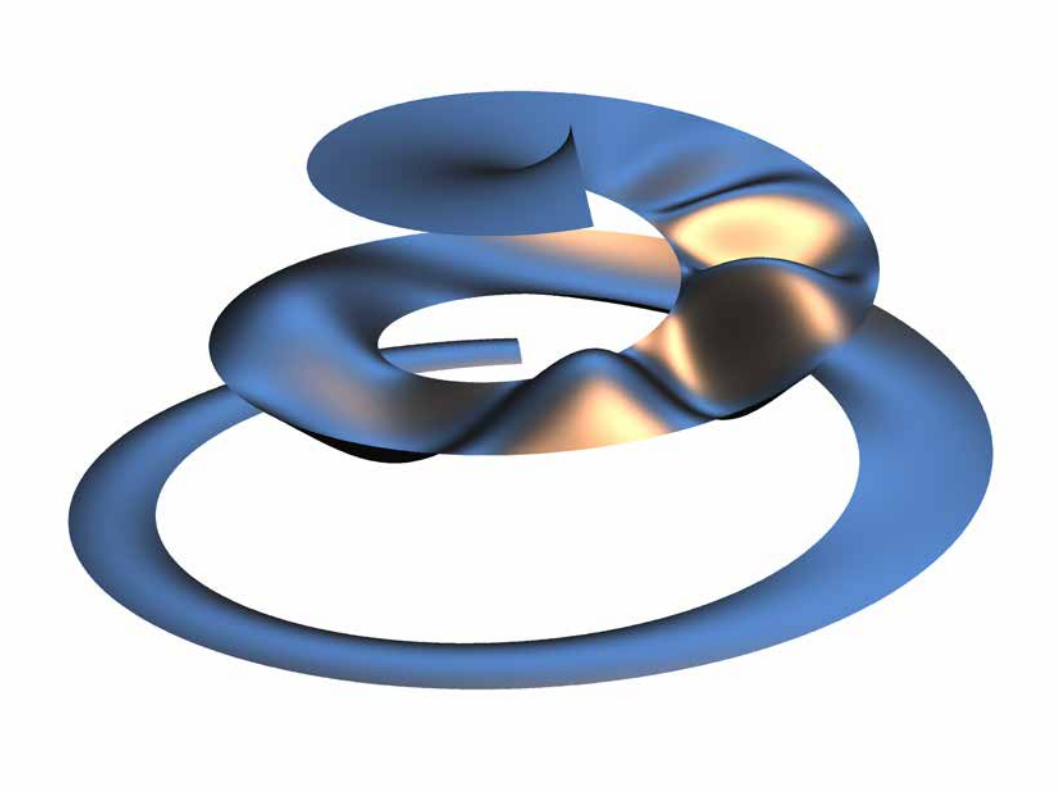
Human and Machine Hearing

—

Extracting Meaning from Sound

– – – –

(Figures and boxes from the 2017 Cambridge University Press book – copyright Richard F. Lyon)



Richard F. Lyon

July 3, 2017

Un beau visage est le plus beau de tous les spectacles ; & l'harmonie la plus douce est le son de voix de celle que l'on aime.

A fine Face is the finest of all Sights, and the sweetest Musick, the Sound of her Voice whom we love.

—Jean La Bruyère (1713) from 1691 French original.

This book is dedicated to my family: my beautiful, smart, cheerful, successful, inspiring, and sweet-voiced wife Peggy Asprey and our awesome children Susan and Erik—they are the loves of my life, and my fortune. Though this book has sometimes absorbed too much of my attention, they have all supported me in writing it, in so many ways. They are my finest of all sights, and sweetest music; they sustain me.

On History and Connection Boxes

While there are historical comments, and comments on connections to related concepts in other fields, throughout many chapters, I have segregated some of them into boxes, both to highlight them and to keep them out of the way. In many cases, my aim is to honor the sources of the ideas we use, while trying to make the literature more accessible by saying a few words about how it connects. I trust that my mention of old technologies such as vacuum tube (valve) amplifiers and Helmholtz resonators and flame manometers will be received as intended: as clues to a very interesting heritage from generations of giants whose shoulders we stand upon, in both human and machine hearing.

My own EE training was in the era of transistors and early integrated circuits, when courses like "Circuits, Signals, and Systems" were all about analog continuous-time technology. In modern times, signals and systems are taught from the beginning with discrete-time concepts, for good reasons having to do both with pedagogy and the modern medium of implementation in digital computers. Although modern engineers may view sound naturally as the kind of discrete-time sampled data that they work with in computers, I have chosen to stick with continuous time as the primary conceptual domain in this work, since sound and the ear really exist in that domain. I hope that readers will not view the continuous-time domain as something out of history—it is the real world.

Online materials

Find errata, and links to code and other resources, at machinehearing.org.

Thanks

There are many people who have cared enough about this work to spend time helping and encouraging me. First among them is Roy Patterson, without whose encouragement I could never have even started, and who has continued to inspire me through the slow process.

Among my readers who have given me actionable feedback, Ryan "Rif" Rifkin stands out; he found me more bugs than everyone else combined. Others who contributed, whether by carefully reading chapters or giving feedback on overall impressions, include: Jont Allen, Peggy Asprey, Fred Bertsch, Alex Brandmeyer, Peter Cariani, Wan-Teh Chang, Sourish Chaudhuri, Brian Clark, Lynn Conway, Achal Dave, Bertrand Delgutte, Dick Duda, Diek Duifhuis, Dan Ellis, Doug Eck, Dylan Freedman, Jarret Gaddy, Daniel Galvez, Dan Geisler, Pascal Getreuer, Chet Gnegy, Alex Gutkin, Yuan Hao, Thad Hughes, Aren Jansen, James Kates, Nelson Kiang, Ross Koningstein, Harry Levitt, Carver Mead, Ray Meddis, Harold Mills, Channing Moore, Stephen Neely, Eric Nichols, Fritz Obermeyer, Ratheet Pandya, Brian Patton, Justin Paul, Manoj Plakal, Jay Ponte, Rocky Rhodes, David Ross, Mario Ruggero, R. J. Ryan, Bryan Seybold, Shihab Shamma, Phaedon Sinis, Jan Skoglund, Malcolm Slaney, Daisy Stanton, Rich Stern, John L. Stewart, Ian Sturdy, Jeremy Thorpe, George Tzanetakis, Marcel van der Heijden, Tom Walters, Yuxuan Wang, W. Bruce Warr, Lloyd Watts, Ron Weiss, Kevin Wilson, Kevin Woods, Ying Xiao, Bill Yost, Tao Zhang, and probably others that I have missed. Many thanks to all!

And finally, huge thanks to Lauren Cowles, my editor at Cambridge University Press, for her years of patience in helping to make this book happen.

Part I

Sound Analysis and Representation Overview

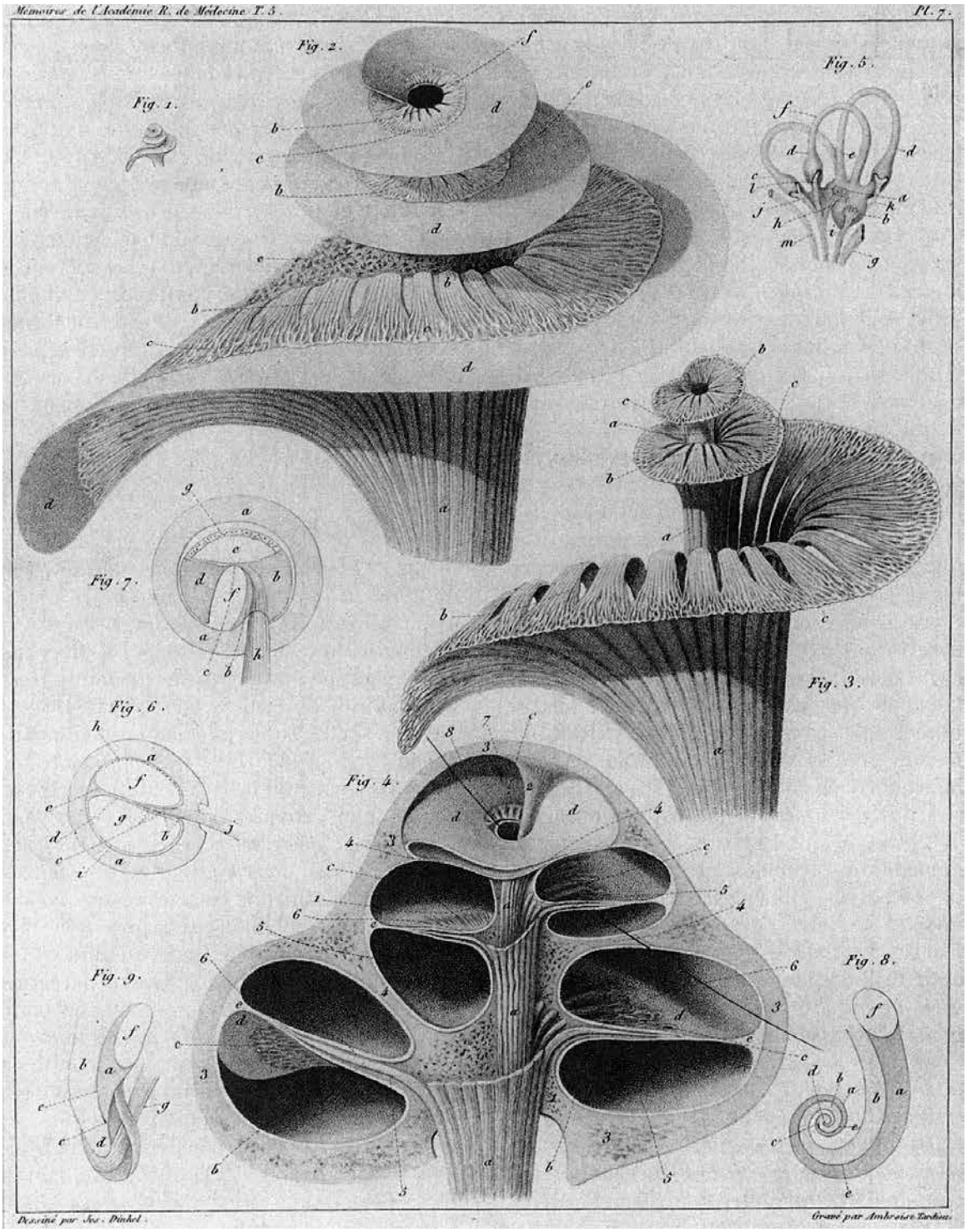
Part I Dedication: John Pierce

This part is dedicated to the memory of John Robinson Pierce (1910–2002). John was a dear friend and mentor for many years, beginning in my undergraduate years at Caltech. He gave me a summer job doing lab work on electronic musical instruments, and then on digital codecs that led to my first journal article. He persuaded his colleagues at Bell Labs to take me on as an intern, even after they had objected to my "less than an A in some important subjects." I owe my knowledge of digital signal processing to this great start with the early researchers and practitioners there. Pierce's work with George Zweig and Richard Lipes at Caltech, after I had left, became one of the most important influences on my thinking in hearing: the wave analysis that led to my filter-cascade approach to modeling the cochlea (Zweig, Lipes, and Pierce, 1976).

Pierce was better known for his work outside of hearing: from his early work in traveling-wave tubes and communication satellites at Bell Labs, his coining of the word *transistor*, his chief technologist role at the Jet Propulsion Laboratory, his science fiction writing under the pen name J. J. Coupling, through his enormous influence on computer music starting at Bell Labs and continuing at Stanford's Center for Computer Research in Music and Acoustics (CCRMA) in the 1980s and 1990s. His regular attendance at CCRMA's weekly hearing seminar provided a huge benefit to many of us in the hearing field. He continued to conduct and publish hearing research at Stanford even in his 80s, for example providing clarity on important issues in pitch perception (Pierce, 1991).

In Part I, we survey our concept of what the machine hearing field is, and how it relates to conventional acoustic approaches to sound processing and to a range of theories of hearing. We include a brief overview of human hearing from the conventional psychoacoustics and physiology points of view, which provide the data and some of the models that we build on.

Throughout the book, but especially in Part 1, I strive to make my point of view clear, describing the relationship of my conceptual framework and models to other concepts, old and new. Partly, this approach is to raise awareness about some older concepts that are still "hanging around," causing unneeded distraction and confusion. Equally importantly, it is to draw attention to ideas that still need more research and exploration, to see how well they hold up when experiments are designed specifically to test them. My hope is that this approach will help others find useful directions in which to extend, or to challenge, what I have gathered here.



Engraving of the structures of the cochlea and spiral ganglion by Gilbert Breschet (1836), before the discovery and description of the microscopic organ of Corti at the interface between the cochlea's ducts in 1851 by Alfonso Giacomo Gaspare Corti.

Chapter 1

Introduction

> . . . things inanimate have mov'd,
> And, as with living Souls, have been inform'd,
> By Magick Numbers and persuasive Sound.
>
> — William Congreve (1697) *The Mourning Bride*

> The ear is a most complex and beautiful organ. It is the most perfect
> *acoustic*, or hearing instrument, with which we are acquainted,
> and the ingenuity and skill of man would be in vain exercised to imitate it.
>
> — John Frost (1838), *The Class Book of Nature: Comprising Lessons on the Universe, the Three Kingdoms of Nature, and the Form and Structure of the Human Body*

Would it truly be in vain to exercise our ingenuity to imitate the ear? It would have been, in the 1800s—but now we are beginning to do so, using the "magick" of numbers. Machines imitating the ear already perform useful services for us: answering our queries, telling us what music is playing, locating gunshots, and more. By imitating ears more faithfully, we will be able to make machines hear even better. The goal of this book is to teach readers how to do so.

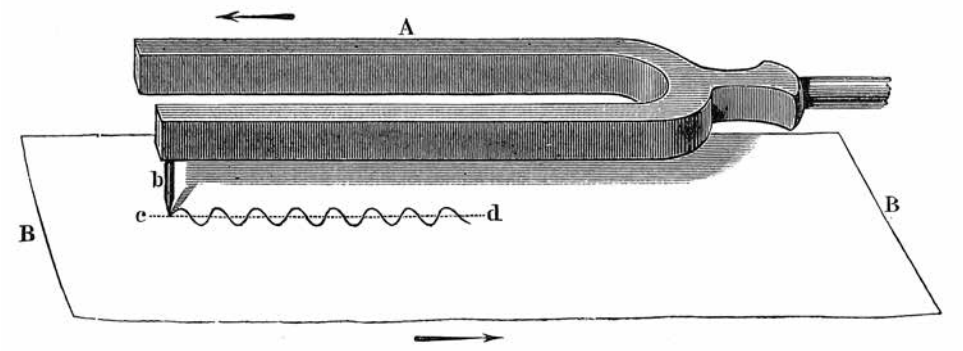


Figure 1.1: Helmholtz explained the idea of a sound's waveform via this diagram of a tuning fork with a stylus point attached, drawing its vibration on a moving piece of paper.

Figure 1.2: A make-it-yourself acoustic siren, much like August Seebeck's, as shown by Alfred M. Mayer (1878). The spinning disk, driven from a crank via string and pulleys, interrupts a stream of air from the tube to make waves of sound pressure that we hear as a tone. Different tones can be made by moving the tube to a different row of holes, or by changing the disk to one with a different pattern of holes. August Seebeck and Hermann von Helmholtz were among the nineteenth-century scientists who used such devices in their research that contributed to connecting the physical and perceptual properties of musical tones to the mechanisms of human hearing—though their theories were somewhat in opposition to each other.

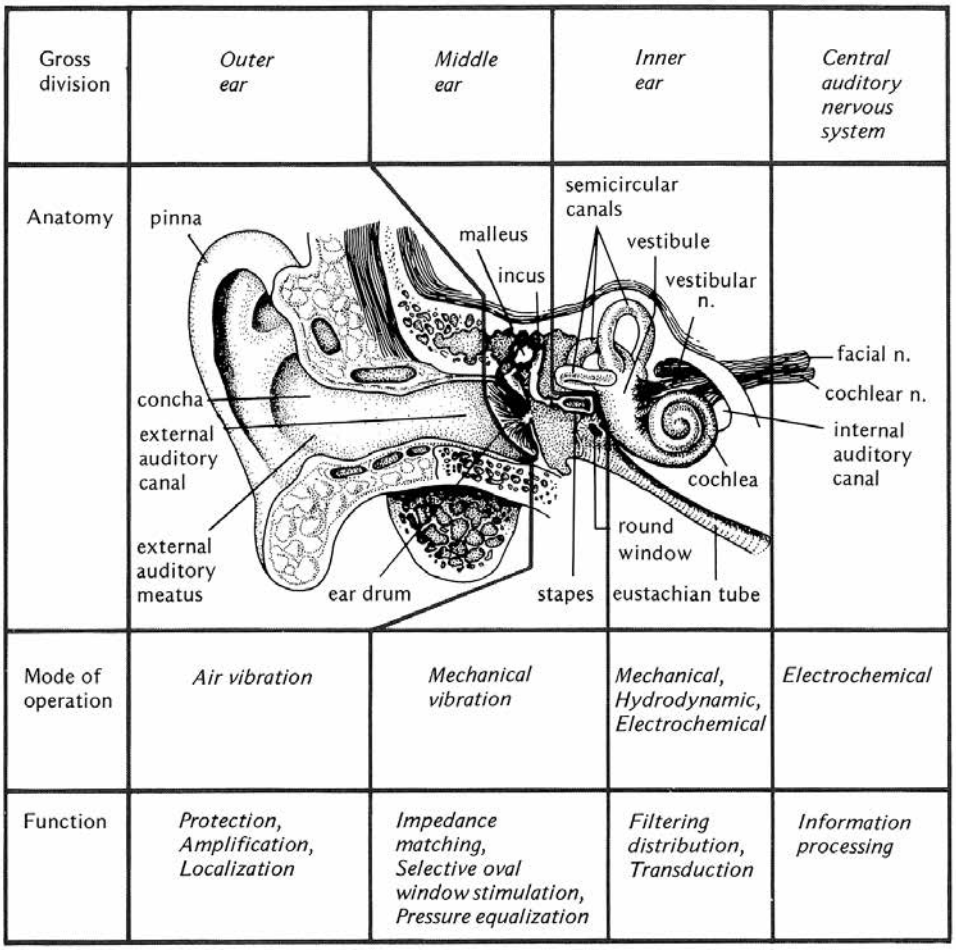


Figure 1.3: Ear diagram by Yost (2007). While the anatomy and modes of operation are important, we are most interested in emulating the *function*, described in the bottom row. The *information processing* in the central nervous system—the bit where meaning is extracted—is the part that remains most open to exploration and speculation. [Figure 6.1 (Yost, 2007) reproduced with permission of Wiliam A. Yost.]

Nomenclature: What to Call This Endeavor

The terms *computer vision* and *machine vision* are in wide use, not quite interchangeably, the former having a more computer-science connotation, and the latter a more industrial or applications connotation. Terms like *computer hearing*, *computational hearing*, and *computer listening* seem awkward to me, especially since I spent a lot of years building analog electronic models of hearing, probably not qualifying as computers. And what about *listening* or *audition* as a better analogy to *vision*? Several of these terms have overloaded meanings: we can convene a hearing, or perform in an audition, or plant listening devices. The term *machine listening* is sometimes used, but mostly in connection with music listening and performance.

The term *machine hearing* has a strong history at Stanford's computer music lab, CCRMA. In their 1992 progress report, Bernard Mont-Reynaud (1992) wrote a section on machine hearing, which noted that "The purpose of this research is to design a model of Machine Hearing and implement it in a collection of computer programs that capture essential aspects of human hearing including source formation and selective attention to one source (the 'cocktail party problem') without tying the model closely to speech, music, or other domain of sound interpretation."

We hope that by calling the space of computer applications of sound analysis *machine hearing*, following Mont-Reynaud, we will leverage this good name and good direction, and help the field build around a good framework, as Marr did with what we refer to as *machine vision*.

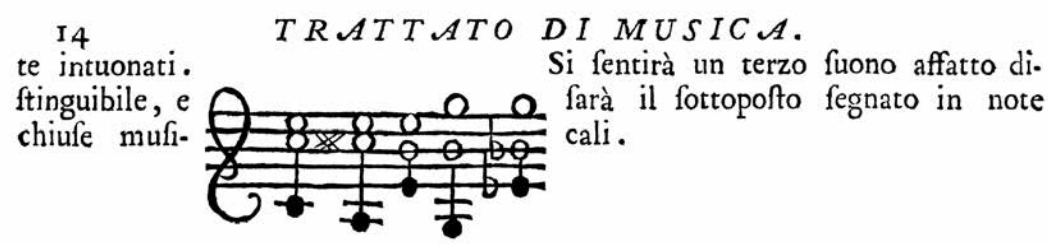

14 TRATTATO DI MUSICA.
te intuonati. Si ſentirà un terzo ſuono affatto di-
ſtinguibile, e ſarà il ſottopoſto ſegnato in note
chiuſe muſi- cali.

Figure 1.4: Tartini's 1754 publication of his observation of *un terzo suono*, a third sound, shown as filled notes below the first two sounds playing on violins or horns—among the earliest recognitions of a nonlinear effect in hearing. The note pitches that Tartini illustrated represent the ratios 4:5:2, 5:6:2, 3:4:2, 5:8:2, and 3:5:2 ($f_1 : f_2 : f_3$, for f_1 being the pitch of the lower played sound and f_2 being the pitch of the upper one, and f_3 being the pitch of the low third tone). The third-tone pitch corresponds to the quadratic intermodulation product $f_2 - f_1$, or the cubic intermodulation product $2f_1 - f_2$, and/or an octave above one of those. As Helmholtz (1863) remarked of these observations, "It is very easy to make a mistake of an octave. This has happened to the most celebrated musicians and acousticians. Thus it is well known that Tartini, who was celebrated as a violinist and theoretical musician, estimated all combinational tones an octave too high." Sorge's 1745 observation of c″ and a″ making an f would be 3:5:1, with *den dritten Klang*, a third-order (cubic) distortion product, at $2f_1 - f_2$.

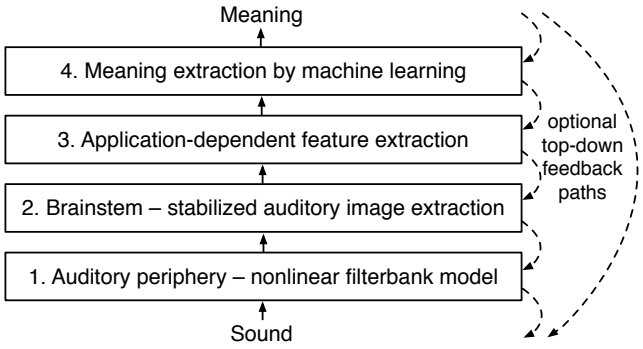


Figure 1.5: The *four-layer model* of machine hearing systems developed in this book—from sound to meaning, and sometimes back the other way. The big feedback loop from meaning to sound is for a system that can make sound and hear itself, for example, a speech conversation system.

Chapter 2

Theories of Hearing

> In respect of the theory of hearing, it seems to me that we need fewer theories and more theorizing. Of theories, focused upon some new finding and seeking to align the entire body of auditory fact with the new principle, we have more than a plenty.
>
> — "Auditory theory with special reference to intensity, volume, and localization," Edwin G. Boring (1926)

> The principle of diversity suggests that a simple description of the auditory process may not be possible because the process may not be simple. Theories that appear at first thought to be alternatives may in fact supplement one another.
>
> — "Place mechanisms of auditory frequency analysis," William H. Huggins and Licklider (1951)

Many theories and models have influenced thinking in this field; here we survey some of these, including those modern theories on which we base machine hearing systems.

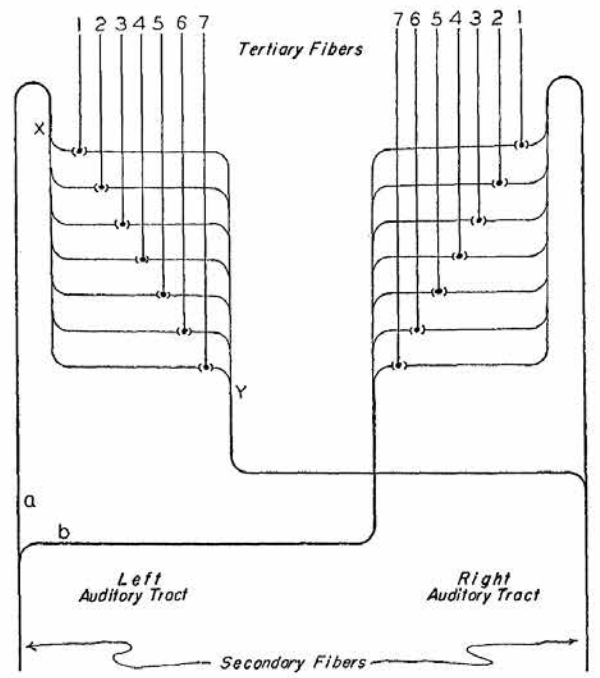


FIGURE 1. HYPOTHETICAL MID-BRAIN MECHANISM FOR THE LOCALIZATION OF LOW FREQUENCY TONES

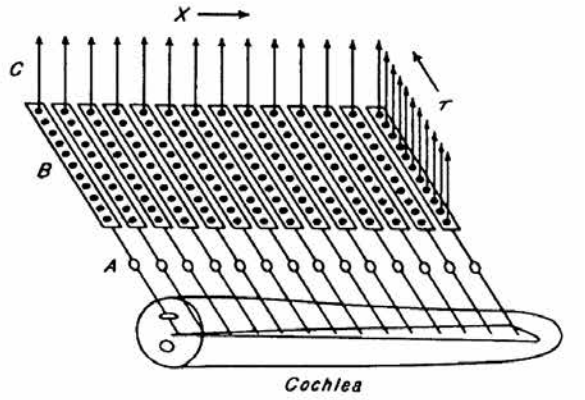


Fig. 2. – *Schematic diagram of overall analyzer.* At the bottom is the uncoiled cochlea. Its lengthwise dimension and the corresponding dimension in the neural tissue above it is designated the x-dimension. The cochlea performs a crude frequency analysis of the stimulus time function, distributing different frequency bands to different x-positions. In the process of exciting the neurons of the auditory nerve, the outputs of the cochlear filters are rectified and smoothed. The resulting signals are carried by the groups of neurons A to the autocorrelators B, whose delay- or τ-dimension is orthogonal to x. The outputs of the autocorrelators are fed to higher centers over the matrix of channels C, a cross-section through which is called the (x, τ)-plane. (Output arrows arise from all the dots; some are omitted in the diagram to avoid confusion.) The time-varying distribution of activity in the (x, τ)-plane provides a progressive analysis of the acoustic stimulus, first in frequency and then in periodicity.

Figure 2.1: Jeffress's (left) and Licklider's (right) drawings of their binaural and pitch models of the neural formation of auditory images (Jeffress, 1948; Licklider, 1951). Coincidence detection between differently delayed neural events, or in Licklider's between delayed and nondelayed events, generates the time-difference dimension of a map. Jeffress does not show a tonotopic axis, but his scheme has generally been interpreted as one frequency slice of a two-dimensional structure like Licklider's (Lyon, 1983; Shackleton et al., 1992; Hartung and Trahiotis, 2001). Jeffress guessed that such a structure might be found in the superior olivary complex—where a mapping of interaural delay was actually found years later. [Reproduced (Jeffress, 1948) with permission of the American Psychological Association; (Licklider, 1951) with permission of Springer.]

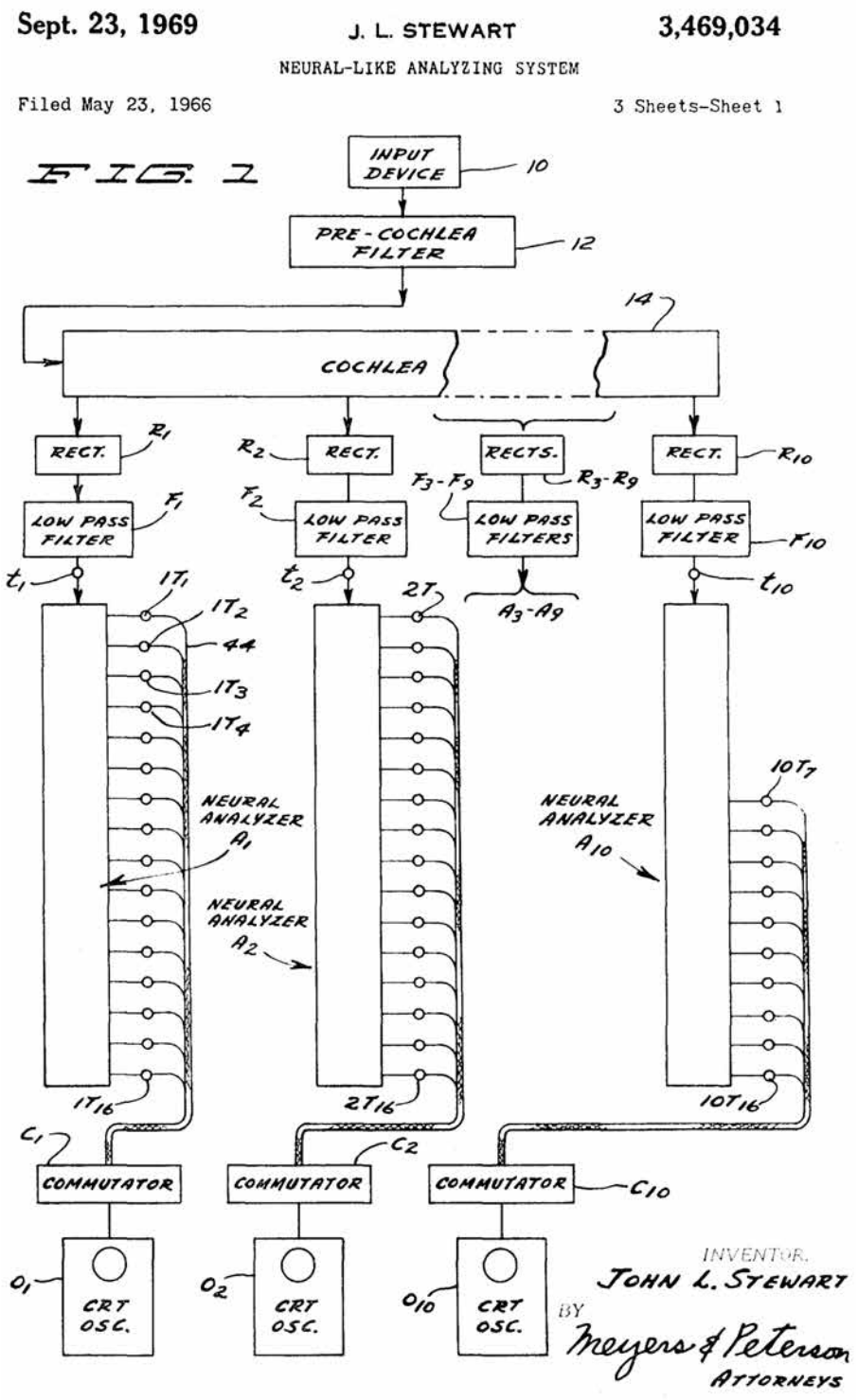


Figure 2.2: A patent drawing from the John L. Stewart (1966) “neural-like analyzer.” The “neural analyzer” stages at each rectified output of the “cochlea” generate a second dimension, mapping the cochlea’s space–time pattern to a space–space pattern, an image-like map, much as in Jeffress’s and Licklider’s theories.

Chapter 3

On Logarithmic and Power-Law Hearing

> The task of clearing the scientific bench top of the century-long preoccupation with the *jnd* [just-noticeable difference], and the consequent belief in logarithmic functions, demands the cleansing power of a superior replacement. My optimism on this score has been recorded in other places, but I would like here to suggest that, if I seem to feel a measure of enthusiasm for the power law relating sensation magnitude to stimulus intensity, it is only because that law seems to me to exhibit some highly desirable features.
>
> — Stevens (1961), "To honor Fechner and repeal his law: a power function, not a log function, describes the operating characteristic of a sensory system"

The Mathematics of Logarithms and Power Laws

The algebraic definition of logarithm leads to several useful relationships. Given a value x, and a base b, the base-b logarithm of the value x is the number y that satisfies the equation:

$$x = b^y$$

The logarithm is essentially a functional inverse of this exponentiation operation. In terms of the logarithm function, we write the "solution" of the above equation as:

$$y = log_b(x)$$

That is, exponentiation maps y to x, and the logarithm function maps x to y, as long as both of them use the same base b. Any positive number other than 1 will work for b, but special numbers like 2 for *binary* logarithmss, e for *natural* logarithms, and 10 for so-called *common* logarithms are most often encountered as bases. The value e is the unique number (about 2.71828) such that the exponential curve e^x has unit slope at $x = 0$ (more generally, $\frac{d}{dx}e^x = e^x$, for this and no other value of e).

Properties of logarithms and different bases are easy to derive from the properties of exponents.

A power law looks similar, but the variables are not in the exponents. Here we base the formulas on an exponent parameter α, usually between 0 and 1, instead of the integer power N and its reciprocal $1/N$ mentioned earlier:

$$y = x^\alpha$$

$$x = y^{1/\alpha}$$

As α approaches 1, we approach the identity relationship between x and y. The other extreme, as α approaches 0, is more interesting, but we'll need to rewrite the relations in a way that makes the power law functions converge to a consistent mapping in that limit. Let's scale and offset x and y to pick the case of converging on the identity function near the point (1, 1)—that is, such that all functions pass through the point (1, 1) with unit slope—while keeping the point of infinite slope at $x = 0$:

$$y = \left(x^\alpha - 1\right)/\alpha + 1$$

$$x = \left(\alpha y - \alpha + 1\right)^{1/\alpha}$$

In the limit of small α, these modified power-law functions approach exponential/logarithm relationships that have been similarly shifted to be tangent to the identity function at (1, 1), as illustrated in Figure 3.1:

$$y = \log_e(x) + 1$$

$$x = \exp(y - 1)$$

In this sense, the power-law functions are good intermediate mappings for many purposes—not linear, but not as extreme as logarithms and exponentials.

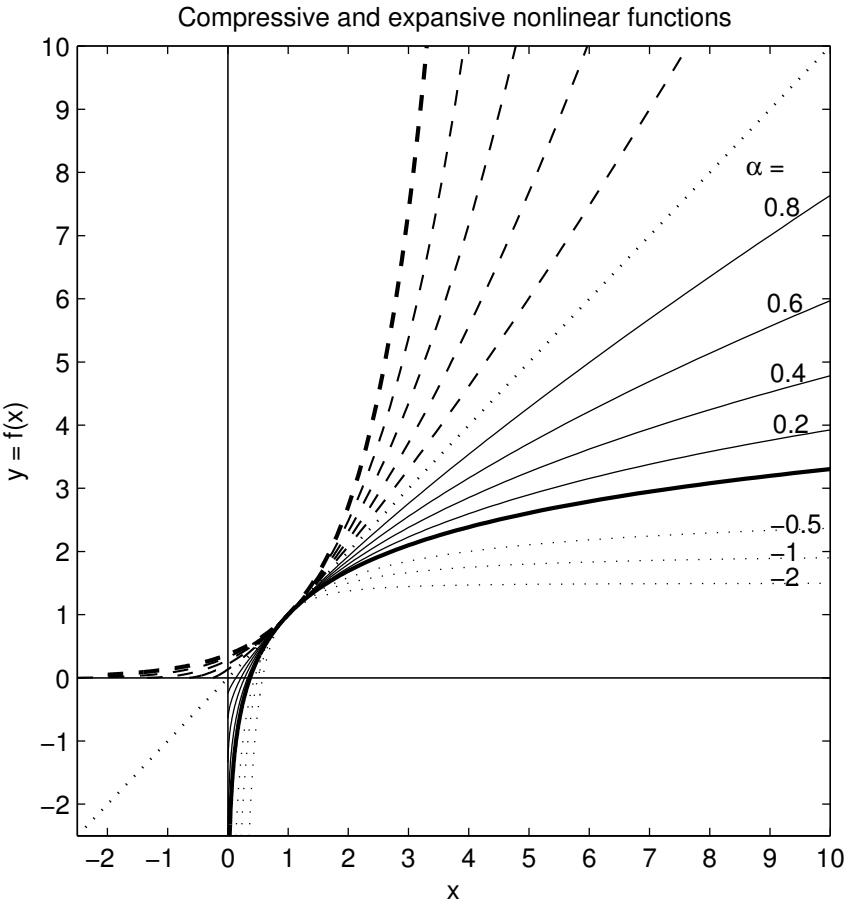


Figure 3.1: Some compressive nonlinearities (solid curves) and the expansive nonlinearities (dashed curves) that are their inverses (compressive means with "diminishing return," that is, slope decreasing as input increases, while expansive means the opposite). The heavy solid curve is a logarithmic compression, and the heavy dashed curve is an exponential expansion; the lighter curves are based on power-law relationships, with exponents $0 < \alpha < 1$ as annotated. As explained in the text, the functions are all adjusted to be tangent to the identity function (dotted) at the point (1, 1), such that the α exponent interpolates the compressive functions between log and linear, for $x > 0$. Also shown (dotted curves) are the even more compressive functions that result from negative exponents—linear tranformations of the reciprocal square root, reciprocal, and reciprocal square. Tukey (1957) discussed the logarithm as the natural limit between these curves with positive and negative exponents.

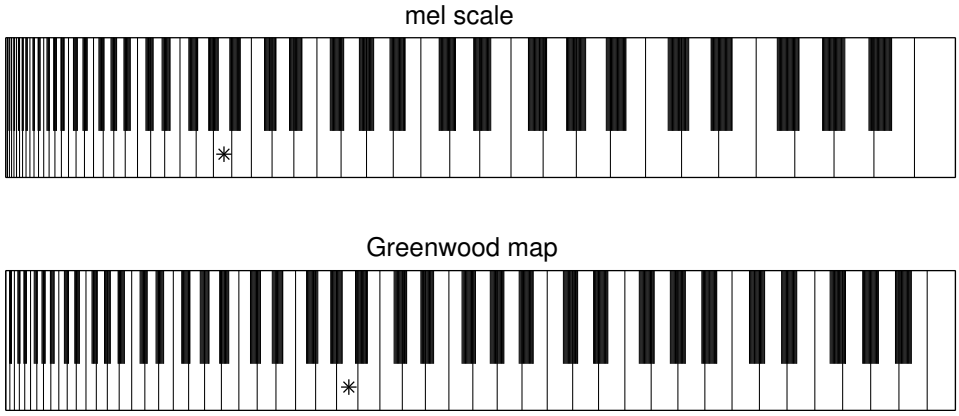


Figure 3.2: A normal piano keyboard has a logarithmic mapping of frequency to place, but these distorted ones have auditory mappings. The top keyboard is based on the mel scale (see Section **??**), which severely squashes the low frequencies. The bottom one is based on the Greenwood map, a more accurate reflection of auditory physiology and psychophysics. In both, the x coordinates of the keys are *stabilized logarithms* of the corresponding note frequencies, with different stabilizing offsets. That is, positions on the distorted keyboard are linear functions of $\log(f + f_{\text{break}})$ for a stabilizing offset f_{break} that we refer to as the break frequency—the approximate breakpoint between a low-frequency linear limit and a high-frequency logarithmic limit. The keys marked "*" are A440, a 440 Hz pitch that is on the linear side of the mel scale's 700 Hz break frequency, but on the logarithmic side of the Greenwood map's 165 Hz break frequency.

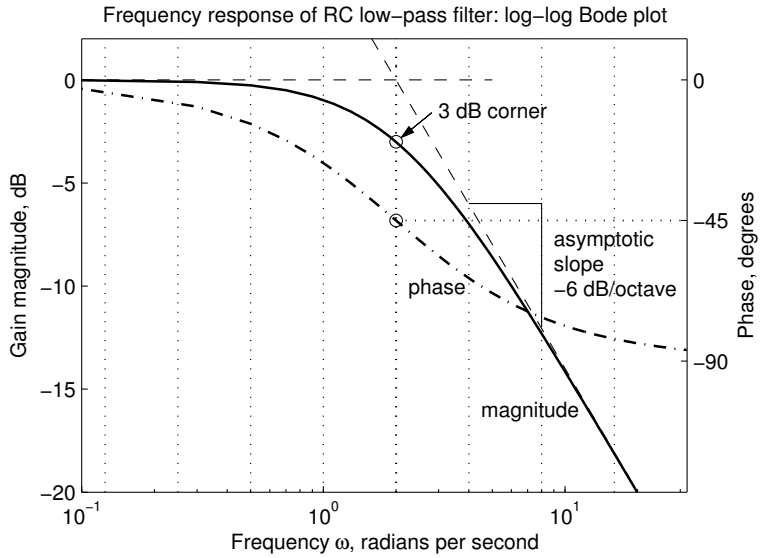


Figure 3.3: Bode plot, or log–log frequency response, of a simple lowpass filter. This is a typical way in which engineers combine logarithmic frequency and amplitude scales to characterize a filter, usually resulting in straight-line asymptotes, such as shown (dashed). The phase response of the filter is also shown (dash-dot). Phase can also be thought of as a logarithmic scale, the imaginary part of the complex logarithm of the filter's transfer function.

Complex Numbers and Euler's Formula

The reader will need basic familiarity with complex numbers (numbers such as $3 + 2i$ that have an imaginary part using the pure imaginary number i, defined by $i^2 = -1$) and with $e = 2.71828\ldots$, the base of the natural logarithms, for any kind of work with sound, waves, hearing, or linear systems. It will also be important to understand some basic properties of the exponential function, $\exp(x) = e^x$, with complex argument. In particular, one needs to be familiar with *Euler's formula* (or Euler's identity, not to be confused with Euler's polyhedron formula or the numerous other mathematical concepts named for Leonard Euler):

$$\exp(i\theta) = \cos\theta + i\sin\theta$$

The formula says that the exponential of a pure-imaginary number, $\exp(i\theta)$, is a point on the unit circle in the complex plane (that is, having an absolute value of 1, or at Euclidean distance 1 from the origin), at an angle from the real axis equal to the value θ in the exponent. Here, angles are measured in natural units, known as radians, equivalent to arc length on the unit circle, counterclockwise from the real axis in the real–imaginary plane; see Figure 3.4. Since the arc length around a complete cycle of the unit circle is 2π, that factor shows up frequently, for example in converting between cycles and radians.

Euler's formula is key to how we express oscillations, or tones. Its oft-quoted amusing special case $e^{i\pi} = -1$ obscures its importance.

The angle θ often represents a phase increasing with time, in which case the derivative $d\theta/dt$ is the frequency, in radians per second, and the exponential represents a rotating complex function of time—an important complex generalization of the *sine wave*.

Notice the notation for the exponential function: we prefer to treat e^x as a function, at least as important conceptually as things like $\cos(x)$ and $\log(x)$, rather than emphasize the number e.

It would work almost as well to use other numbers, such as 10, instead of e—with $\exp_{10}$ as the inverse of $\log_{10}$—representing oscillating signals as $10^{i\theta}$. But that would introduce unnatural conversion factors: rather than the 2π units of phase per cycle that we get with $e^{i\theta}$, we would have about 2.729 units of phase per cycle, and that number, $2\pi/\log_e(10)$, couldn't be expressed exactly without invoking e, or natural logarithms. So *naturally,* we use e, and the natural phase unit, radians, and the corresponding natural definitions of the sine and cosine functions, and frequency in the natural units of radians per second.

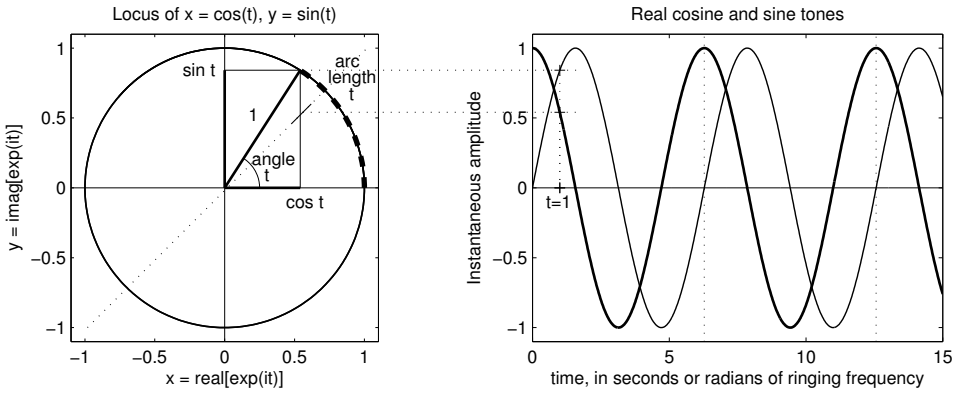


Figure 3.4: Applying Euler's formula to tones. On the left, the points on the unit circle represent pairs $(\cos t, \sin t)$ with parameter t; by Euler's formula, each such point can be interpreted as a complex-plane plot of the value $\exp(it)$. The marked example point and coordinates represent $t = 1$, corresponding to an arc length of 1 and an angle of 1 radian, as measured from the $\exp(0) = 1$ point on the positive real axis. On the right, the cosine function $\cos(t)$ (dark curve) and sine function $\sin(t)$ (light curve) are the x and y coordinates of points on that unit circle, plotted as functions of the parameter t (the dotted diagonal on the left reflects the x coordinate to make the $\cos(t)$ ordinate on the right).

Complex Logarithm History

Before Euler articulated the formula that bears his name, circa 1740, Roger Cotes had already observed the logarithmic form of it in 1714 (Stillwell, 2010):

$$\log(\cos\theta + i\sin\theta) = i\theta$$

where by log he meant the natural logarithm, base e. In so doing, he invented the value e, and was among the inventors of the natural logarithm.

But Cotes's version has a problem that was not appreciated at the time of this first attempt to extend logarithms to the domain of complex numbers: in order to treat the complex logarithm as a function, we need to define a principal value, by picking one *branch*. There are multiple distinct values of θ that give the same results for $\cos\theta$ and $\sin\theta$, and hence for the log, so the equation can't be true for more than one of them. For the others, $i\theta$ is not within the range of the chosen branch. We can extend Cotes's formula to say

$$\log(\cos\theta + i\sin\theta) = i(\theta + n2\pi)$$

for some n that depends on θ, chosen to bring the result into the range of principal values (typically defined as the range where the imaginary part, the angle in radians of the argument of the log function in the complex plane, is greater than $-\pi$ and less than or equal to π).

Euler used the exponential function, the inverse of the logarithm, to sidestep this complication and give a simple and always correct equation. The interpretation involving points on a circle in the Cartesian complex plane was not known until it was discovered and published by Caspar Wessel in 1797 and by Jean-Robert Argand in 1806 (Wessel's article wasn't translated from Danish into French until 1899, so it didn't have much impact) (Fine, 1903). The Cartesian complex plane as a mechanism for visualizing complex numbers geometrically is sometimes called the *Argand plane*.

Electrical engineers know that the breakthrough in analysis of AC circuits, which led to widespread electrical power generation and distribution, was Charles Steinmetz's application of complex numbers to circuit analysis (Steinmetz, 1893); complex logarithms are key to analyzing propagation of telephone and telegraph signals over long wires. This application of Euler's formula also contributed to the techniques of filter design and analysis that led to a revolution in multiplexed wired and wireless telephone and telegraph communications, across continents, and across oceans. These steps have been part of what has so accelerated progress in the technical arts in the last century.

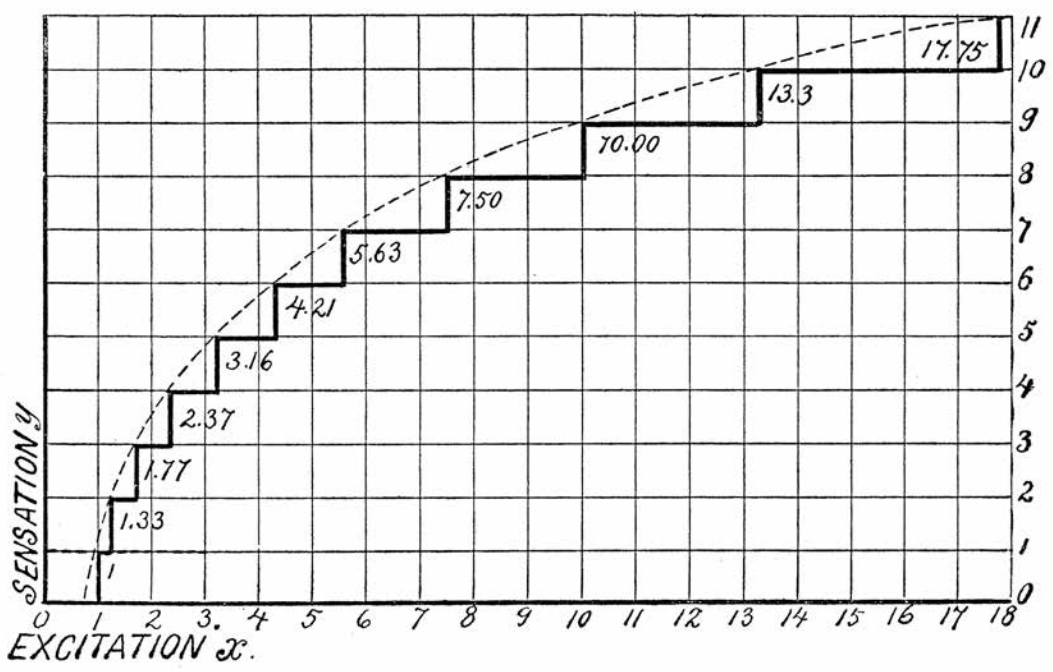


Figure 3.5: The Weber–Fechner psychophysical law: intensities in geometric progression evoke sensations in arithmetic progression; plot from Howell (1915). It is apparent that an intensity (excitation) approaching zero cannot be accommodated in this scheme, and that the zero point of sensation is arbitrary. Logarithmic scales always have such problems.

Chapter 4

Human Hearing Overview

> On the zigzagging road towards wisdom about the human auditory system we collect knowledge from two entirely different sources of experimental information. First, from anatomy and physiology … Secondly, from perception and psychoacoustics … Our ever-wondering mind tries to combine and to explain these findings in terms of some model, law, hypothesis or theory.
>
> — "The residue revisited," J. F. Schouten (1970)

GAFORUS. *Theorica musicae.* Mediolani 1492.

Figure 4.1: The observation by Pythagoras that small integer ratios of string lengths or tensions, or of reed flute lengths, led to consonant notes, and that bells and glasses of water could be tuned to corresponding pitch ratios, was celebrated in this woodcut from Franchino Gafurio's *Theorica Musice* of 1492. Stillingfleet (1771) points out that the traditional story about Pythagoras and the hammer weights that he used to tension the strings is incorrect, since the string tensions would have to be set in the squares of those ratios to get the consonant tone intervals described.

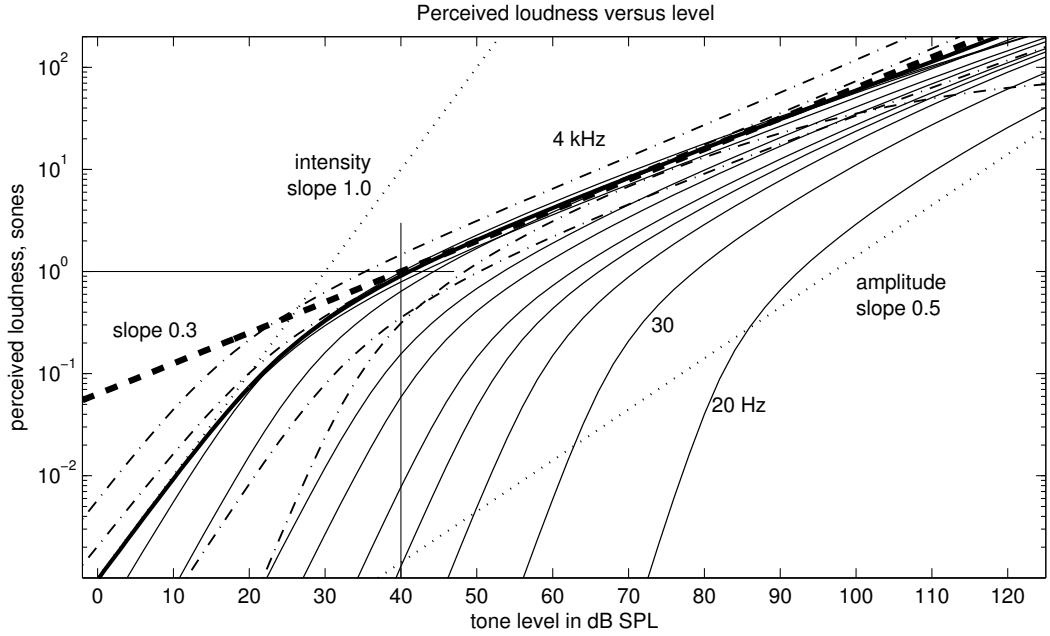


Figure 4.2: Perceived loudness in *sones*, as functions of sound intensity, for selected tone frequencies. The heavy diagonal dashed line (slope 0.3) is based on the conventional approximate definition that loudness in sones is proportional to the 0.3 power of intensity, with 1 sone at 40 dB SPL, at 1 kHz. As the upper right portion of the figure shows, this is a fair approximation at high enough intensities and high enough frequencies. The other curves are based on a *stabilized power law*, using a stabilizing offset corresponding at each frequency to the power level at the 20 phon curve of Figure 4.3; a power-law exponent of 0.28 is used. The solid curves are for frequencies up to 1 kHz (20, 30, 40, 50, 60, 80, 100, 200, 400, and 1000 Hz), with the 1000 Hz curve heavier; the dash-dot curves are for 2, 4, 8, and 15 kHz. The dotted line at the far left (labeled "intensity") shows the slope that would correspond to a linear relationship between intensity and loudness; this slope is approached at low intensities. The dotted line at the far right (labeled "amplitude") shows the slope that would correspond to a linear relationship between pressure amplitude and loudness (intensity power-law exponent 0.5); very low frequencies come close to this slope at high levels.

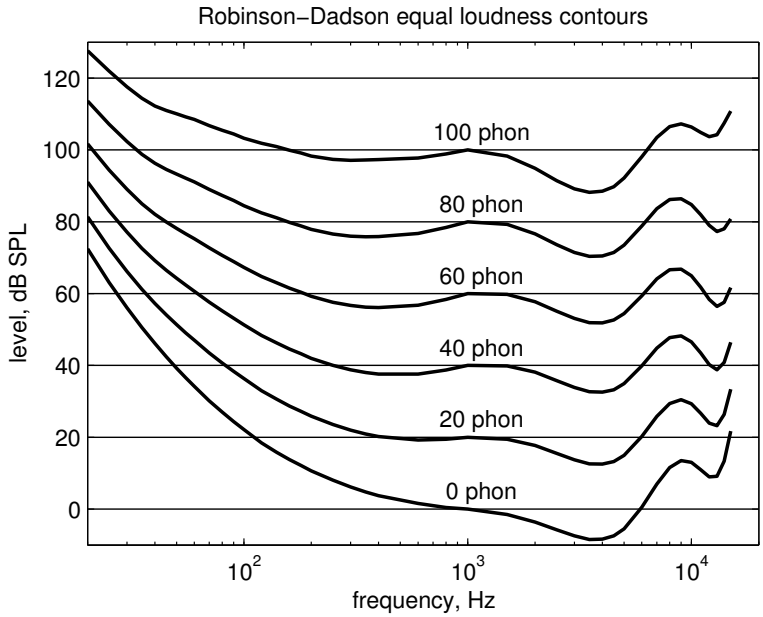


Figure 4.3: The equal-loudness contours known as Robinson–Dadson curves, plotted from their original model parameters (Robinson and Dadson, 1956), map intensity in dB SPL to a loudness-related log-like measure, *phons*. The dB SPL scale is an objective intensity scale, relative to the nominal 1 kHz threshold sound pressure of 20 μpascal RMS, and the phon scale is defined to be equal to dB SPL for a 1 kHz sine wave, but to connect other frequencies and intensities of the same perceptual loudness. The equal loudness contours do not show how the percept of loudness grows with level—that is the function of the *sone* scale.

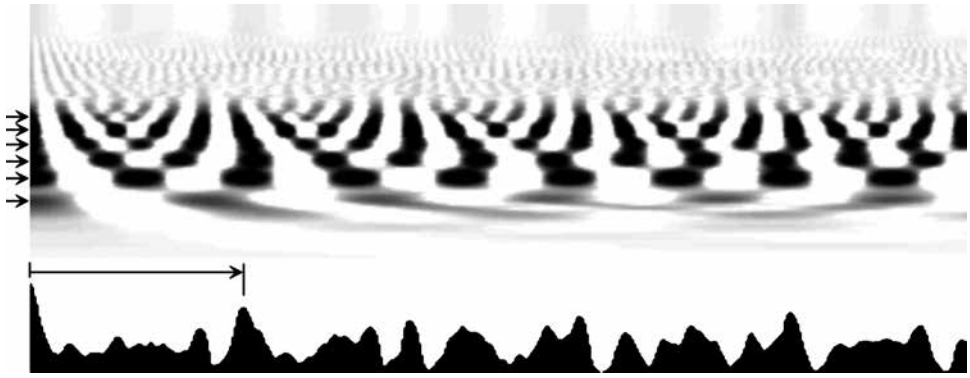


Figure 4.4: One frame of a Licklider-style auditory image of the orchestral chime sound from the Acoustical Society of America's *Auditory Demonstrations* CD (Houtsma et al., 1987), showing several strong partials (nearly-sinusoidal components, from the resonant "modes" of the bell) that contribute to a strong peak at the delay corresponding to the perceived strike note pitch period. The upper part of the figure is the auditory image, or correlogram, with vertical axis corresponding to cochlear place, or frequency, and horizontal axis representing the delay or lag parameter, measured from the left edge; see Chapter 21 for details on such images. The graph at the bottom shows the sum across frequency channels; peaks in this graph are likely pitch periods; the period corresponding to the strike note pitch is indicated by an arrow. The arrows on the left indicate the positions of strong partials along the frequency axis. Several strong partials are close to a 2:3:4:5 relationship, but not quite harmonic.

Weber's and Fechner's and Stevens's Laws and Loudness JND

The threshold for detection of loudness differences, the *just-noticeable difference* (jnd) of loudness, is often expressed as the *Weber fraction* of intensity: $\Delta I/I$. Weber's law says that this fraction is constant, across a wide range of intensity. This law is related to Fechner's law, that sensation scales as the log of intensity: they become equivalent on the assumption that the jnd is an equal increment of sensation level at all intensities.

Experiments find a "near miss to Weber's law," in which the Weber fraction decreases somewhat with increasing level, but not as quickly as the Stevens power law would suggest. For sine waves of mid frequencies, the Weber fraction decreases from about 30% near 20 dB SPL to about 10% near 90 dB SPL. For broadband sounds such as white noise, the fraction is even closer to being level independent. At very low levels, near the threshold of hearing, of course it must be much larger. If we were to predict a Weber fraction on the assumption that sensation scales with a Stevens power law, with exponent 0.3, and assuming a jnd is a given sensation increment, we would have a much further miss from Weber's law. Taken together, these observations suggest that the idea that a jnd is a constant increment of sensation is not plausible. It would be more accurate to say that the jnd corresponds to a constant Weber fraction of sensation: $\Delta S/S$. At least, it would be a near miss, with jnd always around 3–10% loudness change, across a wide range of loudness.

Krueger (1989) proposes this compromise:

> Fechner and Stevens erred equally about the true psychophysical power function, whose exponent lies halfway between that of Fechner (an exponent approaching zero) and that of Stevens. To be reconciled, Fechnerians must give up the assumptions that Weber's law is valid and that the jnd has the same subjective magnitude across modalities and conditions; Stevensians must give up the assumption that the unadjusted (for the use of number) magnitude scale is a direct measure of subjective magnitude.

Loudness versus intensity, and the jnd of intensity, are complex aspects of loudness perception. More interesting, perhaps, is the question of how loudnesses combine when multiple sound signals are added. For signals that are similar enough, adding them is equivalent to a change of intensity, and the loudness follows its usual intensity pattern. But if the signals are different, in that they have different frequency content, such as different sine-wave frequencies or different noise frequency bands, then the loudness increases faster than would be predicted by just considering the intensity, or acoustic power, of the combination—which brings us to critical bands ...

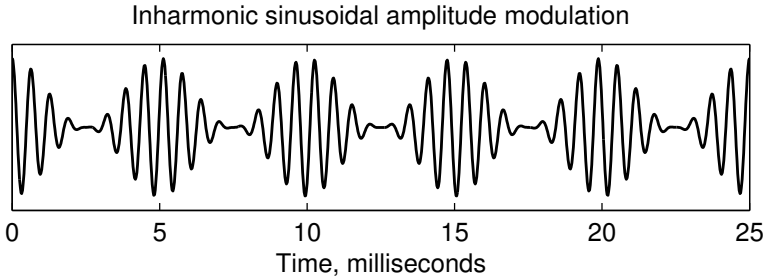


Figure 4.5: A three-component tone complex can be made by modulating a carrier with a slower envelope (here a 1560 Hz carrier and 200 Hz modulator). When the carrier and modulator frequencies are not harmonically related, the resulting signal is not quite periodic. The perceived pitch is usually close to the modulator frequency, but is pulled toward the time interval between a pair of carrier waveform peaks; which interpeak interval is used can sometimes be ambiguous, and pitch matches are sometimes ambiguous. Here the carrier is close to 8 times the modulator, so the interval between corresponding peaks is about 8 cycles of 1560 Hz: $8/1560$ s. According to the *first effect of pitch shift*, the perceived pitch is close to the reciprocal of this interval: $1560/8 = 195$ Hz; but according to the *second effect of pitch shift*, a subject will match it closer to a pitch that is the 7th subharmonic of the lower sideband tone frequency: $(1560 - 200)/7 = 194.3$ Hz. This fractional-hertz difference is enough to show up as a significant effect in pitch-matching experiments.

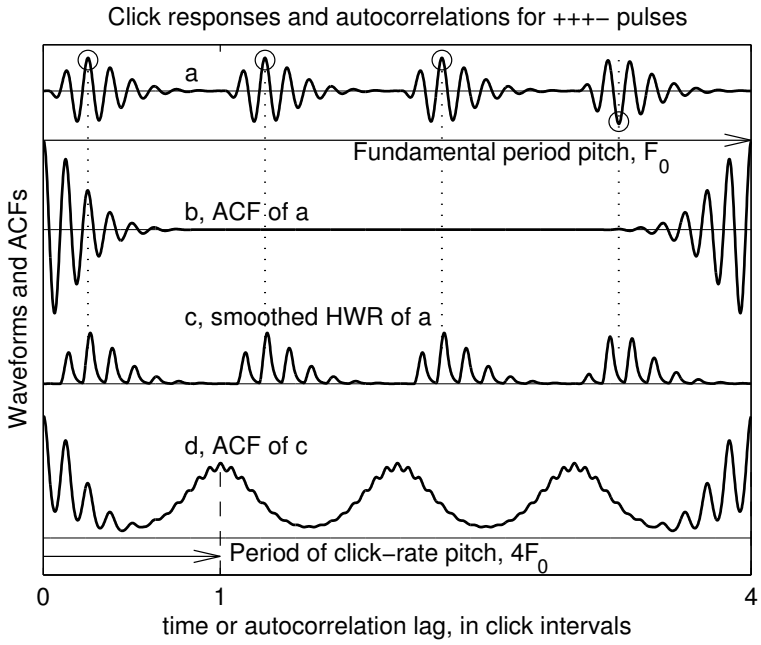


Figure 4.6: The effect of half-wave rectification (HWR) on the autocorrelation function (ACF) of the response to a click train with polarity sequence "+++–": (a) the bandpass-filtered click train; the clicks may be delivered as a sound that looks like this waveform, or as a broadband click train to which a point in the cochlea responds this way; the polarities can be seen by comparing the waveform at the times of the equally-spaced dotted lines. (b) the ACF of (a), showing maxima at zero and at the period, 4 interclick intervals. (c) the result of HWR and smoothing of (a). (d) the ACF of (c). One period of four clicks is shown, since the signals and ACFs repeat with this period. Depending on the parameters, such as the bandpass center frequency and click rate, the perceived pitch may correspond to either the interclick interval (dashed line at lag 1), or the period (at lag 4). The HWR is needed to get the ACF to explain the pitch perception at the interclick interval, since the ACF (b) has no peak at that period.

