Analog VLSI Model of Binaural Hearing

Carver A. Mead, Xavier Arreguit, and John Lazzaro

Abstract—**The stereausis model of biological auditory processing was proposed as a representation that encodes both binaural and spectral information in a unified framework. We describe a working analog VLSI chip that implements this model of early auditory processing in the brain. The chip is a 100 000-transistor integrated circuit that computes the stereausis representation in real time, using continuous-time analog processing. The chip receives two audio inputs, representing sound entering the two ears, computes the stereausis representation, and generates output signals that can directly drive a color CRT display.**

I. Introduction

ARTIFICIAL speech-recognition systems suffer from a paucity of computational resources in comparison with biological auditory systems. As a result, many of nature's methods of speech processing have been abandoned as computationally impractical by engineers building speech-recognition systems. We believe that adaptive analog VLSI is an excellent medium for realizing biological strategies for auditory processing in a computationally efficient manner. This paper describes a working analog VLSI chip that implements a proposed model of early auditory processing in the brain.

Humans hear using two ears. Psychoacoustic experiments show that humans use binaural information to improve speech intelligibility in noisy environments. Most artificial speech-recognition systems use monaural sound input, owing to a lack of computational resources and to the representational difficulties of incorporating binaural information in a recognition system. The stereausis model of biological auditory processing, first suggested by Loeb *et al.* [1] and developed recently by Shamma [2], proposes a representation that encodes both binaural (cross-correlation) and spectral (autocorrelation) information in a unified framework.

II. The Stereausis Algorithm

The stereausis algorithm derives a two-dimensional representation of binaural sound from two sound inputs. This representation depends on the characteristics of the cochlea, the sense organ of hearing.

The cochlea transforms acoustic energy into a mechanical traveling wave and detects the velocity of this wave at several thousand points along its trajectory, converting the velocity into electrical signals transmitted on about 50 000 nerve fibers to the brain. Individual fibers are most sensitive to a restricted range of frequencies. To first order, auditory nerve fibers are the half-wave-rectified, time-differentiated outputs of resonant low-pass filters. The fibers maintain temporal information well; for frequencies under 3 kHz, the fibers encode the shape of the filtered input waveform [3].

The mechanical traveling-wave structure with which the cochlea implements filtering has an exponentially tapered stiffness. A traveling wave excited by an input pulse first excites higher frequency fibers, and later lower frequency fibers, as it traverses the cochlea. The time delay inherent in the structure increases exponentially with distance; the cutoff frequency of fibers connected to the structure likewise decreases exponentially. A detailed description of cochlear function can be found in the work by Lyon and Mead [4].

The stereausis representation is simply the complete cross-correlation of the outputs from the left and right cochleas. Although correlation is a simple mathematical operation, the nonlinear spatiotemporal processing of the cochlea makes nontrivial the mathematical characterization of stereausis output for a given binaural signal. In this paper, we shall explain the stereausis algorithm, give a detailed explanation of the architecture of the chip, and show outputs of the chip for a variety of artificial and speech stimuli.

III. System Architecture

Fig. 1 shows a block diagram of the stereausis chip. Binaural electrical signals provide inputs to silicon models of the left and right cochleas [5], drawn as cascades of boxes containing the symbol Δ. The silicon cochleas are one-dimensional physical models of a traveling-wave structure, implemented as a cascade of 56 second-order sections with exponentially scaled time constants. The array thus contains 3136 correlation elements operating in parallel. The chip is 6.9 mm on a side using 2 μm CMOS technology. For the data shown in this paper, we tuned the silicon cochleas to span approximately two decades, 100 Hz to 12 kHz.

Outputs from each second-order section connect to nonlinear circuit models of sensory transduction in the cochlea [6]. These circuits perform time differentiation, nonlinear waveshaping, and half-wave rectification. The final circuit output uses a two-wire representation, coding opposite waveform polarities on separate wires.

The chip compares every output of the left cochlea with every output of the right cochlea in parallel, using the two-dimensional array of analog processors shown in Fig. 1. The comparison element computes a measure of the correlation of the instantaneous activity of the two inputs.

Fig. 2 shows a block diagram of the comparison element. Each element receives four wires, representing the positive and negative derivatives of the two input signals. Four correlation processors, drawn as wedges, compute the normalized one-

Manuscript received August 21, 1990; revised November 22, 1990. This work was supported by the Office of Naval Research, the California Office of Competitive Technology, and the System Development Foundation.

