A Computational Model of Filtering, Detection, and Compression in the Cochlea

Richard F. Lyon

Fairchild Artificial Intelligence Research Laboratory
4001 Miranda Ave.
Palo Alto, California 94304 USA

ABSTRACT

We claim that speech analysis algorithms should be based on computational models of human audition, starting at the ears. While much is known about how hearing works, little of this knowledge has been applied in the speech analysis field. We propose models of the inner ear, or cochlea, which are expressed as time- and place-domain signal processing operations; i.e. the models are computational expressions of the important functions of the cochlea. The main parts of the models concern mechanical filtering effects and the mapping of mechanical vibrations into neural representation. Our model cleanly separates these effects into time-invariant linear filtering based on a simple *cascade/parallel filterbank* network of second-order sections, plus transduction and compression based on half-wave rectification with a nonlinear *coupled automatic gain control* network. Compared to other speech analysis techniques, this model does a much better job of preserving important detail in both time and frequency, which is important for robust sound analysis. We discuss the ways in which this model differs from more detailed cochlear models.

Introduction

A *computational model* is an algorithm that mimics the relevant behavior of the system being modeled; it differs in this respect from descriptive and analytical models. We present a multi-level sound analysis algorithm which models the behavior of the cochlea, or inner ear, much better than previous sound analysis or speech analysis algorithms; at the same time, it is more computationally practical than previous cochlear models. Our model can also be viewed as simply an approach to speech analysis based on the physiology of hearing, as opposed to the more popular approaches based on the physiology of speech production or on mathematical tricks. The resulting algorithms are suitable for real-time processing of speech and other sounds, since the computational complexity is similar to that of other speech analysis algorithms.

This work is motivated by the observation that there is a large community of researchers who study hearing, and that there is much knowledge that has not been applied seriously by anyone in the speech analysis community; this would seem to be a good place to look for the kind of breakthrough that the speech analysis field so badly needs. As J. B. Allen pointed out, "To understand the hearing process is to understand the cochlea..." [1]; similarly, to implement a hearing machine is to implement a cochlear model.

The specific speech analysis problem that motivates this work is the inability of all current speech analysis algorithms to effectively deal with sounds other than pure simple speech sounds. Because of its superior separation of sounds along time and frequency dimensions, we fully expect that the cochlear model will lead to sound analysis techniques capable of robustly dealing with speech sounds mixed with various noises, and even mixed with other speech sounds. The ultimate performance attainable with this class of techniques should be similar to human performance (when integrated into a system that utilizes the same knowledge sources available to the human on a given task), but these should not be regarded as techniques designed to give super-human performance on tasks such as intelligibility improvement.

The model we present here is really a severe simplification of the complex behavior of the cochlea, designed to preserve the aspects most relevant to sound separation and speech parameterization. The main simplification is the separation of the interacting behaviors of the basilar membrane and the organ of Corti into non-interacting models: simple time-invariant filtering, followed by an almost trivial detection nonlinearity, and finally a rather complex nonlinear mechanism that compresses the huge dynamic range of the mechanical domain into a range appropriate for neural representation. This last stage lumps a number of physiological effects, including mechanical nonlinearities, into one computational mechanism.

Background and Approach

It has long been recognized that sounds are best characterized in a "frequency domain", and that the cochlea performs the job of transforming the incoming time-domain pressure signal into this other domain. The exact nature of this frequency domain has not been well clarified, however. Ohm's acoustic law is particularly misleading, saying that the ear is insensitive to phase implies a misunderstanding, or at least some hidden assumptions, about the nature of the frequency domain. The concept of short-time spectrum provides some clarification, but not enough. Concepts such as smoothed filterbank envelopes, LPC spectra, etc., never quite managed to capture the right combination of time-domain and spectral effects to tell the difference between complex single sounds and separate unfusible sounds with similar short-time spectra.

To resolve such problems, we have to put much more emphasis on the time-domain detail that survives beyond the transformation in the cochlea. We end up needing algorithms that combine the best features of the old "place", "volley", and "telephone" theories of hearing. The neural representation of sounds as patterns of spikes undergoes extensive processing in the central nervous system; those further levels of processing, involving pattern detection by correlation and related techniques, are less well understood, and will be mentioned only briefly in this paper.

