

Sensory Apparatus Cast In Silicon By VLSI Microchip Innovator

Carver Mead's Bottom-Up Analysis
Clears Path To Realtime Networks

By tackling the nervous system from the ground up, this microchip whiz is laying the foundation for the future. From the sensory apparatus on up to the upper cortex, Mead estimates it will take 10 years to decipher all the layers between the eye, or ear, and the brain.



CARVER MEAD IS VIRTUALLY THE founder of very-large-scale integration techniques for digital microchip production. As co-inventor of the silicon compiler — a software system for producing digital microchips directly from high-level descriptions rather than with manual lithography — and the inventor of the gallium-arsenide transistor, Mead has already made a name that will forever be remembered in annals of electronic achievements.

The California Institute of Technology professor is not one to be content resting on his laurels. His newest “life’s work” demonstrates that fact. Mead has turned away from the digital techniques he has already perfected, toward analog microchips based of the biological metaphor. As chairman of Synaptics Inc., a neural-network

microchip maker, Mead has joined with the 4004 and Z-80 microprocessor innovator Federico Faggin who heads that effort..

Simulations Don’t Cut It

For Mead the rapid pace of technological advancement has produced distinct advantages for those setting out to build systems based on biological metaphors. Because of advanced digital semiconductor technology that he largely created, there is an abundance of computing power available to run neural network simulations. But for Mead, simulations will never stand-up to reality, and it is “real data from the real world that systems must ultimately address.”

No matter how successful digital computer simulations of neural systems become, they will

never be able to deal with the flood of real data in real time. The main problem, according to Mead, is the discrete nature of the computer itself. Computers must digitize data into discrete bundles and file it away in memory, a process that Mead contends strips away the the realtime nature of dynamic systems.

Because of the loss inherent in the digitization process, the behavior of a computer simulations cannot truly mirror the functionality of real biological systems. For Mead, substituting simulations for the real thing obscures the fundamental issue of neural-network dynamics beyond recognition. And ultimately all mental processes arise from dynamic systems of neurons.

What we are aware of in daily life is only a small part of what

there is. "There is a vast neural iceberg beneath the cognitive tip of conscious thought," Mead explained. Consciousness arose from an enormously complicated system that operates according to radically different rules than do computers. And for Mead, what makes a computer good at general purpose computing is precisely what makes it poor at imitating the brain.

Even parallel processors with thousands of nodes fail to make even a reasonable approximation to the brain, because they parcel out tasks in the wrong size and have little hope of achieving the communication bandwidth that is appropriate to simulating the brain. The synapse is one of the most complicated elements of living neurons, but still it is just too simple to be simulated by the powerful algorithms available on a parallel processor's node.

According to Mead, computers introduce a spurious dilemma as a result of their computational superiority. They make it necessary to explicitly program the tradeoffs between precision, time, and resolution.

But natural systems use the physical parameters of their components to automatically define those tradeoffs. For example, neurons sum together their inputs as result of Kirchhoff's law. And other physical parameters are set by devices that are exponential in nature, which

results in a wide dynamic range.

What the analog nature of neural systems give away by operating at lower precision, they regain many fold by not aliasing away real time events as do digitizing computers. Their continuous mapping of functions into physical locations make the

**"There is a
vast neural iceberg
beneath
the cognitive tip
of conscious thought."**

functions which are needed also be nearby in physical location, virtually eliminating the communication bottlenecks that plague computer systems which use global connections, such as a bus or backplane.

The Answer — Microchips

The answer to the problems of simulating neural systems with digital hardware, according to Mead, is to cast those analog functions in discrete silicon technology. But this cannot be done in one fell swoop, because there are many levels of dynamic information processing between sensation and the high-level cognitive functions of which we are consciously aware. Many of these levels are prewired at birth, but many other sensory pathways develop from the kind of objects

experienced throughout life, especially during early development.

When an animal is born, a "blooming, buzzing confusion" of data, in William James words, assaults the senses. Evolution has come up with ways to organize that data, but these systems took a long time to develop and proceeded from the simple to the complex. But many system developers ignore the basic fact that evolution proceeded from the bottom up, by attempting to build systems from the top down. The problem, according to Mead is that "no one knows where the top is." Neurologists have acquired some knowledge about the way sensory organs represent information from the bottom up. But the higher levels that are the domain of psychology have almost no connection with our knowledge about these underlying operations. Some of the major pathways have been mapped out and some of the grosser aspects of information transformations are the subject of speculation, but science is a long way from understanding the brain from the top down.

Traditional artificial intelligence (AI) used this top-down approach only to fall into a trap. "The trap was set by grossly underestimating the amount of information processing between the bottom and the top." The correct way to proceed, if one desires to imitate

the brain, is to start at the bottom by building something that can interpret real sensory data. It is easy to build an image recognition system that recognizes what the researchers expected it to see in the first place. But it is extremely difficult to expose a system to real sensory data and then discover what it is seeing.

The AI Trap:

“You announce that you are going to do a really hard problem, then you start working on it and discover that it is orders of magnitude harder than you ever imagined. So you do what every good scientist does, you make up a toy example, something that you think you can solve. The principle is to keep simplifying the problem until it goes away...

So you solve your simplified version, and then make a little demo of that. You can make it look really good, because we have all these powerful computers around that can do wonderful simulations. You then announce to the press that you have solved the problem. But you have to be careful not to reveal what the really hard parts were. And when you look at them, you can't believe that they are really that hard, and you feel really stupid. Then you go back to step one of an even more difficult problem.”