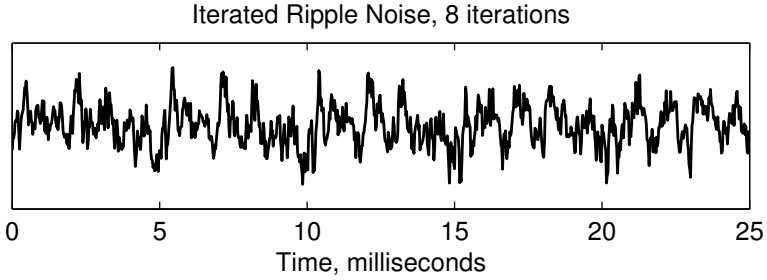


Figure 4.7: Iterated ripple noise (IRN) is another special stimulus signal that has been extensively studied. A noise signal is added to a delayed version of itself, giving a "ripple noise" with a pitch sensation determined by the delay. The signal illustrated here went through 8 iterations of delay and add, with a 5 ms delay; waveform features can be seen repeating with 5 ms separation even though the signal is still a random noise.

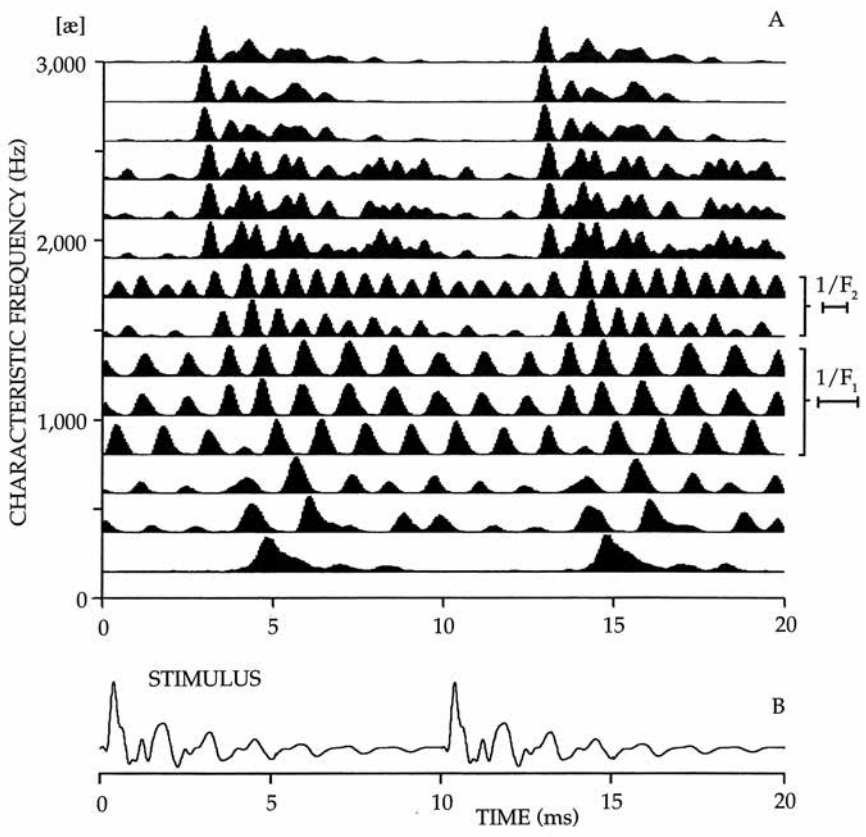
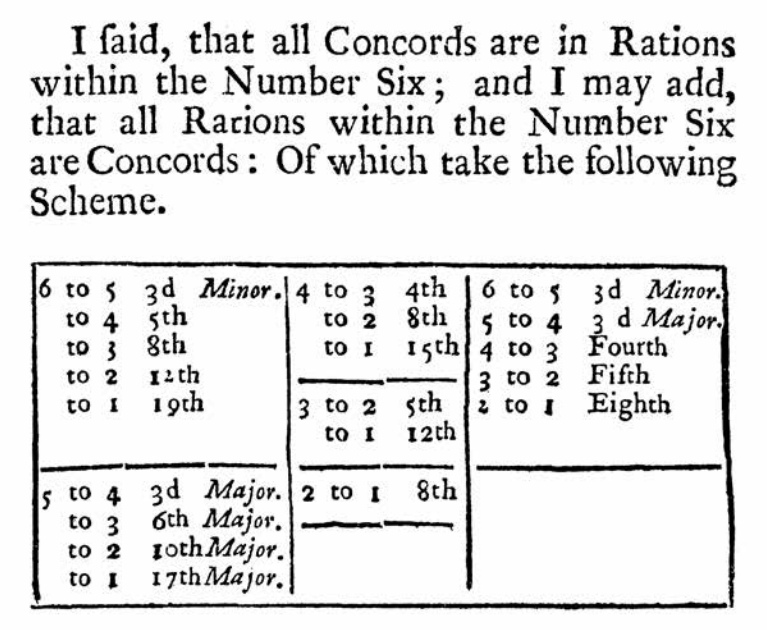


Figure 4.8: Period histograms of auditory nerve fiber firings in response to a periodic vowel sound show pitch-synchronized activity, for fibers of all CFs (Delgutte, 1997). Even fibers that primarily synchronize to the formant (vocal tract resonance) frequencies (here F_1, 8 cycles per pitch period, and F_2, 14 cycles per pitch period) show a pattern that repeats at the pitch rate. Synchrony to the formant frequencies spreads to fibers of higher CF. Fibers with CF above 2 kHz show synchrony to a wide range of lower frequencies, in a pattern prominently synchronized to the pitch rate. The pitch here, 100 Hz, is quite low relative to the cat's auditory-system tuning, so we do not see the resolved low harmonics (2 through 5 cycles per pitch period) that would likely be apparent in human auditory nerve data. [Figure 3 of Chapter 16 (Delgutte, 1997) reproduced with permission of John Wiley & Sons.]



I ſaid, that all Concords are in Rations within the Number Six; and I may add, that all Rations within the Number Six are Concords: Of which take the following Scheme.

6 to 5 3d *Minor.* to 4 5th to 3 8th to 2 12th to 1 19th	4 to 3 4th to 2 8th to 1 15th	6 to 5 3d *Minor.* 5 to 4 3 d *Major.* 4 to 3 Fourth 3 to 2 Fifth 2 to 1 Eighth
5 to 4 3d *Major.* to 3 6th *Major.* to 2 10th *Major.* to 1 17th *Major.*	3 to 2 5th to 1 12th 2 to 1 8th	

Figure 4.9: William Holder explained that “From the Premises, it will be easie to comprehend the natural Reason, why the Ear is delighted with those forenamed Concords ; and that is, because they all unite in the Motions often, and at the least at every sixth Course of Vibration, which appears from the Rations by which they are constituted, which are all contained with that Number, and all Rations contained within that Space of Six, make Concords, because the Mixture of their Motions is answerable to the Ration of them, and are made at or before every Sixth Course. First, how and why the Unisons agree so perfectly ; and then finding the Reason of an Octave, and fixing that, all the rest will follow.” (Holder, 1731).

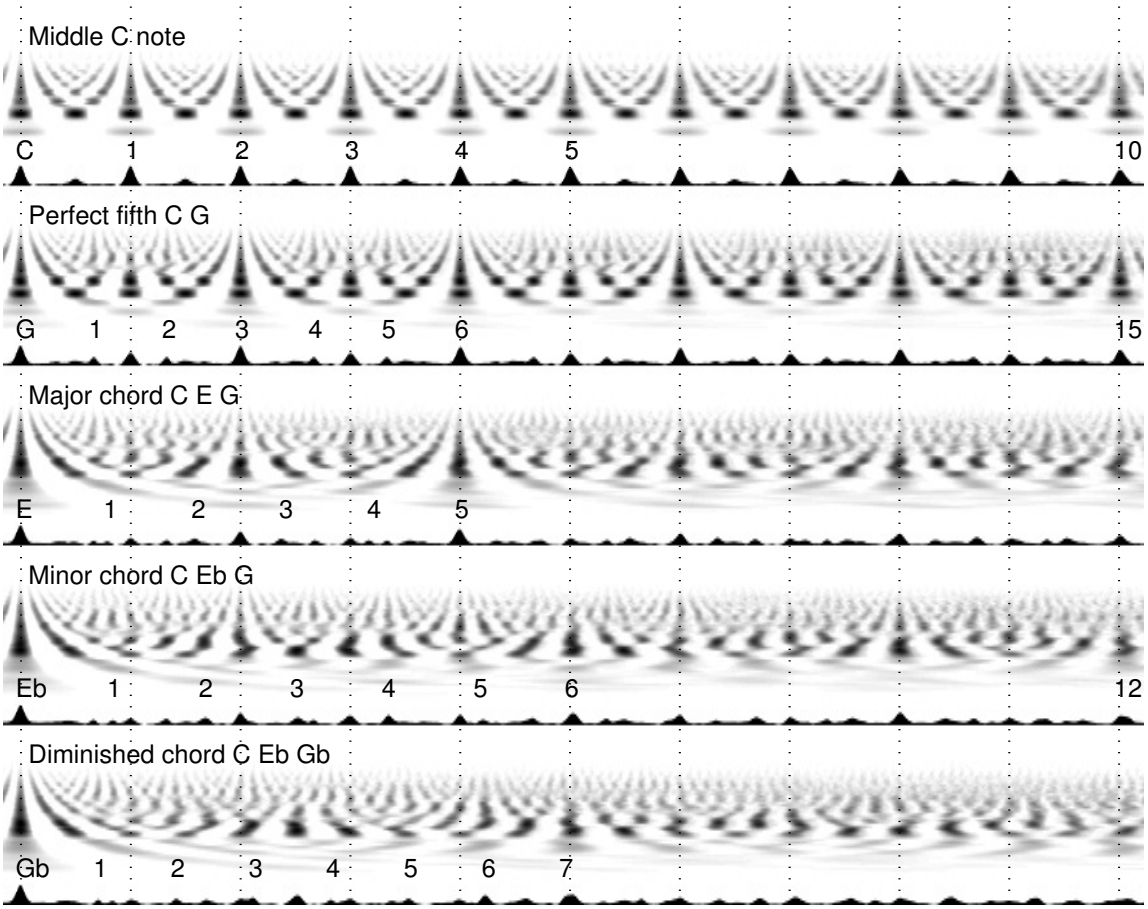


Figure 4.10: The stabilized auditory image (SAI) displays a steady musical tone or chord as a steady image. Each stripe of this figure is an SAI of a musical tone or chord, starting with the middle-C (260 Hz) note of a bassoon played alone (top stripe), followed in the subsequent stripes by four combinations of that C tone with bassoon notes of other pitches, to visualize how harmonicity, consonance, and dissonance appear in this representation. Below each stripe of SAI is a *summary SAI*, a graph of the sum of the SAI over frequency channels, peaks of which correspond to likely perceived pitch periods, including induced root pitches. Between the SAI and the summary SAI, the most recently introduced note is named and its periods along the time lag axis are labeled. The patterns get increasingly complex, from the simple structure of the C note, to the pattern with twice the period when the G is added (3 periods of G aligning with two periods of C, a perfect fifth, making a root pitch of C, down an octave), to four times the period for the major chord CEG (pitches near 4:5:6, root pitch down another octave, at four cycles of C), to a more complex pattern of partial alignments for the minor chord (pitches near 10:12:15), and finally the relative disorder of the more dissonant diminished chord (pitches near 20:24:29 or somewhat near 5:6:7). According to Holder, it is only the alignments out to about 6 cycles that make “concords.”

Chapter 5

Acoustic Approaches and Auditory Influence

> Machines which automatically recognize patterns from a stream of acoustic events, for example a spoken command, would have great utility in both communications and data processing. This paper reviews two applications of an elementary recognizer to the problem of actuating certain logical functions, and indicates how more ambitious recognizers might be utilized. In this regard, the automatic measurement of a talker's voice pitch and voicing dynamics appears fundamental to speech analysis, and hence to many recognition schemes. Visual inspection of spectral data taken from different speakers supports this contention.
>
> — "Artificial auditory recognition in telephony," E. E. David (1958)

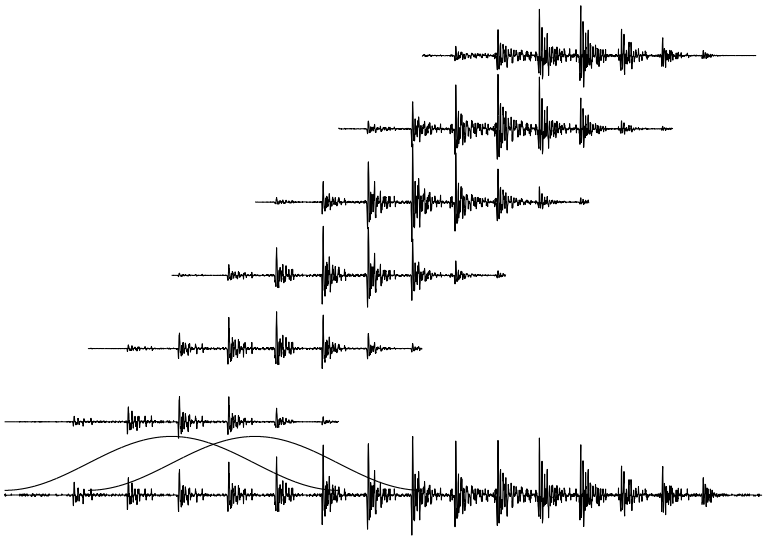


Figure 5.1: Short-time analysis is typically done on windowed segments of a sound waveform. This diagram illustrates a bell-shaped window function, placed in two different positions separated by a "hop," on a waveform of speech (of the word "I" by a low-pitch male voice). Multiplying the window, point-by-point, by the original waveform generates a *windowed segment* at each position; six of them are shown above the positions that they came from. In this example, the window is a *Hamming window* (a *raised cosine*) of 80 ms duration, and the hop size is 20 ms.

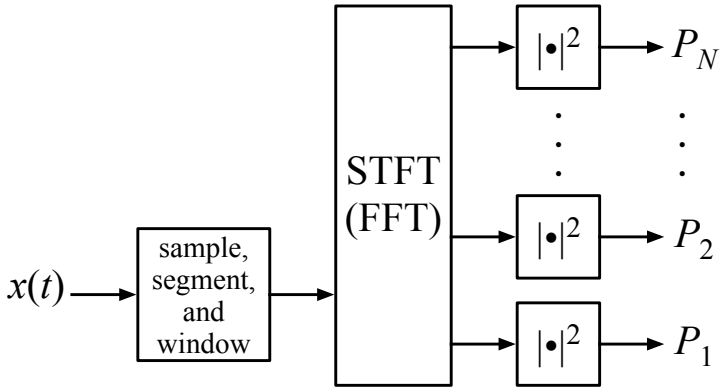


Figure 5.2: The short-time spectrum of a signal $x(t)$ can be estimated by a short-time Fourier transform (STFT, typically using the fast Fourier transform algorithm, FFT). The center frequencies are in arithmetic progression, and the bandwidths all equal, unless a further stage of channel combining is added. A segment length and window function establish the time scale of the STFT. The power outputs from such an analysis are typically next compressed through a logarithmic or power-law function.

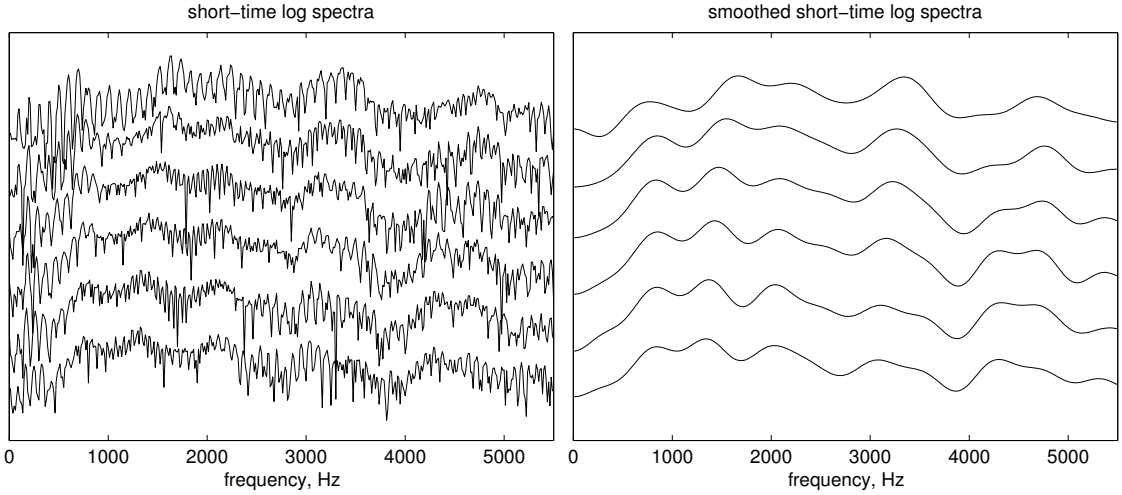


Figure 5.3: Short-time log spectra (starting from zero frequency on the left) from an FFT analysis of the 80 ms Hamming windowed segments reveal too much irrelevant detail (left). The logarithm compresses the dynamic range, but puts too much emphasis on low values (making the down-going spikes) and flattens high values. These spectra correspond, from bottom to top, with the segments shown in Figure 5.1. On the right, each spectrum is smoothed to remove "ripples," or details not very relevant to the acoustic source. Such operations in the log-spectrum domain are explained in Section **??**.

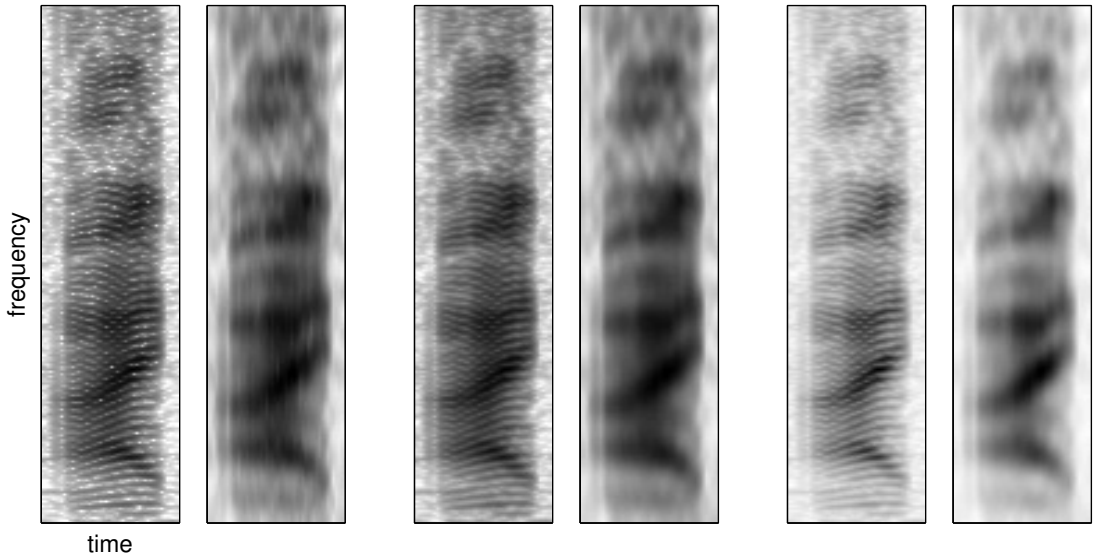


Figure 5.4: Three pairs of spectrograms and frequency-smoothed spectrograms of the word "I" are shown, using a conventional light-to-dark scale (white for silence, dark where there is energy). Each spectrogram is about 0.25 s along the time axis, and the frequency range axis is linear from 0 to 5000 Hz. In the left pair, a log spectrogram with dynamic range from white to black of 60 dB is shown. White specks where estimated power is near zero (like the downward spikes in Figure 5.3) persist as light streaks and raggedness after smoothing across the frequency dimension. In the middle pair, a very slight smoothing across frequencies is applied in the power spectrum domain before the logarithm, which eliminates most of that anomaly by keeping values away from zero. This pre-logarithm smoothing is done with a [0.1, 0.8, 0.1] filter (each spectral energy value is replaced by 80% of itself plus 10% of each of its lower- and higher-frequency neighbors). In the right pair, a power law with exponent 0.15 is used instead of the logarithm, but with the same pre-compression smoothing as in the center, resulting in more contrast in the high-power areas, and less in the low-power areas. The results with the slight pre-compression smoothing (center and right pairs) are much cleaner, with the few very-near-zero values having been fixed before allowing the nonlinearity to spike downward there. The white specks are gone, and the noise or raggedness that these specks add in the smoothed spectrum is removed. Spectrograms and smoothed spectrograms like those on the left are commonly seen in publications, and are a clear indication that the effect of the logarithm was not carefully considered or dealt with. The rightmost spectrograms are most “meaningful” in the sense that they give a better visualization of the aspects of the sound spectrum that encode the spoken word, and as a result may also be better as input to a machine learning system.

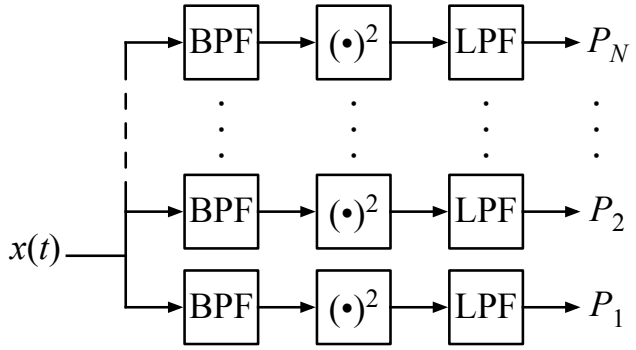


Figure 5.5: The short-time spectrum of a signal $x(t)$ can be computed by a bank of bandpass filters (BPF), followed by squaring to detect instantaneous power, and finally lowpass filters (LPF) that establish the smoothing time scale. An N-point spectrum is made via N different BPFs with different center frequencies, and often different bandwidths, and N usually identical LPFs.

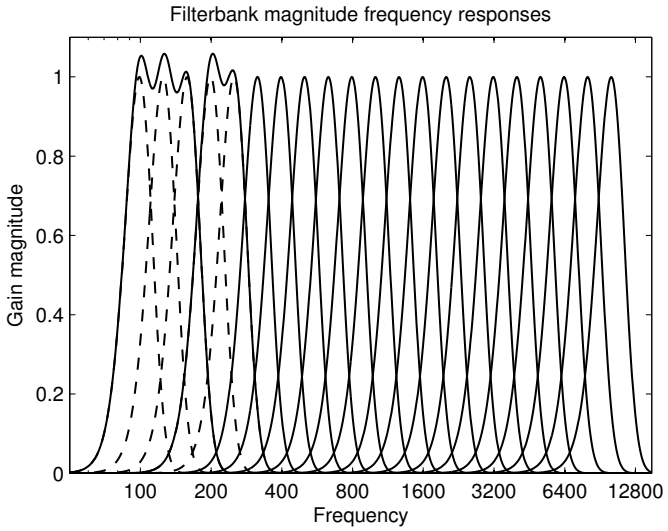


Figure 5.6: A constant-Q filterbank, here represented by 21 similar Gaussian-shaped frequency response gain curves, can give rise to an approximately auditory-scale filterbank analysis by combining some of the low channels to make an 18-dimensional spectrum vector, as Plomp et al. (1967) did; channels that are combined are shown dashed, and their sums solid. Conversely, an FFT-based analyzer, with channel responses of equal width on a linear frequency scale, can give rise to an auditory spectrum by combining increasing numbers of channels at the high-frequency end.

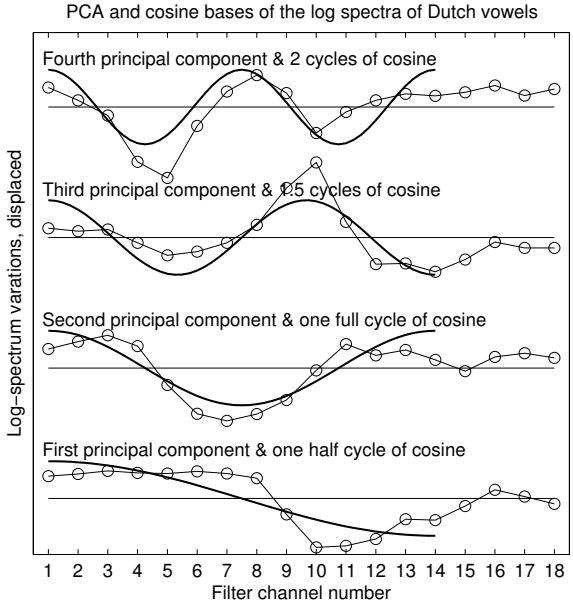


Figure 5.7: The first four principal components of speech log spectra found by Plomp, Pols, and van de Geer (1967), light curves with circles, and the cosine-transform basis functions that approximate them, heavy curves. The filter channels are approximately as shown in Figure 5.6. To get a fair fit to the cosine basis, we have ignored the last four filter channels, with center frequencies of 5 kHz and above. The analyzed speech signals were just vowels at a constant level, so there is no zero-order or constant function in the basis set.

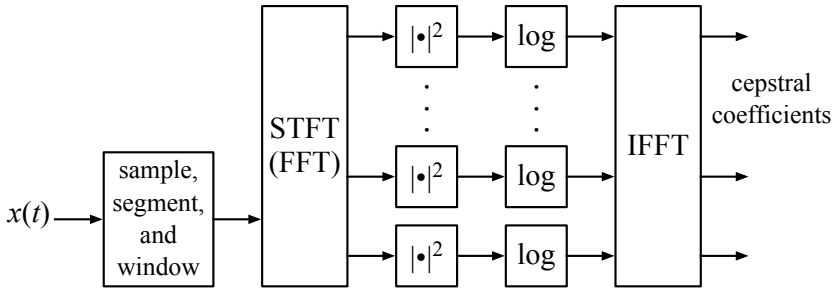


Figure 5.8: A cepstrum analyzer based on fast Fourier transform (FFT) operations. The first FFT produces complex coefficients, so a squaring or absolute value is needed. The second FFT can be a cosine transform if the spectrum is interpreted as symmetric in frequency, since a real symmetric function has no sine-phase components (no imaginary parts in its Fourier coefficients). The outputs are arranged from low to high *quefrency*, representing the smooth part and detailed part, respectively, of the log spectrum.

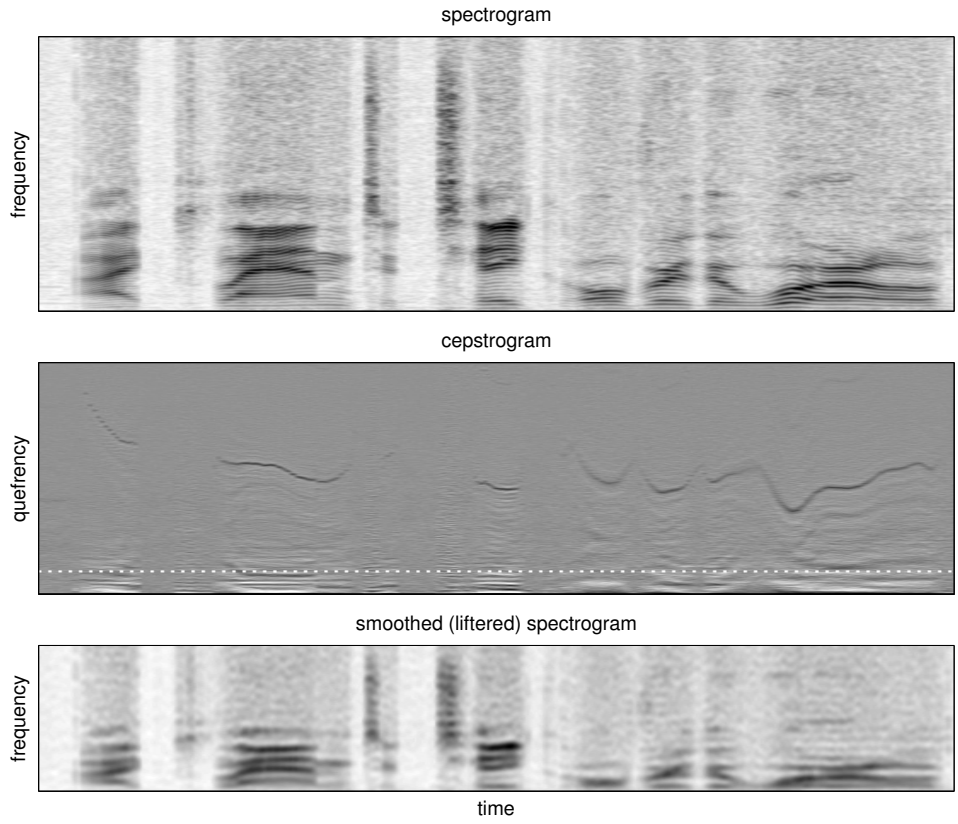


Figure 5.9: Power-law-compressed (exponent 0.15) spectrogram (top), cepstrogram (middle), and liftered spectrogram (bottom) reconstructed from the low-quefrency cepstral coefficients (from below the dotted white line). The first word in this utterance is the "I" illustrated in Figure 5.4. The full 11 kHz frequency range of the voice recording is shown in both spectrograms, at different scales, though this is a larger frequency range than is typically shown on a spectrogram with linear frequency scale. The dark curve visible in portions of the cepstrogram represents the pitch period, mirroring the movement of the pitch harmonics that are resolved in the narrowband spectrogram at top. The cepstrogram vertical axis (quefrency) can be thought of as the lag parameter of an autocorrelation, which is what it would be if the log or power-law nonlinearity were omitted and the second Fourier transform done on the power spectrum; here the lag runs from 0 to about 15 ms; cepstral coefficients are signed, with positive values darker, and negative values lighter, than the gray background. Cepstra are not usually displayed as cepstrograms, since there's not much interesting structure to look at, other than the pitch track, but they are used as representations for further sound processing.

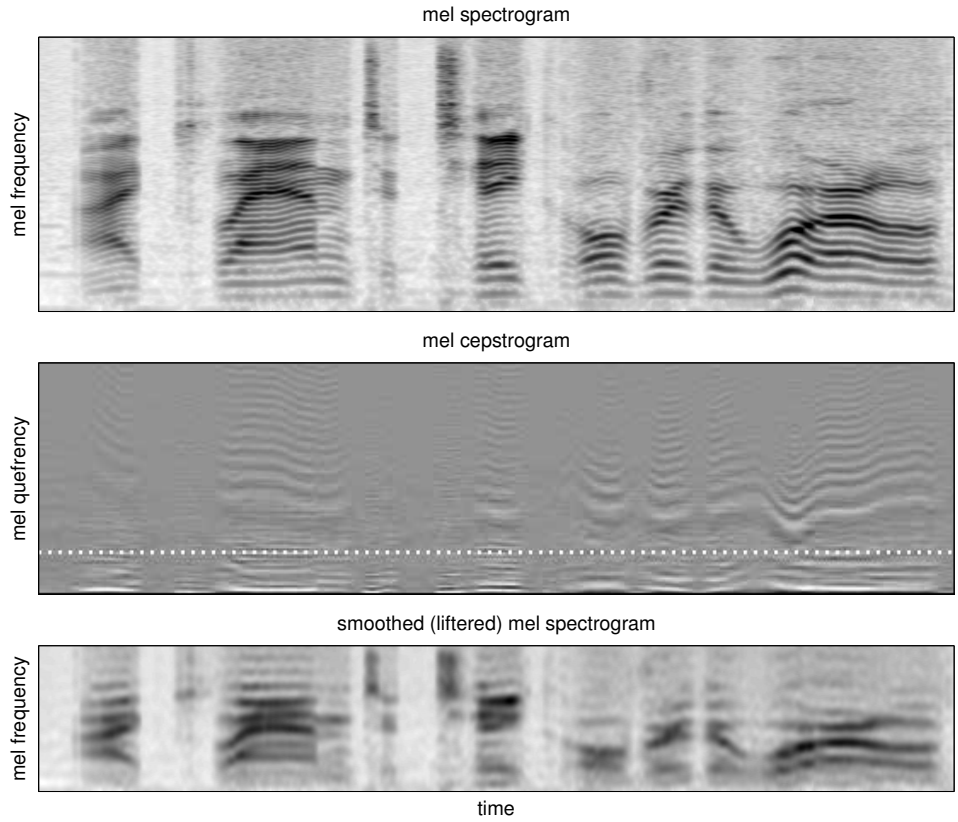


Figure 5.10: Mel-scale power-law spectrogram (top), mel-frequency cepstrogram or MFCCs (middle), and liftered mel spectrogram (bottom) reconstructed from the low-quefrency cepstral coefficients (from below the dotted white line); though this low-quefrency region looks bigger than in Figure 5.9, it is only 22 coefficients instead of the 33 used in the linear-frequency case. The mel frequency warping interferes with the usual cepstrum's simple representation of periodicity in the high-quefrency region, smearing out the pitch curve into ripples, but gives a much better distribution of spectral information.

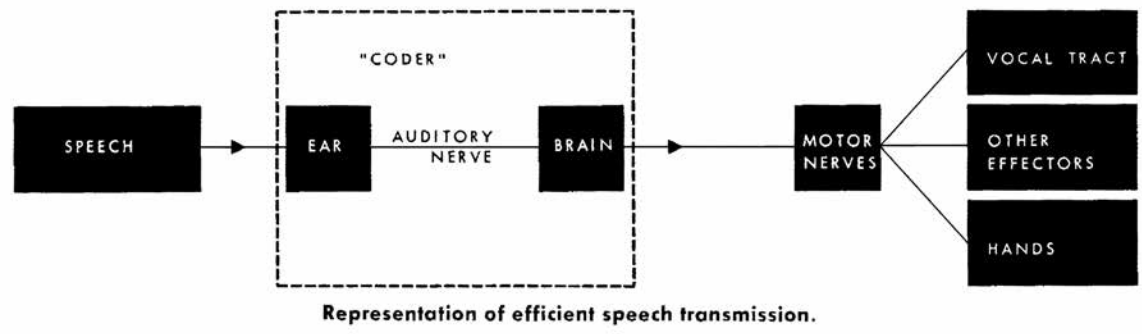


Figure 5.11: An early proposal for a speech analysis and recognition system based on mimicry of human processing (David, 1958). In agreement with our current recommendations, it respects the auditory nerve as a key representational funnel. [Figure 5 of (David, 1958) reproduced with permission of IBM.]

Part II

Systems Theory for Hearing

Part II Dedication: Charlie Molnar

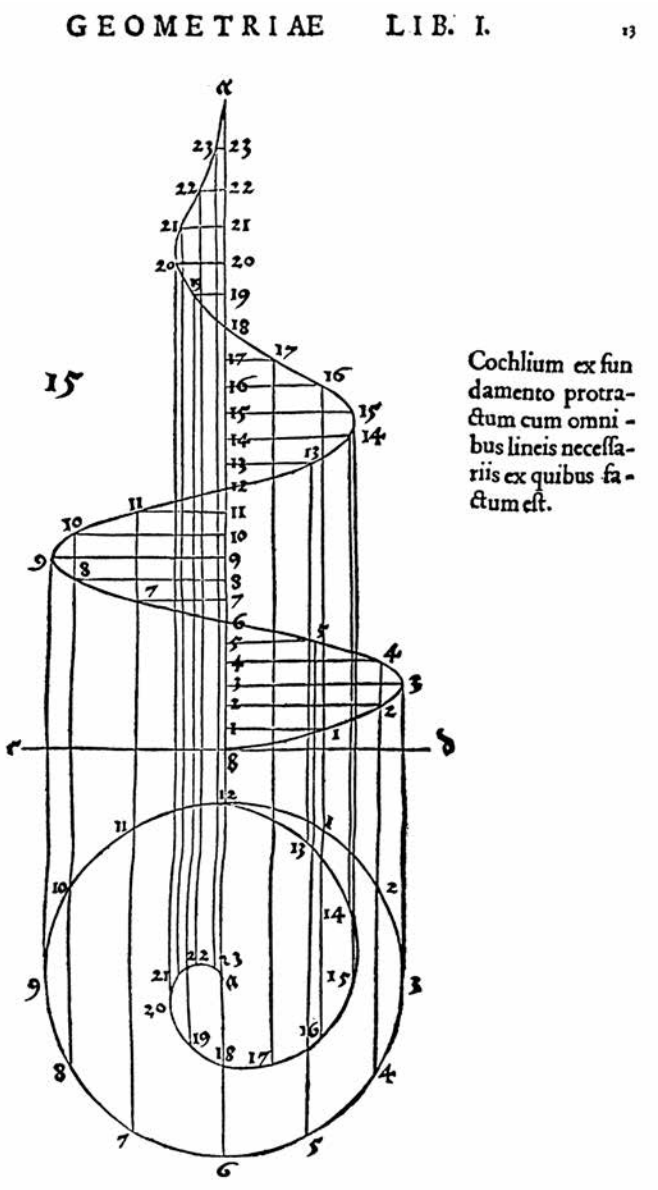
This part is dedicated to the memory of Charles E. Molnar (1935–1996). Charlie is mostly known outside the hearing field for his invention, with Wesley A. Clark, of the LINC—the "laboratory instrument computer"—which was according to many the first personal computer; and for his work in asynchronous and self-timed computer circuits, which is what he was working on when he died in 1996. He was a super generous guy, always willing to discuss and advise, and our talks about hearing and circuit design were very important to me. His "system of nonlinear differential equations modeling basilar-membrane motion" (Kim, Molnar, and Pfeiffer, 1973) was probably the first example of a great way to integrate nonlinearity into a filter-cascade model of the cochlea.

In this part, we develop the mathematical and engineering basics needed to model the ear.

We start with a review of linear systems theory, the body of knowledge that allows the design and construction of efficient and flexible filters. Even for readers very familiar with linear systems theory, a reading of this chapter should be a useful refresher and an introduction to our terminology and approach.

After a chapter on the discrete-time version of linear system theory, we apply the theory to resonant filters and elaborated resonant filters such as the gammatone family. Then we extend into nonlinear systems, with a whole chapter on automatic gain control.

Finally we discuss wave propagation in distributed systems, and how we model it with linear systems of the sort that will lead to good machine models.



Albertus Durer's 1532 *cochlium*, a "spiral extended from the base with all the necessary lines from which it is created," is unrelated to hearing, but generates a waveform that resembles the impulse response of a cochlear filter.

Chapter 6

Introduction to Linear Systems

> The fact that representation of waveforms as a sum of sine waves is useful in the elucidation of human hearing indicates that something involved in hearing is linear or nearly linear.
>
> — "The nature of musical sound," John R. Pierce (1999)

E.E. Connection: Direct and Alternating Current (DC and AC)

Electrical engineers often divide signals between *direct current* (DC) and *alternating current* (AC). The term *current* implied by the abbreviations is largely irrelevant, and it is not considered redundant or contradictory to speak of an AC current, a DC voltage, a DC response in a system that is not even electrical. DC just means steady, unchanging, or at zero frequency, while AC means cycling back and forth between positive and negative values, usually sinusoidally, with frequency as a parameter.

We use the term AC occasionally, in reference to frequency-dependent analysis. We use DC more frequently, to refer to steady-state conditions and low-frequency limits of signals and systems. Sound has no useful information at DC (at zero frequency), but the DC response of a sound-processing filter is often a practical characterization of its low-frequency behavior.

These terms were first popularized in the electric power transmission business in the 1880s, when Thomas Edison took the side of DC and George Westinghouse took the side of AC in their "Battle of the Currents" (Billington and Billington, 2013). Such contentiousness is not relevant now that things are better understood.

E.E. Connection: Linear Electrical Circuits and Filters

Electrical circuits made of linear components, such as resistors, capacitors, inductors, transmission lines, and ideal amplifiers, are what we normally think of when we speak of electrical or electronic filters. Consider the first-order lowpass filter of Figure 6.1: a circuit with a resistor of resistance R connects an input terminal to an output terminal, and a capacitor of capacitance C connects from the output to ground (that is, to a common circuit node, with respect to which the input and output voltages are measured). Typographically, we often use roman letters such as R and C as reference designators for components, and corresponding italic variable names such as R and C for their resistance and capacitance values.

Unlike a resistor, a capacitor is a component that has *state* or *memory*; it holds electrical charge, an imbalance in the number of electrons on its two plates, and the voltage across it is proportional to the charge that it is holding. The charge stored in the capacitor is the time integral of the current into it (the current I in Figure 6.1), and the voltage is proportional to that charge, with a ratio called the *capacitance*. This time integration is what gives it state: the present voltage is a function of the history that determines the stored charge. The other kind of electrical component with simple state is the *inductor*; the current through an inductor tends to persist, proportional to the time integral of the voltage across the inductor.

The resistor–capacitor or "RC" filter of Figure 6.1 is called *first-order* because it has only one *state variable*: the voltage across the capacitor. In general, the order of a system is the number of state variables needed to specify its instantaneous state. In our RC circuit, the voltage across the capacitor, call it V_C, is the filter output signal $y(t)$:

$$y(t) = V_C = \frac{1}{C}\int I dt \quad \text{or} \quad I = C\frac{dV_C}{dt}$$

The resistor, on the other hand, is a much simpler *stateless* element, characterized by *Ohm's law*, which says that current is proportional to voltage (the voltage V_R across the resistor corresponds to $x - y$ in our circuit), at every instant of time:

$$x(t) - y(t) = V_R = IR \quad \text{or} \quad I = \frac{V_R}{R}$$

The effect of connecting these two elements as drawn is that their currents I are equated, causing the output voltage to change more slowly than the input voltage, because it takes time for the integration of the current to change the voltage across the capacitor. The mathematics describing it is worked out in the main text.

The filter passes slow (low-frequency) fluctuations pretty well, but fast (high-frequency) fluctuations less well, so it is called a *lowpass filter*. The output is *smoother* than the input, but follows very low frequencies with a gain of 1, so it is called a *smoothing filter* (lowpass filters in general might have other gains). It may be referred to by such terms as *RC filter*, *RC lowpass*, or *RC smoothing circuit*. The latter names uniquely specify the circuit: the only way to connect one R and one C to make a filter that preferentially responds to low frequencies.

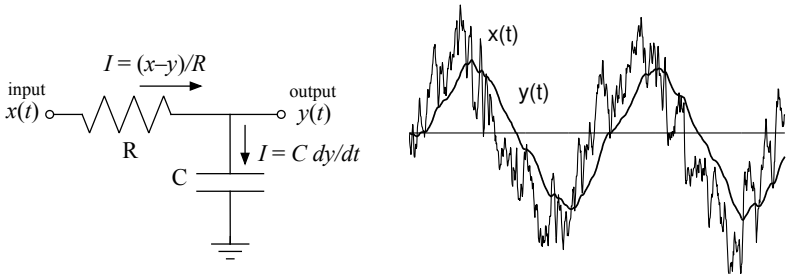


Figure 6.1: An example filter schematic diagram, and example input and output waveforms. The RC filter, with a resistor R in series with the input and capacitor C shunting the output, is the simplest smoothing filter. The differential equation that relates the output voltage $y(t)$ to the input voltage $x(t)$ is found via Kirchhoff's current law, which equates the currents $I(t)$. The graph on the right shows how a noisy input voltage signal $x(t)$ results in a relatively smooth output signal $y(t)$.

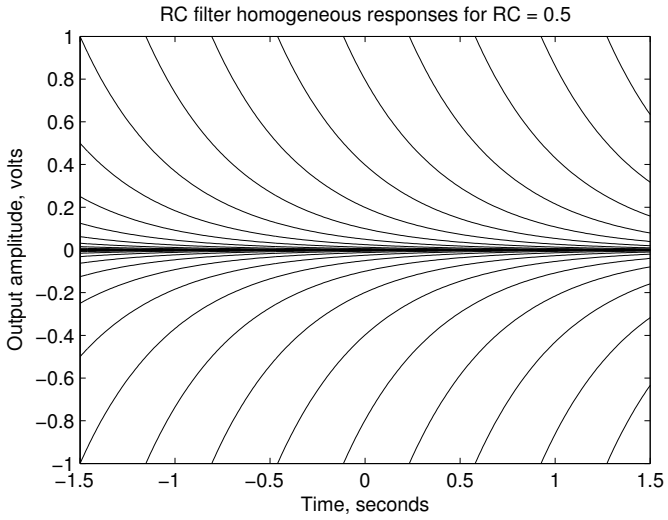


Figure 6.2: The homogeneous responses of the RC circuit are the signals that can appear at the output when the input is zero. The first-order system response is an exponential decay toward zero, from any starting value specified at any time value. For this example, the decay time constant is $\tau = 0.5$ s.

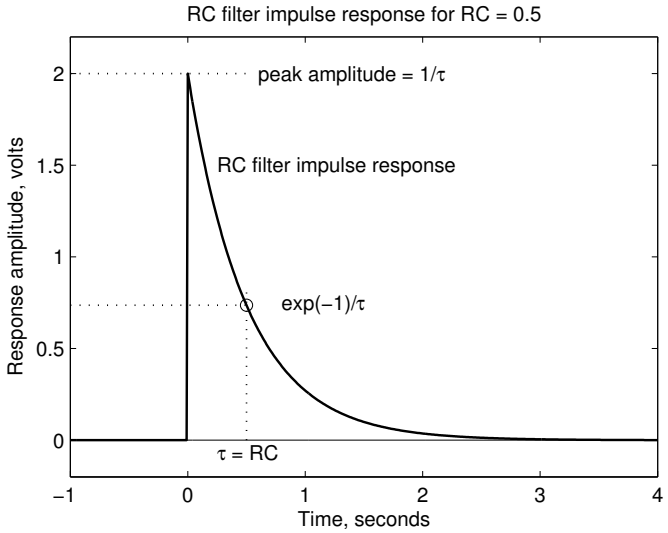


Figure 6.3: The impulse response of the example first-order RC smoothing filter of Figure 6.1 (solid curve), with time constant $\tau = 0.5$ s.

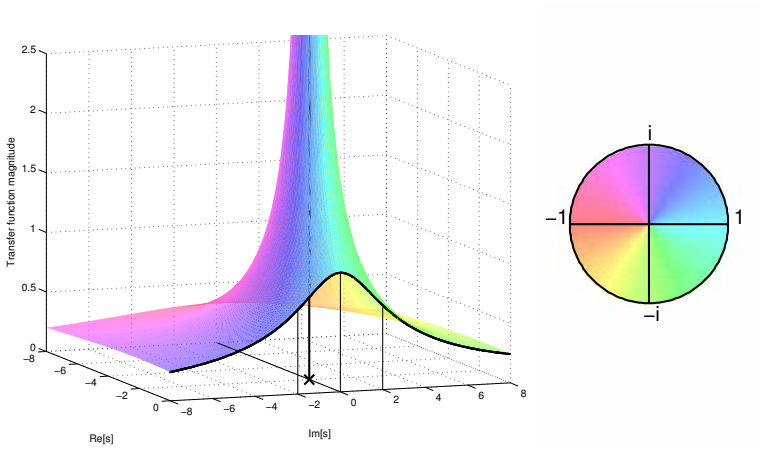


Figure 6.4: The transfer function of the example RC filter of Figure 6.1. The magnitude of $H(s)$ is plotted as a surface height above the complex s plane, while the phase of $H(s)$ determines the hue of the surface color (following the phase–hue legend on the right). For the example filter with $\tau = 0.5$, the transfer function has a singularity at $s = -2 + i0$ (at cross and heavy vertical line). The surface is cut along the imaginary s axis to reveal the frequency response. The frequencies $\omega = 0$ (DC) and $\omega = \pm 2$ (the 3-dB points) are marked by lines.

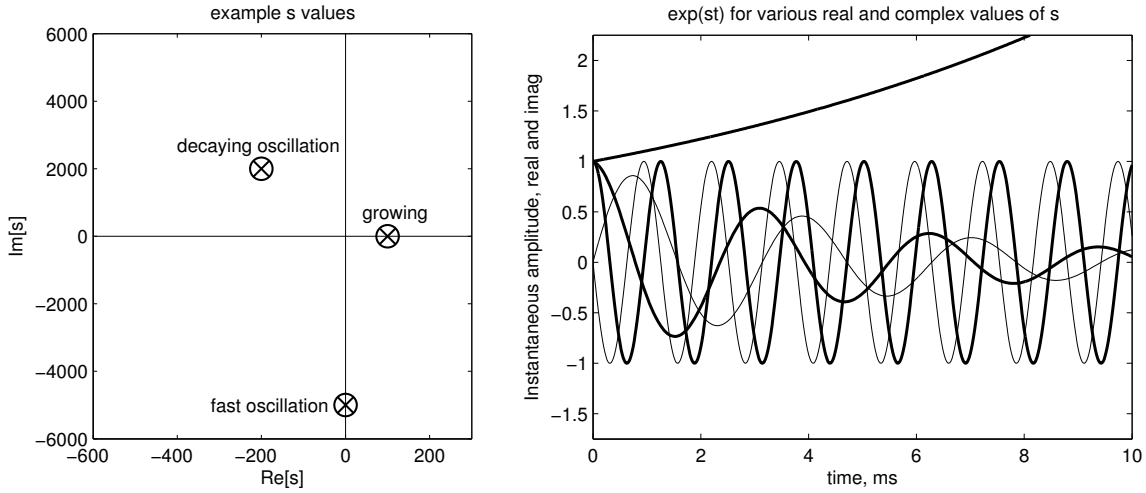


Figure 6.5: Plots of example s values (left) and the real and imaginary parts of $\exp(st)$ (right); for complex s, the imaginary part is shown by light curves, and the real parts are always heavy curves. One real s value, $s = 100$, yields the slow exponentially growing curve on top. A pure imaginary value $s = -i5000$ corresponds to the steady high-frequency sinusoids shown; notice that the imaginary part *leads* the real part, since this one is a *negative frequency* complex sinusoid. The damped sinusoid shown corresponds to $s = -200 + i2000$. In each case, the units of s are inverse seconds, or radians per second. If the input to a linear time-invariant system, or filter, is any of these signal shapes, then the output will be the same, just multiplied by the complex factor $H(s)$ for that filter at that s value. That is, these functions $\exp(st)$ are eigenfunctions of all LTI systems (subject to some region-of-convergence restrictions that we mostly ignore).

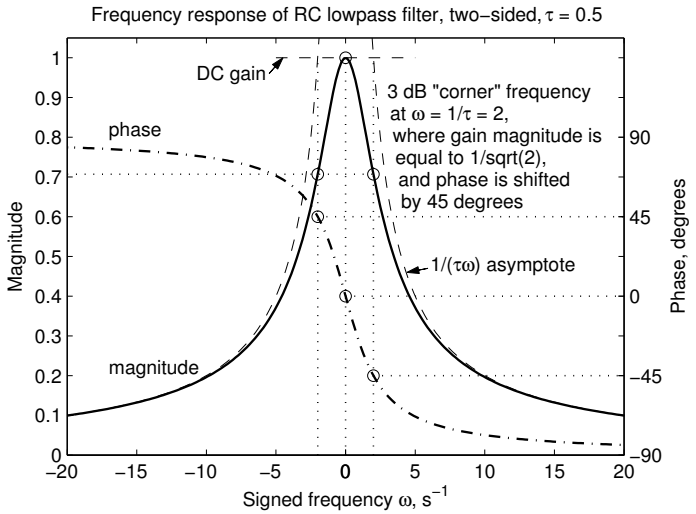


Figure 6.6: The magnitude frequency response (solid) and phase frequency response (dash-dot) of the RC lowpass filter with time constant of 0.5. The phase varies from zero at DC to a lag of a quarter cycle (90 degrees or $\pi/2$ radians) at high frequencies (and the negative of that at negative frequencies). The frequency parameter ω is in radians per second. The 3 dB *corner frequency*, $\omega_C = 1/\tau = 2$, is where the gain is 0.707 and the phase is ±45 degrees (marked with circles). Also shown (dashed) are the magnitude-gain high-frequency asymptotes, the hyperbolas $|1/(\tau\omega)|$, and the low-frequency (DC) limit of gain 1. These power-law asymptotes are shown for comparison with their straight-line versions in a log–log plot in the next figure.

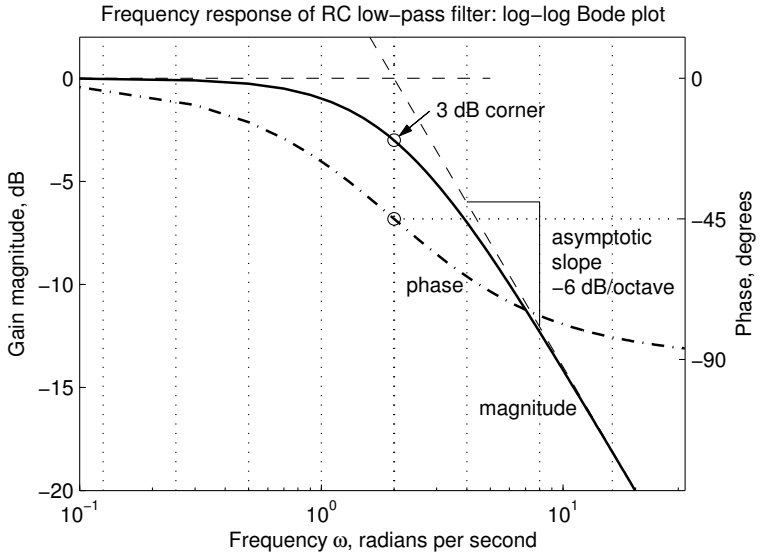


Figure 6.7: Bode plot, or log–log frequency response, of the example RC lowpass filter. Magnitude-gain asymptotes are shown dashed, and octaves are marked by dotted lines. The high-frequency asymptote has a slope of –6 dB per octave (or –20 dB per decade, that is, per factor of 10 in frequency), no matter what R and C values are used, due to the $1/f$ characteristic rolloff of a one-pole system. Note that the same data looks very different in this presentation than in the linear plot: the low-frequency asymptote and the "corner" shown in Figure 6.6 make more sense this way.

M.E. Connection: Mechanical State

Mechanical engineers will note that we have started with electrical examples. There are corresponding linear mechanical systems that have similar equations. Electrical circuits are often used as models of mechanical systems, and we use them that way when we connect them to mechanical filtering in the ear.

In mechanical systems, masses have state, because their velocity tends to persist, with momentum being proportional to the time integral of applied forces that accelerate the mass. And springs have state, since they push back with a force proportional to displacement, where displacement is the time integral of the velocity of whatever is displacing. Like inductors and capacitors, these elements store energy: masses, like inductors, store kinetic energy, while springs, like capacitors, store potential energy. Mechanical systems can be modeled by electrical ones, and vice versa, by making the appropriate analogs, such as current for velocity and voltage for force.

If you have come to hearing via mechanical engineering, acoustics, or applied physics, I trust you will be able to make the appropriate connections.

E.E. Connection: Lumped and Distributed Circuits

Both resistors and capacitors, as well as (idealized) inductors and transformers and amplifiers, are what are known as *lumped* elements, since they lump the effects of physical structures into simple device models that are described by just the voltages and currents at their terminals.

Another important class of linear element is the *distributed* element, most notably the transmission line. Any piece of wire long enough to introduce an appreciable delay, such that points along the wire cannot be treated as all having the same voltage, must generally be treated as a distributed element. The analysis of distributed systems, or transmission lines, was developed to characterize and improve telegraph lines in the nineteenth century (Heaviside, 1892). The treatment of distributed elements has a lot in common with the treatment of lumped elements, but the math is a bit different. In particular, distributed systems involve partial differential equations, to describe functions of both time and space, while systems of lumped elements get by with ordinary differential equations, describing functions of time only. We'll treat distributed systems in Chapter 12 and later chapters, as we get closer to the mathematics of waves in the cochlea.

A continuous-time moving-average filter is perhaps the simplest example of an LTI system that cannot be represented as a lumped system. An exponentially-weighted moving average, on the other hand, is what the RC lowpass filter example of this chapter computes, using the state of a single lumped element, a capacitor.

Math Connection: Complex Exponentials as Eigenfunctions of LTI Systems

Consider the generalization of real sinusoidal inputs to decaying and growing complex sinusoids of the form

$$x(t) = A_x \exp(st)$$

for some frequency-like parameter s, not necessarily pure imaginary like $i\omega$. With this complex exponential as input, manipulation of the convolution integral shows that the output will be a complex exponential with the same parameter s:

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} A_x\, h(u) \exp(s(t-u))\, du \\ &= A_x \exp(st) \int_{-\infty}^{\infty} h(u) \exp(-su)\, du \\ &= A_y \exp(st) \end{aligned}$$

That is, the output is like the input but with a different complex amplitude factor A_y, proportional to A_x in a way that depends on s (for values of s where the integral converges). Functions with this property—that the output function is like the input function times a constant factor—are known as *eigenfunctions*. The ratio of output to input, as a function of s, is called the *transfer function* $H(s)$, and the integral that determines it is known as the *Laplace transform* of the impulse response $h(t)$:

$$H(s) = \frac{A_y}{A_x} = \int_{-\infty}^{\infty} h(u) \exp(-su)\, du = \mathcal{L}\{h(t)\}$$

For the example RC filter, as we'll see in Section **??**, the transfer function is:

$$H(s) = \frac{Y(s)}{X(s)} = \frac{A_y}{A_x} = \frac{1}{\tau s + 1}$$

for values of s where the integral converges, which are $\mathrm{Re}[s] > -1/\tau$.

E.E. Connection: Slightly More General Circuits

The transform variable s can be pushed all the way back into circuit elements. If Ohm's law (his electrical law $V/I = R$ relating voltage, current, and resistance of a resistive circuit element), and the idea of resistance as the ratio, are generalized to complex values, then the simple techniques of DC circuit analysis suddenly become capable of AC circuit analysis: analyzing the response for frequencies other than zero. Impedance (the complex version of resistance) can be thought of as the transfer function from current to voltage: $V(s)/I(s) = Z(s)$. The element impedances come from the usual trick of replacing differentiation by the s operator (and integration by $1/s$) in the definitions of the terminal relationships of the different component types. A capacitor of capacitance C is thus treated as an impedance equal to $1/sC$, and an inductor of inductance L as an impedance sL. This way, the electrical engineer can do a complete analysis of a circuit's frequency response algebraically, without ever looking at a differential equation or an integral. In mechanical systems, masses and springs can be similarly treated.

Systems of masses and springs, or inductors and capacitors, can *ring* in response to an impulse, rather than simply decay monotonically toward zero as our first-order example does. They can *resonate*, or respond strongly to signals of certain frequencies, as investigated in Chapter 8.

To simplify matters, we focus on examples in the form of the common *voltage divider* circuit of Figure 6.8. The impedance of block Z1 between the input and the output (for example, R in our first-order RC filter) is referred to as the *series impedance*, and that of block Z2 between the output and ground (which is $1/(sC)$ in the RC filter) as the *shunt impedance*. The currents I through the two impedance blocks Z1 and Z2 are equal, and the voltages across them (impedance times current, $Z_1 I$ and $Z_2 I$) add up to the input voltage X. The output voltage Y is $Z_2 I$, so the ratio of output to input voltage is easily found by canceling the I factors:

$$H = \frac{Y}{X} = \frac{Z_2}{Z_1 + Z_2}$$

For circuits of resistors, the impedances are just resistances, and the variables are all real scalars. The circuit is known as a *resistive voltage divider* in that case, and the output voltage is always less than the input voltage, by a ratio that does not depend on frequency. For circuits of *reactive* elements (inductors and capacitors), which store energy and induce frequency-dependent phase shifts and gains, we use the complex impedances, with the same algebra, to compute a complex frequency-dependent ratio, the transfer function. The first-order RC lowpass filter's transfer function can thus be found without explicit use of a differential equation. If the impedances Z_1 and Z_2 of the two blocks Z1 and Z2 include two reactive elements of different types (a capacitor and an inductor), the system would be second order and could resonate.

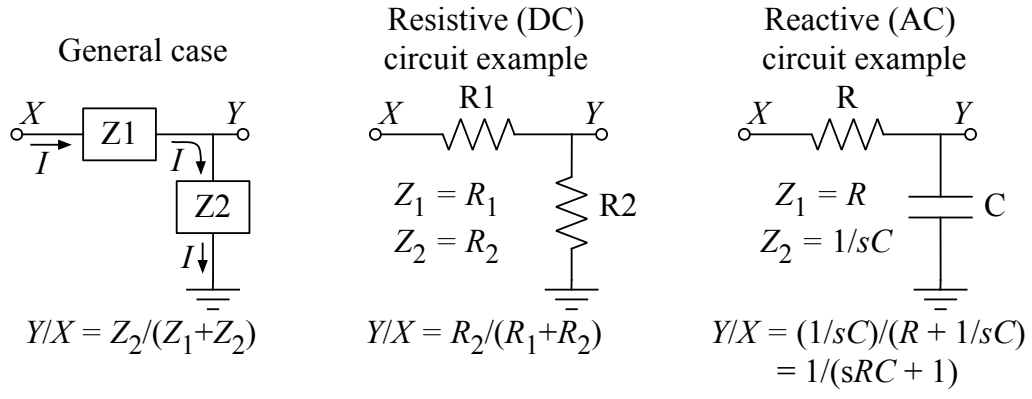


Figure 6.8: Schematic diagram of the general voltage-divider circuit (left), a simple resistive case, and a reactive case.

E.E. Connection: Second-Order Filter Circuits

Second-order filters are more interesting and relevant to hearing than our first-order example is, since they can be *resonant*, that is, particularly responsive to frequencies in a certain range. A resonant system is known as a resonator, or *single-tuned resonator* in the second-order case. Second-order resonant systems are the building blocks of almost all models of cochlear function.

In mechanics, examples of resonant systems are mass–spring systems and pendulums. Resonant systems generally have two different kinds of energy storage mechanisms, and dynamics that make the system's energy oscillate between the two types, for example, between the kinetic energy of a moving mass and the potential energy of a stretched or compressed spring.

In electrical circuits, inductors store energy as the kinetic energy of collectively moving electrons (Mead, 2002) (or in the magnetic field in the Maxwellian conception), and capacitors store potential energy by accumulating charges against a net electrostatic repulsion. The differential equations that describe the motion back and forth between kinetic and potential energy are essentially the same as for mechanical systems.

