SeeHear

Nielsen, Lars; Mahowald, Misha; Mead, Carver

1987

Document Version:
Publisher's PDF, also known as Version of record

CODEN: LUTFD2/(TFRT-7355)/1-15/(1987)

SeeHear

Lars Nielsen
Misha Mahowald
Carver Mead

Department of Automatic Control
Lund Institute of Technology
April 1987

Department of Automatic Control
Lund Institute of Technology
P.O. Box 118
S-221 00 Lund Sweden

Document name

Date of issue
April 1987

Document Number
CODEN: LUTFD2/(TFRT-7355)/1-15/(1987)

Author(s)
Lars Nielsen
Misha Mahowald
Carver Mead

Supervisor

Sponsoring organisation

Title and subtitle
SeeHear

Abstract

Paper at the 5th Scandinavian Conference on Image Analysis, Stockholm, June 2–5, 1987, Sweden.

Key words
Analog VLSI, neural systems, blind prosthesis.

Classification system and/or index terms (if any)

Supplementary bibliographical information

ISSN and key title

ISBN

Language
English

Number of pages
15

Recipient's notes

Security classification

Paper at the 5th SCIA, Stockholm, June 2-5, 1987, Sweden.

SeeHear

Lars Nielsen
Lund Institute of Technology
Lund, Sweden

Misha Mahowald
Carver Mead
California Institute of Technology
Pasadena, California

1. Introduction

The SeeHear is a system designed to help the blind. The heart of the system is a single custom chip upon which an image is projected by a lens. The function of the system is to map visual signals from moving objects in the image into auditory signals that can be projected through earphones to a listener. A sensation is evoked similar to that which the listener would experience if the moving objects were emitting sound. We hope that the auditory signals provided by the SeeHear device, in addition to the sound cues already present in the environment, will enable blind people to create a more detailed internal model of their surroundings than that which can be extracted from naturally occuring sound cues alone.

Information processing on the chip consists of:

1. Calculating the position of a light source in a two-dimensional optical projection, and processing the intensity information to emphasize temporal changes in intensity.
2. Synthesizing a sound having the appropriate psychophysiologically determined cues for a sound source at that position.

The SeeHear device is small enough to be mounted on the head of the listener, so that all source positions will be measured in a head-centered coordinate system natural to visual and auditory localization. The SeeHear system concept is schematically illustrated in Figure 1. It consists of optics, our custom chip, and ear-phones.

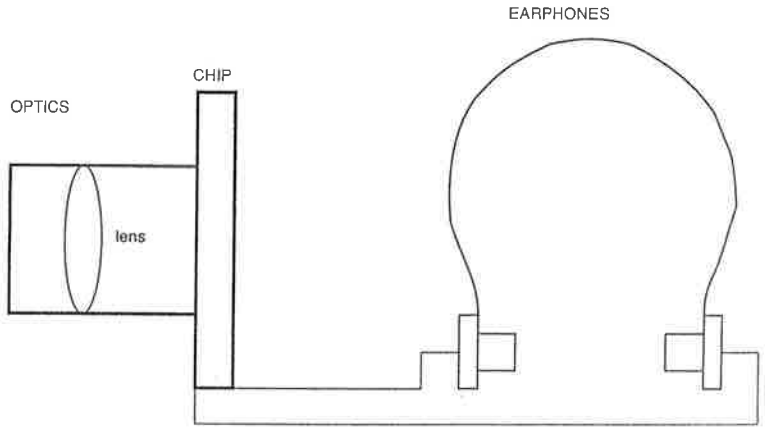

Figure 1. The SeeHear system concept.

We have used biologically inspired representations of visual and auditory information in the inner processing of the chip. The optical processing done on the intensity information is based on signal processing in the vertebrate retina. This processing emphasizes motion information which, in the visual system, is capable of providing depth information when interpreted by higher centers of the brain. Auditory information is generated emphasizing transient events, which are known to provide optimal information for spatial localization.

2. Biological Basis of the SeeHear System

Information about sensory systems comes from a variety of sources such as ethology, human psychology and animal neurophysiology. Although each species of animal has specialized to fill its own ecological niche, underlying structures of peripheral vision and hearing arose in response to fundamental environmental pressures common to all. Human and other vertebrates need to distinguish among a myriad of diverse stimuli in order to perceive objects in the physical world, to avoid obstacles and predators, to procure food, and to care for young. Under these conditions, the evolutionary process has seen to it that the most primitive processing paradigm has been conserved across many species. By focusing our attention on those characteristics which are common to many species of animals, we can come to an understanding of the basic problems facing all higher animals, including humans.

Sensory systems enable animals to gather information about the environment; they transduce input from the environment into neural signals which are processed in the brain in order to create in the animal an internal model of the world. The animal bases all of its actions on this model. Among other things, the internal model must contain information which will enable the animal to find food and avoid danger. It is clear that the ability to sense objects from a distance and maintain a sense of their locations in space is a great selective advantage; it is easier to find food if you don't have to run into it to realize that it's there, particularly if it's mobile. Furthermore, it is easier to delay becoming food if you can localize approaching predators and if you can navigate through obstacles well enough to run in the opposite direction. It is not surprising that the visual and auditory systems of widely diverse species are specialized, beginning at the earliest stages of neural processing, to localize visual and acoustic events. The mapping of light and sound inputs into neural representations appropriate for the localization of the sources of these inputs is performed by peripheral sensory systems. The mapping performed by the visual system differs from that of the auditory system because the physics of light is different than the physics of sound.

Biological Visual Systems

Vertebrates have highly optimized methods for generating visual representations (Shepherd, 1983). The first steps of visual information processing are performed in the retina, upon which the visual scene is projected through a lens. There the light is sensed by a two-dimensional grid of photoreceptors, each of which generate an analog neural potential proportional to the logarithm of the intensity at that point in the image. The logarithm provides a large dynamic range of receptor response, and insures that differences in receptor output will represent a contrast ratio that is independent of absolute intensity with which the scene is illuminated. The light incident on a photoreceptor comes from a local region of space called the *receptive field* of that photoreceptor (Barlow and Mollon, 1982). The location of a photoreceptor on the retina, encodes the location of the light source in real space.