Many cochlear models have been reported in the past. Most are models of only the mechanical motion of the basilar membrane, including nonlinear and active effects, to various degrees of fidelity. Various approximations to the wave mechanics of the cochlea are exploited to give varying degrees of simplification in the models. For example, the 3-dimensional chambers are modeled in one or two dimensions, the wavelengths are assumed to be long compared to lateral dimensions, the longitudinal elastic coupling is ignored, and the properties of the structure are assumed to be time-invariant and passive. These approximations all seem to be reasonably harmless to the

CH 1746-7/82/0000 - 1282 $ 00.75 © 1982 IEEE

kinds of effects we wish to consider. Another important approximation that leads to a simple way to evaluate transfer functions is that the mechanical properties (mass, stiffness, loss) change slowly enough with distance that no significant amount of wave energy is reflected [2, 3]. All that remains is for us to sample the model along the spatial dimension, and approximate the sections with lumped parameter filter structures. For more detail on cochlear mechanical modeling, including recent bibliographies, see [4, 5, 6].

Some hearing models include a "second filter" of various sorts [1, 7], transduction nonlinearities [8, 9], and simple compression mechanisms [8]. The second-filter can be included in the filtering part of our model, as long as it is linear and time-invariant. Almost any detection nonlinearity is adequate for high-frequency bands (only the short-time envelope matters), but correct representation of low frequencies requires that the nonlinearity be primarily half-wave. Compression mechanisms as simple as logarithmic point nonlinearities are popular in speech analysis, but are very inadequate at preserving detail of high-energy signal regions and suppressing noise in low-energy regions. Simple per-channel automatic-gain-control (AGC) mechanisms are a little better, viewed in the time domain, but still don't adequately handle wide variations of energy across the frequency dimension. Hence we propose a *coupled AGC* that adapts in both time and frequency dimensions.

Filtering

As mentioned above, our model assumes that cochlear filtering can be modeled as time-invariant and linear; but there is much evidence that, in detail, these assumptions are not correct [10]. However, for the purpose of speech analysis, active and nonlinear effects can be accounted for adequately by lumping them into the compression mechanism. That is, we assume the purpose of active mechanisms is to boost the level of weak signals, and the purpose of nonlinear loss elements is to reduce the excitation due to strong signals. Other nonlinear effects, such as cubic difference tones, are assumed to be by-products relatively unimportant to normal hearing.

We do not explicitly use a frequency domain, but rather a place domain. We use a discrete-place approximation to the physical structure of the cochlea, indexed by channel number; different channels have different frequency sensitivities, and can be characterized by filter transfer functions.

We start by adopting the conventional RLC transmission-line analogue to the one-dimensional long-wave hydrodynamic model of basilar membrane motion [2, 3]. For a given frequency, a pressure wave propagates with wave-length and attenuation given by a complex wavenumber k, a function of place, without reflection. For a short section of transmission line, of length dx, with nearly constant wavenumber k (the reciprocal of Zweig's parameter χ [2]), the complex transfer function to the pressure wave of the chosen frequency ω is:

$$\frac{P}{P_0} = e^{-ikdx}, \text{ with } k = \frac{c}{\sqrt{\omega_R^2 + i\omega\omega_R/Q - \omega^2}},$$

where ω_R depends on place, and c is a constant that depends upon how x is measured and upon other parameters of the RLC model.

For longer sections, we simply integrate the quantity being exponentiated along the length of the section (i.e. just average k over the section), and, depending on how physical we want to be, perhaps throw in a factor to account for conservation of energy [2]. Since k depends on frequency, we do the integration and exponentiation for many frequencies to plot the transfer function of a section of the transmission-line model.

The result is a notch filter, which for short enough sections can be accurately approximated by a single biquadratic filter transfer function. The notch is formed by a high-Q zero pair near a lower-Q pole pair; see figures 1 and 2.

