ON THE IMPORTANCE OF TIME— A TEMPORAL REPRESENTATION OF SOUND

5

Malcolm Slaney and Richard F. Lyon

Advanced Technology Group
Apple Computer, Inc.
Cupertino, CA 95014 USA

1 INTRODUCTION

The human auditory system has an amazing ability to separate and understand sounds. We believe that temporal information plays a key role in this ability, more important than the spectral information that is traditionally emphasized in hearing science. In many hearing tasks, such as describing or classifying single sound sources, the underlying mathematical equivalence makes the temporal versus spectral argument moot. We show how the nonlinearity of the auditory system breaks this equivalence, and is especially important in analyzing complex sounds from multiple sources of different characteristics.

The auditory system is inherently nonlinear. In a linear system, the component frequencies of a signal are unchanged, and it is easy to characterize the amplitude and phase changes caused by the system. The cochlea and the neural processing that follow are more interesting. The bandwidth of a cochlear "filter" changes at different sound levels, and neurons change their sensitivity as they adapt to sounds. Inner Hair Cells (IHC) produce nonlinear rectified versions of the sound, generating new frequencies such as envelope components. All of these changes make it difficult to describe auditory perception in terms of the spectrum or Fourier transform of a sound.

One characteristic of an auditory signal that is undisturbed by most nonlinear transformations is the periodicity information in the signal. Even if the bandwidth, amplitude, and phase characteristics of a signal are changing, the repetitive characteristics do not. In addition, it is very unlikely that a periodic signal could come from more than one source. Thus the auditory system can safely assume that sound fragments with a consistent periodicity can be combined and assigned to a single source. Consider, for example, a sound formed by opening and closing the glottis four times and filtering the resulting puffs of air with the vocal resonances. After nonlinear processing the lower auditory nervous system will still detect four similar events which will be heard and integrated as coming from a voice.

The duplex theory of pitch perception, proposed by Licklider in 1951 [11] as a unifying model of pitch perception, is even more useful as a model for the extraction and representation of temporal structure for both periodic and non-periodic signals. This theory produces a movie-like image of sound which is called a correlogram. We believe that the correlogram, like other representations that summarize the temporal information in a signal, is an important tool for understanding the auditory system.

The correlogram represents sound as a three dimensional function of time, frequency, and periodicity. A cochlear model serves to transform a one dimensional acoustic pressure into a two dimensional map of neural firing rate as a function of time and place along the cochlea. A third dimension is added to the representation by measuring the periodicities in the output from the cochlear model. These three dimensions are shown in Fig. 1. While most of our own work has concentrated on the correlogram, the important message in this chapter is that time and periodicity cues should be an important part of an auditory representation.

This chapter describes two cochlear models and explores a structure which we believe can be used to represent and interpret the temporal information in an acoustic signal. Section 2 of this chapter describes two nonlinear models of the cochlea we use in our work. These two models differ in their computational approach and are used to illustrate the robustness of the

Published in *Visual Representations of Speech Signals*, Martin Cooke, Steve Beet, and Malcolm Crawford (eds.), pp. 95-116.

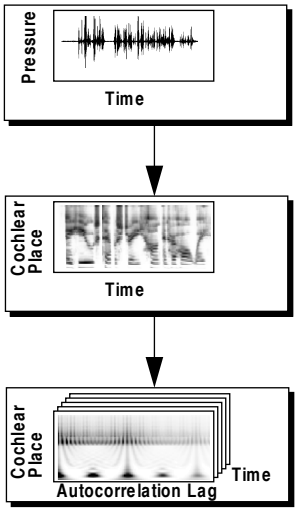


Fig. 1– Three stages of auditory processing are shown here. Sound enters the cochlea and is transduced into what we call a cochleagram (middle picture). A correlogram is then computed from the output of the cochlea by computing short time autocorrelations of each cochlear channel. One frame of the resulting movie is shown in the bottom box.

temporal information in the output of the cochlea. Over the past forty years there have been several ways to summarize this information at the output of the cochlea [11] [22] [36]. Since these representations produce such similar pictures we describe them all with the term correlogram. Correlograms, their computation and implementation, are the subject of Section 3 of this chapter. Finally, Section 4 describes the use of correlograms for sound visualization, pitch extraction, and sound separation.

2 NONLINEAR COCHLEAR MODELS

Two different computational models of the cochlea are described in this work: the older model [12][30], which we refer to as the "passive long-wave model," and the newer model [14], which we refer to as the "active short-wave model." The two models differ in their underlying assumptions, approximations, and implementation structures, but they share three primary characteristics (not necessarily implemented independently or in this order):

- Filtering: A broadly tuned cascade of lowpass filters models the propagation of energy as waves on the Basilar Membrane (BM).
- Detection: A detection nonlinearity converts BM velocity into a representation of inner haircell (IHC) receptor potential or auditory nerve (AN) firing rate.
- Compression: An automatic gain control (AGC) continuously adapts the operating point of the system in response to its level of activity, to compress widely varying sound input levels into a limited dynamic range of BM motion, IHC receptor potential, and AN firing rate.

The several differences between the models are largely independent of each other, so there is a large space of possible models in this family. The main differences between the two models we have experimented with are:

- The passive long-wave model is based on a popular one-dimensional (long-wave) hydrodynamic approximation with a lightly-damped resonant membrane [37]; the active short-wave model is based on a two-dimensional hydrodynamic approximation (emphasizing the short-wave region) with active undamping and negligible membrane mass [15].
- Our passive long-wave model is implemented with complex poles and zeros, while the filters in the active short-wave model have only complex poles. These decisions are based on rational filter approximations to the different underlying hydrodynamic simplifications.

- The passive long-wave model uses time-invariant linear filters followed by a variable gain to functionally model the AGC. The active short-wave model varies the filter pole Q over time to effect a gain variation and to model the mechanical AGC in terms of active adaptive hydrodynamics.

Both models are motivated by the desire to compute a representation of sound that is approximately equivalent to the instantaneous firing rates of AN fibers. By assembling the firing rates versus time for a large number of fibers with different best frequencies (BF), we construct a picture called the "cochleagram." The cochleagram is useful as a visual representation of sound, and as a numerical input to other sound processing functions, such as automatic speech recognition.

The cochleagram has a wealth of fine time structure or "waveform synchrony" driven by the temporal structure of the incoming sound. The extraction and representation of the important perceptual information carried in the temporal structure on the AN is the main topic explored in this chapter. Nevertheless, for the display of cochleagrams, we often just smooth away the details via a lowpass filter, in order to reduce the bandwidth enough to fit a signal of some duration (e.g., a sentence) into the resolution of the display medium. These "mean-rate" cochleagrams would be rather flat looking if they really represented mean AN firing rates. Instead, we follow Shamma [29] in using a first-order spatial difference (a simple Lateral Inhibitory Network or LIN) to sharpen the cochleagram response peaks due to spectral peaks.

2.1 Modeling approach

Sound waves enter the cochlea at the oval window, causing waves to travel from the base to the apex along the BM. The speed at which waves propagate and decay is a function of the mechanical properties of the membrane and the fluid, and of the wave frequency. The most important property that changes along the BM is its stiffness. As a wave of any particular frequency propagates along the BM from the stiff basal region toward the flexible apical region, its propagation velocity and wavelength decrease, while its amplitude increases to a maximum and then rapidly decreases due to mechanical losses. The amplitude increase is due to the energy per cycle being concentrated into a smaller region as the wavelength decreases, and, in the case of an active model, to energy amplification in the traveling wave.

For both one-dimensional and two-dimensional hydrodynamic models, a technique known as the WKB approximation allows us to describe the propagation of waves on the BM one-dimensionally, using a local complex-valued "wavenumber." The wavenumber k (the reciprocal of Zweig's λ parameter [37]) may be thought of as a reciprocal wavelength in natural units, or a spatial rate of change of phase in radians/meter. But, it can also have an imaginary part that expresses the spatial rate of gain or loss of amplitude.

In general, k depends on frequency (ω) and on the parameters of the wave propagation medium (for example, stiffness, mass, damping, height, width). We allow parameters of the medium to depend on x, the distance along the BM measured from the base. Thus the wavenumber is expressed as a function of frequency and x: $k(\omega, x)$. The equation that describes the wave medium and lets us find k from the frequency and the parameters at location x is known as the dispersion relation, and may be derived from some approximation to the hydrodynamic system. The popular long-wave approximation [38] is simplest, but is only valid when the wavelength is very long compared to the height of the fluid chambers of the cochlea. A better approximation to physical (or at least mathematical) reality results from a 2D or 3D model of the hydrodynamics [25][33]. Different models lead to different solutions for $k(\omega, x)$ [37].

