Analog VLSI Hearing Systems

Richard F. Lyon
Apple Computer, Inc.
20525 Mariani Avenue
Cupertino, California 95014

Abstract: Animals hear through complex mechanical–neural signal processing systems. We would like to be able to duplicate this functionality in machines, for purposes of recognizing speech and for other sound-related tasks. Computational models on digital computers offer one method for exploring models of auditory processing, but in this paper we discuss the use of analog VLSI systems as a promising alternative modeling medium. The analog cochlea model that we presented earlier is being modified and extended to include closed-loop adaptation and hair-cell functions. Later processing stages being investigated will exploit information in the fine time structure (waveform synchrony) that is apparent in the cochlea's output signals. Per-channel cross-correlation for binaural localization and auto-correlation for pitch and timbre representation are two ways of utilizing fine time structure that we have previously experimented with in digital models. The analog medium imposes different constraints on these stages, the most important result of which is the need to incorporate adaptation at all levels, just as in the biological system.

Introduction

The future existence of machines that hear depends on our ability to understand hearing and to implement effective real-time models of hearing. Modeling the cochlea (inner ear) has led in recent years to moderate performance gains in several research speech recognition systems, but research in hearing and its applications to such problems has been hampered by the tremendous computational burden of complex signal processing models on digital computers. In order to make progress toward long-sought breakthroughs in this area, we need to be able to more quickly evaluate our theories about hearing by testing them in use with real sounds in real applications, preferably in real time. Analog VLSI technology offers an inherently real-time implementation medium with a variety of interesting properties.

As a first step in building machines that hear, we have implemented an analog electronic cochlea that incorporates much of the current state of knowledge about cochlear structure and function. We are currently extending this CMOS VLSI

hearing system to include closed-loop adaptation and other models of neural processing in the auditory nuclei of the brain stem.

Background

The problems we must solve to build perception machines are mostly similar to those that nature had to solve biologically in the evolution of intelligent animal behavior. The key problems in sound perception are to cope with a very wide dynamic range of loudness and to separate sounds on the basis of their properties, such as frequency content and time structure. We believe that, by developing circuits that solve the same problems using imprecise analog components, we will increase our understanding of how animals hear. The levels of the auditory system beyond the cochlea, all the way through linguistic processing in the cortex, present a range of interesting challenges for this work.

The interpretation of a large and complex literature of conflicting ideas on the physiology of hearing continues to be an important aspect of this research; we present our own view of how some of the controversies in this area can be resolved in a coherent functional framework. The approach used to model the nonuniform fluid-dynamic wave medium of the cochlea as a cascade of filters is based on the observation that the properties of the medium change only slowly and that wave energy is therefore not reflected to any significant degree. The effect of active outer hair cells is included as a variable negative damping term; the variable damping mechanism is shown to be effective as a wide-range *automatic gain control* (AGC) associated with a moderate change in sharpness of tuning with signal level. The system is not highly tuned in the sense of a high-Q resonator, but rather achieves a high-gain *pseudoresonance* by combining the modest gains of many stages. Sharp *iso-output tuning curves* result from the interaction of the adaptive gains and the filters, as has been observed experimentally in both neural frequency threshold curves and basilar membrane iso-velocity curves.

In this view, there is no conflict between the observed sharp tuning curves and the corresponding broad filter transfer functions derived from the hydrodynamic model. Various other nonlinear and masking effects are similarly explained with the same active-adaptive system framework.

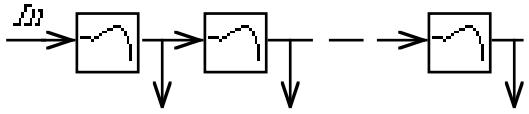
Cochlear Model Summary

The analog cochlear model presented earlier is based on modeling wave propagation in the cochlea as a cascade of simple filters that are linear in the short

term but that can non-linearly adapt their characteristics in response to perceived sound level; this section briefly summarizes the model.