C. A. Mead and X. Arreguit are with the California Institute of Technology, Pasadena, CA 91125.

J. Lazzaro was with the California Institute of Technology, Pasadena, CA. He is now with the Department of Electrical and Computer Engineering, University of Colorado, Boulder, CO 80309.

IEEE Log Number 9042028.

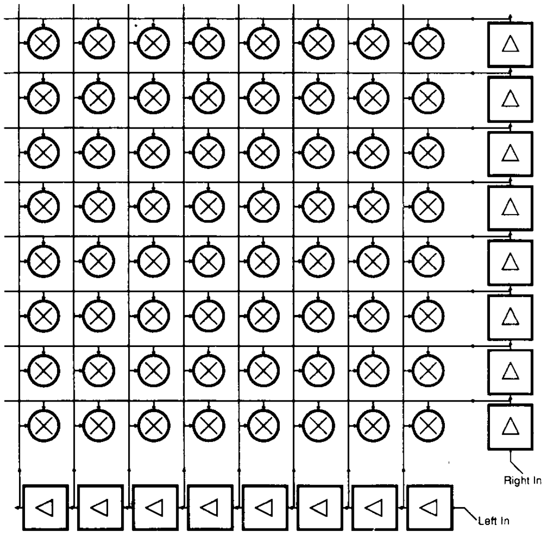

Fig. 1. Architecture of the stereusis chip. Left and right signal inputs enter silicon cochleas at the lower-right corner of the figure. The cochleas are analog circuits that act both to filter the signal and to delay the signal in time; to symbolize this delay, we draw the cochleas as a cascade of boxes labeled Δ, which border the bottom and right edges of the figure. The center of the chip computes the correlation of every left-cochlea output with every right-cochlea output in parallel; there are 3136 analog four-quadrant correlator circuits on the chip, each drawn in the figure as a circle marked with a cross. The outputs of the correlation elements are scanned off the chip and displayed on a video monitor. The chip contains the scanning logic necessary to drive an NEC Multi-Sync monitor, requiring only a few driver transistors and passive components off-chip.

quadrant product of the two inputs,

$$z(t) = \frac{x(t)y(t)}{x(t) + y(t)}.$$

Correlations of the same polarity of the two inputs are summed to produce an aggregate correlation signal; correlations of the opposite polarity of the two inputs are summed to produce an aggregate anticorrelation signal. Because the signals cannot be both correlated and anticorrelated at the same time, a winner-take-all network [7] compares these aggregate correlation and anticorrelation signals, producing the final outputs of the comparison element.

Both outputs of each comparison element are sequentially scanned off the chip and displayed as two colors on a video monitor. The chip contains the scanning logic necessary to drive an NEC Multi-Sync monitor, requiring only a few driver transistors and passive components off-chip. The scanning circuitry, not shown in Fig. 1, is similar to that reported in [8]. For the figures in this paper, the video output for each comparison processor was configured to display black when the signals are correlated, white when the signals are anticorrelated, and gray for equal amounts of correlation and anticorrelation. These outputs form a 56-pixel by 56-pixel image of binaural correlation. The upper-left corner of the image represents 100 Hz frequencies, the lower-right corner 12-kHz frequencies.

IV. Cochlear Response to Pure Tones

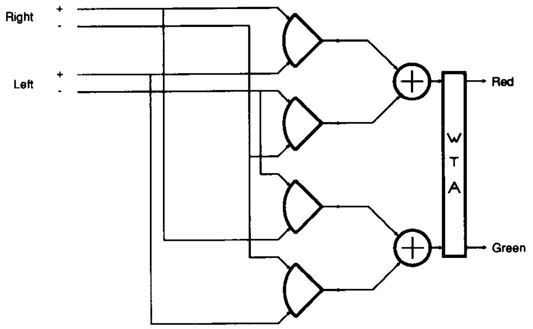

Fig. 2. Functional block diagram of a comparison element. Correlation processors, represented by wedges, compare the positive and negative portions of the two inputs. The sum of these partial computations represents correlation. Correlation processors comparing the positive portion of one input with the negative portion of the other input compute the anticorrelation of the two inputs. A two-input winner-take-all network, drawn as a box labeled WTA, compares these correlation and anticorrelation inputs. The output-scanning circuitry represents correlation as red and anticorrelation as green; for the figures in this paper, extra off-chip circuitry mapped correlation to black and anticorrelation to white.