The WKB approximation says, roughly, that we can describe wave propagation along the x dimension by integrating the rate of change of phase and relative amplitude indicated by k. In a uniform medium, a (complex) wave traversing a distance dx is multiplied by

$$Exp[ik(\omega, x)\,dx]. \tag{1}$$

According to WKB, in a nonuniform medium, as a wave traverses a region from x_1 to x_2, it is multiplied by

$$Exp\left[i \int_{x_2}^{x_2} k(\omega, x)\, dx \right]. \tag{2}$$

The WKB approximation includes an amplitude correction factor as well. This factor depends on whether the wave being propagated represents pressure or displacement and insures the wave amplitude correctly accounts for energy as the wavelength changes. In the short-wave region, under an assumption of constant BM mass and width, no amplitude correction is needed for the pressure wave. On the other hand, an amplitude increase proportional to k is needed for the BM displacement or velocity wave. In the long-wave region, pressure amplitude decreases as $k^{-1/2}$, while displacement and velocity increase as $k^{3/2}$. In the general 2D case, and for more general mass and stiffness scaling, amplitude scaling is more complex [15]. For our models, we choose ad hoc stage gains near unity that provide plausible correction factors and lead to good-looking results.

We model wave propagation using a cascade of filters by noting that the exponential of an integral is well approximated by a product of exponentials of the form

$$e^{ik(\omega, x)\, dx} \tag{3}$$

for a succession of short segments of length dx. We then only need to design a simple filter

$$H_i(\omega, x) = e^{ik(\omega, x_i)\, dx} \tag{4}$$

for each segment of the model corresponding to BM location x_i. For short enough segments, the filter responses will not be too far from unity gain and zero phase shift, and will themselves be well approximated by low-order causal rational transfer functions (i.e., by a few poles and/or zeros).

The conversion of mechanical motion into neural firings is performed by the Inner Hair Cells (IHC) and neurons of the auditory nerve (AN). IHCs only respond to motion in one direction and their outputs saturate if the motion is too large. Thus a simple model of an IHC is a Half Wave Rectifier (HWR), while more complicated models might use a soft saturating HWR such as

$$\frac{1}{2}(1 + \tanh(x + a)). \tag{5}$$

Even more realistic models of IHC and AN behavior take into account local adaptation, refractory times, and limited firing rates [19]. Our work is more interested in the average firing rate of a number of cells, thus we do not need this level of detail. Both cochlear models described in this chapter use a simple HWR as a detector.

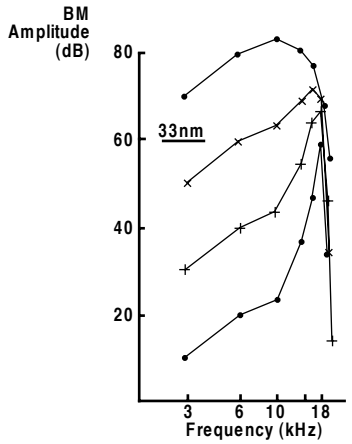
All IHC models share the important property of acting like detectors. This means that they convert the pressure wave with both positive and negative values into a signal that retains both the average energy in the signal and the temporal information describing when each event occurs. Over a period of several cycles, the average pressure at a point on the BM will be zero. But by first using a HWR, or other hair cell model, the average will be related to the energy in the signal yet the fine time structure is preserved. This temporal information will be important later when trying to group components of a sound based on their periodicities [12].

Such a non-linearity is important part of understanding the perception of sounds with identical spectra but different phase characteristics. One such set of sounds was studied by Pierce [24]. In his study, carefully constructed sounds with identical spectra but different phases were shown to have different pitches. A simple HWR dectector is sufficient to turn the phase differences into envelopes whose periodicities explain the different pitches.

Finally, a model of adaptation, or Automatic Gain Control (AGC), is necessary. In its simplest form, the response to a constant stimulus will at first be large and then as the auditory

system adapts to the stimulus the response will get smaller. There are many types of adaptation in the auditory system that respond over a large range of time scales. Some of these adaptations affect the mechanical properties of the BM and thus change the wave propagation equation.

The interaction of sound levels and wave mechanics is clear in the iso-intensity mechanical response data of Rhode [27], Johnstone [10], and Ruggero [28]. Typical data are shown in Fig. 2. In all cases, the peak of resonant response is blunted at high sound levels, resulting in an increased bandwidth, a shift in best frequency, and a reduced gain for frequencies near the characteristic frequency (CF). These effects are qualitatively in agreement with the result due to reducing the pole Q in our active short-wave model. Our passive long-wave model, on the other hand, keeps the mechanics constant and applies a pure gain variation before the IHC. Models that rely on the place of maximum response can not realistically count on the cochlea to map a consistent frequency to a particular place. Using a Lateral Inhibition Network to shift the response peak closer to the sharp cutoff edges gives a more consistent mapping.



SPL	Q_{3dB}	CF
80dB	1	10k
60dB	2.7	16k
40dB	4.8	17k
20dB	8.3	17k

Fig. 2– Mössbauer data shows the non-linearity of the cochlea. This data, measured by Johnstone, shows the motion of the BM at four different sound levels. Note that the response is most highly tuned at the lowest sound levels. Adapted from [10] with permission.

2.2 The Passive Long-Wave Model

Our passive long-wave model was designed by Lyon [12] based on a long-wave analysis of the cochlea by Zweig [37]. The implementation of this model is described by Slaney [30]. This model uses a cascade of second-order sections to approximate the complex, frequency-dependent delay and attenuation a wave encounters as it travels down the BM. This model uses a HWR as a detector and four stages of a multiplicative AGC to model adaptation.

The transfer function for a stage of the model is based on an approximation to the long-wave solution for a short section of the BM. The transfer function, or ratio of complex ouptut amplitude to input amplitude, over a length dx of the BM is a function of frequency, ω, and is written

$$\frac{P_o}{P_i} = A(\omega)\, e^{ik(\omega)\,dx} \qquad \text{with } k(\omega) = \frac{c}{\sqrt{\omega_R^2 + i\omega\omega_R^2/Q - \omega^2}} \tag{6}$$

where $A(\omega) \approx 1$. When the wavenumber k is real-valued, the transfer function contributes just a phase change and there is no change in amplitude. Negative imaginary values of k cause the exponential's mangnitude to be less than one and the wave to decay. The resulting

transfer function, as a function of sound frequency, for a small section of the cochlea with ω_R near 5.8 kHz is shown in Fig. 3.

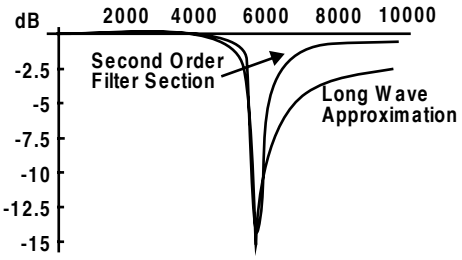


Fig. 3- The frequency response of a small section of a long wave model of the BM is compared to an approximation based on a second order filter. A large number of these sections are combined to form the overall lowpass filter characteristic of the cochlea.

The parameters ω_R and Q are local parameters for each small section of BM being modeled in the cascade. The resonance frequency,ω_R, changes exponentially from 2π 20 kHz at the base of the cochlea to approximately 2π 20 Hz at the apex. The transfer function in Equation (6) is a notch filter, and when the response is combined from the base to any point along the BM the result is a Low Pass Filter (LPF).

Equation (6) describes the transfer function for a pressure wave traversing a section of length *dx*. Pressure is converted to BM velocity by the local BM impedance, which is essentially a simple resonator described by the same ω_R and Q. We approximate the transfer function and the accompanying resonator using a biquadratic filter (two poles and two zeros). The resulting structure is computationally efficient and an adequate model of the cochlea for our purposes. This structure is shown in Fig. 4.

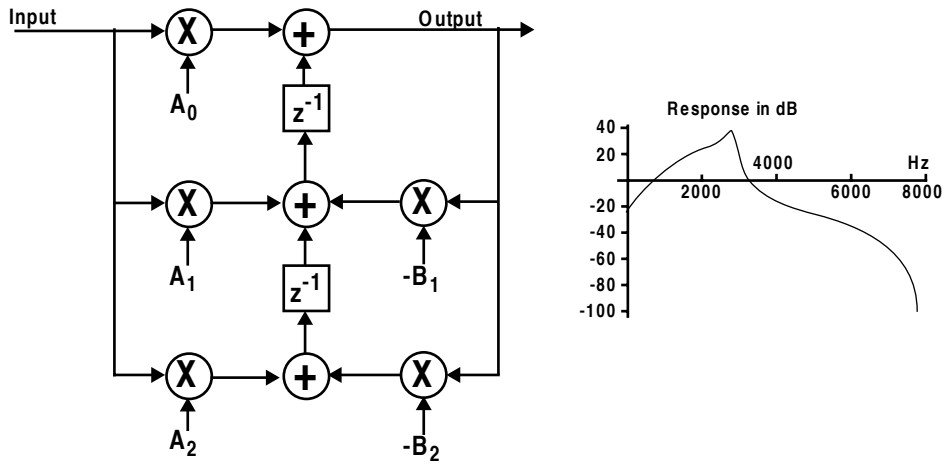


Fig. 4– A single stage of the cascade-parallel model is implemented with the second-order section shown on the left. The frequency response for one typical stage is shown in Fig. 3. The BM is simulated by combining stages into a cascade. The typical response, including a differentiator to convert to velocity, is shown on the right.

