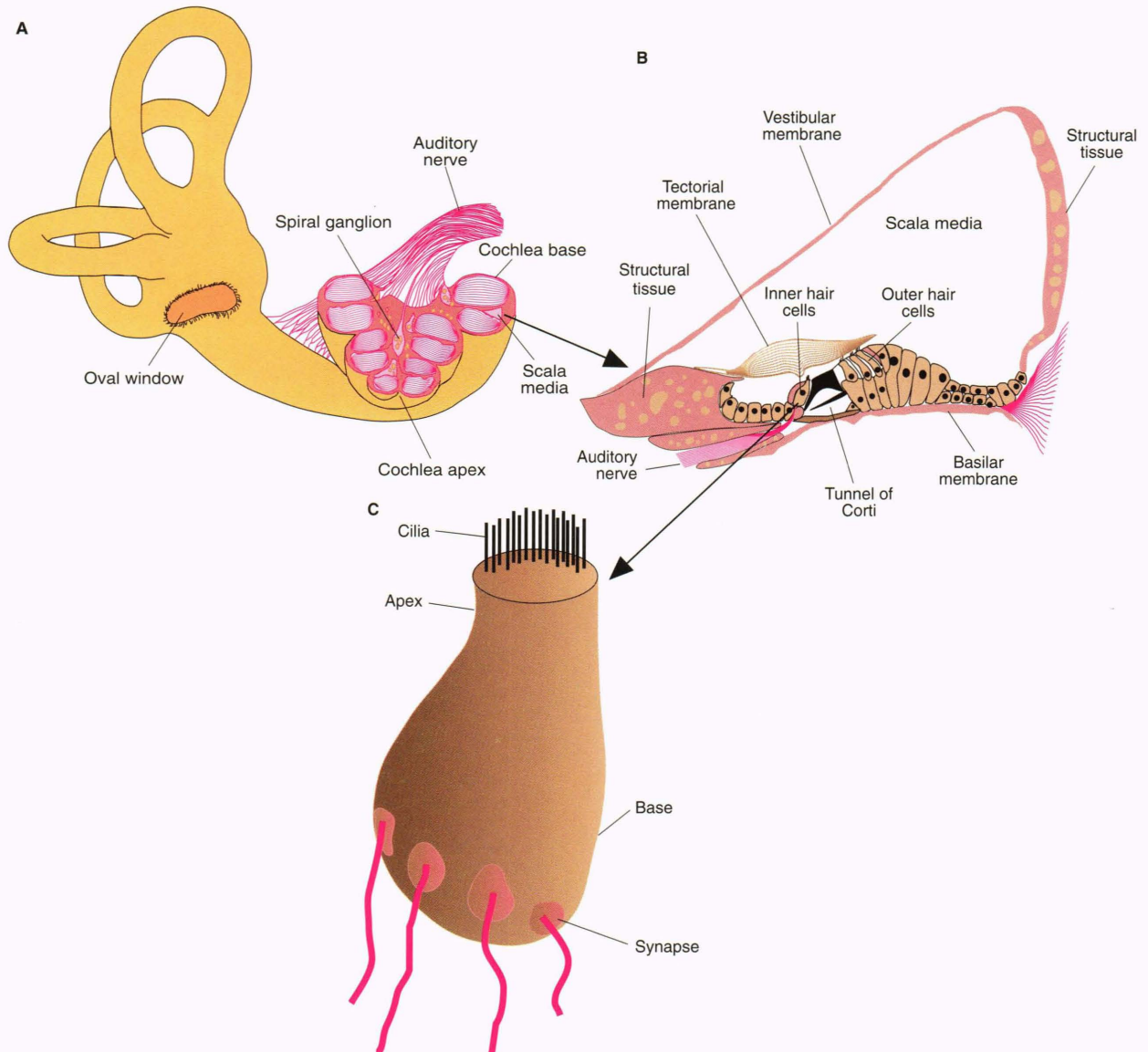
Between 1960 and 1990, experimental results led to the extension and revision of concepts of the physiology and anatomy of the cochlea and the auditory nerve. To the surprise of most who studied audition, Spoendlin<sup>5</sup> showed that nearly all the fibers of the auditory nerve connect only to inner hair cells. One wondered what the function might be of the 20,000 outer hair cells. At least a partial answer was to come.

Another surprising finding concerned the vibration of the basilar membrane. The optical microscope and strobe used by von Békésy to measure the motion of the basilar membrane required high sound intensities, so high that continued stimulation at those levels would cause hearing loss in humans. In addition, virtually all von Béké-

sy's preparations were postmortem specimens. From his research, he had concluded that the system function relating vibration of a given location on the basilar membrane to acoustic input was linear. William Rhode<sup>6</sup> was able to repeat the experiments using the Mössbauer effect, a more sensitive technique than the optical method used by von Békésy and one that allows *in situ* experiments on live animals. For the lower intensities he was able to use, Rhode found that a given location on the basilar membrane exhibited a sharp resonance, and his results indicated an interesting nonlinearity in the response, as shown in Figure 4.

Recent work by Brownell<sup>7</sup> has shown that the outer hair cells change length and shape in response to electric fields such as those observed in the cochlea. The present view holds that these hair cells are involved in an active feedback process that affects the vibration of the basilar membrane and is also responsible for the sharp tuning at resonance and for the nonlinear characteristics. Our understanding of the effect of the active process (and thus our ability to model it) is developing as research continues. When metabolism in the cochlea is greatly reduced (e.g., by making the animal anoxic or by using a postmortem specimen), the sharp tuning and nonlinearities are no longer present, and the mechanical properties are those observed by von Békésy. Since von Békésy used excised, isolated temporal bones, the differences between his and more recent results are not surprising.

Another important feature of the auditory periphery is the effect of inner hair cells on the transduction process by which the vibrations of the basilar membrane result in neural signals. The cilia of the inner hair cells of the mammalian cochlea are like stiff, miniature fork prongs. We have known for some time that these hair cells are



**Figure 2.** The cochlea. **A.** Anatomy showing the cochlea in cross section. **B.** Enlarged view of the scala media showing the basilar membrane, hair cells, and supporting cells. **C.** Enlarged view of an inner hair cell.

morphologically polarized;<sup>8</sup> that is, shear stress on the cilia of a hair cell in one direction causes excitation of the auditory fibers that synapse on the cell, whereas stress in the opposite direction causes suppression. The transformation relating the temporal pattern of nerve impulse activity in an auditory fiber to the basilar membrane motion that leads to excitation of that fiber is strongly nonlinear, but reasonable models exist.

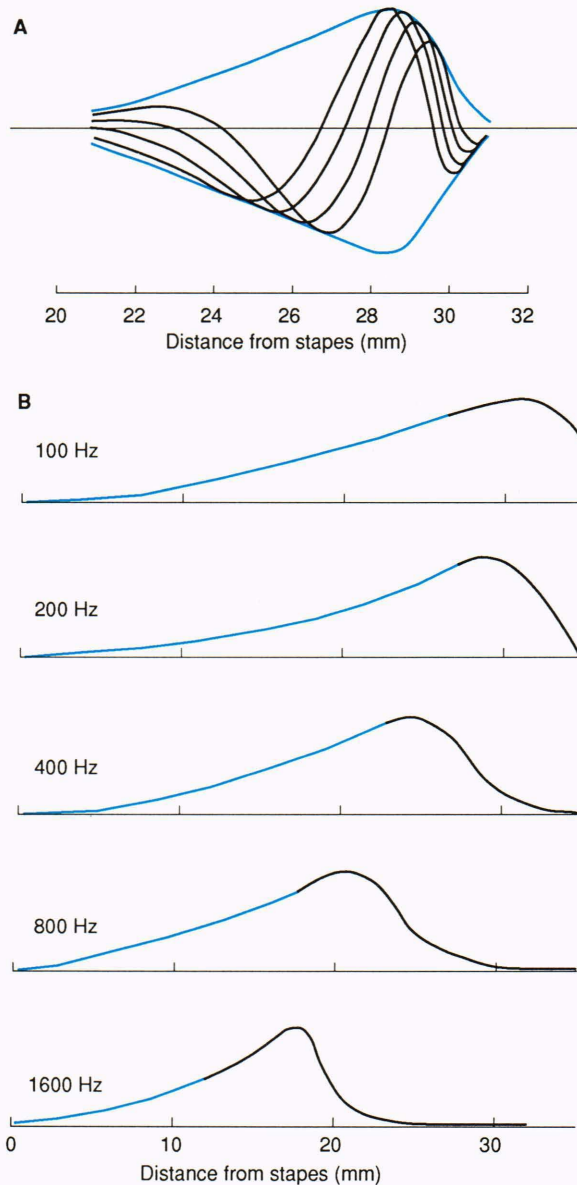
### CODING OF SPEECH IN THE AUDITORY NERVE

The 30,000 fibers of the auditory nerve provide a parallel, distributed coding of the auditory input. Each fiber sends a train of nerve impulses to the cochlear nucleus, the first brain structure to receive information from the cochlea. Thus, the cochlea transforms the acoustic signal to a spatiotemporal pattern of neural activity in the auditory nerve.