As the visual information is transmitted and processed through multiple layers of neurons, the location information is preserved through *retinotopic mapping.* The two dimensional layers of neurons in the retina transmit their output through the optic nerve to two dimensional sheets of neurons in the brain in a conformal mapping that maintains the relative spatial locations of the signals. Physiological investigations of the retina and of the various visual areas of primate brain have shown that the receptive fields of adjacent neurons correspond to adjacent regions of the visual field.

As the neural processing of the visual information proceeds, the representation becomes more and more complex. The transformation from simple intensity to more complex representations is begun in the retina itself; signals from the photoreceptors are transformed through several layers of neural processing before they are transmitted through the optic nerve to the brain. The output cells of the retina, called retinal ganglion cells, encode information about such things as local intensity gradients and time derivatives of intensity. On-center and off-center

ganglion cells are sensitive to stationary edges, responding to spatial derivatives of intensity within their receptive fields, while on-off ganglion cells are motion sensitive, responding only to temporal derivatives of the intensity profile.

The visual centers of the brain must construct a model of three-dimensional space based only on the spatio-temporal pattern of signals it receives from the retina. In all animals, motion signals are an important part of the reconstruction process. A great number of vertebrates derive visual depth information exclusively from the relative motion of objects in the retinal image that result from the animal's own movements. Although humans use binocular stereopsis for detailed information about depth at close range (less than 1-2 meters), parallax induced by head and body movements is an effective cue to depth even with one eye and, at large distances, motion parallax is the only depth cue available.

Motion parallax is a simple geometric phenomena; it is *not* dependent on binocular interaction. Different versions of the phenomena occur dependent on how the eye and the line of sight are moving relative the scene of interest (Carterette and Friedman, 1975). The simplest example can easily be demonstrated by introspection. If the eyes are fixated at infinity, and the head is moved, the apparent velocity of any object in the retinal image is a monotonic function of the distance to the object. The closest object moves fastest, while more distant objects move more slowly. Objects at infinity appear stationary.

In summary, we have learned three important facts from the visual processing in biological systems:

1. The very first step in visual processing is performed by the photoreceptors; their output represents the logarithm of intensity. The logarithm expands the dynamic range of the photoreceptors and insures that differences in receptor output represent a contrast ratio, which is independent of the illumination of the scene.
2. The locations of light sources in a two-dimensional optical projection of real space are coded by the locations of neurons in retinotopic arrays.
3. The depth cues needed to reconstruct the third dimension of real space are provided by motion signals, which are generated early in visual processing, at the level of the retina. The basic motion cue reported by the retina is the time derivative of the intensity profile.

Auditory Psychophysiology

The SeeHear system depends on the ability to synthesize sounds that will appear to the listener to have come from a specified physical location. Fortunately, auditory psychophysiological research has led to an understanding of sound localization by humans that has made such a sound synthesis possible. Bloom (1977a,b) and Kendall and Martens (1985) have succeeded in quantifying the acoustic cues that humans use for sound localization. Using these cues they were able to synthesize sounds appearing to have come from arbitrarily specified directions. The cues used are a consequence of the interaction of sounds with the physical environment. The sound to be sensed by both ears must propagate through the air and around the head. The sound is then reflected by the pinna and tragus of the outer ear before entering the ear canal and arriving at the ear drum and the inner ear. The modifications of the sound in its journey to the right and left inner ears provide the cues to the location of the sound source.

There are two major horizontal localization cues; these cues are both the result of binaural interactions, as illustrated in Figure 2. The first cue is the difference in arrival time of a sound to the right and left ear, due to the difference in path length. A sound source directly in front of the listener is equidistant from the right and left ears, and therefore has no interaural time delay. Maximal delay occurs with a sound source at plus or minus ninety degrees, directly on the right or left of the listener. The maximum delay depends on the size of the binaural separation. Typical values of the interaural time delay are between 350 and 650 μs. The second horizontal localization cue is due to the interaction of the sound with the head of the listener. Higher frequencies are attenuated when traveling around the head, as shown in Figure 2 a. The

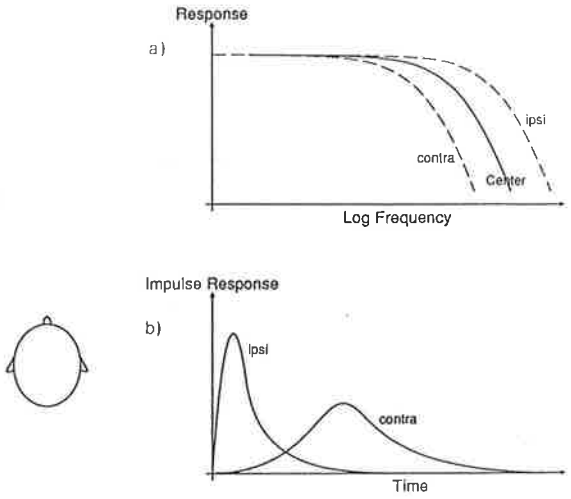

Figure 2. Horizontal cues for auditory localization.

degree of the dispersion of the incoming sound wave depends on the details of the interaction with the listener's head, outer ear, and ear canal. In general, however, sound arriving at the ear farthest from the sound source is more dispersed, having experienced greater high frequency attenuation. We will refer to the high frequency attenuation as *acoustic headshadow*. The impulse response in the left and right ear canal, showing the effects of both cues, is shown in Figure 2 b. The response in the left ear is sharper and less delayed than that in the right ear.