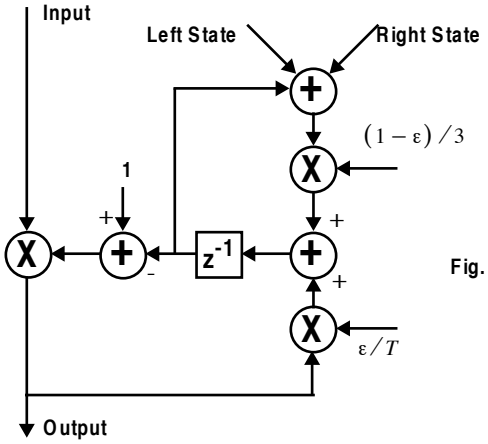
The bandwidth and center frequency of the notch varies as a function of the characteristic frequency corresponding to each position along the BM. The bandwidth of the poles is given by the expression

$$BW = \frac{\sqrt{f_{break}^2 + CF^2}}{Q_{hi}} \tag{7}$$

and corresponds roughly to a critical band. Model parameters Q_{hi} and f_{break} specify the high-frequency limit of the pole Q, and the frequency that separates the nearly constant-Q and constant bandwidth regions. Filter stages are cascaded, with center frequencies from high to low, so that the center frequency of each stage falls a small fraction of the bandwidth below the center frequency of the previous stage.

The filter Q_{hi} and the percentage of filter overlap are parameters of our long-wave model; no one value is correct for all situations. Instead, depending on the use, we often choose one of two values. If we are computing a cochleagram then we use a Q_{hi} of 8, which yields unrealistically sharp filters but produces a picture with good frequency-domain resolution. If instead we are looking at the periodicities of the signal by computing a correlogram, we can use a more realistic bandwidth and a Q_{hi} of 4. To keep the same number of channels per octave, we step the narrow filters by a factor of 25% of *BW* and the wide filters by 12.5% of *BW* (overlap 75% and 87.5% respectively).

An important step in a cochlear simulation is a model of adaptation. In our passive cochlear model this function is performed by time varying gains in an AGC loop. To simulate adaptation the AGC is operating at a point where it is sensitive to new sounds. After an increased sound loudness is detected, the gain is turned down. The structure of this multiplicative AGC is shown in Fig. 5. Four of these stages, each with a different time constant, are used to model the range of adaptation rates found in the auditory system.



AGC	Target	Time Constant
First	.0032	640 ms
Second	.0016	160 ms
Third	.0008	40 ms
Fourth	.0004	10 ms

Fig. 5– The structure of the Automatic Gain Control (AGC) used in the cascade-parallel model of the cochlea. Four of these stages are combined, each with a different value for the time constant (ε) and output target value (T), to make the complete adaptation model. In the passive long-wave model of the cochlea, the AGC follows the detector, so only positive values are seen.

While the passive long-wave model does a reasonably good job of calculating a cochleagram, it differs from the physiology in two areas. Most importantly, there is evidence that an active mechanical mechanism is part of the cochlea and serves not only to amplify low-level traveling waves but also to sharpen the iso-response mechanical tuning curves. Incorporating the AGC into the mechanical properties of the cochlea tends to compress the range of signal levels that need to be represented by neural circuits. This point is minor for high-precision floating-point computer implementations but is critical for the representation and for neurons, which have a limited dynamic range.

2.3 The Active Short-Wave Model

The second model we use in our work is based on two-dimensional hydrodynamics and includes negative damping to model the mechanical amplification of low-level signals by active outer haircells. This approach has been described by Lyon and Mead [14] [15] in terms

of analog circuits. We have implemented their approach digitally, and extended it to include a coupled AGC loop that adapts the filters.

Though the model is based on a full two-dimensional analysis, we characterize it as a short-wave model to emphasize the importance of the short wavelength near the response peak, and to clearly contrast it with our long-wave model. Because of the short-wave behavior and the form of the active undamping and higher-order loss mechanisms we have used, it is possible to get a reasonably sharp response peak even with no BM mass [15]. This massless approach is not possible in a passive long-wave model. Finding a more realistic version of this class of model requires more data on BM mechanics and OHC micromechanics.

Like the passive long-wave model, our active short-wave model is built with a cascade of second-order filter sections that model pressure wave propagation. But in the short-wave model, the filter stages adapt in response to the sound level by lowering their pole Q when the local wave energy is high. The filters have unity gain at DC, for lossless propagation of low-frequency waves. When the pole Q is greater than 0.707 (the usual case), each filter has a frequency range over which its gain is greater than unity, implying active amplification. With quiet sounds, the Q values for each stage can be as high as 2. This results in gain peaks for the cascade near 2^{10} (60 dB), depending on parameters such as the filter overlap and step factor. Adjusting the Q values between 0.707 and 2 can change the overall filter cascade's peak gains by about 60 dB in response to sound levels over a 100 dB range.

The active cochlea has negative damping. This corresponds to a positive imaginary part of the wavenumber k and causes energy to be added to propagating waves. A wave of a particular frequency would grow without bound if the damping did not become positive at some place as the wave propagated. The curve of gain versus place for a particular frequency is also reflected in the curve of gain versus frequency at a particular place. Each filter stage needs to have a gain greater than unity followed by a falloff toward zero; a pair of poles is the simplest way to approximate this shape qualitatively.

In simple dynamic systems, damping is used to quantify the rate at which energy is dissipated over time. Negative damping in a dynamic system, or in a simple second-order filter, results in an instability. But with a propagating wave, damping quantifies the rate of energy dissipation per distance. Negative damping in a uniform medium would be unstable, but in a cochlea with changing parameters, each region can have negative damping for a range of frequencies, and positive damping for higher frequencies, with no instability. The negative damping of a section of the wave medium is captured via the WKB approximation as a gain greater than unity.

In the cochlea, there must be a physical limit to the amount of energy that can be added to a wave by active outer haircells. Therefore, at high sound levels the system must become passive, corresponding to reducing all the Q values to 0.707 or less in our model. Changing the Q values between the small-signal and large-signal limits results in an overall compressive behavior, in agreement with compression seen in the actual cochlear mechanics [16].

In the cochlea, the level of OHC activity is probably controlled both by a fast local nonlinearity and by a slower feedback loop involving the cochlear efferents and the olivary complex. The degree of feedback and activity is not the same at all places, but is dependent on the signal spectrum. We model this control loop using a set of parallel time-space smoothing filters similar to those used in the coupled AGC of our long-wave model. Following the IHC detection nonlinearity, four filters with different time constants and space constants add their outputs. This sum is added to a minimum damping parameter to compute a filter stage's pole damping. The wide range of possible loop characteristics, nonlinearities, and binaural effects that no doubt occur in the olivary complex have not been explored, but this ad hoc AGC gives good compression and qualitatively correct shifts in CF, bandwidth, and phase. Due to the spatial coupling and the fact that cascaded stages interact, there is also a qualitatively reasonable two-tone suppression effect, as there was in the long-wave model [13]. The structure of this active and adaptive model is shown in Fig. 6.

The digital implementation of the active short-wave model uses a novel second-order filter structure, shown in Fig. 7, in which one coefficient directly controls the pole damping, or 1/Q [35]. This allows us to connect the output of the AGC loop filters to the cascaded filter stages, without a block to convert pole CF and Q to filter coefficients. The CF parameter di-

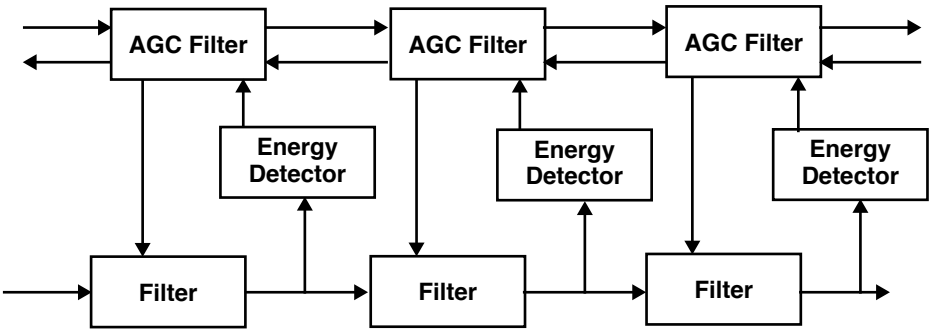


Fig. 6– A more realistic model of the cochlea uses energy detectors that control the parameters of the cochlear filters. The AGC filters integrate the energy over time and space and control the damping of the BM filters.

rectly controls the pole frequency, and is a design parameter that is held constant for each stage. The separation of frequency and damping is exact only in the limit of low pole frequencies, and is usable with care up to only about half of the Nyquist frequency. Thus the high-frequency end of the short-wave model is not as high as it is for the long-wave model, for a given sound sample rate.

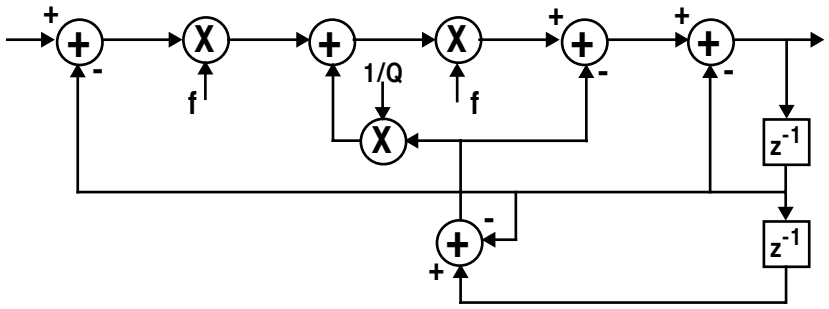


Fig. 7- This structure is used to implement a single stage of our model of the hydrodynamics of the cochlea. Unlike the structure shown in Fig. 4, the center frequency and the filter damping, or 1/Q, can be controlled directly.