By using hyperfine pipette electrodes, researchers have been able to isolate single nerve fibers of the auditory nerve and observe their impulse trains. Most of these fibers exhibit random “spontaneous” activity, that is, one records impulses (also called neural spikes) in the absence of acoustic stimulation. Virtually all of the fibers respond to some sound by a corresponding increase in the rate of impulses.

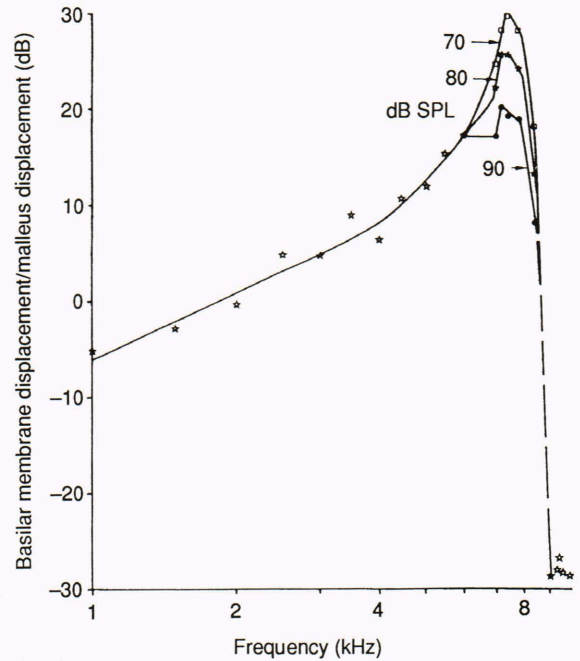
Information is carried in the nerve fibers by a “pulse modulation code,” that is, the information is in average timing of neural impulses. In the fibers of the auditory nerve, the temporal pattern of impulses is highly random for both responding and spontaneous conditions. One may envisage the nature of such random time series by means of an analogy: If each impulse from a fiber made a “pop” sound, the activity of a group of fibers would have the sound of popcorn in the making.

A “tuning curve” may be obtained for a given fiber by plotting its threshold of response to pure tone acoustic



**Figure 3.** Vibration of the basilar membrane from von Békésy's studies of cadaver specimens. **A.** Displacement pattern of several sequential, discrete times within a 200-Hz cycle. The envelope is shown by blue lines. **B.** Envelope of excitation for sinusoidal excitation at five frequencies. Black parts of the curves are based on experimental observation, blue parts on extrapolation. The curves are normalized to have the same peak amplitudes. The horizontal axis for each curve is distance along the basilar membrane from the stapes (i.e., the basal end of the cochlea). The human cochlea is about 35 mm in length. The vertical scale is greatly magnified, since even for the high levels von Békésy used (about 140 dB relative to a reference level of 0.0002 dyn/cm<sup>2</sup>, which is about human threshold at 1000 Hz), the maximum amplitude of vibration was only about  $3 \times 10^{-3}$  mm.

stimulation as a function of the frequency of the tones. At each frequency, the intensity is slowly increased until a discernible increase occurs in the spontaneous firing activity. Typical tuning curves are V-shaped, with a clear minimum point at some frequency. This minimum point is called the characteristic frequency (CF) of the fiber and



**Figure 4.** Ratio of amplitude of vibration of a place on the basilar membrane to cochlear drive as measured by the motion of the malleus. Each curve represents a different sound pressure level (SPL) in decibels relative to a reference level of 0.0002 dyn/cm<sup>2</sup>. The Mössbauer technique was used to measure vibration amplitudes. For a linear relationship, the curves would coincide. Non-linearity is seen at the location of greatest sensitivity, where the sharpest peak in the curve occurs for the lowest SPL. Experiments were performed using live squirrel monkeys. (Data reproduced, with permission, from Ref. 6, p. 1223: © 1971 by the Acoustical Society of America.)

is determined by the position along the basilar membrane of the inner hair cell that drives that fiber. Therefore, the spatial dimension of the spatiotemporal pattern is along the basilar membrane and can be calibrated in terms of frequency (i.e., CF).<sup>2</sup>

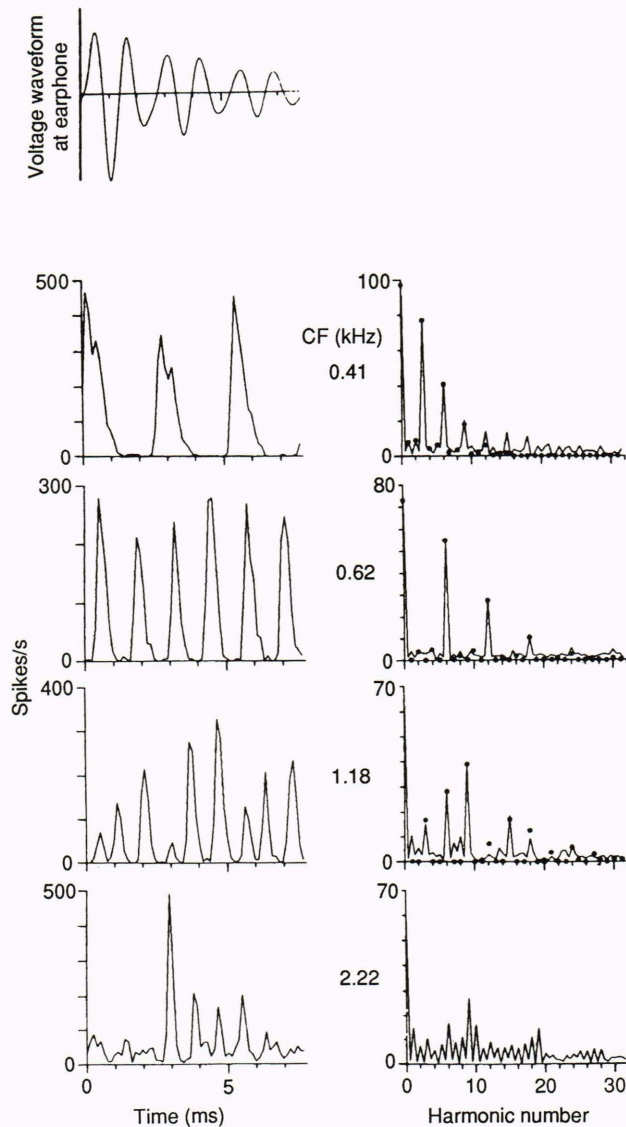
The coding of speech stimuli by the auditory nerve fibers is of great interest in human communication; however, until the late 1970s, virtually all studies of auditory nerve activity dealt with responses to pure tones, clicks, or combinations of tones, but not with speech signals. Considering the strongly nonlinear nature of the auditory periphery, it was difficult to predict speech coding on the basis of responses to these simpler stimuli.

### Steady-State Vowels

In two papers published in 1979,<sup>9,10</sup> Murray B. Sachs and Eric D. Young of The Johns Hopkins University School of Medicine reported experimental findings of the coding of sustained steady-state vowels by the auditory nerve fibers. A key feature of their work was that they recorded from several hundred individual auditory nerve fibers in each experiment. Thus, they obtained a representative sample of the spatiotemporal pattern of neural activity for the stimuli used. Upon isolation of a fiber, the spontaneous firing rate and the CF were determined. Then impulse train responses to the synthesized

vowels / $\epsilon$ / (as in bet) and / $\alpha$ / (as in father) were obtained for a range of sound intensities. The use of synthetic rather than real speech signals is part of the quantitative methodology in the physiological studies. (See Refs. 9 and 10 for details.)