The cascaded filter stages, which may also be viewed as delay-line stages, are second-order lowpass filters (i. e., just a pair of poles, no zeros) with nearly maximally-flat response. Later stages of the cascade have longer time constants, or lower cutoff frequencies, to model the frequency–place mapping of the cochlea. To model wave amplification due to active outer hair cells in the cochleas, the filter stage Q values are increased from the maximally-flat condition of Q=0.707 up to as much as Q= 0.9. The modest gain peaks thereby introduced in each stage combine to yield a significant peak in the overall cascade transfer function. The peak gain of this pseudoresonance is variable over a range of 50 dB or more by small changes in the Q values of the stages.

Figure 1. A Filter Cascade as used in the Cochlear Model.



A standard CMOS VLSI technology (a logic process, not a more specialized memory or analog process) is used to produce transconductance amplifiers, which are used in combination with fixed capacitors to make adjustable second-order filter stages. The time constants (or corner frequencies) and Q values of the filter stages are set by the bias voltages applied to current sources in the transconductance amplifiers. In the micropower subthreshold region where the circuits are operated (to achieve the long time constants needed for audio processing), bias currents and time constants are exponential functions of bias voltages and of transistor offset voltages (usually known as threshold variations).

Small transistor mismatches can lead to severe gain variations in our cascade structure, so the automatic adaptation of the Q values is necessary not only to accommodate a wide dynamic range of sound loudness, but also to accommodate a range a variation in the primitive components. We believe that similar constraints apply to the architecture of biological sensory systems, and that this need to adapt to low-quality components explains the existence of adaptation at all levels of neural systems.

Taps along the cascade (either every stage or every several stages) bring out channels of filtered sound pressure as seen at the inner hair cells, which transduce mechanical vibrations into neural signals. The information beyond this level is represented in the biological system as action potentials on the cochlear nerve, with statistics that faithfully represent the detailed half-wave rectified filtered

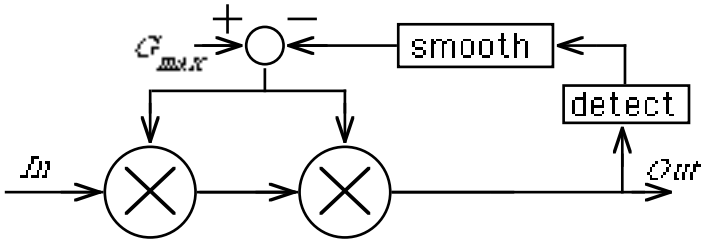
waveforms (at least up to frequencies of several kilohertz). Cochlear models have traditionally ignored this *fine time structure* information, or treated it in very simple ways, and have concentrated instead on reducing the information to a *rate* of nerve firing. There is mounting evidence in the hearing research community that the most important information about sound is to be found in the time structure, rather than in the rate. Therefore we seek new and improved models of the signal processing functions of the hair cells, the primary auditory neurons, and the auditory nuclei of the brainstem, which take input from the cochlea and send signals back to the cochlea to effect adaptation. These levels are fairly tightly coupled and must be jointly modeled; in the sections that follow, however, we try to discuss them separately.

Closed-loop Adaptation

The cochlea chips built and tested so far have not included on-chip closed-loop control of the Q bias voltages, so the Q values were set by off-chip manual adjustment to establish reasonable test and measurement conditions. Our previously described computational model of coupled AGC in the cochlea used four stages of variable gain after a time-invariant filter cascade. Further background and results on the use of AGC in a cochlear model has also been presented . Using a detection and coupled smoothing approach similar to those, but controlling the analog time-varying filter Q's and gains instead of separate gain stages, would yield similar overall amplitude compression results, but possibly problematic dynamics.

The overall amplitude response of sensory systems (perceived or measured response vs. input stimulus level) is often viewed as approximately logarithmic, or as a power law with a low exponent, over a fairly wide range of stimulus levels. An AGC that is built using a forward-path gain element that divides by the smoothed detected output level implements a square-root compressor—that is, a power law with an exponent of one-half, which is not compressive enough to be realistic. By modifying such an AGC to use a cascade of two identical gain elements, each controlled by the same feedback signal from the output level, the overall response becomes a cube-root compressor, which is becoming more realistic. As even more such gain elements are cascaded together, the response quickly approaches logarithmic, which is hard to distinguish from a power law of low exponent. By feeding back spatially smoothed output level signals into the filter cascade, the analog cochlea's coupled AGC uses many filter stages as variable gains, and thereby approaches an overall logarithmic response.