Consider the circuit of Figure 6.9, a second-order filter formed by adding an inductor to the series impedance of the RC lowpass filter. We call it "circuit A," the first of several resonant systems that we analyze in detail in Chapter 8.

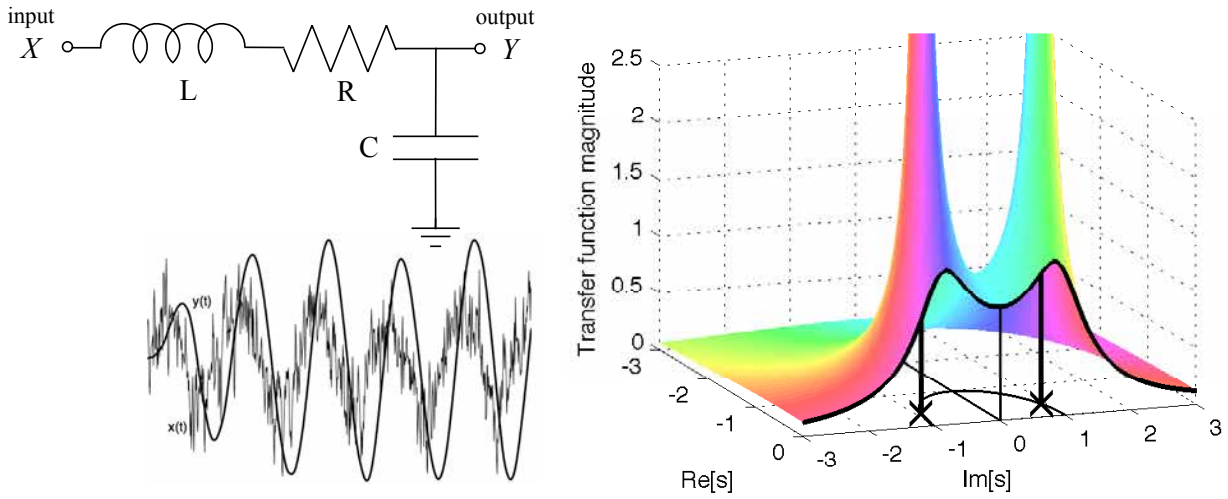


Figure 6.9: Filter A, a second-order resonant lowpass filter, is diagrammed (left) with an example of a noisy input waveform and a corresponding output waveform, which is smooth but has an increased amplitude of the input component that is close to the resonance frequency. The transfer function of filter A (right), plotted as in Figure 6.4, resembles a tent fabric draped over a pair of "tent poles" at the singularities, the two complex pole positions (at crosses and heavy vertical lines).

The Z1 block, or series impedance, has impedance $Z_1 = sL + R$, and the Z2 block, or shunt impedance, has impedance $Z_2 = 1/sC$, so the transfer function, from the voltage-divider approach described in the previous box, is

$$H_A(s) = \frac{1/sC}{sL + R + 1/sC} = \frac{1}{s^2LC + sRC + 1}$$

This filter is termed second-order because it has two independent state variables: the voltage or charge stored on the capacitor, and the current stored in the inductor. In this case, the numerator is zero-order, with no roots, so the filter has two poles but no zeros. The poles can be real, for large enough R, but in the more interesting case, where the circuit is resonant, the poles are a complex conjugate pair.

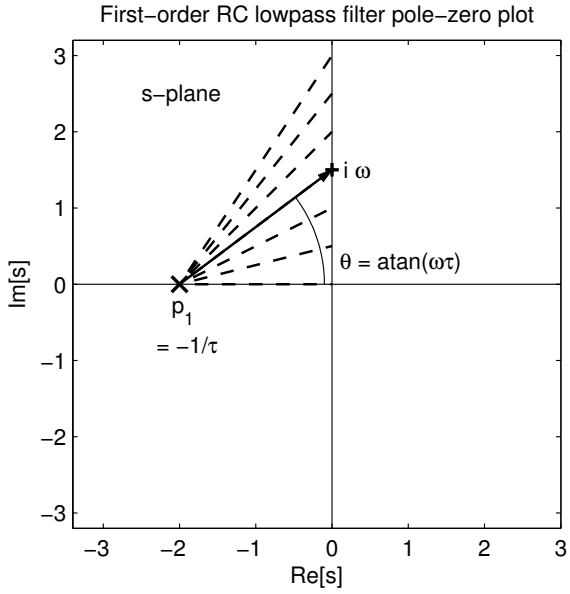


Figure 6.10: Calculating the frequency response of the example RC lowpass filter from an s-plane pole–zero diagram with one pole at $p_1 = -1/\tau$ and no zero: $1/(s-p_1)$. The magnitude response is inversely proportional to the length of $s - p_1$, the line from the pole p_1 to the frequency point at $s = i\omega$ (shown here for $\omega > 0$), and the phase lag is the angle θ of that line from the real axis. Since the angle θ appears in the denominator, it becomes a negative phase shift, which represents a lag, or delay (it would be a positive phase shift for $\omega < 0$, which is still a lag).

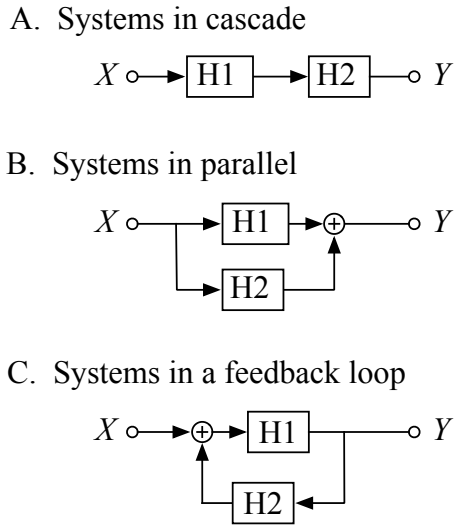


Figure 6.11: Example filter systems interconnected in cascade, parallel, and feedback configurations.

E.E. Connection: On Cascades

The notion of a cascade of filters is so central to our models of hearing that it needs a little extra attention. There are at least three relevant contexts in which cascades appear prominently in linear systems and hearing literature. These contextual meanings are closely related, and we use them all, but discussions in the literature are sometimes limited to narrower interpretations.

First, as a description of a way of interconnecting filter circuits, as traditionally used in the radio, telephone, and television fields, cascades are sometimes mentioned in the sense analyzed in this chapter. Specifically, in a cascade connection, the filter stage circuits are "buffered" (by a unity-gain follower amplifier) such that the output of each one is not disturbed when it is connected to the next; that is, such that there is no backwards influence, coupling, or "loading" from the next connected circuit. Cascades that implement what we now call gammatone filters were described in the 1940s by Eaglesfield and by Tucker. Tucker (1946) explains the difference between the generic concept of "in series" and "cascade," a difference we illustrate in Figure 6.12: "The use of a series of tuned circuits coupled together by mutual inductance, mutual capacitance, or resistance is well known, and the response of such an arrangement is analysed in various textbooks and papers. The use of a cascade of tuned circuits which have transmission coupling but no mutual coupling is not so often referred to, however, although such an arrangement occurs frequently in practice." Eaglesfield (1945) referred to his cascades simply as multistage amplifiers, where by stages he meant the circuit portions isolated from backward coupling by buffer amplifiers.

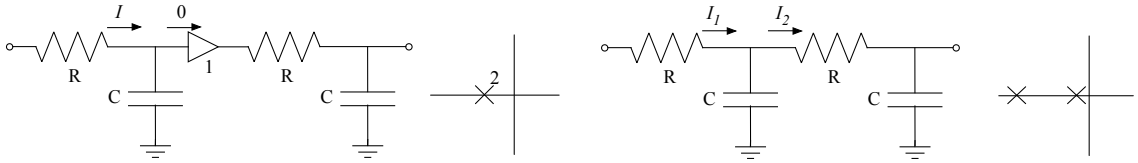


Figure 6.12: A cascade of two identical RC filters separated by a buffer amplifier (left) and without a buffer (right). The buffer (triangle with label "1" indicating its gain), takes zero input current and copies its input voltage to its output, so the second filter stage does not influence the first stage by taking current out of it, or "loading" it. The resulting transfer function H^2, the square of the one-pole stage transfer function, has two poles in the same location, at $s = -1/RC$, as shown in the left pole plot. Without the buffer, some of the current through the first resistor flows into the second, contrary to what we assumed when we computed the transfer function of the RC filter. Such mutual coupling—the second stage affecting the first stage—results in a different filter, with two different pole locations as shown in the right pole plot.

Second, as a description of the propagation of waves along a continuum or sequence of places, a cascade of filters representing each place-to-place transition is a natural description. In modeling the propagation of sound-induced waves in the inner ear, the notion of a cascade was articulated by Licklider (1953) as a description of the traveling-wave model: "...the passage of the traveling wave down the partition–fluid system, which if it is viewed as consisting of resonators, consists of many elements in cascade and not in parallel." Here Licklider's cascade concept may also have been intended to cover the then-current electrical network models of the cochlea that would actually support wave propagation in both directions; that is, he did not necessarily recognize Tucker's distinction quoted above. At least one paper on modeling waves in the cochlea has used *cascade* for simply interconnected lumped circuits (Kletsky and Zwislocki, 1981), but most of us in the auditory modeling field use the term as Tucker suggests, reserving it for models that propagate signals through stages only in the forward direction (Lyon, 1998; Sarpeshkar, 2000).

Finally, cascades are an implementation strategy for many kinds of digital filters. A typical textbook on digital signal processing will mention cascades only in this sense, as the usually most natural and robust way to break up a filter design into easily implementable pieces.

In this book, all of these senses are important, and mostly equivalent. In the case of the cascade model of wave propagation in the inner ear, multiple outputs are also very important. The "multiple-output cascade filterbank" is the natural structure for efficiently implementing a model of sound processing in the cochlea.

Chapter 7

Discrete-Time and Digital Systems

> Since a sufficiently approximate solution of many differential equations can be had simply by solving an associated difference equation, it is to be expected that one of the chief fields of usefulness for an electronic computor would be found in the solution of differential equations.
>
> — "The use of high-speed vacuum tube devices for calculating," John Mauchly (1942)

Statistics and History Connection: Generating Functions and Z Transforms

In the field of statistics, discrete sequences are sometimes described by *generating functions*, which are essentially the same as what we call Z transforms. When the discrete sequence is the probability mass function of a discrete random variable, its generating function is called a probability-generating function.

Generating functions were described by Abraham de Moivre in 1730, in proving his formula for the distribution of the sum of several independent identically distributed random variables (such as the total number of pips showing on six dice), and were named by Pierre-Simon Laplace in 1780 (Hald, 2005).

The theory of discrete-time systems with Z transforms was worked out at the MIT Radiation Laboratory during World War II by the mathematician Witold Hurewicz (1947)—for analyzing predictors of variables such as airplane positions as part of their radar development effort, as recounted by Bennett (1993). The name was provided a few years later, in follow-up work by Ragazzini and Zadeh (1952), who discussed the relationship to generating functions in depth.

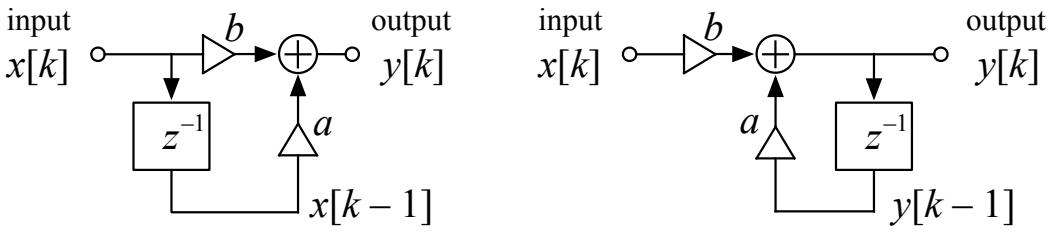


Figure 7.1: The signal-flow diagrams of two different first-order digital filters are shown here: a finite-impulse-response (FIR), or nonrecursive, filter on the left, and an infinite-impulse-response (IIR), or recursive, filter on the right. With appropriate choice of coefficients a and b, namely $a = 1 - b$, with both coefficients between 0 and 1, these are smoothing filters (that is, they suppress fluctuations and have average output equal to average input). The FIR filter on the left can only do a little bit of very local smoothing, while the IIR filter on the right behaves like the RC lowpass filter, with potentially very long time constant, when the coefficients make it a smoothing filter. The label z^{-1} in the box represents the unit delay operator. The triangles represent multiplication by the constant coefficients shown. The two filters' difference equations, $y[k] = bx[k]+ax[k-1]$ and $y[k] = bx[k] + ay[k - 1]$, and z-domain operator equations, $Y = bX + z^{-1}aX$ and $Y = bX + z^{-1}aY$, are apparent from the signal-flow diagrams. Labeling the input $x[k]$ and the output $y[k]$ as here suggests the operation steps on sequences of samples, with z^{-1} being a unit delay, or memory element. We alternatively label them X and Y, which suggests interpreting the z^{-1} as a transform operator. We switch back and forth between such labels and interpretations, as we did in Chapter 6.

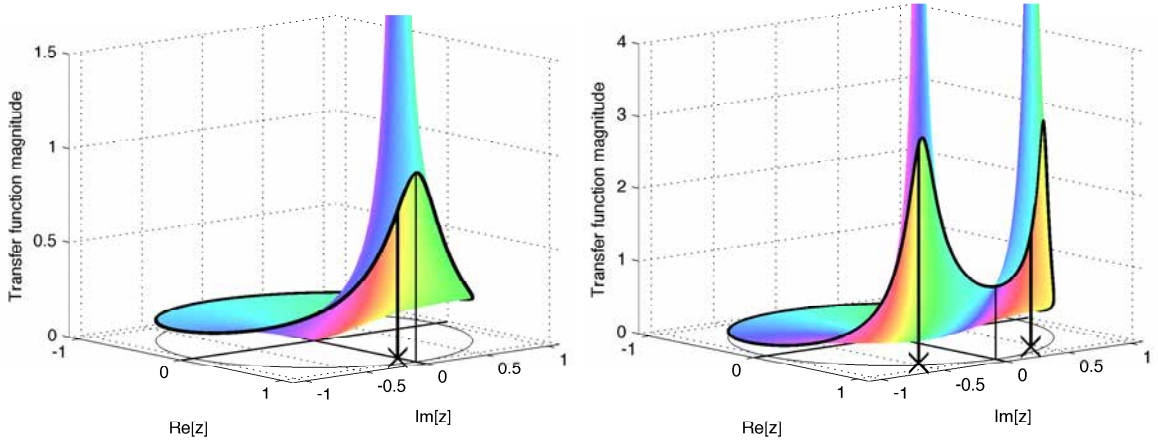


Figure 7.2: Complex transfer functions of one-pole (smoothing) and two-pole (resonator) filters, evaluated inside the unit circle of the z plane. The frequency response is the transfer function evaluated on the unit circle, shown by the dark curves at the circular cut. As in Figure 6.4, phase is mapped to hue (see the color plates); there is one cycle of hue variation around each pole.

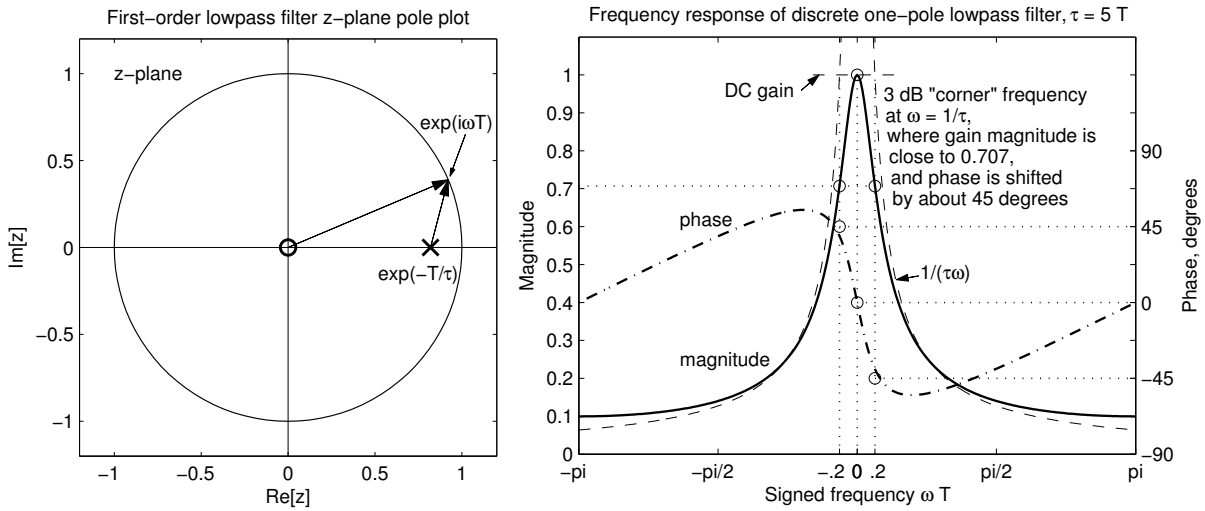


Figure 7.3: Calculating the frequency response of the example first-order discrete-time lowpass filter from a z-plane pole–zero diagram with one pole and a zero at the origin. In this example, the pole is located at $z = \exp(-0.2)$, corresponding to $T/\tau = 0.2$, a smoothing time constant of 5 samples (for example, 5 ms time constant at 1 kHz sample rate, $T = 0.001$ s). For any frequency ω, the magnitude response is inversely proportional to the length of the line from the pole to the frequency point at $z = \exp(i\omega T)$, and the phase lag is the angle between the real axis and that line. Since the zero is at the origin, its distance to the frequency point on the unit circle is always 1, so the zero does not affect the gain magnitude (any other position of the zero would affect the gain magnitude); this zero does provide a phase lead, however, which reduces the net phase lag.

Detail: Zeros at the Origin

Why does the one-pole continuous-time filter correspond to a one-pole–one-zero discrete-time filter?

$$H(z) = \frac{bz}{z - a}$$

The factor of z in the numerator of the example lowpass transfer function represents a zero at the origin. It means the output is advanced by one sample from what a filter without that zero would do, that is, compared to the case where the filter output is taken *after* the delay element in Figure 7.1. After that delay, the output would be $y[k-1]$, or $z^{-1}Y$. The pole at the origin in $1/z$ cancels the zero in z in that case, corresponding to omitting the z in the transfer function.

Being at the origin, the zero in z affects the phase, but not the magnitude, of the frequency response: the factor of z, or $\exp(i\omega T)$, contributes a phase advance of ωT at frequency ω rad/s.

When a discrete-time filter has more poles than zeros, zeros at the origin can be added, by adding factors of z, advancing the output, reducing the phase lag while keeping the filter causal. The filter is not minimum phase without them (the concept of minimum phase was introduced in Section **??**), because the inverse filter would not be causal.

There is no corresponding concept for continuous-time systems with rational-function transfer functions; s-domain points that would map to $z = 0$ are infinitely far to the right side in the s plane, where zeros have no effect at finite frequencies.

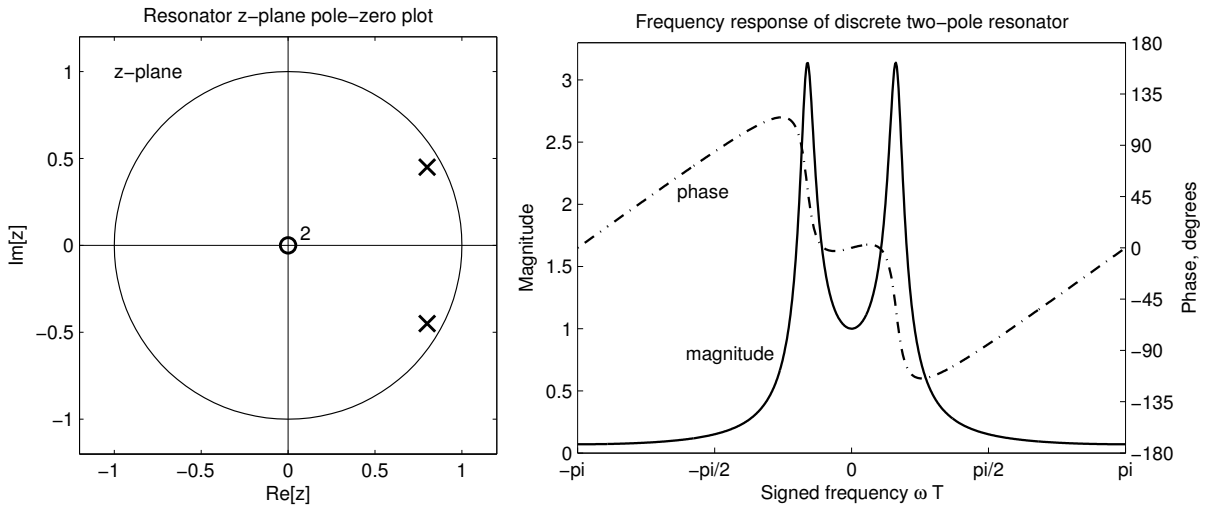


Figure 7.4: The pole–zero plot and frequency response of the example second-order discrete-time resonator of Figure 7.2. The response magnitude is proportional to the product of the reciprocal distances between points on the unit circle and the two poles, so the positive and negative frequencies that are close to the upper and lower poles both produce gain peaks. Following the process described in Figure 7.3 for each pole and zero, the log magnitude gains of the two poles add, and so do the phases contributed by the two poles, as these are the real and imaginary parts of the complex log of the complex gain. The two zeros at the origin add to the phase, but do not affect the gain magnitude.

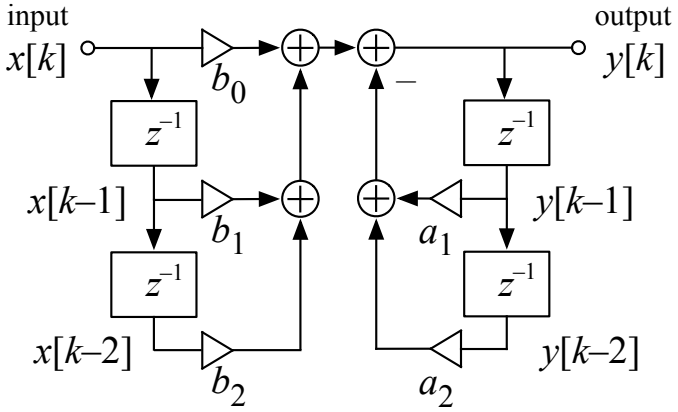


Figure 7.5: The signal-flow diagram of a second-order digital filter with three forward coefficients and two feedback coefficients. The filter's difference equation $y[k] = b_0x[k] + b_1x[k-1] + b_2x[k-2] - a_1y[k-1] - a_2y[k-2]$, which is apparent from the diagram, is also the program step that computes an output sample.

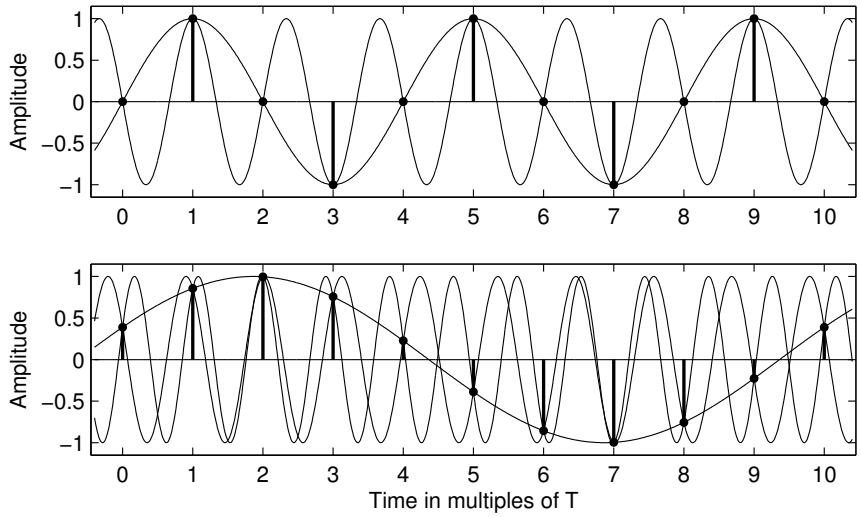


Figure 7.6: Sampled sine waves with frequencies 0.25 and 0.75 times the sampling frequency (top) can give identical sample sequences (dots), as discussed in the text. In light of their identical samples, the signals shown are *aliases* of each other. Another example, frequencies 0.1, 0.9, and 1.1 cycles per sample (bottom), shows how a low frequency has aliases slightly above and below the sampling frequency.

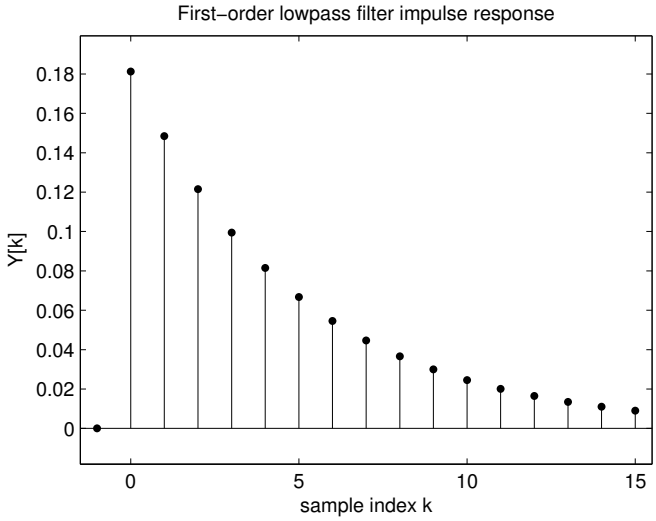


Figure 7.7: The discrete-time first-order smoothing filter's impulse response is a geometric sequence (for $k \geq 0$), a sequence of samples of an exponential decay.

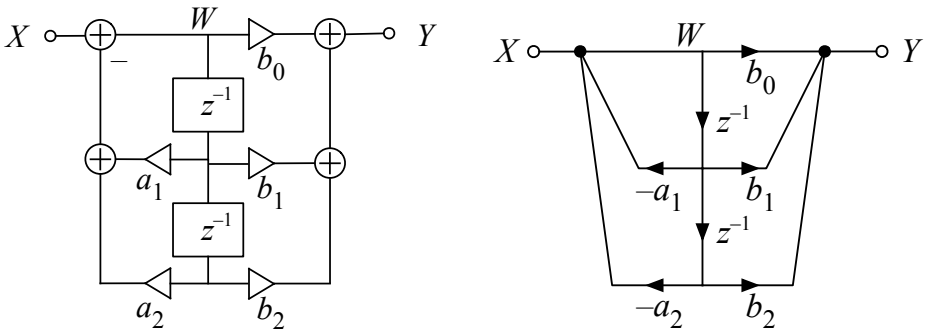


Figure 7.8: In a typical representation of a direct form II discrete-time or digital second-order section (left), the feedback block that makes the poles comes first, followed by the feed-forward block that makes the zeros. These blocks share two delay elements, delaying an intermediate signal, neither X nor Y, which is the output of the first block and the input of the second block. The directions of signal flow are implied by the coefficient multipliers (the oriented triangles), so the generous use of arrows that we saw in Figure 7.5 is typically avoided. In the even more concise form (right) arrows on lines are used to indicate multiplicative operators, including the delays, and dots represent additions.

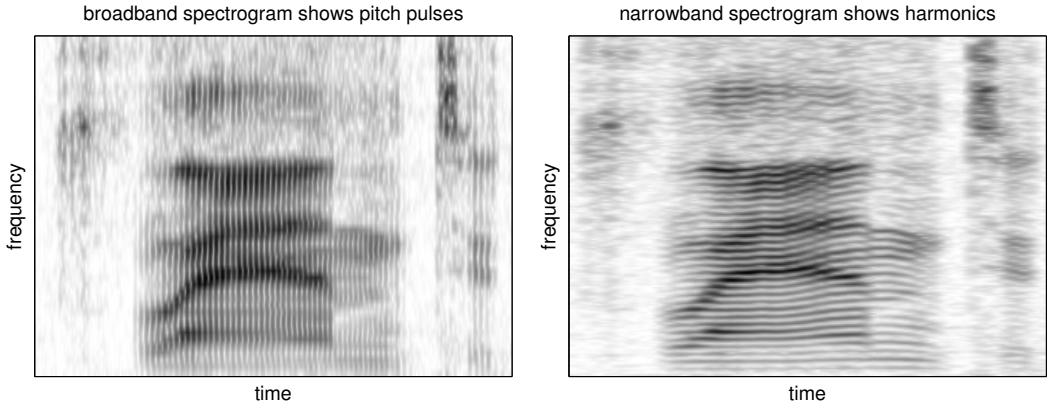


Figure 7.9: A wideband spectrogram (left) and a narrowband spectrogram (right) show the short-time power levels out of a bank of bandpass digital filters. When the filter bands are wide enough, the temporal response is fast enough to show individual glottal pulses of speech, as on the left. Conversely, when the filter bands are narrow enough, they can resolve individual harmonics of the pitch frequency, as on the right. But neither version captures the much finer temporal structure, the phase information or individual waveform peaks that the auditory nerve can represent. The spoken words analyzed here are “plan to.”

Chapter 8

Resonators

Another experiment should be adduced. Raise the dampers of a pianoforte so that all the strings can vibrate freely, then sing the vowel *a* in *father, art,* loudly to any note on the piano, directing the voice to the sounding-board of the piano; the sympathetic resonance of the strings distinctly re-echos the same *a*. On singing *oe* in *toe*, the same *oe* is re-echoed.

— *On the Sensations of Tone*, Hermann Ludwig F. Helmholtz (1863)

In the classes of circuits discussed in the following chapters, the pole-zero patterns show at a glance the general form of the frequency characteristics with the important features placed clearly in evidence; they display the effects of varying the circuit parameters; and they reveal the key approximations that permit certain groups of complex circuits to be treated as equivalent circuits of less complexity.

— *Pole-Zero Patterns: In the Analysis and Design of Low-order Systems*, Angelo and Papoulis (1964)

EE Connections: Alternative Resonant Circuits

We analyzed a simple two-pole resonant circuit in Chapter 6 (see Figure 6.9). Three other circuits with the same three components in series (that is, connected such that the same current goes through all of them) are shown in Figure 8.3.

Two of these new circuits, filters B and C, are of the same form as filter A, namely the generalized voltage divider form shown in Figure 6.8. Since they have the same sum of impedances (the denominator of the voltage-divider equation), we only need to update the Z2 impedances (the numerators of the voltage-divider equation) to get their transfer functions:

$$H_{\mathrm{A}}(s) = \frac{1/sC}{sL + R + 1/sC} = \frac{1}{s^2LC + sRC + 1}$$

$$H_{\mathrm{B}}(s) = \frac{R + 1/sC}{sL + R + 1/sC} = \frac{sRC + 1}{s^2LC + sRC + 1}$$

$$H_{\mathrm{C}}(s) = \frac{R}{sL + R + 1/sC} = \frac{sRC}{s^2LC + sRC + 1}$$

The first two, filters A and B, have unity gain at DC (at $s = 0$); the average output voltage will be equal to the average input voltage. The third is *AC coupled*, that is, with zero gain at DC (a zero at $s = 0$), due to the capacitor that blocks any steady current from input to output; its average output will be zero. All three filter transfer functions have the same poles, since they have identical denominators, so they have identical homogeneous solutions, including identical dynamics of energy decay when the input is not being driven. The fourth, filter D, also has the same poles and homogeneous solutions, since it is a parallel combination of filter A and a *straight-through path* without poles; we come back to this one in Section **??**.

Using the quadratic formula to write the poles, the roots of the denominator, gives:

$$p_1, p_2 = \frac{-RC \pm \sqrt{R^2C^2 - 4LC}}{2LC}$$

We can factor filter A into a cascade of two one-pole filters having these two poles. This cascade is a pair of RC filters if the poles are real. But if the poles are complex, then such factors in isolation do not correspond to real (real-valued or realizable) systems or circuits. This is the interesting case here, as it represents resonance. Real resonant circuits have such poles in complex-conjugate pairs, as the quadratic formula suggests, so it takes a second-order filter to represent a real circuit with resonance.

Suppose we vary the resistance R when L and C are fixed, keeping R small enough that the poles are a complex-conjugate pair (that is, the quantity $R^2C^2 - 4LC$ under the square root is negative). In that case, the formula represents points on a circle of radius $1/\sqrt{LC}$ in the complex s plane, as shown in Figure 8.4. We call this radius the *natural frequency* ω_{N} of the circuit. The resistance R determines where the poles are on the circle of that radius, by setting their real part to $-\gamma = -R/(2L)$. The pair of poles can be parameterized in several ways, including these two based on natural frequency and either *decay rate* γ (gamma) or *damping factor* ζ (zeta):

$$p_1, p_1^* = -\gamma \pm i\sqrt{\omega_{\mathrm{N}}^2 - \gamma^2} = \omega_{\mathrm{N}}\left(-\zeta \pm i\sqrt{1 - \zeta^2}\right)$$

The damping factor is a nondimensional measure (between 0 and 1 for the resonant case) of how "lossy" the system is, or of the rate at which the stored energy is dissipated, relative to the system's natural frequency:

$$\zeta = \frac{\gamma}{\omega_{\mathrm{N}}} = \frac{R}{2}\sqrt{\frac{C}{L}}$$

The amplitude of resonance decreases by a factor of e in the time $1/\gamma$, corresponding to $1/\zeta$ radians of oscillation at frequency ω_{N}; the stored energy (amplitude squared) decays by a factor of e in half that time.

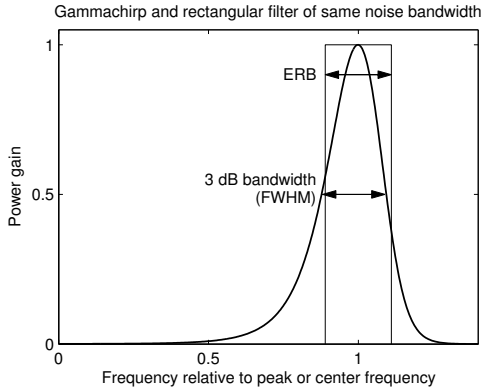
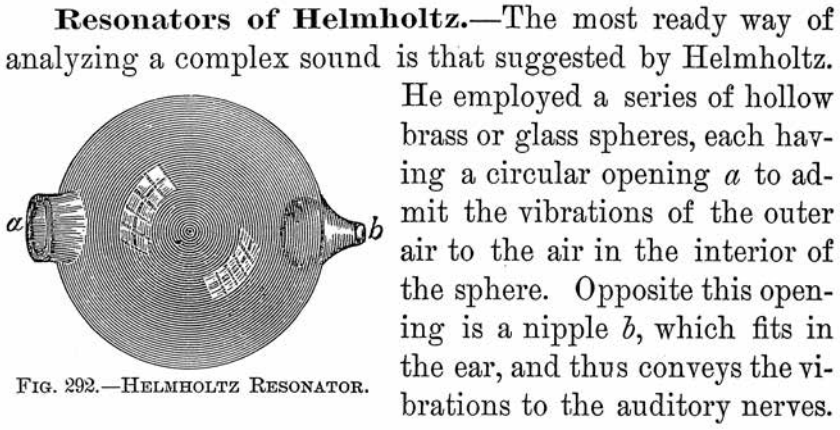


Figure 8.1: The power response (the square of the magnitude response) of an asymmetric bandpass filter, and of a rectangular filter that passes the same total noise power when the input is a white noise (that is, with the same area under the curve). The equivalent rectangular bandwidth (ERB, also known as equivalent noise bandwidth, ENB) of the asymmetric filter is the width of the rectangular filter shown. The ERB of bandpass filters that we encounter in hearing, such as this *gammachirp* filter, is typically slightly greater than the 3-dB (half-power) bandwidth (also known as the full width at half maximum, FWHM), depending on the filter shape. Both the ERB and the FWHM are popular characterizations of a filter's bandwidth, which, as shown here, are not generally equal.



Resonators of Helmholtz.—The most ready way of analyzing a complex sound is that suggested by Helmholtz. He employed a series of hollow brass or glass spheres, each having a circular opening *a* to admit the vibrations of the outer air to the air in the interior of the sphere. Opposite this opening is a nipple *b*, which fits in the ear, and thus conveys the vibrations to the auditory nerves.

FIG. 292.—HELMHOLTZ RESONATOR.

Figure 8.2: Before we had electrical resonators, fluid-mechanical *Helmholtz resonators* were used for sound analysis. Helmholtz developed and used a set of these resonators, tuned to the frequencies of musical notes, to help him "hear out" the sinusoidal components of complex tones. These resonators are tuned by the interaction of the springiness of the air in the globe with the momentum of the mass of air in the neck. Figure from Quackenbos et al. (1891).

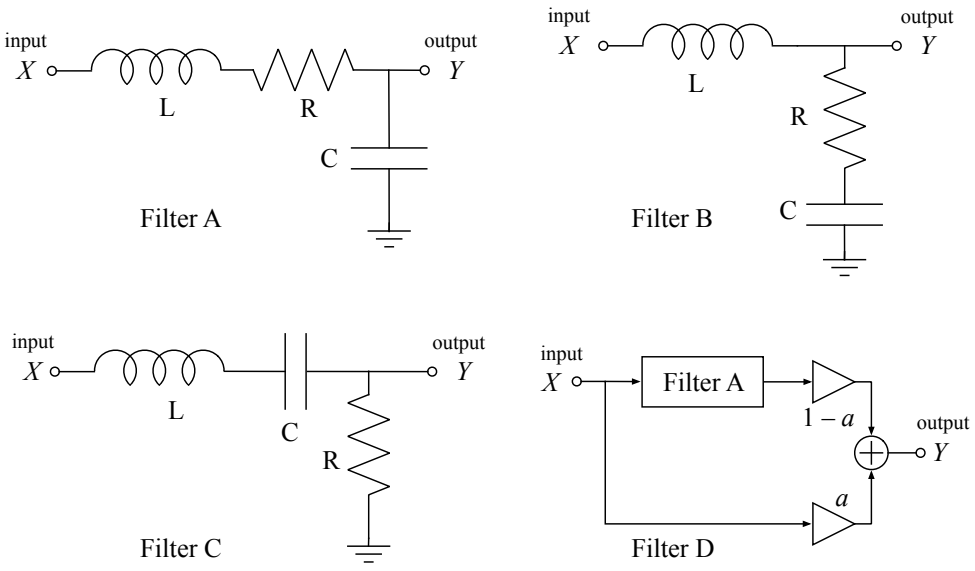


Figure 8.3: Circuit diagrams of four resonant filters. Filter A (top left) is the same as previously shown in Figure 6.9. Filter B (top right) has the resistor, the energy-dissipating element, moved from the series impedance (the impedance of the elements connecting the input to the output in the voltage divider circuit) into the shunt impedance (the impedance of the elements connecting the output to ground). Like Filter A, Filter B has unity gain at DC, since the DC impedance of the capacitor in the shunt leg is infinite. Filter C (bottom left) is a second-order filter with capacitive coupling, or zero response at DC (at zero frequency), since the capacitor is now in series. Filter D (bottom right) uses a pair of adjustable-gain buffer amplifiers (shown as triangles) to mix the output of a filter A circuit with its input. All of these filters have the same pair of poles, and all are relevant as basic building blocks and limiting cases in our study of auditory filters in subsequent chapters.

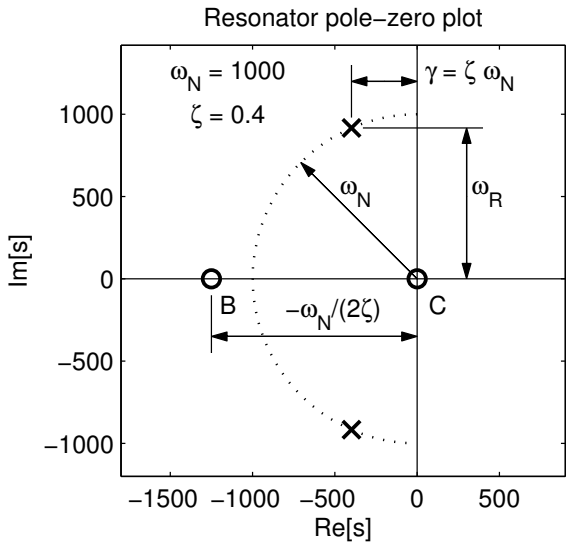


Figure 8.4: The s-plane pole–zero plot for three resonators, filters A, B, and C, illustrated for $\omega_N = 1000$ rad/s and $\zeta = 0.4$. The poles (crosses) at $-400 \pm i\,917$ rad/s (which is $-\gamma \pm i\omega_R$) are the same for the three filters. The dotted semicircle shows the locus of pole positions at radius ω_N for other values of damping factor ζ between 0 and 1: when the damping factor is near zero, the poles are near the imaginary axis, and when it is near 1, the poles approach each other at the negative real axis. Filters B and C each have a zero in addition to the poles, at the positions shown with circles.

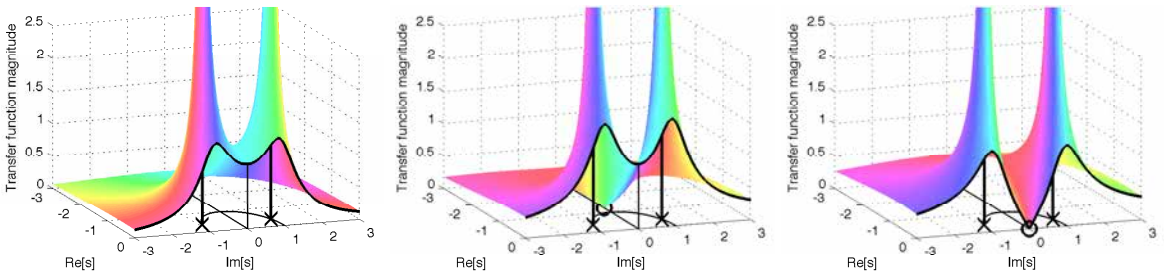


Figure 8.5: The transfer functions of the resonator filters A, B, and C, for natural frequency 1 and damping factor 0.4. See the color plates.

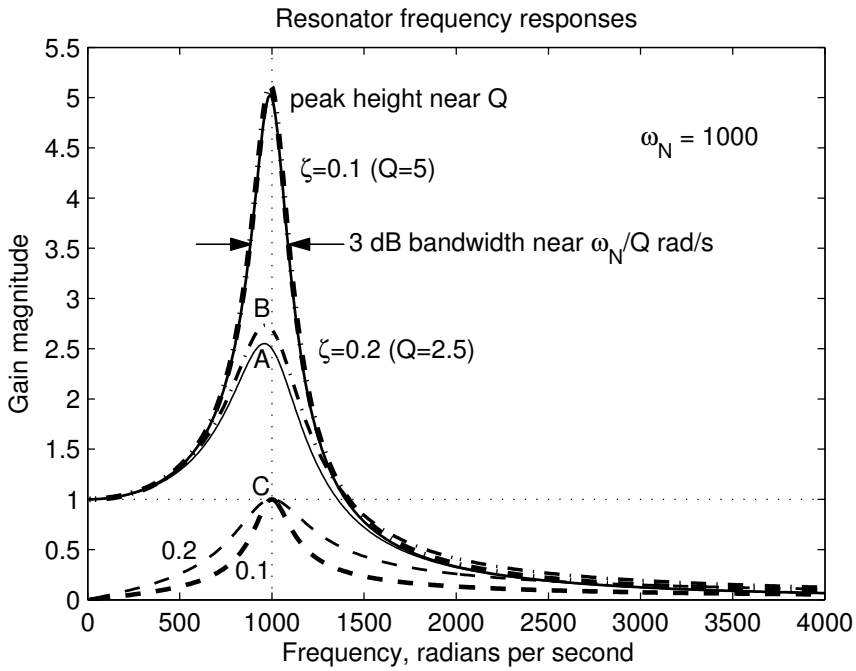


Figure 8.6: The amplitude frequency response of the three resonators, for a natural frequency of 1000 rad/s and damping factors 0.1 and 0.2. Filter A (solid curves) and filter B (dash-dot curves) have unity gain at DC, and higher gain near resonance; filter C (dashed curves) has zero gain at DC, and peaks at unity gain at exactly the natural frequency of the resonance. At low damping, the zero in filter B has little effect, so $\zeta = 0.1$ curves for A and B are very similar.

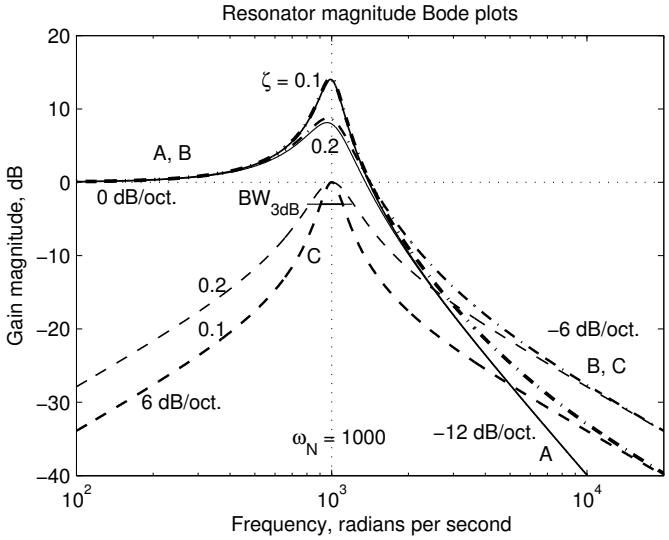


Figure 8.7: Bode plots (dB gain versus logarithmic frequency) of three resonators, for a natural frequency of 1000 rad/s and damping factors 0.1 and 0.2. The Bode plot makes it clear that filter A (solid curves) and filter B (dash-dot curves) are very similar at low frequencies; both have unity gain at DC and higher gain at resonance; in the case of low damping ($\zeta = 0.1$), their peaks overlie each other almost exactly, too. At high frequencies, filter B approaches filter C (dashed curves) which has zero gain at DC and has a peak at unity gain at exactly the natural frequency of the resonance. With twice the damping, the response at 3 dB down is about twice as wide, as marked on the filter C curves.

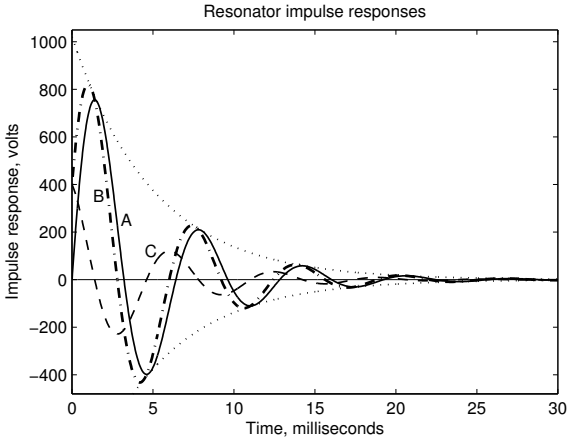


Figure 8.8: Impulse responses of the three resonators, and the exponential envelope (dotted curve) for filter A, for $\zeta = 0.2$. Filter A (solid) and filter B (dash-dot) impulse responses each have an integral of 1, while filter C's impulse response (dashed) integrates to zero. Only filter A shows no step at $t = 0$ (a step has a high-frequency spectrum falling at 6 dB per octave), therefore, only filter A's response falls at 12 dB per octave. The ringing frequency is about $1000/2\pi$ Hz, for a period of about 2π ms.

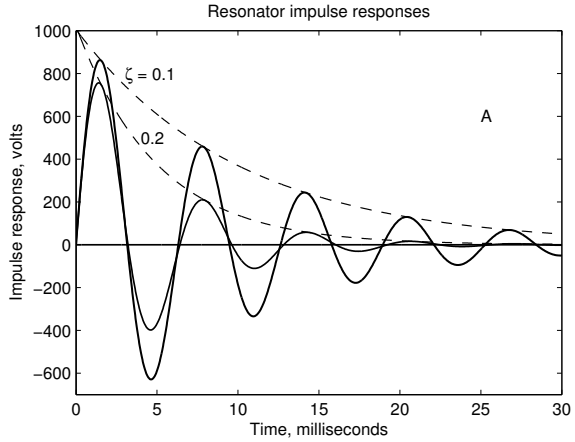


Figure 8.9: Impulse responses of filter A, for $\zeta = 0.1, 0.2$ (solid), and the corresponding different exponentially decaying amplitudes, or envelopes (dashed).

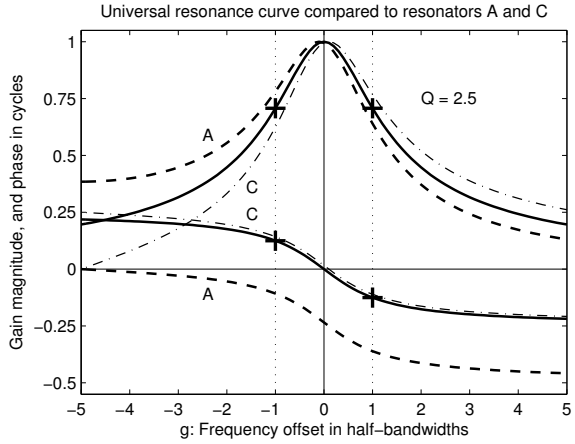


Figure 8.10: Amplitude and phase responses of the universal resonance curve (solid), on a normalized frequency deviation scale that makes it independent of Q, compared to the responses of resonator filters A and C, for $Q = 2.5$ (the A and C curves would be closer together, and the approximation much better, at higher Q than this). The peak gain of filter A has been normalized to 1 to match the other curves, but its phase has not been adjusted to match the condition of zero phase at resonance. The deviation $g = -5$, or 2.5 times the 3 dB bandwidth, corresponds to zero frequency (DC) for the $Q = 2.5$ filters. The 3-dB points of the universal resonance curve, at deviation $g = \pm 1$, gain $\sqrt{2}/2$, and phase ±45 degrees (0.125 cycle) are marked with crosses.

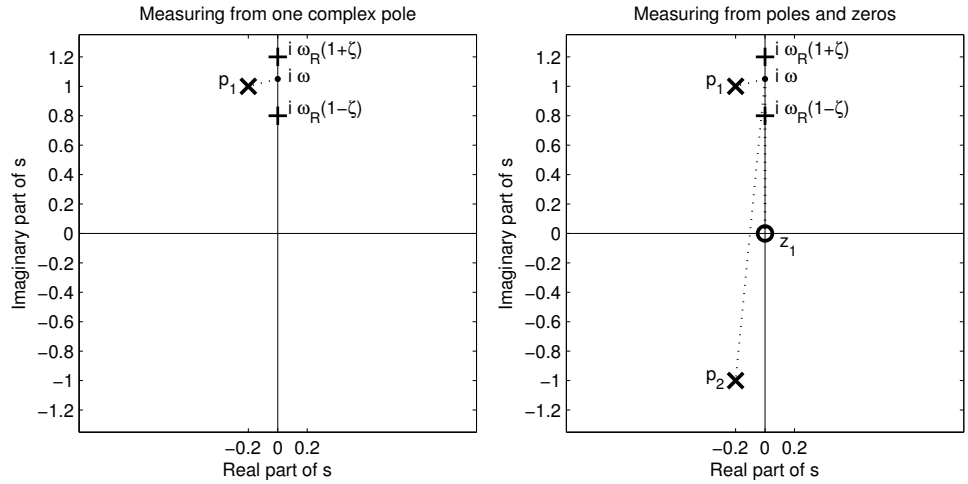


Figure 8.11: Graphical construction of the frequency responses of resonators. Using the left-hand plot with a single complex pole, the gain magnitude of the universal resonance curve, or of the one-pole complex resonator, is inversely proportional to the distance between the pole p_1 and the frequency point $i\omega$ (as a function of ω), here illustrated for ω just above the ringing frequency, ω_R. Frequencies are scaled such that $\omega_R = 1$. The cross marks "+" at $i\omega_R(1 \pm \zeta)$ indicate the points where the distance from the pole to the frequency point increases by $\sqrt{2}$ (the 3-dB points, deviations $g = \pm 1$ in the universal resonance curve, also marked by "+" in Figure 8.10). The real filters have a second pole, p_2; and may have a zero, z_1, as shown in the right plot. The relative distance to the second pole doesn't vary much near the resonance peak, but does move and tip the peak of Filter A a bit when it is included. For filter C, there is also a zero at $s = 0$; the distance to the zero appears in the numerator, so moves and tips the response in the opposite direction. The pole and zero positions shown represent $\zeta = 0.2$ or $Q = 2.5$, corresponding to the curves in Figure 8.10, where DC is at $g = -5$. Pole–zero plots such as these were once commonly used for measurement and calculation of frequency responses, but with modern computers we no longer need such aids; on the other hand, they are still very useful as tools to suggest at a glance the form of the transfer function surfaces of Figure 8.5.

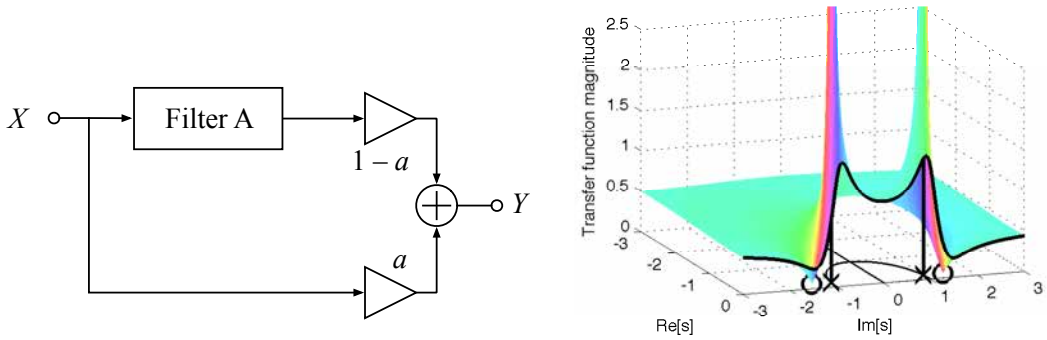


Figure 8.12: Filter D: an asymmetric resonator—schematic and complex transfer function. Adding a straight-through path in parallel to the two-pole resonator of filter A results in a strongly asymmetric peak in the frequency response, involving a complex pair of zeros in addition to the poles inherited from filter A. The ratio of the path gains sets the zero positions. The DC gain is the sum of the path DC gains; as shown, the net DC gain is 1. The illustrated transfer function is for $a = 0.5$ and $\zeta = 0.2$, half the damping of the poles in the illustrations of Figure 8.5, since the zeros near the poles would make the frequency response fairly flat with the higher damping. See the color plates.

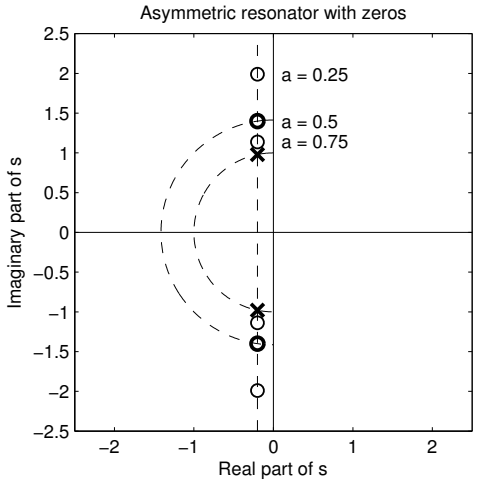


Figure 8.13: The s-plane pole–zero plot for filter D. Using mixing weight $a = 0.5$ as shown in Figure 8.12, the zeros (shown bold) are on a circle of larger radius than that of the poles, by a factor of $\sqrt{2}$, and at the same x coordinate. For other mixing weights, the zeros move to other positions at the same x coordinate; examples for weights $a = 0.25$ and $a = 0.75$ are shown.

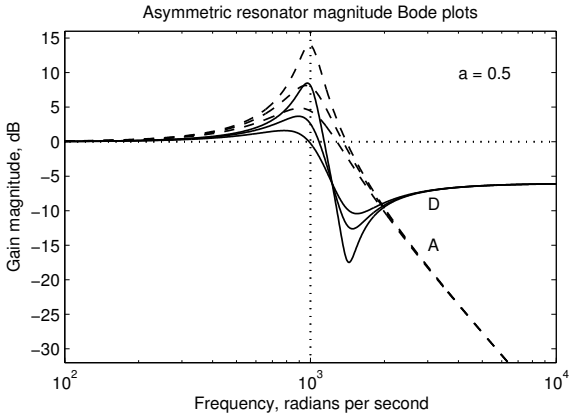


Figure 8.14: Bode plot for the asymmetric resonator (filter D, solid), compared to the all-pole resonator (filter A, dashed), for damping factors 0.1, 0.2, and 0.4, with mixing weight $a = 0.5$. The zeros cause an *anti-resonance* or *notch* about a half octave above the resonance. The high-frequency asymptote is flat, which means the impulse response will contain an impulse.

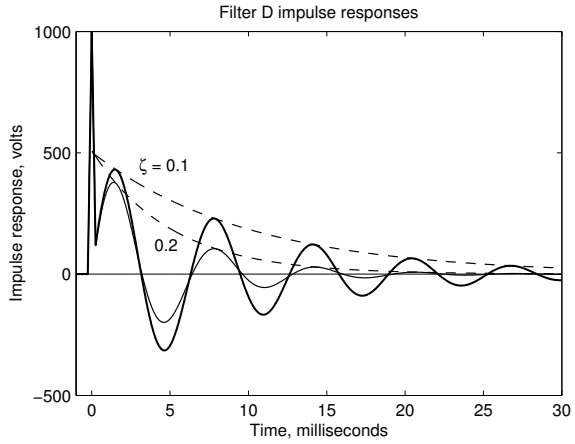


Figure 8.15: The impulse responses of the asymmetric resonator, filter D, contain an impulse of weight a (illustrated here with $a = 0.5$) at $t = 0$, representing the straight-through path. Since an impulse is infinitely tall and narrow, it is approximated in this plot by a triangle, of base width 0.001 s and height 1000, with the correct total area 0.5. Impulse responses for two dampings are shown, with the envelopes (dashed) of the exponentially decaying portions.

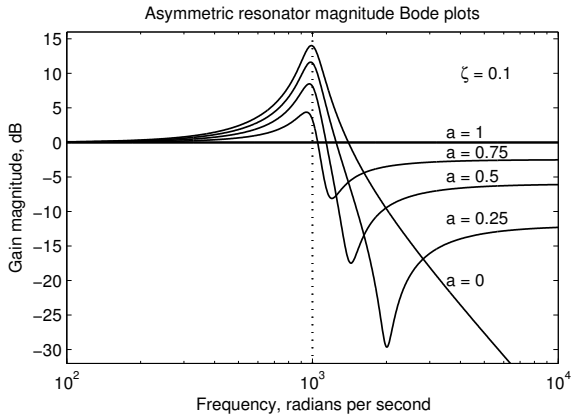


Figure 8.16: Bode plot for the asymmetric resonator with variable mixing ratio. The weight of the "straight through" signal is a, and of the two-pole resonator is $1-a$. This parameter interpolates between a flat response at $a = 1$ and the response of the simple resonator at $a = 0$, resulting in a cancellation dip above the peak.

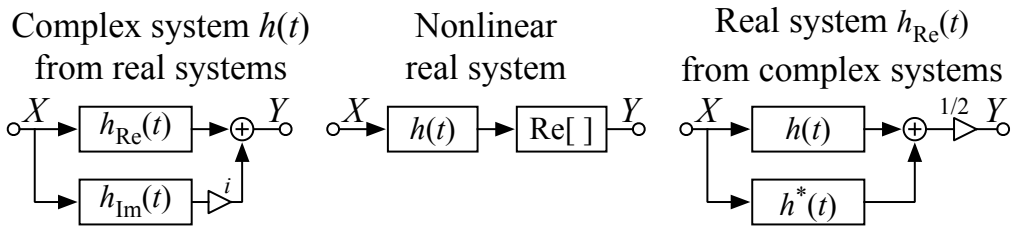


Figure 8.17: Real and complex systems: the complex system on the left, made from two real systems, can be followed by a real-part operator to make a nonlinear system, center; when the input is real, this nonlinear system is equivalent to the linear system on the right. Therefore, we can model the real system on the right via the simpler (lower-order) complex system on the left and a real-part operator as in the middle; or we can implement a discrete-time system such as the complex one on the left, using complex numbers in a computer, and take the real part of its output, as a way to implement the system on the right for real inputs.

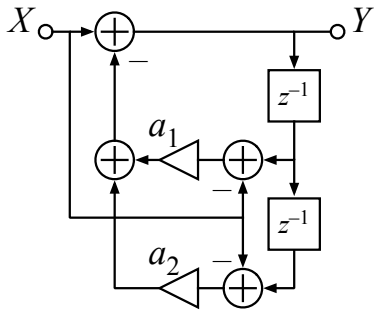


Figure 8.18: A direct-form two-pole filter stage with the input connected as shown here has unity gain at DC. It is evident by inspection that if the input is constant and the output is equal to the input, then the subtracted feedback will be zero, no matter what the coefficient values a_1 and a_2 are, due to the differences being zero where the input is subtracted from the two delayed outputs; therefore output equal to constant input is an equilibrium point, so the DC gain is 1. The transfer function to Y as shown also has two zeros at $z = 0$, and is a minimum-phase transfer function (assuming the poles are inside the unit circle, making it stable); an output taken after one or both of the z^{-1} delay elements has one or two samples of extra delay, canceling one or both zeros, so would not be minimum phase.

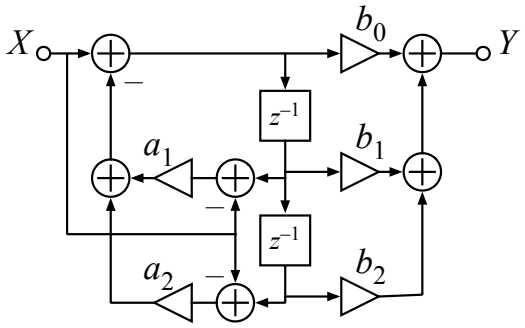


Figure 8.19: The two-pole filter of Figure 8.18 is easily modified to include zeros. In this form, the poles can be moved, while the zeros are held fixed, without the DC gain changing. The relationship of coefficient values to pole and zero locations is as discussed in Section **??**, except that pole-location-dependent numerator of the A factor there is subsumed in the input gain $1 + a_1 + a_2$ provided by the way the input is connected here.

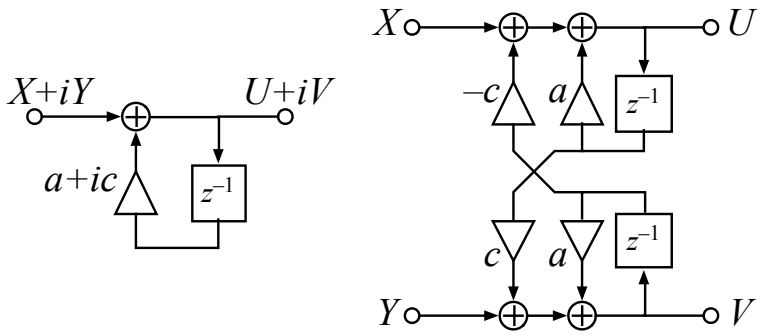


Figure 8.20: A one-pole complex-valued filter (left) is equivalent to the real-valued two-pole filter known as a *coupled-form* filter (right), if X, Y, U, and V are real, as is evident by expanding the complex multiplication into its four real terms. The system on the right can be viewed as a complex system with one pole at $z_p = a+ic$, or as a two-input–two-output linear system with real outputs whenever both inputs are real. The two-input–two-output system can nevertheless be analyzed like other LTI systems in terms of complex inputs, outputs, and eigenfunctions. If the input to the complex system on the left is real, then that system is equivalent to the one on right with the Y port unused (zero imaginary part). For that one-input system, taking only the real-part output, U, gives two poles, at z_p and z_p^*, and one real zero at $z_z = a$ (and a zero at $z = 0$ that keeps it minimum-phase). Similarly, the transfer function from the X input to the real V output has only the pole pair (and the zero at $z = 0$), as can be verified with a little algebra.