The waveform at the top of Figure 5 is one period of the synthesized / $\alpha$ /. The steady-state vowel signal is periodic, a string of many repetitions of this waveform. The fundamental frequency of the signal is about 120 Hz. Auditory fibers (in fact, nearly all nerve fibers) do not sustain average firing rates over about 250 per second. Thus, observation of the impulse train of a particular fiber would show that each period of the sustained vowel



**Figure 5.** Period histograms (left) and their power spectral densities (right) for four fibers; characteristic frequencies (CF's) shown in the center are applicable to both columns. The stimulus was a synthesized steady-state vowel / $\epsilon$ /. One pitch period of the vowel is shown at the top of the left column; the time scale is the same as that for the period histograms. (Data reproduced, with permission, from Ref. 10, p. 1383; © 1979 by the Acoustical Society of America.)

stimulus would have, at most, one or two impulses. The observer could estimate the fiber's average firing rate, but could hardly discern the temporal structure of the periodic waveform. Fine structure is, however, evident in period histograms obtained as follows: One period of the stimulus is divided into several intervals. The time of each impulse is recorded according to the interval of the period in which it falls. This measurement is made for numerous periods of the stimulus, and a histogram is constructed that gives an estimate of the average instantaneous firing rate of the fiber during a period of the stimulus.

Since, at most, a few periods of a vowel are adequate for recognition by a listener, what meaning can we attribute to averages of such sparse activity taken over hundreds of periods? One appealing interpretation is that the random firings within the auditory nerve bundle exhibit an ergodic-like property. We may thus assume that the period histogram of a single fiber averaged over time is the same as the instantaneous average of a group of fibers. Therefore, if a sample of a few hundred fibers is used to represent about 30,000 fibers, each recorded fiber represents 100 or more fibers that are driven by the same or nearby hair cells. If the spike trains in these fibers are assumed to be statistically independent, and if the responses of the recorded fiber to the repeated presentations of the vowel are assumed to be statistically independent, then the period histograms such as those shown in Figure 5 (left) may be considered to be an ensemble average. Since these records are based on time averages for many stimulus periods, they may be considered equal to the average activity of the 100 or more fibers represented by the recorded fiber for one stimulus period. The assumptions of independence are reasonable and are supported by data. A more difficult question is whether such group activity is used in the neural processing of vowels and other speech signals at higher levels than the auditory nerve.

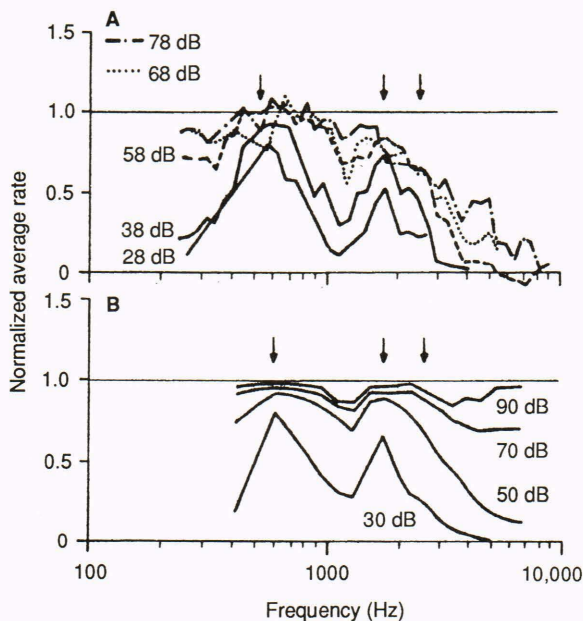
In voiced speech, the configuration of the vocal tract is important in determining the envelope shape of the speech signal's power spectrum. That configuration is usually well represented by the vocal tract resonances called formants, which are seen as peaks in the spectrum of the speech signal. In steady-state vowels, the frequencies of the formants, especially the first (F1) and second (F2), largely determine a listener's judgment of which vowel is presented.

Sachs and Young set out to determine from their data how well the formants were represented in the activity of the auditory nerve fiber population. The first candidate code they considered was average firing rate as a function of CF.<sup>9</sup> Results for the vowel / $\epsilon$ / are shown in Figure 6A. At very low sound levels, we see fairly good representation of F1 and F2 by peaks in the plot of the normalized average rate versus CF. But at higher intensities, these peaks are lost because of the small dynamic range of the fibers, which causes them to saturate. Since normal conversational levels are in the range of 55 to 65 dB, rate coding appears to be a poor coding candidate.

An alternate coding scheme for the vowels takes account of the fine temporal structure seen in period

histograms.<sup>10</sup> The period histograms shown on the left of Figure 5 are dissimilar for the different fibers; the right column of the figure displays the discrete Fourier power spectral densities of the period histograms for each fiber. Thus, the graph displays the magnitude of each fiber's response at each harmonic of the fundamental frequency (F0) of the steady-state, periodic vowel. Sachs and Young devised an algorithm, called the ALSR (average localized synchronous response), to show the representation of the vowel's spectrum in the population of auditory nerve fibers. The algorithm computes a global response to each harmonic by averaging the power spectral densities over fibers. The averages, however, are weighted for each harmonic to include only those fibers with CF's within 1/2 octave of that harmonic. Thus, the averages are "local" for each harmonic, that is, they only include those fibers with a CF close to the harmonic.

Stated more concisely, for each harmonic, an average is taken for fibers with CF's local (by 1/2 octave) to the harmonic's frequency of activity synchronized to the harmonic—thus, the name of the algorithm. The powerful insight of Sachs and Young in devising the ALSR is apparent in Figure 7A, which shows the ALSR records for the vowel /e/ for an intensity range of 28 to 78 dB. Note the vertical displacement of the curves; otherwise, they would overlap extensively. Clearly, the ALSR record for steady-state vowels provides an accurate and precise representation of F1, F2, and F3, and exhibits very strong amplitude compression without distortion. This coding scheme was also shown to represent vowels well under conditions of noise contamination.<sup>12</sup>



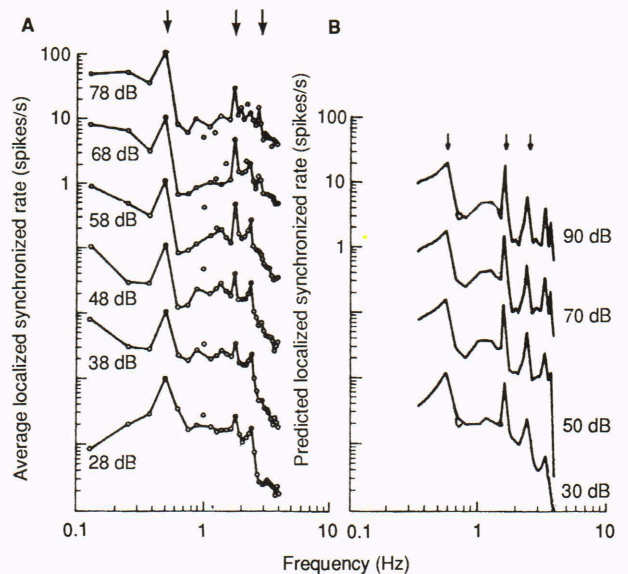
**Figure 6.** Rate coding of the steady-state vowel /e/ for various input intensities in dB SPL (see Fig. 4 caption). **A.** Normalized, measured average firing rate response plotted as a function of characteristic frequency. The arrows indicate the first three formants, F1, F2, and F3 of the vowel. **B.** Simulation results. (Data reproduced, with permission, from Ref. 11, p. 151; © 1988 by the Acoustical Society of America.)

The ALSR computation may seem complex and particular to steady-state vowels. In terms of the coding of speech signals by the population of auditory nerve fibers, however, it suggests that, in addition to ergodicity, (1) meaningful collective temporal behavior within groups of fibers exists, and (2) the temporal behavior provides an excellent representation of the formant structure of synthesized, steady-state vowels. As we discuss in the following section, localized synchrony can be computed by means other than the digital Fourier transform, and the role of temporal coding in representing the power spectrum of speech is not restricted to synthesized, steady-state vowels.