Mead is afraid that the neural-network field is going to succumb to the same trap and “that is the fastest way for a budding field with a lot of promise and interdisciplinary interaction to go under.” The way to avoid the AI Trap is

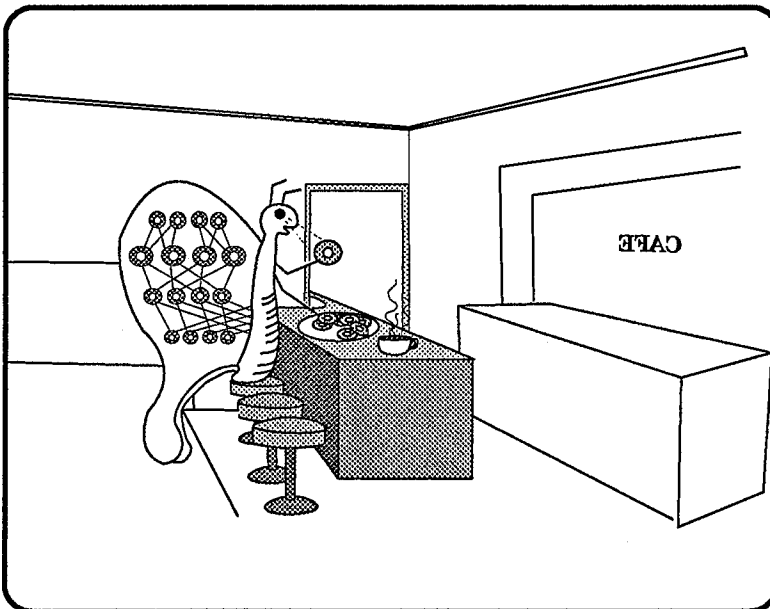
to make getting real data the number one priority. A good start in this direction was made in 1952 when Egon L obner, then at RCA, built an artificial retina that worked from real data. Its only major problem was size. Because it had to use non-solid-state technology it was housed in a module four feet by four feet by three feet. It was a wonderful technological start, according to Mead, but the implementation was just too cumbersome.

But with the advent of the integrated circuits all that has changed. Even with current technologies there is room for yet another factor of 100 in speed enhancement and size reduction before the limits of the physical devices must be circumvented with new technologies.

Just as nature evolved a system to handle sensory data in real time, the neural network industry, according to Mead, is also evolving a technology. The constraints of the technologies being used will set the parameters of the tradeoffs between precision, time and resolution, as do living systems. By learning that lesson from biology, analog technologies can be evolved along similar lines. By studying the constraints of biological systems, its functions can be cast in microchip technology.

Technical Evolution

One of the major constraints used by so-called “wetware” are its constraints on local wiring



The senses shape the character of cognition itself by converting raw sensation into an internal representation that can be called upon to recognize similar instances.

schemes. The complexity of circuits, whether they are cast in silicon or in the wetware of the brain, depends upon a close match to wiring. Just as there is not enough information in DNA to specify the multitude of connections in the brain, there are not enough programmers in the world to explicitly specify all wiring topologies needed for neural networks.

The nervous system uses a wiring geometry that is slightly greater than two-dimensional, since opening up all the folds and creases in the cortex results in just a square meter of area that is about one millimeter thick. Silicon technologies already have a similar slightly greater than two-dimensional constraint. But because both systems are basically two-dimensional, the major problem is also the same — namely the cost of its wiring. Highly local wiring schemes in the brain must be copied by silicon versions. That means that processing must take place within a context that is mapped locally, just as in the nervous system.

Bad News & Good News

According to Mead “the bad news” is that designers are going to have to consider the constraints of their medium in a way that they have only vaguely considered up to now, particularly with respect to the locality issue. Systems that learn from experience must have an underlying

structure that enables that learning to take place. Genetic codes perform that function in wetware, but wiring and the architecture on which a problem set is mapped, does it in hardware.

Systems must also be self adjusting, “normalization is the first order of business,” both for

Designers must consider the constraints of their medium in a way only vaguely considered up to now.

voltage levels and for gain. Every part of the nervous system is self adjusting, according to Mead, in order that it might accept a wide dynamic range of inputs. The other major characteristic of the nervous system is its ability to perform time-domain processing without a global system clock.

So much for the bad news. “The good news is that the result will be worth the effort.”

Today chips can be built with about 10^8 th synapses on a wafer and soon as many as 10^{10} th will be feasible. Other niceties are that the analog device's processing primitives are well suited to neural information processing, such as exponential and hyperbolic functions — just the functions that are difficult for digital technology. The analog represen-

tation is not as precise as digital, but it can handle current and time resolutions of over seven orders of magnitude. Additionally, wafer-scale integration is easier with analog design, because of its low power consumption, a result of operating devices below their threshold voltage. An artificial retina, for instance, consumes power down in the microwatt range, since the technology is inherently current limited. Also connections can be easily time multiplexed, like television bands. Optical devices like photodetectors are also easy to built into silicon.

But the good news may be “hard to swallow” because designers will have to go back to developing a technology from the ground up. It is easier to just simulate neural systems with highly developed digital technologies. But the electronics community is resilient. It is also chocked full of people who hunger for new ideas to start wiring up. As a result, “I believe that ten years from now, there are going to be a whole lot of neural networks in analog silicon.”

First Cognizer Chip

Mead's first excursion into analog silicon was to improve upon a colleague's circuitry at CalTech, John Hopfield's associative memory. He fabricated his associative-memory chip based on Hopfield's theory, but he used several clever analog design

techniques to improve upon it. With his intricate understanding of processing technology, Mead was able to represent the trickier aspects of the Hopfield's neuron model with standard fabrication techniques.