Localization of sound in the vertical direction is made possible by the pinna and tragus of the outer ear. Incoming sound may enter the ear canal via two paths, as shown in Figure 3.

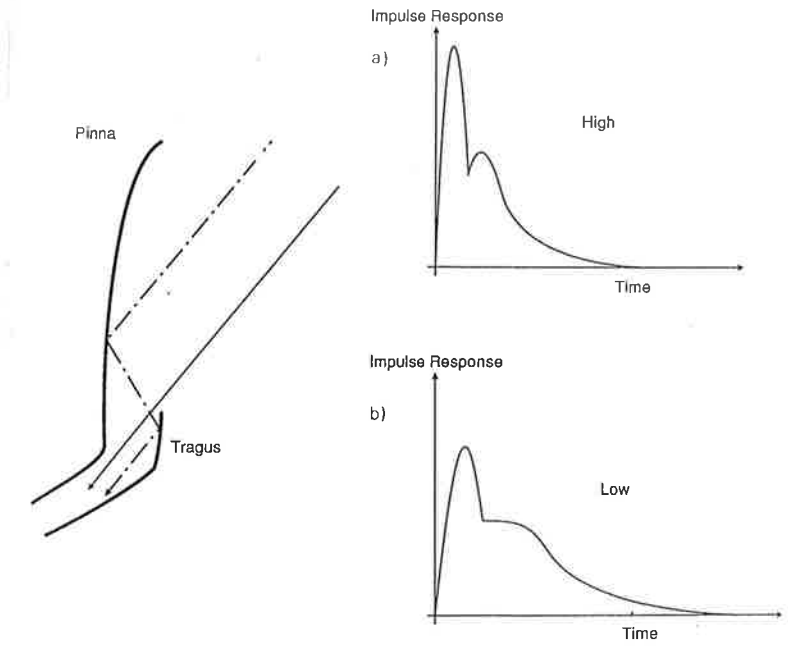

Figure 3. Vertical cues for auditory localization.

One path is as direct as possible given the shape of the outer ear. The other path is longer; the incoming sound bounces from the pinna to the tragus and then into the ear canal. The signals traversing the two paths combine in the ear canal. The difference in path length between the two signals can be measured in psychoacoustic experiments as a notch in the spectral sensitivity function, due to destructive interference when the increase in path length due to pinna-tragus

reflection is one-half wavelength of the incoming sound. The time delay between the two paths is a function of elevation. The shape of the outer ear is unique in every individual, so the absolute values of the time delays vary. For all people, however, the size of the time delay is a monotonic function of elevation, with small delay at high elevations and larger delay at low elevations. The impulse responses for sources at high and low elevations are shown in Figure 3. For humans, typical values of time delay are 35-80 μs.

The sufficient set of cues for sound localization are thus:

1. Interaural time disparity—Horizontal
2. Acoustic headshadow—Horizontal
3. Pinna-Tragus characteristics—Vertical

Main principles

The basic effects and principles of vision and hearing may now be summarized. Visual processing at the retinal level detects motion mainly by taking time-space derivatives of the intensity profile of the visual scene. The transient nature of the time derivative localizes visual events in time. Motion of the observer causes events to move over the retina. The brain is able to construct a model of the three-dimensional world from these events moving on the retina by the use of motion parallax. Hearing is also mainly concerned with events; transients of sound are far easier to detect and localize than repetitive sounds.

The basic difference between visual localization and auditory localization is the way position is represented in the peripheral sensory processing stages. In vision, location in a two dimensional array of neurons in the retina corresponds to location of objects in a two-dimensional projection of the visual scene. The location information is preserved through parallel channels with retinotopic mapping. The auditory system, in contrast, has only two input channels; location information is encoded in the temporal patterns of signals in the two cochlea. These temporal patterns provide the cues which the higher auditory centers will use to build a two-dimensional representation of the acoustic environment, similar to the visual one, in which the position of neuron corresponds to the location of the stimulus.

3. The SeeHear design

The needed *function* of the SeeHear system can now be formulated, in the context of the principles outlined in Section 2. Signals representing visual events must be transformed into acoustic events. The *location* (direction (θ, ϕ)) of each the of the visual events must be encoded and that encoding used to synthesize an acoustic signal that will provide the hearing cues appropriate for an acoustic event at that location. Events occurring simultaneously in a two-dimensional projection of a visual scene must be transformed into an acoustic signal coded in only two channels. Information that would allow the reconstruction of depth (motion parallax cues, for example) must be preserved. If the SeeHear system successfully performs the overall function, users can create an internal model of the visual world using their auditory system.

Designing a chip performing the function needed is an interesting architectural challenge. Processing elements and their interconnections must be designed and spatially arranged on the chip. The key problem is to find specific subsystems which, in aggregate, will perform the overall function. We will present our subsystems leading up to a complete design.

Vision

The SeeHear visual system is similar to the vertebrate eye. A lens maps the scene onto a two-dimensional array of pixels. Each pixel contains a photosensor and associated local processing. The light falling on one pixel in the array comes from a direction in real space. The location of a pixel in the array corresponds to location of a particular feature in a two-dimensional

projection of the visual scene. As in biology, the visual system first takes an analog logarithm of the intensity and then takes an analog time derivative. This computation is done locally on the chip, thereby maintaining the direction information. The visual processing in the SeeHear system thus incorporates two main features: the location of light sources is encoded in the retinotopic array of pixels, and emphasis of motion information in the visual processing is done by computing the analog time derivative of an unsampled time varying signal. Preliminary results indicate that this signal, when used as input to the sound synthesis machinery, will generate appropriate sound signals for the auditory system.

Sound synthesis

Consistent with the continuous nature of the visual part of the SeeHear design, we use continuous methods for doing the sound synthesis that will conserve the analog nature of the signal from the pixels, and that will take advantage of the place information in the retinotopic visual array.