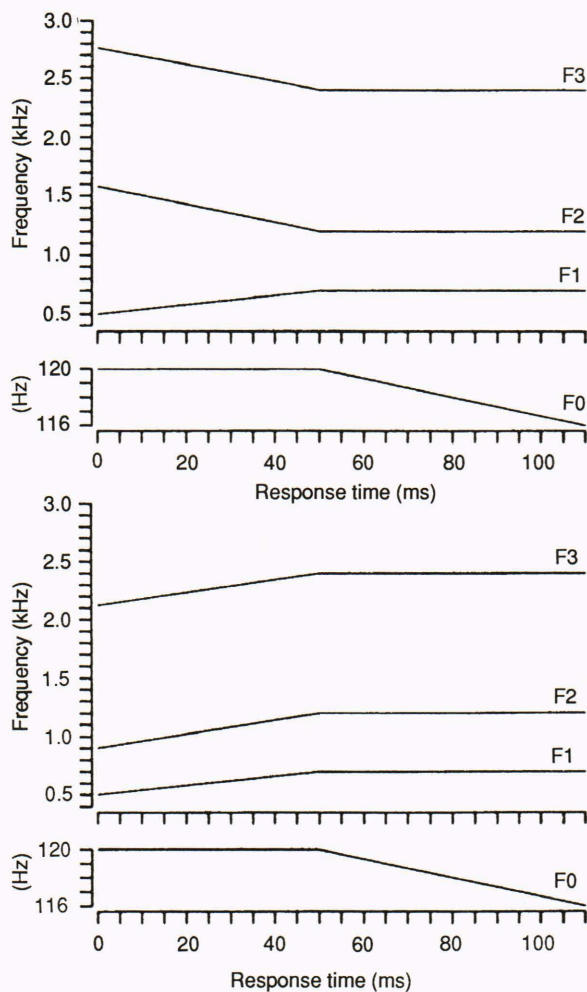
### Syllables

Steady-state vowels are the simplest speech sounds. In further experiments by Miller and Sachs,<sup>13</sup> the temporal coding of the consonant-vowel syllables /dα/ and /bα/ was recorded. Again, they used a large population of individual fibers in each experiment and recorded many responses from repeated stimuli to allow averaging. Figure 8 shows the time course of formants F1, F2, and F3 for the two synthesized syllables. The consonants are differentiated by the slopes of F2 and F3 in the first part of the syllables.

The ALSR algorithm tracked the time course of the formants reasonably well; however, for our purposes, a more direct picture of the neural firing rate data is nicely displayed by neurograms, a method used recently by Secker-Walker and Searle.<sup>14</sup> From the raw data collected



**Figure 7.** The average localized synchronized rate (ALSR) response to the vowel /e/. **A.** Neural ALSR responses determined by using the same data as those used for Figure 6A. **B.** Simulation results. Arrows indicate the first three formants, F1, F2, and F3. Both plots are successively shifted one order of magnitude for clarity; otherwise, they would overlap extensively. Numbered scales refer only to the top curve. (Data reproduced, with permission, from Ref. 11, p. 153; © 1988 by the Acoustical Society of America.)



**Figure 8.** Trajectories of fundamental frequency (F0) and the first three formants (F1, F2, and F3) for synthesized syllables /dα/ (top) and /bα/ (bottom) used by Miller and Sachs.<sup>13</sup> (Data reproduced, with permission, from Ref. 14, p. 1428: © 1990 by the Acoustical Society of America.)

by Miller and Sachs,<sup>13</sup> they computed an average instantaneous firing rate for each recorded fiber as a function of time since syllable onset. Unlike vowels, syllables are not steady-state signals; hence, period histograms are not appropriate. Instead, the averages were computed over hundreds of repeated presentations of the stimulus.

Figure 9 shows the neurogram for /dα/. Each trace of the neurogram is derived from the data for one fiber, the CF of which is given at the left margin. As we move down the vertical axis, we are sampling fibers excited by different hair cells along the cochlea from apex to base.

Secker-Walker and Searle noted that the pattern of instantaneous firing fell into groups that are labelled G0 through G3. Within each group, the traces are quite similar, and the edges of the groups are quite distinct. In addition, they noted that the rate of vocal excitation (voice pitch, F0) was well represented by group G0, and the formants F1, F2, and F3 by G1, G2, and G3, respectively. Clearly, the averaged instantaneous firing rate forms a high-resolution temporal encoding of the speech for-

nants. These averages appear to be representative of the group-averaged activity of clusters of fibers with the same or close CF's.

A system with a single input and a single output obeys an "uncertainty principle" concerning time and frequency resolution, whereby very fine frequency resolution is gained only with poor time resolution and vice versa. In the auditory nerve fibers, a representation appears to have excellent frequency and time resolution.

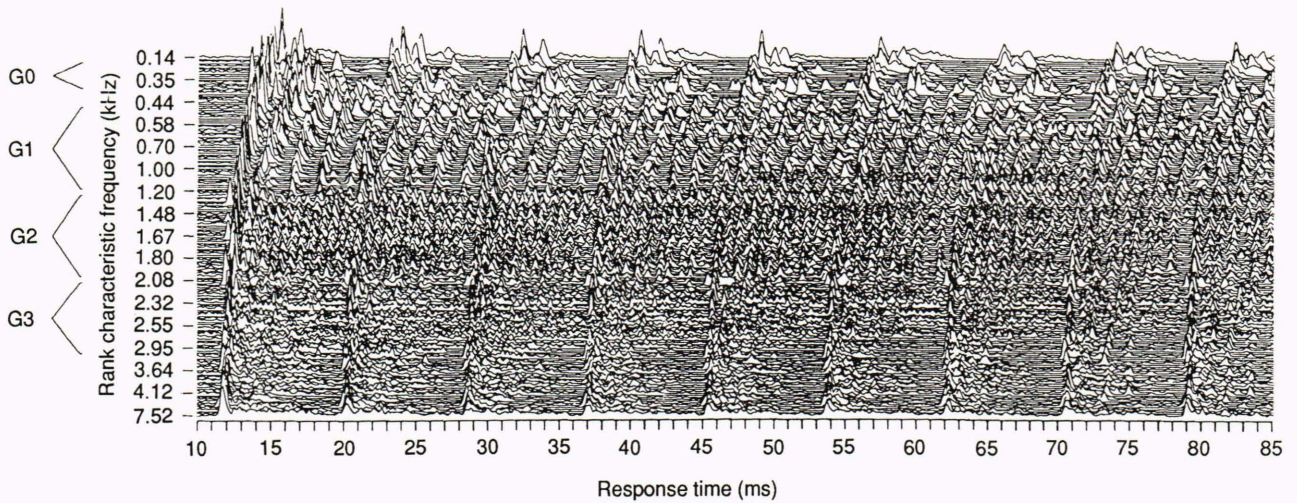
What is responsible for this phenomenon? The first notable feature is the dense sampling of the cochlea by the auditory nerve fibers, so that a location on the basilar membrane (with its CF) is represented by many auditory nerve fibers. Secondly, there is the random (Poisson) code of impulses in the auditory nerve fibers. The combined activity of groups of fibers produces a modulated random train, where the modulation is capable of time resolution considerably better than 1 ms. Finally, the filter characteristics of the basilar membrane are sufficiently broad that input spectral properties dominate the temporal pattern of response rather than filter characteristics.<sup>10,14</sup> By these mechanisms, the ear can encode the spectral content of signals accurately and respond to rapid changes. Interestingly, the pulse code in this system that is the usual carrier of information by nerve fibers is actually converted by the collective behavior of the group to an analog code especially suited to the transmission of speech information.