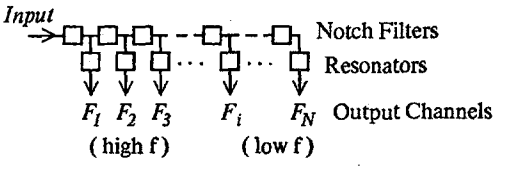

Fig. 1. Block diagram of the cascade/parallel filterbank

Now we have a model of traveling pressure waves, computationally expressed as a cascade of biquad filters, each modeling a short section of the cochlea. In the cochlea, the resonance frequencies that determine the notch locations change approximately geometrically, starting with 20 kHz at the input end, and terminating at about 50 Hz. If we look at any representative place on the cochlea, or at any stage output in our notch filter cascade, and we ignore frequencies over 20 kHz, we see a very sharp low-pass function. The cascaded notch filters conspire to make a collection of low-pass filters with very steep rolloff.

To convert pressure waves to basilar membrane motion or velocity, we still need to add a resonator. In the RLC transmission line model, we see that the current (membrane velocity) through a shunt leg is related to the voltage (pressure) by a single-tuned series resonator. The resonator is just another second-order filter, with a zero at DC and a high-Q pole-pair located between the previous and next notch filter zero pairs. The composite transfer function from sound input to velocity at a place is the "tuning curve" of our model; it is an asymmetric bandpass function that simultaneously provides good frequency resolution by having a sharp rolloff, and good time resolution by being relatively wide bandwidth.

We have defined an unusual general filter structure, consisting of a cascade of second-order filters plus a parallel collection of second-order filters connected after each stage of the cascade. We call this one-input many-output structure, illustrated in figure 1, a cascade/parallel filterbank. It has the useful property that the sum of the orders of the transfer functions from input to outputs greatly exceeds the sum of the orders of the component sections. That is, it achieves an economy of computation by using the same filter sections in many high-order transfer functions, by directly modeling the structure of a sectioned cochlear transmission line.

Several models of cochlear mechanics include a "micromechanical" second filter, which is a resonance in the organ of Corti that contributes a zero pair about an octave below the basilar membrane resonance [1]. We can easily include this in our computational model by putting this zero pair in the resonator section; this becomes another biquad section, if the zero at DC is separately implemented with a simple first-order-difference filter. This first-order section can be in front of the cascade/parallel filterbank, rather than being duplicated in each channel.

All of the filters and transfer functions being discussed can be equally well implemented (computed) with either continuous-time or discrete-time techniques, in either analogue or digital technologies. We illustrate them in continuous-time s-plane notation for simplicity, since the pictures of poles and zeroes are identical except for scale as frequency is changed, as long as Q is constant. See figure 2 for s-plane pole-zero plots and transfer functions with typical parameters.

Of course, with this general filterbank structure, the frequencies are not confined to a geometric spacing, and the Q's need not be constant. For better cochlear modeling, or to concentrate resolution in the region of maximum speech information, resonant frequencies should be spaced further apart, and Q's reduced, near the extremes of the band of interest.

The main function of this filtering section is to separate complex mixtures of sounds into high-signal-to-noise-ratio regions, mainly by separating different frequencies into

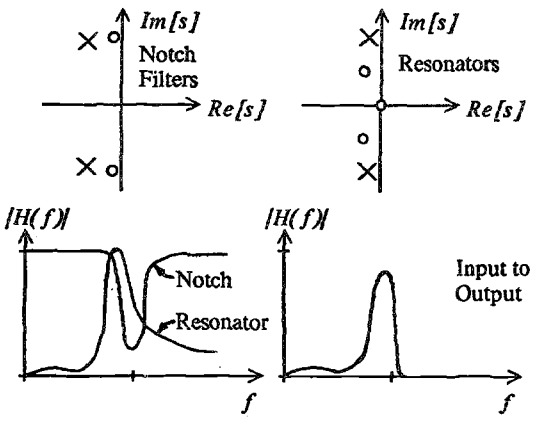

Fig. 2. Pole-zero plots and transfer functions of filters used in the filterbank.

different places, but also by preserving enough time resolution to, for example, separate the responses to separate pitch pulses. Thus, simultaneous voiced speech sounds that differ in some formants and in pitch will be separated into recognizably distinct patterns of activity at the output.