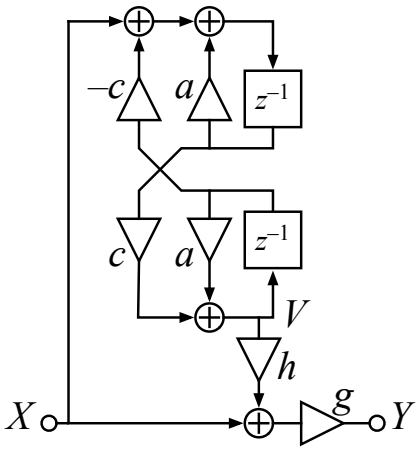


Figure 8.21: A pair of zeros is added to the coupled form by mixing the input with a minimum-phase all-pole filtered version, as in filter D.

Chapter 9

Gammatone and Related Filters

> The form $m(t)$ appears both as the integrand in the definition of the Gamma function $\Gamma(\gamma)$ and as the density function of the Gamma distribution, therefore we propose to use ... the term "Gamma-tone" or "γ-tone."
>
> — "Spectro-temporal receptive fields of auditory neurons...," Aertsen and Johannesma (1980)

Math Connection: Shifting Property of the Laplace Transform

Laplace transforms have many interesting mathematical properties. One that we use in this chapter is the *shifting property*: a shift in the s-plane corresponds to a multiplication by a complex exponential in the time domain. That is, if $X(s)$ is the Laplace transform of $x(t)$, then $X(s-d)$ is the Laplace transform of $x(t)\exp(dt)$, for any real or complex constant d.

If $X(s)$ is a rational function, then the shifted $X(s-d)$ is a rational function with the poles and zeros of $X(s)$ shifted in the s-plane by adding d to their locations. For example, the first-order smoothing transfer function $1/(\tau s + 1)$ shifted to $1/(\tau(s - d) + 1)$ has its pole moved from $-1/\tau$ to $d - 1/\tau$. The corresponding impulse response changes from $\exp(-t/\tau)$ to $\exp(-t/\tau)\exp(dt) = \exp(t(d - 1/\tau))$. If d is real, this shift is equivalent to a change of time constant (and a change of gain), but still gives a first-order filter, with time constant moved from τ to $\tau/(1 - d\tau)$, which is stable if the shift is not too great ($d < 1/\tau$ leaves the pole in the left half plane).

If d is pure imaginary, the pole at $1/(\tau(s - d) + 1)$ is off the real axis. The impulse response retains its original decay rate, and the factor $\exp(dt)$ is oscillatory, so the filter becomes a complex resonator. We encounter both real and imaginary shifts in analyzing gammatone-family filters.

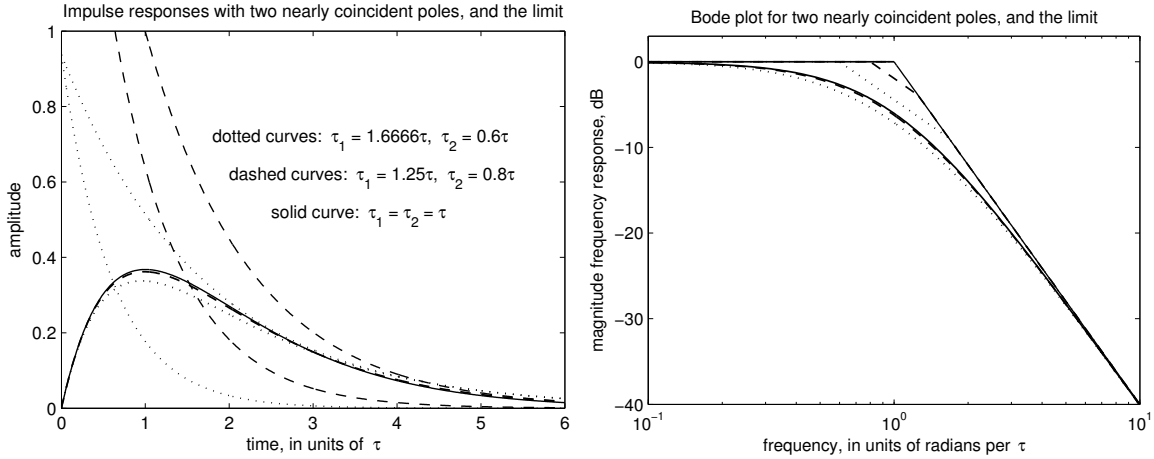


Figure 9.1: In the left panel, the solid curve is the limit of the difference-of-exponential curves that are the impulse responses of smoothing filters with two real poles with a geometric mean time constant τ. The dotted curves show the two scaled exponentials, and their difference, for pole time constants that differ by almost a factor of three. The dashed curves correspond to closer time constants. In the right panel, the same curve styles are used for the Bode plots of the corresponding amplitude frequency responses. With two poles, the filters yield −12 dB/octave (−40 dB/decade) rolloff; a short section of −6 dB/octave near the corner barely affects the shape.

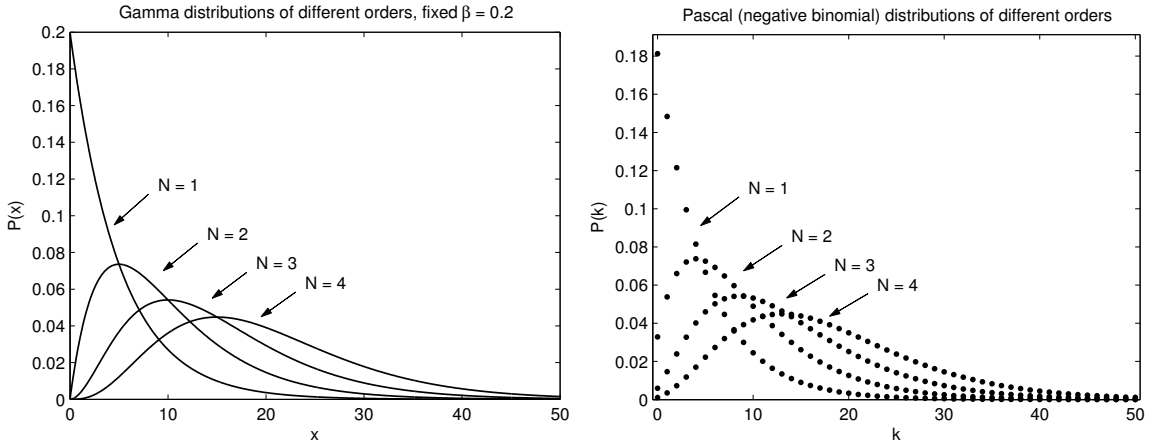


Figure 9.2: On the left are the impulse responses of cascades of N identical continuous-time one-pole smoothing filters, which are also gamma or Erlang distributions, for several values of N. On the right are impulse responses of cascades of N identical discrete-time one-pole smoothing filters with time constants of about 5 samples, which are also Pascal or negative binomial distributions of nonnegative integers.

Statistics Connection: The Gamma Distribution

In statistics, the *probability density function* (PDF) of a continuous random variable is analogous to the impulse response of a continuous-time smoothing filter in linear system theory, in that it has an integral of one and is subject to transforms and convolutions. The analogy is best for smoothing filters with nonnegative impulse responses, as opposed to those that ring, since PDFs are nonnegative everywhere. The PDF of the sum of two independent random variables is the convolution of their individual PDFs, so it is analogous to the impulse response of a cascade of two smoothing filters with the individual PDFs as their impulse responses. Causal impulse responses correspond to PDFs of nonnegative random variables.

It is common in statistics to consider the PDF of a sum of N *independent identically distributed* (i.i.d.) random variables—analogous to the impulse response of a cascade of N identical filters. When the individual PDFs are one-sided exponential distributions of mean $1/\beta$, $P(x) = \beta \exp(-\beta x)$, $x > 0$ (like the impulse response of the one-pole smoothing filter), the resulting PDF is well known as the gamma distribution (also known as the Erlang distribution or Pearson type III distribution). Its formula is:

$$P(x) = \frac{\beta^N}{(N-1)!}\, x^{(N-1)} \exp(-\beta x)$$

The Erlang distribution is applicable only to integer values of N, which is usually all we need; for the more general gamma distribution, the denominator $(N-1)!$ is typically written in terms of the gamma function as $\Gamma(N)$ to apply as well to noninteger N.

Calculations on PDFs are often done via Fourier or Laplace transforms. These transforms, known respectively as *characteristic functions* and *moment-generating functions* of the distributions (Miller and Childers, 2012), are analogous to linear system frequency responses and transfer functions. It is easy to find formulas for the distribution and its characteristic and moment-generating functions in tables. Similar formulas are also found in tables of Fourier and Laplace transforms in linear systems books, typically with slightly different terminology. In statistics, the sign convention for the moment-generating function is different; with parameter t corresponding to $-s$, the moment-generating function of the gamma distribution converges to the left of the poles in t:

$$M(t) = \frac{1}{(1 - t/\beta)^N} \quad \text{for} \quad \mathrm{Re}[t] < \beta$$

This moment-generating function corresponds to the Laplace transform of the impulse response, which converges to the right of the poles in s. Statisticians sometimes write $t < \beta$, as they usually consider only real values of the parameter t, which are enough for generating the moments of the distribution.

Statistics Connection: Scale-Space Smoothing Filters

Not all smoothing filters are analogous to PDFs, since a PDF must be nonnegative. But some important families of smoothing filters, such as those used in scale-space analysis (Witkin, 1983), do respect a nonnegativity constraint. Well-known properties of the PDFs of sums of random variables are then immediately applicable to the problem of successive smoothings at different scales. In particular, the mean of a random variable is analogous to a delay (the low frequency group delay, or how far the center of mass of an impulse response is displaced from $t = 0$); and the standard deviation is analogous to a temporal spread, a measure of the time over which signals are smoothed. For sums of random variables, the means add, and the variances add (the variance being the square of the standard deviation). Therefore, in cascades of smoothing filters, delays add and smoothing time constants combine via the square root of sum of squares; these properties hold even if the filters are not unity gain at DC, so not quite PDFs.

For the one-pole smoothing filter, the delay and temporal spread are both equal to the time constant τ, so a cascade of N of them has a delay of $N\tau$ and a spread of $\sqrt{N}\tau$. The mean and standard deviation of the gamma distribution, in terms of the rate parameter β, are correspondingly N/β and $\sqrt{N}/\beta$.

There are corresponding analogies between discrete-time impulse responses and *probability mass functions* of discrete random variables, with Z-transforms known as *probability generating functions*. When Lindeberg (1990) worked out the details of smoothing filters suitable for discrete scale-space filtering, he did so in the language of generating functions. A cascade of N discrete-time one-pole smoothing filters has an impulse response analogous to a Pascal distribution or negative binomial distribution, the distribution of a sum of N geometrically distributed integers, as illustrated in Figure 9.2.

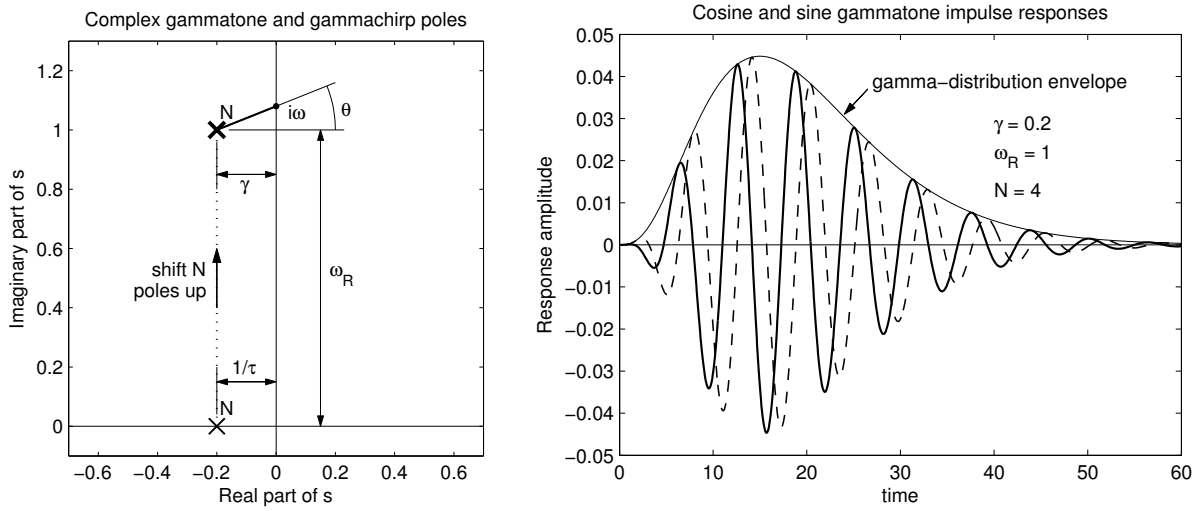


Figure 9.3: The complex gammatone is a system with N coincident poles (plot in the s plane, left), *shifted* from the position of the poles of an N-pole smoothing filter with decay rate γ. As a function of frequency ω, the magnitude frequency response is inversely proportional to the Nth power of the length of the line from the poles to the frequency point $i\omega$, and the phase lag is N times the angle θ shown, going through zero at $\omega = \omega_R$. The real and imaginary parts (right, solid and dashed curves, respectively) of the complex gammatone impulse response are found by multiplying the gamma-distribution envelope (smooth curve), the impulse response of the lowpass, by the oscillatory $\exp(i\omega_R t)$ caused by the shifting of the poles by $i\omega_R$. These real- and imaginary-part curves are real gammatones, of cosine and sine phase respectively. For these plots, the order is $N = 4$ and $\gamma/\omega_R = 0.2$.

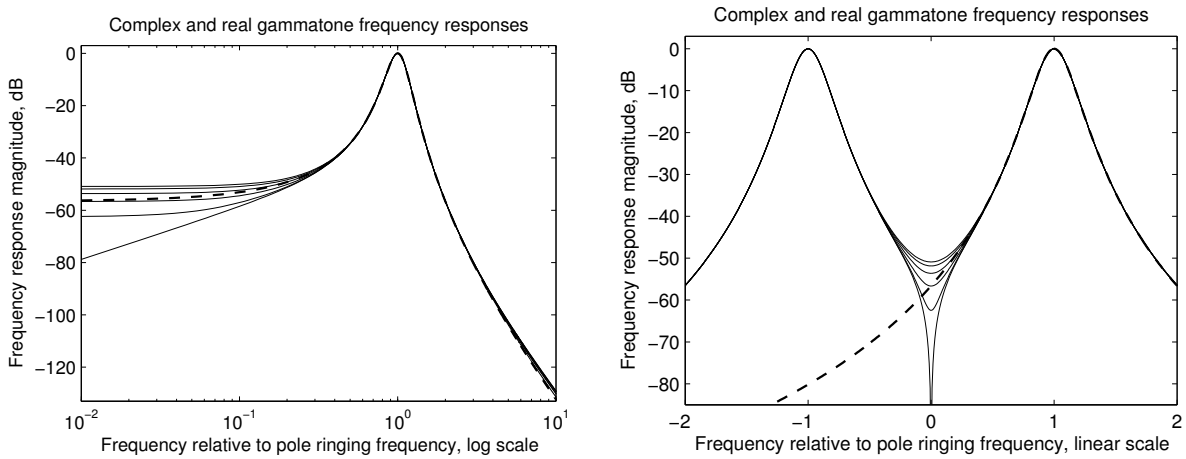


Figure 9.4: The log-magnitude frequency responses of complex (dashed curve) and real (solid curves) gammatones, with $\zeta = 0.2$, $N = 4$, and a variety of gammatone phases, on both log (left) and linear (right) frequency scales (the gains of the real gammatones have been adjusted to match the unity peak gain of the complex gammatone). The complex gammatone response is exactly symmetric about its peak frequency on a linear frequency scale, while real filters must be symmetric about zero frequency, as the right plot shows.

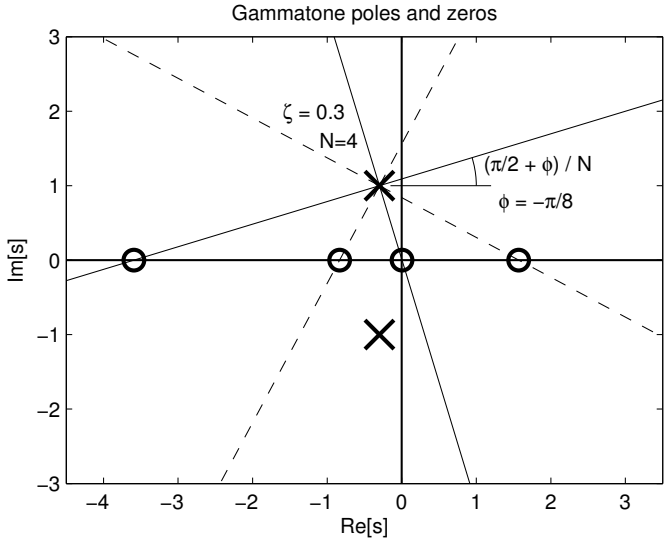


Figure 9.5: The s-plane locations of the real gammatone's N zeros can be constructed from the poles, as described in the text. Solid lines show where the top pole cluster contributes $\pi/2$ radians or 90 degrees of phase, and dashed lines show where it contributes $-\pi/2$ radians or 270 degrees. Where these rays meet the real axis, the other pole cluster provides the opposite phase, and the contributions cancel, making the four zeros shown. In this example, the gammatone phase parameter is $\phi = -\pi/8$; other phases rotate this pattern, moving the zeros along the real axis. Notice also that for these particular parameters, the damping $\zeta = 0.3$ shown puts one of the zeros very close to $s = 0$, making a very low gain in the low-frequency tail response; certain other combinations of phase, damping, and order will do the same. Other special phase values ($\phi = \pm\pi/2$) will move one of the zeros out to infinity, leaving $N - 1$ real zeros.

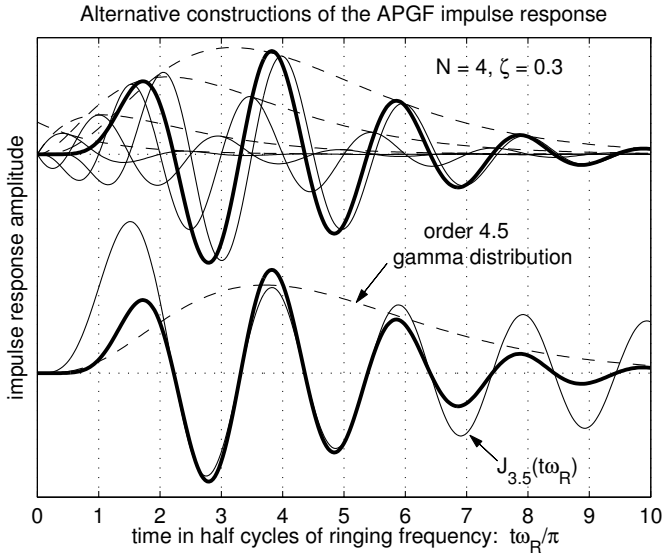


Figure 9.6: The impulse response of the Nth-order all-pole gammatone filter (heavy curves) can be constructed as the sum of gammatones of orders 1 through N, appropriately scaled and phased, as shown on the top; gamma-distribution envelopes are shown dashed. Below, the same impulse response is constructed as the product of a gamma distribution of order parameter $N + 0.5$ times a Bessel function of the first kind, of order $N - 0.5$ (for this illustration, the amplitude of the gamma distribution factor has been scaled down by a factor of 5, and the Bessel function has been scaled up by a factor of 5, for clarity). The damping parameter affects only the exponential time constants of the envelopes.

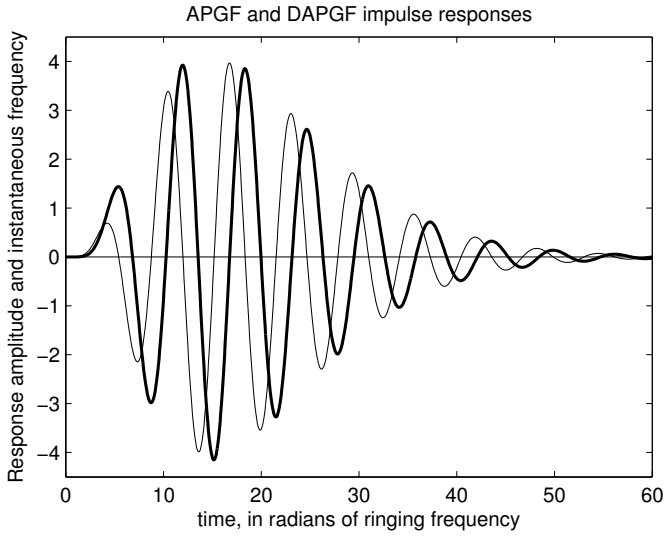


Figure 9.7: Impulse responses of the 4th-order all-pole gammatone filter (APGF, heavy curve) and its derivative, the DAPGF (light curve), with pole damping $\zeta = 0.2$ ($Q = 2.5$, $Q_{3\text{dB}} \approx 5$).

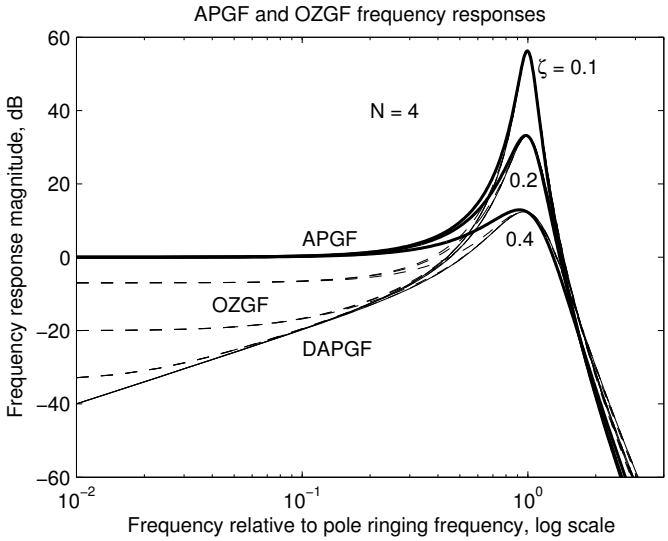


Figure 9.8: Amplitude frequency responses of 4th-order APGF and several OZGFs, including the limiting DAPGF, for three damping factors. Note that the low-frequency tails do not depend on the damping, unlike the case with (real) gammatone filters. The position of the one real zero in the OZGF interpolates between the APGF (zero at infinity) and the DAPGF (zero at $s = 0$), whether the zero is at positive or negative s. The APGF responses are the fourth powers of the resonator A responses shown in Figure 8.7, so the curves are precisely the same as in that figure, but with the dB scale changed by a factor of 4.

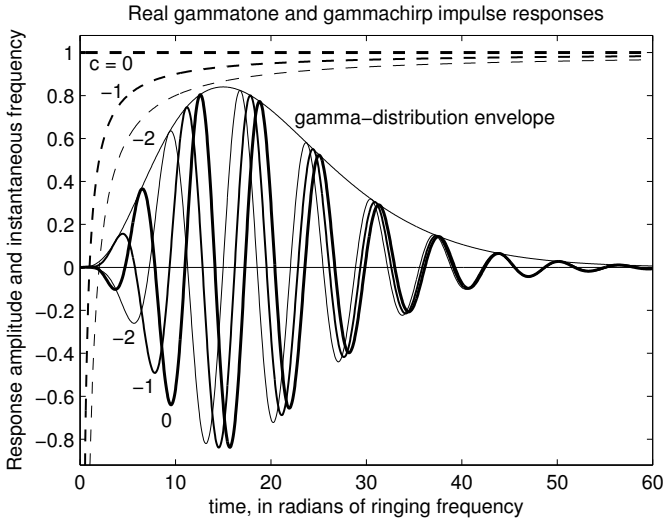


Figure 9.9: Impulse responses (lower curves) and instantaneous frequencies (upper curves) of 4th-order complex gammachirp filters with c values of 0 (gammatone case), -1, and -2. The amplitude scale is arbitrary, the frequency scale is normalized relative to ω_R, and the phases have been chosen to align near one of the later peaks. The pole Q is 2.5, which makes the effective filter $Q_{3\,\mathrm{dB}}$ about 5. The gamma-distribution envelope is also plotted. Notice that with negative c, the zero-crossing times are stretched out more near the beginning, relative to the equally spaced zero crossings of the $c = 0$ gammatone. The zero-crossing times don't change if the envelope is changed by changing the pole Q.

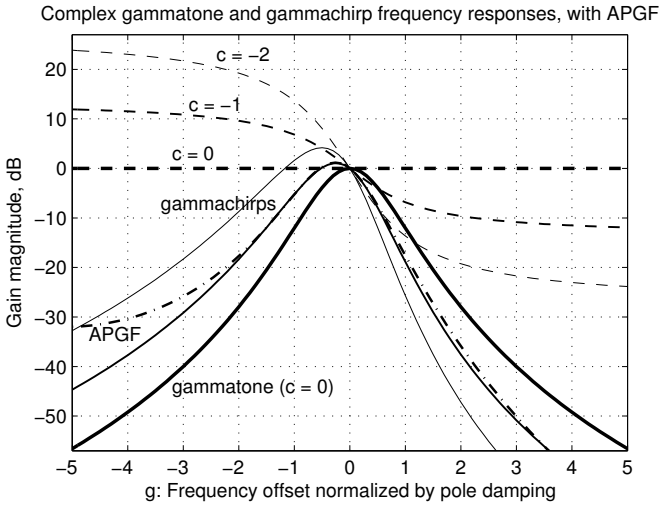


Figure 9.10: Amplitude frequency responses of 4th-order complex gammachirp filters with $c \leq 0$ (solid curves), including the complex gammatone (heavy symmetric curve). The dashed curves show the anti-symmetric log-gain function that converts the gammatone to the gammachirp. An all-pole gammatone filter (APGF) with $\zeta = 0.2$ is also shown (dash-dot line), approximately aligned with the $c = -1$ gammachirp, for comparison of their asymmetries. For real gammatones and gammachirps, the low-frequency tails can be somewhat above or below the tails illustrated for the complex filters.

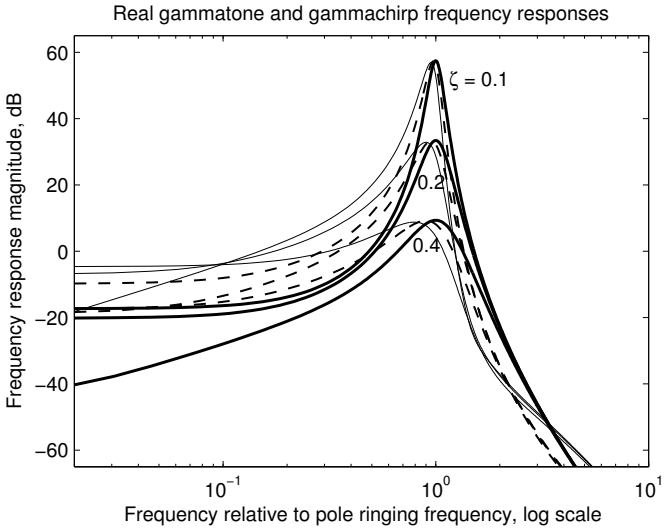


Figure 9.11: The Fourier transforms of the impulse responses are displayed here, for the three impulse responses of the previous figures, plus modifications with half and double damping values. The $c = -1$ gammachirp (dashed curves) has a nonmonotonic damping dependence in the tail, as a zero happens to move close to $s = 0$ when the damping is reduced. In several cases shown, a zero moves to near $s = 0$, pushing the low-frequency tail down; in some of the gammachirp cases, there's also a dip on the high side, due to another spurious zero that comes from interference between the complex gammachirp and its conjugate. These behaviors also have a complicated dependence on the ϕ values used; here we use the same values as in Figure 9.9.

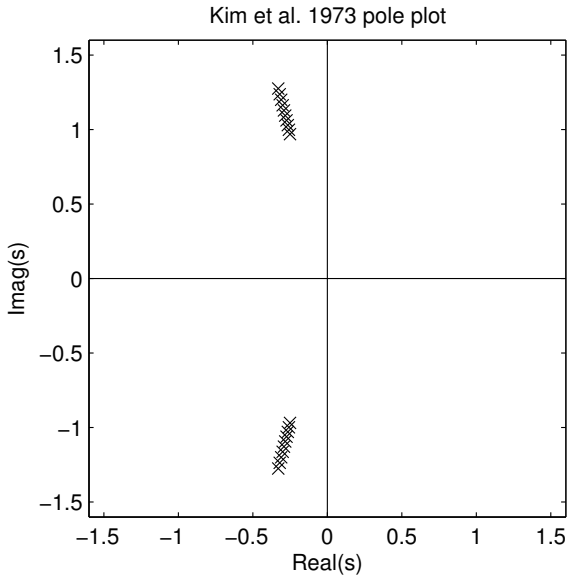


Figure 9.12: The ten pole pairs of Kim, Molnar, and Pfeiffer's model of hydromechanical filtering in the inner ear, normalized to the lowest pole natural frequency.

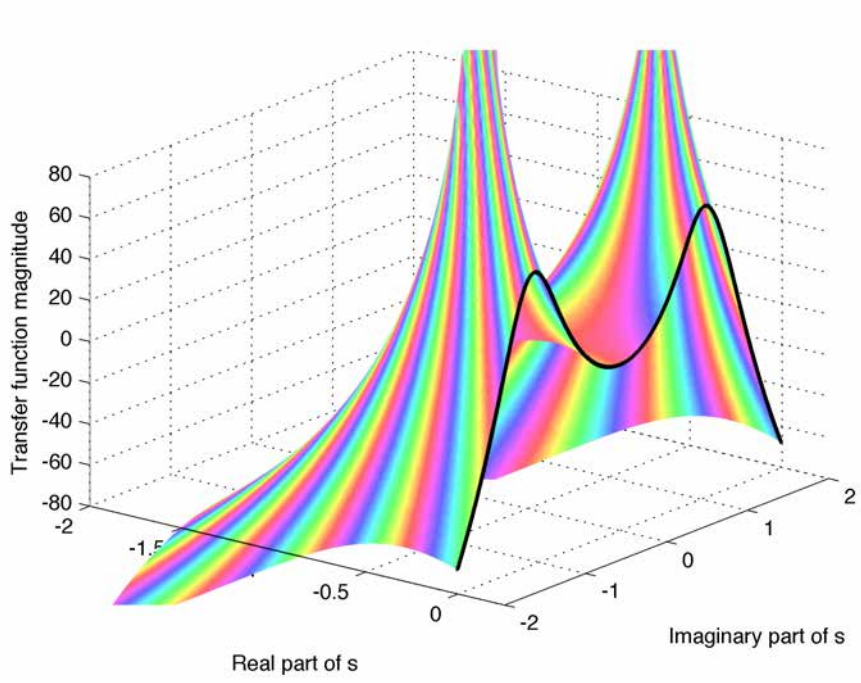


Figure 9.13: The complex transfer function for the Kim et al. filter. The cut line on the imaginary s axis shows the log-magnitude transfer function, with zero frequency in the center. The phase (hue) goes through 10 cycles around each cluster of 10 poles.

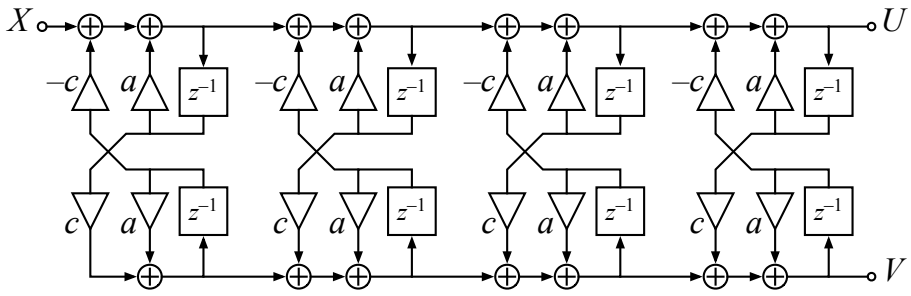


Figure 9.14: A cascade of N complex one-pole filters, with a real input and a real output, would make exactly a real gammatone filter if they were continuous-time filters. With discrete-time filters such as these coupled-form stages, it is an excellent digital approximation to the gammatone, including the zeros that come from taking the real part at the output. The outputs U and V are digital real gammatones of different phases, while $U + iV$ is a complex gammatone. A proportionate change in all of the a and c coefficients moves all the poles together, changing the damping without changing the ringing frequency.

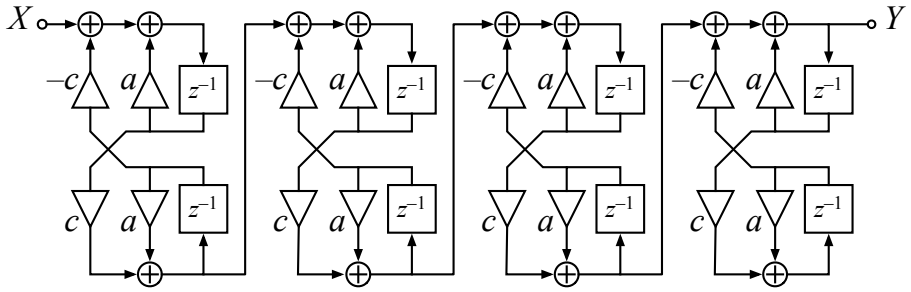


Figure 9.15: A cascade of N identical real two-pole digital filters is the impulse-invariance digital implementation of the all-pole gammatone filter—its impulse response is exactly a sampled version of the continuous-time APGF impulse response. Each coupled-form filter stage of this structure can be interpreted as a one-pole complex filter, with only the imaginary part of the output being used; equivalently, each stage is a two-pole real-valued filter with no zeros (except at $z = 0$, as explained in Section **??**). The same structure, with graduated pole frequencies instead of identical, will implement the linearized Kim et al. (1973) filter or other all-pole filter cascade.

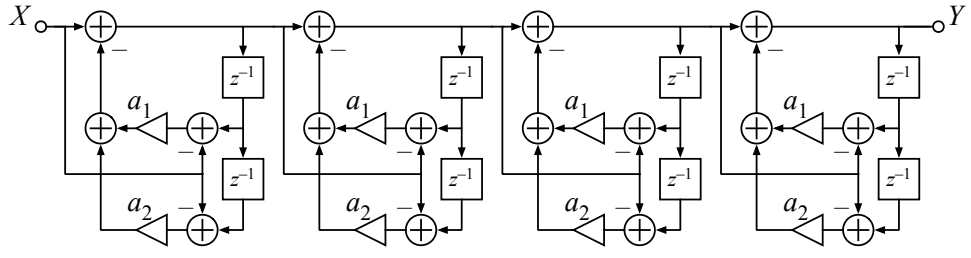


Figure 9.16: A 4th-order APGF constructed as shown, using the modified direct-form two-pole stage of Figure 8.18, always has unity gain at DC, no matter how the poles are moved via the a_1 and a_2 coefficients.

EE Connection: Cascades of Similar or Identical Filters

Transfer functions and impulse responses of coincident-pole filters are not unique to the hearing field, where they are called gammatones; similar functions are known in other fields, including electronics, physics, and statistics.

Papoulis points out that the gamma-distribution impulse response shape arises, very nearly, for cascaded smoothing filters even with noncoincident poles, as the result of a *causal central limit theorem*. Fitting the form of a gamma distribution to such a system's impulse response can yield noninteger N when the poles are not equal, but using the nearest integer N still gives an excellent approximation, effectively representing the system as a cascade of N identical one-pole filters (Papoulis, 1962).

Much of the mathematics of these filters has parallels in the field of statistics. Karl Pearson (1916) developed a number of probability density functions with simple parameterizations and with properties useful in statistics problems. The power-gain frequency response curve of a complex gammachirp filter is exactly a Pearson type IV distribution. Its symmetric special case, the gammatone, is a Pearson type VII distribution, or a Student's t-distribution, which has the Cauchy–Lorentz and Gaussian distributions as its low-order and high-order limits, corresponding to the universal resonance curve and the Gaussian filter.

Cascades of identical resonant filter stages were analyzed for use in wireless television and radio systems in the 1940s (Eaglesfield, 1945; Tucker, 1946), with the observation that their envelope step responses are *incomplete gamma functions* (integrals of the gamma distribution), connecting them, at least tenuously, to the current *gammatone* terminology. Tucker's cascade implementation is shown in Figure 9.17.

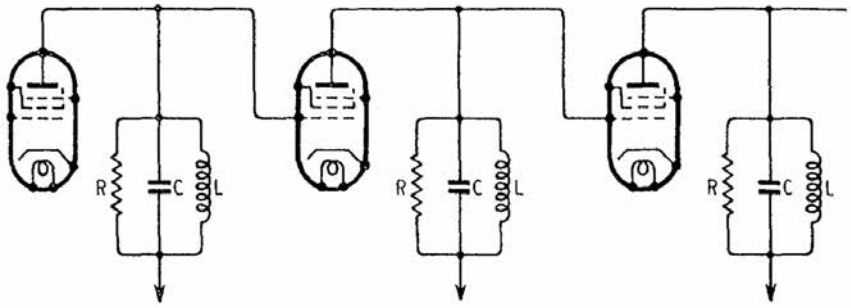


Fig. 1. *Cascade of tuned circuits without mutual coupling, and with blocking capacitors or coupling windings omitted for simplicity.*

Figure 9.17: In his gammatone-like filter cascade, made of parallel-RLC resonators with buffer amplifiers between them, Tucker (1946) represented the buffer amplifiers as vacuum pentodes. This reduction of the mathematical topic to a concrete realization would have made it easy to grasp for engineers of that time, who were familiar with such slightly-abstracted schematics. The pentodes act as transconductors, converting a grid voltage to a plate current, and the filter is the impedance of the parallel circuit, which converts the plate current out of the pentode to a grid voltage into the next pentode. At DC, the inductor shorts the current to ground, making a zero at DC, so the cascaded filters are like our filter C of Chapter 8. [Figure 1 (Tucker, 1946) reproduced by permission of SJP Business Media.]

Chapter 10

Nonlinear Systems

> These results indicate that cochlear mechanics incorporates an essential nonlinearity, so that linear superposition for neighboring spectral components does not apply even at low sound levels.
>
> — "Auditory nonlinearity," J. L. Goldstein (1967)

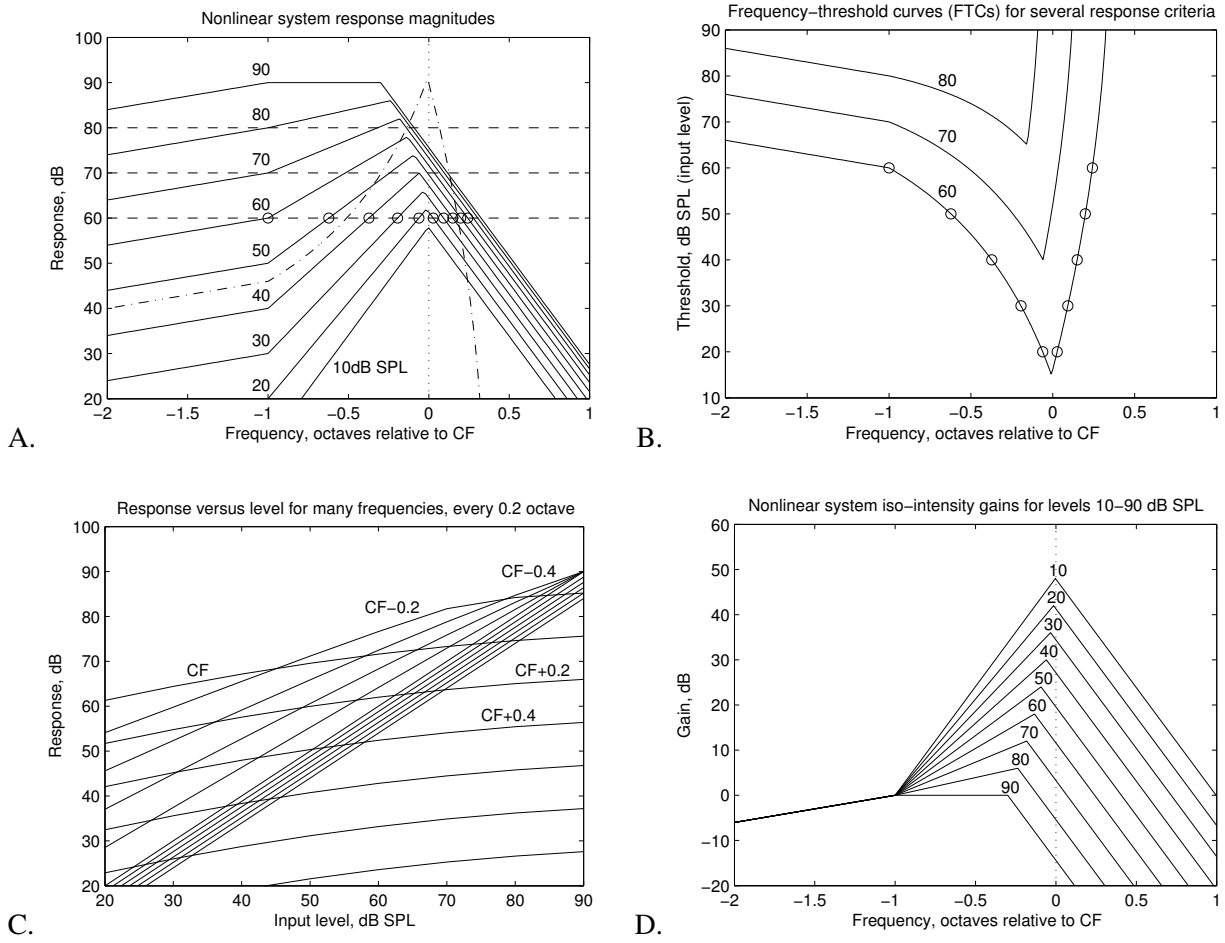


Figure 10.1: Four views of the response of a nonlinear system to sine-wave input at various levels. If we specify a system with the curves in one plot (completely enough), then the other three are generated from it mechanically.

A. Iso-level or iso-intensity curves plot the response versus frequency when the input level is held constant (at levels indicated by parameters on the curves); the dotted vertical line indicates the CF, the frequency of greatest sensitivity at low levels.

B. Iso-response curves, or frequency–threshold curves, plot the input level needed for a given output level, or response threshold, criterion; the 60, 70, and 80 dB response criteria correspond to levels shown by horizontal dotted lines in A, and corresponding points on the 60 dB curve at circled. The dash-dot line in A is a reflection of the 60 dB curve in B, to indicate how much "sharper" the iso-response curve is than the iso-level curves.

C. Iso-frequency curves plot the response level versus input level, for various frequencies. For frequencies near or above CF (CF, CF + 0.2, CF + 0.4), the system is very compressive: the curves have a low slope.

D. Iso-intensity gain curves resemble the magnitude frequency responses of linear systems, except that they are different at different input levels.

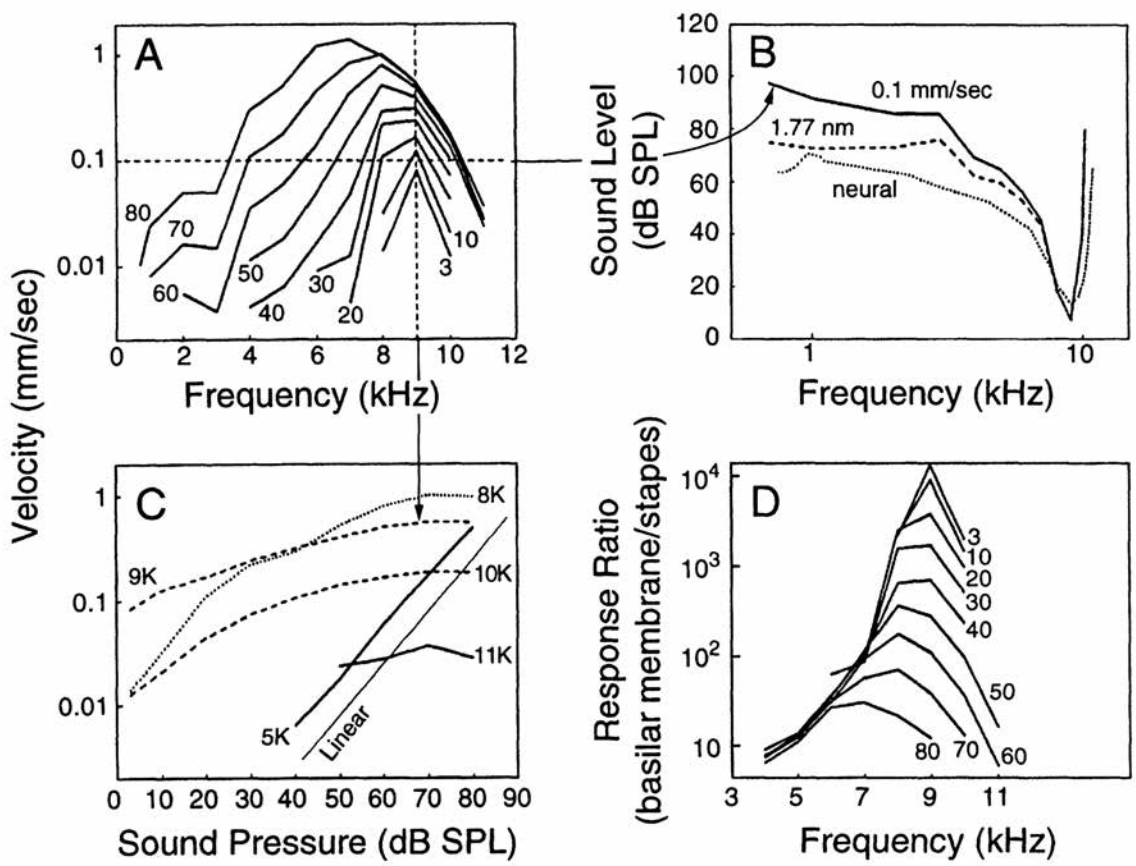


Figure 10.2: Measurements on live cochleas, plotted in the four ways described in Figure 10.1, showing a hugely nonlinear response. The response is quantified in terms of basilar membrane velocity, using laser doppler velocimeter data from Ruggero (1992). Though panel A shows a broad response region at moderate and high levels, panel B shows mechanical frequency–threshold curves (FTCs) at least as sharp as neural FTCs. Panel C shows response approaching linear for frequencies well below CF, and most compressive for frequencies above CF. Panel D shows a gain change at CF of more than 50 dB. [Figure 5.8 (Geisler, 1998) based on Figure 1 (Ruggero, 1992) reproduced with permission of Dan Geisler, Mario Ruggero, and Elsevier Science and Technology Journals.]

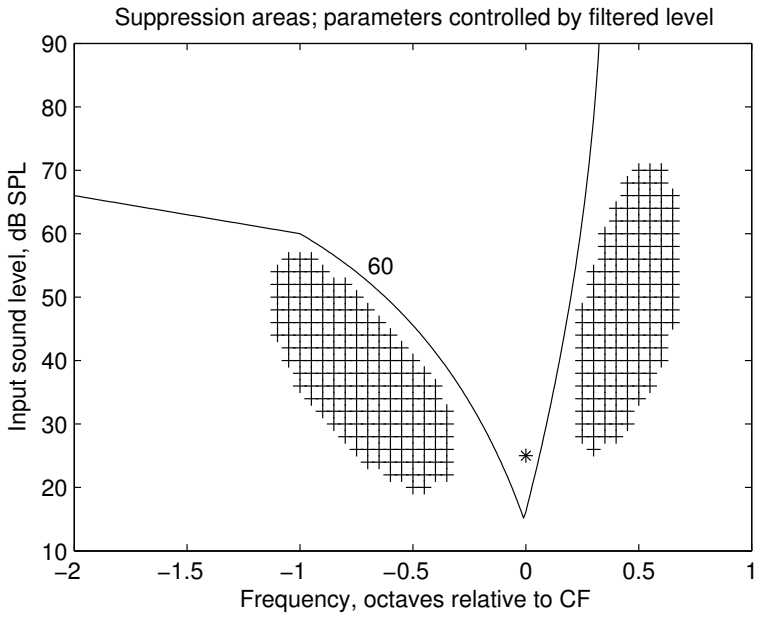


Figure 10.3: Two-tone suppression areas. When a first tone is presented at the frequency and amplitude signified by the '*' (near CF, at a low but detectable level, as shown here relative to the 60 dB response criterion), the system output can actually be reduced by addition of a second tone in the above-CF or the below-CF suppression area. The shape and size of these areas depends on the suppression criterion (here, a 1 dB drop in total output power), and on how the system's parameters depend on the spectrum of the input (here via the power detected in a broad curved filter centered at CF).

Example: AM Radio Demodulation

Consider an amplitude-modulation (AM) radio receiver as an example—a desired sound signal is broadcast as variations in the amplitude of a radio-frequency carrier wave, and we want to get the sound back. Digital radios sometimes work by having an analog continuous-time front end that down-converts radio frequencies around the channel of interest to a fixed *intermediate frequency* (IF), and then sampling the IF signal and doing the rest of the processing digitally. Suppose the IF frequency (the center of the band, to which the carrier frequency is converted) is 30 kHz, and we sample at 100 kHz. For AM radio channels spaced at 10 kHz, the signal bandwidth is 5 kHz (possibly somewhat higher, but use 5 kHz as an example), and there are two sidebands, so we care about frequencies in 25–35 kHz. We can start with a digital bandpass *IF filter* to pass these frequencies of interest and remove everything else. If we then detect the modulated signal's amplitude as a power, by squaring, we generate second-order difference signals, between the carrier and the sideband components, in the frequency range 0–5 kHz, which are the demodulated audio signal components that we want. We also generate sum and double-frequency components in the range 50–70 kHz; these are aliases of frequencies in the range 30–50 kHz, but that doesn't bother us much, as we can remove those alias frequencies by using a lowpass filter, without bothering the low frequencies that we want.

But this square-law detector is not really what we want, as it distorts the audio, which is proportional to amplitude, not power. We could next take a square root (since the powers are always positive, coming from amplitudes modulated up and down from the carrier level that represents zero sound signal). But square roots are expensive. So instead of square-law, we might use an absolute value, or full-wave rectification operator. This nonlinearity generates fourth-order and sixth-order distortion, and so on, in addition to the second-order. So it makes a band at 100–140 kHz, which are aliases of 0–40 kHz, and 150–210 kHz, aliasing to 0–50 kHz, etc., thereby adding some unwanted junk into the audio band of interest. In situations like this, we may want to use a much higher sample rate, or a different demodulation technique, to get a cleaner result. The biggest aliased components are from even multiples of the carrier itself (60, 120, 180, 240, 300 kHz, aliasing to 40, 20, 20, 40, 0 kHz) which are fixed and easy to keep outside the band of interest up to eighth order, so it's not terribly bad. But if the carrier is off from 30 kHz by just 10 Hz, the tenth-order nonlinear component will alias to 100 Hz, and will make an audible hum.

These kinds of issues, particularly *intermodulation product frequencies* between sample rates and carriers, are analyzed carefully in radio design, and are sometimes relevant in machine hearing as well, especially if a high-quality reconstructed sound is needed, as in a hearing aid. Wherever a strong nonlinearity is used, it's a good idea to consider where harmonics of a signal will alias to. Softer nonlinearities such as time-varying gains, if carefully applied, are less likely to cause audible distortion and aliasing, since they generate much smaller distortion product amplitudes.

Chapter 11

Automatic Gain Control

> In recent years, devices for the automatic control of gain have increased in importance in various areas of amplifier technology. One class of such devices is based on the following principle: a portion of the output signal current of a valve amplifier is extracted, amplified and fed to a rectifier; the resulting rectified signal voltage is then used to vary the grid voltage of an amplifier valve. In this manner an increase in output power leads to a reduction in gain.
>
> — "On the Dynamics of Automatic Gain Controllers," Karl Küpfmüller (1928)

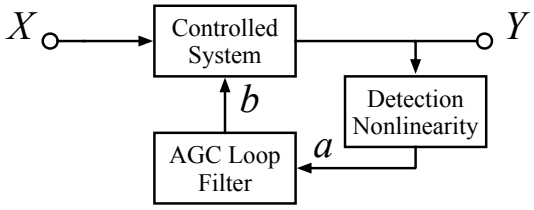


Figure 11.1: A system with automatic gain control (AGC). The loop filter output, b, can control any parameter of the controlled system that affects its gain. The loop filter, in combination with the properties of the controlled system and the detector, determines the dynamics of the response to a change in input level.

On "Level"

The concept of *level*, frequently found as *intensity level* or *loudness level*, is usually expressed on a logarithmic scale, in decibels. In the 1960s, standards organizations actually began to *define* level to be the logarithm of the ratio of an intensity to a reference intensity, so that they could cast the decibel as a unit of level, making the dB behave more like a conventional unit than as a logarithm; for example, ANSI (1960) defines *level*: "In acoustics, the level of a quantity is the logarithm of the ratio of that quantity to a reference quantity of the same kind. The base of the logarithm, the reference quantity, and the *kind* of level must be specified." Most engineers have not been taught this definition of level, though, and use level more informally as a general notion of a measurement of how big a signal is, whether they represent it logarithmically or not.

In an automatic gain control loop, we typically feed back some measurement of output level to control the system gain. Some treatments in the literature assume that output level is measured logarithmically, but this model is difficult to get to work right at very low signal levels, so is more often avoided.

Wheeler (1928) speaks of "maintaining the desired signal level in the detector or rectifier," which is much like how we treat it here. That is, we let the detection nonlinearity (the rectifier) provide a signal that we take to represent level, with no prejudice about whether it is proportional to power, or amplitude, or log power, or something else.

In a real AGC system with signals representing sounds, level is a derived quantity, or even an abstraction, of what the system adapts to. A detector or rectifier produces a derived signal whose short-time average can be taken as level. But the rectified signal—whether positive part or absolute value—also contains fine temporal structure that is not part of what we call level. There may be no clean separation between the frequencies or time scales of level fluctuations and the frequencies or time scales of fine structure. But we can pretend.

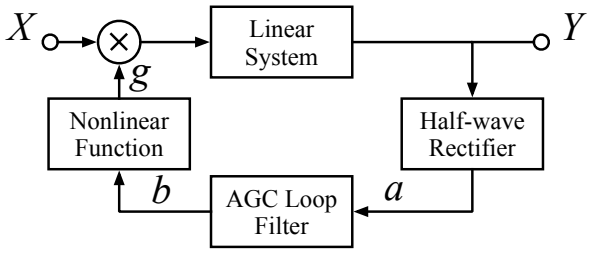


Figure 11.2: A loop with multiplicative gain is a tractable special case of the model of Figure 11.1. The "controlled system" is expanded as a linear system with a variable gain at its input, the gain g being a decreasing nonlinear function of the control parameter b. In this diagram, we also specialize the detection nonlinearity to a half-wave rectifier. The system can be approximately analyzed by linearizing the loop, treating it as a linear system with signal levels, as opposed to the signals X and Y themselves, as its input and output variables.

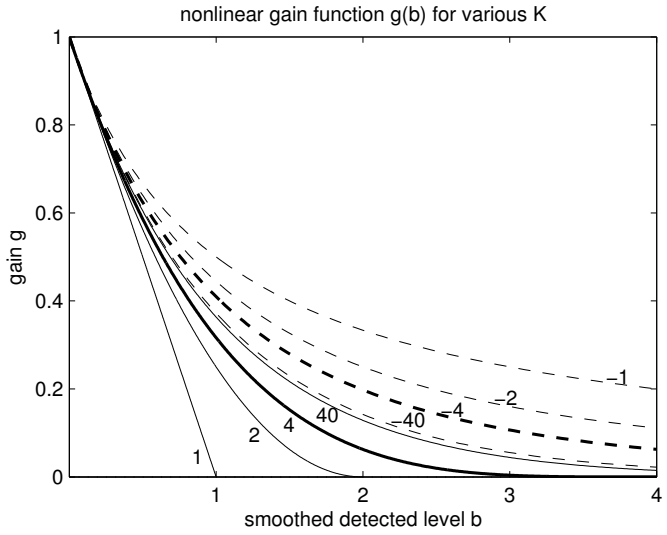


Figure 11.3: The family of nonlinear gain functions $g(b)$ used in the AGC analysis, for various values of the K parameter. All functions are near $1 - b$ at low levels, but their high-level behavior depends on the parameter. Negative K values are represented with dashed curves; $|K| = 4$ curves are bold, representing typical system choices. For high $|K|$, the functions approach $g(b) = \exp(-b)$.

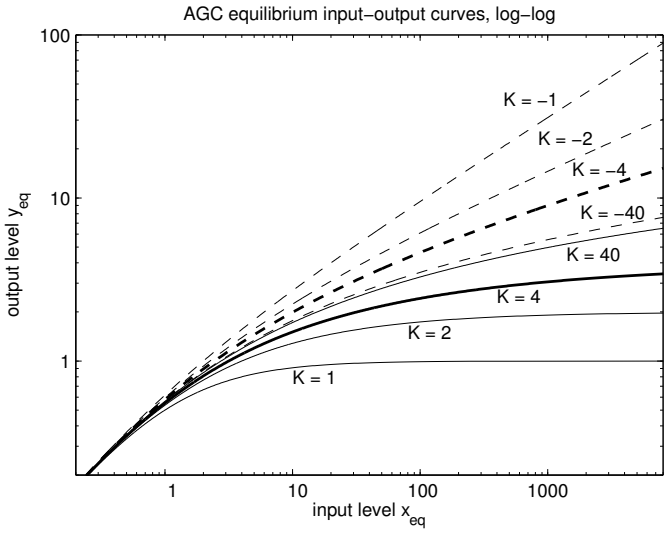


Figure 11.4: The input–output level curves for an AGC system with nonlinear gain function $g(b) = (1-b/K)^K$. Curve styles are as in Figure 11.3. At low levels, all of the curves are approximately linear; at high levels they compress to varying degrees. Negative values of K tend toward power law (root) compression, with straight asymptotes of slope $1/(1 - K)$ in this log–log plot, while positive values of K cause compression toward a constant output level K (horizontal asymptotes). For high $|K|$, at the divide between positive and negative K, the high-level response approaches logarithmic compression (no straight-line asymptote).

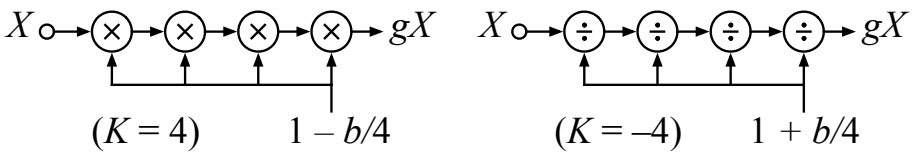


Figure 11.5: Variable-gain amplifiers (multipliers or dividers) can be cascaded to give a gain control over a fairly wide dynamic range, compared to the dynamic range of their control variable. Here the multipliers (left) correspond to our nonlinear gain function with $K = 4$, and the dividers (right) to $K = -4$.

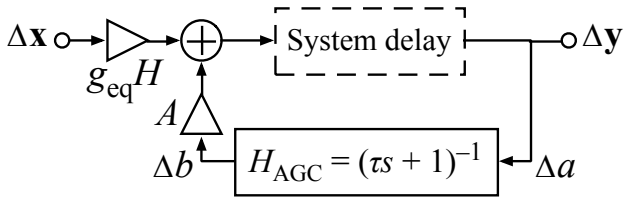


Figure 11.7: This linear system is a small-signal model of the AGC's dynamic response to changes in the input level about an equilibrium condition. The inputs and outputs of this linear model are the perturbations of input and output levels, $\Delta\mathbf{x}$ and $\Delta\mathbf{y}$. The dashed box represents the delay of the controlled system to level fluctuations; we ignore it initially, or assume that the delay is negligible compared to the time constants of the loop. The negative gain A expresses the slope with which changes in the control parameter b affect the output level at this equilibrium. For definiteness in subsequent analysis, a first-order smoothing filter is used here as the AGC loop filter.

History Connection: Wheeler's Automatic Volume Control

Harold A. Wheeler (1928) of the Hazeltine Corporation, a manufacturer of radios, analyzed a cascade of several variable-gain amplifiers in the automatic volume control (AVC) of an AM broadcast radio receiver. His resulting input–output level curves are as plotted in Figure 11.6.

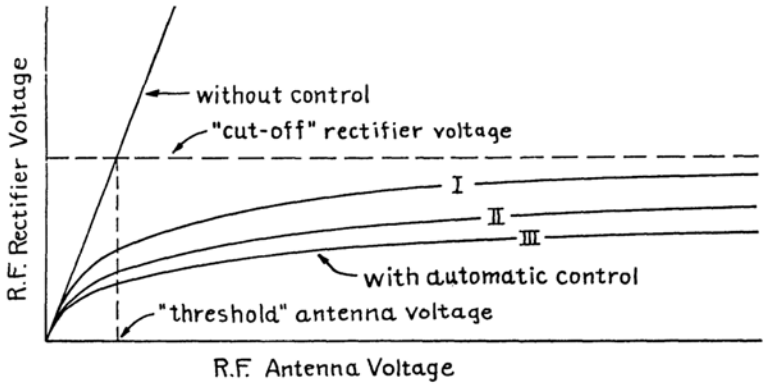


Figure 11.6: The input–output level curves of Wheeler (1928), for systems with one, two, or three cascaded variable-gain amplifiers, corresponding approximately to our model of Figure 11.5 with K values of 1 to 3. The radio-frequency output level ("R. F. Rectifier Voltage") is almost independent of input level ("R. F. Antenna Voltage") when the input level is well above the "threshold." [Figure 3 (Wheeler, 1928) reproduced with permission of the IEEE.]

Our family of nonlinear gain functions is a generalization of this idea of a cascade of several variable-gain amplifiers. The number of cascaded variable-gain multipliers is $|K|$. For example, four stages may multiply four times by $1 - b/4$, or divide four times by $1 + b/4$, as illustrated in Figure 11.5. The nonlinear gain functions are chosen to have $g(0) = 1$, and less gain otherwise (because b is nonnegative), with the initial rate of gain reduction being $dg/db = -1$, independent of the parameter K.

Wheeler's amplifier stages—vacuum pentodes—had the property that their gain could be controlled by a grid voltage, with the gain decreasing approximately linearly toward zero at a cutoff voltage. Thus with K stages his nonlinear gain function was something like $g(b) = (1 - b)^K$ in our terminology—like our $g(b)$ with positive K, and b representing the output level fed back as grid control voltage, but without the divisor of K that we use to keep the initial rate of gain reduction independent of K.