A description of the spatiotemporal response of the silicon cochlea provides a framework for understanding the operation of the stereausis chip. Fig. 3(a) shows the output as a function of time at five positions along the silicon cochlea, in response to a low-frequency sinusoidal input.

The waveform travels without significant damping through the high-frequency regions of the cochlea, shown at taps t_0, t_1, and t_3. As the waveform reaches the region tuned to its frequency, its output is at a maximum, as shown at tap t_5. For cochlear positions past this point, the traveling wave is damped rapidly as shown at tap t_7. Plotting the value of the signals at t_0, t_1, t_3, t_5, and t_7 at a particular moment in time yields the spatial response of the cochlea, shown in Fig. 3(b). This plot highlights the exponentially decreasing velocity of propagation of the traveling wave.

Fig. 3(c) shows the results of the hysteretic differentiator [9], which provides the nonlinear wave-shaping model of sensory transduction. The circuit produces a large voltage change whenever the time derivative of the input changes sign. Other circuits differentiate this signal with respect to time, with an adjustable time constant, and perform half-wave rectification, producing the final outputs shown in Fig. 3(d). This signal, and a companion signal coding the opposite polarity, are the outputs that connect to the two-dimensional array of comparison processors. Any signals over 20 mV in amplitude are sufficient to provide sharply defined output.

V. Chip Response to a Pure Tone

Fig. 4 shows the output of the stereausis chip in response to a 100 Hz sinusoid, presented binaurally with different interaural phase delays (IPD's). Fig. 4(a) shows the chip response to a 0 radian IPD. The cochleas receive identical inputs and send identical patterns to the array of comparison processors. A black stripe of correlation along the diagonal of the display reflects this similarity.

The off-diagonal response is the result of the correlation of the two spatial waves at different spatial shifts. Let us compare the output of the first tap of the left cochlea with the taps of the right cochlea (first column) for a traveling wave at time $t = 0$ (Fig. 3(b)). Within a certain distance ($x < t_1$ in Fig. 3(b)), the phase shift of the sinusoid with respect to tap t_0 is smaller than

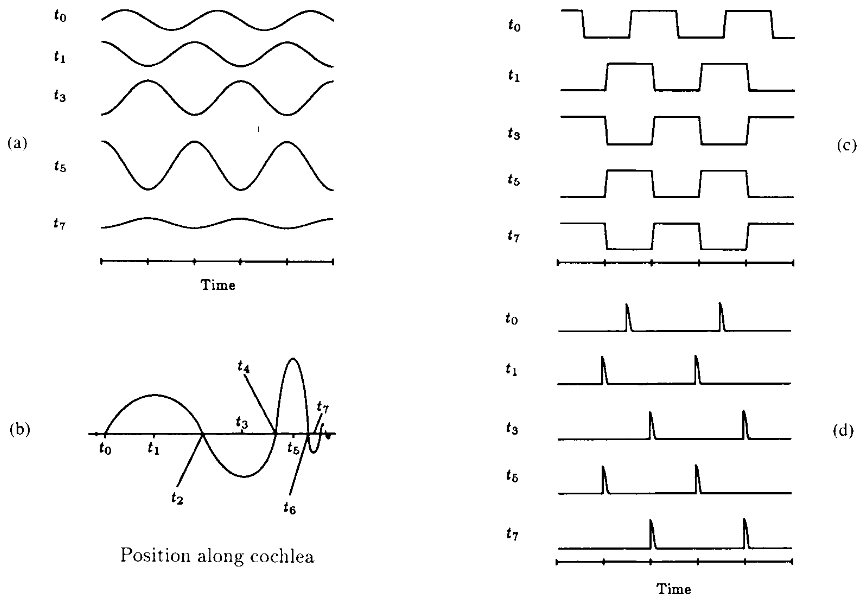

Fig. 3. Plots detailing the operation of the silicon cochlea. (a) Cochlear response to a sinusoid, at five taps (t_0, t_1, t_3, t_5, t_7, not equaly spaced) along the traveling wave structure, chosen to correspond to significant points on the waveform at a given time. (b) Spatial response of the cochlea to a sinusoid, at a particular moment in time. Note that the markers (t_0, t_1, t_3, t_5, t_7) in (b) refer to the position of the waveforms in (a). (c) Intermediate and (d) final processing of the cochlear signals by the circuit model of sensory transduction.