Detection

The outputs of the filtering model are bandpass functions of the original input waveform, and thus may be thought of as zero-mean "carrier" signals. To convert them to a more useful form, we need to "amplitude demodulate" them by using some kind of a detection nonlinearity, such as the diode in an AM radio. Although the exact nature of the nonlinearity may not be critical, handling this stage reasonably correctly does require some special attention to low frequencies, and in the case of discrete-time implementations, to very high frequencies.

We need to recognize that the neural representation of signals, which the output of this model is trying to mimic, has a bandwidth at least as high as the full range of voice pitch, and probably exceeding 2 kHz. This allows us to easily represent the time structure of formant-frequency carriers which are AM modulated at the pitch rate. But it also means that there will be a range of low-frequency "carriers" that can be synchronously represented in the output bandwidth; signals in this range are conveyed more nearly as direct signals than as envelopes, thereby preserving phase information. If we want to use the same nonlinearity for these low frequencies as we use for higher frequencies, and we want the apparent pitch of the result to agree for a fundamental and for an AM modulated carrier, then the nonlinearity has to be half-wave. A full-wave or square-law nonlinearity would preserve the pitch of an AM modulated signal, but double the pitch of a fundamental; this would be unacceptable. There is also considerable physiological evidence for a half-wave detection function in the hair cells of the organ of Corti [11]. The exact shape of the half-wave nonlinearity is not obvious; proposals include "soft" half-wave [8], and exponential [9]. We propose to use instead a simple "ideal" half-wave rectifier, which is very easy to implement and to understand, and whose "gain" is independent of the input signal amplitude.

In discrete-time implementations, the use of a nonlinearity produces harmonics which may lie outside of the Nyquist bandwidth. The non-bandlimited distorted signal will alias back into the baseband. Most of the high-frequency energy from half-wave rectification is in the second harmonic, and should be kept in-band by over-sampling by at least a factor of two. Higher-order distortion products are less important.

After the detection nonlinearity, we can lowpass to a bandwidth consistent with the neural domain, and decimate. If we are not doing binaural processing, 1 kHz is probably an adequate bandwidth; for the benefit of nonlinear processing that follows, it would be a good idea to keep the signal oversampled by a factor of two.

Compression

Consider the problem of producing a high-quality printed spectrogram, maintaining locally high contrast in the face of tremendous variations in the average input power level across time and frequency. This requires compression of a large dynamic range signal into a halftone pattern; the problem is very similar to that faced by the human auditory system in converting sounds to neural firing patterns. The output rates vary over only about two decimal orders of magnitude as the input power varies over twelve or more orders of magnitude from threshold of hearing to threshold of pain. What are the properties of this compression, and what physiological mechanisms achieve it? These are intriguing questions which presently have only very sketchy answers.

The concept of an automatic gain control, which controls the forward path gain of a system in an attempt to keep the output level nearly constant, has been in use in electronic systems for a long time. However, no AGC is able to handle the kinds of signal ranges and achieve the degree of compression that our ears can, without severely distorting the signal quality. Another common compression technique is to use a compressive nonlinearity, such as the logarithm, to effectively reduce the instantaneous gain applied to large signals, while increasing the gain applied to small signals. Applying this to speech spectra gives the familiar effect of rather flattened peaks and severely unstable or noisy behavior in the valleys. In printed spectrograms, peaks are so flattened that it is often difficult to localize formant tracks more accurately than a few hundred hertz. What is needed is an adaptation mechanism that can apply a varying gain across time and frequency dimensions, maintaining sharp peaks and clean valleys, emphasizing onsets and offsets, and de-emphasizing overall spectral tilt and gradual loudness changes.

There are actually a rather large number of suspected adaptive mechanisms in the human auditory system, operating in different domains, at different rates, and covering different parts of the entire 120 dB range of sound levels. For example, the gain applied to very low level signals (0 to 40 dB SPL) may be effectively enhanced by active mechanisms in the organ of Corti; efferent signals stimulate the outer hair cells, causing stereocilia to exert forces, just like muscles, which might be the source of the "superregenerative" active mechanisms. At higher levels the same mechanism, operating in a different phase, may actually reduce the bending of the cochlear partition, causing reduced sensitivity and lower frequency selectivity. At very high levels, the stapedial reflex reduces the mechanical coupling efficiency of the middle ear, protecting the cochlea from harmful levels of vibration. Other mechanisms within the cochlea may include a varying "DC bias" in the basilar membrane position, caused by hair cell interactions, that affect the operating point on their detection nonlinearity; and changes in the concentration of K^+ ions in the endolymph in the cochlear duct may change the sensitivity of the inner hair cells.