Horizontal cues: One auditory source We start with the simplest sound synthesis problem in which we wish to generate only one auditory event from a single location. The output of the device can be used directly to produce sound that can be presented through earphones. The auditory cue generator for one source is illustrated in Figure 4. An event generating input device is connected to two delay lines. The delay lines are analog; one leading to the left and the other leading to the right. Each line delays an input signal by an amount per unit length specified by a variable control.

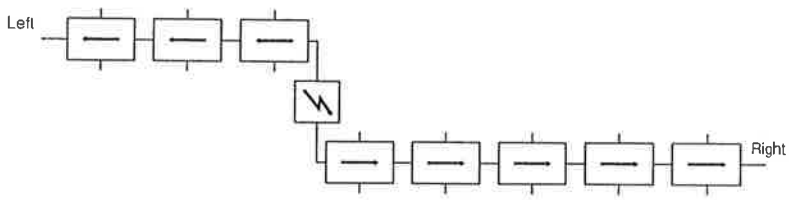

Figure 4. A generator of horizontal auditory cues for one acoustic source. A binaural headshadow model (delay lines) that generates the signals in Figure 2.

The input device drives a pulse into both delay lines. If the left delay line is shorter than the right, the signal will reach the left output first. In addition to providing a delay, the analog processing in the delay lines filters the higher frequencies from the signal. The farther the signal travels, the more dispersed it gets. We thus get both the binaural time disparity cue and the acoustic headshadow cue from the delay lines. The horizontal direction is determined by the difference in length of the left and right delay lines. We therefore call the delay lines a *binaural headshadow model.* We have now synthesized exactly the signals present in the ear canals as shown in Figure 2.

Horizontal cues: A set of acoustic sources An auditory cue generator for multiple sources is illustrated in Figure 5. The system has a set input devices, each representing events at different horizontal locations, spaced at regular intervals along two delay lines. As before the delay lines are analog, and one is leading to the left and one is leading to the right.

The extended system is able to synthesize sounds appearing to have come from multiple sound sources simultaneously. The sound sources represented by the input devices are distributed along a horizontal arc; the horizontal positions of the sound sources correspond to the positions of the input devices along the delay lines. This extended system can generate sounds for multiple sources using only two delay lines which are shared by several input devices. This sharing is possible because, at the signal values we are using, the analog processing in the delay lines is linear. We can have more than one input device and the signal from each device will be superimposed on whatever signal is already present in the delay lines at that position. The

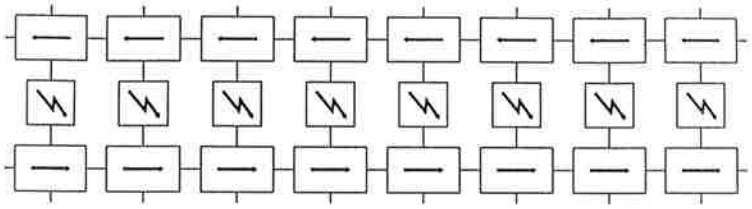

Figure 5. A generator of auditory cues using a set of signal sources. The generator is a binaural headshadow model (delay lines) that generates the signals in Figure 2. Observe that the sources can share the delay lines in this design due to linear superposition.

different inputs are added linearly in the delay lines just as sound pressure waves at moderate levels are combined in air.

Vertical cues The vertical cues are obtained by analog delay and add sections, as illustrated in Figure 6. These sections provide multiple paths, and the size of the delay determines the length of the longer path relative to the direct path. Setting the delay in the delay and add section thus gives the elevation, and for this reason we call them a *pinna-tragus model.* It is capable of generating signals as those in Figure 3.

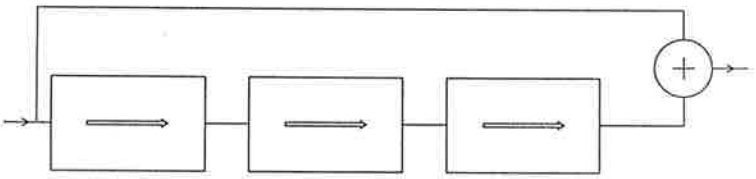

Figure 6. A pinna-tragus model that can generate the signals in Figure 3.

Combining horizontal and vertical cues We are able to generate sounds appearing to come from multiple sound sources at the same elevation by combining a binaural headshadow model (Figure 5) with two pinna-tragus models (Figure 6), so that at the end of each delay line there is a delay and add section. The delay of the delay and add sections, which must vary as a function of elevation, is controlled by the time constants of the sections, which is controlled separately from that of the headshadow delay line. This system is capable of giving the auditory direction cues in Figure 2 and Figure 3. Assume that a single input device slightly to the left of center signals an event with a pulse. The three directional hearing cues for a single sound source to the left of the listener will then be obtained. Sounds generated from these signals, presented through earphones inserted in the ear canals, will give the illusion of a sound source slightly to the left of the listener at the elevation set by the pinna-tragus model.

The size of the auditory field (and, consequently, the horizontal positions of the equidistant sound sources) is controlled by setting the total delay of the delay lines. The maximal total delay of one delay line (corresponding to an auditory field of 180 degrees) is the time it takes a sound, originating directly on one side of the listener, to travel from the closer ear to the ear on the far side of the head. Setting the delay in the pinna-tragus model gives the elevation ϕ. The relative simplicity of the implementation is due to the fact that the analog processing of the delay lines imitates the superposition of sound pressure waves in air.