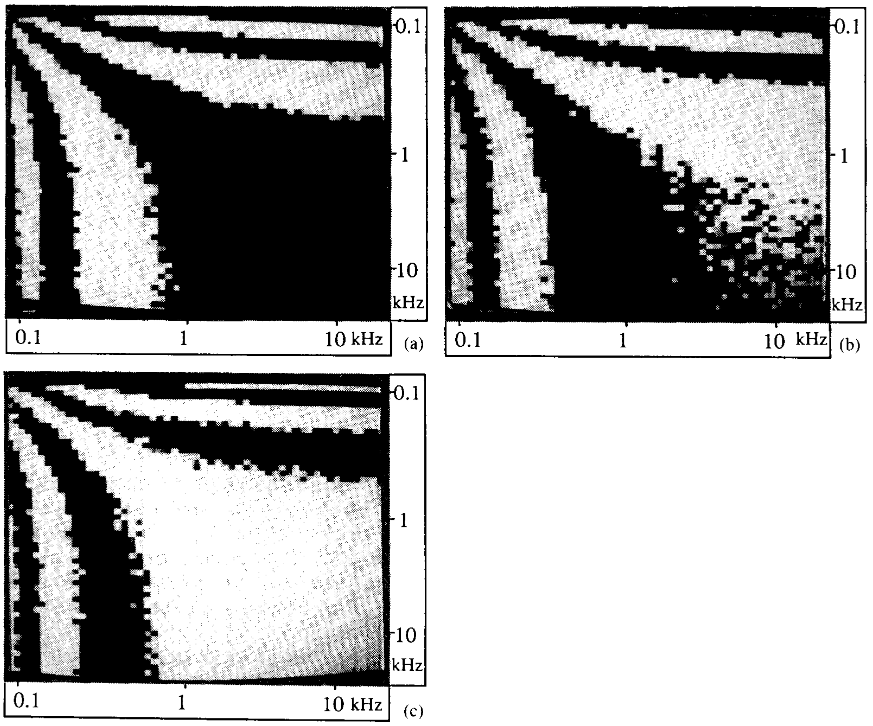

Fig. 4. Photographs showing the output of the stereausis chip, in response to a 100 Hz sinusoid, presented binaurally with different interaural phase delays (IPD's) of (a) 0 radian, (b) $\pi/2$ radian, and (c) π radian. Numbers on the edge of the photographs show the best frequency at different cochlear places along the chip.

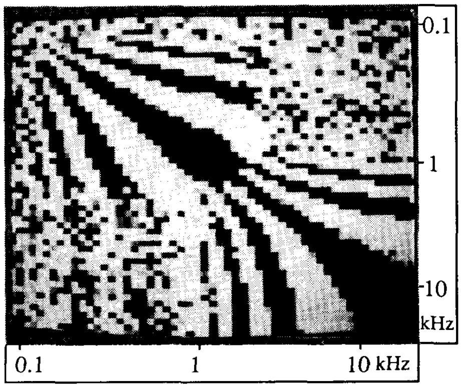

Fig. 5. Photograph showing the output of the stereausis chip in response to the simultaneous monaural presentation of 200 Hz and 2000 Hz sinusoids of equal amplitude. The lower-right pattern encodes the 2000 Hz sinusoid, whereas the upper-left pattern encodes the 200 Hz sinusoid.

$\pi/2$. The comparison elements within that distance display black, indicating correlation. Between the taps t_1 and t_2, the sinusoid is shifted progressively from $\pi/2$ to π with respect to the sinusoid at tap t_0. In this region, the comparison elements display white, indicating anticorrelation. In the region of strong damping, a progressive phase shift is observed (between t_2 and t_3, a shift from π to $3\pi/2$; between t_3 and t_4, a shift from 2π to $5\pi/2$; etc.) resulting in alternate correlation and anticorrelation stripes. Due to the exponential decrease of the traveling-wave velocity, the distance needed to shift the sinusoid by $\pi/2$ decreases ($t_1 - t_0 > t_2 - t_1 > t_3 - t_2 > t_4 - t_3$), resulting in the curvature of the stripes. The output image is thus a pattern of curved stripes that is symmetric with respect to the diagonal, for a 0 radian IPD.