Mead's (addition) to the architecture borrowed from Hopfield was to use alterable connections to his artificial neurons, so that the device was programmable rather than fixed in function like a read-only memory (ROM.) Hopfield's microchips had to be programmed once-and-for-all with masked resistors. The programmability problem was tricky, because the system had to be able to represent both the positive and negative resistances that represent excitatory and inhibitory connections found on real neurons. Positive resistances are difficult enough to build in silicon, but negative resistances are impossible to build. Mead's insight was to use a dual-rail voltage technique to represent this problem elegantly. He used the constraints of the medium to an advantage instead of using brute-force techniques to surmount the problem, as is common for digital designers.

Mead and his students at Cal-Tech built a programmable associative memory consisting of 22 artificial neurons with 462 connections. The circuit was fabricated in 1984 by the MOSIS service, run by the Defense Advance Research Projects Agency (DARPA). It used an standard

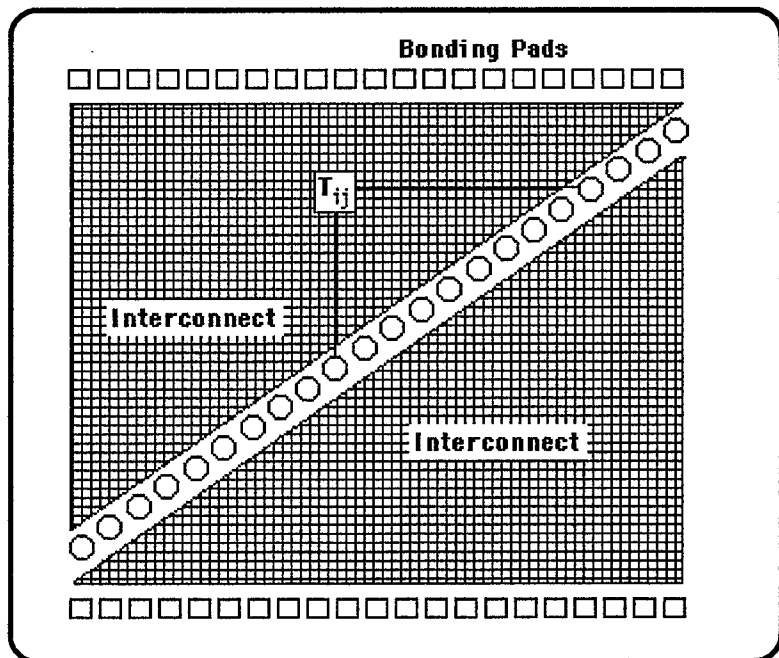
silicon process with relatively relaxed four-micron minimum design features. Amplifiers were used to model the neurons and were placed along the diagonal of a 22-by-22 cell array.

Unlike Hopfield's chips, Mead's implementation was fully programmable for different purposes. Mead's chip was necessarily more complicated, though, since each cell needed over 40 transistors plus a double set of wires for the dual-rail supply voltages. The chips exhibited the remarkable "fail-soft" property characteristic of brains. Though brain cells die every minute, one's conscious mind is not effected in the slightest. Similarly Mead's chips con-

tinued to store two 22-bit vectors and remember them by association even with component failure rates of over fifty percent.

Silicon Eyes And Ears

But Mead's first excursion into neural networks, with his associative memory chip, was just to get his feet wet. Since then his commitment to building systems from the bottom up has turned his interests to building silicon versions of human sensory apparatus. The senses are the first step in a hierarchy of information processing that stretches from the outer periphery of the body up to the highest level of the cerebral cortex. While most other neural



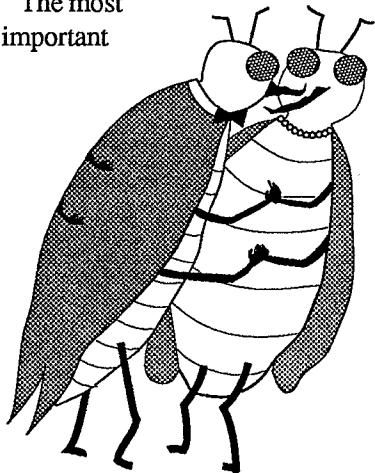
Mead's earliest attempt at crafting neural networks from analog silicon wafers was a programmable version of professor John Hopfield's associative memory.

network researchers are *starting* at the cortex, Mead is starting with the senses — silicon senses.

For Mead, the most amazing thing about the nervous system is its ability to cope with the “blooming, buzzing confusion” of sensory information that would swamp any known digital system. “There is no time to swap data out to disk,” Mead mused, “the nervous system simply always processes information at an enormous rate.”

To learn how that system works, Mead is building sensors that work in real time and which employ the principles used by the human visual and the auditory systems. He claims that his systems are not as good as the systems in real animals, but that the very act of building them is revealing how neural systems work in real time.

The most important



Mead used “bugs,” usually avoided parasitic transistors, to act as photodetectors. Their small size enables many more to be fabricated on a single microchip.

thing that is being revealed is how information is represented in the body. “When it is done, the entire neural network business is going to be a business about representations, make no mistake about it,” Mead contends. He maintains that this simple fact has been true of computer science since its beginning, that is, the proper data structure is half the job. He claims that it is just as true of the nervous system. “Getting the representations right is what is going to make the business fly or die.”

The reason that one must start from the bottom, is that the bottom is where the necessary invariants get built into the internal representation of information. And achieving the proper representation will allow systems to recognize the objects of vision or speech as people do. For speech, the proper invariants will allow recognition systems to be built “without having to do all the crazy stuff we do today to get it all tweaked up for the particular speaker and so forth.”

Mead has pledged his next ten years of research to getting the representation right so that those who are building object recognition systems will have a representation on which recognition will be worth doing.