Perhaps the most important adaptation mechanism in sensory systems is lateral inhibition. Sensory neurons with a large response reduce their own gain as well as the gain of others nearby, by way of lateral distribution of their outputs to inhibitory synapses on neighboring sensory neurons [12]. Of all senses, probably only hearing and vision require mechanisms beyond lateral inhibition to accommodate their large input range; for a description of the role of lateral inhibition in vision, see [13].

The closest model in the literature to the one we propose is the transduction model of [8], which includes a single-channel model of the adaptive response of a hair cell and

its associated primary auditory neuron. A collection of single-channel AGC's of this sort has the same problem as the logarithm: peaks are flattened as all channels force their outputs to about the same level. If we simply take that model, and add some kind of coupling between nearby channels so that gains are somewhat interdependent, we get a reasonably good model. The trouble is that the time constant of this coupled AGC, like most AGC's, is strongly dependent on signal level. For the range of signals we need to deal with, this effect can be reduced enough by using a controlled-gain element with a super-linear control function; that is, the gain should be proportional to perhaps the cube or the exponential of the control signal level.

Instead of a cube-law controlled gain we can use a cascade of three stages of bilinear elements (simple multipliers), with possibly separate control signals, time constants, and degrees of coupling on each. If the slowest variable-gain stage operates on a slow "syllabic" time scale and with complete coupling, we can move it out to in front of the filtering (like the stapedial reflex), reducing the dynamic range required in the whole system without introducing much distortion. The next two stages of gain control can operate more locally and more quickly after the filtering and detection, in just about any way we choose. The only hard part is to pick the details. For example, we still probably need to include a compressive nonlinearity (limiter) somewhere in the system, so that an unbounded input will produce a bounded output; a hard limiter may be just the thing; or, in Schroeder's model, adding a current-limiting resistor is the simple solution.

We propose the following discrete-time algorithm as a straw-man version of the coupled-AGC compression network (see figure 3).

$$\left.\begin{aligned} Output_i &= \text{Limit}[Detect_i\, Gain_A\, Gain_{B,i}\, Gain_{C,i}] \\ Excess_i &= Output_i - Target \\ Gain_{C,i} &= Z^{-1}[(1-\varepsilon_C)Gain_{C,i} - \varepsilon_C(Wt_{C,i}\cdot Excess)] \\ Gain_{B,i} &= Z^{-1}[(1-\varepsilon_B)Gain_{B,i} - \varepsilon_B(Wt_{B,i}\cdot Excess)] \end{aligned}\right\} \text{for all } i$$

$$Gain_A = Z^{-1}[(1-\varepsilon_A)Gain_A - \varepsilon_A(Wt_A\cdot Excess)]$$

Output is the final vector of signals that represent the high-quality spectrogram, with place index i; *Detect* is the vector of outputs of the detection model. *Excess* is a vector for feedback in the AGC loop; and *Target* is approximately the desired output level.

$Gain_A$ is the gain control signal that adjusts the overall signal level, independent of channel index; this gain can be moved to before the filtering, with little effect. $Gain_B$ and $Gain_C$ are vectors of two levels of per-channel gains. Wt_A is a vector of weights from all channels to the overall gain; most likely these weights are all equal. $Wt_{B,i}$ and $Wt_{C,i}$ are vectors of cross-coupling weights from all channels to channel i. The vector inner product function is designated by dot. The slowest AGC filter time constant is T/ε_A, for sampling interval T. The faster AGC filter time constants are T/ε_B and T/ε_C.

Limit is the compressive nonlinearity that produces a bounded output; its maximum output level should be at least an order of magnitude higher than *Target*, the desired average output. With this scheme, an average output of $0.9\,Target$ is consistent with a steady-state gain reduction of 1000 relative to the small-signal gain, corresponding to a 60 dB accommodation of input level. Another 60 dB of accommodation occurs as the average output rises to $0.99\,Target$. Peaks localized in time or place can be very much higher than *Target*, especially at onsets before the gain adapts.