Parts (b) and (c) show the network result for IPD's of $\pi/2$ and π, respectively. When the signal is delayed at the input of one cochlea, the black correlation stripes are shifted from the diagonal toward the cochlea where the signal is delayed. Since the spatial disparity between the two traveling waves is proportional to the temporal delay between the two ears [2], the amount of shift is a measure of the interaural time delay. If the signal is delayed by an amount corresponding to a phase shift of π, the spatial waves are in counterphase and the solution is again symmetric with respect to the diagonal, with an anticorrelation stripe on the diagonal. Because the cochleas are acting like delay lines, the patterns reflect lateralization information using only the cochlear outputs, without requiring explicit neural delays.

Nonlinearities of the circuit model of sensory transduction act to compress the input signals; as a result, the output of the stereausis chip is relatively insensitive to the absolute and relative amplitudes of the signal inputs. In addition, the winner-take-all network in each comparison element introduces a logarithmic compression, which further reduces the amplitude information in the output of the stereausis chip. Alternative circuits for sensory-transduction models and comparison that preserve amplitude information may be preferable for certain applications.

Fig. 5 shows the response to the sum of two pure tones (200 Hz and 2 kHz) presented with zero IPD. There are four distinct regions in the chip output. The upper-left corner shows the strong autocorrelation of the 200 Hz sinusoid, the lower-right corner the strong autocorrelation of the 2 kHz sinusoid. The other regions show the weak cross-correlation of the 200 Hz and 2 kHz sinusoids. In this way, the stereausis representation naturally segments monaural signals into distinct spectral regions.

VI. Response to Complex Sounds

Binaural processing is known to improve sound perception in complex acoustic environments [10]. In this section we show the chip response to white noise and to speech sounds, two important classes of complex sounds.

The autocorrelation function of white noise has a single peak at $\tau = 0$, and has no significant energy for $\tau \neq 0$. If the silicon cochleas were simple delay lines with no resonant behavior, the stereausis-chip response to monaural white noise would be a single black correlation stripe on the diagonal, and random responses everywhere else. In fact, for low-amplitude white noise, the chip produces this single diagonal stripe. Fig. 6(a) shows chip response to high-amplitude monaural white noise; in addition to the black diagonal correlation stripe, there are white bands of anticorrelation that border the diagonal. These anticorrelation bands are a result of the slightly resonant, band-limited cochlear response. The exponential nature of the cochlear delay line results in a natural invariant of the representation: at any position, the distance along the line representing half-cycle at the resonant frequency at that position is constant. For that reason, the resonant anticorrelation sidebands are equally spaced from the central correlation band, independent of position.

Fig. 6(b) shows the response of the chip to white noise input delayed in one cochlea relative to the other by 700 μs. As expected, the stripe is shifted with respect to the diagonal according to the delay between the two ears.

Psychoacoustic experiments show that we perceive a faint, but distinct, pitch in response to a sum of two time-delayed, correlated noise signals [11]. This signal is also known as comb-filtered noise. Fig. 6(c) shows the response of the stereausis chip to this stimulus, with a 700 μs delay. In addition to the diagonal black correlation band, there are curved correlation bands symmetrically displaced from the diagonal. The position of these stripes changes with delay between the two signals, and thus encodes the pitch of the noise.