The Visual System

Carver Mead’s first attempt at building silicon senses was to model the human eye. The human retina uses an array of neurons

lining back of the eye to extract features and condense a vast amount of raw sensory data. By studying the real retina, Mead and graduate student M.A. Mahowald were able to mimic its operation with a photo-detection microchip that possessed built-in invariants similar to those of the biological eye. A tiny photo detector array that was sensitive to motion, an invariant built into the human retina, was used to recognize a moving object directly without any computerized image processing. That operation would have required a supercomputer to compute directly from a dumb sensor like a television camera.

By modeling the basic operations of the eye, Mead’s microchip combined both sensing and image processing onto a single 6mm by 8mm microchip. The silicon device held 2304 photo-sensitive receptors together with their associated processing elements, each of which measured only 100 microns by 125 microns. The circuit was built entirely from transistors.

A feedback loop using competitive shunting networks damped the response of the photo-detector so that the microchip did not respond to absolute light intensities, but to ratios of available light. The main detection system registered the rate that the intensity of light changed, in effect deriving the derivative of the light intensity in its neigh-

borhood.

In contrast, conventional image processing takes snapshots of visual scenes at discrete time intervals and then attempts to correlate points among these frames in order to discover the boundaries of objects and their movements. That operation takes a supercomputer to handle the extremely complex mathematical correlations required. Mead showed that the eye is not doing correlations at all, rather it handles light variations as a continuous stream.

Mead used the tiny delays among his feedback loops to provide the time it takes to derive a rate of change in the intensity at the point of detection. Thus rather than perform complex correlation calculations, the eye senses motion directly.

The digitization of visual images by computers is, for Mead, the largest obstacle to their success at recognizing objects. Rather than discrete sequences of time steps, his microchips use chains of dynamic timing loops to set the parameters of precision, time and resolution so that the proper invariants make the recognition process easy.

Mead used bipolar parasitic transistors that are usually avoided in standard microchip processing for his photo-detectors. When photons struck the surface of his microchip they absorbed it to produce an electron-

hole pair. That generated a current which was amplified by the operation of the transistors. This effect, usually unwanted by microchip fabricators, was used by Mead to produce his very effective photodetector array.

The output of this parasitic detector was then passed through

**The entire
neural network
business is going to be
about representations,
make no mistake
about it.**

two transistors and fed back into the device's input just as in the peripheral vision neurons of the retina. That feedback attuned the detector to motion. By replicating that simple circuit over the surface of the chip an exceedingly small sensor was formed that could readily detect the movement of images focused on the surface of the chip. The circuit also could distinguish foreground objects from background objects a feat that has been difficult for digital computers.

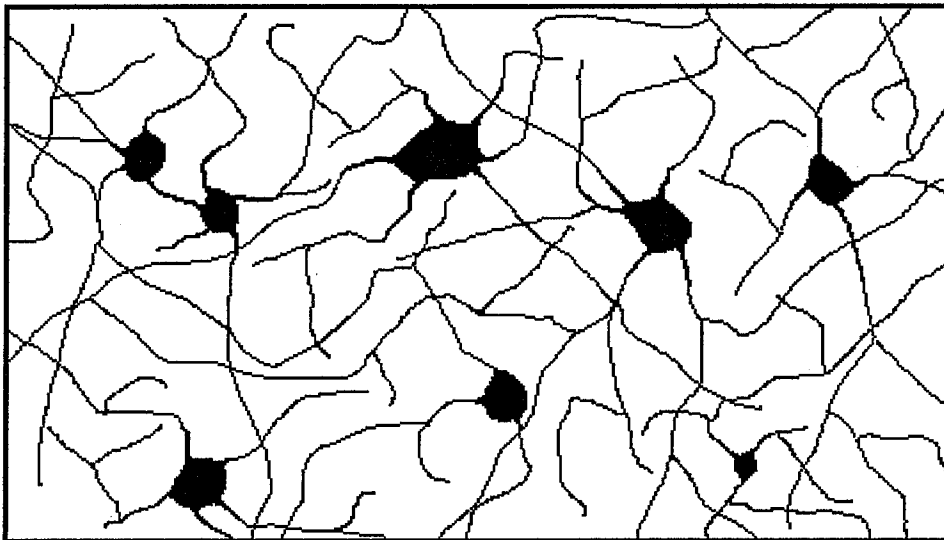
Having built the first level of his retina, Mead is moving on to model the next layer of the retina — the amicroine layer. Amicroine cells are inter-neurons, passive cells with an underlying layer of variable resistance connections.

The function of this layer, according to Mead, is to add temporal enhancement to the spatial derivative of the first layer. This second processing step will enhance the region around a point where movement is detected. Each photocell's response will modify the response of its neighbors in the array through a simulated amicroine layer. The two layers operating in concert will locate the median of the overall signal intensity allowing the system to scale its response to the average level of input.

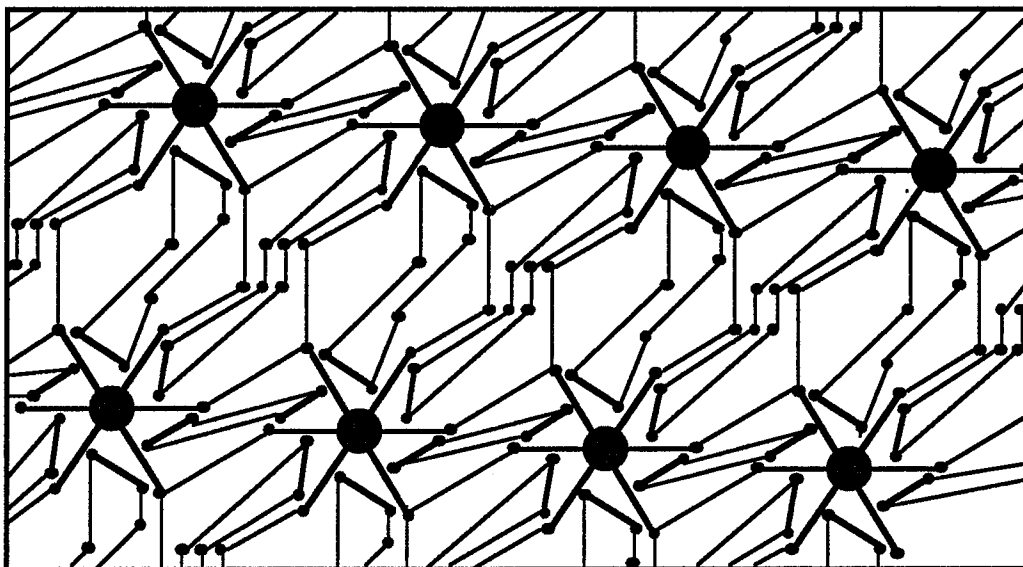
The amicroine system solves a difficult problem for any sensory system by scaling its response. Any signal processing system can be adversely affected if the median of the input signal is close to its upper or lower bounds.