## COMPUTER MODEL OF THE EAR

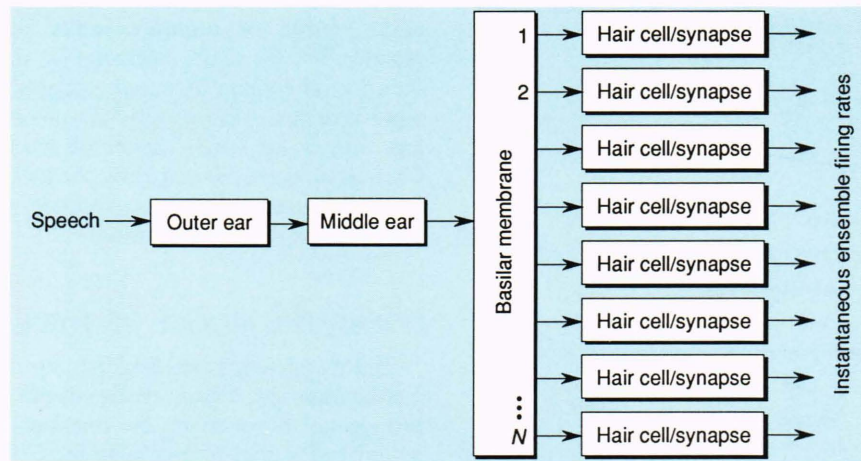
Before one considers building a processor incorporating the attractive characteristics of the ear, a model of the processing of sound by the ear must be achieved and neural and model results compared. This approach was the basis of recent dissertation research by Karen L. Payton in the Electrical and Computer Engineering Department of The Johns Hopkins University.<sup>11,15</sup> In the foregoing discussion, the auditory periphery was described in sections: outer ear, middle ear, basilar membrane, and hair cell/synapse. Payton's model is modular (as are most models of the ear), incorporating these same sections, as shown in Figure 10. The design of each module was based on published data and models.

Each module was designed to fit sinusoidal input-output characteristics over a wide frequency and amplitude range. If the system as a whole were linear, the response to complex stimuli could be derived from a frequency representation of the stimulus and the system transfer function. The basilar membrane and the hair cell/synapse modules are nonlinear, however. Payton designed the tuning properties of the basilar membrane model to fit those obtained experimentally, but all nonlinearity of the simulation was in the hair cell/synapse module.

Considering that the nature of the hair cell/synapse module is strongly nonlinear and that sinusoidal inputs were used to study and model the module, agreement between the simulation and neural results for the broad spectra, steady-state vowel stimuli was not certain. Results of the simulation are shown beside the neural results in Figures 6B and 7B. The correspondence is



**Figure 9.** Neurogram of 178 fibers in response to /dα/. (See the text for a full explanation under Syllables.) (Data reproduced, with permission, from Ref. 14, p. 1429: © 1990 by the Acoustical Society of America.)



**Figure 10.** Block diagram of a speech processor based on the physiology of the auditory periphery.

close. The temporal code (Fig. 7A) gives an excellent representation of formants F1, F2, and F3 and of total amplitude compression. The rate code, like the experimental results (Fig. 6A), represents F1 and F2 well at low intensities, but saturates quickly at higher intensities.

Payton's computer simulation was successful, but was tested for a very limited set of stimuli, two steady-state vowels over a range of intensities. One approach to the direction of further work would be to test the simulation with more stimuli and possibly to expand the model to include nonlinearity in the vibration of the basilar membrane. This path is not attractive, however, because simulation of this complex system on a digital computer is slow, and although larger, faster computers would speed the work, real-time simulation is not in view.

Andreas Andreou, a faculty member in the Electrical and Computer Engineering Department of JHU, was familiar with Payton's dissertation research and had been working with a new technology, analog very large scale integration (VLSI), with support and collaboration from Robert Jenkins and his APL group. Andreou suggested

fabricating a real-time, low-power speech processor using analog VLSI to implement the model. The potential usefulness of the approach as a modeling tool was evident. With Payton's simulation and several other efforts to model the auditory periphery to guide us, we embarked on a project aimed at simulating the auditory periphery in silicon. The next section describes the project in progress.

## SILICON MODEL OF THE EAR

### Subthreshold Analog VLSI Approach

In recent years, fabrication of VLSI circuits has become widely accessible to universities through programs established by the Defense Advanced Research Projects Agency and the National Science Foundation. One ingredient of the availability of VLSI is MOSIS (the metallic-oxide-on-silicon implementation system), established at the University of Southern California to serve as a silicon broker for fabricating integrated circuits quickly and



at low prices. Another ingredient is the availability of computer-aided design tools to facilitate the design of such complex circuits.

Power dissipation is a chief limitation in any large-scale implementation of neural architectures containing many neural units. The massive concurrence attainable in computing with large neural networks imposes serious limits on the amount of power that each circuit can dissipate. In any practical scheme, the devices, like neurons, must operate with currents in the nanoampere range or smaller. Carver Mead<sup>16</sup> has suggested that silicon analog processing be performed by operating the circuits with the transistors essentially turned off (subthreshold operation). In this operating regime, transistors behave like nearly ideal transconductance devices with very small but highly controllable diffusion currents in the nanoampere range. This technique offers the possibility of achieving ultra-low-power analog processing at the high integration density of silicon transistors with fully customized, fast-turnaround chips.

The arbitrary processing precision of digital systems is limited only by the analog-to-digital (A/D) converters, and such systems are immune to device mismatch and fluctuation problems. The achievement of real-time simulation digitally for complex systems can be difficult and expensive, however. In contrast, analog systems do not require A/D or D/A conversions and have the advantages of efficient use of silicon area, ultra-low power consumption, and truly real-time concurrent operation. Thus, our computing paradigm aims to map biological computations onto silicon, where the physics of the underlying circuits is intimately related to the computational algorithm.

An important circuit element of an analog cochlear processor is the continuous-time filter for representation of the middle ear and the basilar membrane. Continuous-time filter designs often require good control and matching of transistor properties to achieve the desired characteristics. Silicon transistors, although fabricated closely together on the same chip, may have significant parametric variations. A successful filter implementation requires attention to the geometric mismatch compensation of the device and the proper tuning strategy. These factors represent a major technical challenge (see the boxed insert).

### Model Elements

The middle ear model characterizes the displacement or velocity at the stapes in response to sound pressure at the eardrum and is usually modeled as a linear low-pass filter. A middle ear filter approximating the frequency response of stapes displacement was implemented on silicon using low-pass filters built from transconductance amplifiers and capacitors. This design takes a small area (about  $600 \mu\text{m} \times 60 \mu\text{m}$ ) on the chip and is stable in the hearing frequency range. Figure 11 shows the design of the filter and the measured frequency response from a test cell included on our first Mosis test chip. The response closely approximates experimental results.

Cochlear dynamics relates basilar membrane motion to the sound stimulus coupled into the cochlear fluids.

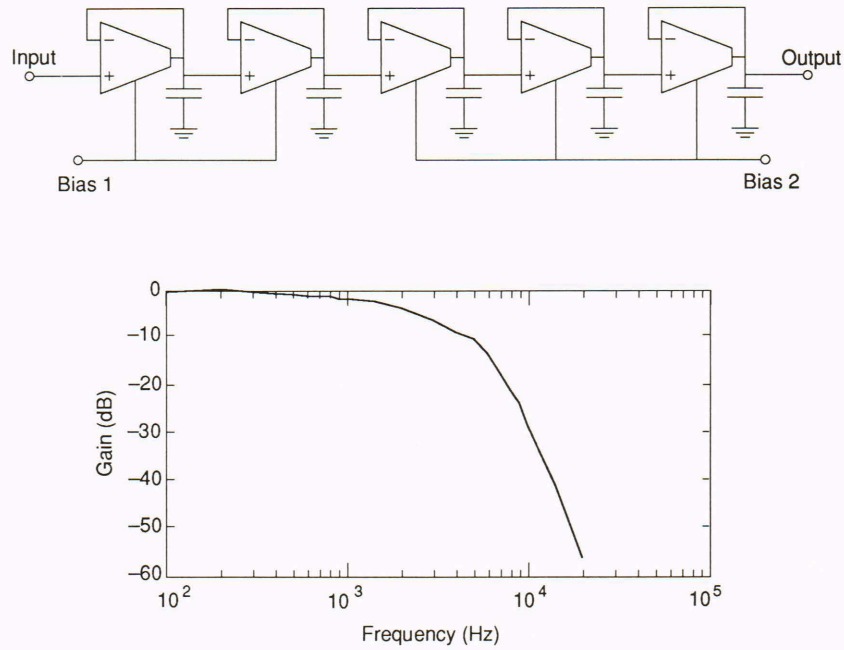
### MONOLITHIC CONTINUOUS-TIME FILTERS

Transistors and capacitors can be easily realized on metallic-oxide-on-silicon (MOS) chips. Passive resistors can also be implemented, but the designer has no control over the exact resistance. One approach to implementing continuous-time integrated filters is to achieve desired time constants with transconductances and capacitances. Filters of this type, called transconductance-C filters, do not require passive resistors and are built solely out of transconductors and capacitors. One advantage of this approach is that time constants and other parameters of a fabricated filter can be electronically tuned by adjusting voltages that control the transconductances. A saturated, subthreshold MOS transistor itself is a transconductor. With several MOS transistors, we can design the operational transconductor amplifier (OTA), which is an active-circuit building block that outputs a current proportional to the difference between the two input voltages. The output/input ratio (i.e., the transconductance) is adjustable by means of a bias voltage. With fixed capacitances of on-chip capacitors, all time constants in a monolithic transconductance-C filter are determined by the transconductances. The transconductance of an OTA can be designed to have an exponential dependence on the bias voltage. This scheme provides a wide range of transconductances that result in a filter parameter range of several orders of magnitude. A monolithic filter using subthreshold transconductances and easily obtained capacitances of 1 to 2 pF can be tuned to any frequency in the hearing range of 20 Hz to 20 kHz.