2.4 Pictures

A time-frequency representation is often used to analyze and display speech signals. Four representations of the utterance "Fred can go, Susan can't go, and Linda is uncertain" from the ESCA Sheffield Workshop are shown in Fig. 8 and Fig. 9. Fig. 8 shows the conventional wide and narrow band spectrograms using the short-time Fourier Transform. These two spectrograms use two different window sizes and thus differ in their resolution in the time and frequency domains.

Fig. 9 shows two cochleagrams of the same utterance. The cochleagram, much like the spectrogram, is a function of time along the horizontal axis and cochlear place, or frequency, along the vertical axis. The darkness of the picture at each point represents the LIN enhanced average of the auditory nerve firing rate at each position along the BM. The spectrograms and cochleagrams show a remarkable similarity. The most noticeable differences in the pictures are the change in the scaling of the frequency axis and some enhancement of the onsets in the cochleagrams.

A more important difference can not be seen in these pictures. Because of the limited space on the printed page, each pixel in these cochleagrams represents the average cochlear firing rate over a period of approximately 5 ms. But the cochlea and the IHCs are exquisitely

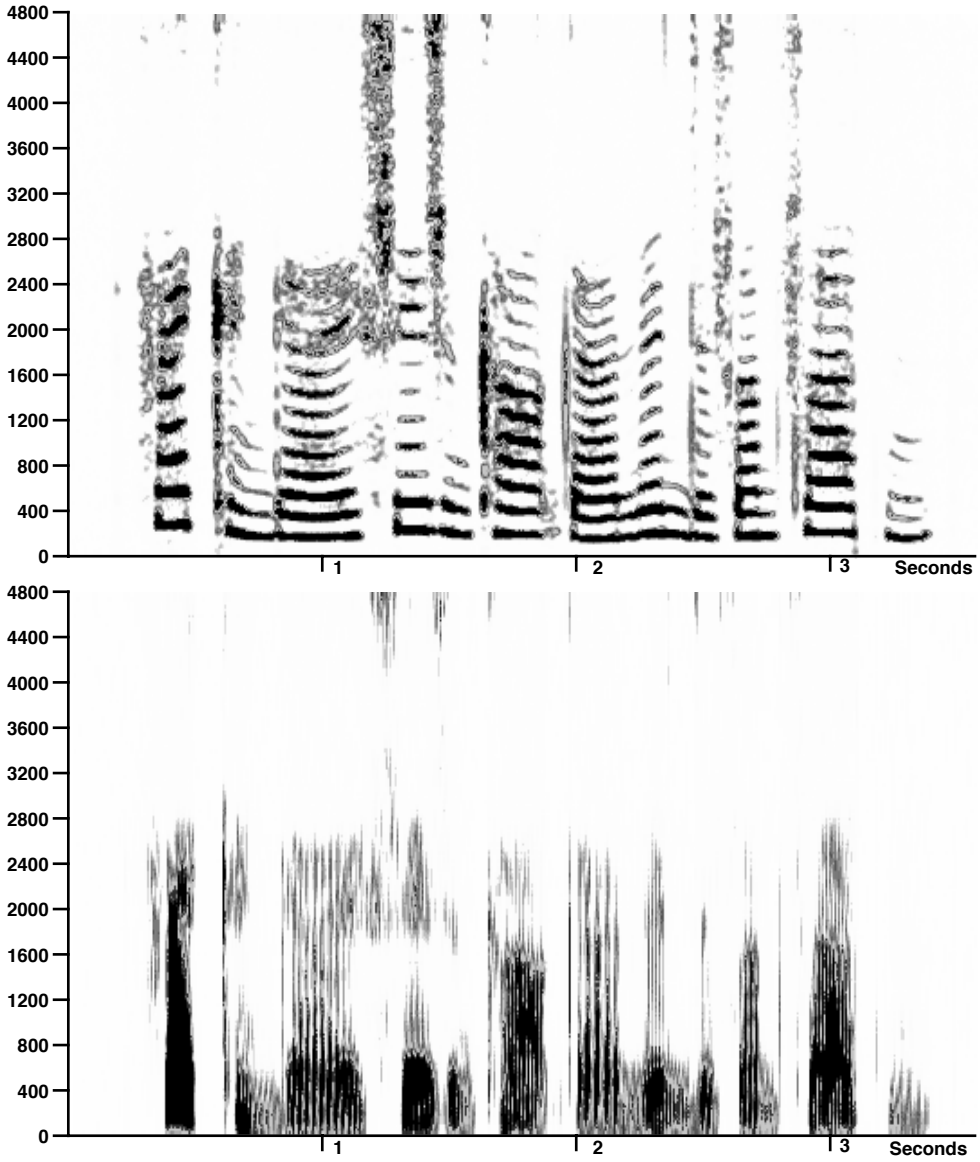


Fig. 8- Narrow (top) and wide-band (bottom) spectrograms of the Sheffield 'clean.wav' utterance, "Fred can go, Susan can't go, and Linda is uncertain." These spectrograms were computed with the Signalyze program for the Macintosh using bandwidths of 20 and 200Hz, respectively.

sensitive to the time structure of each component of the sound. Fig. 10 shows an expanded view of the dipthong "rea" from the word greasy. At this time scale each glottal pulse and each waveform peak is visible. One can still follow the formant tracks, but in addition the glottal pulses that trigger the formant information allow one to group frequency channels that come from the same source.

We argue that this temporal information is important. Conventional models of audition base all performance on a suitably narrow resolution in the frequency domain. We feel this is

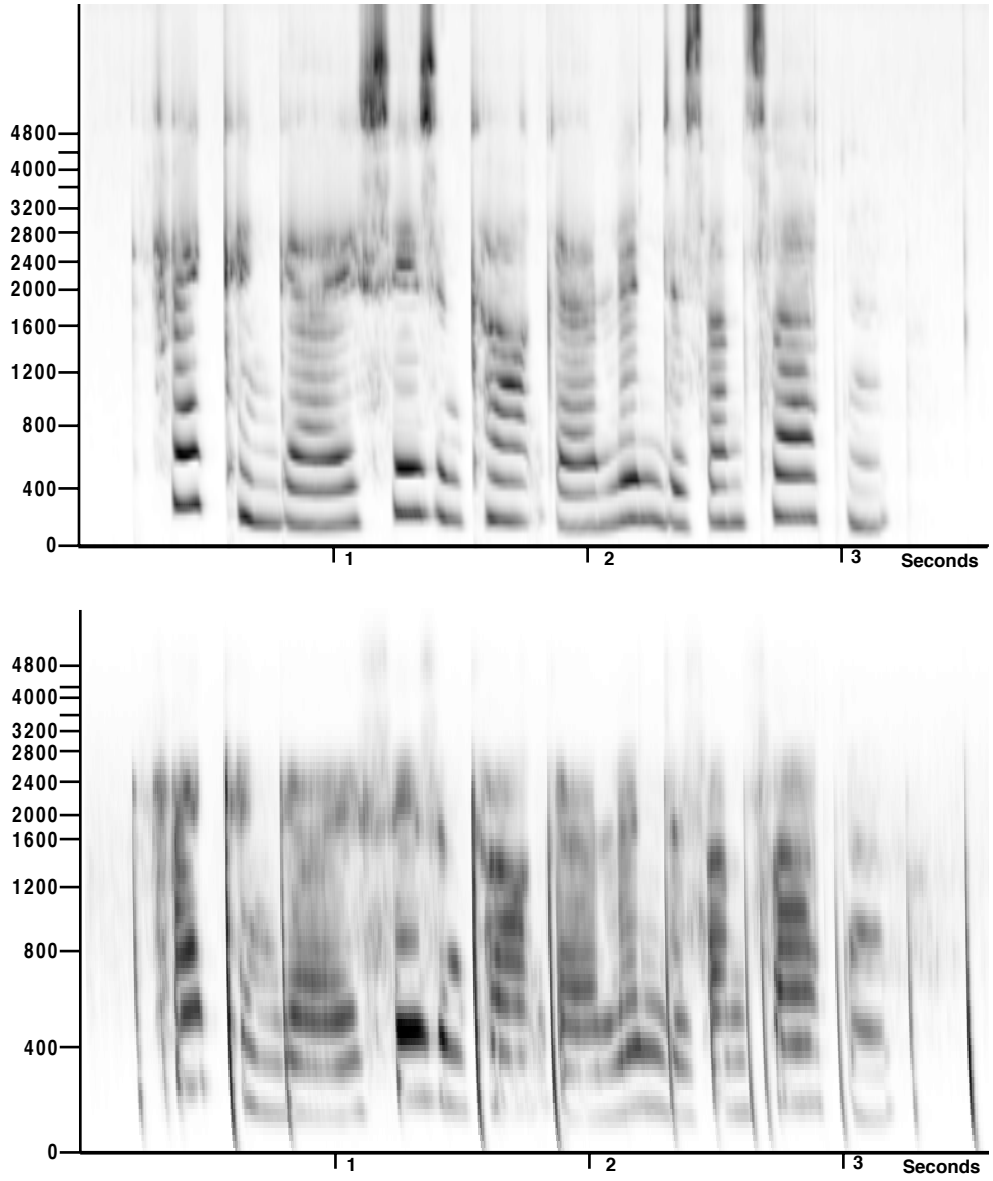


Fig. 9 - Passive long-wave (top) and active short-wave (bottom) cochleagrams of the Sheffield 'clean.wav' utterance. The passive long-wave cochleagram was computed using the default MacEar parameters and "df=100 tau=1". The active short-wave cochleagram was computed using "df=100 tau=1 gain=.001".

unrealistic since the bandwidth and center frequency of the mechanical system change with level. Instead, if the auditory system is based on the temporal information in the signal then the performance of the system is relatively insensitive to each filter's bandwidth and center frequency. The correlogram is one way to capture the temporal structure in the cochlear output and is the subject of the remainder of this chapter.

