Computational Models of Neural Auditory Processing

Richard F. Lyon
Fairchild Laboratory for Artificial Intelligence Research
4001 Miranda Ave.
Palo Alto, CA 94304

Abstract

Explicit neuron firing models are investigated for use in computational modeling of auditory processing. Models for primary auditory neurons are driven by the receptor signal from a hair cell model, which is driven in turn by a filtering model of basilar membrane motion. The output of the primary auditory neuron model is a times-of-firing representation of neural signals. Certain types of processing, such as auto-correlation and cross-correlation, are very simple with this representation, not requiring multiplication. The neuron model used is a leaky-integrate-to-threshold model with a refractory period. Several neurons are modeled at each hair cell, or filter channel. It is found in experiments with these models that the detailed time-of-firing information contains most of the cues of speech formants, pitch, direction, etc. The more conventionally studied firing rate *vs.* place representation misses important aspects of these cues. Models of pitch perception, binaural directional perception, and sound separation are being based on the cochlear and neural models. The models are all implemented as computational algorithms, and are used in support of related speech recognition and hearing research.

1. Introduction

The brain receives all of its inputs in the form of neuron firings on specific fibers at specific times. We explore the application of this type of discrete-event signal representation to the problem of hearing, with a view to using it as a practical representation of sound signals for use in speech recognizers. The problem naturally fits time-and-place domain algorithms, with place representing specific fibers or groups of fibers. The algorithms we want are simulations, or computational models; hence, we reject at the outset any model that depends heavily on a mathematically abstract concept such as frequency, or instantaneous firing rate, or probability of firing. We discuss our experience with simulations of cochlear models, models of primary auditory neuron firings, and models for further processing of nerve firings to access cues such as pitch, formants, binaural time of arrival difference, etc.

2. Motivation and Approach

The mammalian auditory system represents an amazing technical mechanism that can detect, separate, and recognize an amazing variety of complex waveforms carried as waves in air. Over the last century or so, modeling and describing our mechanisms of hearing has been an active scientific pursuit. There has been tremendous progress in understanding hearing, especially from the points of view of psychophysics and physiology. Despite this progress, scientists and engineers studying speech and hearing still do not typically use hearing models much more sophisticated than the Fourier analysis model originally espoused by Ohm and Helmholtz. This crude model makes it difficult for them to relate new models and experimental results to details of reality. The present line of investigation seeks to apply the modern tools of signal processing and discrete simulation of physical systems to provide a new substrate of front-end analysis techniques that are more realistically related to hearing. These new techniques provide an alternative to Fourier analysis that will hopefully allow more progress to be achieved in studies of hearing. Having a complete runable and testable set of model algorithms will provide a place for plugging in and testing new and improved models for the various mechanisms of hearing. These algorithms, or computational models, are also expected to be very good candidates for the front end of a high performance speech recognition system.

The basic approach in this work is to look at what sounds do to you when they impinge upon your ears. There are many layers of acoustic, mechanical, electrochemical, and neural processes that can be examined and modeled *from the front in*, or as *bottom-up, data-driven* algorithms. As each layer is modeled by an algorithm, experimented with, and compared with what is known of reality, new insights can be gained about the representation of information at each layer; eventually, the synergies between the layers of processing become visible. The systems of algorithms that result are not easy to describe in the conventional language of mathematical signal analysis, since the layers are not generally linear, or time-invariant, or otherwise ideal. These algorithms are termed computational models, as they represent computations that attempt to mimic the hearing system; they are not primarily useful for *describing* or *analyzing* the hearing system. Previous reports on this line of work [1, 2] cover some of the basic functions of the ear, and some models of internal processing, but do not cover any explicit modeling of neuron firings. This paper describes some explorations of explicit neuron models and the implications of using the discrete neural event representation of signals.

A specific goal of this work is to investigate how to effectively use the fine time structure of the signals transduced by the cochlea. A major thesis is that this information is very important to signal separation; that is, the brain uses more than short-time power spectrum information—it is sensitive to phase in useful ways.

We are also motivated in these efforts by the observation that regions of brain tissue, which are basically two-dimensional sheets with parallel wiring through the third spatial dimension, are tonotopically (or cochleotopically) organized in one dimension; it is fascinating to consider how the other dimension is utilized. The two specific tech-

CH1945-5/84/0000-0392 $1.00 © 1984 IEEE

niques that have been investigated, auto-correlation for periodicity perception and cross-correlation for binaural lateralization, both produce two-dimensional "images" which are reasonable candidates for the kind of information that could be projected onto cortical auditory areas, or other neural structures at intermediate levels of the nervous system. These images change relatively slowly, compared to the bandwidth of the outputs of the primary auditory neurons. Perhaps there are many more types of processing, at all levels of the nervous system, that convert fine time structure to slowly changing spatial structure.