For positive K, an input level growing without bound reduces the gain $g(b)$ toward a limit of zero as the output level approaches K is our formulation, or as the output level approaches 1 in Wheeler's. Therefore his plot, reproduced in Figure 11.6, has all the curves approaching the same "cut-off" level, while our positive-K curves in Figure 11.4 approach different levels. Besides this subtle difference, our use of the divisor K also lets us generalize to negative K values, which do not lead to a finite output level bound, and which better model the compression in cochlear models.

Wheeler notes that in light of certain limitations with the amplifier tubes, "it is undesirable to reduce the amplification ratio per stage below about 1/10 of its normal value. When controlling several tubes, these limitations become unimportant." With three amplifier tubes cascaded, he therefore achieves a gain range of about a factor of 1000, or 60 dB. In cochlear models, we get similar benefits from distributing large gain changes over multiple cascaded filter stages.

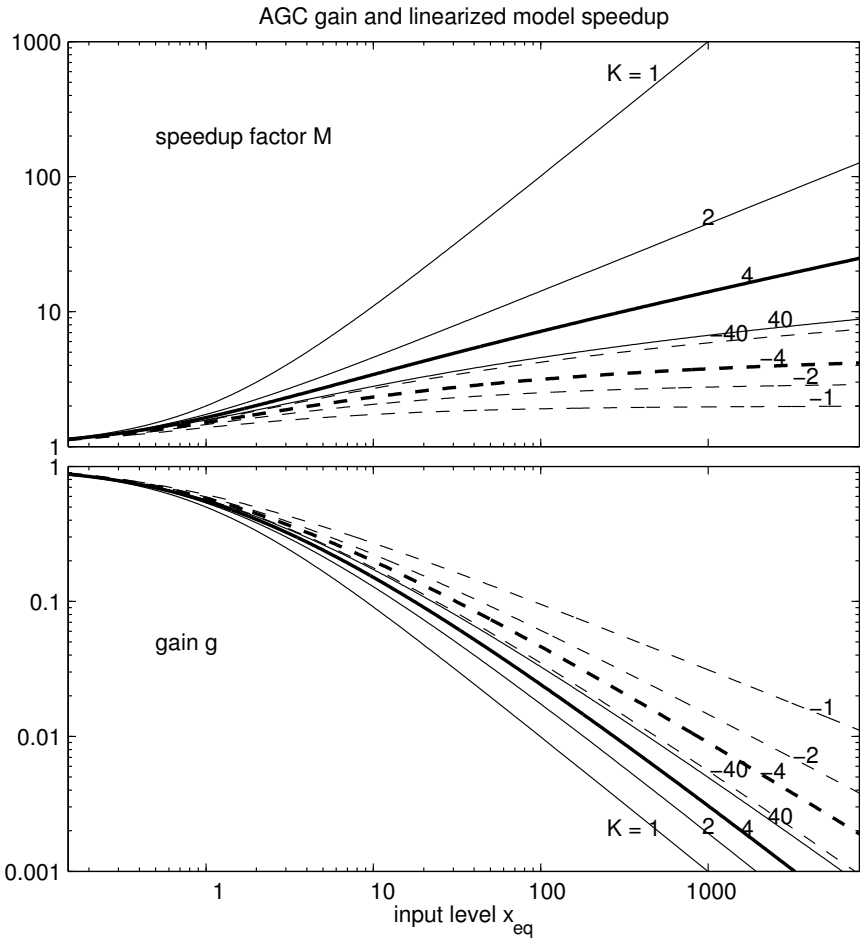


Figure 11.8: As the equilibrium input level increases, the gain g (lower panel) decreases, and the linearized closed-loop system gets faster by a speedup factor M (upper panel), for the example AGC loop with $H = 1$ (curve styles and K values as in Figure 11.3). Both sets of bold curves, $|K| = 4$, show moderate speedups. Lower positive K values lead to very low gains and corresponding very high loop speedups—a challenge to stability in the real system that includes additional delay in the loop. The cochlear models that we build are not quite as simple as this model, but their AGC behavior is approximately modeled by $K = -4$ (heavy dashed curves), with well-controlled loop time constants and compressed but not tightly controlled output level.

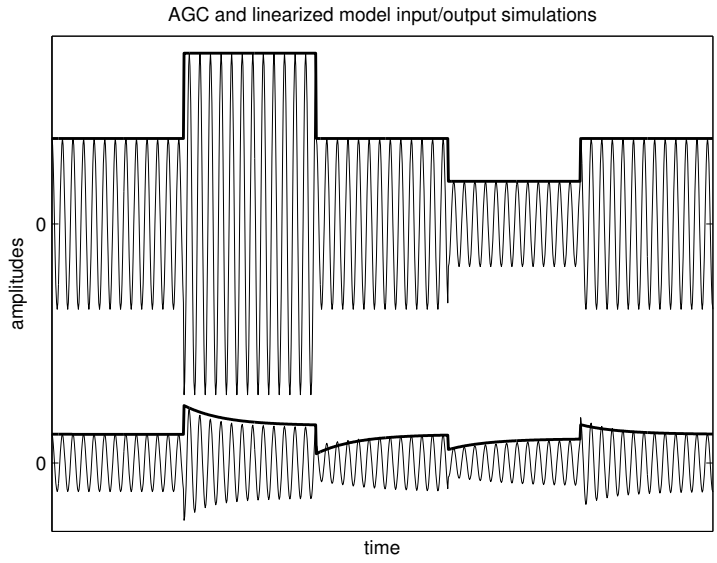


Figure 11.9: Inputs (top) and outputs (bottom) of a simulation of the multiplicative AGC system of Figure 11.2, and its small-signal linear model (see Figure 11.10 for parameters), as the input amplitude is stepped up and down by factors of 2. The amplitude-modulated sinusoids are the inputs and outputs of the nonlinear AGC simulation, while the bold curves are the inputs and outputs of the linearized model added to the equilibrium levels, scaled up by a factor of π for comparison to the peaks of the sinusoids. The fact that the output curves do not perfectly match is an indication that the linearization about the equilibrium condition is not a perfect model of the nonlinear dynamics. The equilibrium gain from input to output is $g_{\text{eq}}H = 0.34$, which is why the output curves are so much smaller than the input curves.

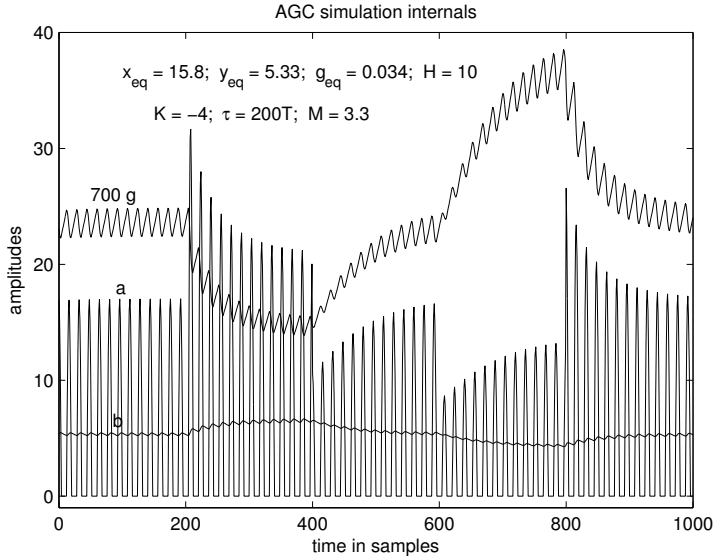


Figure 11.10: The internal signals a, b, and g of the simulation shown in Figure 11.9, of the AGC system of Figure 11.2, show how the half-wave-rectified output is smoothed, imperfectly, resulting in a gain g with considerable ripple. The parameters of the simulation are indicated on the plot; the g signal has been scaled up to reduce confusion. Notice that the speedup factor (in the linear approximation) is a moderate $M = 3.3$, even though the equilibrium gain reduction is fairly severe at $g_{\text{eq}} = 0.034$ (about −30 dB relative to the gain at low level).

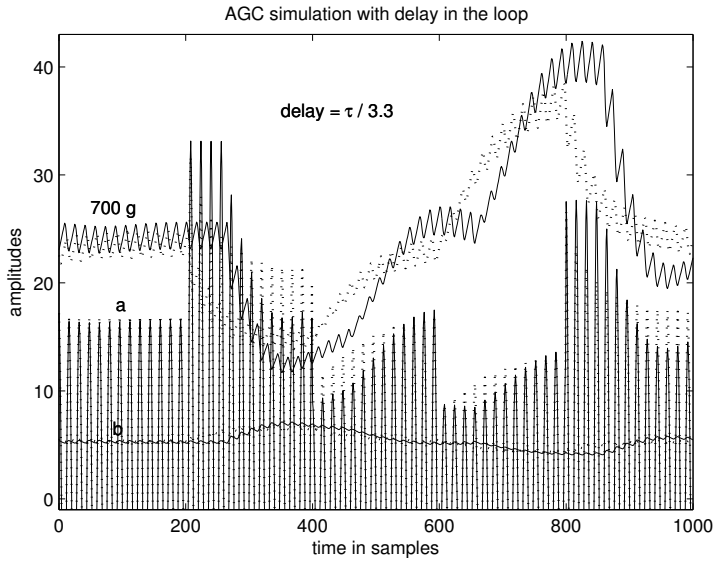


Figure 11.11: The simulation shown in Figure 11.10 has been repeated with a delay equal to $\tau/3.3$ in the forward path. Solid curves show the new simulation, while the dotted curves are copies of the response from the previous figure, from the simulated system without delay. This delay corresponds to τ/M using $M = 3.3$ from the linearization about the initial equilibrium. With the delay, the smoothed output estimate b is slow to react, and the gain adjustment overshoots. The output level excursions are slightly greater. With this much delay, even though it is a fairly small fraction of the loop filter's time constant, a larger speedup factor M would lead to considerably worse behavior.

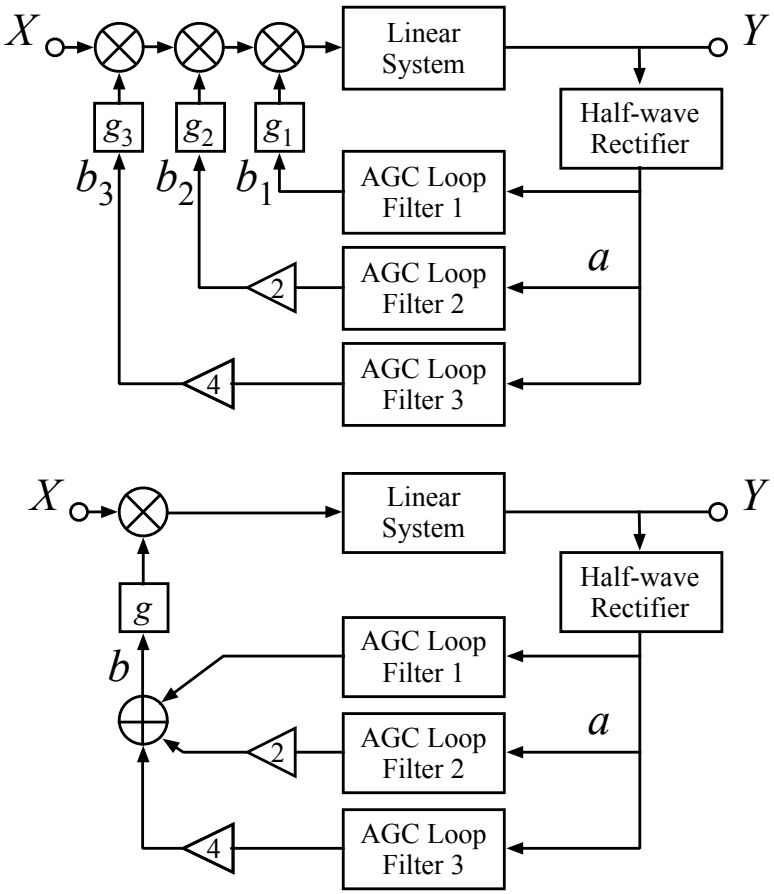


Figure 11.12: An AGC with multiple filters in the feedback path, each controlling a separate gain through separate nonlinear gain functions (top), and an alternative way to configure a multi-time-constant loop with the outputs of the loop filters summed to control a single gain (bottom). If the outermost loop is very slow, it can reduce slow level variations, leaving the more inner loops to deal with faster and smaller level variations. If the nonlinear gain functions g_i are approximately exponential, even with moderate $|K|$, the effects of the two schemes will be similar.

Chapter 12

Waves in Distributed Systems

> The movement of waves down the basilar membrane is analogous to the propagation of light waves in a medium of continuously changing index of refraction. While the velocity of light varies as it travels through the substance, substantial reflections will not occur as long as the index of refraction changes slowly enough.
>
> — "The cochlear compromise," Zweig, Lipes, and Pierce (1976)

Example: Delay Lines and Moving-Average Filters

A *delay line* is a linear system whose output is a copy of its input from an earlier time, with delay T:

$$y(t) = x(t - T)$$

The transfer function is not representable as a rational function, and has no poles or zeros:

$$H(s) = \frac{Y(s)}{X(s}) = \exp(-sT)$$

and the frequency response is just a phase lag proportional to frequency:

$$H(i\omega) = \exp(-i\omega T)$$

In the 1940s, J. Presper Eckert built delay lines for use in a radar moving target indicator, and then as memory systems for early digital computers (Galison, 1997). The delay lines in the UNIVAC used acoustic compression waves in cylindrical tubes of mercury, each storing 720 distinguishable pulses (10 words of 12 6-bit characters each) (Bell and Newell, 1971); that is, the system had a least a 720-dimensional state space. It would have taken thousands of parts to approximate such a system with lumped electrical elements. A hundred such delay lines constituted the 1000-word recirculating memory of the UNIVAC.

Consider a continuous-time moving-average filter: the output at any time is the average value of the input over an interval of length T ending at that time. This system cannot be described in terms of ordinary differential equations, nor implemented in terms of lumped circuit elements, because its state must represent all the details of the input over the preceding interval of duration T, as via a delay-line memory. If the input functions of time are band-limited—have a bound on the highest frequencies present in them—then the moving average can be well approximated by lumped circuits, or by discrete-time systems using samples of the signal to be smoothed. Or such systems can be built as physical analogs, using wave-propagating devices or magnetic-tape delay loops or some other distributed mechanism to hold the continuous-time distributed state. Whether by a distributed delay line or a lumped approximation to one, a moving-average filter can be implemented as shown in Figure 12.1.

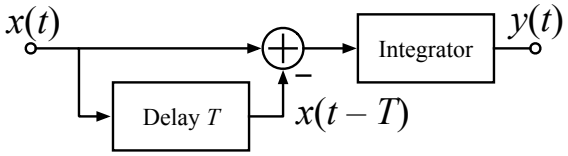


Figure 12.1: This diagram represents a linear time-invariant system that computes a moving average: $y(t) = \int_{t-T}^{t} x(t)dt/T$. The delay operator holds the details of the function $x(t)$ over an interval of duration T, so cannot be described via a finite number of state variables. Due to the subtraction $x(t) - x(t - T)$, the integrator outputs the difference between the integral up to time t and the integral up to time $t - T$ (except for a potential constant offset that can be dealt with by resetting the integrator state to zero when the input has been at zero long enough). In a physical implementation, the delay might be implemented by propagating waves through a lossless uniform medium, or by an approximation such as a suitable length of coaxial cable or a string under tension.

EE Connection: Linear Electrical Transmission Lines

It is common to model hydrodynamic wave systems such as the cochlea by electrical circuit analogs. Figure 12.2 shows a *ladder filter* of series inductors and shunt capacitors, approximating an electrical transmission line's inductance and capacitance per unit length. A lossless transmission line is essentially a pure delay (due to its wave equation, developed below), and its discretization into this circuit of inductors and capacitors acts approximately as a delay up to a bandwidth near the section resonant frequency $1/\sqrt{LC}$.

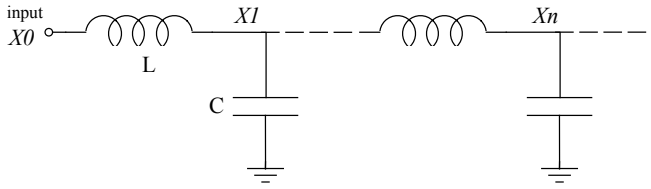


Figure 12.2: Waves travel at speeds somewhat less than the speed of light along wire transmission lines, including telephone lines, coaxial cables, power lines, etc. Such lines can be approximated or modeled by LC delay lines, circuits of iterated lumped elements as shown here.

For low enough frequencies, the responses at points Xn in the circuit of lumped elements are very much like the responses at a set of points on the distributed line, if those points are separated by comparable total amounts of series inductance and shunt capacitance. At higher frequencies, near $\omega = 1/\sqrt{LC}$, where $k(\omega)$ is large and the wavelength $2\pi/k(\omega)$ is not long compared to the section spacing, the approximation breaks down. Due to the local resonance, such high frequencies will not propagate through the lumped circuit, which is why a lowpass filter can be made this way. Such filters were traditionally called *electric wave filters* (Campbell, 1922; Zobel, 1924).

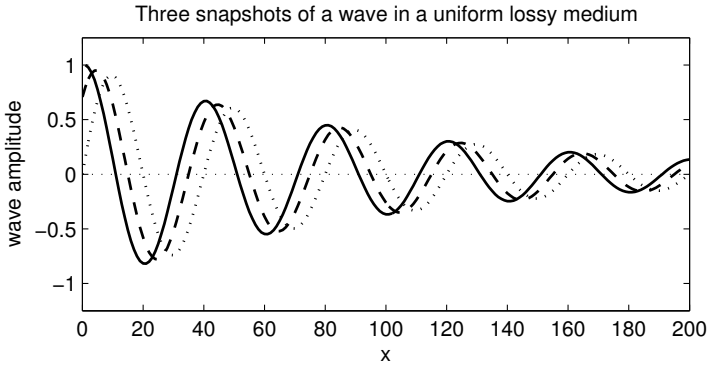


Figure 12.3: Three snapshots of a wave in a uniform lossy medium: the real part of $\exp(-ikx+i\omega t)$. The wave is sinusoidal in time, and a decaying sinusoid in space (due to the complex wavenumber k of the medium). The three snapshots (solid, dashed, and dotted curves) are separated by time intervals Δt corresponding to 1/8 cycle of the sinusoid, or $\pi/4$ radians (45 degrees) of phase (the period in time is $T = 8\Delta t = 2\pi/\omega$). The wave has a wavelength ($2\pi/k_{\mathrm{Re}}$) of 40 distance units (real part of k is therefore $k_{\mathrm{Re}} = 2\pi/40$ radians per distance unit). The wave amplitude is decaying as the wave propagates, by a factor of e per 100 distance units (imaginary part of k is therefore $k_{\mathrm{Im}} = -1/100$).

Physics Connection: Plane Waves in Multiple Dimensions

In uniform systems of more than one dimension, we can represent *plane waves* in any direction by a simple generalization: replace the dimension x by the *space vector* (location in 2D or 3D space) $\mathbf{x}$, the wavenumber k by the *wave vector* $\mathbf{k}$, and their product by the *dot product* (sum of products of coordinate dimensions) $\mathbf{k} \cdot \mathbf{x}$:

$$W(\mathbf{x}, t) = A \exp(-i\,\mathbf{k} \cdot \mathbf{x} + i\omega t)$$

The wave vector points in the direction of wave propagation. Planes perpendicular to this direction are called wavefront planes. Moving from a position $\mathbf{x}_1$ to a position $\mathbf{x}_2$ changes $\mathbf{x}$ by the vector difference $\mathbf{x}_2 - \mathbf{x}_1$. If this difference is orthogonal to the wave vector $\mathbf{k}$, then the phase at a given time does not change, because the dot product $\mathbf{k} \cdot \mathbf{x}$ does not change. Such positions are within a wavefront plane of the plane wave (it is also possible to have waves that are not plane, such as spherical waves emerging from a point source, but that's beyond what we'll need to consider here).

This generalization of k to a vector, plus the generalization to complex k, is the reason that waves are more often written in terms of wavenumber than in terms of wavelength.

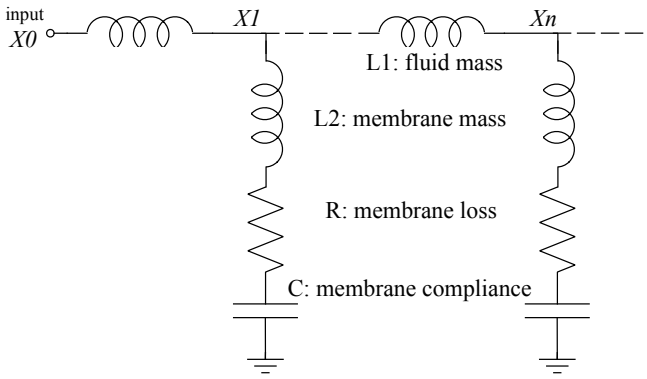


Figure 12.6: A transmission line of the type that has often been used for modeling wave propagation in the cochlea, starting with Wegel and Lane (1924), assuming a basilar membrane with significant mass and frictional loss. The mass and loss, modeled by inductance and resistance in the shunt-admittance legs, have little effect at low frequencies, where the membrane compliance, modeled by capacitance, limits the shunt current and makes the system act like the delay line of Figure 12.2; at higher frequencies this model exhibits local resonance and loss.

EE Connection: More General Transmission Lines

In the cochlea, the series inductance L in the model is analogous to fluid *mass*, while the shunt capacitance C is analogous to membrane *compliance*. Compliance is springiness, displacement per force or strain per stress—the reciprocal of stiffness. For more general media, such as lossy transmission lines and active cochleae, more general series impedances and shunt admittances are used, as shown in Figure 12.4, to represent combinations of mass, springiness, and loss or gain.

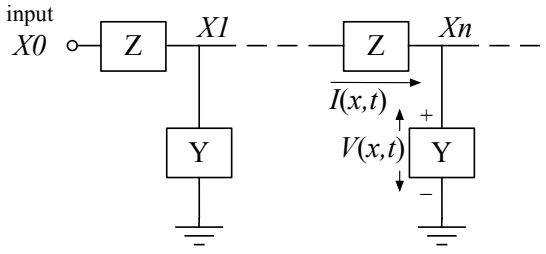


Figure 12.4: The general transmission-line model uses series impedances Z proportional to the distributed line's series impedance per unit length, and shunt admittances Y proportional to the line's shunt admittance per unit length. The electrical signals used to analyze it are the currents through the series elements and voltages across the shunt elements, signed as shown. The elements Z and Y are more general than the inductors and capacitors of Figure 12.2; each may contain series or parallel connections of several lumped elements, for example.

When the impedance of the series and shunt elements of a transmission line—or more precisely the series impedance $Z(\omega)$ per unit length (ohms per meter) and shunt admittance $Y(\omega)$ per unit length (siemens per meter)—are known and fixed (not varying with location), the wavenumber and characteristic impedance (ratio of wave voltage and current amplitudes) are simple functions of Z and Y.

The wavenumber solutions are found by solving Heaviside's *telegrapher's equations*, or *coupled time-harmonic transmission line equations*, that follow from elementary circuit analysis (Steinmetz, 1910; Mohamed, 2006):

$$\frac{dV}{dx} = -ZI, \quad \frac{dI}{dx} = -YV$$

These coupled equations imply a pair of similar wave equations for voltage and current waves, as can be seen by substituting each into a derivative of the other:

$$\frac{d^2V}{dx^2} = ZYV, \quad \frac{d^2I}{dx^2} = ZYI$$

Since the spatial second derivative of $\exp(-ikx + i\omega t)$ provides a factor of $-k^2$, the wave equations are satisfied by sinusoidal voltage and current waves of frequency ω whenever the wavenumber satisfies:

$$k(\omega)^2 = -Z(\omega)Y(\omega)$$

The ratio between the resulting wave amplitudes is the line's characteristic impedance Z_0, given by:

$$\frac{V}{I} = Z_0(\omega) = \sqrt{Z(\omega)/Y(\omega)}$$

In the case of a pure LC line, with $Z = i\omega L$ and $Y = i\omega C$ per unit length, k satisfies the relation $k^2 = \omega^2 LC$, so k is proportional to ω, signifying a frequency-independent velocity of $v = \omega/k = 1/\sqrt{LC}$, or a pure delay of $\sqrt{LC}$ per unit length.

EE Connection: Single- and Double-Ended Lines

The transmission lines that we have seen so far are known as *single-ended* lines, as a single wire carries the wave signal as a voltage relative to ground. Transmission lines are often built or drawn as *balanced lines*, as in Figure 12.5, with equal amounts of series impedance in two paths—like the 300-ohm twin-lead TV antenna wire some of you may be familiar with. The two chambers of the cochlea make a balanced or *differential* structure, too. In the analysis of such transmission lines, however, it is common to transform to a single-ended line, with only one line of series impedances, and to take voltages relative to a *ground* rather than differentially between the two sides. The resulting single-ended line is equivalent, for the differential wave; that is, for any wave that is antisymmetric about the center.

The same transformation is used in analyzing cochlear hydrodynamics, and in converting that analysis to an electrical equivalent, by assuming the pressures and longitudinal velocities (like currents) in cochlear fluid–membrane waves are opposite on the two sides of the cochlear partition (this is a good approximation as long as the wavelength is long compared to the dimension of asymmetries of the cochlear partition).

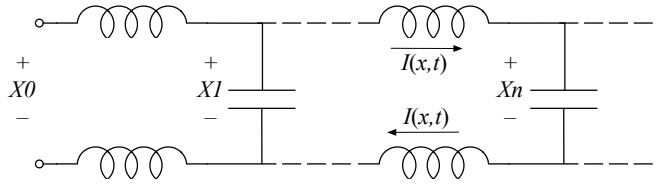


Figure 12.5: In a *balanced* or *differential* delay line, the signal of interest is the difference between the voltages on two sides, rather than with respect to a global *ground* potential. The symmetry allows such systems to be reduced to equivalent *single-ended* lines. The approximate symmetry of the hydrodynamic system of the cochlea across the cochlear partition is usually treated as good enough that the cochlea can be modeled by a single-ended transmission line circuit.

Physics Connection: Dispersion Relations and Bidirectional Waves

The relationship between wavenumber k and frequency ω is known as the *dispersion relation* for the medium. For a given frequency, it will typically have two (sometimes more) solutions for k, representing waves traveling in the $+x$ and $-x$ directions. We sometimes ignore the latter, the wavenumber for the backward-traveling wave, and represent k as simply a function of ω. Some systems will also have multiple *modes* of wave propagation, each with its own wavenumber for a given frequency, representing multiple solutions of a single dispersion relation (Watts, 2000); again, we sometimes ignore such complications, but keep in mind that they may become important second-order effects when evaluating some models of the cochlea.

Suppose a medium propagates waves in the $+x$ direction. For mechanical or hydrodynamic media, such as media that sound waves propagate through, these waves may be characterized by displacements and velocities (for example, of points on a guitar string, or points in air). Displacement and velocity patterns propagate as sinusoidal waves $W(x,t)$ of the form described above. For electrical lines, waves are usually described in terms of voltages and currents, which propagate similarly:

$$V_+(x,t) = V_1 \exp(-ik(\omega)x + i\omega t)$$

$$I_+(x,t) = \frac{V_+(x,t)}{Z_0(\omega)}$$

where ω is the frequency being considered, $k(\omega)$ is the propagation constant or wavenumber, V_1 is the complex amplitude of the voltage wave propagating in the $+x$ direction (assuming k is positive or has a positive real part), and $Z_0(\omega)$ is the characteristic impedance of the medium.

A wave can also propagate in the other direction, corresponding to the other solution of the dispersion relation; consider the electrical line's dispersion relation:

$$k^2 = -Z(\omega)Y(\omega)$$

$$k = \pm\sqrt{-Z(\omega)Y(\omega)}$$

When we treat $k(\omega)$ as a (single-valued) function of frequency, the reverse wave typically has wavenumber $-k(\omega)$, as implied by the square-root form of the dispersion relation (in other media than this simple transmission line model, multiple solutions of the dispersion relation may not be so simply related).

Besides negating the wavenumber in the phase expression, the reverse wave negates the relationship between voltage and current (since we want to continue to measure current consistently as flowing in the $+x$ direction):

$$V_-(x,t) = V_2 \exp(+ik(\omega)x + i\omega t)$$

$$I_-(x,t) = \frac{-V_-(x,t)}{Z_0(\omega)}$$

When waves exist in both directions at once, these currents and voltages add linearly to make consistent solutions.

Physics Connection: Reflections and Standing Waves

When waves of one frequency are propagating in both directions (for example, due to reflections from the far end of a finite line), the resultant waves in the medium are just sums of the forward and backward wave voltages and currents, which we can write in a way that emphasizes that the temporal pattern is everywhere sinusoidal:

$$V(x,t) = \exp(i\omega t)\left[V_1 \exp(-ik(\omega)x) + V_2 \exp(ik(\omega)x)\right]$$

$$I(x,t) = \exp(i\omega t)\frac{\left[V_1 \exp(-ik(\omega)x) - V_2 \exp(ik(\omega)x)\right]}{Z_0}$$

Due to the sum in one and difference in the other, the voltage-to-current ratio is no longer simply the characteristic impedance Z_0, as it is for the forward wave alone, nor $-Z_0$ as it is for the backward wave alone.

When the amplitudes V_1 and V_2 of the forward and backward waves are equal, the result is a pure standing wave, with sinusoidal time variation at every point, but with nonmoving sinusoidal spatial envelopes, showing zero current at the voltage envelope maxima and zero voltage at the current envelope maxima:

$$V(x,t) = |V_1| \exp(i\omega t)\cos(k(\omega)x - \phi)$$

$$I(x,t) = \frac{|V_1|}{Z_0}\exp(i\omega t)\sin(k(\omega)x - \phi)$$

for some ϕ that depends on the phases of V_1 and V_2.

Standing waves can come from reflections in lossless media. More generally, partial reflections lead to unequal forward and backward wave amplitudes.

A transmission line is not usually infinite; if it is terminated by a load resistance or impedance, or a short or open circuit, that termination will constrain the voltage-to-current relationship at that point, making a boundary-value problem that the V and I of the sum of forward and backward waves must satisfy. The solution will give the compatible amplitudes and phases of forward and reflected waves. Only in the case of termination by the characteristic impedance will there be no reflected wave: the forward wave will transfer all of its energy into the termination impedance, just as if it was propagating its energy down an infinite transmission line. Conversely, if the termination impedance is lossless (a short or open circuit, or a purely reactive impedance), all the energy will be reflected, and the backward amplitude will be equal to the forward amplitude, with just a phase shift that depends on the impedance, making a standing wave. In between these cases, some energy will be transferred into the termination and some will be reflected.

A similar analysis applies at boundaries between media, electrical or otherwise. For example, where sound waves in the ear canal hit the ear drum, some energy is transferred in and some is reflected; when the middle ear pushes on the cochlear fluid, some energy is transferred in and some is reflected. A high efficiency of energy transfer corresponds to a matching of characteristic impedances of the wave propagation media. In an electrical system, a *transformer* is used to change the voltage–current ratio to interface efficiently between different impedances; in the ear, the leverage of the middle-ear bones does this job.

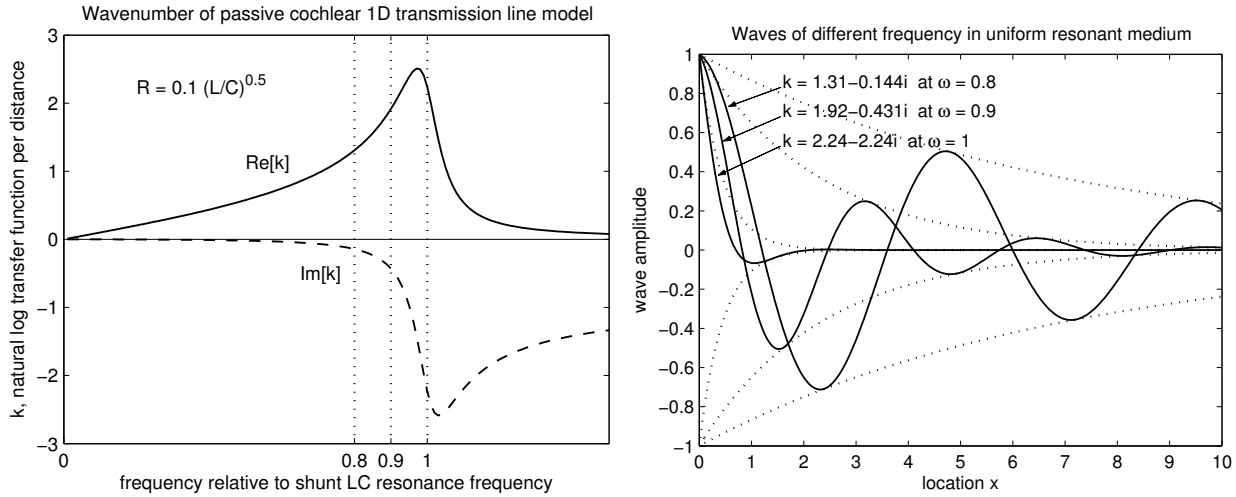


Figure 12.7: The complex wavenumber of the transmission line of Figure 12.6 is plotted in the left panel (real part solid, imaginary part dashed). As discussed in this chapter, these plots can also be interpreted as the log transfer function of a unit-length segment of the line (phase lag per distance solid, log gain dashed). Snapshots of waves at the marked frequencies (0.8, 0.9, and 1.0 times the shunt resonance frequency) are plotted in the right panel, and annotated with corresponding numerical values of k, to illustrate the relative phase shift of $\mathrm{Re}(k)$ radians per unit distance, and attenuation by a factor $\exp(\mathrm{Im}(k))$ in a unit distance, at these frequencies; spatial envelopes are shown dotted. Lower frequencies, where the imaginary part of k is negligible, propagate easily, while frequencies near and above resonance are strongly attenuated.

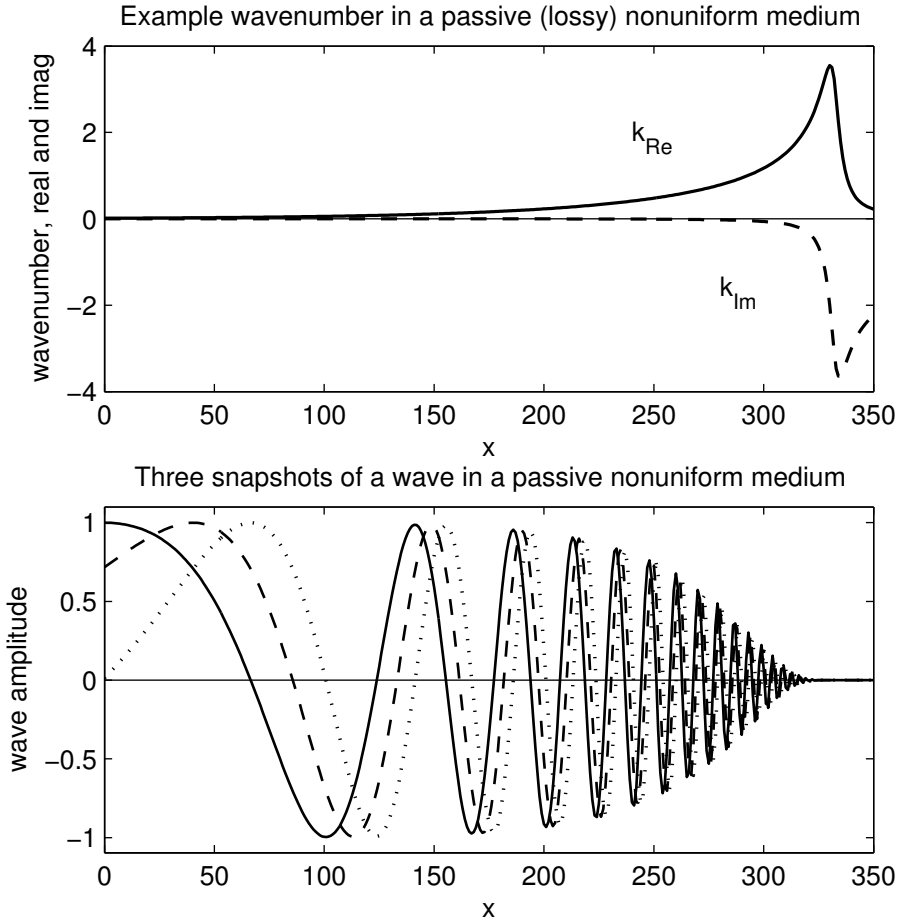


Figure 12.8: The wavenumber real and imaginary parts (top curves, solid and dashed respectively) of a hypothetical nonuniform medium, for a given frequency corresponding to the resonant frequency of the medium near its right-hand end, and corresponding wave snapshots (lower curves). The wave is sinusoidal in time, but not in space, due to the spatially varying properties of the medium that reduce the resonant frequency by a factor of 2 every 50 distance units. The wavenumber curves resemble those of Figure 12.7, since the same form of underlying model is used here, but with spatially varying parameters. The three snapshots (solid, dashed, and dotted curves) are separated by time intervals equal to 1/8 cycle of the sinusoid, or $\pi/4$ radians (45 degrees) of phase. By comparing the positions of peaks or zero crossings, it can be seen the wave is moving rapidly in the region of low k (left side), and slows down as k increases (right side). As in Figure 12.7, the attenuation becomes high when the wavelength gets down to about 6 distance units (the wavenumber gets up to about 1). The wave energy is fully dissipated shortly before it reaches the location with shunt resonance frequency equal to the wave frequency. The wave power is proportional to the square of the amplitude shown; no amplitude correction has been applied to account for either the slowing of the wave or the varying physical parameters of the system.

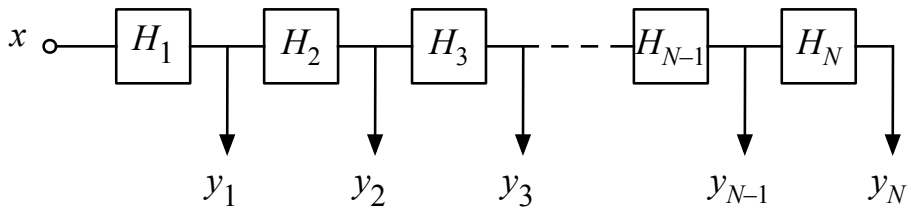


Figure 12.9: A cascade filterbank models wave propagation in a distributed system if each stage transfer function is a good approximation to the effect of the system's local dispersion relation, via the relation $H_j(\omega) \approx \exp(-ik(\omega, x_j)\,\Delta x)$

Part III

The Auditory Periphery

Part III Dedication: Georg von Békésy

This part is dedicated to the memory of Georg von Békésy (1899–1972), winner of the 1961 Nobel prize in medical science and physiology for his work on hearing. I never met Békésy, who died when I was still an undergraduate. I so admire his persistent work against tough odds in hearing research, and giving us the basic observations and model of how the cochlea works via traveling waves, that I felt this part of the book should be dedicated to him. My favorite quote is from his 1974 paper, published after his death (thanks to his punctual submission), in which he recalled his realization that "... dehydrated cats and the application of Fourier analysis to hearing problems became more and more a handicap for research in hearing" (Békésy, 1974). I think we have gotten beyond the cat problem, and it is my hope that, through this book, we make progress toward the day when Fourier analysis no longer handicaps hearing research.

In this part of the book, we survey the auditory periphery, and develop a machine model of its sound-processing function. We review auditory filter models, and the fitting of auditory filter models to human and animal data, based on some of the filter types developed in Part II.

We extend these filter models to develop the CARFAC description of the cochlea based on filter cascade models of wave propagation, and present a machine implementation in detail, through the final output step of the cochlea: the release of neurotransmitter by the inner hair cells to stimulate the auditory nerve.

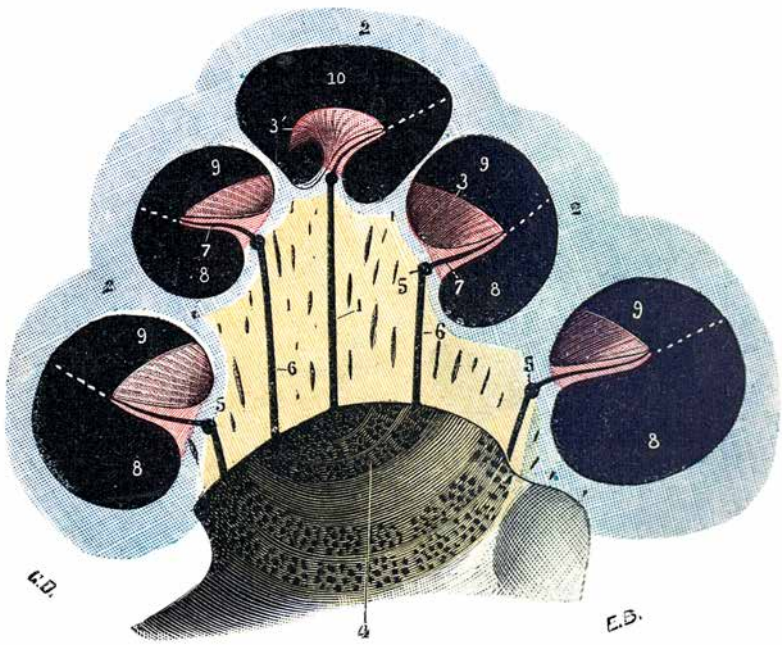


Fig. 890.

Coupe transversale du limaçon osseux : l'un des segments, vu par sa surface de coupe (*demi-schématique*).

A semischematic transverse cut of the "bony snail," the cochlea, published in color by Leo Testut (1897) (see color plates).

Chapter 13

Auditory Filter Models

> The original aim of this research was to obtain a mathematical expression for the amplitude characteristic of the hypothetical auditory filter that could be used to predict the power that a tone must have to be just audible in the presence of a given noise.
>
> — "Auditory filter shape," Patterson (1974)

> The auditory filter may be considered as a weighting function representing frequency selectivity at a particular centre frequency. Its shape can be derived using the power-spectrum model of masking which assumes: (1) in detecting a signal in a masker the observer uses the single auditory filter giving the highest signal-to-masker ratio; (2) threshold corresponds to a fixed signal-to-masker ratio at the output of that filter.
>
> — "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Moore and Glasberg (1987)

On Quasi-Linear Filters

When is linear not linear? What kind of nonlinear system lets us use linear systems descriptions effectively? How can a level-dependent filter with a strongly compressive input–output relationship be described as a linear filter, with a frequency response? These questions are addressed by the notion of a *quasi-linear filter*.

A quasi-linear filter is really a family of filters, with typically one parameter that chooses between them. In the case of auditory filters models considered here, the parameter is a signal level (an input level or an output level or some such parameter). Each filter in the family is an ordinary time-invariant linear system, described by a frequency response and other conventional descriptions, such as impulse response, poles and zeros if rational, transfer function, etc.

When the input to the auditory system is a broadband noise-like signal of a particular level, the behavior (whether observed physiologically or psychophysically) is often described fairly well in terms of linear filtering. But for different input levels, different filters are needed. Across a wide dynamic range of levels, significant changes of gain, bandwidth, and such are needed to fit the data, signifying a strongly nonlinear process. Yet at any particular level, a linear filter model fits well.

A description of the auditory system in terms of a quasi-linear filter consists of a combination of a family of linear filters and the relationship that controls the parameters of that family in response to a level measurement. Dynamic variation of level and parameters, as would occur in an AGC loop when the level is changing, is not part of the quasi-linear model.

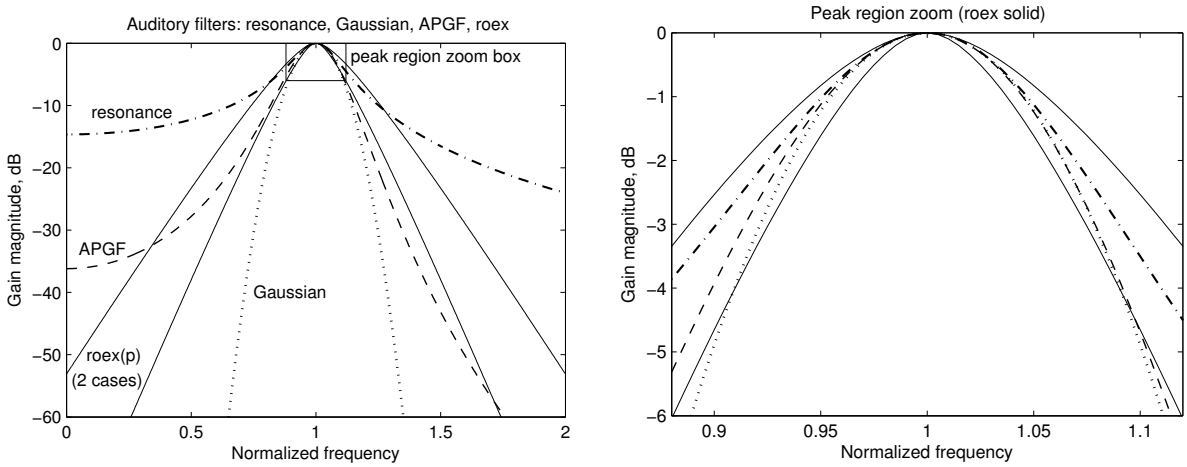


Figure 13.1: A few auditory filter model shapes, showing the large room for possibilities between the simple two-pole resonator and the Gaussian filter. The asymmetric 4th-order all-pole gammatone (APGF), and the symmetric roex(p) are included as examples of intermediate shapes. The illustrated filter shapes are matched for equal curvature at the peak, except for the sharper roex(p) case. The peak region zoom on the right shows how the matched roex peak curvature leads to rather wide skirts, and more reasonable skirts lead to a rather sharply curved peak, compared to the other filter shapes.

	roex			gammatones				cascades		
	(p)	(p,r)	(pl,pu,r)	GTF	GCF	APGF	OZGF	APFC	PZFC	PZFC+
1. Simple	fd	fd	fd	td	td	ld	ld	ld/s	ld/s	ld/s
2. BW control	+	+	+	+	+	+	+	+	+	+
3. Peak/skirts	–	*	*	+	+	–	+	*	+	+
4. Asymmetry	–	–	+	–	+	+	+	+	+	+
5. Gain variation	–	–	–	–	*	+	+	+	+	+
6. Stable tail	–	+	+	–	*	+	+	+	+	+
7. Runnable	–	–	–	+	*	+	+	+	+	+
8. Waves	–	–	–	–	–	–	–	+	+	+
9. Impulse resp.	–	–	–	–	+	+	+	–	–	+
10. Dynamic	–	–	–	–	*	+	+	+	+	+

Table 13.1: Scoring various auditory filter models on the ten criteria. The domains of simple description are frequency domain (fd), time domain (td), Laplace pole–zero domain (ld), and Laplace per stage (ld/s). The * represents partial credit: for the roex and APFC peak/skirts shape criterion, some control but not a great fit; for the gammachirp filter (GCF), various criteria have been met by useful pole–zero filter approximations.

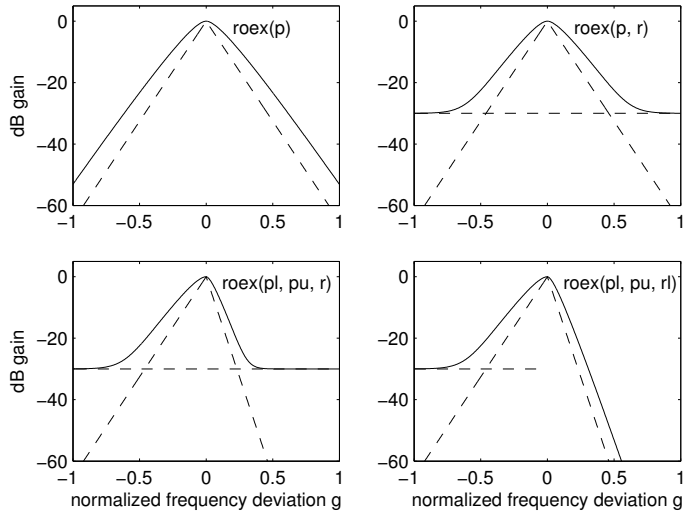


Figure 13.2: The roex family can take various parameters, including some illustrated here. In the roex(p) model (upper left), the factor $1 + p|g|$ rounds the filter's power gain above a symmetric triangular skeleton. In the roex(p, r) model (upper right), a floor at gain r is imposed (here at $r = 0.001$). In the roex(pl, pu, r) model (lower left), the upper and lower sides have different slopes, but the same floor. Allowing separate floor levels, or using no floor on the upper side, as Rosen and Baker (1994) did, allows more realistic asymmetric filter shapes (lower right) (we do not treat this as a separate model here, just a different parameterization of the asymmetric version). All of these are still too peaky (not rounded enough) near the peak, compared to more realistic models.

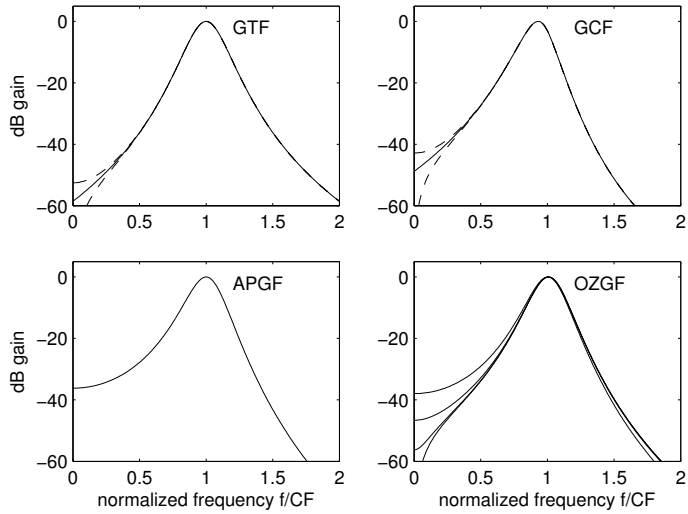


Figure 13.3: Auditory filter model shapes in the gammatone family include the real and complex gammatones (upper left) and gammachirps (upper right)—the real versions at several phases shown dashed—the all-pole gammatone filter (lower left), and the one-zero gammatone filter (lower right). The OZGF has explicit control of the low-frequency tail shape by the zero parameter, whereas the real GTF and GCF tail shapes vary as a side effect of other parameters.

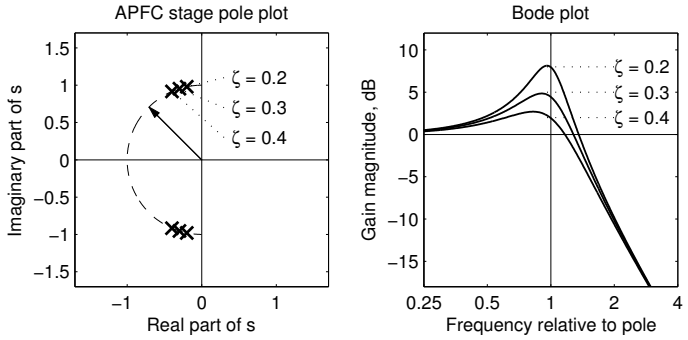


Figure 13.4: Diagram of the level-dependent motion of the poles of an APFC stage in response to a gain-control feedback signal, and the effect on the resonator frequency response. The positions indicated by crosses in the s-plane plot (left) correspond to pole damping ratios (ζ) of 0.2, 0.3, and 0.4. Corresponding transfer function gains (right) of this resonator stage do not change at low frequencies, but vary by several decibels near the pole frequency.

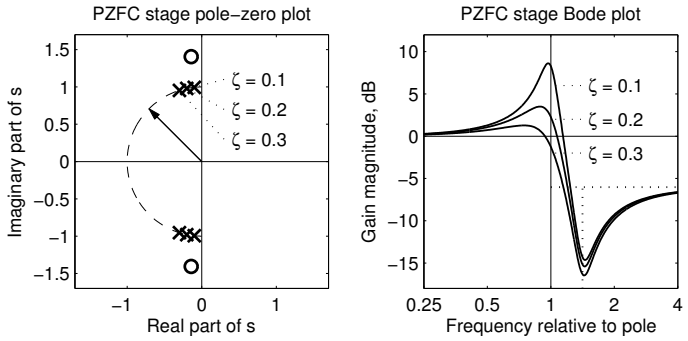


Figure 13.5: Diagram of the motion of the poles of a PZFC stage in response to a gain-control feedback signal, and the effect on the resonator frequency response. The positions indicated by crosses in the s-plane plot (left) correspond to pole damping ratios (ζ) of 0.1, 0.2, and 0.3, while the zero's damping ratio remains fixed at 0.1. Corresponding transfer function gains (right) of this asymmetric resonator stage do not change at low frequencies, but vary by several decibels near the pole frequency. The fact that the stage gain comes back up after the dip has little effect in the transfer function of a cascade of such stages.

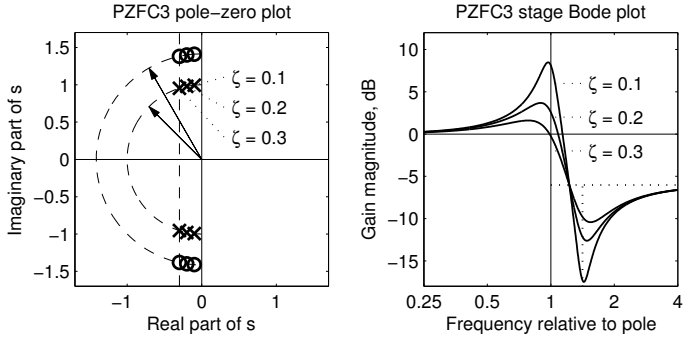


Figure 13.6: Diagram of the motion of the poles and zeros of a PZFC3 stage in response to a gain-control feedback signal, and the effect on the resonator frequency response. The zeros move along with the poles in the PZFC+; they are constrained to the same real-part value in this PZFC3 variant.

Human Notched-Noise Masking Experiments

Human auditory filter shapes can be inferred from the results of an ingenious family of experiments on the detection of tones in noise, especially by using noise bands arranged asymmetrically above and below the tone frequency, as illustrated in Figure 13.7.

A *notched noise* consists of two frequency bands of noise with a quiet frequency band (the notch) between them. Such noises have been used as maskers in tone-detection experiments, to get at the filtering that the auditory system does, since the 1950s (Webster et al., 1952); the method became more important in the 1970s (Patterson, 1976; Patterson and Nimmo-Smith, 1980), after it became clear that listeners were employing an "off-frequency listening" strategy to detect masked tones. That is, the interpretation of experimental data was that in trying to detect tones, listeners were effectively paying attention to the filter channel with best signal-to-noise (SNR, or tone-to-masker) ratio, rather than to the channel with the filter's peak frequency matching the probe tone's frequency. Taking this effect into account, experiments with *asymmetric notched noise*, that is, using probe tones placed off-center in the notches, provided a way to better assess the effects of different parts of the auditory filter shape.

Detection thresholds in these tests are based on *two-alternative forced-choice* (2AFC) experiments, in which a probe tone is present in one of two stimuli, and the subject has to say which one. The tone detection threshold, for a given noise spectrum and level, is taken as the tone level for which subjects are 75% correct. Filter model parameters can then be fitted by optimizing, in a squared-error sense, the prediction of these experimental thresholds given the noise parameters.

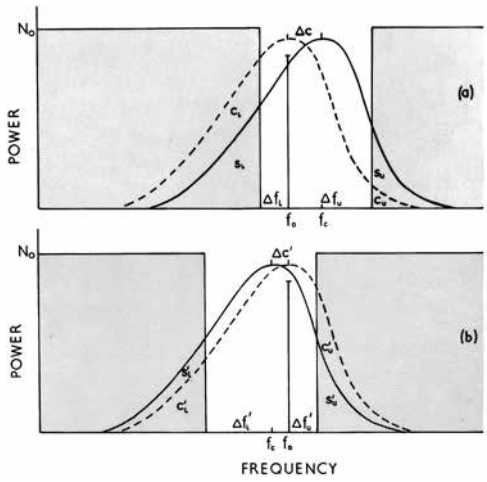


Figure 13.7: In the asymmetric notched-noise masking experiment, the presumption is that the auditory filters that are shifted to pick up the most favorable ratio of probe tone power to total noise power (the solid curves shown, as opposed to the dashed curves with their filter peaks at the probe tone frequency) are the filters that determine the tone detection threshold via a threshold signal-to-noise ratio. By using various different notch widths on the low and high sides of the probe tone frequency, as in the top and bottom plots here, the experiment's detection threshold data from human listeners provide indirect information about both sides of the filter shape. [Figure 1 (Patterson and Nimmo-Smith, 1980) reproduced with permission of AIP Publishing.]

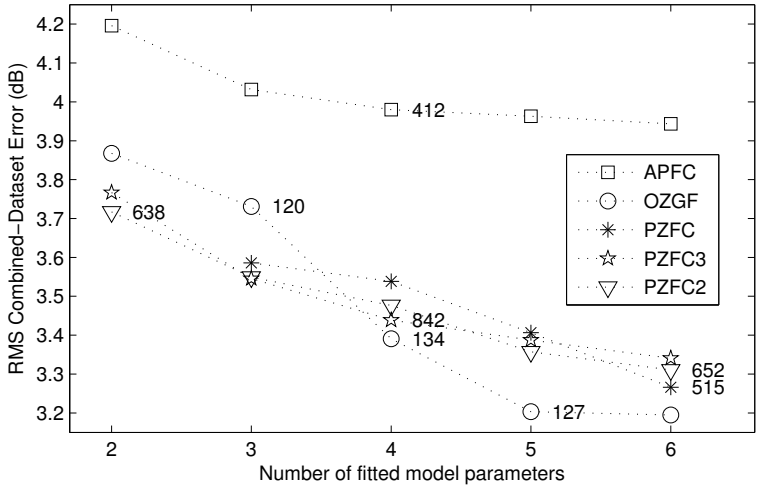
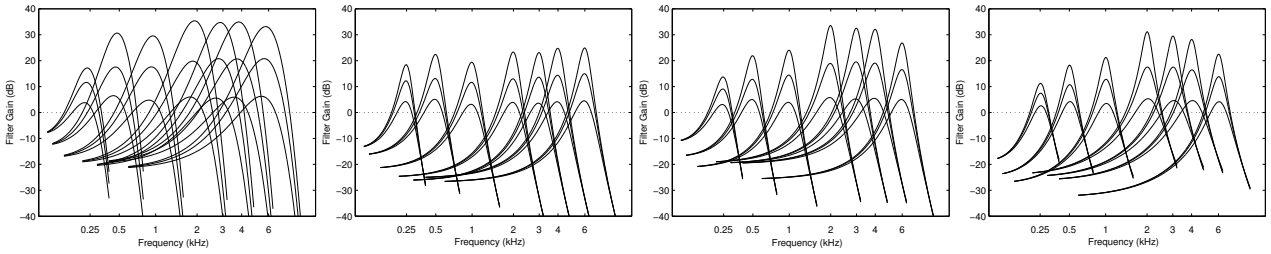
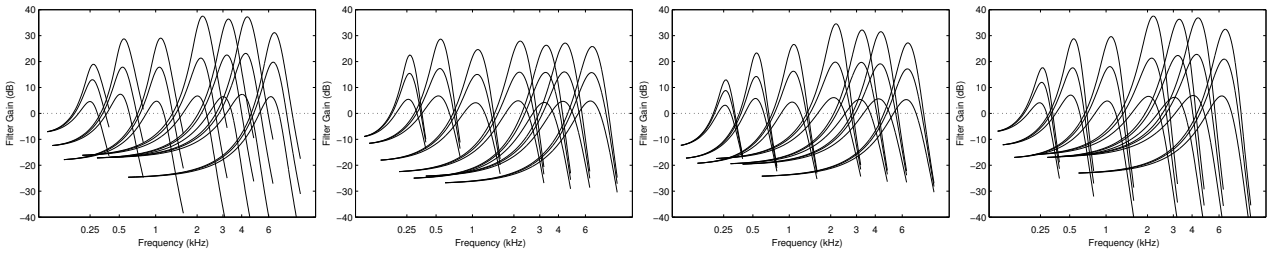


Figure 13.8: Threshold-prediction rms errors for several auditory filter models, versus number of fitted parameters, on the combined human masked-threshold datasets of Baker et al. (1998) and Glasberg and Moore (2000). The fit numbers are for reference only; different filter models are identified by different symbols, as shown in the legend. For each model type, only the fit with lowest error at each number of parameters is shown (at each number of parameters, several different parameterizations are possible within the model fitting framework). The errors are monotonically decreasing, since adding a free parameter never increases the error. The PZFC+ variants (PZFC3, star, and PZFC2, triangle) are the PZFC modified to have the zeros move with level, parallel with the poles, as opposed to the original PZFC (*) for which the zeros are fixed.



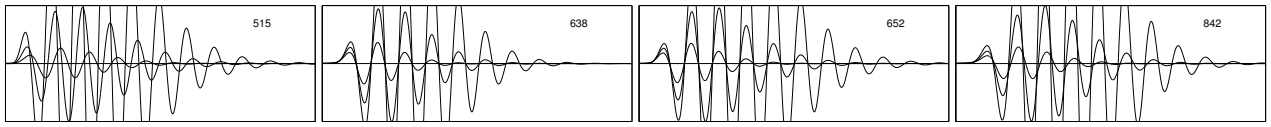
(a) APFC 412 (b) OZGF (APGF) 120 (c) OZGF (APGF) 134 (d) OZGF 127

Figure 13.9: Auditory filter gain plots for an all-pole filter cascade and for some OZGF model types, including APGF (the one zero moved out to infinity to make it an all-pole filter). The frequency axes are on the ERB-rate scale. In each case, the curves represent filter gain when the tone detection thresholds are 30 dB (highest curves), 50 dB, and 70 dB (lowest curves). The curve spacing is related to the input–output compression: curves close together, as at 250 Hz, correspond to a response that is only slightly compressive, while curve tips 15 dB apart represent a 4:1 compressive response. The APFC is not competitive in terms of prediction error, since its bandwidth is too large and its high-side rolloff too steep. Compare these shapes with the conceptual level-dependent filter description of Figure 10.1 D, and with the measured mechanical responses of Figure 10.2 D.



(a) PZFC 515 (b) PZFC2 638 (c) PZFC2 652 (d) PZFC3 842

Figure 13.10: Auditory filter gain plots for several PZFC model types, including the PZFC2 and PZFC3 versions of the PZFC+ with different constraints on the zero positions relative to the pole positions.



(a) PZFC 515 (b) PZFC2 638 (c) PZFC2 652 (d) PZFC3 842

Figure 13.11: The impulse responses for the 1 kHz channel of three versions of the PZFC, at noise levels corresponding to the three tone threshold levels 30, 50, and 70 dB SPL. The large (off-scale) curves are for the noise level that leads to 30 dB SPL tone threshold, the medium (full-scale) curves for 50 dB, and the small curves for 70 dB. The base PZFC (left) has nonmovable zeros, and up to about 180 degrees of phase shift of the zero crossings between the high and low levels. The PZFC2 models allow the zeros to move more than the poles do, keeping the zeros' Q or relative damping the same as the poles, and achieves stable zero-crossing times. The PZFC3 variant (right) constrains the zeros to follow the poles horizontally, at the same real part of s-plane location, as in filter D of Section **??**; this much zero movement limits the zero-crossing level dependence to about 45 degrees.