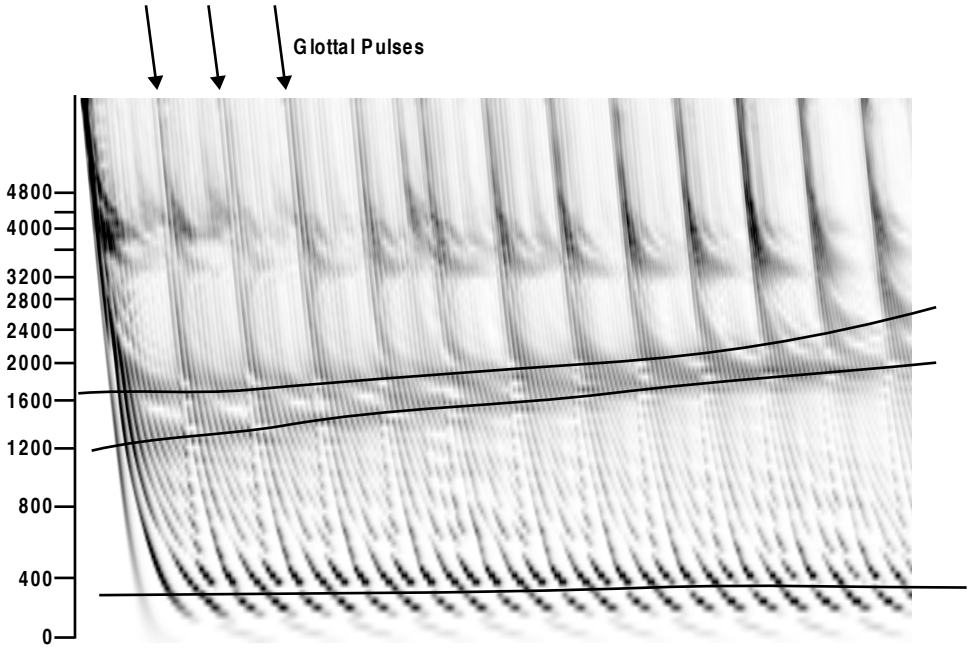


Fig. 10- Expanded cochleagram of the dipthong "rea" in greasy from the Sheffield 'timit.dip' utterance. The first three formant tracks are shown (the lowest formant is excited with two harmonics). The vertical lines, each of which represents a glottal pulse, are tilted slightly due to the natural delay through the cochlea.

3 REPRESENTING TIME—THE CORRELOGRAM

We use the correlogram to summarize the temporal information in the sounds we hear. This chapter argues that the correlogram is biologically plausible and serves as a representation that higher level processes can use to form auditory objects. The cochlea separates a sound into rather broad frequency channels yet retains the timing of the original sound. How is it that the brain extracts this information from the acoustic signal and uses it to group sounds?

The first step is to gather evidence of events that repeat. There are many ways to do this and in this section we will describe a range of techniques from those that are biologically plausible to those that can be efficiently implemented on a digital computer. We use the term correlogram, literally "picture of correlations," to describe the resulting sequence of images. While other researchers have described different implementations, we believe that each of these representations have the same goal: to represent the time structure of a signal.

In its ideal form, a correlogram is computed by measuring the short-time autocorrelation of the neural firing rate as a function of cochlear place, or best frequency, versus time. Since an autocorrelation is itself a function of a third variable, the resulting correlogram is a three-dimensional function of frequency, time, and autocorrelation delay. For display, we assemble a frame of data, all autocorrelations ending at one time, into a movie which is synchronized with the sound.

An idealized structure to compute the correlogram is shown in Fig. 11. Sound enters the correlogram array from the cochlea, a picture is computed, and is then sent to higher level structures in the brain.

We can't show a correlogram on paper but we can show individual frames and talk about the significant features. More examples are available in a video report we have published [31]. Fig. 12 shows several correlograms and illustrates how the correlogram changes as the

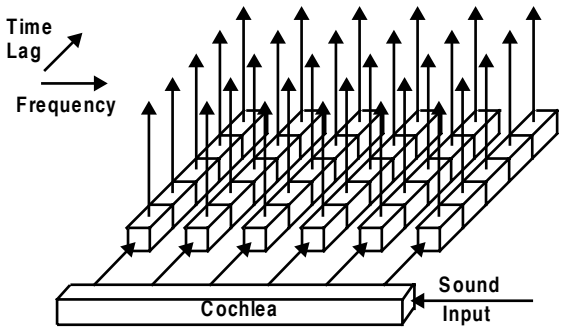


Fig. 11– This structure can be used to measure the temporal information in a sound. Sound enters the cochlea on the bottom right and is analyzed into broad frequency channels. Each channel is then correlated with itself and the resulting picture is passed to higher structures in the brain for further processing.

pitch and formant frequencies of the sound change. Distance along the BM is shown on the vertical axis of a correlogram. Since each section of the BM is most sensitive to a single frequency the vertical axis of a correlogram roughly corresponds to frequency, with the base of the cochlea, or the part that is most sensitive to high frequencies, at the top and the apex, or the low frequency portion, at the bottom.

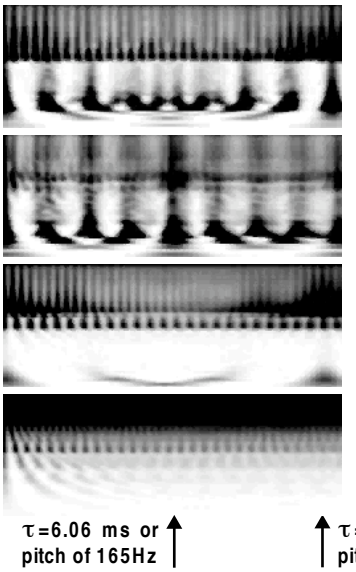


Fig. 12- Four frames of a correlogram of a voice. The first three frames are voiced sounds. When the pitch is raised the vertical structures become closer together (first and second frame). As the formant frequencies change the horizontal bands move (first and third frame). Finally, if the sound is unvoiced (last frame) then there is no vertical structure.

Autocorrelation time delay, or lag, is shown on the horizontal axis of a correlogram. The width of a correlogram is chosen to include time delays long enough to include the lowest expected pitch. Generally this is at least 10ms.

The activitiy of the correlogram is displayed as darkness in the image. As in a conventional spectrogram, dark areas represent autocorrelation lags and cochlear frequencies where there is a large response.

Voiced sounds, as shown in Fig. 12a-c, best show the utility of the the correlogram. Strong vertical lines at particular autocorrelation lags indicate times when a large number of cochlear channels are firing with the same period. This is a strong indication of a pitch which has a frequency inversely related to the autocorrelation lag. When the pitch increases, as shown between Fig. 12a and b the dominant line moves to the left, to a lag equal to the reduced period.

Horizontal bands are indications of large amounts of energy within a frequency band. The correlogram frames shown in Fig. 12a and c illustrate how the correlogram changes as the

vowel /a/ is changed to /i/. Note that the first formant drops while the second and third formants move much higher. Finally, only voiced sounds have a pitch. Unvoiced sounds, like the letter /s/, do not contain any periodic information and thus the correlogram is uniformly black for all BM channels that contain energy. This is shown in Fig. 12d.

An autocorrelation of $x(t)$ is defined by the following integral

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)\, x(t-\tau)\, dt. \tag{8}$$

For dynamic signals, we are interested in the periodicities in the signal within a short window ending at time t. This short-time autocorrelation can be written

$$R_{xx}(\tau, t) = \int_{0}^{\infty} x(t-s)\, x(t-s-\tau)\, w(s)\, ds = [x(t) \cdot x(t-\tau)] * w(t) \tag{9}$$

where $w(t)$ is an arbitrary causal window which limits the autocorrelation to a neighborhood of the current time. As indicated by the convolution form above, one way to calculate such a running autocorrelation is to filter the instantaneous correlation through a smoothing filter whose impulse response is a window [11].