3. Cochlear Filtering and Transduction Models

This investigation is based on a version of the cochlear filtering, compression, and detection models presented in [1]. The filtering stage is an unusual form of filterbank, consisting only of a single cascade of second-order canonic sections (this is an optimization of the cascade/parallel filterbank of [1], with particular parameters that allow rearrangement of the poles). For a given amount of computation, this model allows about three times as many channels as a typical sixth-order per channel parallel filterbank. The transfer functions are also much better fits to the real transfer functions of basilar membrane motion, since this filter form is derived from a physical model of wave propagation in the cochlea.

The transduction process is modeled by a compressive network (several stages of coupled automatic gain controls) and an explicit mechanistic model of detection, adaptation, and smoothing in the hair cell. The coupled AGC network is greatly simplified from that presented in [1], involving only connections between nearest neighboring places. Several cascaded stages (as opposed to the nested stages proposed in [1]) serve to approximately model various possible mechanisms of adaptation, with different time constants and spatial extents, in the middle and inner ear. These stages reduce the dynamic range seen by the hair cell model, and hence reduce the amount of point-wise compression that each hair cell must do; the result is that time-domain contrast between loud and very-loud are maintained, whereas that contrast would be lost without the compression stages.

4. The Hair Cell Model

The hair cell model is taken almost exactly from Allen [3], by assuming that that model applies identically to every channel, with no interactions. Since that model's adaptation behavior depends on cell membrane currents, both resistive and capacitive, it might be a better assumption to add some membrane currents between adjacent cells. This would have much the same effect as a fast fairly-local coupled AGC stage, which was used instead. Allen's model attempts to explain adaptation only in intervals less than 50 msec, and so ignores the need for adaptation stages in addition to the hair cells. His hair-cell model converts mechanical displacement to a variable resistance through a hyperbolic cosine (soft half-wave) nonlinearity; this resistance controls the flow of current between an external battery (endolymph potential) and the cell's internal "receptor potential". Charge inside the cell is stored on the membrane capacitance, and leaks out through the membrane resistance. As the voltage inside the cell increases with stimulation, the voltage across the variable resistance declines, reducing the instantaneous local gain of the transduction process. Note that the gain change is after the nonlinearity; very large inputs will drive the variable resistance in almost an on/off mode, so that timing will be preserved, but amplitude information will be suppressed.

According to the model, the stimulus seen by the primary auditory neurons is the current out of the cell membrane, including both resistive and capacitive components. The resistive component of this receptor current is proportional to the so-called receptor potential, which has been measured experimentally and conforms to the model. The capacitive component is proportional to the derivative of the receptor potential, and hence contains more high frequency information, as needed to account for the observed levels of nerve firing synchrony. Finally, the current output is smoothed with a time constant of about 150 microseconds, modeling the diffusion of ions within the hair cell, and reducing waveform synchrony above several kilohertz.

5. Primary Auditory Neuron Models

There are many types of neuron models, including simple logic-gate models, dendritic tree circuits, and stimulus/rate functions. But to generate explicit firing data from a receptor signal, a time-domain stochastic simulation type of model is needed. The most popular such model is the "integrate to threshold" (ITT) model, in which a stimulus input is integrated until the integral reaches a threshold (which may have been randomly chosen); then the neuron fires, the integral is reset to zero, and integration continues immediately. For such a model, average firing rate and instantaneous probability of firing are well-defined, and simply proportional to the (non-negative) stimulus input. A post-stimulus-time or period histogram of firings from such a model will always have exactly the same shape as the input stimulus, plus Poisson noise.

To enforce a refractory period, or minimum delay between firings, a common practice is to simply hard-limit the input to a level that corresponds to the appropriate maximum firing rate. Thus the delay from the onset of a strong stimulus to the first firing can be as long as the refractory period; this is unrealistic, and causes a loss of important information about the exact time of onset of strong stimuli. A better model of refractoriness is to simply wait for a prescribed (possibly randomized) time after each firing before continuing to integrate from the reset level. Thus, all neurons that happen to not be in their refractory state can fire arbitrarily soon after the onset of a strong enough stimulus. The result is that neurons typically become phase-locked (or injection-locked) to some component of the fine time structure of strong stimuli. Neuron firing histograms become sharpened, as stimulus peaks "capture" firing times. The history dependence introduced by this simple refractory mechanism makes it difficult to say much meaningful about instantaneous probability of firing, or even average firing rate; we happily abandon these concepts in favor of studying the actual firing behaviors.