But two obstacles stand in the way of modeling the amicroine layer with analog microchip technology. First the amicroine level is highly interconnected requiring an astronomical level of wiring density. Secondly it requires variable resistances. Mead has built variable resistances into his associative memory using a dual-rail voltage approach, but that took 40 transistors for a single connection. A more economical solution than that was needed.

To solve the wiring problem for the amicroine layer Mead has worked out a novel hexagonal wiring layout. Mead derived aspects of his approach from



Carver Mead makes every attempt to accurately portray biological systems in the circuits he fabricates. The biological "amacrine" cells of the eye, (top,) compare closely to his silicon version, (bottom.)



This artist's conception of Mead's wiring topology for the amacrine layer of neurons in the eye is actually quite close to the real chip. At Synaptics Inc., where Mead is chairman, several similar microchips are being developed.

seminal work done by Boston University's Stephen Grossberg. (See related story page 33.)

The Auditory System

Mead's second foray into sensory apparatus was to synthesize a silicon "ear." To do so, Mead went back to the anatomy books. His first discovery was that the auditory system evolved before there were speaking animals. Thus the auditory system could not have been designed to understand speech, rather it was "engineered" to localize and identify sounds. Mead discovered that an abundant amount of auditory research revealed several accepted facts regarding *what* the auditory system does to localize sounds, but an understanding of *how* it does it was yet to come.

Nevertheless, the mechanisms appeared to be well understood. Sounds appear to be localized separately in the horizontal and vertical planes. The most well known mechanism for how sounds are localized, according to Mead, is horizontally. In the horizontal plane localization is done with stereo cues from transients that come to one ear before the other, known as the interaural time-delay cue. The distance between the two ears horizontally yields a time delay of about 700 microseconds.

Another less well known, but equally important cue for horizontal localization, comes from the fact that the ear aimed toward

a sound gets more high frequencies than the one facing away from the sound. This is called the head-shadow cue. The head-shadow cue is a very powerful mechanism, often being as strong as the interaural time-delay cue. The interaural time-delay cue and the head-shadow cue together

**The auditory system
could not have been
designed
to understand speech,
rather it was
"engineered"
to localize
and identify sounds.**

yield localization in the horizontal plane.



Vertical localization of sound is more difficult, but comes about because there are two paths into the ear canal, one that goes directly to the ear drum and one that deflects sound off the lobes of the outer ear as it enters. The delay due to the deflected path depends upon the vertical angle from which the sound come. But the time delay cue for vertical is only about 70 microseconds, making it a weaker, but still a clearly perceptible, cue.

Modeling the Ear

In order to duplicate this localization machinery in silicon, Mead has built several chips

based on a detailed analysis of the ear's structural dynamics.

The transduction between the sound waves that come into the head and the electrical impulses that feed the brain occurs in the cochlea, a snail-shaped structure which is filled with fluid. It is divided into three chambers by membranes, the most important of which is called the basilar membrane. The fluid is incompressible so that pressure on one side of the membrane displaces the fluid and therefore the surface of the membrane.

The basilar membrane divides the cochlea along its length. Fluid goes down the length of the cochlea to its end where it is open to return along the other side. The springyness of the basilar membrane together with the mass of the fluid makes a traveling wave structure down which signals can propagate like an electronic delay line.

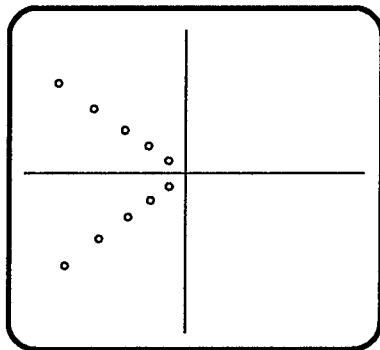
The stiffness of the basilar membrane is the key to its functionality. Its stiffness decreases by about two orders of magnitude over its length of about 3 or 4 cm. The velocity of propagation, which is inversely proportional to the mass and density of fluid and directly proportional to the stiffness of the membrane, also decreases along its length.

If one drives the membrane with a sine wave, a pressure difference is set up between each side of the membrane and that starts a traveling wave down its

length. The wave travels very rapidly at the beginning and then slows down due to the decreasing stiffness of the membrane.

An inevitable part of slowing down is that the energy per unit time is constant, but the energy density per unit distance gets larger. To compensate, the amplitude of the wave grows as the wavelength gets shorter. The membrane is very close to lossless until the waveform gets very short, at which point the second order properties of the membrane damp it out. That shearing loss in the membrane rolls off the high frequencies very rapidly.