The SeeHear Chip

The principle SeeHear chip concept is shown in Figure 7. The pixel array is oriented so that the rows correspond to the horizontal direction and the columns to the vertical direction. The processing in each pixel strongly emphasizes time derivatives, so edges moving in the visual scene moving over the photoreceptors generate large, pulse-like transient signals. The pixel outputs in each row act as input devices for a hearing cue generator like the one in Figure 5. Each pixel provides input to both delay lines of the hearing cue generator at the points in the

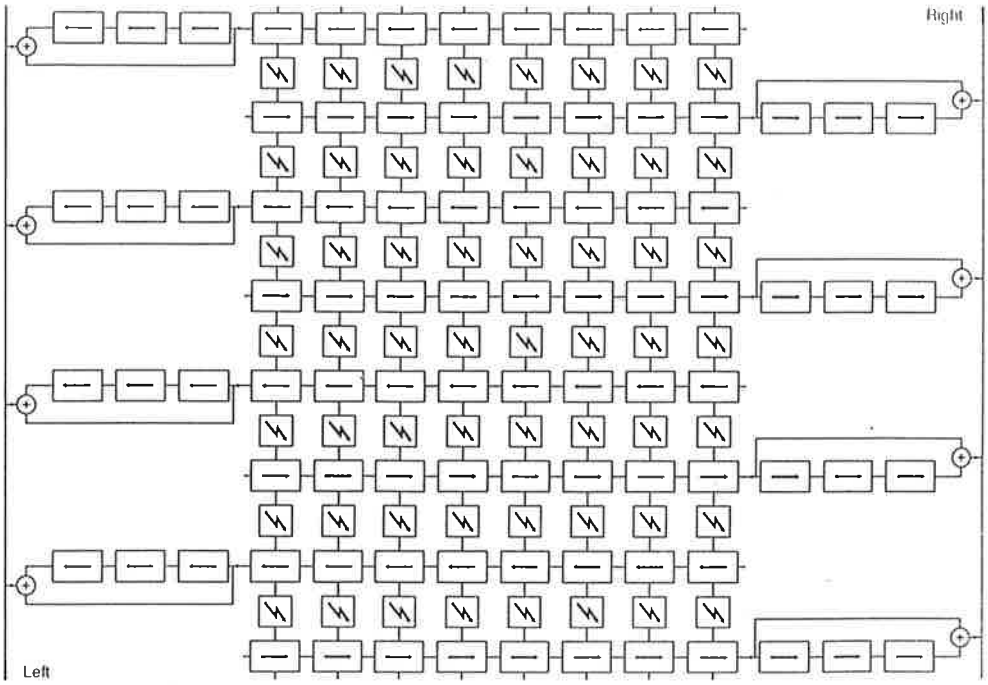

Figure 7. An overall diagram of the concept of the SeeHear chip.

lines which are physically adjacent to the pixel. The signals from the pixels along the row are delayed and filtered by an amount which is a function of the horizontal position of the pixel. This delay and filtering generates the horizontal hearing cues. For the vertical cues, we observe that all of the pixels in one row of the array correspond to the same elevation angle in real space. The pinna-tragus model (delay and add section) at the end of each row is therefore tuned to have a delay corresponding to its elevation. This delay gives the vertical hearing cue. To form the output to the earphones, the outputs from the pinna-tragus models to the left are combined, as are the outputs from the pinna-tragus models to the right. Once again, the analog design automatically provides linear superposition.

The function of the SeeHear device, outlined at the beginning of this section, has been realized with elegance in this chip. The visual processing, all done in a local computation, emphasizes time derivatives of intensity, and so provides transient events as inputs to the sound generation system. Transient events are more easily localized by the auditory system than continuous signals. The conversion from a visual representation of position to an acoustic one is achieved in a simple and natural mapping in this architecture. In the visual system of the SeeHear, the position of a pixel corresponds to a location. The position of a pixel on the chip determines the delay and degree of filtering an input signal will undergo in the two delay lines and the delay of the delay and add sections. These delays generate the acoustic cues that the auditory system uses to determine the location of a sound source. The analog nature of the processing in the SeeHear leads to an economy of computational elements. The superposition of signals in the auditory cue generator allows the parallel computation of sound cues for events occuring simultaneously in the visual scene, and encodes the information in only two output channels. A multiple channel visual representation is thus transformed into a time-varying acoustic representation appropriate for perception by the auditory system. Thus, not only are computational elements of the auditory cue generator shared by several pixels, but no additional computational machinery is introduced to provide clocking or sampling; time is its own representation and memory is spatially encoded.

4. Implementation

Having presented the conceptual framework of the SeeHear system in Section 3, we will now describe the silicon implementation of the SeeHear system. Primitive circuit elements are combined to form building blocks, which can be composed in the final layout.

Layout

An overall diagram of the layout of the SeeHear chip was shown in Figure 7. A row includes a row of pixels, a single delay line and a pinna-tragus section at the output of the delay line. Alternate rows are mirrored to generate a roughly hexagonal array of photoreceptors and alternating rightward propagating and leftward propagating delay lines. The signals for the right and left channels are generated by summing the current outputs of the pinna models onto output wires running vertically along the right and left sides of the chip respectively. The floor plan for the chip exactly reflects the arrangement in Figure 7. The actual chip contain 19 rows of 21 pixels each.

Primitive circuits

Before we can describe the building blocks that correspond to the conceptual units previously discussed, some explanation of the primitive circuits that we use in these analog designs must be provided. The SeeHear is one project in a current effort to use CMOS technology for analog design (Mead, book in progress). A family of circuits suited for neural computation has been developed. The circuits operate in the sub-threshold range and hence require extremely small currents (10^{-12} - 10^{-6} A) for their operation. One circuit, which the conceptual units here are based on, is the transconductance amplifier, shown in Figure 8.