A slightly different definition is useful on a digital computer. By windowing the data first, we can implement the correlation using an FFT algorithm and reduce the computations by an order of magnitude or more. Now assume

$$w(t) = 0 \text{ for} t < 0 \text{ and} t > T \tag{10}$$

and form a windowed signal ending at a particular time t:

$$y_t(s) = x(t-s)\, w(s)\,. \tag{11}$$

The windowed autocorrelation can now be written

$$R_{xx}(\tau, t) = \int_{0}^{T} y_t(s)\, y_t(s+\tau)\, ds = F^{-1}\|F(y)\|^2 \tag{12}$$

where F and F^{-1} indicate the forward and the reverse Fourier Transform. This equation can be rewritten to make it more like Equation (9)

$$R_{xx}(\tau, t) = \int_{0}^{T} x(t-s)\, w(s)\, x(t-s-\tau)\, w(s+\tau)\, ds. \tag{13}$$

The correlogram is also a function of BM position or frequency. Using Equation (9), we can write the following equation for the correlogram as a function of the cochlear firing rate $x_f(t)$ at the position along the BM most sensitive to a sinusoid of frequency f. The most general form of the correlogram is written

$$C_f(\tau, t) = \int_{0}^{\infty} x_f(t-s)\, x_f(t-s-\tau)\, w(s)\, ds. \tag{14}$$

Autocorrelations are often normalized so that the value for zero lag is equal to one. Such normalization reduces the dynamic range required for display, but completely eliminates any indication of the relative power in different frequency channels. Since autocorrelation doubles the dynamic range required to represent varying signal levels, we partially normalize by the square root of the power. This serves as a compromise so that a correlogram can be displayed with a dynamic range comparable to the cochleagram. This is written

$$\hat{C}_f(\tau, t) = \frac{C_f(\tau, t)}{C_f(0, t)^{1/2}} \quad . \tag{15}$$

Since the autocorrelation of a non-negative function is also non-negative, the resulting normalized correlogram will have values between 0 and a maximum value that we scale for the display technology.

The correlogram as described above is a continuous function of time, frequency, and autocorrelation lag. We have already sampled the auditory input as a function of time and the BM as a function of place or frequency. Sampling the auditory input means that only discrete time lags are possible in the autocorrelation without interpolating to a higher sample rate.

Still, the value of the correlogram at any one frequency and lag changes at every sample time. There is no way to display 16000 or more correlogram frames per second, so instead we subsample the correlogram to a more manageable rate of 10 to 30 frames per second. Thus to prevent temporal aliasing it is necessary to lowpass filter the correlogram. The easiest way to accomplish this filtering is to choose the correlogram window so that it is an appropriate lowpass filter. Without much biological evidence to base a window length on, we instead choose a Hamming window twice as long as the frame sampling interval. This serves to average the correlogram over a long enough interval to prevent aliasing.

There are many ways to measure periodicites and implement a correlogram. The correlograms in this chapter were computed on a digital computer using an FFT to efficiently implement the correlation operation. But there is little reason to think that neurons would use an FFT. Instead a direct solution, like that shown in Fig. 13 is more plausible. In this implementation a neural delay line, perhaps using a combination of axonal delays and neural resonators, delays a copy of the signal. For each time delay, a neuron fires when both the delayed and undelayed inputs are active at nearly the same time. A second neuron then sums the number of coincidences and remembers them over a small window in time. This second neuron can be called a leaky integrator or lowpass filter. Structures similar to the correlator shown in Fig. 13 have been found in the owl [5] for doing binaural cross-correlation and in the bat [34] for echo location.

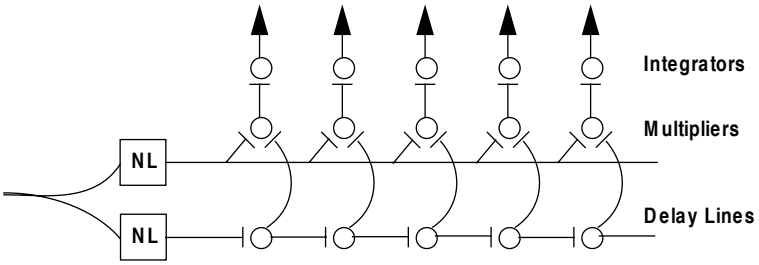


Fig. 13- This simple structure, first proposed by Licklider in 1951, can be used by the brain to calculate a correlogram. For each time lag, delayed (bottom line) and undelayed versions of the auditory signal are multiplied together and integrated. The two boxes labeled "NL" are optional non-linear processing steps.

While the correlogram was first proposed in 1951, only recently has it become feasible to explore the use of the correlogram with multiple seconds of sound. Using an FFT on a Cray YMP supercomputer we can compute one second of a correlogram with about a second of CPU time. The same calculation takes about a half hour on a small personal computer.

An even more efficient implementation is possible using analog silicon VLSI as described by Lyon [17]. Using low-power sub-threshold CMOS transistors, the chip computes a real-time video correlogram from a microphone input. This implementation combines a cascade of analog filters, simulating the cochlear transmission line, with an array of Charge-Coupled Device (CCD) delay lines. At each position in the correlogram array (frequency versus lag) there are four CCD gates, a transistor multiplier, a capacitor to sum the current output, and

video scan-out circuitry. A separate gate array generates the video timing and addresses the correlogram pixels in the proper order. The two chips produce a recognizable correlogram, not as precise as the digital versions, but which can be computed using a single 9 volt radio battery for power.

There are many ways to compute variations of a correlogram. One way to describe these implementations is shown in Fig. 13. This is a generalization of the basic correlator described above and includes two optional non-linearities that modify the neural input. In a method proposed by Patterson [23], only the non-linearity in the undelayed signal path is present. This non-linearity is an adaptive peak picker and produces a binary output when it sees a major peak. As each peak occurs the delayed input signal is transferred to the leaky integrators. An approach first described by Weintraub [36] uses identical non-linearities in each path. Each non-linearity replaces the original waveform with an impulse train that represents the location of each peak in the waveform. In addition, the impulses of the signal are scaled by the energy in the original peak. In both cases, the large amounts of data that are combined to form a single frame of a correlogram help to average out the noise caused by these approximations.

Another technique which might be used to generate a correlogram is to model chopper cells in the Cochlea Nucleus [8] and to count their output spikes. Chopper cells prefer to fire at a fixed rate and tend to lock to sound periodicities. It is easy to imagine that these cells could be used to measure the periodicities in an auditory signal. We have not yet tried to generate a correlogram using this approach.

4 APPLICATIONS

Let us review our progress to date. We believe we have a good understanding of how to make a cochlear model. The models we describe here are a severe simplification of the complex behaviour of the cochlea, designed to preserve the aspect most relavent to auditory processing. This we believe is the temporal information in the signal. While there are many details that remain to be worked out, one can now choose any number of models that can be used to model various aspects of the cochlea.

One aspect that is clear, at least to us from our review of cochlear mechanics, is that the tuning curves can not be sharp enough to account for all the exquisite properties of the human auditory system. But yet the system is quite good at preserving the temporal information in the signal. Even above 3khz, where phase locking to high frequencies is lost, auditory nerves preserve the envelope and thus the timing of the glottal pulses. The correlogram is one way to capture this temporal information.

Given a temporal representation of sound one certainly wonders what it is good for. This section describes the use of the correlogram as a tool for visualization, a model of pitch perception, and our efforts to perform sound separation using this representation.

4.1 Sound Visualization

The most striking property of a correlogram movie is that the visual and acoustic experiences are so similar. It is intuitively appealing to be able to see sounds in much the same way that we hear them. It is, of course, hard to share this kind of experience in a book, but we can illustrate some of the things we have seen.

A simple example is provided by “Strike Note of a Chime,” Demonstration Number 24 from the Acoustical Society of America’s *Auditory Demonstrations* CD [9]. Bells are interesting because they are inharmonic, with several different mechanical modes. Each mode corresponds to a resonance at a different frequency and the inharmomic relationship between these resonances accounts for the rich sound associated with a bell.

Fig. 14 shows several frames of a correlogram of a bell. At first, there are many harmonics and the sound is quite rich. Different overtones decay at different rates as is seen in the second

and third frames. Finally after two seconds, or the last frame, there are only two (inharmonic) components left.

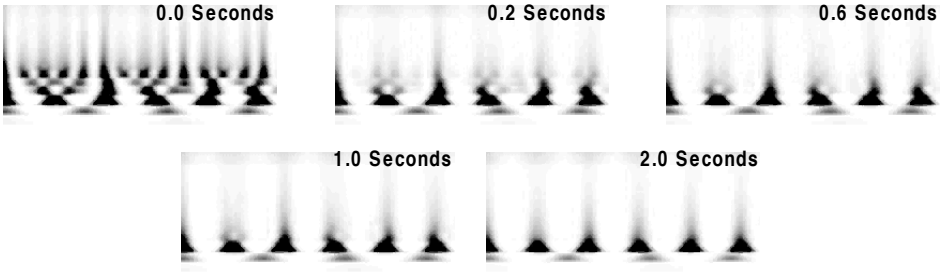


Fig. 14- Correlogram of the Strike Note of a Chime. Five frames show the different decay rates of the resonating modes of an orchestral chime. This example is Demonstration 24 from the Auditory Demonstrations CD [9].