By studying the membrane itself, Auditory researchers had found that the law by which the stiffness varies is exponentially decreasing with distance. That causes an exponential decrease in the velocity of the wave accounting for the subjective perception of sound as logarithmic. That is, if the frequency of a wave is a multiple of the original, then



A pole plot of the electronic delay line shows how it simulates the exponential change in thickness of the ear's membrane, (both axes are log scales.)

exactly the same wave is formed, it is just shifted over to the right or left. Mead contends that this is the very first step in a piece of representation that is taken for granted, the octave relationship.

According to Mead's analysis, hearing has a frequency invariance built into it. Harmonic frequencies are due to the very characteristics of the cochlea's factor-change in frequency which corresponds to a constant shift in position. The spatial pattern on the membrane for waves that are twice as fast are identical, just shifted over. If it is doubled again, then that wave is also identical, just moved over again by the same amount. Thus the octave relationship, that piece of invariance, is built into the aural representation at the very lowest level of mechanical transduction in the hearing system.

Basilar Membrane Details

After the sound pressure from the outside world is transformed into a traveling wave along the basilar membrane, four rows of hair cells above the fluid-filled cochlea sense the rate at which the membrane vibrates. The tips of these hair cells which are called cilia, move as small as an angstrom to change the firing rate of the neural cells to which they are attached. These neural cells then feed the auditory nerve, which goes up to the brain. If these hairs are ever dislodged, by loud noises (such as gunshots or

shrill guitar solos,) they do not grow back.

The outer three rows of these hair cells are fastened in a stationary position above the membrane to exert a force on it like a muscle. As the membrane rocks back and forth, the mechanical motion stimulates the force transduction mechanism of those cilia to apply a force back onto the membrane in such a direction that it lowers the mechanical damping of the system. It is a negative mechanical resistance, (positive feedback.)

There are enough hair cells that the mechanism can become so active as to cause a spontaneous mechanical oscillation (ringing in the ear) such as when one stretches in the morning.

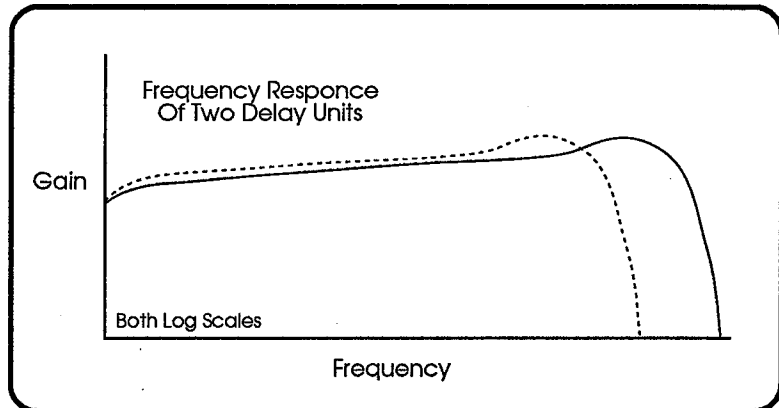
This active undamping system, from an engineering point of view, demonstrates that the ear is set up to hear transients. According to Mead, the ear is *not* set up like a series of band-pass filters, as nearly every other engineering model suggests. If it was, then whenever a transient came along it would set the bandpass filters ringing and the information in the transients would be lost. According to Mead, that is not what the cochlea is doing, rather the cochlea creates a traveling wave structure that preserves the transient nature of sound.

When the sound level is very low, the system prefers to go "boing," as Mead puts it, rather

than not hear a sound at all. Consequently, at very low sound levels there is no feedback to turn down the gain and the basilar membrane becomes very resonant. The bottom two orders of magnitude of hearing are dominated by this very resonant behavior where the ear is not as good at hearing transients, but it is extremely good at hearing sound. Because the bandwidth gets increasingly narrow, the noise does not go up as fast as the gain does, making it a very effective way of increasing the signal-to-noise ratio at very low sound levels.

The mechanical undamping is controlled by efferent fibers coming down from the brain to the cochlea. The efferent fibers are a very complex and sophisticated feedback system that can turn down the mechanical gain of the outer hair cells. When there is lots of sound information coming in, the automatic-gain signals are cognized in the olivary complex and feedback signals, which are very slow compared to the auditory signal, are sent back to turn down the gain.

The compressive nonlinearity in the transduction mechanism of the inner hair cells is approximately a square-root function. There are two different kinds of compressive nonlinearities in the automatic-gain control as well as some others further up that are not as well understood, Mead



Here the frequency response of just two out of hundreds of delay units is shown to be very similar. Each of the curves above roughly corresponds to an inner ear hair.

contends.

The frequency response of the basilar membrane over most of its working range is quite smooth and is not at all resonant without the undamping at very low sound levels. However, it does have a very very sharp cut off at high frequencies and slow roll-off on the low end.



Since the velocity of propagation is fast at the beginning and slow at the end, the poles of the filters are exponentially separated, starting with very high frequencies being very fast with the velocity of propagation then going down.

Chips Like The Ear

The chips to emulate these functions of the ear were built in conjunction with Richard F. Lyon of Schlumberger Labs. The first to be built was a traveling-wave structure that is a silicon analog of the cochlea. A transmis-

sion line was made out of transconductance amplifiers that output a current which is proportional to the voltage at their inputs.

Each delay element used two capacitors as the dynamic components, being charged by two amplifiers going forward and one going backward. The one going backward provided the positive feedback corresponding to the action of an outer hair cell by decreasing the damping of the traveling wave structure. The positive feedback was about half enough to make the system oscillate.

All the circuits were cast in metal-oxide semiconductor, (MOS,) and operated in the subthreshold range which means that the currents through the transistors was exponential relative to the voltage on their gates. In that region, Mead maintains that MOS transistors are perfectly

well behaved.