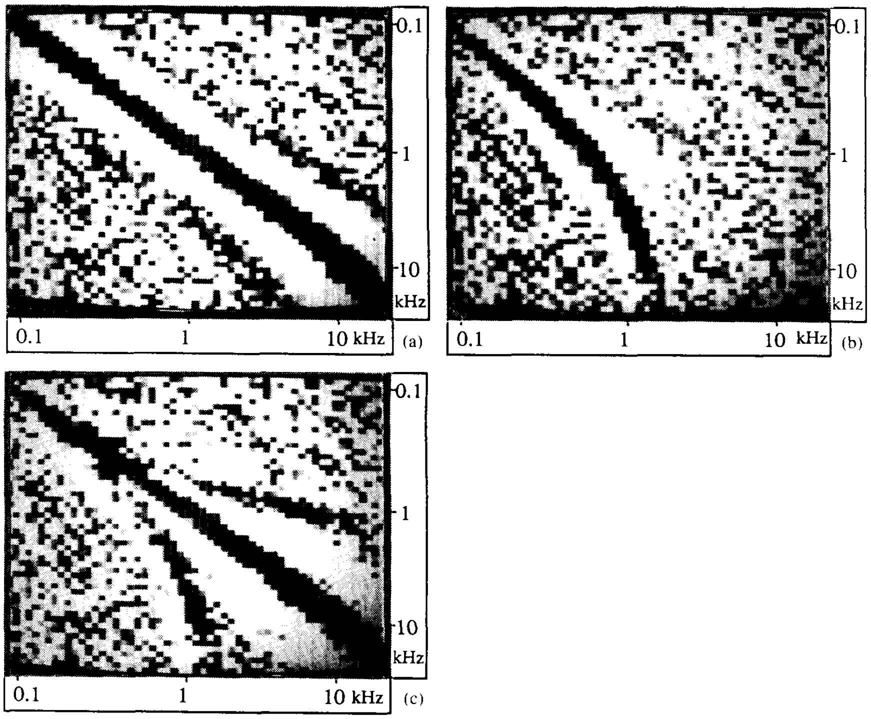

Fig. 6. Photographs showing the output of the stereausis chip in response to white noise. (a) Chip response to the presentation of identical white noise signals to both inputs. (b) Chip response to the binaural presentation of white noise, delayed 700 μs into the left input. (c) Chip response to the presentation of identical comb-filtered white noise to both inputs, created by the addition of a white noise signal to a delayed (700 μs) replica of itself.

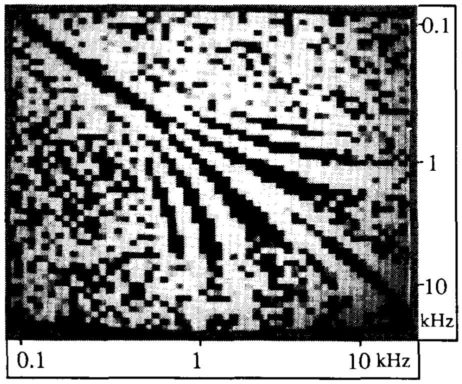

Fig. 7. Photograph showing the output of the stereausis chip in response to zero-delay white noise and a 1 kHz sinusoid presented binaurally with a π-radian IPD. The sinusoidal stimulus dominates in the range of its peak response, giving an anticorrelation band along the diagonal, and a symmetric off-diagonal correlation structure. The broad-band noise dominates the lower right and upper left regions with a correlation band along the diagonal.

An interesting property of correlation representations generally, and the stereausis representation in particular, is that the pattern obtained for the white noise (one stripe) is much simpler than the one obtained for simpler stimuli (several stripes for a sinusoid). The stereausis chip enchances the detection of a pure tone in noisy environments, as shown in Fig. 7. Here the sinusoidal response is clearly visible in the middle of the display, while the noise response dominates at higher and lower frequencies. The output patterns depend both on the spectral information of the different sounds arriving at the two ears and on the localization of these sounds in the environment.

Speech sounds either are periodic signals (e.g. voiced speech) consisting of harmonics of the fundamental frequency that is generally perceived as the pitch or are noisy and aperiodic (e.g. unvoiced speech). Each sound produces a different pattern at the output of the stereausis network, depending on its composition (identity of the sound) and on the localization information. Figs. 8 and 9 show the response of the stereausis network to natural speech sounds (vowel /i/ and three fricatives, /s/, /sh/, /h/). Real-time, continuous speech results in patterns that continuously change between sounds.

The curved stripes in the vowel patterns correspond to resonances in the vocal track called speech formants. Differences between the vowel patterns are encoded in the position of the different formant stripes. The patterns result from the combination of harmonics in the sound composition; therefore, stripes extending across the display correspond to the perceived pitch. When the vowel is whispered (Fig. 8(c)), the sound corresponds to cavity resonances excited by turbulant flow in the vocal track. There are no pitch pulses, so the low-frequency portion of the signal is lost.