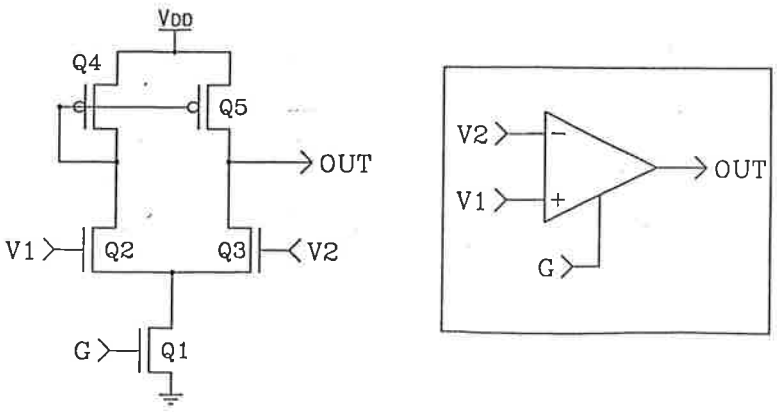

Figure 8. Transconductance amplifier. Schematic transistor diagram and symbol.

The transconductance amplifier generates a current that is a function of the difference between two input voltages and proportional to the transconductance G set by a control voltage. For small input voltage differences, the current out of the transconductance amplifier is given by $I_{out} = G(V_1 - V_2)$. The transconductance amplifier is frequently used in a configuration as shown in Figure 10 below. The output of the transconductance amplifier is tied to the negative input and to a capacitor that is implemented by a large-area transistor. The capacitor integrates the difference between the input signal and the value on the capacitor: $\frac{dV_{cap}}{dt} = \frac{G}{C}(V_{in} - V_{cap})$. For high frequencies on the input, the voltage on the capacitor represents a smoothed, integrated version of the input. For lower frequency changes of the input, the capacitor voltage simply follows the input with a small time lag, determined by the size of the capacitor and the transconductance of the amplifier.

Building blocks

The three building blocks which comprise the system are: the retina model (pixel), the binaural headshadow model (analog delay line), and the pinna-tragus model (delay and add sections).

Retina model (Pixel) The pixel circuit is composed of a receptor (the light transducing element), and a differentiator that emphasizes the temporal derivative of the intensity at that receptor. Each photo-receptor provides an output voltage that is logarithmic in the light intensity over 4 to 5 orders of magnitude. The intensity range covered is comparable to that covered by the *cones* in human visual systems. The logarithmic characteristic provides an output voltage difference proportional to the *contrast ratio*, independent of the absolute illumination of the scene.

The schematic of the receptor is shown in Figure 9. The operation is similar to that described in (Mead, 1985). Bipolar transistor structures are a natural biproduct of the bulk CMOS process used to implement the SeeHear system. They are usually considered a parasitic device, and can lead to certain problems in standard logic circuits. They are, however, excellent photo-detectors, giving approximately 1000 electrons out per absorbed photon in. We use a large-area bipolar phototransistor of this type as our primary receptor. MOS feedback transistors operating in sub-threshold are used to create an output voltage that is a logarithmic function of the photo-current. This output voltage changes about $320mV$ for each decade change in light intensity. The output voltage range is 1V to 2.5V below V_{DD}, where direct coupling to subsequent stages is readily accomplished.

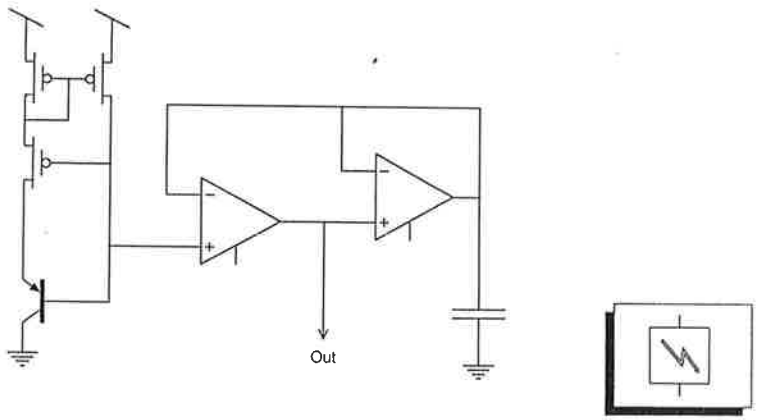

Figure 9. Retina model (Pixel).

The differentiator is made up of two transconductance amplifiers and a storage capacitor, as the circuit diagram in Figure 9. We approximate a delayed version of the signal by the time averaged signal stored on the capacitor. The output of the differentiator is a voltage determined by the first amplifier whose positive input is driven by the photoreceptor and whose negative input is the voltage stored on the capacitor. This circuit has a gain at short times set by the open-circuit gain of the first amplifier—usually 50 to 100. The gain for long times is 1, since the entire arrangement acts as a follower. The output of the pixel to a steady illumination with a small superimposed square wave modulation is shown in Figure 10. The output of the differentiator is at a voltage level that is convenient for the next level of implementation of the SeeHear system. The control G of the second amplifier sets the maximum current that may flow into or out of the storage capacitor. The time-constant is set by the control of the first amplifier.

In the configuration shown, maximum outputs will occur when high contrast features move over the retina. This emphasis on derivatives provides maximum opportunity for users to generate information by body movement, as seeing people do with normal visual processing.

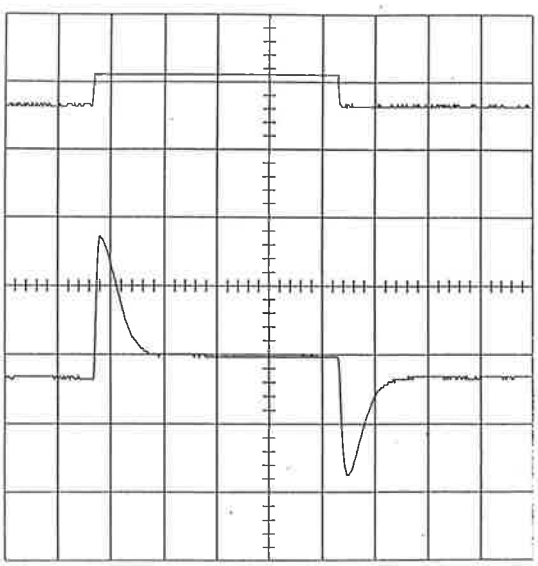

Figure 10. Pixel response to steady illumination with a small superimposed square wave.