Figure 2. A simple AGC loop.



For this kind of AGC response, it is not important that the individual variable gains be the reciprocals of the detected output level. To make the system more realistic, especially at low levels, the gains must have some fixed maximum value when the output level is zero. The output of the system will grow initially as the square-law response of a non-coherent power detector for very weak inputs, and the effective power-law exponent will quickly decrease until at high input levels the system response is approximately logarithmic.

Therefore, the circuits that we are experimenting with determine the maximum Q or maximum gain from fixed (or manually adjusted or very slowly adapting) bias voltages, and subtract off an amount in proportion to the detected output level. The detected output level is fed back through a time-space smoothing filter made of capacitors and a spatially-coupled resistor network such as the ones being used in Mead's retina chips .

The dynamics and stability of such AGC systems are tricky due to the inclusion of the filter dynamics in the AGC loop. We do not yet have a satisfactory analysis or conclusion on this issue.

Hair Cells and Primary Auditory Neurons

Each output tap from the filter cascade provides a voltage representing the pressure wave in the cochlea. These signals must be further processed to yield a representation of basilar membrane motion as seen by the inner hair cells, and to then produce a representation of the hair cell's output. In the biological system, the hair cells detect some combination of displacement and velocity, depending on the characteristic frequency and the viscosity of the fluid that is bending the hairs. The hair cell's output is a release of neurotransmitter to the primary auditory neurons (of the spiral ganglion), which then convert the signal into a statistical form as action potentials. We may choose to have the output of the electronic cochlea either as an analog representation of the neurotransmitter release, or as a pulse representation, depending on the plan for later processing stages.

Conversion of the tap outputs to hair cell stimuli involves a spatial differentiation (difference between adjacent taps) to convert pressure to acceleration, followed by a time integration or two to convert to velocity or displacement. A direct analog implementation in this sequence would be very problematic, however, due to offsets in the differentiator that would then be integrated. A better approach is to combine these two linear operations and the following nonlinear hair cell detection and adaptation function into a single functional approximation that has a simple circuit implementation.

John Lazzaro of Caltech has already built and tested several generations of hair cell and neuron circuits, based on differentiator-like circuits with hysteresis, which will be the subject of a later work. His neuron circuits are those developed by Carver Mead of Caltech for his retina model—they are essentially integrate-to-threshold models of retinal ganglion cells that would be appropriate for use as spiral ganglion cells as well. Inclusion of ideas such as leaky integrators and refractory periods longer than pulse lengths has been seen to improve neuron realism, and could easily be adapted to the analog systems as well. In any case, a rapid (< 1 ms) adaptation (gain control) within the hair cell model is essential to emphasize onsets and produce the precise timing needed for both binaural and monaural processing.

The relative advantages of basing later stages of processing on pulse representations vs. continuous analog representations are not yet well understood. The pulse representation lends itself to logic-based correlation and pulse or digital delay techniques, which may be more compact but more noise-producing and power-consuming than continuous analog techniques. The analog techniques employ continuous representations, with delays implemented as filter stages and with correlation implemented as bilinear multipliers. The pulse representation may need several pulse streams per hair-cell channel in order to achieve a reasonable signal-to-noise ratio in the statistical representation. Mead and Lazzaro, in collaboration with the author, have already experimented with a few versions of analog and pulse correlators; these stages of processing are discussed briefly in the following two sections.

Binaural Processing

Multiple sound signals, such as speech and interfering noises, can be fairly well separated out, localized, and interpreted by human listeners with normal binaural hearing. In order for machines to approach this level of performance, a rich representation of binaural sounds is essential. Brain structures of the auditory brainstem are essentially two-dimensional sheets with different properties of the

signal mapped along different spatial dimensions—we need to implement model systems that produce representations of similar dimensionality, which is also the dimensionality of a silicon chip. Cross correlating corresponding channels from two cochleas to produce a two-dimensional image, as originally suggested by Jeffress in 1948 , is the obvious first step.