4.2 Pitch

Pitch is an obvious quantity to measure with a correlogram. Licklider originally proposed the correlogram as a pitch model and only recently has it been studied and compared to human performance [20][21][32]. The results closely match the published literature for all experiments except those based on loudness changes.

Pitch is measured from a correlogram as shown in Fig. 15. After the correlogram is computed, evidence for a pitch at each lag is found by summing across channels. The resulting function is called a summary correlogram. It measures how likely a pitch would be perceived with the given time delay. Inverting this time delay gives the resulting pitch frequency.

It is important to realize that pitch is not a single valued function. Pitch is conventionally defined as "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale." But, for many sounds any number of frequencies can be called the pitch. Most engineering solutions reduce pitch to a single valued quantity, but the correlogram pitch detector described in Fig. 15 estimates the likelihood that a pitch exists at the corresponding time delay. If a single pitch estimate is desired then one solution is to choose the largest peak and call this the pitch.

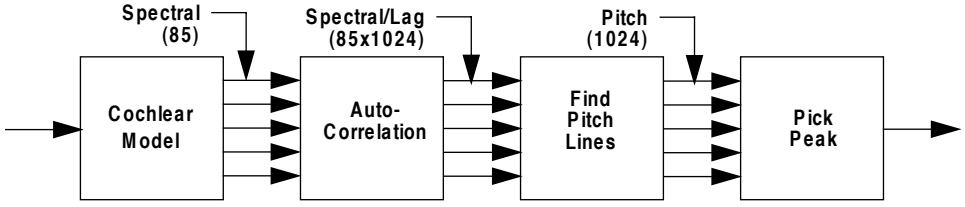


Fig. 15- Human pitch perception can be modeled with this correlogram technique. After computing the correlogram, a summary correlogram is computed (third box) by summing the correlation across channels, or along vertical lines. The numbers in parenthesis show the typical amount of data at each time step.

Fig. 16 shows the processing involved in a pitch detector we have built [32].This pitch detector adds two additional ad-hoc stages to improve the system's performance with real-world sounds. We have not found these stages to be necessary with synthetic sounds, but with real sounds we have found they improve the performance of our pitch detector. Fortunately, neither step is hard to implement with neural circuits.

To compute a pitch, a correlogram of the sound is first non-linearly filtered to emphasize the vertical structures in the correlogram. This is equivalent to biasing the pitch detector so that it will emphasize sounds that are harmonic. The summary correlogram is computed, and then a final stage of sub-harmonic processing is performed. In our pitch detector this is implemented using the narrowed-autocorrelation idea proposed by Brown [3]. This type of pro-

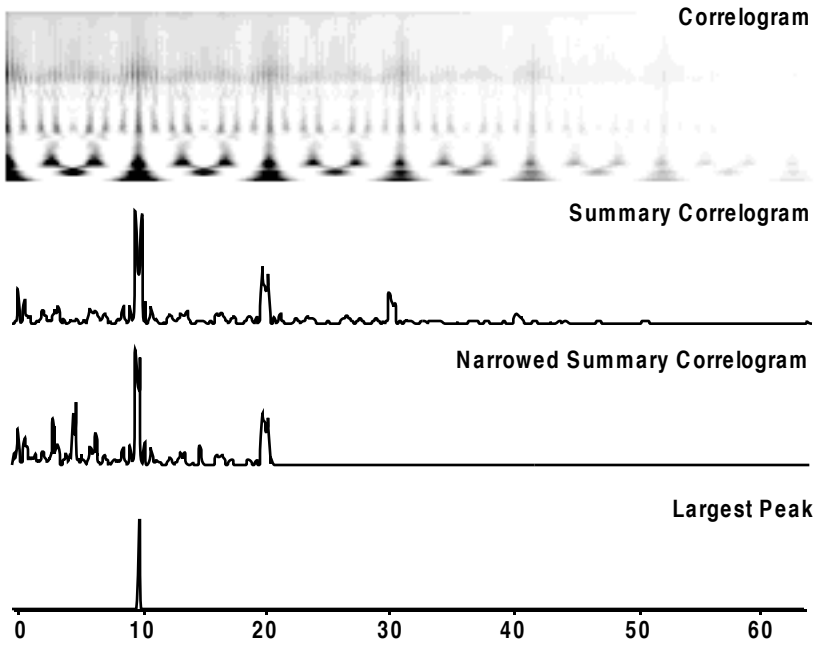


Fig. 16- Pitch of a vowel. Data processing steps in a correlogram pitch detector are illustrated here. After computing the summary or integrated correlogram, subharmonics are considered using the narrowed autocorrelation technique. Finally, if desired, the highest peak can be chosen and considered the pitch.

cessing is equivalent to the sub-harmonic analysis proposed by Hermes [7] and the pitch spiral proposed by Patterson [22].

In our pitch detector, a single pitch value is independently chosen at each frame time (30 times a second). This pitch detector has no history so it is quite happy to choose a completely different pitch at each frame. Humans do not work this way: instead we use the pitch at recent times to help us to choose the most likely pitch in the future. The result is that if two pitches are equally likely then this pitch detector will oscillate between the two possible choices. A better choice would be to model the dynamics of pitch perception, perhaps based on the data for pitch just noticable difference (JND) as a function of time interval [1].

Fig. 17 shows the pitch measured from a sound with an ambiguous pitch, the continuous Shepard tones by Jean-Claude Risset from the ASA *Auditory Demonstrations* CD (Demonstration 27). In this example the pitch is heard to constantly fall. But analysis by correlogram shows that at each frame a number of pitches are possible, each separated by an octave. Our

pitch detector is happy to oscillate between likely pitches but humans tend to follow a single pitch track, perhaps over many octaves.

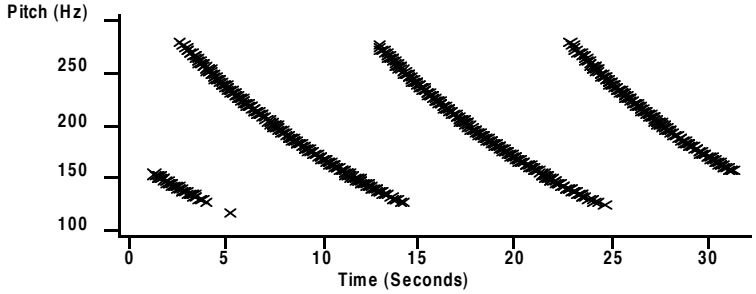


Fig. 17- Pitch of the continuous Shepard tones. Note that the correlogram pitch detector described in this chapter does not enforce any frame-to-frame coherence. Thus it is equally likely to choose either pitch if the summary correlogram assigns the two periods similar magnitudes.

4.3 Sound Separation

Our ultimate goal with correlogram processing is to understand how humans separate the sounds in our environments. Even with a monaural recording we are quite good at separating out the vocals from an instrumental track, hearing a bird as a separate object outdoors, or even listening to a single conversation in a noisy room.

There are many cues [2] that we use to group pieces of a sound into a whole auditory object. Some of the cues we have studied are onsets, pitch, and common modulation. A good example of the power of common modulation is the Reynolds-McAdams oboe [18][26]. In this sound a single oboe sound was analyzed into its even and odd frequency harmonics. Then the harmonics were put back together, but each set of harmonics was independently jittered. At first, the harmonics are fixed and the sound is heard as the original oboe sound. After a few seconds the vibrato is turned on and the two sets of harmonics are heard as separate objects. The odd harmonics sound like a clarinet since clarinets have most of their energy in the odd harmonics. The even harmonics go up an octave in pitch and sound like a soprano.

Fig. 18 shows correlogram frames that represent this sound. Over time the movie shows the even harmonics moving left and right. The odd harmonics are moving independently and the original oboe sound splits into two sounds. The pitch tracks for the two sets of harmonics

are shown in Fig. 19. At this time we do not know whether this grouping is based on detection of the FM modulation or synchronous onset detection in the correlogram domain.

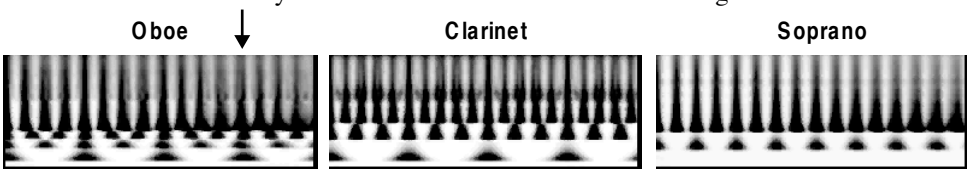


Fig. 18- Three frames illustrating the correlogram of the Reynolds-McAdams oboe. The left frame shows the correlogram of the combined sound (even+odd harmonics) when the sets are slightly inharmonic. Note that the right most pitch line (at arrow) is no longer straight. The middle frame shows the correlogram of just the odd harmonics, or the clarinet. The right frame shows the correlogram of the even harmonics, or soprano. When this soud is played for human listeners, the independent vibrato clearly causes the sound to split into two objects.