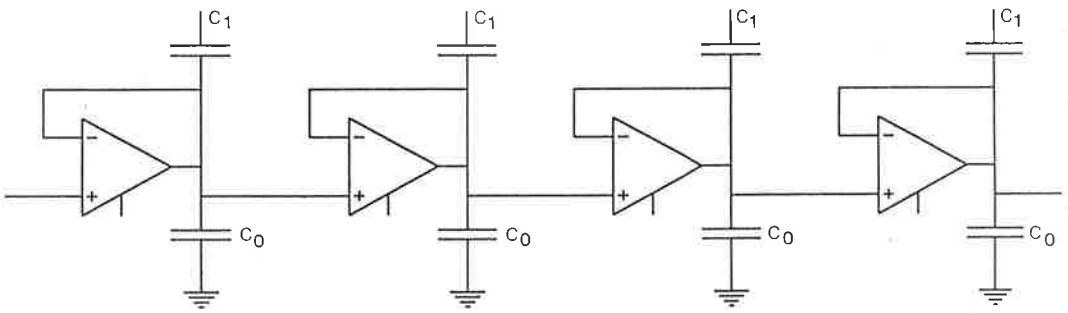

Figure 11. Binaural headshadow model (Delay line).

Binaural Headshadow Model All auditory processing is accomplished by means of analog delay lines. Each analog delay line is simply a long string of follower-integrator sections as shown in Figure 11. Each section delays the incoming signal and broadens it slightly. The time it takes a signal to reach the output of the line, and the degree of attenuation of high frequencies in the signal are both monotonic functions of the number of sections through which that signal must propagate. The delay of the line from one end to the other (which determines the range of interaural time disparities that the SeeHear system can encode) is controlled by the number of sections and the time constant of each section.

Coupling from individual pixel outputs is accomplished by means of the C_1 capacitors. The capacitive voltage divider ratio $C_1/(C_1+C_0)$ is chosen to inject about $200mV$ into the line for a $5V$ excursion of the pixel differentiator output signal. The inputs to the delay lines are supplied with a DC level from off-chip. The optimum coupling occurs when the time-constant of the differentiator roughly matches that of the delay line. Figure 12 shows the output of one delay line excited by a small step in light intensity. In part (a) of the figure, the light was focused on a pixel near the output end of the delay line. In part (b), the light was focused on a pixel in the middle of the chip, and in (c) near the input end of the delay line. Note that the qualitative features required by the headshadow model are nicely embodied in the delay line implementation.

Pinna-tragus model (Delay and add sections) The combined, processed, outputs of two adjacent rows of pixels appears at the output of each headshadow model delay line. Each

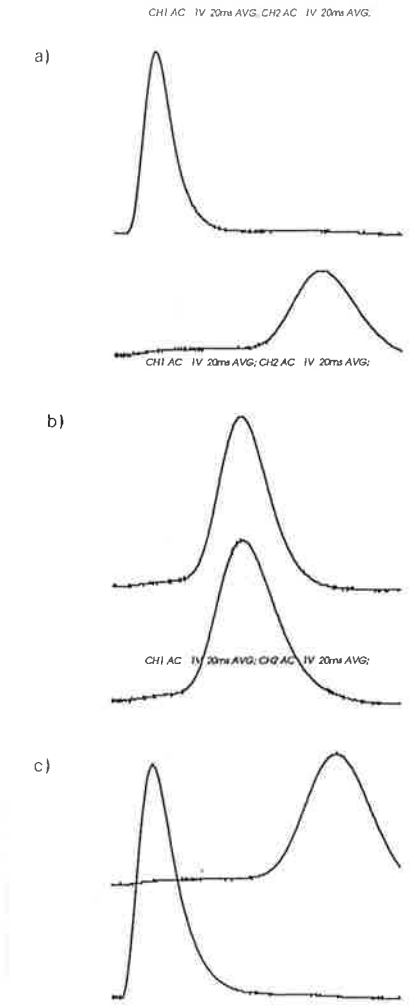

Figure 12. Horizontal cue measurements

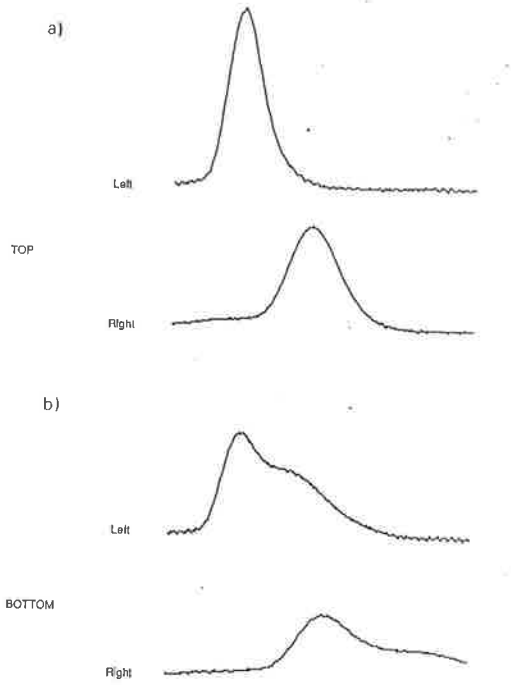

Figure 13. Elevation cue measurements

such output is equipped with its own pinna-tragus model. The addition is accomplished by summing the output currents of two transistors onto a wire. The voltage output of the head-model is fed to the gate of one of the transistors, and the output of the pinna-tragus model is to the gate of the other. It is thus approximately the headshadow signal and a delayed version of that signal. These currents are summed directly onto the appropriate output line. The output transistors are operated above threshold, where their non-linear behavior does not cause unacceptable distortion problems. In order to realize the decrease in pinna-tragus delay with elevation, the transconductance controls of the amplifiers in the delay and add sections are connected to polysilicon lines (which have a high resistance) that runs vertically along the edge of the chip. Each end of the line are brought out to an off-chip pad. By applying different voltages to the two ends of the line, any desired gradient in pinna-tragus delay can be achieved.