A popular refinement of the ITT model is the leaky-ITT model, in which the integrator is replaced by a leaky integrator (a one-pole lowpass filter with large finite gain at DC). In addition to the gain/threshold and refractory-time parameters of the ITT model, the leaky-ITT model also has a time-constant parameter. In studies of populations of neurons, all these parameters are chosen randomly within a small range, to avoid totally deterministic identical behaviors; we also sometimes introduce a randomized reset level in our studies, though this is of little effect.

Just as the refractory model led to a way to "clean up" information about timing of strong stimuli by sharpening histograms, the leaky-ITT model can help clean up the representation of weak stimuli. When no signal is present, the rest-level receptor signal, the output of the hair-cell model, is a small positive value. If the leak is such that the rest-level times the DC gain of the leaky integrator is very near threshold, then the rest firing rate of the neuron will be zero or slightly positive. Under this condition, the neuron will be exquisitely sensitive to slight variations above and below the rest stimulus level, as its leaky integrator output is usually sitting just below threshold. The degree of phase locking will be much higher than in a straight ITT model with its higher rest firing rate. It is not known to what extent real auditory neurons conform to this model. Perhaps the observed population of low-spontaneous-rate neurons manage to adjust themselves to operate in this sensitive mode, while high-spontaneous-rate neurons do not.

In our simulations, we use a single randomized population over a wide range of rest rates (0 to 60 firings per second, averaging about 25). At high stimulus levels, the average firing rates are around 150 per second; peak rates, averaged over 5 msec after a sound onset, are about 300 per second. Most experiments are done with 20 kHz sample rate, 92 cochlear model channels, and 24 neurons per channel, for a total population of 2208 neurons; this is around a factor of 20 fewer than the number of afferent primary auditory neurons in humans, so our results are somewhat noisy. A short section of a typical neuron firing pattern is shown in figure 1, along with the speech waveform that elicited it; each dot represents one or more neuron firings at a particular channel and time index.

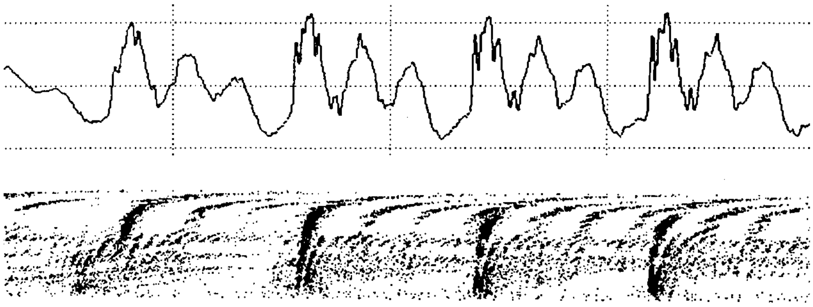

Figure 1. Speech waveform and the resulting *neurogram*, with the basal (high frequency) region at the bottom.

6. Software Representation of Models and Firing Data

All the models mentioned are implemented as *signal types* in the signal representation language SRL [4]; all programming and experimentation are done within the SRL/ISP [5] integrated signal processing environment on Symbolics model 3600 Lisp machines. A class of signals called array-valued-signal-type, or arrsig for short, were implemented to allow a time-domain next-state-simulation approach to be embedded within the structure of SRL; all these signals are functions of a single time index into a range of arrays of numbers, as opposed to a range simply of numbers. For example, the fetched value of an instance of hair-cell-bank-arrsig-type at a particular time index is an array of floating-point receptor current values; and for a primary-neuron-bank-arrsig-type, an array of 1-bit values indicating whether each neuron fired at that time (within a 50 micro-second sample interval).

For many kinds of processing, it is not interesting exactly which neurons fired at each time, but only how many neurons fired at each frequency channel at that time. For these applications, an instance of count-neuron-bank-arrsig-type returns an array of 8-bit firing counts. If desired, a multi-neuron-bank-arrsig-type can be used to represent multiple neurons as 2208 separate channels of 1-bit values.

Correlation (or coincidence) arrsigs return two-dimensional arrays of counts of coincident firings at various delays for each channel, within a selected slower sample interval. Internally, they save a history of input firings from a count-neuron-bank as a linked list of events; times and channels with no firings use no storage and no further processing. Correlation is implemented with only simple counting—multiplication is not used. Allocation and freeing of history lists is handled explicitly, with the garbage collector turned off, for speed and space economy.