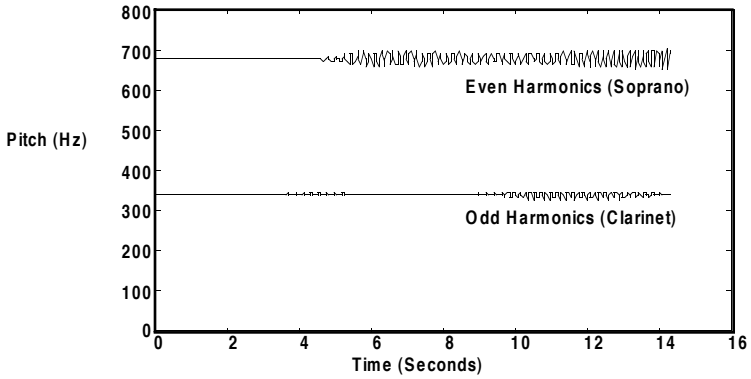


Fig. 19 - Pitch of the odd harmonics (clarinet, bottom) and even harmonics (soprano, top) are shown here as a function of time. At first, the harmonics are held fixed and the original oboe sound is heard. After three seconds the independent vibrato is turned on and the sound separates into a clarinet and a soprano.

More work needs to be done to build models of sound separation that take into account the dynamics of the auditory system. The correlogram can quantify the short term periodicities (less than 25 ms) in the signal but does not capture the information at longer time scales. For example, a voiced signal can be thought of as a vocal tract signal modulated by the glottal pulses. The correlogram does a good job of representing the amplitude modulation or pitch of the voiced signal as activity in a spatial map. But modulations with even lower frequencies, such as the 6 Hz tremelo of a human voice, are not explictly represented. Higher level models of the auditory processing will need to represent these longer time scales in order to understand the dynamics of real sounds.

Acknowledgements: Over the years many of our colleagues have helped us as the ideas presented here evolved. We especially want to acknowledge the discussions we have had with the people at the Hearing Seminar at Stanford's CCRMA, and at Caltech's computational and neural systems group. We would specifically like to thank our colleagues Bill Stafford, Daniel Naar, Richard Duda, and Steve Greenberg for their support and encouragement.

References

[1] Johan 't Hart, René Collier, and Antonie Cohen. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge, England: Cambridge University Press, 1990.

[2] Albert S. Bregman. *Auditory Scene Analysis*. Cambridge, MA: Bradford Book, MIT Press, 1990.

[3] Judith Brown and Miller S. Puckette. "Calculation of a 'narrowed' autocorrelation function." *J. Acoustical Soc. Amer*. 85 (4 1989): 1595-1601.

[4] Laurel H. Carney and Tom C. T. Yin. "Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model." *Journal of Neurophysiology* 60 (5 1988): 1653-1677.

[5] C. E. Carr and M. Konishi. "A circuit for detection of interaural time differences in the brain stem of the barn owl." *The Journal of Neuroscience* 10 (10 1990): 3227-3246.

[6] Richard O. Duda, Richard F. Lyon, and Malcolm Slaney. "Correlograms and the separation of sounds." In *Conference Record. Twenty-fourth Asilomar Conference on Signals, Systems, and Computers* in Pacific Grove, CA, Maple Press, 457-461, 1990.

[7] D. J. Hermes. "Measurement of pitch by subharmonic summation." *J. Acoustical Soc. Amer*. 83 (1 1988): 257.

[8] Michael J. Hewitt, Ray Meddis, and Trevor M. Shackleton. "A computer model of a cochlear-nucleus stellate cell: responses to amplitude-modulated and pure-tone stimuli." *J. Acoustical Soc. Amer*. 91 (4 1992): 2096-2109.

[9] A. J. M. Houtsma, T. D. Rossing, and W. M. Wagenaars. "Auditory Demonstrations." Woodbury, NY: Acoustical Society of America, 1987.

[10] B. M. Johnstone, R. Patuzzi, and G. K. Yates. "Basilar membrane measurements and the travelling wave." *Hearing Research* 22 (1986): 147-153.

[11] J. C. R. Licklider. "A Duplex Theory of Pitch Perception." *Experentia* 7 (128-13 1951). Also reprinted in *Pysiological Acoustics*. E. D. Schubert (ed.). Dowden, Hutchinson and Ross, Inc. Stroudsburg, PA 1979.

[12] Richard F. Lyon. "A computational model of filtering, detection, and compression in the cochlea." In *Proceedings of the 1982 International Conference on Acoustics, Speech and Signal Processing* in Paris, France, IEEE, 1282-1285, 1982.

[13] Richard F. Lyon and Lounette Dyer. "Experiments with a computational model of the cochlea." In *Proceedings of the 1986 International Conference on Acoustics, Speech and Signal Processing* in Tokyo, Japan, IEEE, 1975-1978, 1986.

[14] Richard F. Lyon and Carver Mead. "An analog electronic cochlea." *IEEE Trans. on ASSP* 36 (7 1988): 1119-1134.

[15] Richard F. Lyon and Carver Mead. *Cochlear Hydrodynamics Demystified*. Caltech Computer Science Technical Report, 1989. Caltech-CS-TR-88-4.

[16] R. F. Lyon. "Automatic gain control in cochlear mechanics." in *The Mechanics and Biophysics of Hearing*, ed. P. Dallos, C. D. Geisler, J. W. Matthews, M. A. Ruggero, and C. R. Steele. Springer-Verlag, 1990.

[17] Richard F. Lyon. "CCD correlators for auditory models." In *Twenty-fifth Asilomar Conference on Signals, Systems & Computers* in Pacific Grove, CA, IEEE, 775-789, 1991.

[18] Steve McAdams. "Spectral fushion, spectral parsing and the formation of auditory images." Technical Report STAN-M-22, Center for Computer Research in Musics and Acoustics, Department of Music, Stanford University, Stanford, CA, May, 1984.

[19] Ray Meddis, Michael J. Hewitt, and Trevor M Shackleton. "Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse." *J. Acoustical Soc. Amer*. 87 (4 1990): 1813-1816.

[20] Ray Meddis and Michael Hewitt. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II. Phase sensitivity." *J. Acoustical Soc. Amer*. 89 (6 1991a): 2883-2894.

[21] Ray Meddis and M. J. Hewitt. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification." *J. Acoustical Soc. Amer*. 89 (6 1991b): 2866-2682.

[22] Roy D. Patterson. "A pulse ribbon model of monaural phase perception." *J. Acoustical Soc. Amer*. 82 (5 1987): 1560-1586.

[23] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. "Complex sounds and auditory images." In *Auditory physiology and perception*, ed. Y. Cazals, L. Demany, and K. Horner. 429-446. Oxford: Pergamon, 1991.

[24] J. R. Pierce. "Periodicity and pitch perception." *J. Acoustical Soc. Amer.* 90 (10 1991): 1889-92.

[25] O. F. Ranke. "Theory of operation of the cochlea: A contribution to the hydrodynamics of the cochlea." *J. Acoustical Soc. Amer.* 22 (1950): 772-777.

[26] Roger Reynolds. "Archipeligo." New York: C. F. Peters, 1983.

[27] W. S. Rhode. "Observations of the vibration of the basilar membrane insquirrel monkeys using the Mössbauer technique.", *J. Acoustical Soc. Amer.* 49 (1971): 1218-1231.

[28] Mario A. Ruggero. "Responses to sound of the basilar membrane of the mammalian cochlea." *Current Opinion in Neurobiology* 2 (1992): 449-456.

[29] Shihab Shamma. "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve." *J. Acoustical Soc. Amer.* 78 (5 1985): 1612-1621.

[30] Malcolm Slaney. *Lyon's Cochlear Model*. Apple Computer Technical Report #13. Corporate Library, 20525 Mariani Avenue, Cupertino, CA 95104, 1988.

[31] Malcolm Slaney and Richard. F. Lyon. *Apple Hearing Demo Reel*. Apple Computer Technical Report #25. Corporate Library, 20525 Mariani Avenue, Cupertino, CA 95104, 1991.

[32] Malcolm Slaney and Richard F. Lyon. "A perceptual pitch detector." In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing* in Albuquerque, NM, IEEE, 357-360, 1990.

[33] Charles R. Steele and Larry A. Taber. "Comparison of WKB calculations and experimental results for three-dimensional cochlear models." *J. Acoustical Soc. Amer.* 65 (1979): 1007-1018.

[34] Nobuo Suga. "Cortical computational maps for auditory imaging." *Neural Networks* 3 (1 1990): 3-21.

[35] Clive D. Summerfield and Richard F. Lyon., "ASIC implementation of the Lyon cochlear model." In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing* in San Francisco, CA , IEEE. Volume V. 673-676. 1992.

[36] Mitchel Weintraub. "A theory and computational model of auditory monaural sound separation." Ph.D. Dissertation, Electrical Engineering Department, Stanford University, Stanford, CA, 1985.

[37] George Zweig, R. Lipes, and J. R. Pierce. "The cochlear compromise." *J. Acoustical Soc. Amer.* 59 (1976): 975-982.

[38] J. J. Zwislocki. "Theory of the acoustical action of the cochlea.," *J. Acoustical Soc. Amer.* 22 (1950): 778-784.