The SeeHear output signals to the two ears, in response to a flashing stimulus, are shown in Figure 13. In part (a) of the figure, the light was near the top of the field of view, whereas in part (b) of the figure, the light was near the bottom of the field of view. In both cases, the stimulus was near the left edge of the field in view. At high elevations, the pinna-tragus delay was short enough that the indirect signal is lost in the width of the direct signal. For lower elevations, a distinct indirect version of the signal is visible in the output to both ears. The SeeHear chip is thus capable of generating the three principle auditory cues in response to visual signals anywhere in the field of view.

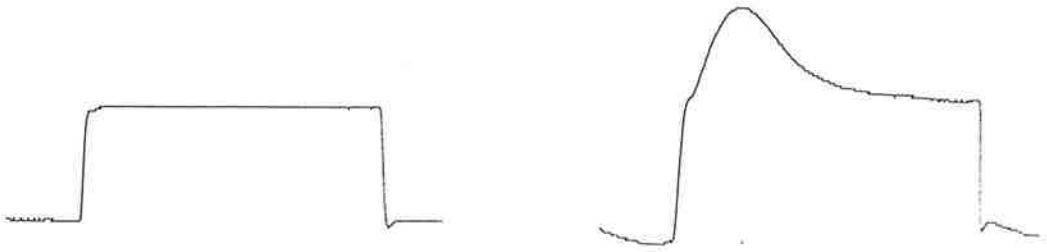

Figure 14. Measurement of the horizontal cue in auditory time scale.

5. Results

Experiments with a working system are, of course, the only way to determine how the SeeHear device will be used. We have performed a number of listening tests to evaluate the realism of the localization cues generated by the device. Experiments were complicated by a design defect that can plague any device that uses on-chip photosensing. In the actual SeeHear layout, not only were the photoreceptors exposed to incoming light, but many other areas of the actual chip were as well. In particular, the delay lines are operating at very low current levels, and are sensitive to extraneous currents produced by collecting the minority carriers generated by the light. The response of the chip with its differentiator disabled allows us to evaluate the magnitude of the effect. The differentiator can be disabled by setting the control on the first amplifier to zero. Under these conditions the response of the chip to a flashing stimulus is as shown left in Figure 14. There is, indeed, a large response induced directly into the delay lines. The response to the same stimulus with the differentiator set properly is shown to the right in Figure 14. The response of the differentiator is nicely superimposed on the background signal due to minority carriers. The amplitude of the desired response, relative to the background, increases as the time constant of the delay line is increased. The responses shown earlier in Figure 12 were obtained by operating the chip with delay times longer than the auditory delay between two ears.

In spite of the confusing effects due to runaway minority carriers, the SeeHear chip creates a remarkably realistic auditory image of a flashing stimulus when the delays are set to correspond to the distance between two loudspeakers. Due to the headshadow cue, the illusion of horizontal localization is much more convincing than that obtainable with time disparity alone.

The pinna-tragus cue, however, cannot be finessed in the same way, and we were not able to achieve the correct auditory illusion of elevation with the current chip. There are other minority carrier effects that were simply averaged out in Figure 13. Minority carriers are generated over the entire surface of the chip, and must diffuse to the nearest junction where they can be collected. This diffusion process takes many milliseconds for the distances involved. The time course of the response due to these distant carriers blurs the signal enough to obscure much of the detail required for horizontal auditory localization, and more than enough to remove any chance of vertical localization. There is another reason that the achievable vertical localization effects are not more realistic. In the implementation described, there were only four stages in the pinna-tragus model. The delay-bandwidth product increases as the square root of the number of stages, and therefore a much better pinna-tragus reflection can be created by using more stages of shorter delay.

Needless to say, the improvements suggested by these observations have been factored into new chip designs, which will materialize shortly.

6. Conclusions

This paper has described a complete system: one whose organizing principles were inspired by those found in biological systems. The system has been implemented in a neural paradigm. The result is a vast amount of both auditory and visual computation in a few square millimeters, and a few milliwatts. The understanding of the mapping of sensory input to internal representations have been crucial, especially the encoding of space. The natural mapping of space and time in an analog VLSI system parallels the kind of processing found in the brain, because both systems use the properties of physical devices as computational primitives, and both technologies are limited by the interconnections rather than by the computations themselves.

As we evolve the system, any improvement in either subsystem should be immediately noticeable to a user. For that reason, the chip should be a valuable proving ground within which to learn a great deal about the early stages of both the visual and auditory systems.

7. References

BARLOW, H.B. and J.D. MOLLON (1982): *The Senses*, Cambridge University Press, Cambridge.

BLOOM, P.J. (1977a): "Determination of monoaural sensitivity changes due to the pinna by use of minimum-audible-field measurements in the lateral vertical plane," *J. Acoust. Soc. Am.* **61**, No.3, 820-828.

BLOOM, P.J. (1977b): "Creating source elevation illusions by spectral manipulation," *Jour. of the Audio. Eng. Soc.* **25**, No.9, 560-565.

CARTERETTE, E.C. and M.P. FRIEDMAN (Eds.) (1975): *Handbook of Perception, vol. 5 Seeing*, Academic Press, New York.

KENDALL, G.S. and W.L. MARTENS (1985): "Simulating the cues of spatial hearing in natural environments," Computer Music Studio, Northwestern University, Evanston, IL 60201, USA.

MEAD, C.A. (1985): "A sensitive electronic photoreceptor," Chapel Hill Conference on VLSI.

MEAD, C.A., *Analog VLSI and Neural Systems*, To appear.

SHEPHERD, G.M. (1983): *Neurobiology*, Oxford University Press, Oxford.