We did not attempt to model cochlear fluid dynamics, but lumped the entire filtering properties of the cochlear mechanics into the transfer functions of the basilar membrane module of our model. In our initial design, the basilar membrane module has linear characteristics, since such a representation worked reasonably well in several simulation studies. Nonlinearities, such as those seen in Figure 4, which are more difficult to implement, may be introduced as the work proceeds and as the role of basilar membrane nonlinearities becomes better understood.

Although the basilar membrane (together with associated structures called the cochlear partition) is a distributed system, it can be represented by a discrete approximation in the form of a filter bank. Secker-Walker and Searle<sup>14</sup> suggested that the asymmetry of the cochlear filter characteristics—low slope (6 to 12 dB per octave) below the CF and steep slope (50 to 500 dB/octave) above the CF—is important in representing the formant structure of voiced speech. The gentle roll-off on the low-frequency side results in a broad response characteristic, so that a group of fibers responds to each formant. The sharp cutoff on the high-frequency side prevents each fiber group representing a formant from responding to the next-higher formant. Hence, much effort went into obtaining these precise properties.

To obtain the desired asymmetric frequency responses, we designed a basilar membrane filter using on-



**Figure 11.** Middle ear circuit filter: design (top) and measured response from a test chip (bottom).

ly first-order low-pass and second-order band-pass building blocks (see Fig. 12). The frequency response of the overall basilar membrane, the driving function of the hair cells, is obtained by a bank of second-order band-pass filters with varying center frequencies. The sharp high-frequency cutoff is obtained by concatenation of low-pass filters whose cutoff frequency is matched to the band-pass frequency.

A basilar membrane filter bank of thirty stages was implemented in a small test chip. The whole bank was tuned through two uniform passive-resistive lines and four bias voltages to ensure a constant-ratio decrease of center frequencies from one stage to the next. The measured frequency response from the test chip is shown in Figure 13A, which has a shape similar to those seen in Figures 3B and 4.

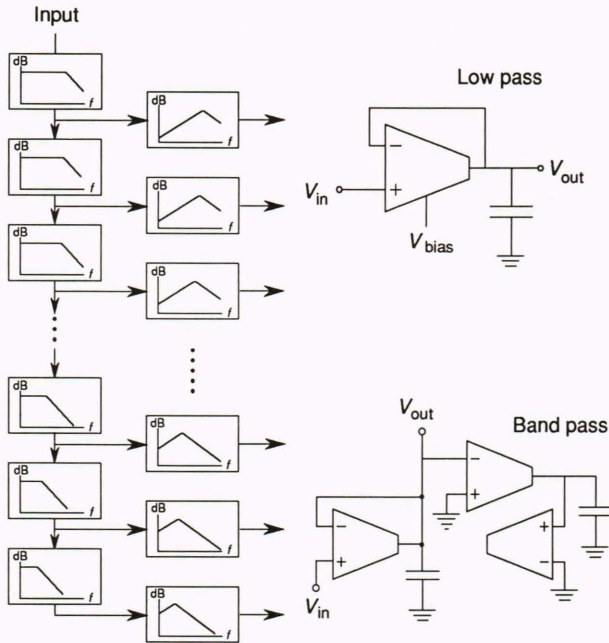
As discussed earlier, the hair cells and synapses play an important role in the functioning of the cochlea. Besides their compressive nonlinearity in electromechanical transduction, they also seem to have rapid and short-term additive adaptation properties, as seen in the discharge patterns of auditory nerve fibers. The realization of an inner hair cell/synapse model is not only challenging but also crucial to the performance of the speech processor. A reservoir model, shown schematically in top view in Figure 14, was proposed by Smith and Brachman<sup>17</sup> to characterize the end-to-end functioning of inner hair cells, from input excitation by motion of the basilar membrane to instantaneous firing rates of the auditory nerve fibers they drive (see the boxed insert).

A crucial nonlinear component in the model is the analog switch, which models the opening and closing of synaptic ionic channels. Hundreds of these switches are required to model a single hair cell in the reservoir mod-

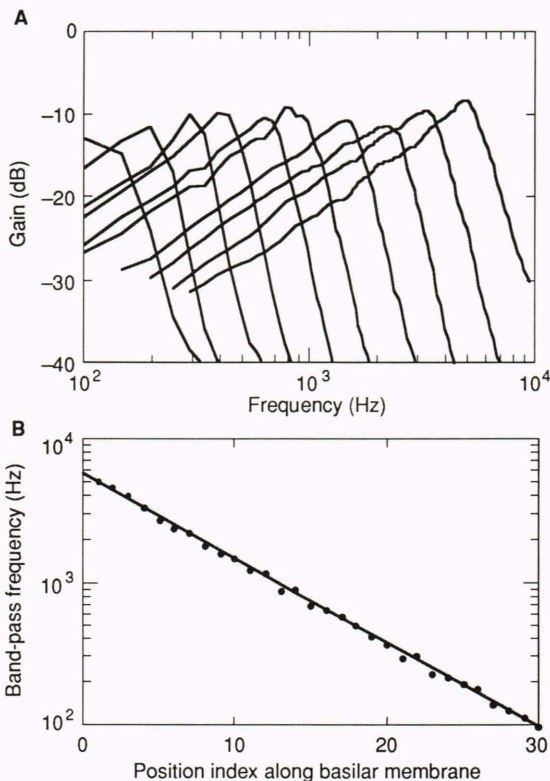
el. A simple and efficient way to implement an analog switch is to use a single MOS transistor, with its source and drain as the switch terminals and the gate as the

#### RESERVOIR MODEL FOR THE FUNCTIONING OF THE HAIR CELL

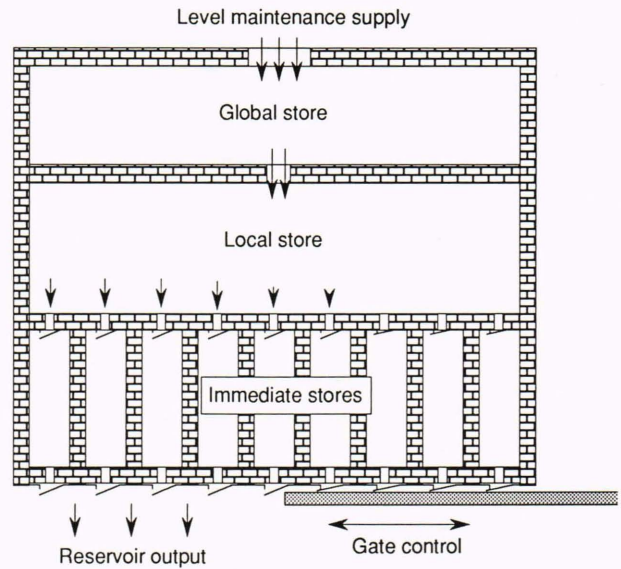
The reservoir model attempts to model the movement within hair cells of the store of ionic material that acts as the neurotransmitter for communication through the synapses. Changes in hair cell receptor potential caused by the motion of the basilar membrane control the opening and closing of immediate-store output gates. The instantaneous firing rate of auditory nerve fibers is modeled to be proportional to the volumetric rate of neurotransmitter release from the immediate stores. The transmitter substance in the global store is maintained at a constant level, which results in a potential difference between it and the local store, so that a partially depleted local store can be refilled. The local store in turn tries to replenish the immediate stores that have been discharged by the opening of output gates. The input gate into an immediate store is open only when the level of the local store is higher; it is closed otherwise, so that the immediate store will maintain its level when inactivated. This highly nonlinear model exhibits the desired additive, rapid, and short-term adaptation and depletion characteristics. The permeability of the gates and sizes of the stores determine the adaptation time course of the model. The operation of the reservoir model can be described in detail by a set of differential equations for which there is an equivalent circuit, shown in Figure 15. The instantaneous reservoir output of the model is the current labeled  $I_{out}(t)$ .