The most promising technology for binaural processing presently seems to be to use the same delay-line building blocks that make up the cochlea as the delay elements in a cross-correlator to extract time-of-arrival-difference cues for lateralization. Separate circuits, also based on the continuous analog representation, can be used for spectral amplitude comparisons, which are useful for both lateral and vertical localization (assuming appropriate *external ears* for microphones). Just as in the cochlea, adaptive AGC functions need to be integrated into these circuits, not because the signal has a wide dynamic range at this level, but because the system must adapt out its own wide range of gain variation.

In the biological system, binaural comparisons are done in the superior-olivary complex (time difference in the medial superior olive and magnitude difference in the lateral superior olive). The superior-olivary complex is the first place in the auditory nervous system that receives inputs from both ears, and is also the origin of the neural signals that return to the cochlea to adjust the activity of the outer hair cells. This arrangement probably helps to keep the gains of the two ears changing together, rather than separately, to make it easier to do amplitude comparisons. The same technique should be employed in our analog models, though each of two cochleas should also do some of its own slow adaptation to adjust for differences between them.

Monaural Processing

A structure similar to the binaural cross-correlator can be used monaurally to extract pitch and timbre cues. The rich two-dimensional representation that such a model produces has been shown by Weintraub to be quite useful for separating out the sounds of two people speaking at once, even without the advantage of binaural signals.

An important difference between the binaural and monaural correlators is the range of delay values that must be implemented. In the binaural case, delays of only one millisecond or less are adequate to span the range of interaural delays for a human-size head; a delay–bandwidth product of about 5 cycles is probably adequate to maintain reasonable lateral cue sharpness. In the monaural case, our experiments with correlagram representations lead us to want a delay-bandwidth

product of 10 to 20 cycles, with total delays of at least 20 ms. The delay-bandwidth product of a cascade of second-order delay stages is about $0.25 \bullet N^{3/4}$ for a delay line of N identical stages (or for a delay line of about N stages per octave in a variable-delay structure). Thus only about 50 stages (per channel) of delay are needed for the binaural cross-correlator, while several hundred stages are needed for each channel of auto-correlation to achieve high-resolution pitch and timbre cues. This may be sufficient motivation to look for alternative representations and delay structures—the cost of delay lines for digital or pulse representations grows only linearly with delay-bandwidth product.

In order to cover a wide range of delays in the monaural correlator (for representation of formant periods, pitch periods, rhythms, and other time-domain effects of various scales in a uniform representation), the delay-line could be built of stages with exponentially increasing time delays, just as in the cochlea, rather than as a constant delay per stage which might be more appropriate in the binaural correlator. This variable-delay scheme is simple to implement with analog filter stages, and is not consistent with the above-mentioned pulse and digital techniques. Hence, we have an interesting trade-off space to explore in designing these structures. Little is known about how the biological system deals with these issues, or even about what monaural processing functions are done in what brain structures.

Conclusions

This short paper has rambled through some ideas that we hope will help outline the issues involved in using analog VLSI techniques to build machines that hear. This line of work is continuing with the involvement of several people at Apple and Caltech and elsewhere, in both simulations and working real-time chips. We expect to be able to demonstrate exciting real-time auditory correlagram displays soon, so stay tuned. Machines that understand what they hear will follow eventually.

References

[1] Richard F. Lyon and Carver Mead. "An Analog Electronic Cochlea", *IEEE Trans. ASSP* **36**:7 pp. 1119–1134, July 1988.

[2] Richard F. Lyon and Niels Lauritzen, "Processing Speech with the Multi-Serial Signal Processor", *Proc. ICASSP 85*, Tampa, March 1985.

[3] Richard F. Lyon and Lounette Dyer, "Experiments with a Computational Model of the Cochlea", *Proc. ICASSP 86*, Tokyo, April 1986.
[4] Carver Mead. *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley (in press).
[5] Lloyd A. Jeffress. "A Place Theory of Sound Localization" J. Comp. Physiol. Psychol., **41**, pp. 35–39, 1948.
[6] Mitchel Weintraub. "The GRASP Sound Separation System" *Proc. ICASSP 84*, San Diego, March 1984.
[7] Richard F. Lyon, "Computational Models of Neural Auditory Processing" *Proc. ICASSP 84*, San Diego, March 1984.