Discussion and Conclusion

We have presented a simple and somewhat flexible speech analysis algorithm based on cochlear models, which is computationally attractive. If second-order sections are implemented with five multiplies per sample, and we sample the speech signal at 20 kHz, then the filtering

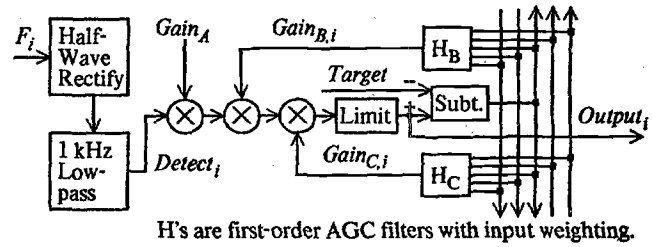

Fig. 3. Block diagram of one channel of the detection and compression models.

complexity is 200K multiplies per second per channel. With 64 channels, the resulting 12.8M multiplies per second can be handled with one or a few modern chips. The corresponding data memory of 256 words (by 32 bits, say) also fits on a chip. Similar numbers apply for the compression network, depending on what sample rate reduction is done, and how many nonzero coupling coefficients are implemented. Only conventional time-domain signal flow-graph kinds of computations are needed, so these algorithms are suitable for almost any general-purpose or special-purpose computing architecture.

The properties of this algorithm are only now being evaluated, in the context of speech recognition and display. We expect that the improved relation of frequency-domain and time-domain information will lead to a more readable spectrogram-type image of speech sounds, and, in conjunction with further levels of neural processing, will eventually achieve a radically better version of real-time visible speech.

Obviously, this work is very preliminary. We hope by this publication to interest other researchers in this exciting new direction in sound analysis.

References

[1]Allen, J. B., "Cochlear Modeling - 1980," ICASSP 81, pp. 766-769, Atlanta, 1981.

[2]Zweig, G., R. Lipes, and J. R. Pierce, "The Cochlear Compromise," JASA 59, pp. 975-982, 1976.

[3]Schroeder, M. R., "An Integrable Model for the Basilar Membrane," JASA 53, pp. 429-434, 1973.

[4]Zwislocki, J. J., "Sound Analysis in the Ear: A History of Discoveries," American Scientist 69, pp.184-192, 1981.

[5]Matthews, J. W., "Mechanical Modeling of Nonlinear Phenomena Observed in the Peripheral Auditory System," D.Sc. thesis, Washington University, St. Louis, Missouri, 1980.

[6]Neely, S. T., "Fourth-order Partition Dynamics for a Two-dimensional Model of the Cochlea," D.Sc. thesis, Washington University, St. Louis, Missouri, 1981.

[7]Nilsson, H. G., "A Comparison of Models for Sharpening of Frequency Selectivity in the Cochlea," Biol. Cybernetics 28, pp. 177-181, 1978.

[8]Schroeder, M. R. and J. L. Hall, "Model for Mechanical to Neural Transduction in the Auditory Receptor," JASA 55, pp. 1055-1060, 1974.

[9]Kim, D. O., and C. E. Molnar, "A Population Study of Cochlear Nerve Fibers: Comparison of Spatial Distributions of Average-Rate and Phase-Locking Measures of Responses to Single Tones," J. of Neurophysiology 42, pp. 16-30, 1979.

[10]Kim, D. O., C. E. Molnar, and J. W. Matthews, "Cochlear Mechanics: Nonlinear Behavior in Two-Tone Responses as Reflected in Cochlear-Nerve-Fiber Responses and in Ear-Canal Sound Pressure," JASA 67, pp. 1704-1721, 1980.

[11]Pfeiffer, R. R., and D. O. Kim, "Response Patterns of Single Cochlear Nerve Fibers to Click Stimuli: Descriptions for Cat, " JASA 52, 1972.

[12]von Bekesy, G., *Sensory Inhibition*, Princeton University Press, 1967.

[13]Werblin, F. S., "The Control of Sensitivity in the Retina," Scientific American, pp. 70-79, January 1973.