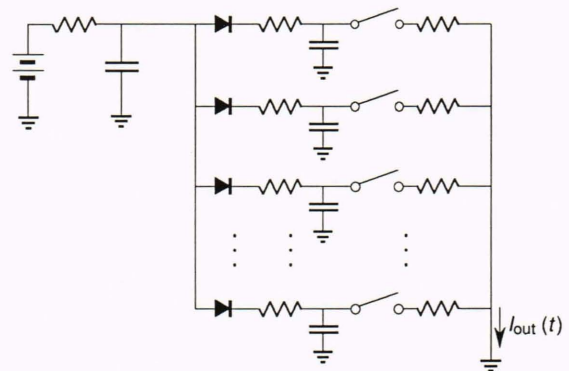
**Figure 12.** Cascaded low-pass and band-pass filters on the left form a lumped-element model of the basilar membrane response. The implementation of the filters is shown on the right. A logarithmic change in center frequency ( $f$ ) along the bank is obtained by linearly varying the bias voltages ( $V$ ) controlling the transconductances of the amplifiers.



**Figure 13.** Results from a test chip for the basilar membrane model shown in Figure 12. **A.** Response of selected outputs along the bank. **B.** Observed center frequency as a function of position along the simulated basilar membrane.



**Figure 14.** Top view of a reservoir model to simulate the hair cell/synapse (see boxed insert). The reservoir models the neurotransmitter substance within hair cells that causes activation of attached auditory fibers.<sup>17</sup>



**Figure 15.** Equivalent circuit for the hair cell reservoir model. The neurotransmitter is represented by electronic charge, and the reservoirs are now capacitors. Diodes and switches model the input and output gates of the immediate stores.

switching control. Small transistors (usually  $2 \mu\text{m} \times 3 \mu\text{m}$ ) are used to minimize the charge injection problem (a common difficulty in switching circuits, where some charge is coupled onto the circuit during switching through parasitic capacitances). The analog switches, each gating one immediate store, are controlled by comparators and a potential divider that simulates the effect of receptor potential on fiber firings.

A monolithic hair cell/synapse circuit is shown in Figure 15; 128 immediate stores implemented with  $2\text{-}\mu\text{m}$  complementary metallic-oxide-on-silicon technology occupied an area of about  $1.4 \text{ mm}^2$  on silicon. In testing the circuit, continuous tones of three different frequencies across a 50-dB amplitude range were used as input. The circuit output was compared with the response of auditory nerve fibers in cats. Figure 16A shows the period

histograms of auditory nerve fibers in cats in response to sinusoidal stimuli of various intensity levels.<sup>18</sup> In comparison, Figure 16B shows the simulated instantaneous firing rates generated by the test circuit during each input period; very good agreement is seen.

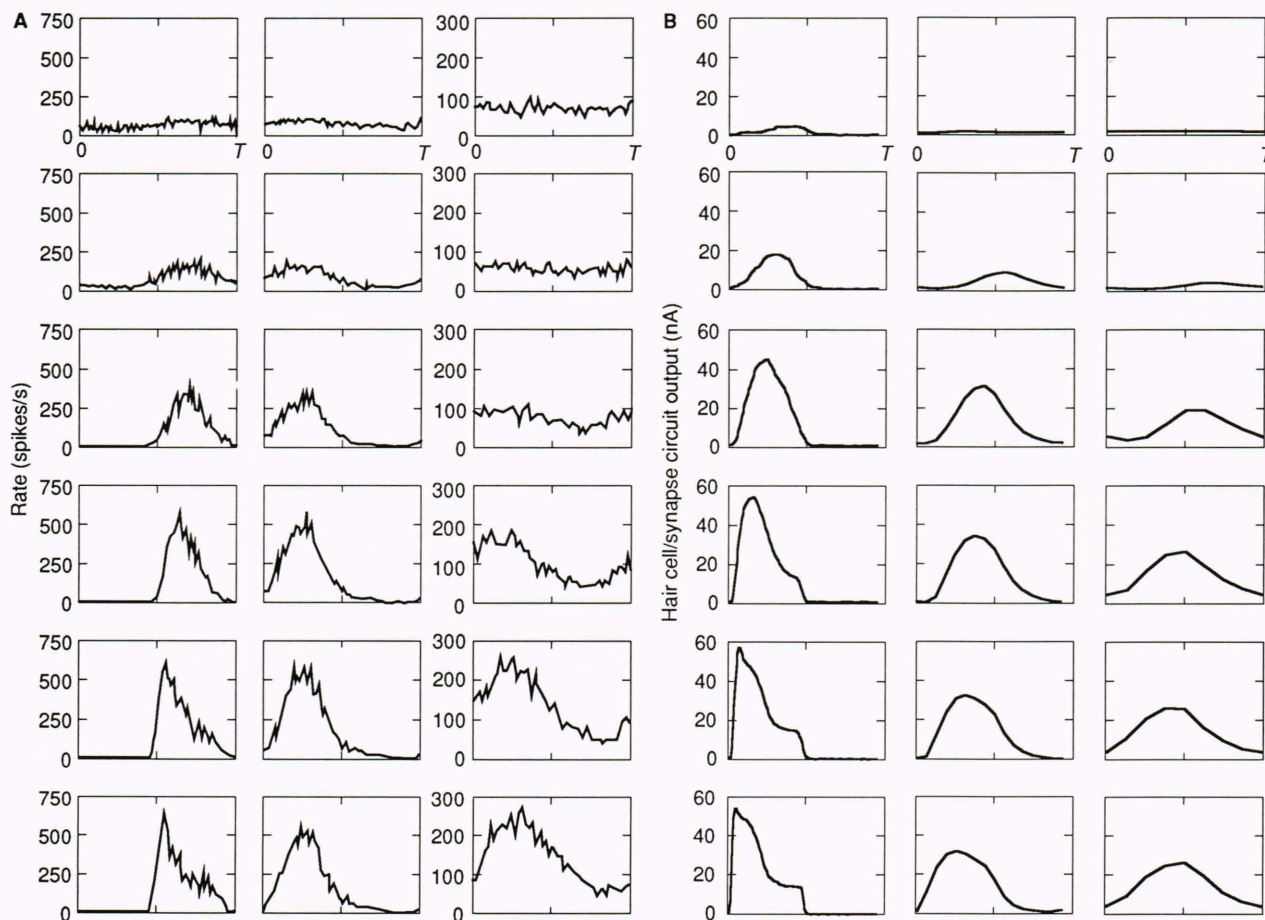
### Neurogram Output by the Silicon Model

Since the auditory periphery is a nonlinear system, its response to complex speech sounds cannot be predicted from responses to simple stimuli such as tones and clicks. Again, the neurogram displays the space-time pattern of cochlear nerve activities in response to speech. Neurograms are derived from experimental recordings of the firing of individual fibers by averaging response activity over repeated applications of speech stimuli. The desired output of our processor is a spatiotemporal pattern similar to the neurogram. The parallel outputs of the bank of hair cell/synapse units, each representing an ensemble average of instantaneous firing by a group of auditory nerve fibers, provide this output. This parallel set of analog signals is derived in real time, without the averaging that would be required if individual nerve fiber firings were simulated in the silicon ear.