Chapter 14

Modeling the Cochlea

> . . . the assumption of a 'passive' cochlea, where elements are brought into mechanical oscillation solely by means of the incident sound, is not tenable. The degree of resonance of the elements of the cochlea can be measured, and the results are not compatible with the very heavy damping which must arise from the viscosity of the liquid. For this reason the 'regeneration hypothesis' is put forward, and it is suggested that an electromechanical action takes place whereby a supply of electrical energy is employed to counteract the damping.
>
> — "The physical basis of the action of the cochlea," Thomas Gold (1948)

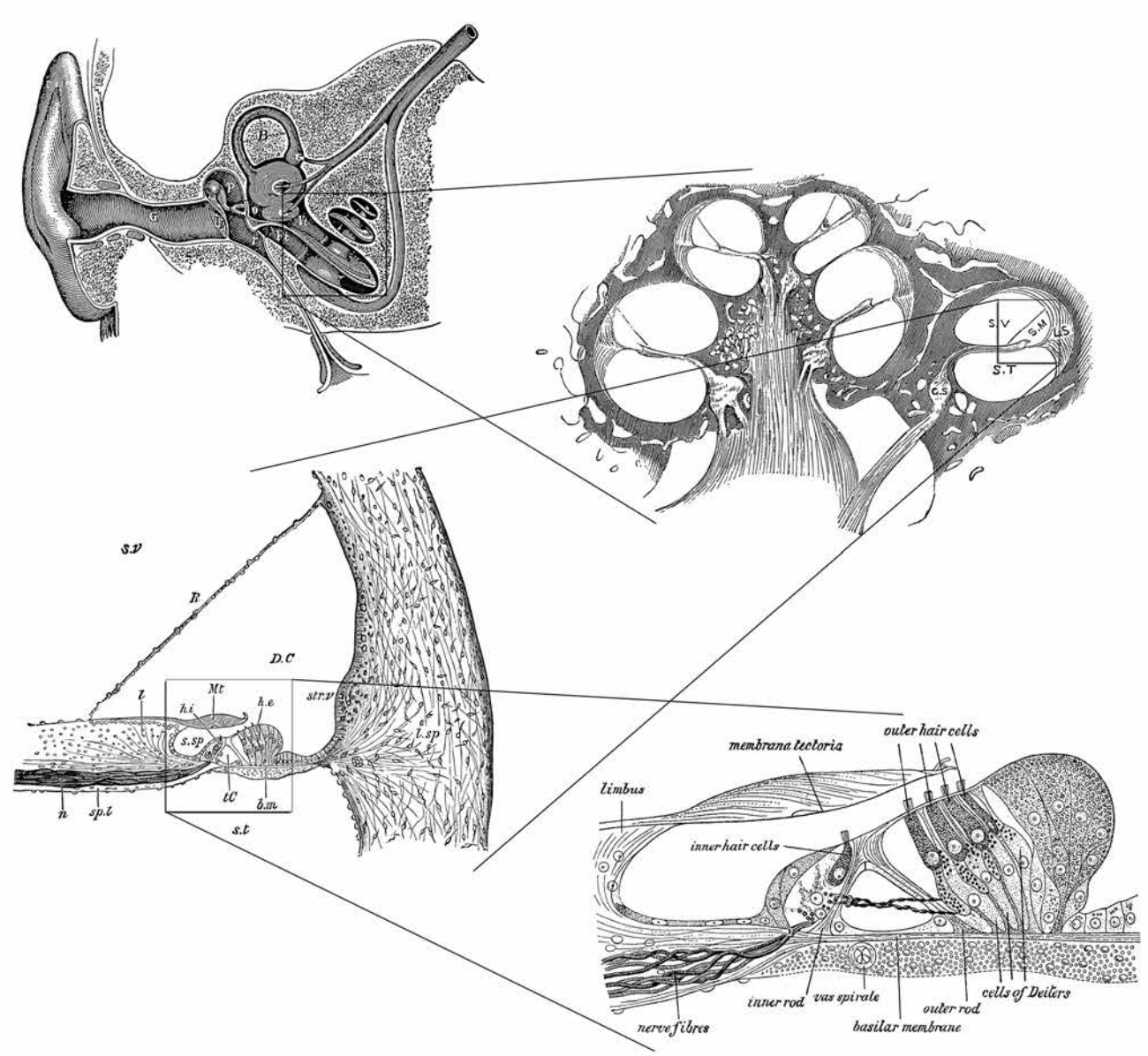


Figure 14.1: Four classic cross sections of the cochlea, from the macroscopic to the microscopic, with boxes and lines to show approximately how they relate.
Upper left: Leo Testut (1897) includes this drawing by Johann Czermak of the outer ear's sound path through the ear canal (G) to the eardrum, or tympanic membrane (T), and the middle ear bones that couple sound into the cochlea of the inner ear, via the oval window (O).
Upper right: *Gray's Anatomy* section through the cochlea. The structures that separate scala vestibuli (S. V.) from scala tympani (S. T.), in the region highlighted, are detailed in the next figure.
Lower left: This cross section through part of one turn of the mammalian cochlea, by Anders Retzius (1884), shows the cochlear duct (D.C, shown as scala media, S. M., in previous figure), scala vestibuli (s.v), scala tympani (s.t), basilar membrane (b.m), Reissner's membrane (R), tectorial membrane (Mt), nerve fibers (n), and the organ of Corti.
Lower right: This *Gray's Anatomy* drawing by Retzius shows a section through the organ of Corti, pointing out one inner hair cell and four outer hair cells.

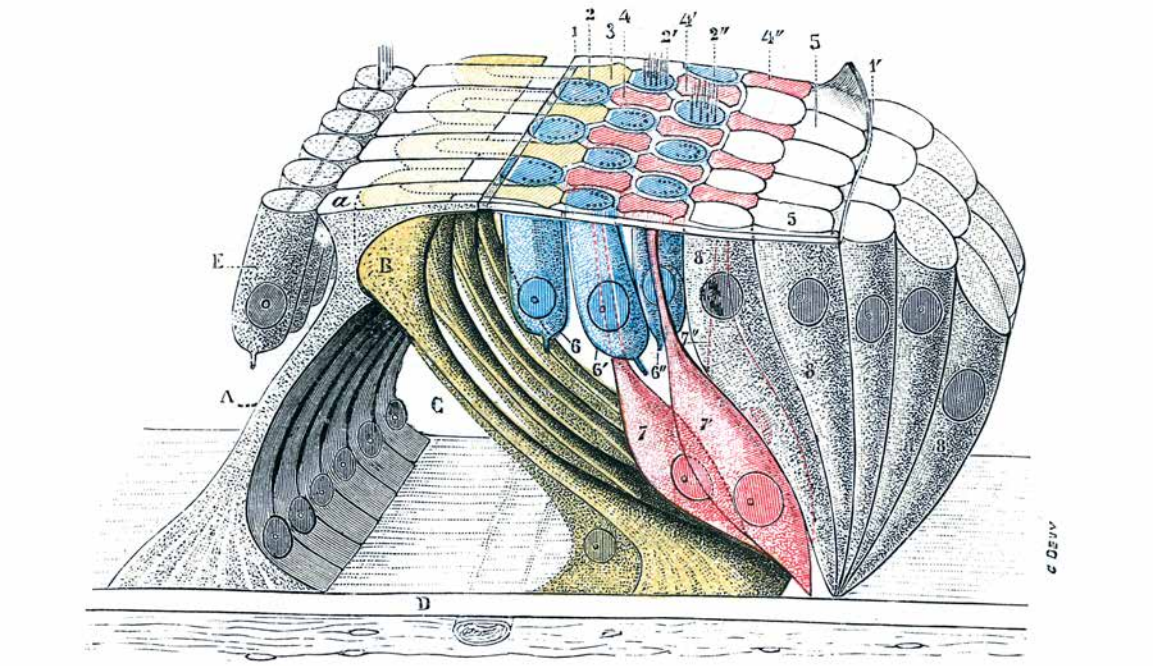


Fig. 918.

La même membrane, avec les cellules qui lui servent de substratum et dont l'empreinte lui donne son aspect réticulé (*schématique*).

Figure 14.2: The three rows of outer hair cells (blue) and one row of inner hair cells (E) sit with their upper ends and hair bundles exposed to endolymph in the scala media through the reticular lamina, but otherwise surrounded by a sealing barrier made up of the pillar cells (A and B), cells of Dieters (7), and other cells of the organ of Corti. This beautiful colored image was published more than 115 years ago (Testut, 1897).

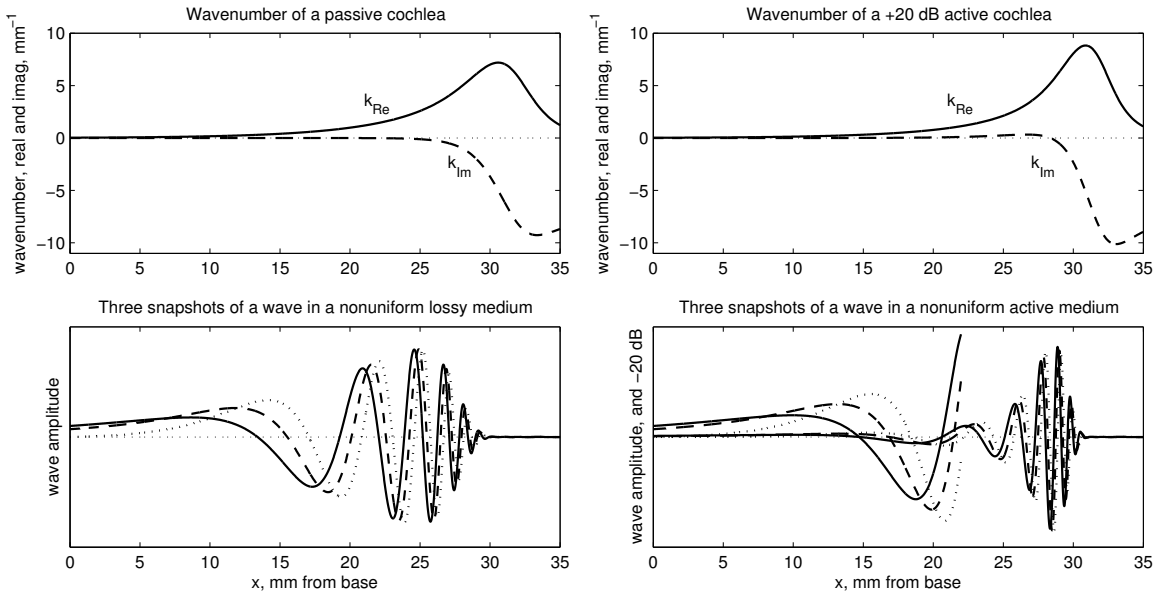


Figure 14.3: Three snapshots of a traveling wave in a passive cochlea (left), and in an active cochlea (right), responding to a sinusoid. The wavenumber, top, is estimated using the methods of Chapter 12, to correspond with our cascade of asymmetric resonators (CAR) model of Chapter 16. The slightly positive imaginary part of the one on the right corresponds to the active gain. The wave is calculated via the WKB approximation, at many more points than we would typically model in a filterbank. In the passive case, the amplitude peak is not very localized. To display the amplified signal in the active case, we cut its gain by a factor of 10 (–20 dB) after showing the part near the base that nearly matches the passive case. To get the large number of cycles from base to apex, we use 10 filter stages per mm, or 350 total, which is more than we would typically use in a machine hearing system (that is, the wavenumbers as plotted in mm^{-1} units are 10 times the natural log of the filter stage transfer functions).

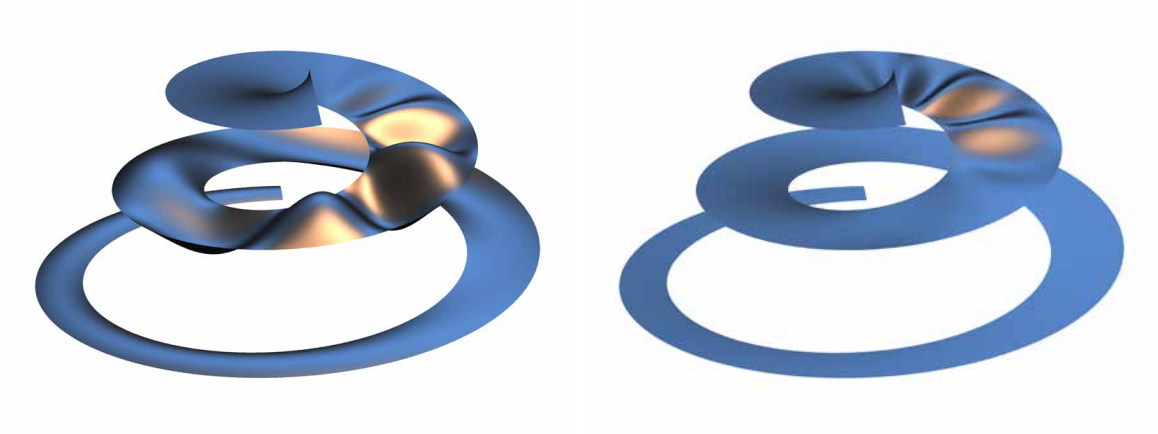


Figure 14.4: The traveling waves shown in Figure 14.3 are here mapped onto a 3D model of the basilar membrane, greatly exaggerated and stylized with colored lights. The active case with 20 dB more gain (right) is rendered for a 30 dB lower input level, so it represents the response on the same scale with a factor of 1000 less input power, corresponding to a cube-root-compressive system (10 dB output level change for 30 dB input level change).

Early Cochlear Resonance and Wave Concepts

While the BM starts narrow and gets wider, the other part of the cochlear partition, the bony shelf (as well as the partition as a whole) starts wide and gets narrower. This bony structure misled the seventeenth-century French anatomist Joseph-Guichard Duverney (1683) to conclude that the cochlea was tuned to low frequencies near the base and high frequencies near the apex; see Figure 14.6. The eighteenth-century Italian Domenico Cotugno realized that the structure responsible for tuning was more likely to be the basilar membrane, and turned this around to the scheme that persists today; he also discovered that the cochlea was normally filled with fluid, not air as previously believed. Other eighteenth-century scientists who worked on the idea of frequency–place tuning include Valsalva, Boerhave, Zinn, Haller, and Goeffry (Shambaugh, 1910). In the nineteenth century, Hermann von Helmholtz tied up the local resonance theory with psychoacoustic and mathematical support.

Almost immediately, the Helmholtz concept of independent resonators, like the stretched strings of a harpsichord, came under attack from others who thought it seemed physically unlikely. It is a testament to Helmholtz's lucid description and analysis, and to his stature and authority, that the idea persisted as long as it did, and that even today it colors the thinking of many people about how the cochlea works. Alternative explanations of cochlear function had a hard time taking hold, with lots of early half-baked ideas, before Békésy observed the cochlear traveling wave in 1928. Even after that observation, there were continued difficulties, since models that fit Békésy's broadly-tuned wave observations could not explain sharp psychophysical and neural tuning.

Charles Herbert Hurst (1895) proposed a nonresonant traveling-wave theory, relying on coincidence of reflections to sort out different pitches. As it was described shortly thereafter (McKendrick, 1899; McKendrick and Gray, 1900),

> Hurst has suggested that with each movement inwards and outwards of the stapes, a peculiar wave is generated which travels up the scala vestibuli, through the helicotrema into the scala tympani, and down the basilar membrane to the fenestra rotunda. This wave is not a mere undulation of the basilar membrane, but it causes movements of fluid to and fro in each scala, and these produce a peculiar wave of pressure. As the one wave ascends while the other descends, a movement (or pressure) of the basilar membrane occurs at the point where they meet, and the movement is chiefly in the direction of the tectorial membrane, so that this membrane strikes suddenly on the hair cells and thus irritates the nerves. The point at which the waves meet will depend upon the pitch of the note, or, in other words, upon the time interval between the two waves. In this way, and without sympathetic resonance, the cochlea would, within limits, respond to tones of different pitch. The intensity of the movement of the tectorial membrane against the hair cells would, of course, correspond to intensity of tone.

This idea was a step toward wave theories, but not a realistic one.

Development of Cochlear Wave Concepts

Emile Kuile (1900) proposed an alternate nonresonant traveling-wave theory that, depending on frequency, sounds would set different lengths of BM in motion, with low frequencies affecting more length than high (Stewart, 1901; Fletcher, 1922).

Max Meyer (1907) published an account of the mechanics of the inner ear in which he rejected the local resonance hypothesis, based on his analysis of the properties of the basilar membrane. He argued that the BM was not under tension and thus would not behave elastically, and that any wave propagating by its displacement would have a wavelength long compared to the cochlea, such that the BM would move essentially as a whole. He appears to have not considered stiffness as alternative to tension as a way to get an elastic displacement; he treated the BM as having a nonlinear limit of displacement, such that larger portions would be displaced by louder sounds, but at all frequencies. He illustrated the antisymmetric motion of fluids in the scalae, as driven by the stapes, but didn't quite get to a wave response on the BM, so missed the opportunity to replace the resonance theory with a more physical wave theory.

About the same time, George Shambaugh (1910) and others developed a theory of the effective stimulus to the hair cells being a resonance of the overlying tectorial membrane (TM). Shambaugh felt that the TM was resonating "in response to the impulse of sound waves in the endolymph." His notion of a wave in the cochlea was a fast sound wave, like some others at that time. He supported the Helmholtz resonance theory while denying that the basilar membrane could be "a vibrating structure." Luciani (1917) supported this view in his eminent physiology textbook.

The move toward a more mathematical and physical model of traveling waves in the cochlea started with H. E. Roaf (1922), who wrote,

> Mass movement of the liquid can take place in one of two ways: Liquid may pass up the *scala vestibuli* through the *helicotrema* and down the *scala tympani*, or the *scala media* may be pushed towards the *scala tympani*. The resistance to these movements is in the former case the inertia of the mass of liquid to be moved, and the friction of the liquid against the walls of its containing tube, and in the latter case the tension of the basilar membrane (Reissner's membrane is usually represented as being flaccid).

This approach was further detailed using membrane stiffness (elasticity) instead of tension, and converted to an electrical analog, by Wegel and Lane (1924), and was given a good impetus when Békésy (1928) reported traveling waves that he observed on the BM.

Otto Ranke (1931) showed that Wegel and Lane's 1D or long-wave model would not be accurate at the point of maximum response, where the predicted wavelength was less than the duct height, and that a 2D model or a simplified short-wave model would work better.

Wever (1962) reviews the development of the traveling wave models of the cochlea, but never mentions the concepts of long and short waves, nor of linearity and nonlinearity. It wasn't until the later experimental observation of nonlinear active amplification (Kemp, 1979) that models of the cochlea began to be able to explain the subtle psychophysics of hearing. A range of historical overviews are available for the interested reader (Shambaugh, 1910; Luciani, 1917; Fletcher, 1922; Wever, 1949, 1962; Hawkins, 2001; Hachmeister, 2003).

Development of the Concept of AGC in Cochlear Mechanics

William Rhode (1971) observed a very nonlinear input–output relationship in cochlea mechanics, using his newly developed Mössbauer technique. In the same year, Rose et al. (1971) were among the first to suggest that observations on auditory nerve spike train patterns strongly suggested a mechanical "sensitivity control" in the cochlea:

> The capacity of a fiber to reflect the waveform of the stimulus when the latter greatly exceeds that sound pressure level which elicits a saturation discharge rate suggests the existence of a cochlear sensitivity control mechanism which may, but perhaps need not be, mechanical in nature. ... acceptance of nonlinearity drastically revises the classical, but nonetheless quite incredible, conclusion that at threshold the receptors are sensitive to displacements as small as a tiny fraction of the diameter of a hydrogen atom. It is also tempting to think that the receptors are not exposed to enormous variations in the amplitude of vibration as is postulated by the orthodox view. In fact, there is recent direct evidence [Rhode 1971] that the motion of the cochlear partition, in the region of maximal amplitude, is markedly nonlinear and therefore a very substantial error may be introduced in calculating the amplitude of the displacement at threshold by linear extrapolation of values observed at very high sound pressure levels.

By the end of the decade, modelers were taking note. Jont Allen (1979) made the case for AGC, in terms familiar to engineers, and began to connect it to efferent feedback:

> Given the opinions which we have so strongly expressed up to this point, the reader might reasonably ask what overall purpose cochlear nonlinearities serve. For those familiar with the data, one answer seems almost obvious: The nonlinear damping (as proposed in nonlinear cochlear models) acts to compress (attenuate) the frequency components of ... the neural excitation, near [CF] ... in order to increase the dynamic range of the filters. Thus the nonlinear damping acts as a mechanical automatic gain control.
>
> ...
>
> The outer hair cells are coupled to the efferent system, and COCB [crossed olivocochlear bundle] stimulation (stimulation of the outer hair cells through the efferent system) also gives rise to broadened tuning about CF in a manner very similar (as best we know) to the nonlinear level dependent mechanical damping. This experimental fact seems to be an important clue toward an understanding of the cochlear nonlinearity.

Allen (1981) continued to explain in the next of his sequence of papers on the state of cochlear modeling:

> A very significant feature of Rhode's data was that he found a compressive nonlinearity at frequencies neighboring the cutoff frequency. As a result, the output (BM displacement or velocity) varies much less than the stapes input displacement or velocity, for frequencies near the best frequency. The significance of this important finding will become clearer as we proceed, but, in my opinion, it is a precurser to an automatic gain control system which seems to be built into the cochlear filters. ... The automatic gain control nonlinearity also explains why the harmonic distortion is always below the primaries in intensity and does not grow large at large input levels as would be predicted from a power-law nonlinearity.
>
> ...
>
> It presently seems clear that this source of distortion is not the byproduct of some poorly engineered component. It is rather perhaps the negligible residual of a sophisticated local feedback mechanism in the mechanical motion of the properly operating cochlea, such as the automatic gain control system mentioned previously.

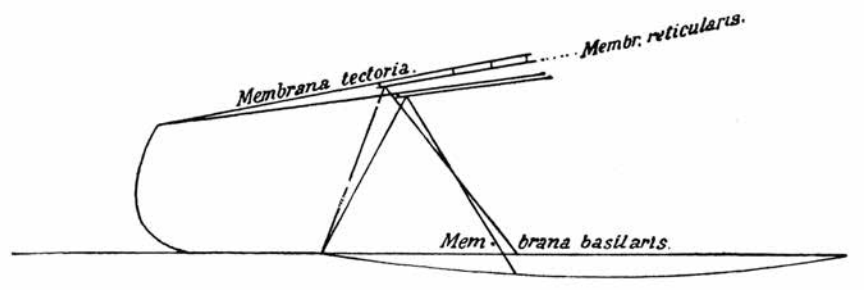


Figure 14.5: Diagram of how the hinged edge of the BM tilts the organ of Corti, causing a shear displacement between its top, the reticular lamina, where the hair cell cilia are, and the tectorial membrane (Kuile, 1900). The triangle represents the *pillar cells*, or *rods of Corti*, surrounding the tunnel of Corti.

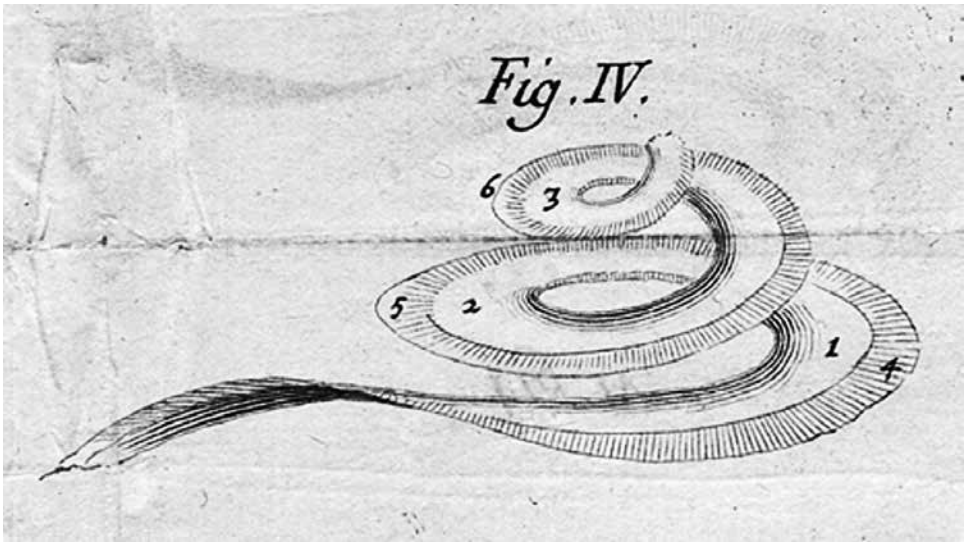


Figure 14.6: Duverney's 1683 drawing of the cochlea's spiral tuned structure. The inner lane (near the axis, his numbers 1–2–3) presumbly represents the bony shelf, the inflexible part of the cochlear partition that starts wide near the base and get narrower near the apex. The outer part (4–5–6), if it represents the basilar membrane, should start narrow and get wider, but he did not see it that way.

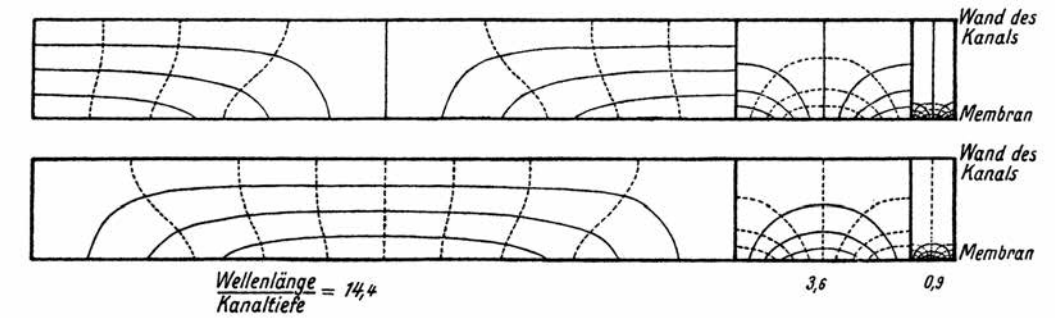


Abb. 2. Verteilung der Potentialströmung über die Kanaltiefe bei verschiedenen Wellenlängen.

Figure 14.7: Otto Ranke (1931) calculated these streamlines and iso-pressure lines for 2D waves in two narrow channels separated by an elastic membrane, at three wavelengths. He concludes, "Thus, while at long wavelengths almost all the pressure amplitude reaches the wall of the channel, at short wavelengths, the pressure at the channel wall remains nearly constant, and all the processes take place only in the immediate vicinity of the membrane." This is how the cochlear wave focuses sound energy into the vicinity of the organ of Corti. The left and middle conditions, with wavelength-to-channel-depth ratio (*Wellenlänge / Kanaltiefe*) of 14.4 and 3.6—corresponding to the wavenumber–height product $kh = 2\pi/14.4 = 0.44$ and $kh = 2\pi/3.6 = 1.75$—straddle the nominal $kh = 1$ boundary between long-wave and short-wave behavior.

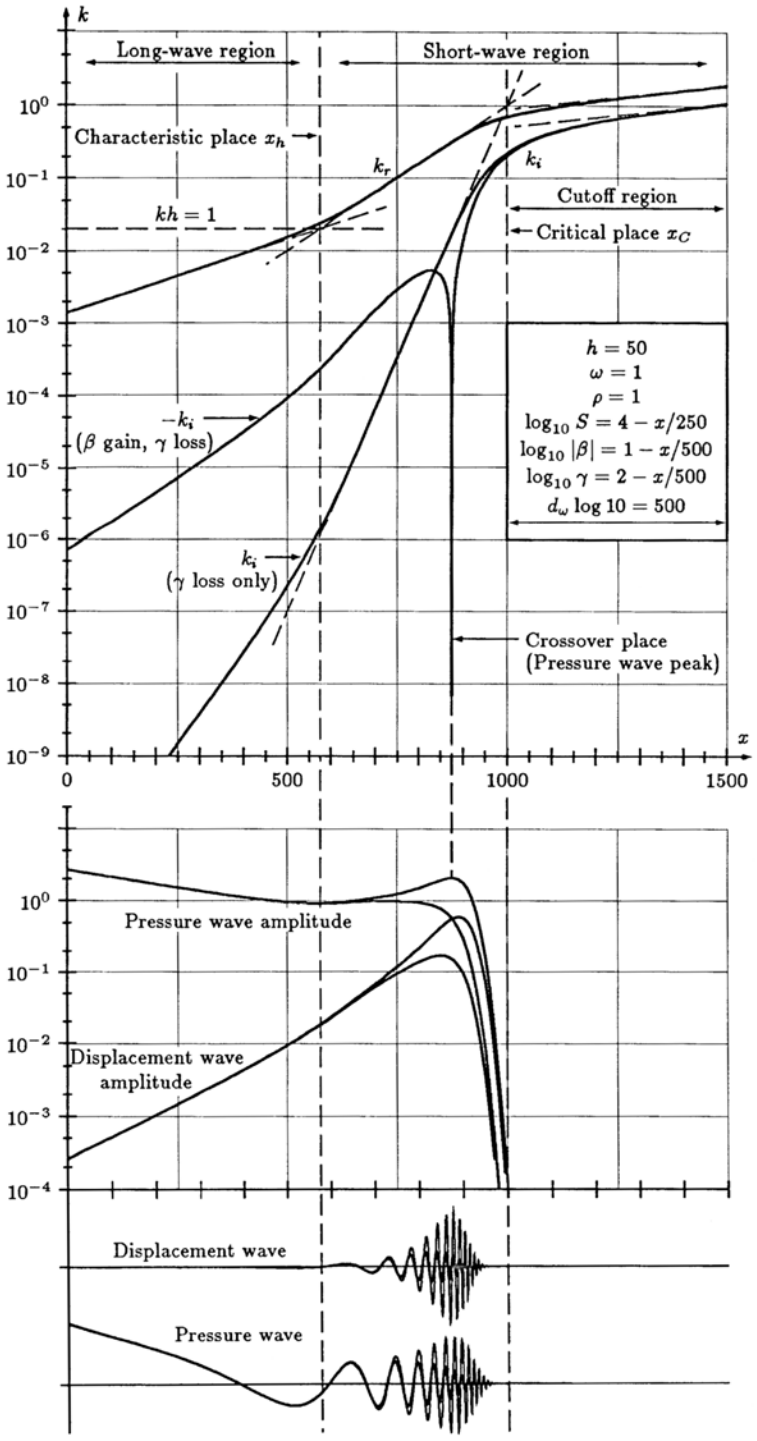


Figure 14.8: Wavenumber calculations from a 2D model as a function of place x, plotted as real and imaginary parts of k, with and without active gain, along with pressure and displacement waves with and without active gain; from Lyon and Mead (1988). The difference between pressure waves and displacement waves is mostly in the base region, where the frequency is low compared to CF and the membrane is very stiff, so the energy propagates with relatively low displacement and high pressure. For a detailed view of active and passive cases, compare Figure 14.3, which is not based directly on a 2D model but on a filter cascade that gives a wavenumber that is comparable except in the cutoff region.

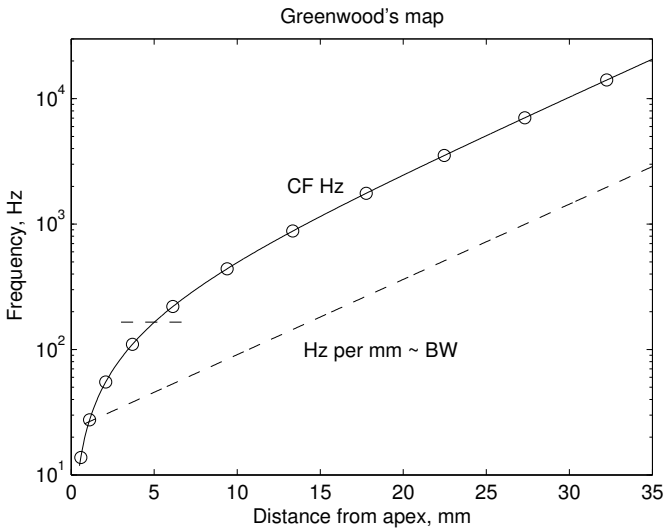


Figure 14.9: Greenwood's frequency–place map (Greenwood, 1990), showing the relation between places and their characteristic frequencies (CF). Points corresponding to frequencies of octaves (powers of 2 times A-440) are marked with circles. For most of the distance, the mapping is approximately geometric, or logarithmic. The dashed line shows the rate of change of frequency with place, in Hz per mm, which is proportional to the nominal bandwidth at each place. One way to get the Greenwood map is to integrate this exponential-in-place bandwidth, with distance from the apex, starting at zero center frequency but nonzero bandwidth.

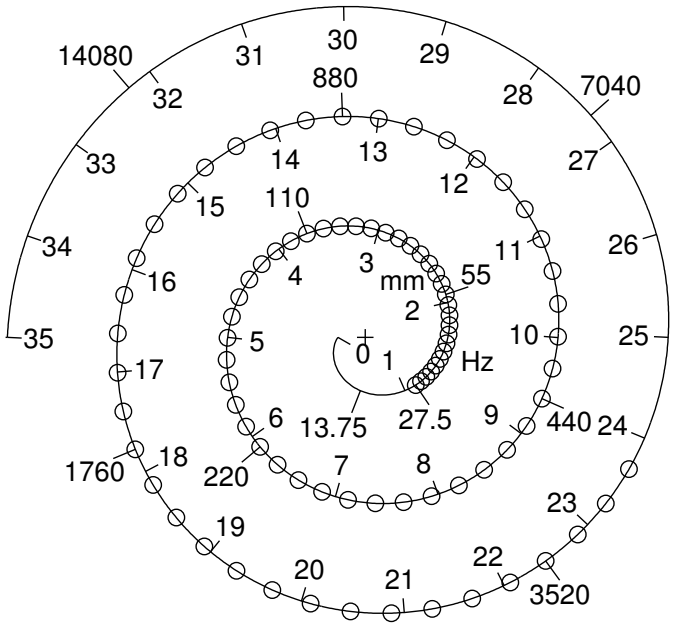


Figure 14.10: Greenwood's frequency–place map, illustrated on a spiral that approximates the shape of the human cochlea. Distances from the apex in mm are labeled inside the spiral, and frequencies of octaves on the outside. The fundamental frequencies, or pitches, of the notes of the 88 keys of a piano are marked by circles. Notice that geometrically spaced frequencies—octaves and notes—are about equally spaced, at nearly 5 mm per octave, in the basal and mid regions, but are bunched up near the apex, with only about 1 mm for the lowest octave of the piano. The human cochlea has about two and three-quarter turns; the final quarter turn shown in the center (the last 1 mm), which maps frequencies down to zero, should be interpreted as the helicotrema.

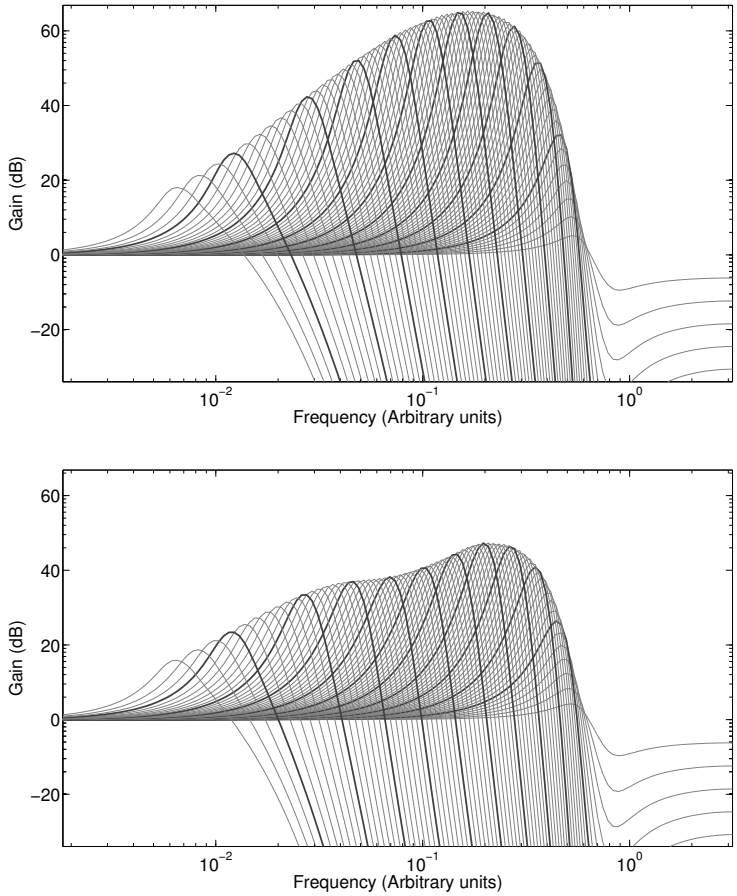


Figure 14.11: Adaptation of the overall filterbank response at each output tap, for the PZFC model of Lyon et al. (2010). The upper plot shows the initial response of the filterbank before adaptation. The lower plot shows the response after adaptation to a human /a/ vowel of 0.6 sec duration. The plots show that the adaptation affects the peak gains (the upper envelope of the filter curves shown), while the tails, behaving linearly, remain fixed.

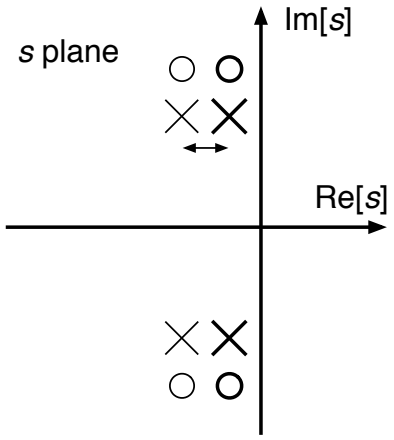


Figure 14.12: Diagram of the motion of the filter-stage poles and zeros in response to the CARFAC's gain-control parameter. The low-damping positions (heavy symbols) provide high gain near the pole frequency, compared to the high-damping positions (lighter symbols).

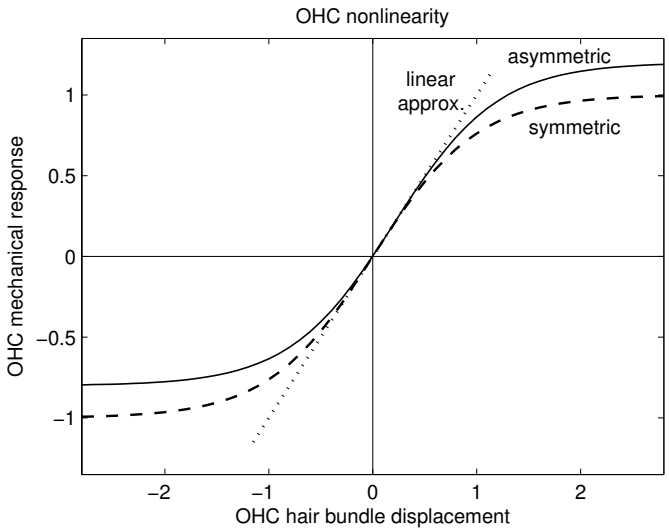


Figure 14.13: The transduction nonlinearity of the outer hair cells is a somewhat asymmetric *sigmoid* (solid), and is sometimes modeled as a symmetric sigmoid, such as a hyperbolic tangent (dashed). The slope of this curve is effectively a gain or active-undamping parameter, which is maximum near the rest (zero displacement) position.

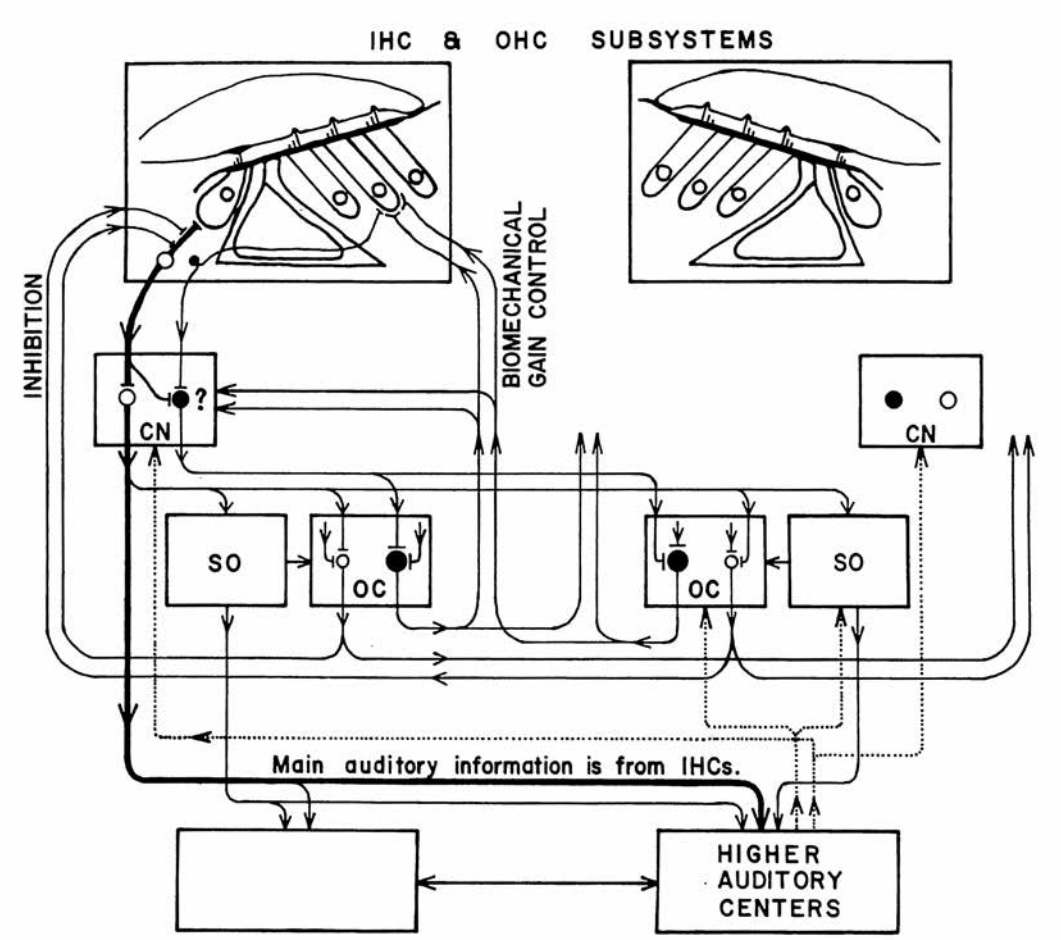


Figure 14.14: Duck Kim (1984) created this "block diagram for the hypothesized IHC and OHC subsystems in the cochlea and the brainstem up to the superior olivary complex and their connections to the remainder of the auditory system." The superior olivary complex (SO) drives the olivocochlear neurons (OC) that provide feedback from the brain to the cochlea, both to control the biomechanical gain and to inhibit the response of the primary auditory neurons that send auditory information from the inner hair cells (IHCs) to the cochlear nucleus (CN). Much of the feedback is crossing between left and right via the crossed olivocochlear bundle (COCB, not labeled), which is a convenient location for injecting electrical signals to directly control the cochlea's gains, both mechanical and neural. Filled circles represent neurons in the outer hair cell (OHC) subsystem. The hypothesized CN neuron with the question mark has since been identified in the marginal shell of the anteroventral cochlear nucleus (AVCN) (Ye et al., 2000). [Figure 7.3 of (Kim, 1984) reproduced by permission of Duck On Kim.]

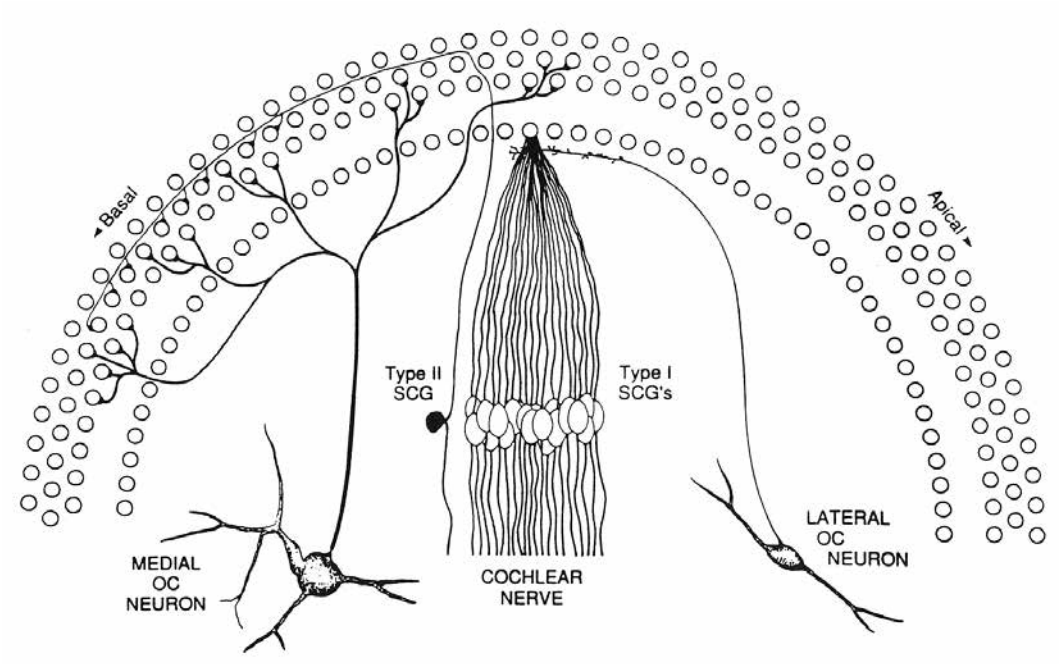


Figure 14.15: Bruce Warr's 1992 "hypothetical isofrequency unit of afferent and efferent innervation from the middle of the cochlea" shows the collection of different neurons types sharing a common CF, and how they relate to cochlear place (Warr, 1992). The small circles represent the one row of inner and three rows of outer hair cells in the spiral organ of Corti. For a given CF, the efferent feedback neurons from the medial olivary complex (OC) control outer hair cells over about a half-octave range of places toward the base from the place that drives the cochlear nerve afferents (Type I spiral ganglion cells), so that they can modulate the outer hair cell activity in amplifying traveling waves that are coming from the base (from the left in this drawing). [Figure 7.12 (Warr, 1992) reproduced with permission of Springer.]

Chapter 15

The CARFAC Digital Cochlear Model

> The modified transmission-line implementation, like the low-pass filter version, is an active system, with adjustments to the filter Q values changing the filter shapes and gains. ... This functional variation of Q with level gives a nearly uniform 2.5:1 compression ratio in the cochlear output for inputs ranging from 0 to 100 dB SPL.
>
> — "Accurate Tuning Curves In a Cochlear Model," James Kates (1993)

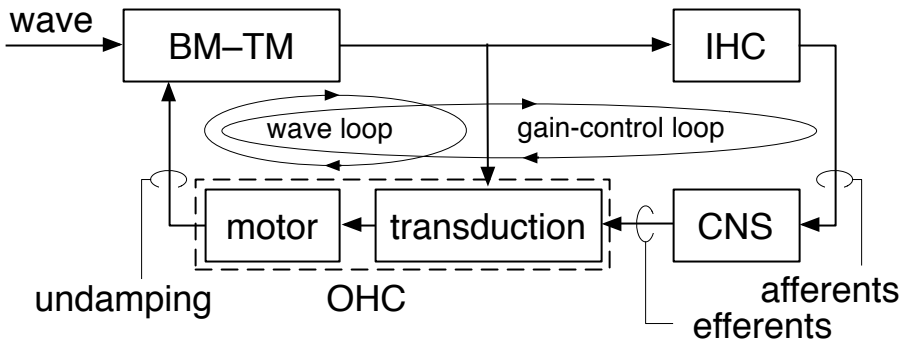


Figure 15.1: This diagram, adapted from Dallos (1992) and elaborated, shows the functional physiological elements of one location in the cochlea, which can be seen as a pair of feedback loops. The short loop, defining the hydrodynamic wave filtering system, involves the basilar and tectorial membranes (BM–TM) and active feedback from the outer hair cells (OHC), working at audio frequencies. The longer and slower loop, the afferent/efferent loop from the inner hair cells (IHC) through the brainstem of the auditory central nervous system (CNS) and back, controls the activity level of the OHCs, automatically adapting the system to the sound level. The instantaneous wave and the slower efferent feedback interact in the OHC, the nonlinear element whose "motor" action provides gain via active undamping of the wave mechanics.

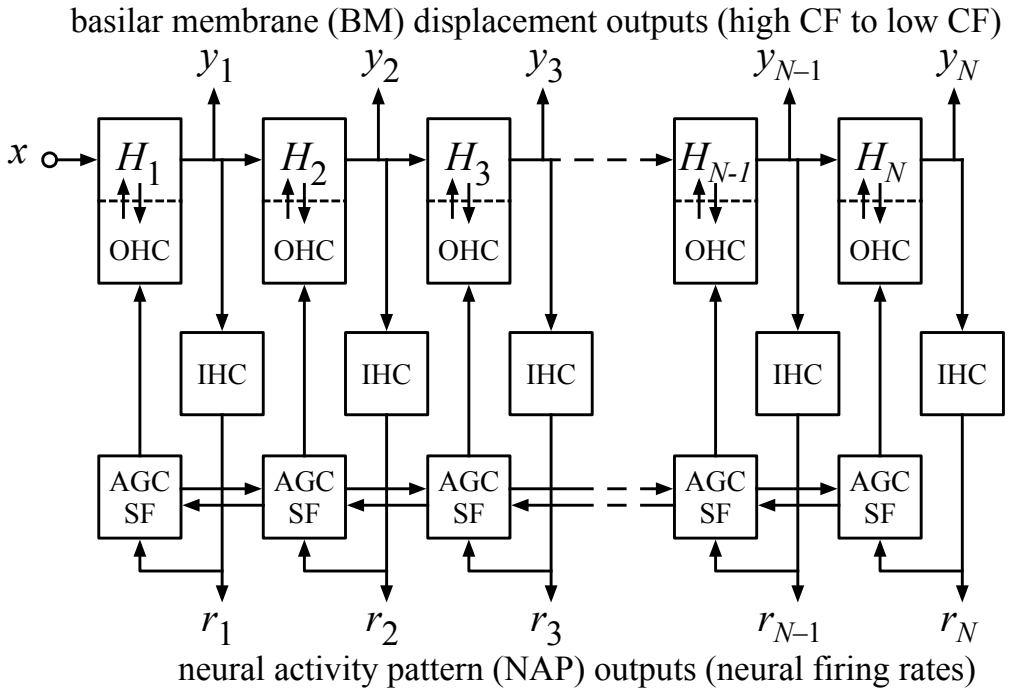


Figure 15.2: The *Cascade of Asymmetric Resonators* consists of the not-quite-linear transfer functions H_1 through H_N that model BM motion based on the cascade structure of Figure 12.9. *Fast-Acting Compression* is implemented by the other elements, including the OHC model that is tightly integrated with the filter stages and gives them their nonlinearity, and the coupled AGC smoothing filters (AGC SF) that modulate how the OHCs control the parameters of the filters. Between these main parts is a detection nonlinearity, such as an IHC model, which can have some compression and adaptive state of its own. The lateral interconnections of the smoothing filters allow a diffusion-like smoothing, or coupling, across both space and time. Outputs from the CARFAC include BM motion y_i (a set of compressed nearly-linear filterbank outputs) and an estimate of average instantaneous auditory nerve firing rate r_i, the NAP.

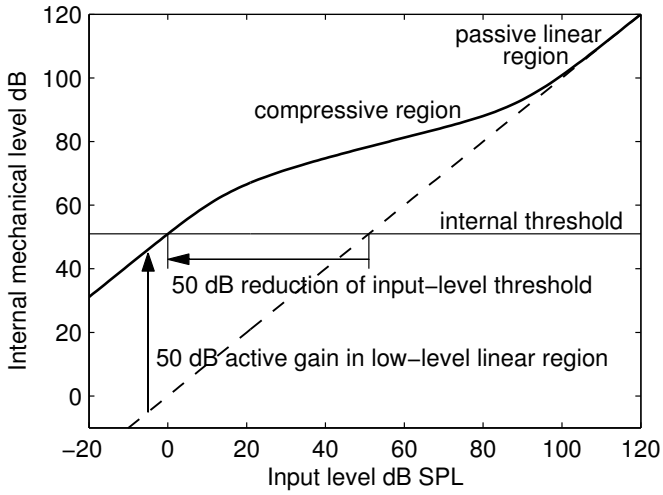


Figure 15.3: The compressive input–output curve exhibited by cochlear mechanics and emulated by the CARFAC model (solid) is compared with the passive linear or "dead" cochlear response (dashed), to show how extra gain at low levels reduces the input level needed to reach a threshold level of mechanical response. Here the mechanical threshold is chosen to correspond to 0 dB SPL with 50 dB of gain at low level. The curve is representative of the middle of the place or CF range, as opposed to very basal and apical regions that exhibit less active gain and less compression.

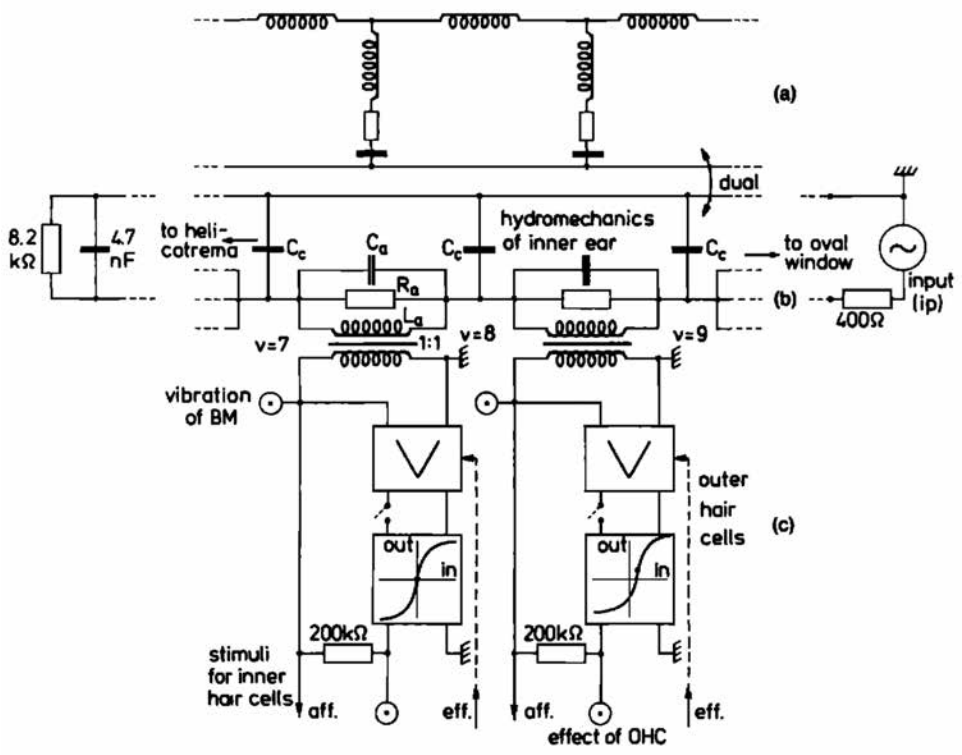


Figure 15.4: The analog bidirectional transmission-line model of Zwicker (1986), with saturating OHC nonlinearity and efferent feedback control, foreshadows the digital CARFAC functionality. Note the efferent ("eff.") control of the OHCs. [Figure 1 (Zwicker and Peisl, 1990) reproduced with permission of AIP Publishing.]

Chapter 16

The Cascade of Asymmetric Resonators

> Up to the threshold of the nervous system, the general outline of the process of frequency analysis is fairly clear. There is little room for doubting that the first main step of the process is essentially a matter of filtering. True, when one encounters electrical filters with one input and a number of outputs they are likely to consist of parallel selective networks fed from a common source. However, cascaded networks with taps at their junctions are not unfamiliar, and they provide a fairly exact analogue, insofar as general lay-out is concerned, of the cochlear analyser. The input is at the basal end of the cochlear partition, and the taps are the receptors or nerve terminals disposed along the length of the partition.
>
> — "Auditory frequency analysis," J. C. R. Licklider (1956)

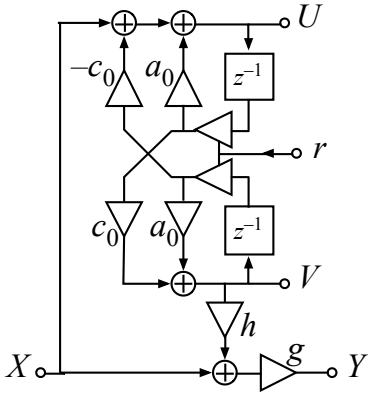


Figure 16.1: A pair of zeros is added to the coupled form by mixing the input with the filtered V output, as in filter D of Chapter 8. The resulting filter has zeros at the same radius in the z plane as the poles, which is a good place for them and gives the coordinated motion to keep the zero crossings of the impulse response from moving too much. The h coefficient controls the ratio of the zero frequency to the pole frequency, and the g coefficient is used to adjust the overall gain. In the arrangement shown, a factorization of coefficients to include an explicit pole radius parameter r is used to enable dynamic control of the damping. The pole radius is related to damping factor ζ by $r = \exp(-\zeta\omega_N T) = \exp(-\gamma T)$, as in Section **??** where the $\gamma = \zeta\omega_N$ is the negative real part of the s-plane pole position, mapped to the z plane via $z = \exp(sT)$. The a_0 and c_0 parameters, the cosine and sine of the pole angle, respectively, represent the pole positions $z = a_0 \pm ic_0$ in the zero-damping ($r = 1$) case.

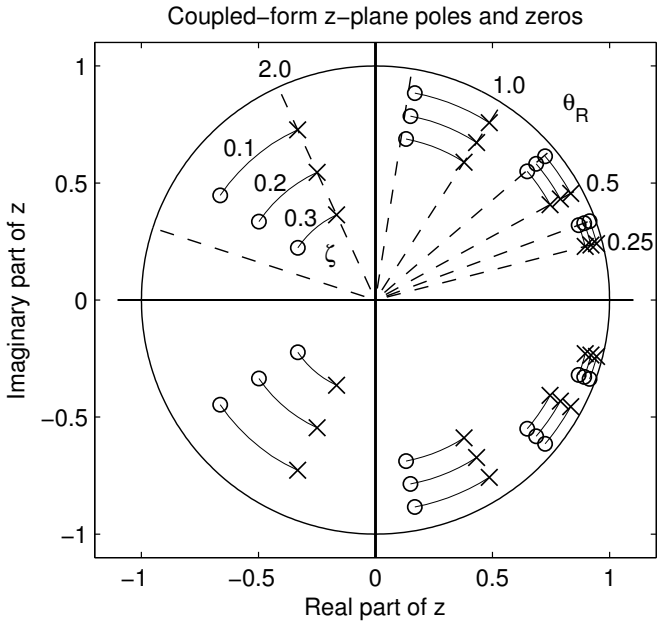


Figure 16.2: Pole–zero plot for the filter stage of Figure 16.1, illustrated for pole frequencies octave-spaced at $\theta_R = 0.25$, 0.5, 1.0, and 2.0 radians per sample, and damping factors $\zeta = 0.1$, 0.2, and 0.3, for the case $h = \sin\theta_R$; zeros are connected to their corresponding poles by solid thin arcs. This h value puts the zeros about a half octave above the poles, except at the highest pole frequencies, as shown by comparison with the radials (dashed) shown at $\sqrt{2}$ ratios (at the higher pole frequencies, the zeros squash closer to the poles, so they miss the dashed lines). Varying the a and c coefficients proportional to $r = \exp(-\zeta\omega_N T)$, or approximating that by $r = 1 - \zeta\omega_N T = 1 - \gamma T$ as we do here, moves the poles and zeros exactly along radial lines. We refer to γ/ω_R as the damping, even though it is not exactly.

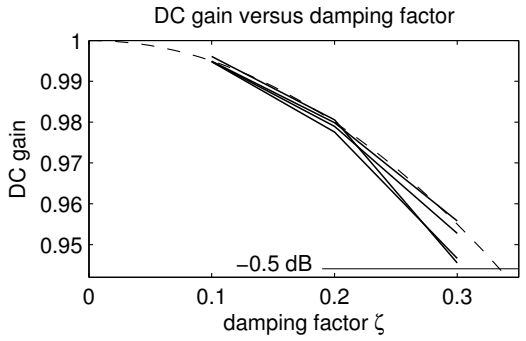


Figure 16.3: DC gains of the filter stages with pole and zero locations shown in Figure 16.2, when the gain coefficient g is fixed at the value that gives unity gain for the undamped case. The approximation $1 - \zeta^2/2$ (dashed) is most accurate at low θ_R. The thin line near the bottom indicates a loss of one-half dB.

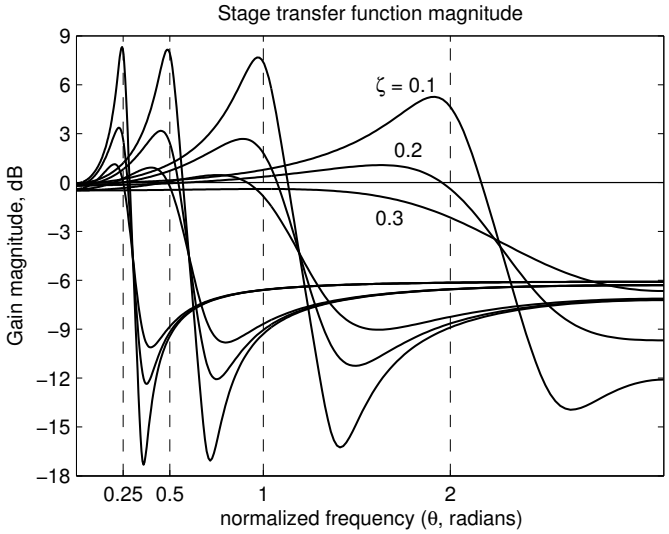


Figure 16.4: The CAR stage frequency-response gains for the four pole frequencies and three damping factors illustrated in Figure 16.2. For these plots, g is fixed, allowing the DC gain to deviate a bit from unity as damping increases. The deep "notch" behavior in the lower half of the plot leads to a very steep high-frequency slope in the cascade response; the fact that the gain comes back up some after the notch has little effect on the cascade filter shape, since the cascade gain is essentially zero in that region.

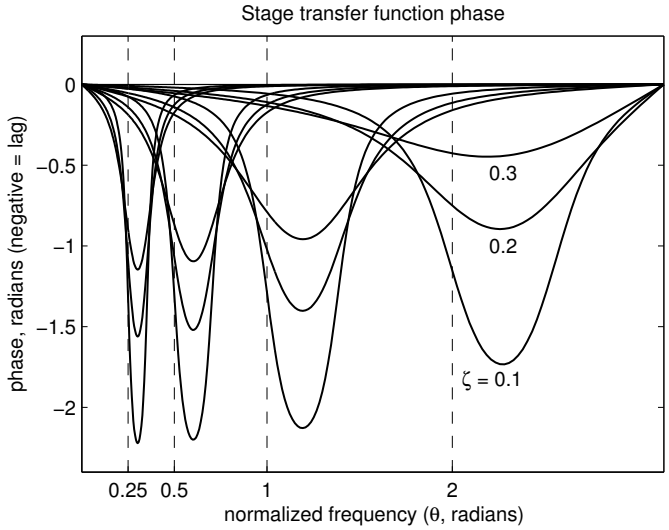


Figure 16.5: The phase responses of the stages with parameters illustrated in previous figures. As the damping changes, the phase stays approximately constant at a frequency just below the pole frequency, but goes through CF with a variable slope, indicating a variable group delay. Beyond CF, where the response is getting small, the phase lag is moving back toward zero, so the group delay there is negative.