Because the current through the transistors was exponential with respect to the voltage on their gates, the transconductance of the amplifier was exponential with respect to the control voltage that sets the gain and therefore the transconductance of the amplifier. Accordingly the velocity of propagation was exponential in the voltage on the control input.

To harness that exponential characteristic in the service of simulating the cochlea, Mead laid down a resistive polysilicon line along the length of the delay line and put a voltage on one end that was lower than the voltage on the other end. That produced a linear gradient on the control voltage terminals of the amplifiers which was converted by the MOS transistors into an exponential gradient in the velocity of propagation of the signals down the line. This silicon "trick" models perfectly the exponential decrease in stiffness along the basilar membrane.

Thus Mead built a transmission line whose velocity of propagation varied over several orders of magnitude along the length of the line that was nevertheless perfectly well behaved. "That is hard to do with traditional technol-

ogy over orders of magnitude. With coils or capacitors that have to get bigger and bigger, three orders of magnitude would take a room full of stuff," Mead explained.

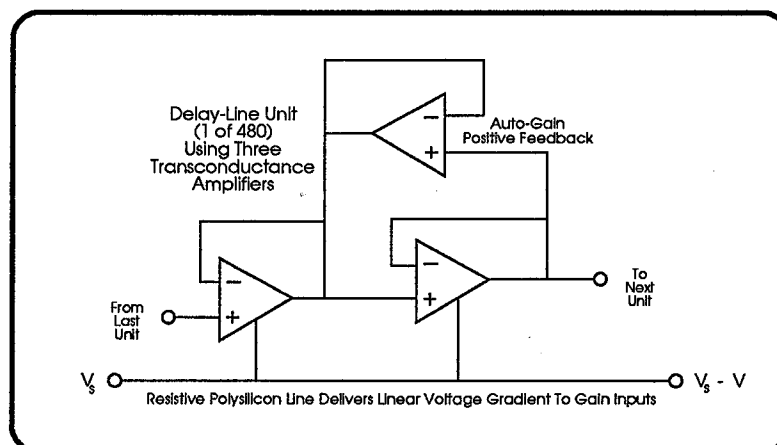
In terms of frequency response plots, the curve for each section of the delay line is only very slightly resonant and extremely broad. But there are lots of them, just like there are lots of hair cells, and they are spaced very closely. But to get the total response one must add up the response of all the curves. Though each curve has only a very slight bump when they are added up a very large gain results. And as one changes the height of the bump on each section a very small amount, the height of the composite changes a lot. "That is the same trick that the real cochlea is doing," Mead explained.

The membrane itself is not undamped very much, but because the signal propagates through a lot of sections each one of which

is only very slightly resonant, the result is a big increase in gain for a very small amount of positive feedback (that is the trick). There is more than a factor of 10 change in gain in the overall system due to that very small change in the resonant peak in each individual section.

Another result of using all those sections is that the high-frequency roll-off gets extremely steep, up to 200 db per octave. That piece of the hearing system's representation makes extremely good use of the fact that the amplitude of a given cochlea channel falls very rapidly with frequency. Namely, it explains how vibrato works on a stringed instrument: an imperceptible change in frequency is actually perceived as a change in amplitude, because that small change in frequency runs up and down the extremely steep curve of one part of the cochlea. Because the cochlea channels have

such a very sharp increase and decrease in amplitude, as one rocks their finger a little bit on the violin, the imperceptible change in frequency is interpreted as a change in amplitude — what musicians call *vibrato*. There are also other



Mead pulled another silicon trick out of the hat when he put a linear voltage gradient on the exponential control inputs to simulate the "log" characteristics of the ear.

perceptual artifacts that come from the fact that the cochlea is using this very very rapid roll off in frequency.

What Mead's detailed analysis of the cochlea makes clear, is that good frequency discrimination is not dependent upon narrow peaks, (bandpass filters), rather it depends on mechanisms with a very sharp roll-off. That is precisely what the auditory system uses.

On the chip Mead built, there are 20 sections per octave in frequency, each of which is about a third of an octave wide in its response. The peaks are broad so that there is considerable overlap among the delay elements. The difference in voltage at each end of the polysilicon line feeding the gain inputs of the transconductance amplifiers determines the peak frequencies over the range of the artificial cochlea. Two capacitors per unit provide the time delay for each of 480 delay sections. The DC gain in each section is slightly less than one.

Right Side Of The Law

Each MOS amplifier on Mead's chip has a gain of about 2000, but unfortunately the MOS fabrication line used had variations of up to 2X among transistors that were supposed to be identical. Nevertheless, this case is on the "right side of the law of large numbers," Mead said. It is possible to be on the wrong side of the law of large

numbers too, Mead elaborated. Being on the wrong side means that if any one thing will mess it up, you are bound to have one wrong. "But in this case it all averaged out better than it had any right to...it must be that the things that have survived in the nervous system were also on the

Luckily we were on the right side of the law of large numbers

right side of the law of large numbers."

Mead is ambivalent over whether nerve pulse trains or simple signal levels should be used as the representation for neural network information processing. To resolve those mixed feelings Mead set out to learn about neural pulse trains by actually building a synthetic hair cell. "That has turned out to be orders of magnitude more difficult than making the cochlea itself," Mead revealed.

In his preliminary studies, Mead concluded that real hair cells preserve "time synchrony" very well. In other words, hair cells fire in synchrony with the signal coming in "a little before top-dead-center, just like a good internal combustion engine." But unlike an engine, hair cells do not

fire on every cycle. Graduate student John Lazzaro and Mead set out to model that action by building synthetic hair cells on silicon microchips

Their primary objective was to model the characteristics of the auditory system that seemed to depend on the timing of the arrival of sound. "The nerve signals are digital in amplitude, but not digital in time...the time of arrival is the critical part of the representation," Mead elucidated.