Such two-dimensional arrsigs can be displayed as animated sequences of frames by converting to a movie-arrsig-type, which keeps dithered 1-bit arrays in a form convenience for quick display. These movies are fascinating to watch, and are plausible candidates for the kinds of representation that a hearing-impaired person might be able to learn to recognize visually in real time.

7. Correlation Processing of Neural Events

So far, the uses we have made of the neural firing information include the auto-correlation-based pitch perception model of Licklider [6], and the cross-correlation-based binaural lateralization model of Jeffress [7]. Both models turn out to be easier to implement with the nerve firing data than with "analog" numeric values that represent something like instantaneous probability of firing; at least, this is true given the relatively small number of neurons that we model. As far as we are aware, these models have not previously been implemented and seriously experimented with.

As mentioned above, we can compute exact correlation functions, on waveforms that are integers and mostly zeroes, by simple counting and adding operations. Since the implementation is equivalent to using coincidence detectors, or AND gates, instead of multipliers, we also refer to the results as auto-coincidence functions and cross-coincidence functions. Each frame of the coincidence function output is an array of counts of how many pulses coincided for each delay and channel index. Correlation is inherently a square-law type of operation, so it doubles the dynamic range of an input; therefore, it is good that the neuron firing rates have relatively little dynamic range, so the coincidence outputs do not need too much dynamic range.

The outputs of these models are further processed, though not currently through neural models of any sort, to try to separate out sounds based on directional, periodicity, and frequency continuity cues, among other things. Two of our separation efforts with earlier representations are reported in [2] and [8].

8. The Place-Invariance Principal

In many models of auditory processing, stages that follow the cochlear filtering are *tuned* in some way to the specific CF of the channel that they process. For example, in [9], each channel has an accurate delay equal to the resonant period of the filter channel whose output it processes. Since it is hard to imagine how neural mechanisms could be tuned as accurately and with as much stability as the mechanical resonances in the cochlea, we prefer to avoid such models, and adopt instead a strongly contrary principal. The place-invariance principal simply states that no channel of the model beyond the cochlear filtering stage is allowed to use any parameter that explicitly represents its place or frequency; each channel is, however, allowed to know its neighbors above and below.

This principal is in direct conflict with the traditional notions of frequency mapping into place, and place being directly perceived on a high-low scale through the "principal of specific nerve energies", which holds that a perception is determined primarily by which nerves fire. Preliminary supportive evidence for the place-invariance principal comes from the cochlear implant project at Stanford (private communication), where a patient who had never heard with one of his ears was found to have no ability to rank place of stimulation in that ear; this patient also had better than usual ability to rank pulse rate. Other patients, who had heard before, generally had a clear perceptual ranking of place into a degree of "shrillness", or some such frequency-like concept distinct from pitch, indicating that the association between place and frequency may be learned, based on time patterns.

9. Cues for Speech Recognition

The relatively sharp phased-locked response of the neuron models leads to good sharp correlation peaks, especially at the pitch period; experiments show them to be much sharper than the correlation peaks of the receptor currents, or other "analog" signals. In addition to pitch structure, the two-dimensional auto-coincidence function patterns are also quite good at showing formants as groups as channels with similar patterns across the delay dimension; that is, energy from a strong resonance "recruits" neurons from places of higher CF to fire in synchrony with the resonance, and this time-pattern is easily seen in Licklider's pitch model. We therefore regard his "duplex theory of pitch perception" to be applicable to many more perceptual effects than just pitch. Bursts, fricatives, etc., all have characteristic patterns in this representation. Figure 2 shows a sample of the coincidence function for the vowel /i/ of figure 1; correlates of periodicity, formants, harmonics, and high-frequency envelope modulation are apparent.

The more traditional rate *vs.* place representation of spectral shape turns out to be less useful. Just as in physiological observations, the model's rate *vs.* place function flattens out very quickly at moderate to high loudness, and hence carries very little information about formants or other prominent spectral features.

The strength of this approach in a speech recognition application is not expected just on the basis that cues are present in the coincidence functions; rather, it is the hoped that cues about interfering sounds are present in a form that allows them to be recognized as non-speech, or as different speakers.

10. Other Applications of the Models

An explicit computational model of hearing can be a very useful tool to someone doing research in mechanisms of hearing, as it provides a framework in which to evaluate the effect of a proposed model change or new feature. Studies of the hearing ability of animals, such as bats and barn owls, could also benefit from the availability of such models.