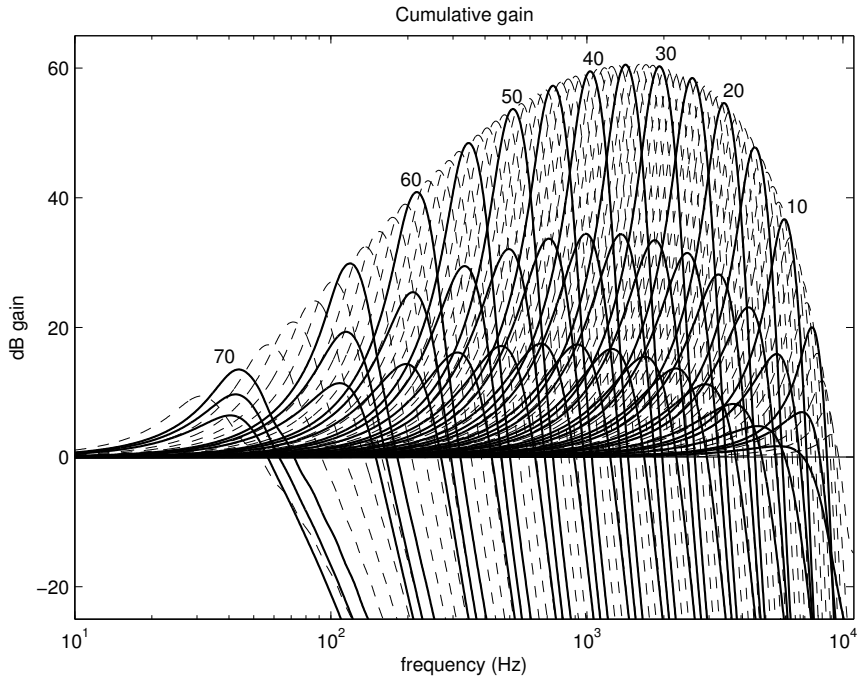


Figure 16.6: The cumulative frequency response (Bode plot) of a cascade of 71 pole–zero CAR stages, with 12 stages per octave at the high-frequency end. Every fifth output tap (or channel) is shown with heavy solid curves, for the same three damping factors as before; at the middle damping, all channels are plotted, with light dashed lines. The pole frequencies range from about 9900 Hz (2.818 radians per sample) down to about 30 Hz, based on equal spacing on a Greenwood map and a 22050 Hz sample rate. Peak locations of responses at the lowest damping define the characteristic frequency (CF) values used in subsequent plots.

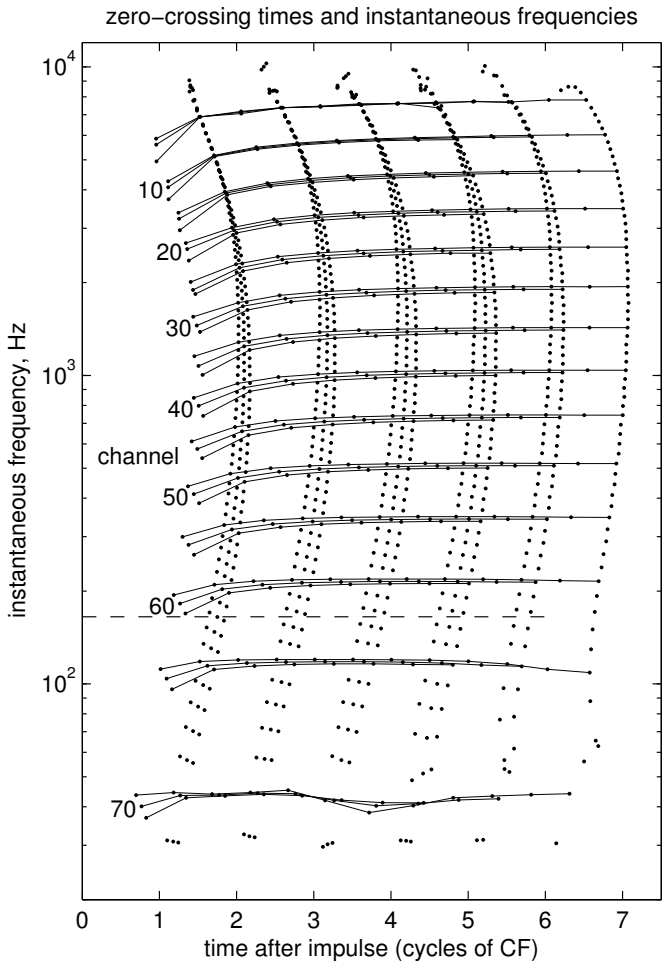


Figure 16.7: The instantaneous frequencies of the impulse responses of the 71 CAR channels, at the three damping levels, as a function of normalized time (cycles of CF after the impulse). Dots mark positive-going zero crossings of every channel, and negative-going zero crossings of every fifth channel. Fewer zero crossings are plotted for impulse responses with higher damping, since they decay sooner. Instantaneous frequencies are estimated near each zero crossing via Hilbert transforms of the impulse responses. Upward glides of about 20% are apparent for higher-CF channels. Lower-CF channels show less upward glide, but not much of the downward glide reported in the auditory nerve (Carney et al., 1999). The dashed horizontal line marks the break frequency in the Greenwood frequency map (see Figure 14.9), below which the channel spacing approaches linear instead of geometric. The zero-crossing times are seen to move through less than 1/4 cycle as the system adapts its gains through about a 40 dB range.

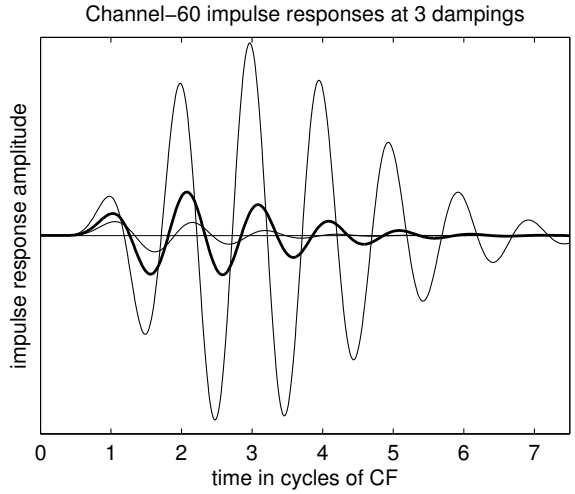


Figure 16.8: The impulse responses of channel 60 of the 71-channel linear CAR model at the three different dampings. The not-quite-aligned zero crossings are apparent. The smaller impulse responses correspond to higher dampings, as would be used at higher levels. The domain spans 7.5 cycles of CF, so the zero crossings align with those plotted in the previous figure. The group delays range from about 2 cycles at high damping (high level) to about 3.5 cycles at low damping (low level); see next figure.

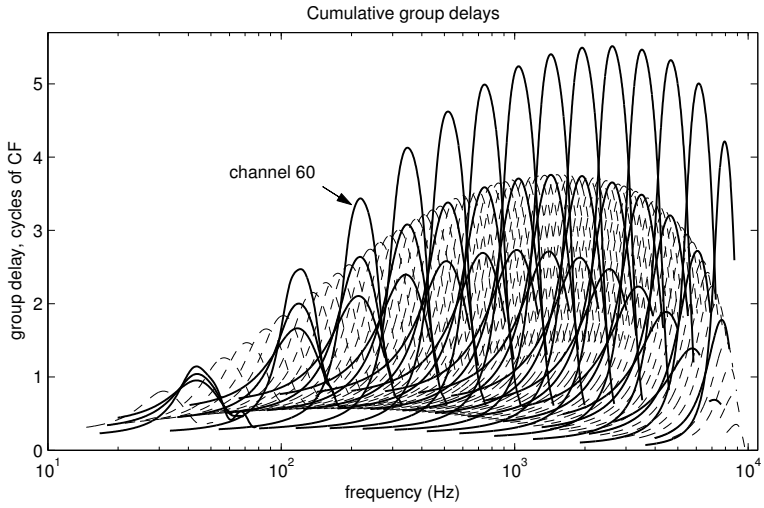


Figure 16.9: Group delays of the linear filter cascade, in units of cycles of the channel's CF, for damping factors 0.1, 0.2, and 0.3, plotted as in Figure 16.6. Channel 60, whose impulse response was plotted in the previous figure, is pointed out; it has a CF near 220 Hz. Delays peak near CF, as can be seen by comparison with Figure 16.6. To reduce clutter, the curves are cut off where the cascade gain is below 1 dB on the low-frequency side or below -3 dB on the high-frequency side. Near their peaks, the filters have about a cycle of delay per 10 dB of gain. The largest absolute time delay, near the apex, or low-frequency, end of the cochlea, is about 20 ms—one cycle of 50 Hz or two cycles of 100 Hz. As in Figure 16.6, every fifth channel is shown, except at the middle damping, where other channels are shown dashed.

Chapter 17

The Outer Hair Cell

The CA (cochlear amplifier) model explains the detection of small differences in time as well as in frequency, the dual character of the electrocochleogram, recruitment of loudness in cochlear hearing impairment, the long latency of normal neural responses near threshold, acoustic emissions (both stimulated and spontaneous) and the locus of TTS (temporary threshold shift) in the frequency range above the exposure tone. Both the classical high-intensity system and the active low-level CA system are highly nonlinear and they combine to compress the great dynamic range of hearing into a much narrower range of mechanical movement of the cilia of the inner hair cells.

— "An active process in cochlear mechanics," Hallowell Davis (1983)

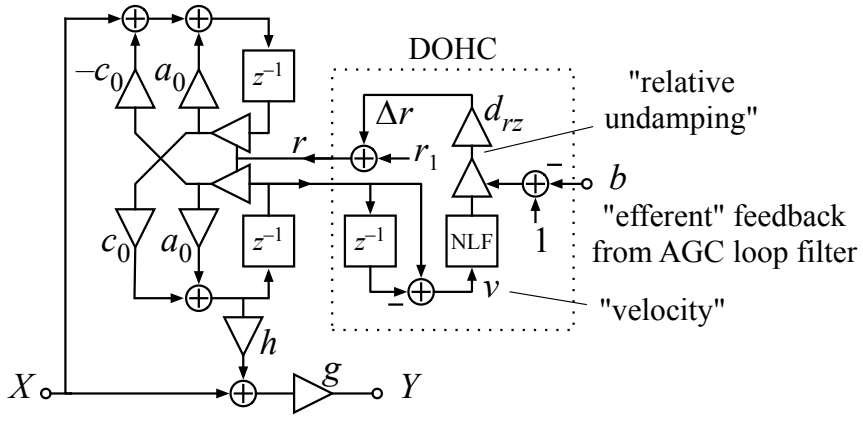


Figure 17.1: The linear filter stage of Figure 16.1 is here extended to incorporate nonlinearity via dynamic variation of the pole and zero radii (r) through functions localized into a digital outer hair cell (DOHC) block. The block computes a velocity (difference across a one-sample delay), then computes a damping (or a *relative undamping*, really), and finally computes and applies a corresponding r coefficient, incorporating both a local instantaneous nonlinearity based on the velocity, as well as "efferent" feedback from an AGC loop filter.

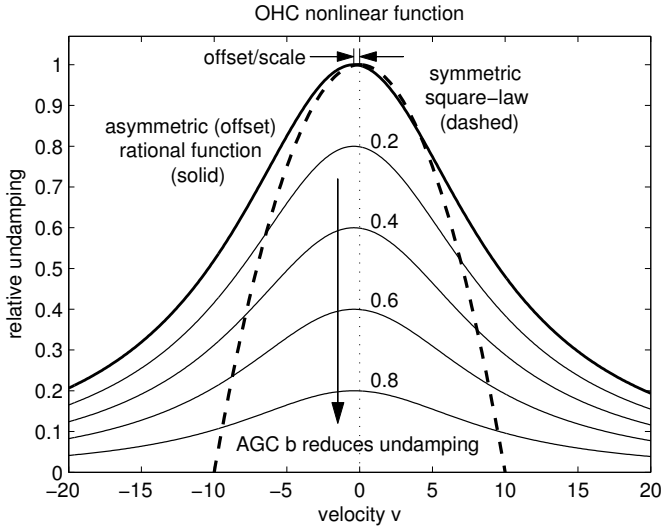


Figure 17.2: The NLF of the DOHC block in Figure 17.1 is shown here as the heavy solid curve. The dashed curve illustrates the sort of symmetric quadratic nonlinearity often used in Hopf oscillators, the Kim model, and various other cochlear models. The lighter solid curves show how feedback from the AGC loop filter multiplies the NLF output by $1 - b$, reducing the relative undamping that the DOHC supplies via this NLF.

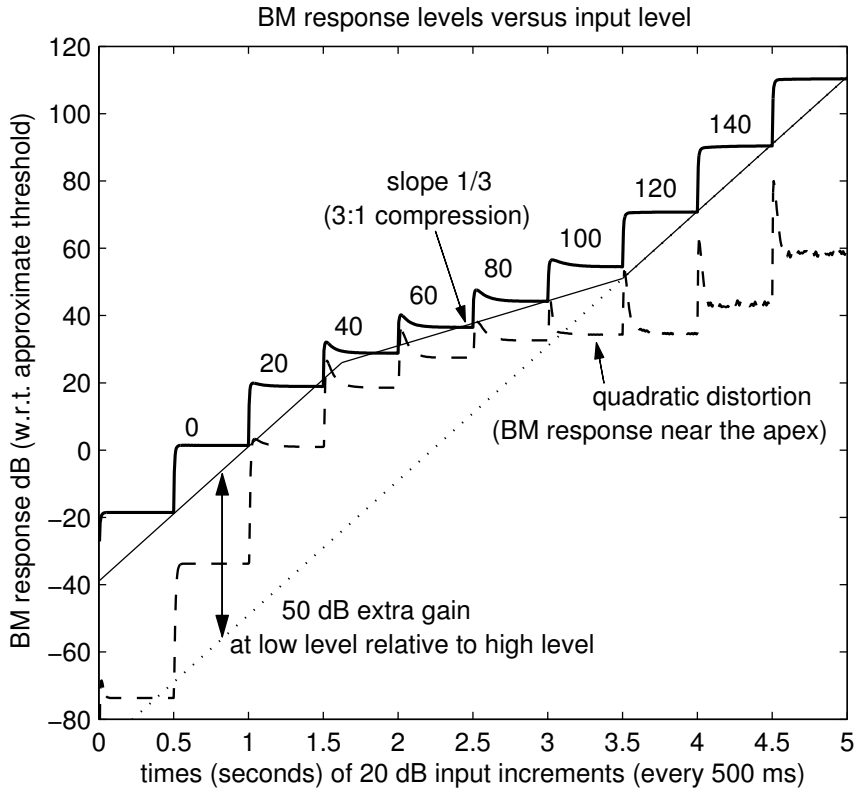


Figure 17.3: CARFAC responses versus stepped input level, at two places (solid and dashed), for a 4-tone input (1.6, 1.8, 2.0, and 2.2 kHz). The response for a place with CF near 1.7 kHz (solid) is compressive, but approaches linear at both very high and very low levels; thin lines approximate the steady-state response levels at which each input level step settles. The low-level linear region has 50 dB of gain compared to the high-level linear region; compare Figure 15.3. The DOHC scheme with offset asymmetry leads to a fairly high level of response to quadratic distortion products at a place with CF near 200 Hz (dashed), approximately tracking the level of response to the primary tones through most of the normal compressive range of hearing. The low-level and high-level linear regions generate relatively less distortion, as indicated by the relative response level at the QDT place. In a real ear, other parts of the system would likely distort strongly at high levels.

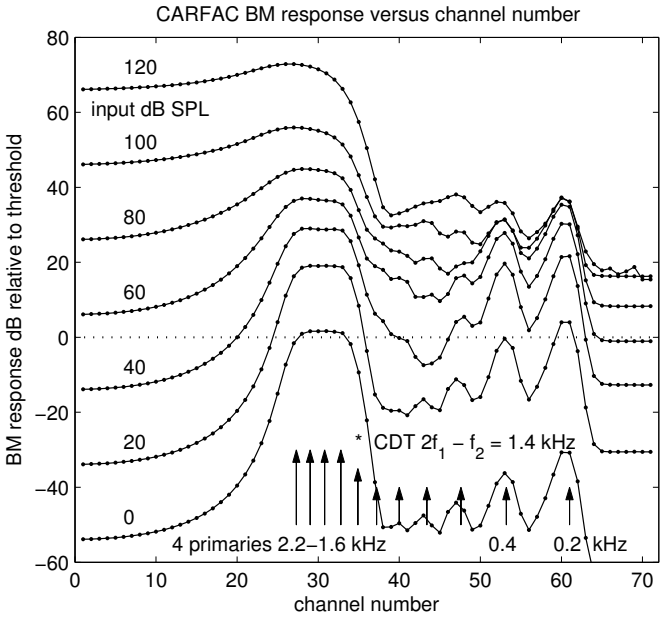


Figure 17.4: The CARFAC's steady-state BM response level at all places (channels), for some of the input levels used in Figure 17.3. The input is a four-tone complex at 1.6, 1.8, 2.0, 2.2 kHz (longer arrows mark places with CFs corresponding to these primaries). Places with CFs at lower multiples of 200 Hz (shorter arrows) also respond, especially to quadratic distortion at the 200 Hz and 400 Hz places. The first low-side odd-order distortion frequency, whose 1400 Hz place is marked with an asterisk, can be interpreted as the $2f_1 - f_2$ CDT of the two lowest primaries. Though it is not spatially resolved, the response at this place is dominated by a 1400 Hz component. The horizontal dotted line at 0 dB response level represents an approximate detection threshold, suggesting that quadratic distortion may be audible even for an input level as low as 20 dB SPL.

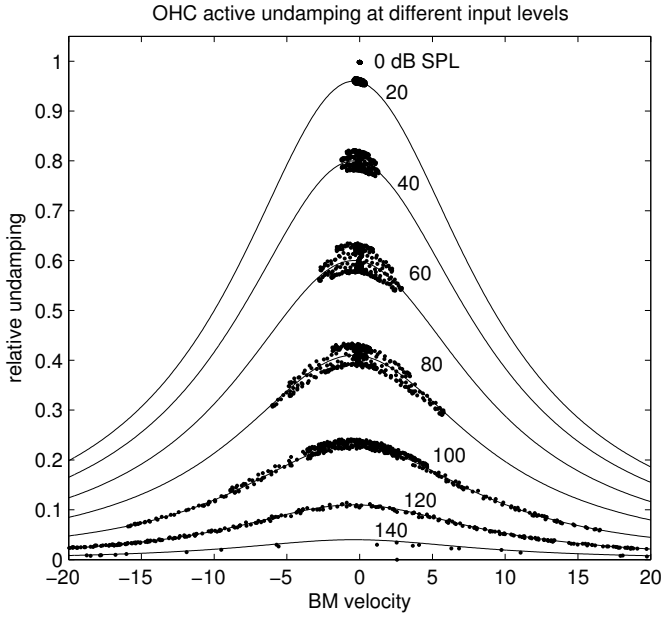


Figure 17.5: The OHC active undamping effect at various input levels, with 0–140 dB SPL input levels labeled. The input is a four-tone complex at 1.6–2.2 kHz, and the OHC effect is sampled near the most responsive place. Thin solid curves are scaled copies of the NLF that the points approximately fall on; that is, where the points might be for steady values of b, as in Figure 17.2. At the highest and lowest levels, the damping is nearly constant throughout the period of the stimulus, so relatively little distortion is generated (at 140 dB SPL, the BM velocity extends far outside the domain plotted, so the small bump in the middle has little effect).

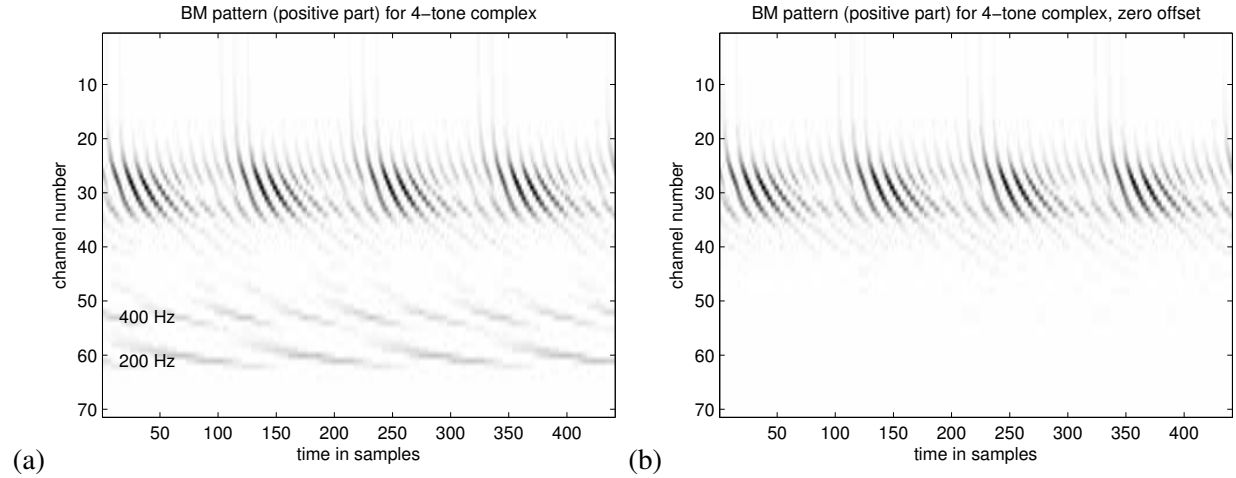


Figure 17.6: These cochleagrams show the positive part of the BM motion (filter outputs) for the 4-tone complex stimulus at 60 dB SPL, the level at which the relative amounts of 200 Hz and 400 Hz quadratic distortion tones is highest with the default NLF offset parameter. The left (a) image is for the default CARFAC, and the right (b) is with zero offset in the NLF. The 20 ms segment encompasses four cycles of the 200 Hz missing fundamental. The relatively small offset asymmetry shown in Figure 17.2 and Figure 17.5 is enough to cause relatively large QDTs.

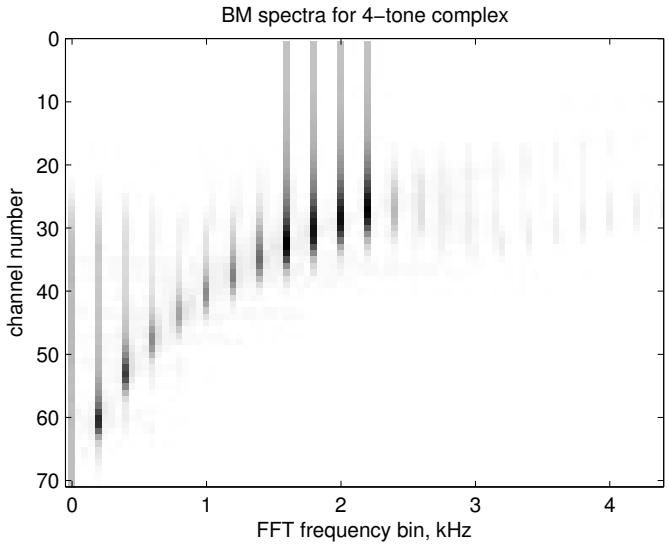


Figure 17.7: The CARFAC BM spectra for all channels, in response to the 4-tone complex at 60 dB SPL; that is, FFT magnitudes of the 20 ms segment shown in Figure 17.6(a), with each channel's spectrum plotted as a row. Visible distortion components include QDTs, including a DC response, and CDTs of all orders. For example, channels 35–40 show strong CDT $2f_1 - f_2$ (1400 Hz) and $3f_1 - 2f_2$ (1200 Hz) components (relative to the two lowest-frequency primaries). The tails above the peaks show where each DT component propagates from: primaries from the base, and distortion products from the region of strong response to the primaries. High-side distortion tones are very weak, as they have no chance to propagate through a region that amplifies them. The amplitude scale is cube-root compressed (sixth root of power) to make weak components visible in this plot.

Chapter 18

The Inner Hair Cell

> The sensitivity of the hair cells is extraordinary: the slope of the input–output curve can reach 20 mV per micrometer of displacement. If hair cells, like photoreceptors, can synaptically transmit statistically significant signals corresponding to 10 μV receptor potentials, the threshold sensitivity of the amphibian sacculus would approximate 500 pm (5 Å).
>
> — "Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli," Hudspeth and Corey (1977)

Biophysics Connection: How Hair Cells Work

Hudspeth and Corey (1977) found that deflection of hair bundles causes a change in the ionic currents into hair-cells, following a sigmoidal curve of the sort described in the text. It was later conjectured (Hudspeth, 1982), and eventually accepted, that these currents are primarily through the stereocilia, via *mechano-electrical transducer* (MET) channels at their tips (Jaramillo and Hudspeth, 1991; Lumpkin and Hudspeth, 1995). In particular, current flows into each stereocilium where it is tied to the next longer one by a *tip link*, a thin chain of proteins that is tensioned when the cilia bundles are bent in one direction, and loosened when they are bent in the other direction. See Figure 18.1.

The tip link mechanically opens and closes the MET channel, through which positive ions (potassium and calcium, mostly) in the endolymph enter the hair cell in this first step of the mechanical to neural transduction. Many details of how this transduction works, including tip-link molecular mechanisms, have been worked out (Gillespie and Müller, 2009; Sakaguchi et al., 2009), though details remain elusive (Fettiplace and Kim, 2014; Zhao and Müller, 2015).

At rest, the MET channels rapidly *flicker* between open and closed, under thermal agitation, with a probability of about 0.1 or more of being open. For high displacements, at some point most of the channels are open and more displacement will not further increase the current; for displacements in the other direction, most are closed and the current will approach zero. Though there are only about two channels per stereocilium, each can pass a large ion current. The statistics of these conductance fluctuations give rise to a sigmoidal (*s*-shaped) detection nonlinearity.

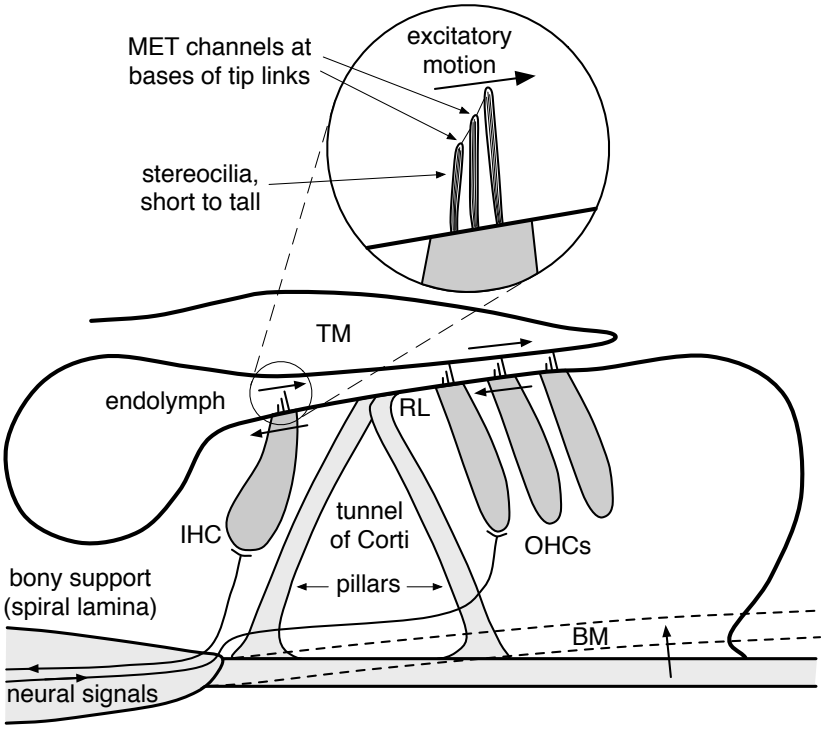


Figure 18.1: In the organ of Corti, the hair bundles of the inner and outer hair cells (IHC and OHC) are displaced by a shearing motion between the reticular lamina (RL) and the tectorial membrane (TM) when the organ of Corti pivots about the inside corner of the tunnel of Corti due to displacement of the basilar membrane (BM; exaggerated displaced position shown dashed). When the motion is in the direction of the arrows, the tip links between adjacent stereocilia (hairs) pull the mechano-electrical transducer (MET) channels open, allowing a positive-ion current to flow from the endolymph into the hair cells at the tips of the shorter cilia. The OHCs feed energy back into the hydromechanical wave, to a degree controlled by neural signals from the brain, while the IHCs send neural signals toward the brain.

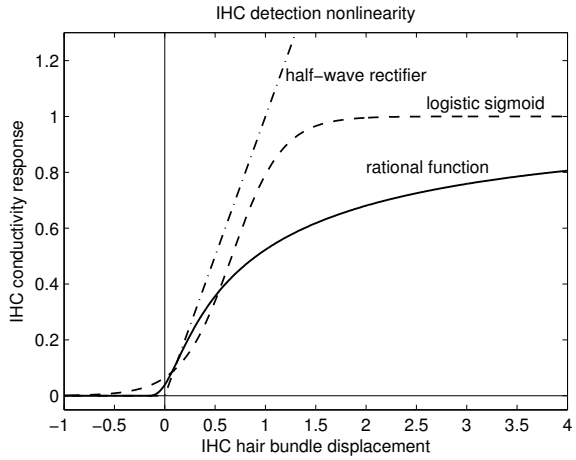


Figure 18.2: The transduction nonlinearity of the IHCs is some kind of a *sigmoid*, such as a displaced logistic function (dashed), and is sometimes modeled as simply a half-wave rectifier (dash-dot line). Other functional forms can also be used; for example, a constant segment at zero response, connected to a rational function (ratio of cubic polynomials) for the rest (solid curve), giving a cubic foot shape, a nearly linear middle region, and a slowly saturating shoulder.

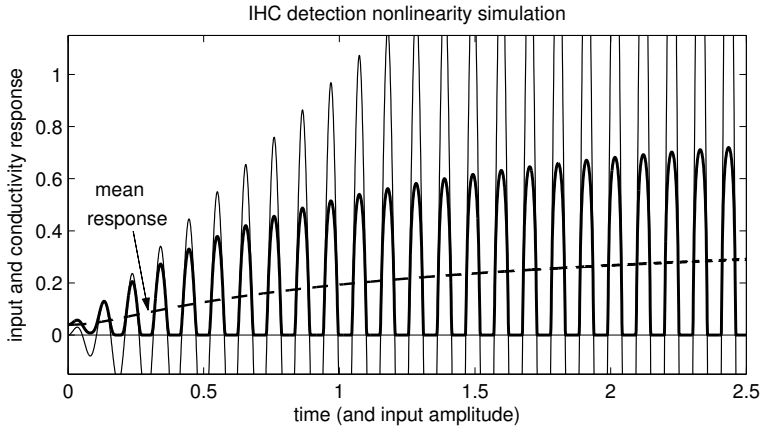


Figure 18.3: A simulation of the output of the rational-function detection nonlinearity of Figure 18.2 (heavy solid curve), when it is driven by an increasing-amplitude sinusoidal input (thin solid curve); this output function of time is the conductance $g(t)$ used in the digital IHC model of Section **??**. The mean response (dashed curve) is also shown; for the purpose of this illustration, the mean is taken over many phases of the input sinusoid.

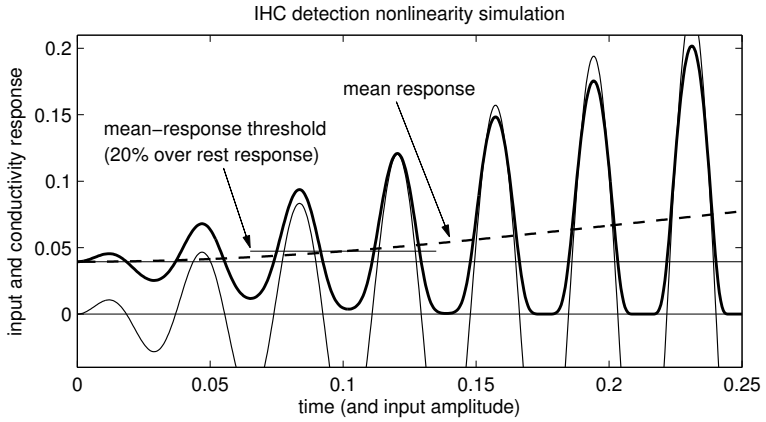


Figure 18.4: A simulation as in Figure 18.3, but for a more limited time and amplitude range, and using a higher frequency, to better illustrate the transition from a nearly linear response at low amplitude to a rectifying response at higher amplitudes. The mean response (dashed curve) increases very slowly (initially quadratically) where the response is nearly linear. It increases from the rest level to 20% higher as the input amplitude increases to about 0.1, where the response is distorted enough to start to look rectified.

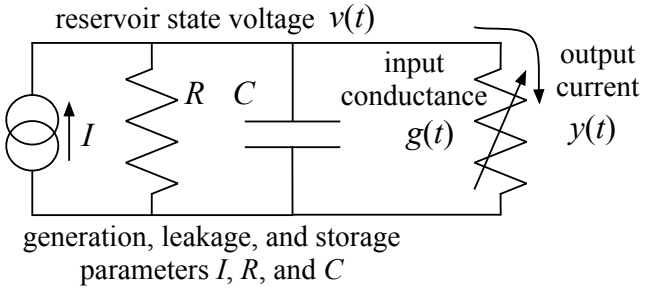


Figure 18.5: The Schroeder–Hall hair-cell model is described by this circuit schematic. The state variable is the voltage $v(t)$ across the capacitor, which is charged up by the current source I and discharged by currents through the fixed resistance R and through the input-controlled variable resistor with conductance (reciprocal of resistance) $g(t)$. The current through the variable resistor (the resistor symbol with an arrow through it) is the output signal. The same schematic can describe the Allen model, but there a saturating nonlinearity is used instead of the HWR for the $g(t)$ detection nonlinearity, and an output smoothing filter is added to reduce synchrony to high frequencies.

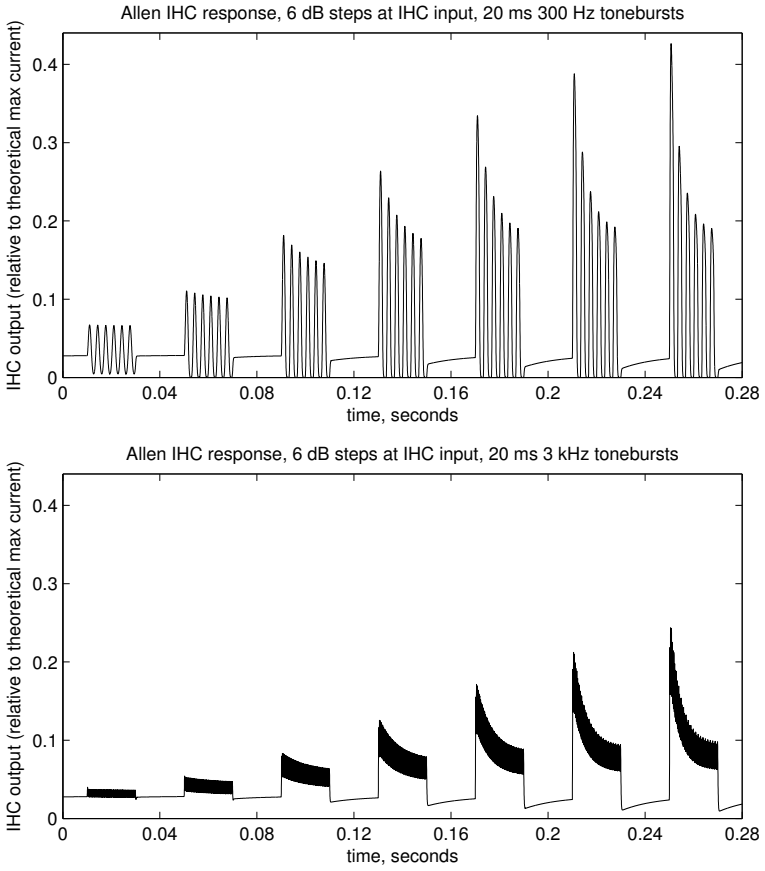


Figure 18.6: The response of the Allen IHC model to 20 ms 300 Hz (top) and 3 kHz (bottom) tone bursts in 6 dB increments, starting 6 dB below the mean-response threshold (with no cochlear filtering). A fairly strong onset emphasis at high levels, such as that exhibited here, is a key property of inner hair cells and their models. The 3 kHz synchrony is attenuated by the lowpass filter. There remains very good synchrony to onsets.

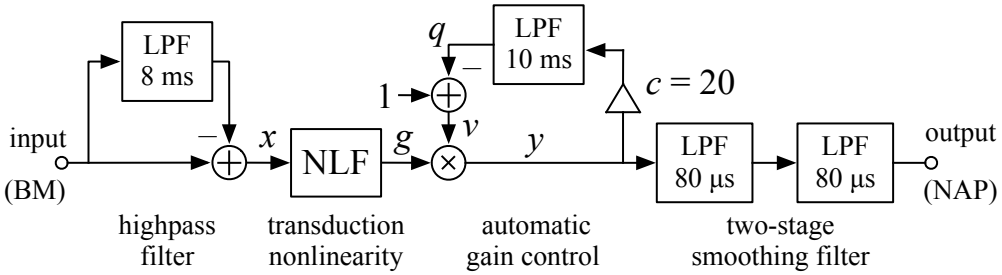


Figure 18.7: The Digital IHC block diagram, an adaptation of the Schroeder–Hall and Allen IHC models. The model uses four instances of the first-order IIR digital filter of Figure 7.1, configured as smoothing filters (lowpass with unity gain at DC). In the diagram, the lowpass filters are labeled LPF, with their respective time constants The first LPF is subtracted to make a highpass filter to suppress frequencies below 20 Hz that are generated from quadratic distortion in the BM wave propagation. The second LPF is the loop filter in an automatic-gain-control loop like the one of Figure 11.2, with $K = 1$ in the nonlinear gain-control function, but with rectifying nonlinearity before the variable gain rather than after. The rectifying nonlinear function (NLF) that converts BM motion to a membrane conductance is a soft rectifying rational-function sigmoid like the one shown in Figure 18.2. The variable gain v models the capacitor voltage of Figure 18.5. The final two LPFs smooth the output.

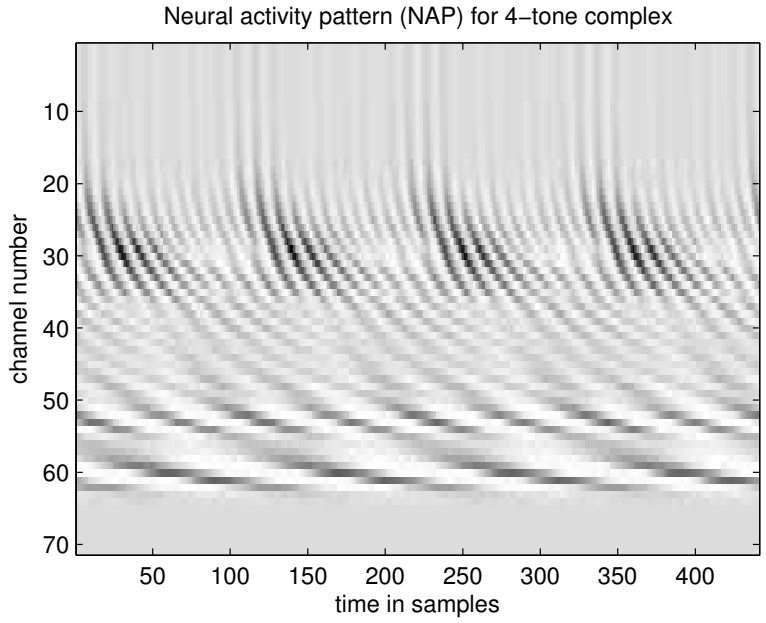


Figure 18.8: The neural activity pattern (NAP), the response of a bank of DIHCs in the context of the CAR-FAC, with the 4-tone stimulus of Section **??**, at 60 dB SPL. The NAP represents, at least conceptually, the instantaneous firing rates of the groups of primary auditory neurons attached to each IHC.

Chapter 19

The AGC Loop Filter

> ... the output (BM displacement or velocity) varies much less than the stapes input displacement or velocity, for frequencies near the best frequency. The significance of this important finding will become clearer as we proceed, but, in my opinion, it is a precursor to an automatic gain control system which seems to be built into the cochlear filters.
>
> — "Cochlear modeling – 1980," Jont B. Allen (1981)

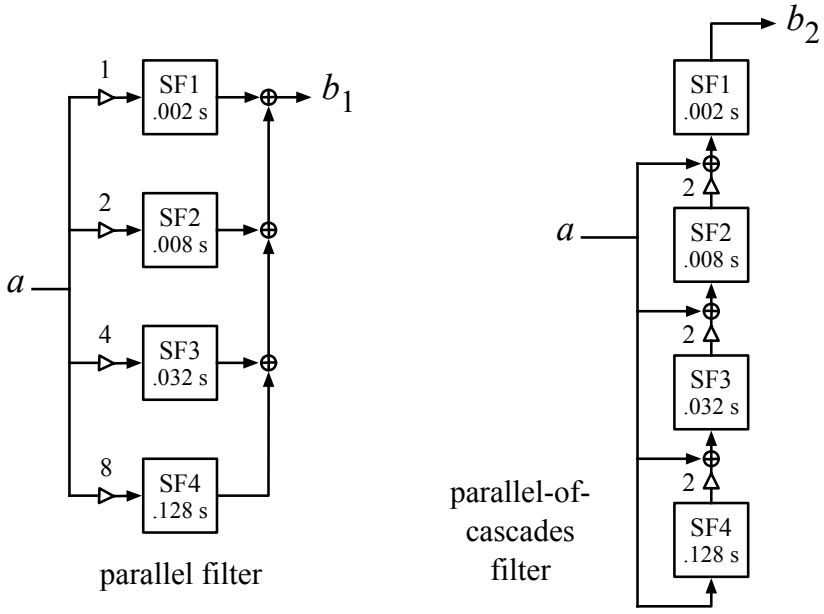


Figure 19.1: The filters in each AGC channel are conceptually made from four first-order lowpass smoothing filters with different gains and time constants, in parallel, as shown on the left, and as introduced in Figure 11.12. The individual smoothing filters (SFi) have transfer functions $1/(\tau_i s + 1)$, with time constant $\tau_1 = 2$ ms, and increasing by factors of four up to $\tau_4 = 128$ ms. But we use the variant shown on the right: parallel combinations of cascades of these first-order filters, sharing pieces in cascade, because this arrangement makes it easier to run the filters with long time constants at lower sample rates.

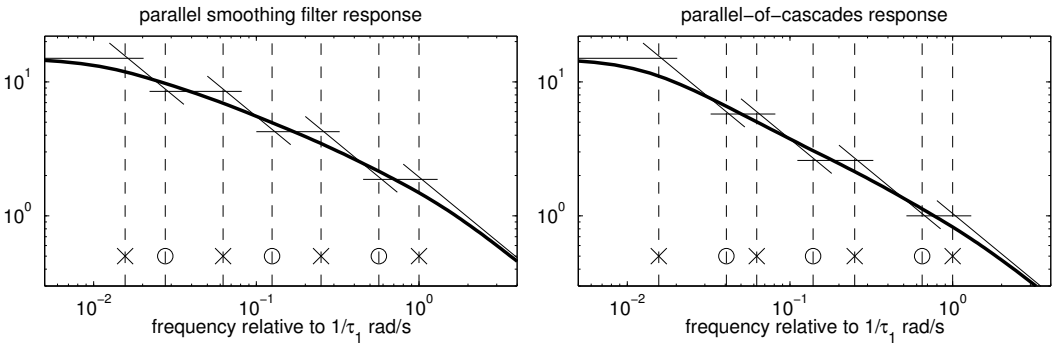


Figure 19.2: Since real poles and zeros induce a slope change of 6 dB per octave in Bode-plot asymptotes, we can use the calculated pole and zero frequencies to directly draw the skeletons of Bode plots for the four-pole smoothing filters. Each pole or zero corresponds to a corner between intersecting segments of slope 0 and –6; since the slopes are known, the frequencies are enough to easily construct the skeleton of the Bode plot as shown. The transfer function will be a smooth curve bounded alternately above and below by these corners, with slope between these asymptotic slopes. For the given time constants and gains, the zero frequencies and resulting slopes are slightly different between the parallel filter (left) and the parallel-of-cascades filter (right).

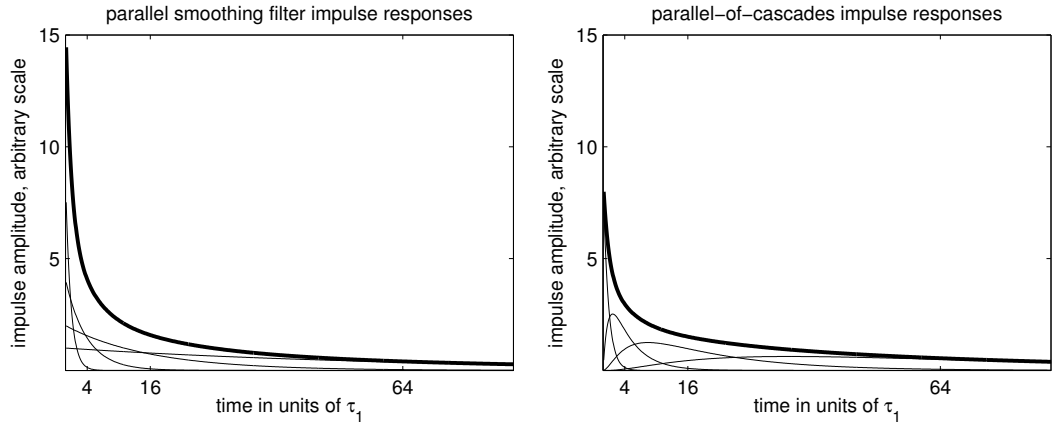


Figure 19.3: The impulse responses of the four-pole smoothing filters are easily found by adding up the impulse responses of their four paralleled parts, as shown here. On the left, the parts are the exponential decays of the first-order filters alone, while on the right they are the impulse responses of cascades of 1, 2, 3, or 4 first-order filters. Of course, since the two systems share the same poles, the total impulse response of either can be described as a weighted sum of those exponentials in the left plot, but with different weights. Therefore, the two structures can be completely equivalent if we generalize the gains; but for the default CARFAC we choose the parallel-of-cascades with gain factors of two.

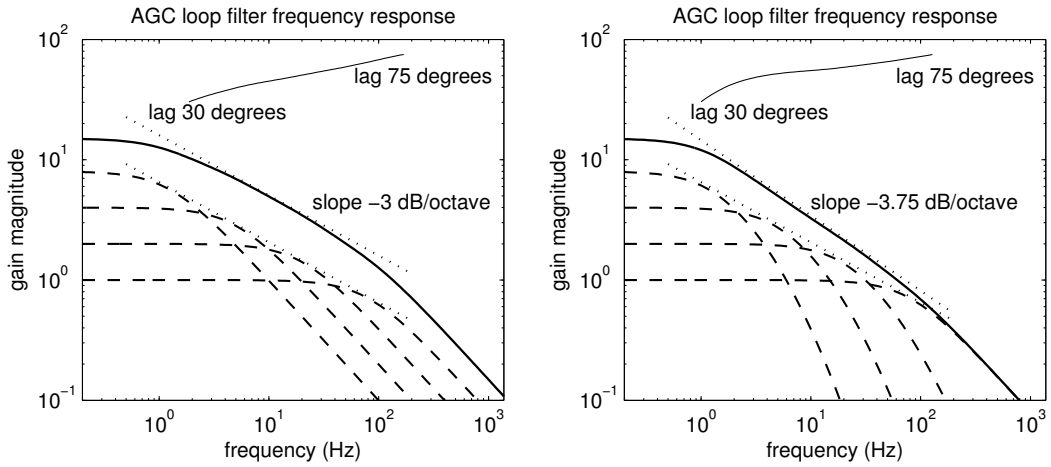


Figure 19.4: The frequency responses of the AGC smoothing filters (solid curves) can be found directly by adding up the complex gains of the four paralleled filters (dashed curves show their magnitudes), rather than via a pole-zero analysis leading to a Bode plot as in Figure 19.2. The purely parallel interconnection of first-order filters (the leftmost filter in Figure 19.1) gives the response on the left, while the parallel-of-cascades variants give the response on the right. The phase lag stays within a moderate 30 to 75 degrees over more than two orders of magnitude of frequency (upper curve, degrees of phase lag, on log scale). The results match Figure 19.2 using $\tau_1 = 0.002$ s.

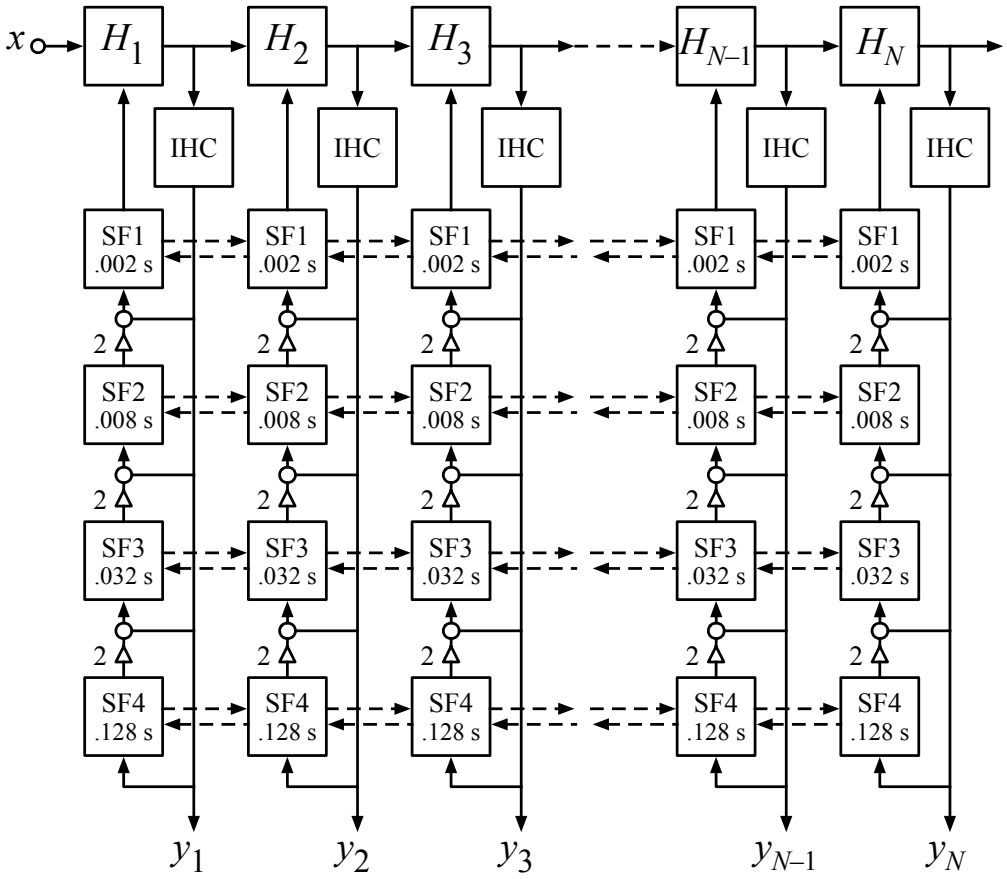


Figure 19.5: The filters in each AGC channel are based on the parallel-of-cascades version in Figure 19.1. In this configuration, the faster filters define a shorter (more local) loop, and it is easier to run the more-remote slower filters at lower sample rates, since their output steps will be smoothed by other filters before being applied to control the CAR filter stage damping. Neglecting the sampling approximations, the loop-filter transfer functions are as shown on the right in Figure 19.4. The lateral interconnections are shown dashed; the mechanism for spatial cross-coupling is detailed in the next figure.

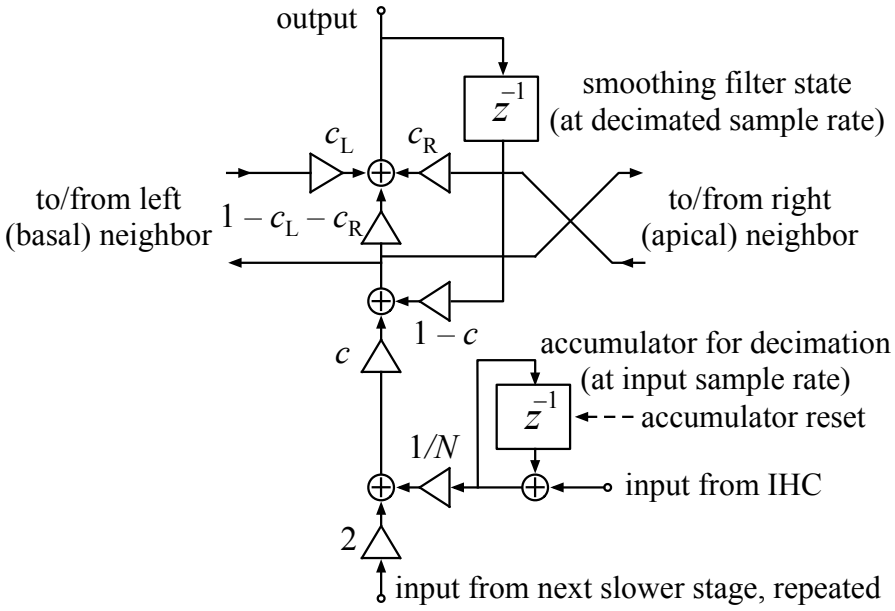


Figure 19.6: The unit smoothing filter (one stage, one channel) of the coupled AGC filter is drawn here in a bottom-to-top flow arrangement, as it is used in Figure 19.5. At the bottom right, input values at a high sample rate (from the IHC, at the CAR's audio sample rate) are accumulated until it is time for the unit to operate. After accumulation of N samples (N being the decimation factor), the accumulator value is divided by N and used as input to the smoothing filter, and the accumulator is reset. The input from the next slower stage, if there is one, is also added in, with a weight of 2; if that stage has a higher decimation factor, each of its output samples may be used as input more than once. The c coefficient controls the time-domain smoothing time constant. In the spatial smoothing part, the 3-point FIR filter $[c_L, (1 - c_L - c_R), c_R]$ applies weight c_L to the value from the left neighbor, c_R to the value from the right neighbor, and enough gain to the current channel to keep the total mixing gain equal to 1.

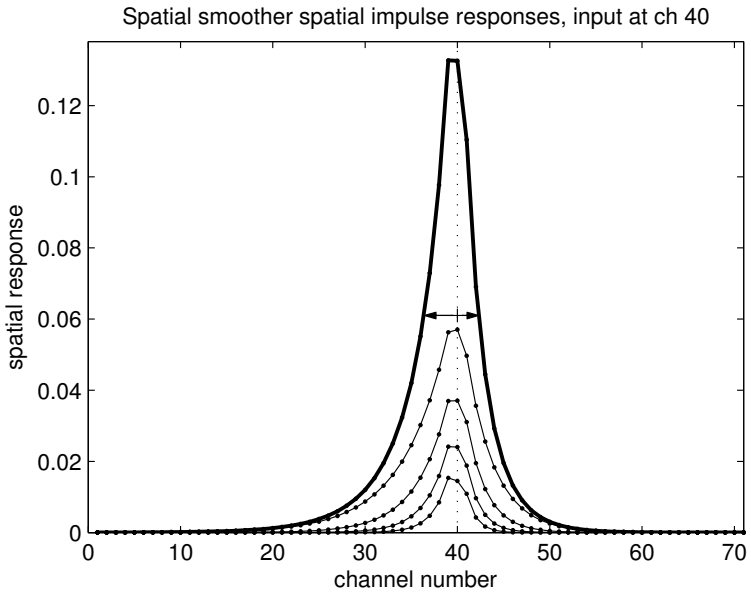


Figure 19.7: Spatial impulse responses of the AGC loop filter's spatial smoothing filters (four lower curves, which include their respective power-of-2 weights), and their sum (heavy upper curve), with an input at channel 40 only. The filters are designed with a moderate asymmetry of spread toward earlier channels (toward the base, or higher-CF end, of the cochlea) as indicated by arrows. The filters are 3-point FIR filters, applied to the AGC lowpass filter state arrays at each AGC update time, and effectively iterated many times by the long time constants of the temporal smoothing. For this figure, and the next one, the parallel paths were kept separate so that we could see the response of each part. For the fastest and most local stage, the coefficients used for this response are: $c_{\mathrm{L}} = 0.286$, $c_{\mathrm{R}} = 0.404$, $c = 0.166$. This c value represents a time constant of about 6 samples at the decimated rate, so the 3-point FIR smoothing is effectively applied about 6 times per time constant, resulting in the Gaussian-like spread seen on the lowest curve. Slower stages have more time to spread further.

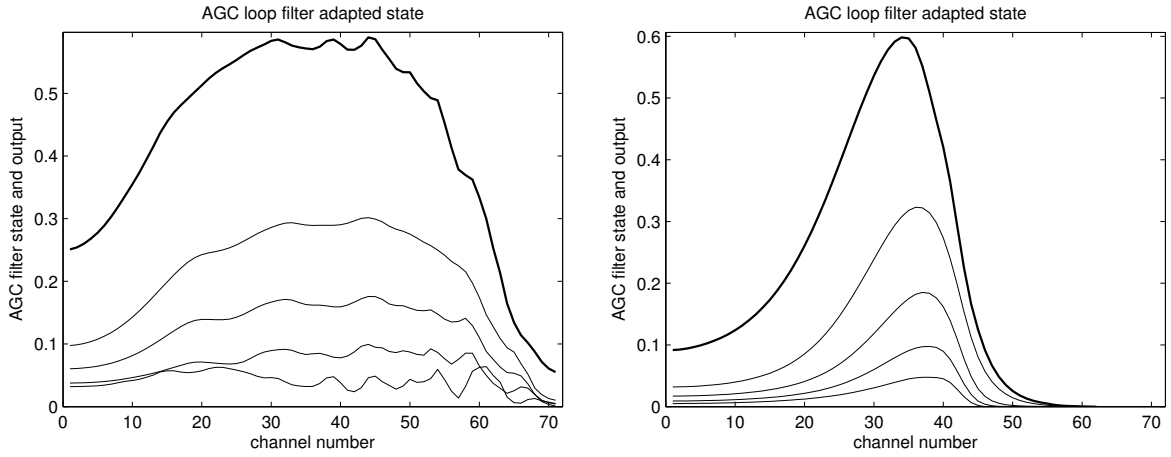


Figure 19.8: Typical states of four AGC parallel smoothing filters (from the left filter of Figure 19.1, lower curves), and their sum (the y_1 signal, upper heavy curves), for a speech sound (left), and a 1 kHz tone (right). We illustrate states of the parallel filter here, rather than the parallel-of-cascades form, so that each curve represents a time scale and its weight; the actual states in our parallel-of-cascades form of Figure 19.5 are roughly scaled cumulative sums of these, with a similar final result. The spatial smoothing is least on the lowest curve, the state of the fastest filter, and greatest on the state of the slowest filter. The spatial smoothing has been designed to be somewhat asymmetric, spreading more to earlier (more basal, higher-CF) channels than to later (more apical, lower-CF) channels, modeling the spread of MOC efferents toward more basal locations (see Figure 14.15). In particular, the place most responsive to the 1 kHz tone is channel 37, but the strongest AGC filter feedback comes at channel 34.

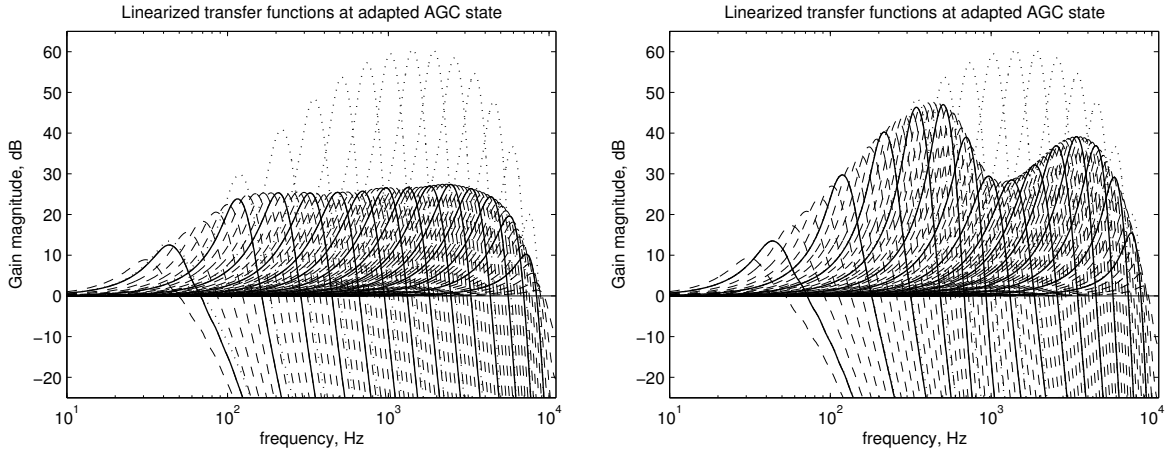


Figure 19.9: The linearized transfer functions of the CARFAC at the adapted AGC states shown in Figure 19.8. The gains of the middle high-gain channels have come down by more than 30 dB in reaction to the speech signal (left), compared to the gain in quiet (selected channels shown as upper dotted curves). The gain reduction is more localized to near the 1–2 kHz region when adapted to a 1 kHz tone (right).

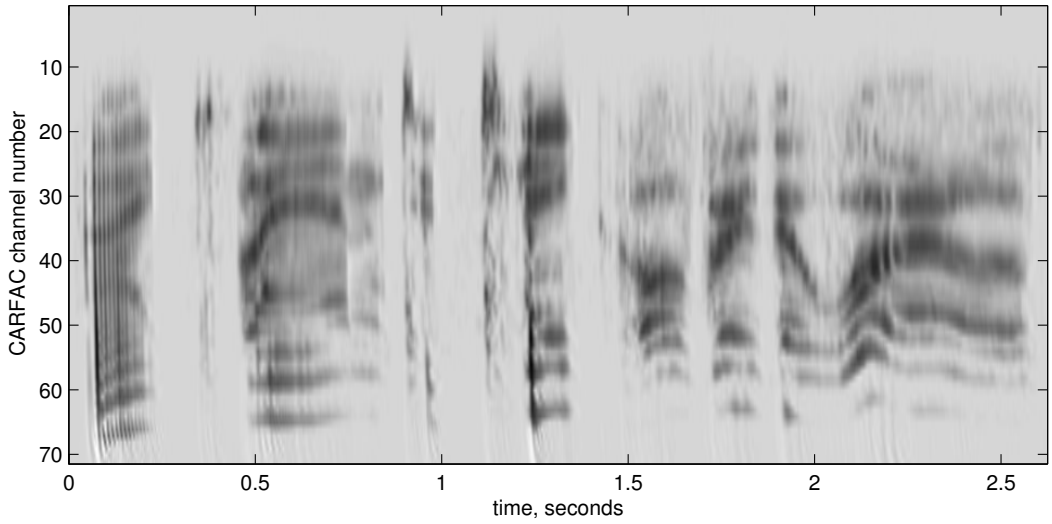


Figure 19.10: The average-rate (time-smoothed) neural activity pattern (NAP) of a few seconds of speech shows how the CAR, OHC, IHC, and AGC work together to make a clear alternative to a spectrogram. We sometimes subtract off the rest response level and clip to white to remove the gray background.