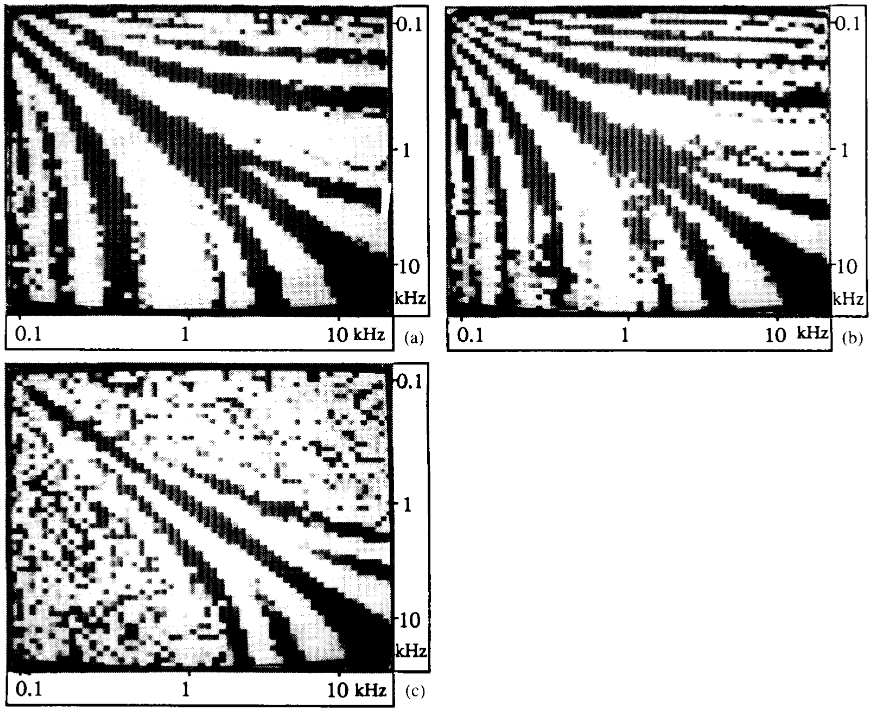

Fig. 8. Photographs showing the output of the stereausis chip in response to the natural speech (vowel /i/) (a) pronounced by a male speaker, (b) pronounced by a female speaker, and (c) whispered by a male speaker. Information about the formant (vocal-track resonance) is contained in the lower-right quadrant, and is similar for both voiced and unvoiced speech ((a) and (c)). Pitch information is captured in the upper-left corner of the display. For the female speaker, the higher pitch frequency is evident, as is the complex interaction of pitch information with resonances in the off-diagonal regions.

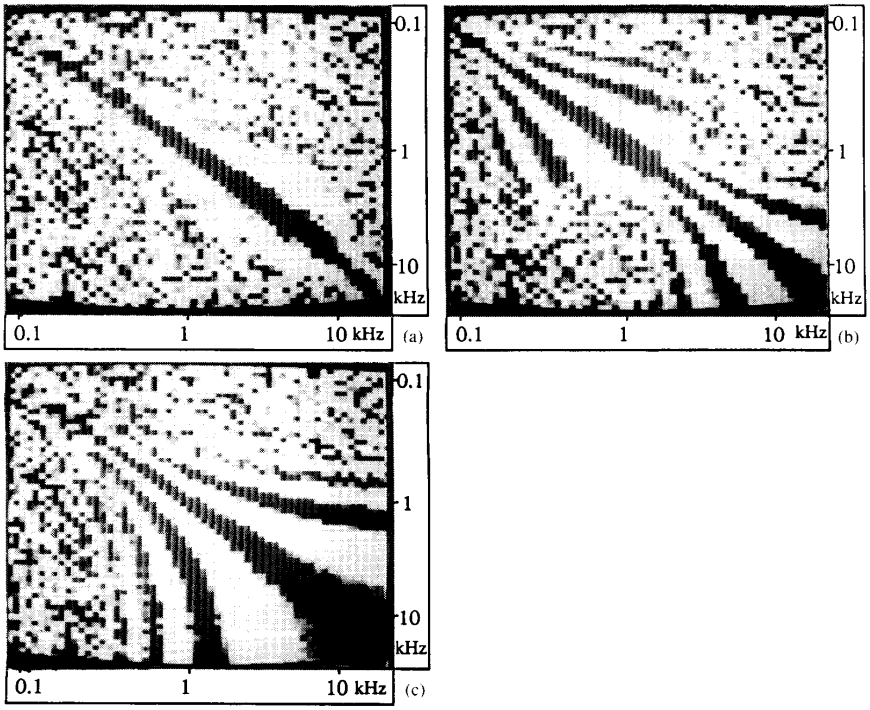

Fig. 9. Photographs showing the output of the stereausis chip in response to the natural speech (fricatives) for a female speaker: (a) /s/, (b) /sh/, and (c) /h/. The lip resonance is just visible in the lower-right corner of (a). Both lip and formant resonances are conspicuous in (b). Only the formant resonance is visible in (c).