We have experimented with applying these models, with adjusted parameters, to simulated echolocation signals of bats, which are downward-sweeping FM chirps. Behavioral studies show a time-delay jnd of as low as 0.5 microsecond in some bats [10]; outputs of our models show auto-correlation peaks almost sharp enough to account for this, and help to illustrate why the particular choice of a downward-sweeping chirp optimizes timing accuracy.

Barn owls show good phase-locking to signals as high as 9 kHz, and are the only animal known with angular binaural lateralization accuracy as good as humans [11]; they need this higher frequency phase locking because their ears are closer together than ours. In both owls and humans, the geometry of the outer ear also plays an important role in localization of sound sources; the effects of models of the outer ears could profitably be examined through good computational hearing models.

11. Hardware Architecture Impact

At Fairchild, we are also interested in VLSI computing architectures appropriate for implementing hearing models in real time. Our first experimental architecture, the multi serial signal processor (MSSP), appears to be very good for implementing cochlear filtering, compression, and hair cell models, and several primary auditory neuron models per channel, since it was designed to operate with conventional 32-bit fractions. Coincidence processing of the neural firing data will require another very different architecture; but the job will be much easier, and will require much less memory, than it would if conventional numeric approaches were used. It appears to make sense to put an architectural boundary at the same place as a data representation boundary.

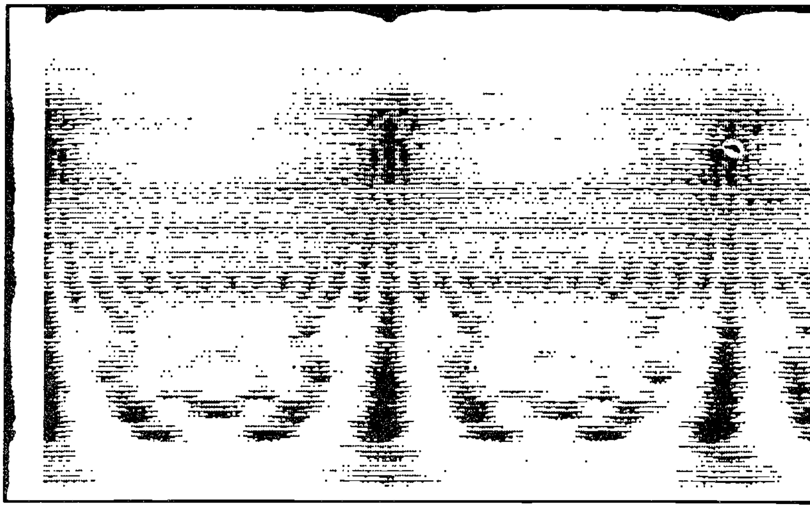

Figure 2. A sample of the *auto-coincidence function* of the neuron firings from figure 1, with high frequencies at the top.

References

[1] R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea", *Proceedings, 1982 IEEE ICASSP*, Paris (May 1982).

[2] R. F. Lyon, "A Computational Model of Binaural Localization and Separation", *Proceedings, 1983 IEEE ICASSP*, Boston, MA (Apr. 1983).

[3] J. B. Allen, "A Hair Cell Model of Neural Response", submitted for publication to *J. Acoust. Soc. Am.*, 1983.

[4] G. E. Kopec, "The Signal Representation Language SRL", submitted for publication to *IEEE Trans. on ASSP*.

[5] G. E. Kopec, "The Integrated Signal Processing System ISP", *Proceedings, 1984 IEEE ICASSP*, San Diego, CA (Mar. 1984).

[6] J. C. R. Licklider, "A Duplex Theory of Pitch Perception", *Experientia* **7**:128-133 (1951), reprinted in *Psychological Acoustics*, E. Schubert editor; Dowden, Hutchinson, and Ross, Inc., Stroudsburg, PA, 1979.

[7] L. A. Jeffress, "A Place Theory of Sound Localization", *J. Comp Physiol. Psychol.* **41**:35-39 (1948).

[8] M. Weintraub, "The GRASP Sound Separation System", *Proceedings, 1984 IEEE ICASSP*, San Diego, CA (Mar. 1984).

[9] S. Seneff, "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model", paper F3, Acoust. Soc. Am. fall meeting, San Diego, 1983; *J. Acoust. Soc. Am.* **74**:supplement 1 page S9 (1983).

[10] J. A. Simmons, "Perception of Echo Phase Information in Bat Sonar", *Science* **204**:1336-1338 (1979).

[11] M. Konishi, "High Frequency Phase Coding in the Cochlear Nucleus of the Barn Owl", *Caltech Biology Annual Report 1983*, 139-140.