Also for different amplitudes of input voltage, the hair cells appeared to skip more cycles for low voltages than for high ones. Hair cells also tended to saturate out while nearly always maintaining synchrony with the input waveform.

To model all those actions, Mead chose a silicon hair cell built with a differentiator driving a capacitor and using the current that goes into the capacitor to fire a neuron circuit with a threshold. Each time the signal exceeded the threshold it fired the neuron in synchrony with the input waveform. But because the current that went into the capacitor was the derivative of the waveform, whether it fired on any given cycle was stochastic, (random,) because it could not be known ahead of time how close the neuron was to its firing threshold.

The firing rate, as a function of the frequency, was found to mirror the original mechanics of

the cochlea — that is, the peak got sharpened because the circuit was taking the derivative. Over most of its operating range, the cochlea is not particularly resonant, but after taking the derivative it *appears* that way because the traveling-wave structure is feeding a differentiator.

Between Cochlea and Cortex

The auditory representation is a very old system. "Back when we were lizards, we didn't have a cortex and visual processing was done in the superior colliculi," Mead explains. Birds, in fact, still use that system. The lower parts of the systems between us and birds look the same. "This is an amazing case of convergent evolution," since similar cochlea were evolved independently in totally separate species.

From the cochlea, neural signals go to three different places in the cochlea nucleus, each of which have radically different anatomies. The output fibers from those locations in turn go to yet other places. The next stop after that is the olivary complex from which the automatic-gain control signals return down the olivo-cochlear bundle and back to the cochlea to turn down the gain on the outer hair cells. Then fibers go up to the inferior colliculi both from the olivary complex and the

top end of the colchlea nucleus. Then, surprisingly there are some fibers that go up to the superior colliculus which was traditionally thought to be a visual area. About 5% of the nerves in the superior colliculi are auditory. Therefore, "there must be some explicit coordination between eye and ear,"


It will take 10 years to map out the information processing steps between the ear and the cortex.

since when the eyes move the auditory map in the brain shifts around to keep it in sync with the visual scene. That must mean that it is extremely important that the representations of the spatial map be in registry with both visual and auditory signals. The outputs of the colliculi go up to the genicular body and on up to the cortex.

The point of enumerating all these levels of information processing is that speech understanding is done way up in the cortex, but the cochlea is setting up the representation way down below. Hence, systems that concentrate on the cortex, like most traditional neural networks, are bound

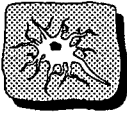
to get the representation wrong.

"It is a long way between the cochlea and the cortex. When we have learned all these stages of processing, then we will have enough information about the representations that we will have something worth doing speech recognition on," Mead contends. But there are a lot of stages yet to be explored and mapped out.

In contrast, the visual system was once totally confined to lower areas, but now it has moved almost exclusively up to the cortex, but the auditory system did not do that, "that tells us something very important... whatever the system did to be so very good at localization and identification of sound is the right front end for a speech understanding system, because it has persisted so long." For Mead that is why one must understand all the stages of processing so that the appropriate representation with the right number of invariants are built in from the start, "so that we don't have to train our speech understanding systems on every individual person and have the words isolated and all the other crazy stuff we have to do today," Mead predicts that it will take 10 years to map out all the information processing steps between the ear and the cortex. 

REFERENCES

- Mead, Carver, "Silicon Models of Neural Computation," *Proceedings of the IEEE First Annual International Conference On Neural Networks*, San Diego, Calif., June 21-24, 1987.
Mead, Carver, Plenary Remarks on the Function of the Ear, *IEEE Conf. on Neural Information Processing Systems — Natural and Synthetic*, Denver, Colo., Nov. 8-12, 1987.



Cognizer Co.

333 S. State St., Suite 141
Portland, Oregon 97034
(503) 246-6464 U.S.A.

JUN 14 1989

June 12, 1989

Carver Mead
CalTech
MS 2567-80
Pasadena, Calif. 91125

Dear Carver,

It was nice talking with you in Portland. Hope you enjoyed your stay at your ranch. We love Oregon too. (Lisa sends her regards.)

I wrote a story from your lecture at the Portland IEEE Circuits and Systems conference for *EE Times* and mentioned your new book (Incidentally, *Analog VLSI and Neural Systems*, is now on the top of my recommended list of books for serious neural enthusiasts. Good Job!)

I am currently updating the *Neural Network Almanac* for the 1990 edition. The feature story on your work will be updated and a listing of *Analog VLSI and Neural Systems* will be added. I will be sending you the updated text sometime this fall for your approval.

In the meantime, I would like to ask if you could help us with the 1989 *Almanac* by giving us a comment on your story in it. For instance, John Hopfield has written for us about his story:

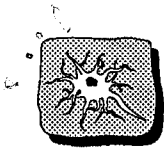
"I am somewhat astonished! The writer has focused on the deepest of issues and has ignored the surface glitz." John Hopfield, Aug. 3 1988.

Enclosed is the story from the 1989 *Neural Network Almanac* on which you are commenting, a sheet on which to write your comment and a self-addressed stamped envelope.

Thanks again and remember, the 1989 story enclosed is only for your convenience when writing your comments. This fall you will be given the opportunity to edit your 1990 *Neural Network* story and write new comments.

Sincerely,

R. Colin Johnson



Cognizer Co.

333 S. State St., Suite 141
Portland, Oregon 97034
(503) 246-6464 U.S.A.

Enclosed is the text from a story about *you* as it appears in the *1989 Neural Network Almanac*.

Please take the time to write a comment about your story that might help us spread the word.

Example comment: "The author has focused on the deepest of issues..." John Hopfield

Comments: _____