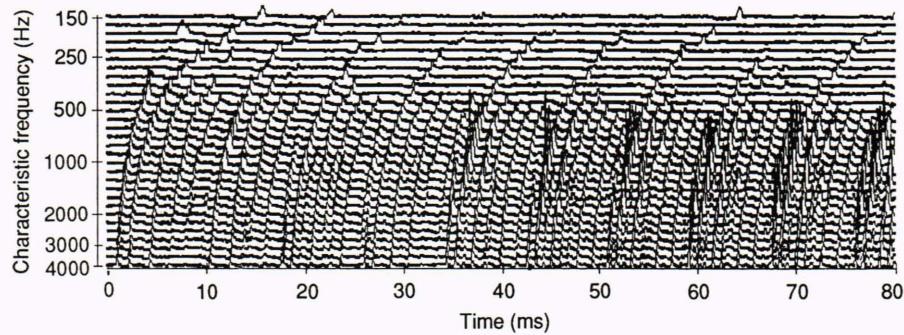
Figure 17 shows a neurogram generated in our first end-to-end test. The individual output traces were generated by reconnecting a single hair cell test circuit to thirty basilar membrane modules one by one to simulate a parallel bank of cells. The input was the syllable /ba/ generated by a Klatt synthesizer.<sup>19</sup> Enough similarity to the neurogram of a real ear exists to justify optimism that the HEEAR project is on track.

### CONCLUSION

In this work we are attempting a computing paradigm whereby computations are mapped onto silicon so that the dynamics of the underlying circuits closely matches the computational algorithm. The approach is in stark contrast to the traditional approach of software modeling on digital machines. With such an approach, however, one can obtain truly massively parallel operation at current levels close to those of biological processing. A successful silicon model of the auditory periphery, that is, one that accurately emulates its human counterpart, is the very beginning of an information processor approaching the incredible processing power of the human sensory system.



**Figure 16.** Period histograms of auditory nerve activity in cats in response to six different intensity levels at three different pure-tone frequencies. The three columns represent responses at 213, 1815, and 3409 Hz. As you move down each column, intensity increases 10 dB per curve. **A.** Experimentally observed results. **B.** Hair cell test circuit output for the same input frequencies in A. ( $T$  = period of the tone.)



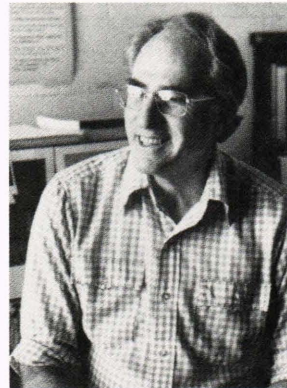
**Figure 17.** Neurogram output of the silicon hair cell in response to the synthesized syllable /ba/ (which lasts about 80 ms). The similarity to Figure 9 is striking and hence quite encouraging.

Much difficult engineering is required for success. As in a related Hopkins project involving a silicon retina, we are in the middle of the learning process. Future work will focus on an improved inner hair cell design and include the effects of the outer hair cells. In addition to trying a new approach to sensory information processing, the project is driven by the faith that such a chip will eventually find use as a preprocessor for a real-time speech recognition system or as an aid to the deaf.

## REFERENCES

- <sup>1</sup> Bruce, R., *Alexander Graham Bell and the Conquest of Solitude*, Little, Brown, Boston, Mass. (1973).
- <sup>2</sup> Dallos, P., *The Auditory Periphery: Biophysics and Physiology*, Academic Press, New York (1973).
- <sup>3</sup> Goldstein, M. H., "The Auditory Periphery," in *Medical Physiology*, Mountcastle, V. B. (ed.), C.V. Mosby, St. Louis (1960).
- <sup>4</sup> von Békésy, G., *Experiments in Hearing*, Weber, E. G. (ed.), McGraw-Hill, New York (1960).
- <sup>5</sup> Spoendlin, H., *The Organization of the Cochlear Receptor*, Karger, Basel (1966).
- <sup>6</sup> Rhode, W. S., "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys Using the Mössbauer Technique," *J. Acoust. Soc. Am.* **49**(4), 1218-1231 (1971).
- <sup>7</sup> Brownell, W. E., "Outer Hair Cell Electromotility and Otoacoustic Emissions," *Ear Hear.* **11**(2), 82-92 (1990).
- <sup>8</sup> Flock, A., "Structure of the Macula Utriculi with Special Reference to the Directional Interplay of Sensory Responses as Revealed by Morphological Polarization," *J. Cell Biol.* **22**, 413-431 (1964).
- <sup>9</sup> Sachs, M. B., and Young, E. D., "Encoding of Steady State Vowels in the Auditory Nerve: Representation in Terms of Discharge Rate," *J. Acoust. Soc. Am.* **66**, 858-875 (1979).
- <sup>10</sup> Young, E. D., and Sachs, M. B., "Representation of Steady-State Vowels in the Temporal Aspects of the Discharge Patterns of Populations of Auditory Nerve Fibers," *J. Acoust. Soc. Am.* **66**, 1381-1403 (1979).
- <sup>11</sup> Payton, K. L., "Vowel Processing by a Model of the Auditory Periphery: A Comparison of Eighth-Nerve Responses," *J. Acoust. Soc. Am.* **83**, 145-162 (1988).
- <sup>12</sup> Sachs, M. B., Voight, H. F., and Young, E. D., "Auditory Nerve Representation of Vowels on Background Noise," *J. Acoust. Soc. Am.* **50**, 27-45 (1983).
- <sup>13</sup> Miller, M. I., and Sachs, M. B., "Representation of Stop Consonants in the Discharge Patterns of Auditory Nerve Fibers," *J. Acoust. Soc. Am.* **74**, 502-517 (1983).
- <sup>14</sup> Secker-Walker, H. E., and Searle, C. L., "Time-Domain Analysis of Auditory-Nerve-Fiber Firing Rates," *J. Acoust. Soc. Am.* **88**, 1427-1436 (1990).
- <sup>15</sup> Payton, K. L., *Vowel Processing by a Model of the Auditory Periphery*, Doctoral dissertation, The Johns Hopkins University (1986).
- <sup>16</sup> Mead, C., *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, Mass. (1989).
- <sup>17</sup> Smith, R. L., and Brachman, M. L., "Adaptation in Auditory Nerve Fibers: A Revised Model," *Biol. Cybernetics* **44**, 107-120 (1982).
- <sup>18</sup> Johnson, D. H., *Responses of Single Auditory Nerve Fibers in the Cat to Single Tones: Synchrony and Average Response Rate*, Doctoral thesis, Massachusetts Institute of Technology (1974).
- <sup>19</sup> Klatt, D. H., "Review of Text-to-Speech Conversion in English," *J. Acoust. Soc. Am.* **82**(3), 737-793 (1987).

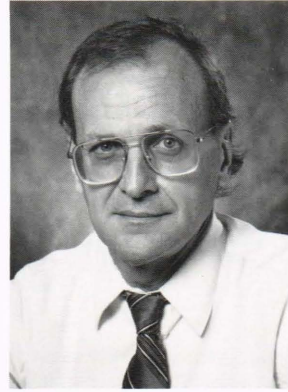
## THE AUTHORS



MOISE H. GOLDSTEIN, Jr., received a B.S. degree from Tulane University in 1949, and M.S. and Dr.Sci. degrees from MIT in 1951 and 1957, respectively, all in electrical engineering. He was a faculty member at MIT from 1955 until 1963, before coming to The Johns Hopkins University. He is Professor of Electrical Engineering and has a joint appointment in the Biomedical Engineering Department at the School of Medicine. The focus of his activities is basic research into speech processing and the development of devices to aid profoundly deaf children.



WEIMIN LIU received a B.S. degree in computer science and engineering from the Beijing Institute of Technology in 1982 and an M.S. degree from the Chinese Academy of Sciences in 1985. In 1990 he received an M.S.E. degree in electrical and computer engineering from The Johns Hopkins University, where he is currently working toward a Ph.D. degree. From 1985 to 1987, he worked on projects in speech processing and recognition as an Assistant Research Fellow at the Institute of Automation, Chinese Academy of Sciences. His research interests include speech signal processing and analog VLSI.



ROBERT E. JENKINS received an M.S. degree in physics from The University of Maryland in 1965. He joined APL in 1961 and is supervisor of the Computer Science and Technology Group, an elected member of the APL Advisory Board, a member of APL's IRAD Committee, program manager for the Space Department's IRAD, a member of the Electrical Engineering Program Committee for JHU's Whiting School of Engineering, and a lecturer in electrical engineering at the Homewood Campus and the APL Education Center. In 1985, he was awarded the Dunning Professorship at the Homewood Campus, where he introduced a new course in VLSI design and conducted research in cellular automation processing.