Although the position of the formants varies slightly from one person to the other, the patterns are similar, suggesting that the stereausis-network response encodes a set of the distinctive features of speech. We would envision the stereausis chip followed by at least one level of feature extraction network which is tuned for spatiotemporal behavior of the chip output (duration of the pattern, direction of change of the stripes, etc.) using temporal and spatial derivative operations. The recognition of speech elements might be based on patterns of these extracted features and of their sequence. This architecture provides a promising approach to preprocessing for speech recognition systems.

Acknowledgment

The authors thank R. Lyon for many helpful discussions. Chip fabrication was provided by DARPA through the MOSIS fabrication service.

References

[1] G. Loeb, M. White, and M. Merzenich, "Spatial cross-correlation: A proposed mechanism for acoustic pitch perception," *Biol. Cybern.*, pp. 149-163, 1983.

[2] S. A. Shamma, N. Shen, P. Gopalaswarmy, "Stereausis: Binaural processing without neural delays," *J. Acoust. Soc. Amer.*, vol. 86, pp. 989-1006, 1989.

[3] E. F. Evans, "Functional anatomy of the auditory system," in *The Senses*, H. B. Barlow and J. D. Mollon, Eds. Cambridge, England: Cambridge University Press, 1982, p. 251.

[4] R. F. Lyon and C. Mead, "Cochlear hydrodynamics demystified," Caltech Computer Sci. Tech. Rep., Caltech-CS-TR-88-4, California Institute of Technology, Pasadena, CA, Feb. 1988.

[5] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 1119-1134, 1988.

[6] J. P. Lazzaro and C. Mead, "Circuit models of sensory transduction in the cochlea," in *Analog VLSI Implementations of Neural Networks*, C. Mead and M. Ismail, Eds. Norwell, MA: Kluwer Academic Publishers, 1989, pp. 85-101.

[7] J. P. Lazzaro, S. Ryckebusch, M. A. Mahowald, and C. Mead, "Winner-take-all networks of $O(n)$ complexity," *Advances in Neural Information Processing Systems 1*, D. Tourestzky, Ed. San Mateo, CA: Morgan Kaufmann, 1988, pp. 703-711.

[8] M. A. Mahowald and C. Mead, "Silicon retina," in *Analog VLSI and Neural Systems*, C. A. Mead. Reading, MA: Addison-Wesley, 1989, pp. 257-278.

[9] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[10] N. I. Durlach, "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory*, vol. II, J. V. Tobias, Ed. New York: Academic Press, 1972, pp. 371-462.

[11] A. J. Fourcin, "The pitch of noise with periodic spectral peaks," in *Proc. Fifth Int. Congress Acoust.*, (Leige), vol. Ia, 1965, B 52.

*

Carver A. Mead, Gordon and Betty Moore Professor of Computer Science, has taught at the California Institute of Technology for more than 30 years. He has contributed in the fields of solid-state electronics and the management of complexity in the design of very large scale integrated circuits, and has been active in the development of innovative design methodologies for VLSI. He has written, with Lynn Conway, the standard text for VLSI design, *Introduction to VLSI Systems*. His recent work is concerned with modeling neuronal structures, such as the retina and the cochlea, using analog VLSI systems. His new book on this topic, *Analog VLSI and Neural Systems*, has recently been published by Addison-Wesley.

Prof. Mead is a member of the National Academy of Sciences and the National Academy of Engineering, a foreign member of the Royal Swedish Academy of Engineering Sciences, a Fellow of the American Physical Society, and Life Fellow of the Franklin Institute. He is also the recipient of a number of awards, including the centennial medal of the IEEE.

*

Xavier Arreguit was born in Oviedo, Spain. He received the M.S. and Ph.D. degrees in electrical engineering from the Federal Institute of Technology, Lausanne, Switzerland, in 1985 and 1989 respectively, where his research dealt with compatible lateral bipolar transistors in CMOS technology. In January 1990, he joined the California Institute of Technology as a Research Fellow. His current research interests include low-power integrated analog circuits and analog VLSI models of neural networks.

*

John Lazzaro received the B.S. degree in electrical engineering and computer science from the University of Pennsylvania in 1984 and the M.S. and Ph.D. degrees in computer science from Caltech in 1986 and 1990, respectively.

He is an Assistant Research Professor at the University of Colorado at Boulder. He did the research shown in this paper while he was a postdoctoral fellow in the computer science option at Caltech. His research involves analog VLSI models of biological auditory, visual, and cardiovascular processing. His research articles have appeared in *Neural Computation*, *The Proceedings of the National Academy of Sciences*, and the proceedings of several neural network conferences.