Part IV

The Auditory Nervous System

Part IV Dedication: J. C. R. Licklider

This part is dedicated to the memory of Joseph Carl Robnett Licklider (1915–1990). "Lick" is best known as one of the "fathers of the Internet" (Poole et al., 2005), based on his ARPA leadership and his writings such as "Man–Computer Symbiosis" and "The Computer as a Communication Device."

But before he was a computer network and systems guy, Lick was an auditory psychologist and modeler (November, 2012). His work on pitch perception, as represented in the "duplex theory," is the basis for much recent work in hearing, including my own, connecting the output of the cochlea to perception and neural processing of complex sounds.

I had the pleasure of meeting Lick just once, in 1984 at a Navy-sponsored workshop on "Artificial Intelligence and Bionics." I think he was a little surprised to see his duplex theory coming back as a practical computational approach, three decades after he came up with it. It has become even more practical since then, thanks partly to his computer innovations.

In this part, we discuss the levels of processing in the auditory nervous system. We develop the idea of auditory images, of the sort that are thought to be extracted by brainstem and midbrain for projection to auditory cortex.

We start where the last part left off, with the "cable" for the telephone theory of hearing, the auditory nerve, which transmits the vibrations as detected by hair cells in the cochlea to the first stop in the brainstem, the cochlear nucleus.

Several kinds of processing in the cochlear nucleus support both binaural hearing and the extraction of properties such as pitch and timbre that can be monoaural or binaural. We cover the extraction of such properties into the *stabilized auditory image*, a basis for sound representation in machine hearing systems as well as a model of representations in the inferior colliculus of the midbrain. We cover binaural spatial processing and the brainstem's olivary complex. We finish this part with a discussion of *auditory scene analysis* as the main aim of the auditory brain, and some ideas for how such analysis might be done in the thalamus and cortex to finally "extract meaning."

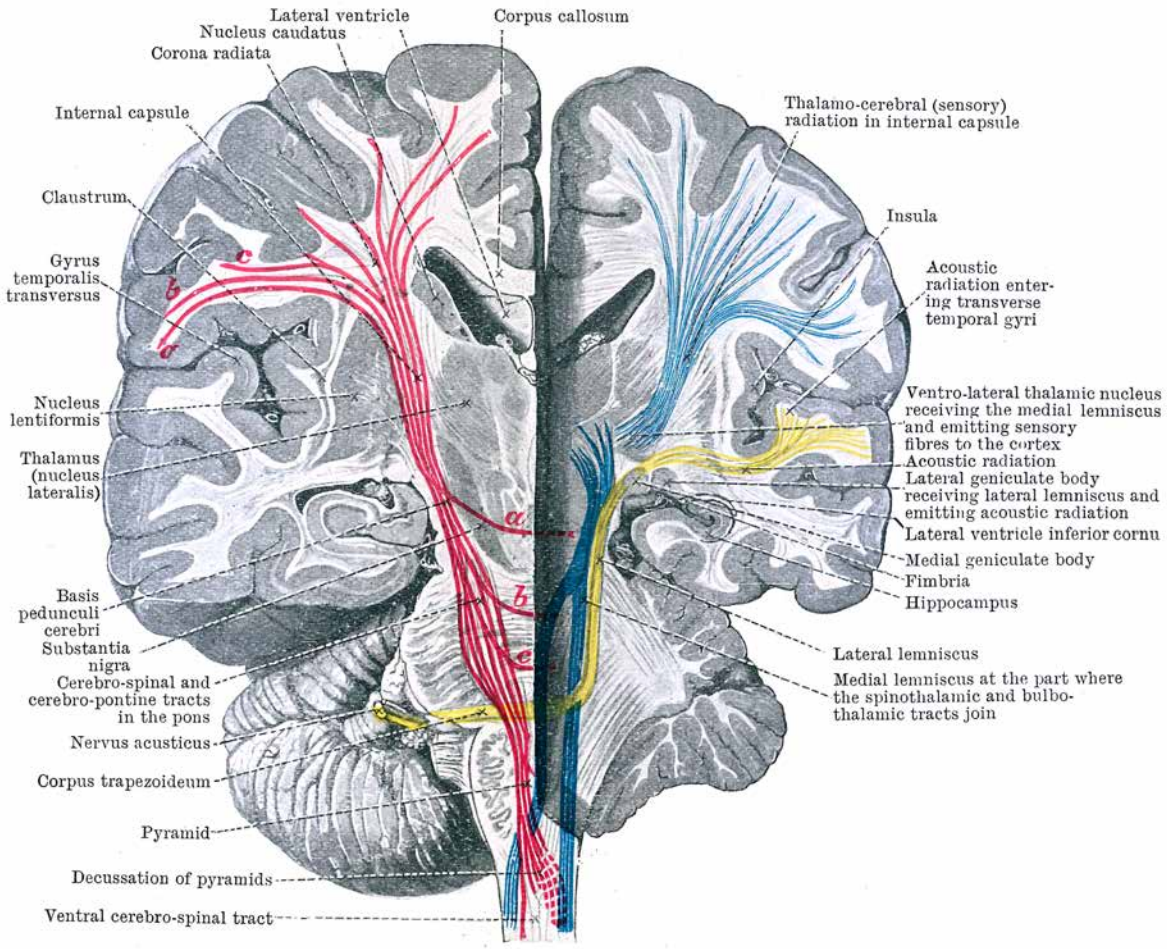


FIG. 506.—A VERTICAL TRANSVERSE SECTION OF THE BRAIN TO SHOW THE WHOLE OF THE CENTRAL ACOUSTIC PATH. The left hemisphere (right side of the figure) is cut on a plane posterior to that of the right. Motor fibres red. Sensory fibres blue. Acoustic fibres yellow.

The auditory nervous system was already fairly well mapped out a hundred years ago, as this color illustration from *Cunningham's Text-Book of Anatomy* shows (Cunningham and Robinson, 1918). Auditory fibers are dark gray here, yellow in the color plate.

Chapter 20

Auditory Nerve and Cochlear Nucleus

> I experimented in this way, and eventually found that I could send as many as 352 impulses per second along the nerve of a rabbit and get a note from the muscle of the pitch of 352 vibrations per second . . . but when I tried by more rapid stimulation of the nerve to get a higher note from the muscle, I failed. . . . Now, am I to conclude that, because I failed to get a higher note than one of 352 vibrations from the muscle, it is not possible to send more than 352 vibrations per second along a nerve? By no means . . .
>
> — "A lecture on the sense of hearing," Rutherford (1887)

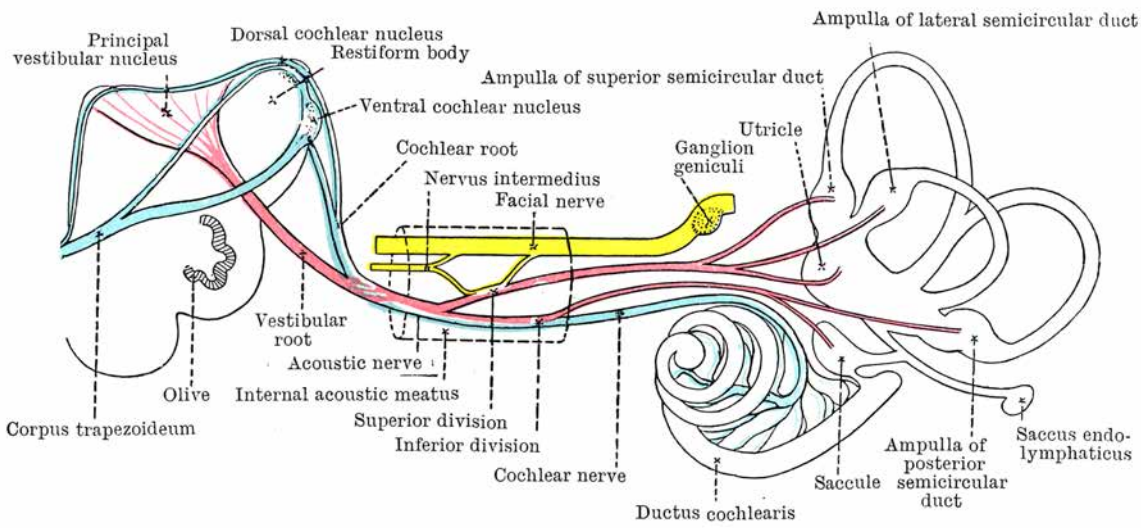


FIG. 587.—SCHEME OF THE ORIGIN AND DISTRIBUTION OF THE ACOUSTIC NERVE.

Figure 20.1: As shown in this color illustration from Cunningham and Robinson (1918), the acoustic nerve, or eighth cranial nerve, includes the cochlear division (dark gray here, blue in the color plate) that serves hearing, and the vestibular division (red in the color plate) that serves balance functions. After a stop at the dorsal and ventral divisions of the cochlear nucleus, the auditory pathway branches into the three acoustic stria, one of which, the ventral acoustic stria (the lower one here) goes to the superior olive on both sides, crossing via the trapezoid body. The facial nerve (yellow in the color plate) takes efferent signals back to the stapedius muscle in the inner ear, via the geniculate ganglion, to serve the protective acoustic reflex.

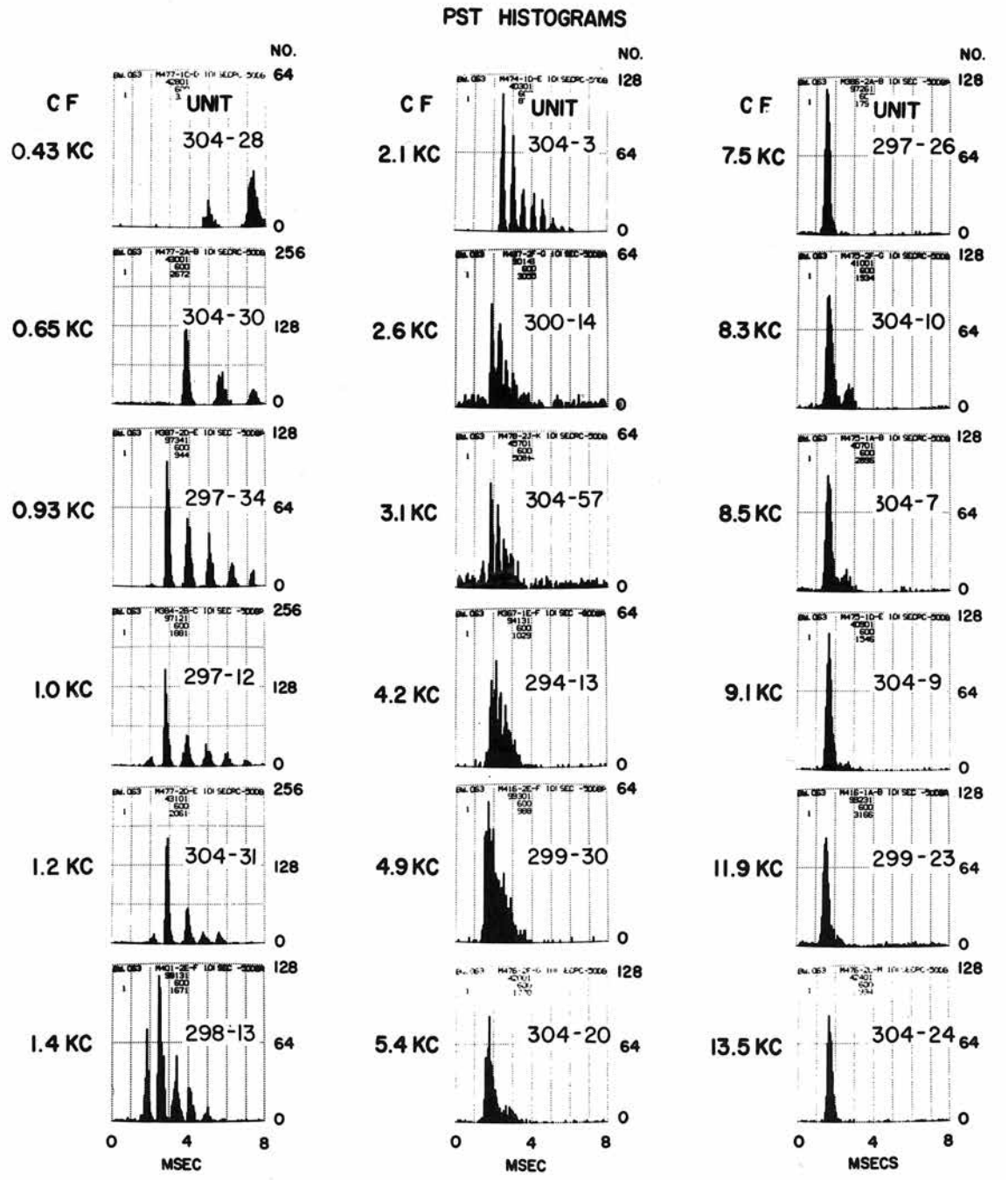


Figure 20.2: Kiang (1965) recorded times of action potentials on cat auditory nerve fibers, in response to brief clicks presented 10 per second, and summarized those firing times as post-stimulus-time (PST) histograms. Each histogram is labeled with the CF of the nerve fiber (the "unit") in KC, which is 1960s terminology for kHz (the unit numbers in each plot represent the animal number and the particular neuron). Notice that the fine time structure of the ringing of cochlear bandpass filters is reflected in the PST histograms for units with CF up to about 4 kHz. With plots sorted by cochlear place, or CF, as here, the graded latency to first response is also apparent, at approximately 1 ms plus 2 cycles of CF. [Figure 4.2 (Kiang, 1965) reproduced by permission of The MIT Press.]

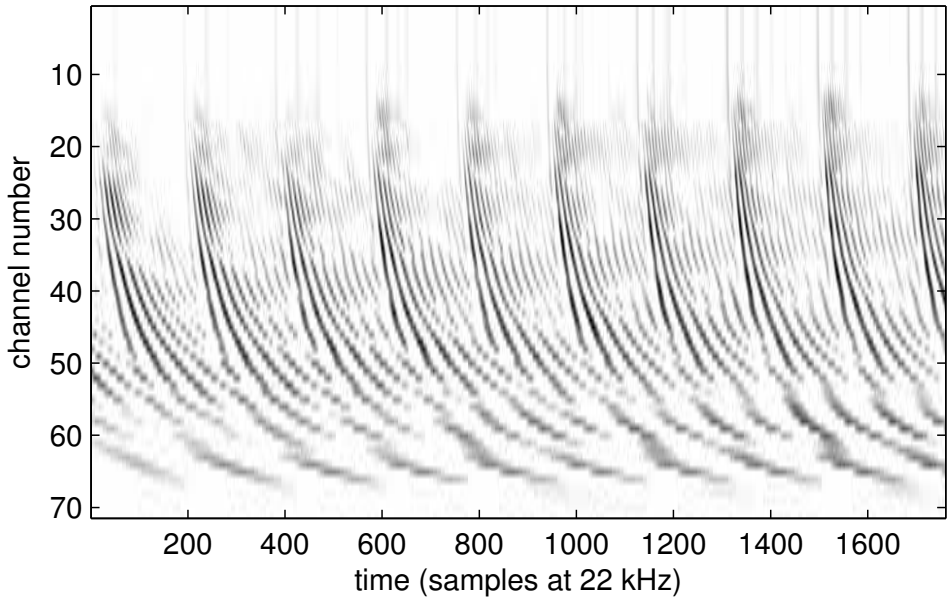


Figure 20.3: A segment of a cochleagram, showing 71 frequency channels as functions of time, in response to a spoken vowel. The IHC output is offset to put the rest response level at zero (white); positive excursions plot as dark regions, and below-rest excursions are clipped to white. This cochleagram spans less than a tenth of a second of sound, not even enough for one syllable of speech.

Chapter 21

The Auditory Image

We must think of the neural arrangement, therefore, as extended in two spatial dimensions. The one corresponding to frequency is the x-dimension, or the dimension of the nervous tissue into which the lengthwise dimension of the cochlea projects. The whole arrangement for determining autocorrelation functions is replicated in the x-dimension. The τ-dimension is functionally orthogonal to the x-dimension, and we can think of it, at least for convenience of graphical representation, as being spatially orthogonal, also. The over-all system, then, yields a representation of the stimulus $f(t)$ in two spatial dimensions and time, a running autocorrelation $\phi(t, \tau, x)$ of the components in each of many frequency bands.

— "A duplex theory of pitch perception," J. C. R. Licklider (1951)

[ht] In contrast to the two-dimensional visual and somatosensory receptor surfaces, the cochlea provides only a one-dimensional rendition of the impinging acoustic energy distribution along the organ of Corti. Consequently, cortical frequency maps can expand along the second dimension of the cortical sheet, providing additional territory for signal processing while closely preserving receptor-related neighborhood relationships.

— "Auditory cortex mapmaking: principles, projections, and plasticity," Schreiner and Winer (2007)

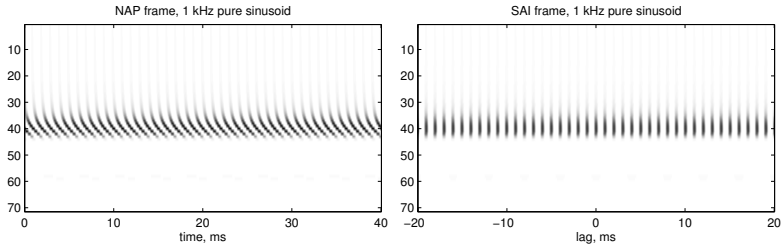


Figure 21.1: The NAP (left) and SAI (right) of a 1 kHz pure tone have a lot in common. Both are simple patterns with period matching the 1 ms tone period. The top part of each image shows a weak response of high-CF channels (low channel numbers) to the 1 kHz tone as it propagates through the cochlea; the middle shows a strong response for places with CF near 1 kHz; and the lower part shows essentially no response for channels with lower CFs. The NAP shows curvature patterns from propagation delays of the cochlear filtering, as we saw in Figure 18.8, and has no certain time origin that could be used to make a stable spatial pattern. The SAI's stabilization process straightens the pattern and stabilizes it, aligning peaks at the 0 lag point.

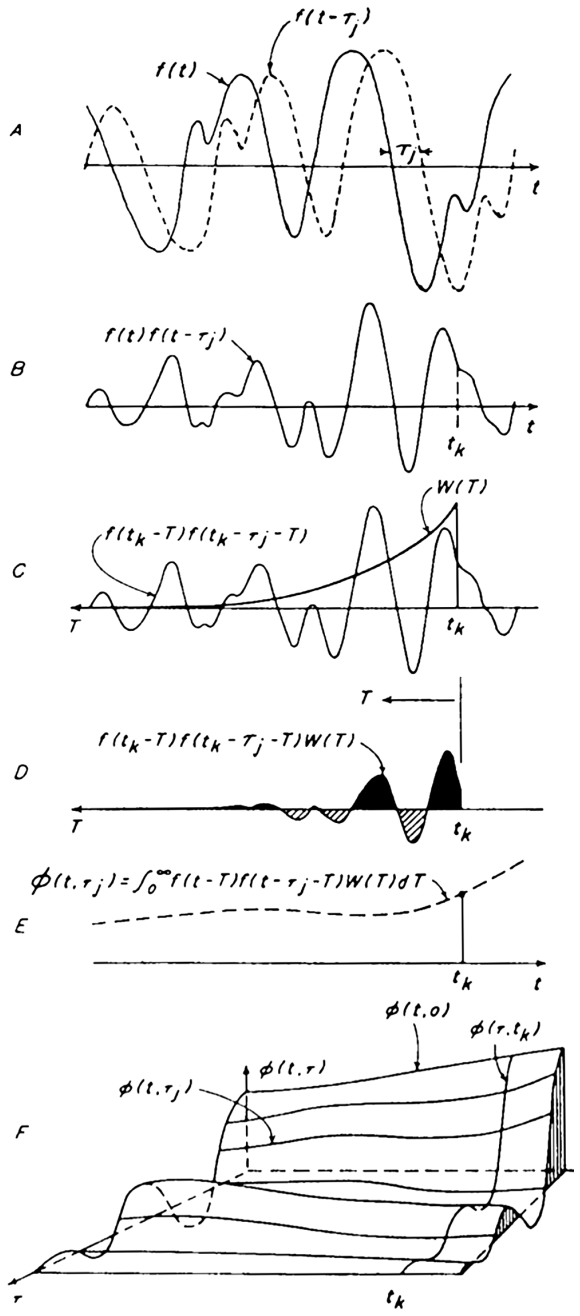


Figure 21.2: Licklider's illustration of the computation of a running autocorrelation function by smoothing the product of a signal $f(t)$ times a delayed version of the same signal (he uses T as the dummy integration variable, where we used u). The panels show: (A) the input signal and delayed input signal, for a particular delay, or lag, τ_j; (B) the product of the signals from panel A; (C) the product waveform, relabeled in terms of time offset T from a particular time t_k, superimposed with the exponential weighting function $W(T)$ (the time-reversed impulse response of a first-order smoothing filter); (D) the weighted product, in this case showing larger areas under the positive parts than under the negative parts; (E) the integral of the signal in panel D, $\phi(t,\tau)$ for the particular τ value, as t changes; (F) the two-dimensional surface $\phi(t,\tau)$, showing slices at various values of t (time) and τ (time lag). Notice that this surface is changing slowly in time, but captures fine temporal information in its lag dimension. [Figure 4 (Licklider, 1951) reproduced with permission of Springer.]

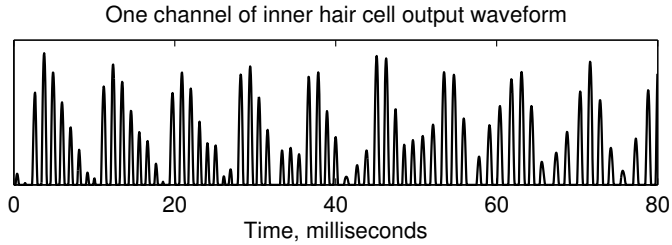


Figure 21.3: The waveform that comes from a single inner hair cell is the nonlinearly detected view of the basilar membrane motion at one place corresponding to one characteristic frequency: one frequency channel. Here we show an 80 ms segment, for channel 42 (a place with CF near 800 Hz) of the cochleagram in Figure 20.3, responding to a spoken vowel of about 120 Hz pitch. Several levels of temporal structure are apparent. This signal represents the input to the calculations illustrated in subsequent figures, analogous to Licklider's input signal, the solid curve in Figure 21.2(A).

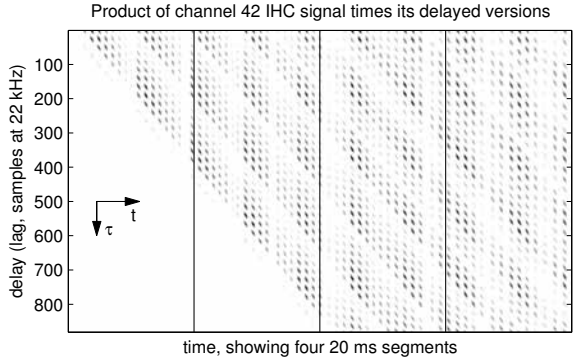


Figure 21.4: The instantaneous products of the channel-42 signal times its delayed versions, for up to 880 samples of delay (out to 40 ms). Vertical lines indicate boundaries between 20 ms segments. This image is analogous to Licklider's Figure 21.2(B), except that we show it in two dimensions for many values of lag, with more-positive values plotting darker, as opposed to his single illustrated lag value.

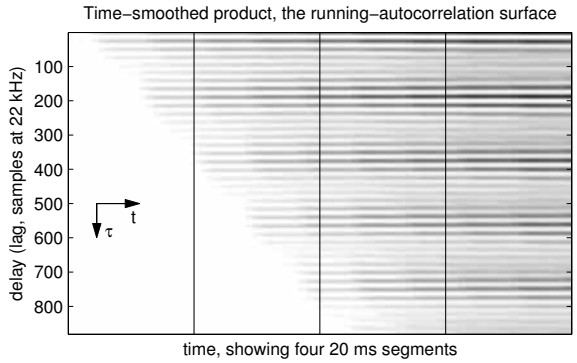


Figure 21.5: Smoothing the products along the time dimension, using a first-order filter with 60 ms time constant, results in this $g(t, \tau)$ running autocorrelation image; the fine time structure is now in the τ dimension, while the function changes only very slowly in time. This image is analogous to Licklider's slowly changing short-time autocorrelation surface of Figure 21.2(F), which is the all-lags version of the single slowly changing correlation coefficient of Figure 21.2(E), which is the signal from (B) smoothed using exponential weighting as in (C) and (D).

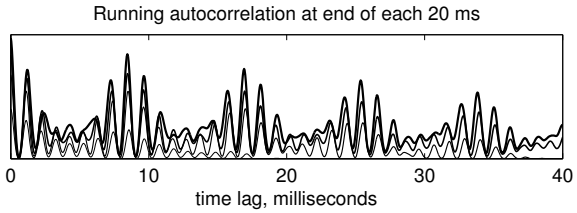


Figure 21.6: Four slices of the running $g(t, \tau)$, at the end of each 20 ms segment, show how the function of lag changes slowly. Later slices are shown with heavier lines. Each slice is a good estimate of the one-sided short-time autocorrelation function of the channel-42 signal. The τ values shown as positive here represent correlations of the current time with the past. In some other figures, we turn the lag axis around and put the past on the left. These four slices, taken at marked times spaced 20 ms apart, are analogous to time slices at times t_k in Licklider's Figure 21.2(F). They all have maxima at the zero-lag position.

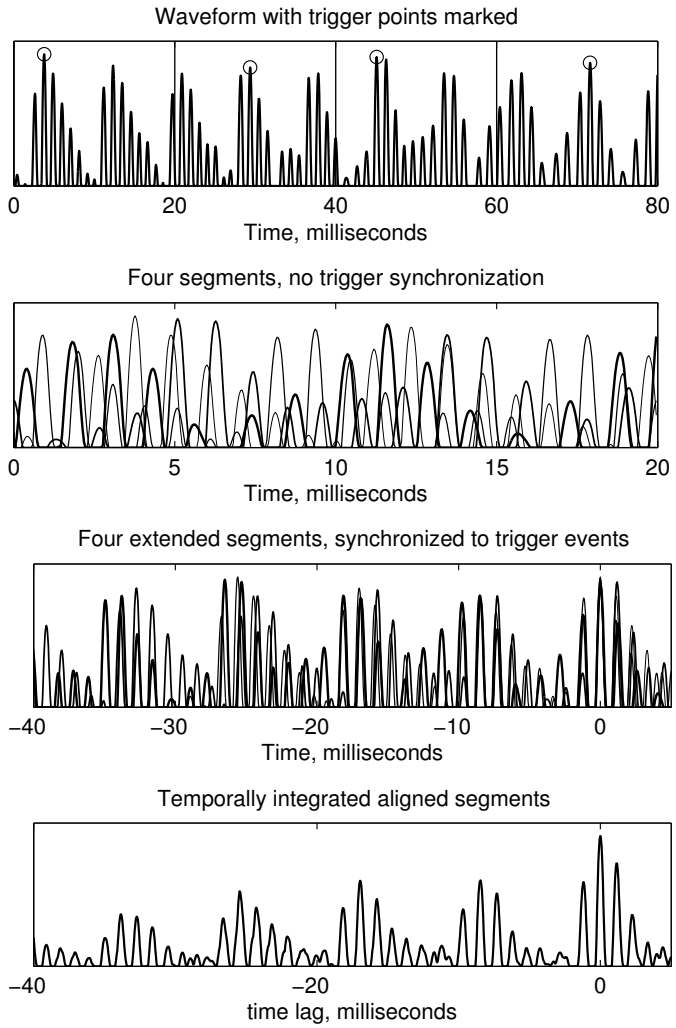


Figure 21.7: The signal of Figure 21.3 is shown divided into 20 ms segments, with the highest value in each segment indicated by circles (top panel). We take these time points as trigger events. If we display the 20 ms segments of the waveform together, they do not line up, and look like a mess (second panel). If instead we summarize the waveform's changes in time by displaying together segments that are aligned based on the trigger events, the picture is much less confusing (third panel). Here, the relative maximum in each 20 ms segment was aligned to the τ origin, and the input signal from 40 ms before to 5 ms after each trigger event was plotted. The signal is not exactly periodic, so the segments do not quite stay aligned away from the time lag origin, but the approximate repetition is clear in the approximately consistent pattern. The aligned segments are then *temporally integrated*, or averaged across the four different trigger times to make one row of one frame of an SAI (bottom panel).

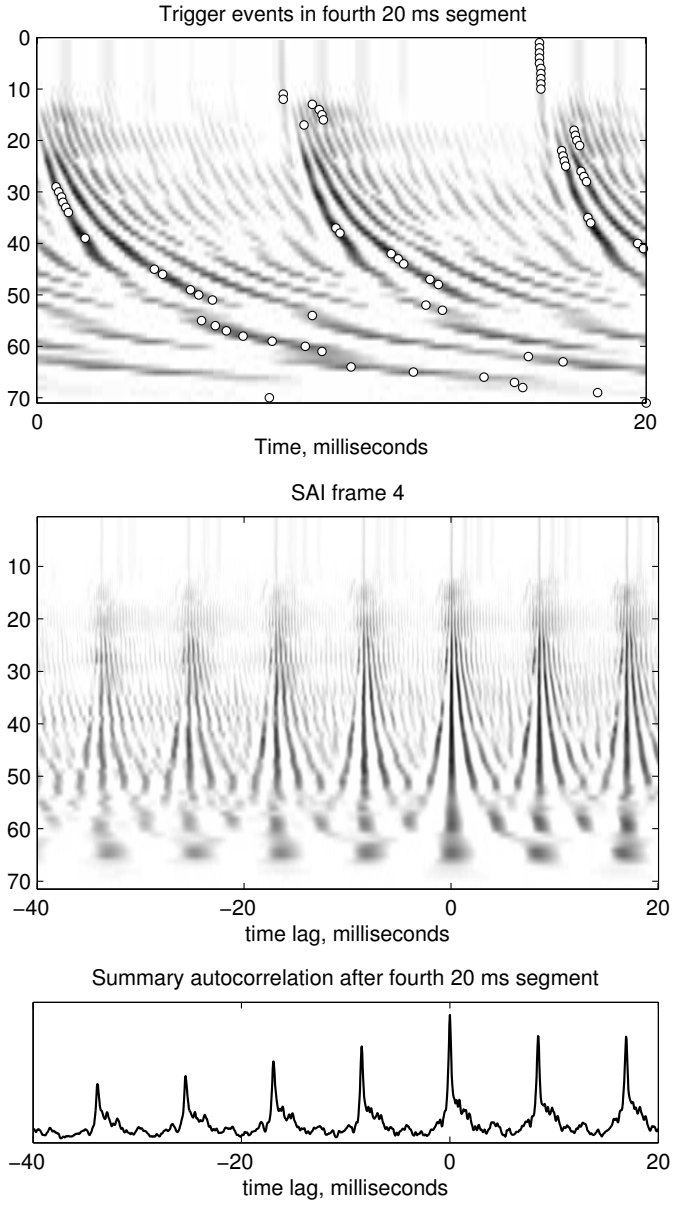


Figure 21.8: The simple triggering algorithm of picking the maximum point in a segment, for each channel, results in these irregular trigger events, shown as circles overlaid on one 20 ms segment of cochleagram (top). The SAI made with the simple trigger algorithm (middle) shows the pitch clearly, but also shows some discontinuities between rows. Even with more sophisticated trigger algorithms, some of this effect will be seen. The average along columns of the auditory image is the summary SAI (bottom), sometimes called the summary autocorrelogram (SACG) especially if the rows are computed by autocorrelation. The trigger irregularities have little effect on the summary.

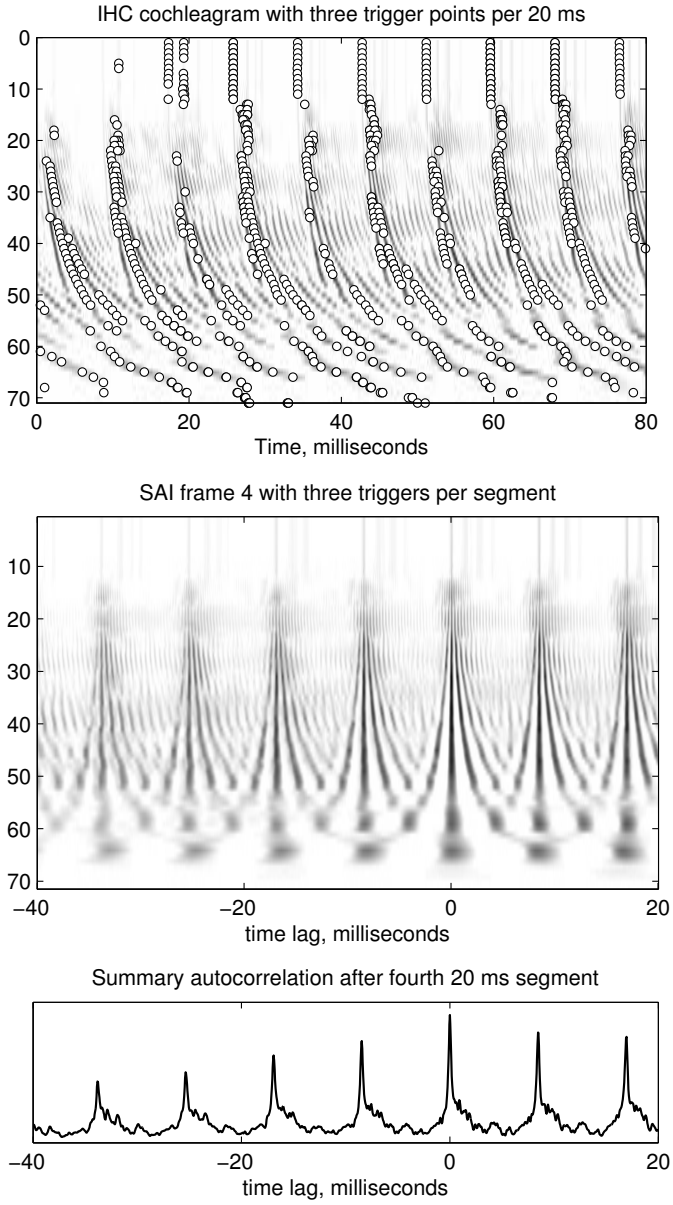


Figure 21.9: The trigger points for 80 ms of the cochleagram (top), using three peaks per 20 ms segment, selected as points of maximum value after weighting with overlapping sine windows, the windows being two segments wide and spaced one-third segment apart. With this method, the points chosen are sometimes not exactly at peaks of the original signal (due to the slope of the window), and the same time point is sometimes chosen more than once (due to the overlapping windows). In the resulting SAI (middle) discontinuities are less prominent with this larger number of trigger points contributing to the temporal integration. The summary SAI (bottom) is not much different from the simpler one shown in Figure 21.8.

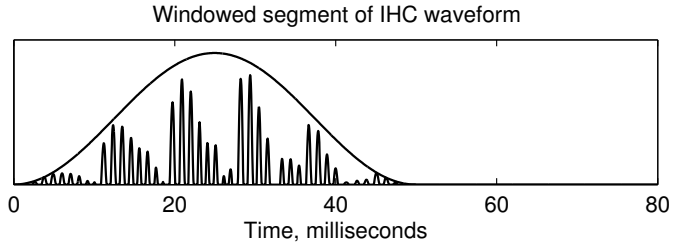


Figure 21.10: The 80 ms segment in Figure 21.3 has been multiplied by the 50 ms raised-cosine (Hann) window to make this windowed segment. Such a segment can be made again 20 ms later, or on whatever frame times are desired.

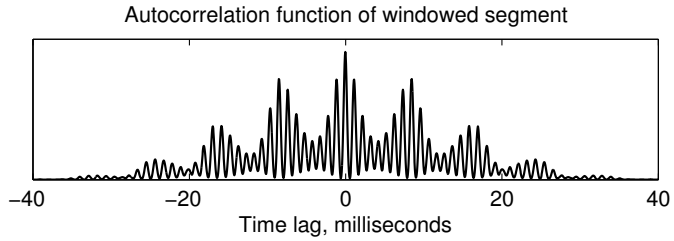


Figure 21.11: The autocorrelation function of the windowed segment in Figure 21.10 is this symmetric function of the time lag parameter. Notice that it resembles the result of TTI in Figure 21.7, at least for moderate lag magnitudes, but the TTI version is sharper and not symmetric.

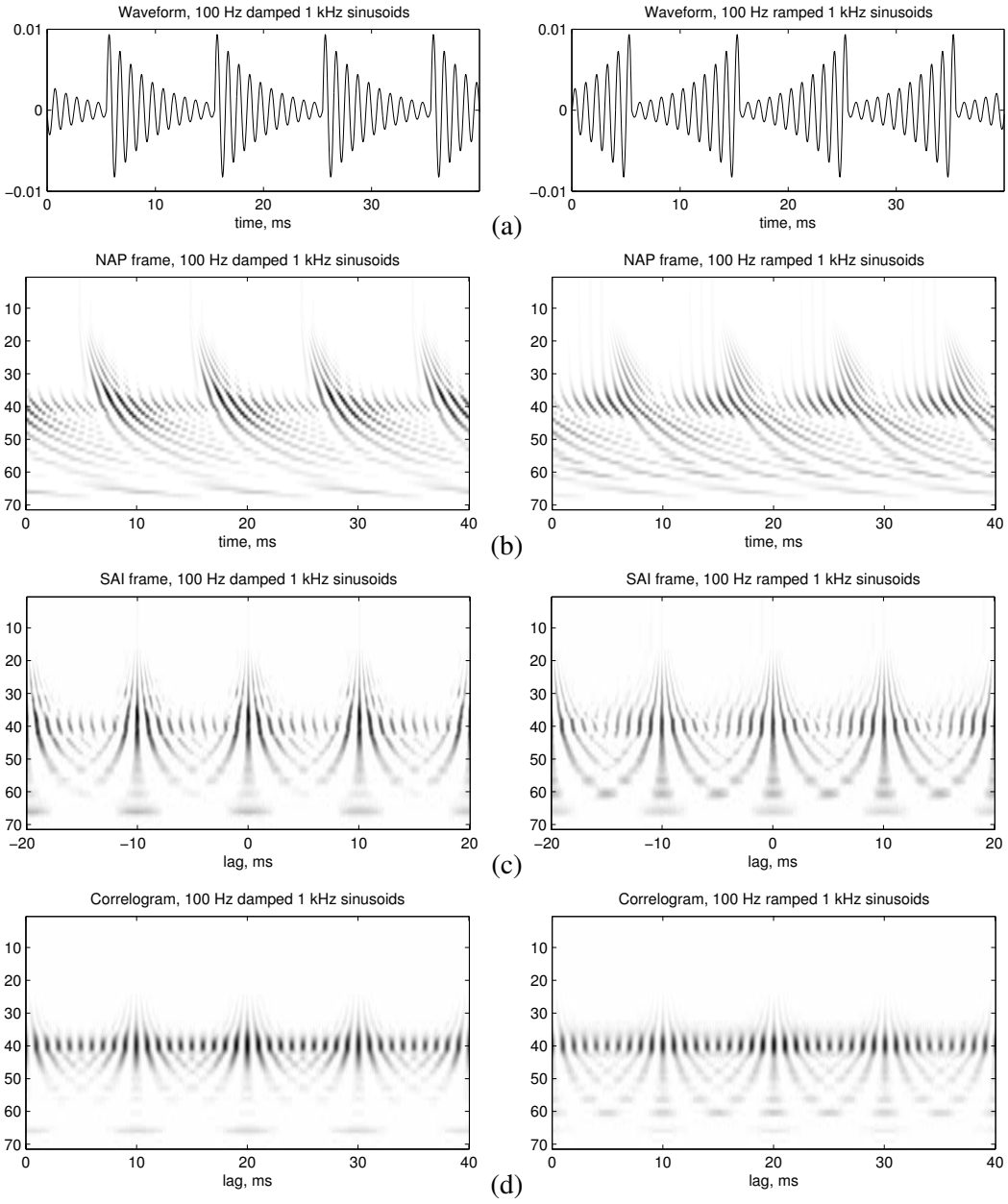


Figure 21.12: Damped (left) versus ramped (right) sinusoids (a), snippets of the resulting NAPs (b), SAI frames (c), and autocorrelogram frames (d). The damped and ramped signals are time reversals of each other and thus differ only in the phase of their Fourier components. Their SAIs show subtle differences that correspond to subtle perceptual differences. Compare with the SAI of a 1 kHz pure tone in Figure 21.1. The correlograms, though symmetric in lag, also show some differences in ramped versus damped, because the NAP rows are not time-reversals of each other. This difference is due to the AGC's lagging gain variation, which results in a stronger fundamental-frequency distortion tone in the damped case, but stronger harmonics in the ramped case.

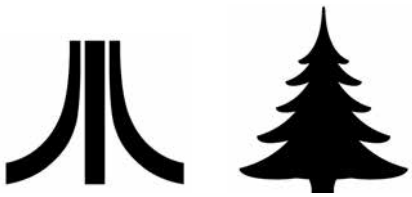


Figure 21.13: The local structures in SAIs are sometimes said to resemble Atari logos or Christmas trees. The wider fringe spacing toward the bottom is caused by the decreasing ringing frequency as waves propagate through the cochlea from the base (channels near the top of the picture) toward the apex (channels near the bottom of the picture).

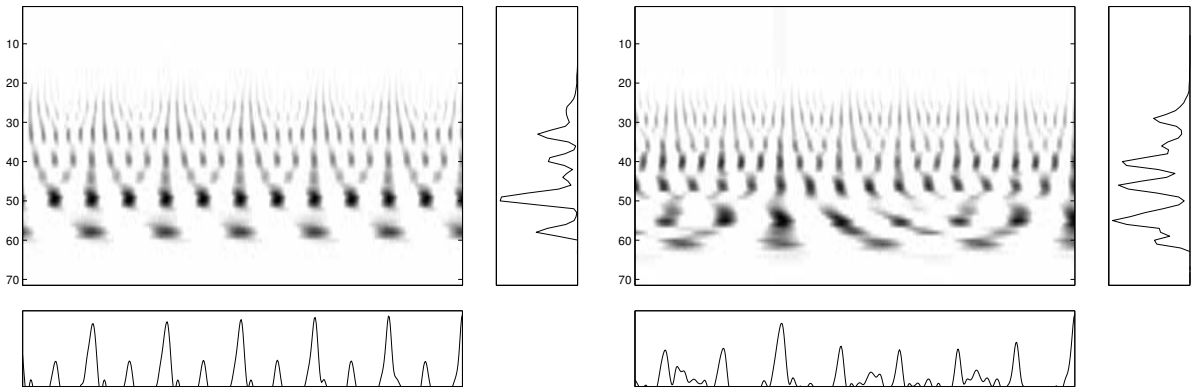


Figure 21.14: These SAIs of piano notes—an isolated note on the left and a chord on the right—show the activity *before* the trigger events, so the trigger times are aligned at the right edge. In the chord, the temporal profile (the average along columns) plotted at the bottom shows the root pitch period at 5 times the period of the highest note. The "auditory spectrum" plotted on the right is indicative of a sort of overall timbre. Compare with the SAIs of one, two, or three steady notes shown in Figure 4.10. The slightly higher pitch of high harmonics, a characteristic of piano strings, is apparent in the uppermost parts of these SAIs.

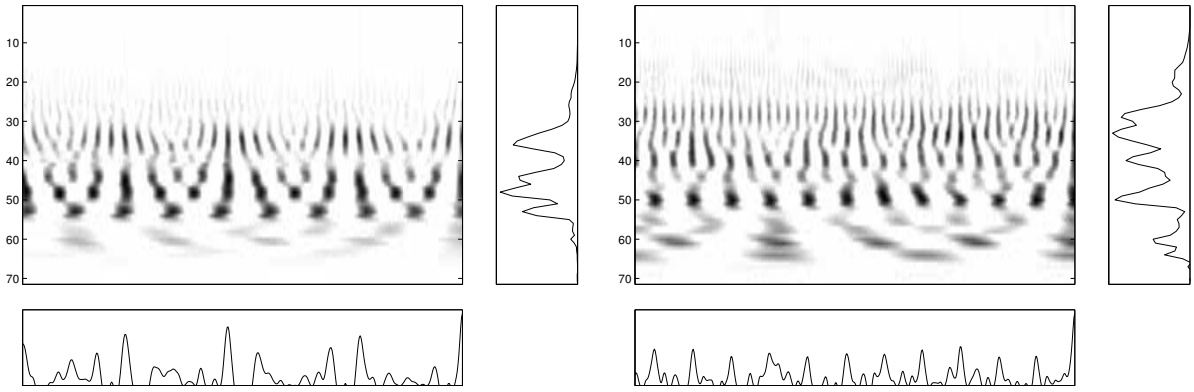


Figure 21.15: Two SAI frames of a jazz music piece. The sound represented is primarily a cowbell-like percussion note on the left, and a more complex mixture with multiple pitches on the right.

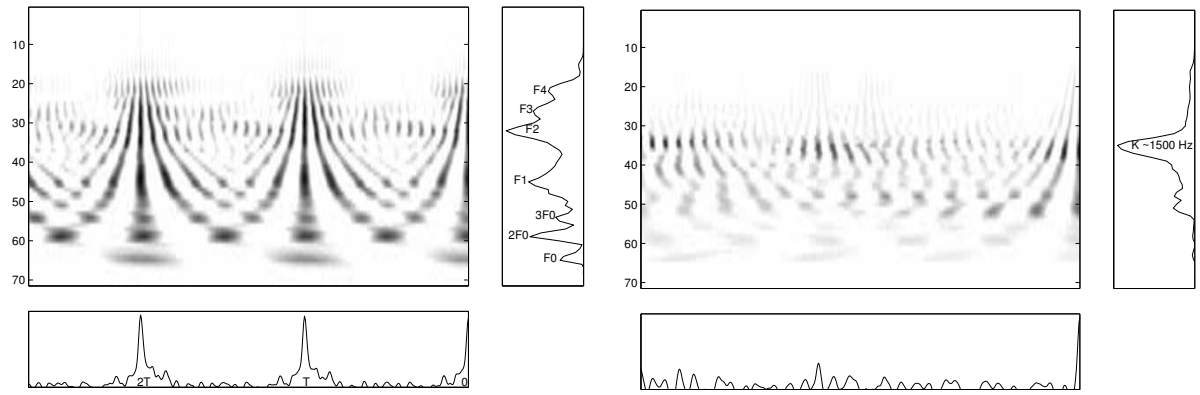


Figure 21.16: SAI of a spoken vowel /æ/ (in "plan") with a pitch of about 122 Hz, on the left, and of a /k/ (voiceless velar stop consonant) release burst, on the right. In the left panel, both the row and column averages show the pitch (frequency F0, period T) of the vowel, but the temporal profile (the average along columns, at the bottom) shows it more clearly and explicitly. The auditory spectrum clearly shows the formants, especially the first and second formants F1 and F2, which are most important in determining the perceived vowel category. In the right panel, the temporal profile of the /k/ shows no periodicity. The auditory spectrum shows the compact k-release burst resonance in the F2 region, typical of a velar stop release.

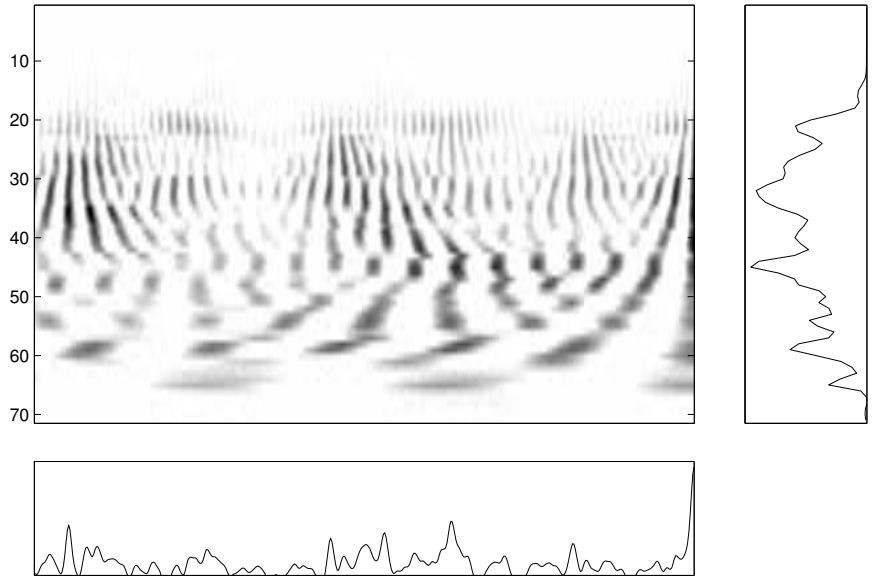


Figure 21.17: SAI of two concurrent vowels: the /æ/ of Figure 21.16 added to an /aɪ/ diphthong from the same speaker but at a lower pitch. The row and column summaries are no longer very informative, but the partial "Atari logo" structures in the image provide locally separated responses to the two vowels—along with some ghost responses at time lags equal to the intervals between pitch pulses of the two vowels. Viewed as a movie, the SAI shows coherent motion of each vowel. See Figure 21.18 for an analysis of this mixture SAI image.

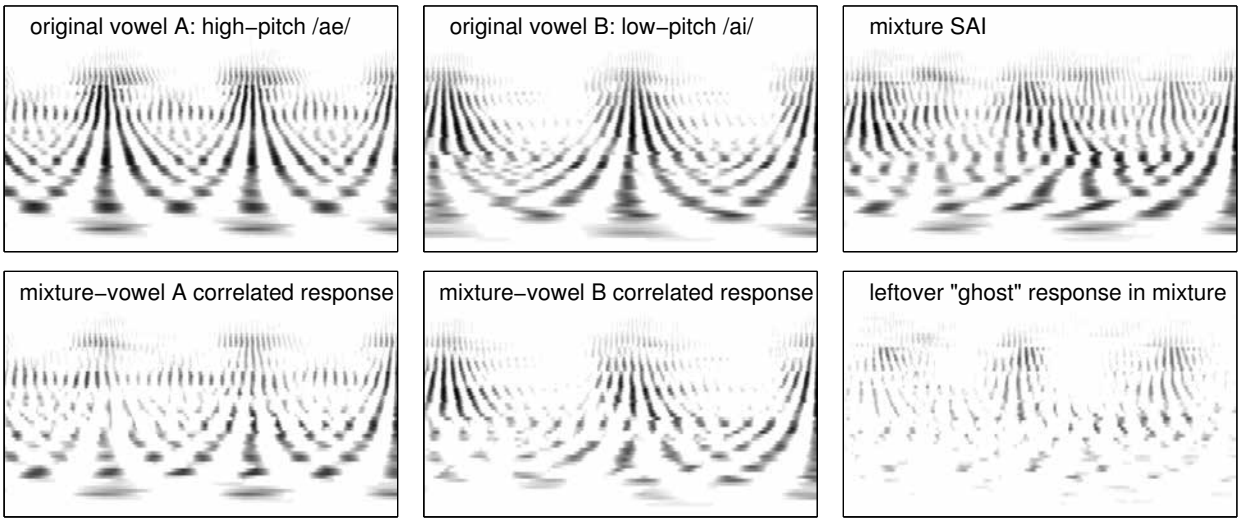


Figure 21.18: The SAI of the concurrent vowel mixture in Figure 21.17 is analyzed to show how it relates to the SAIs of the original clean vowels that were mixed. The SAIs of the two vowels and the mixture are shown on top. Below the originals, the portions of the mixture that match are shown (extracting as the max of each point in the pair original and mixture SAIs). Below the mixture SAI is shown the difference between the mixture SAI and the max of the two original SAIs, which yields the pattern that does not match either original sound's pattern—the "leftovers" or "ghosts" caused by one vowel's pulses corrrelating with the other's.

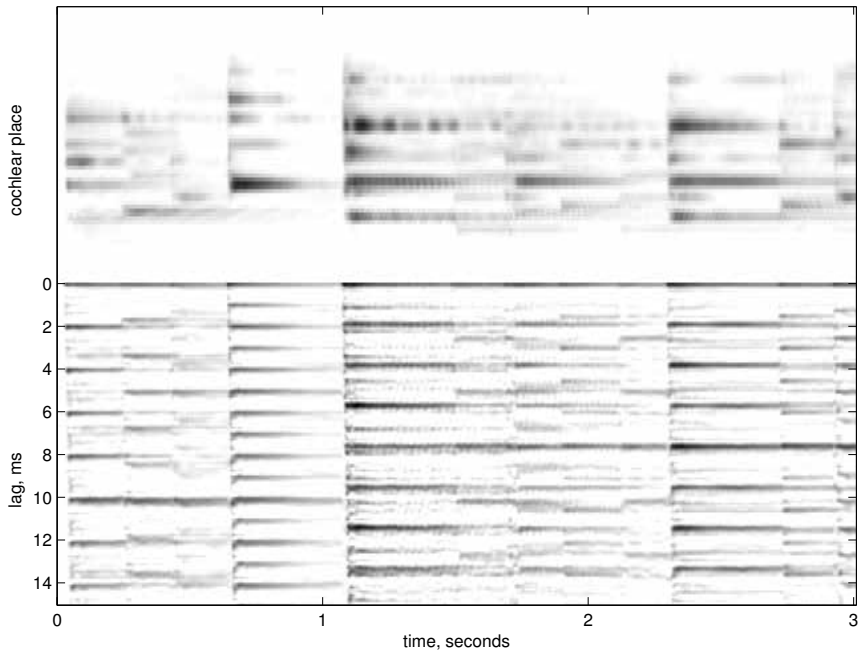


Figure 21.19: Combination cochleagram and pitchogram of 3 s of piano music. The period relationships that correspond to harmonic chords and consonant pitch intervals are more apparent in the lower part, the pitchogram, than in the cochleagram.

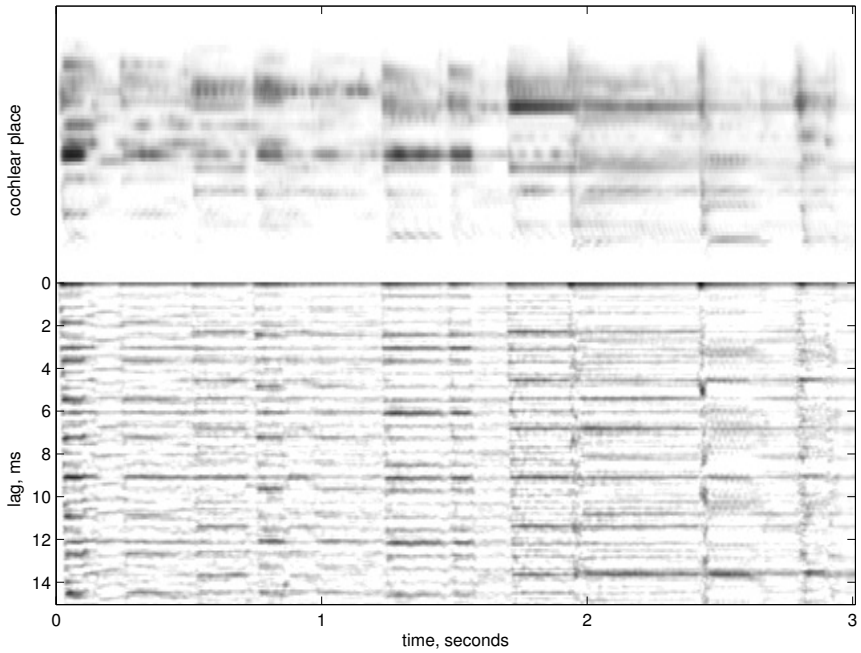


Figure 21.20: Combination cochleagram and pitchogram of 3 s of guitar music. A common period of 9 ms is apparent for many of the notes and chords.

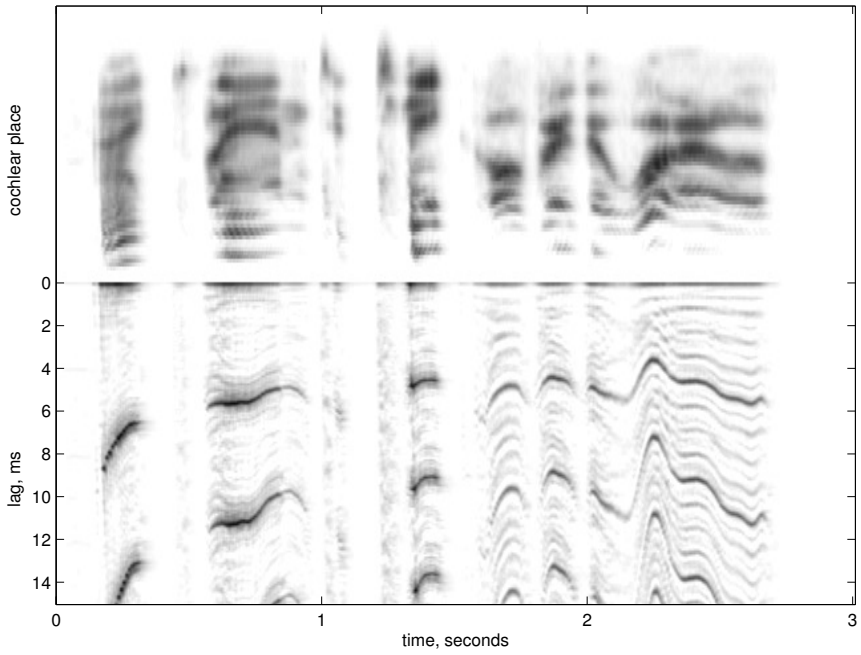


Figure 21.21: Combination cochleagram and pitchogram of 3 s of speech. The top part, the cochleagram, resembles a speech spectrogram, while the pitchogram on the bottom clearly shows the pitch contours of the voiced segments.

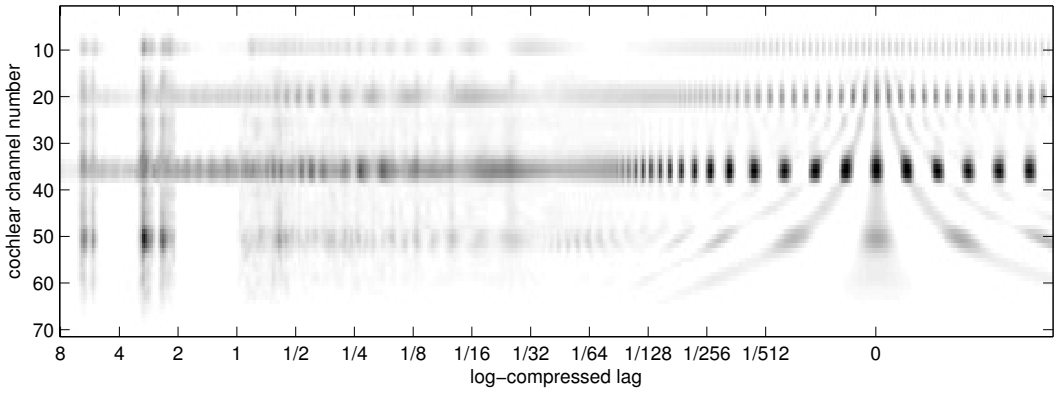


Figure 21.22: An SAI frame of a telephone ringing (file BelgiqueBellPhone.mp3 from freesound.org), with longer lags logarithmically compressed. The telephone sound exhibits structure on many time scales, out to its 3 s period and 6 s double period, and has somewhat different structure in different frequency regions.

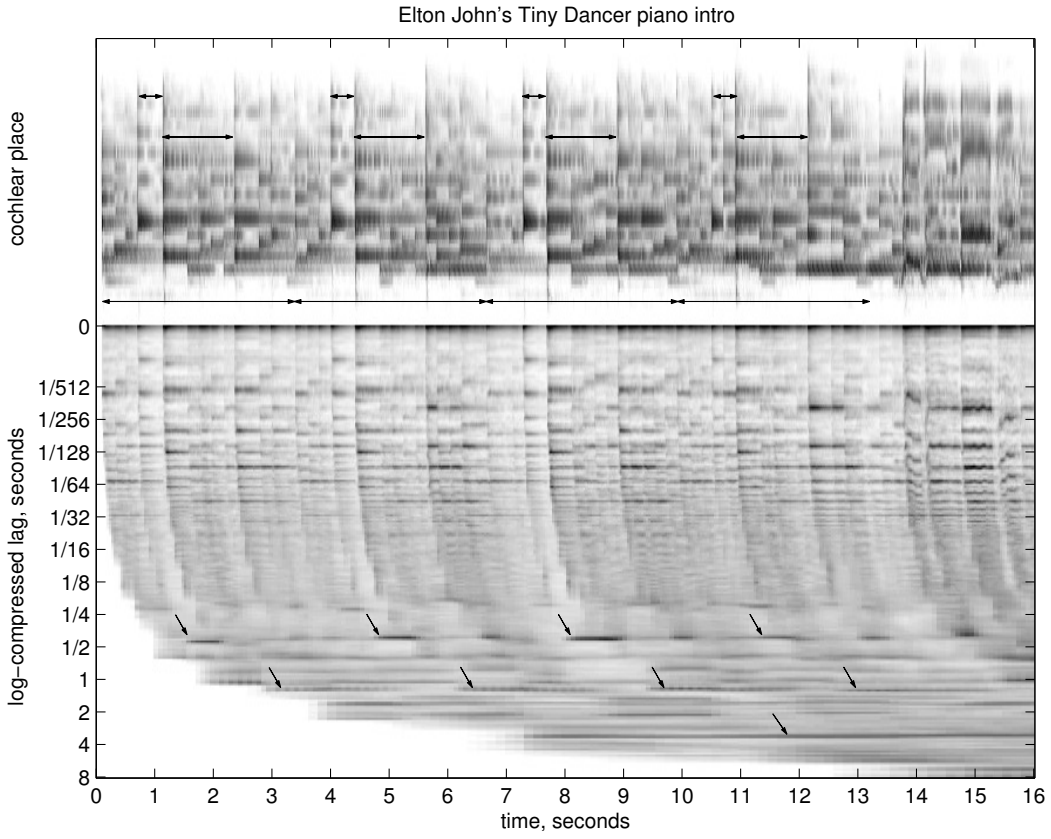


Figure 21.23: Combination cochleagram and log-lag pitchogram/rhythmogram of the piano opening of Elton John's "Tiny Dancer," with vocals in the final 2 s. Prominent time intervals between strong chord onsets, indicated by arrows from top to bottom, correspond to the durations of eighth notes (0.4 s) and of three eighth notes (1.2 s); measures (3.2 s) are also marked. The sweeping curves at the bottom reflect the delay of causal buffering, correlation, and averaging at exponentially increasing time scales.

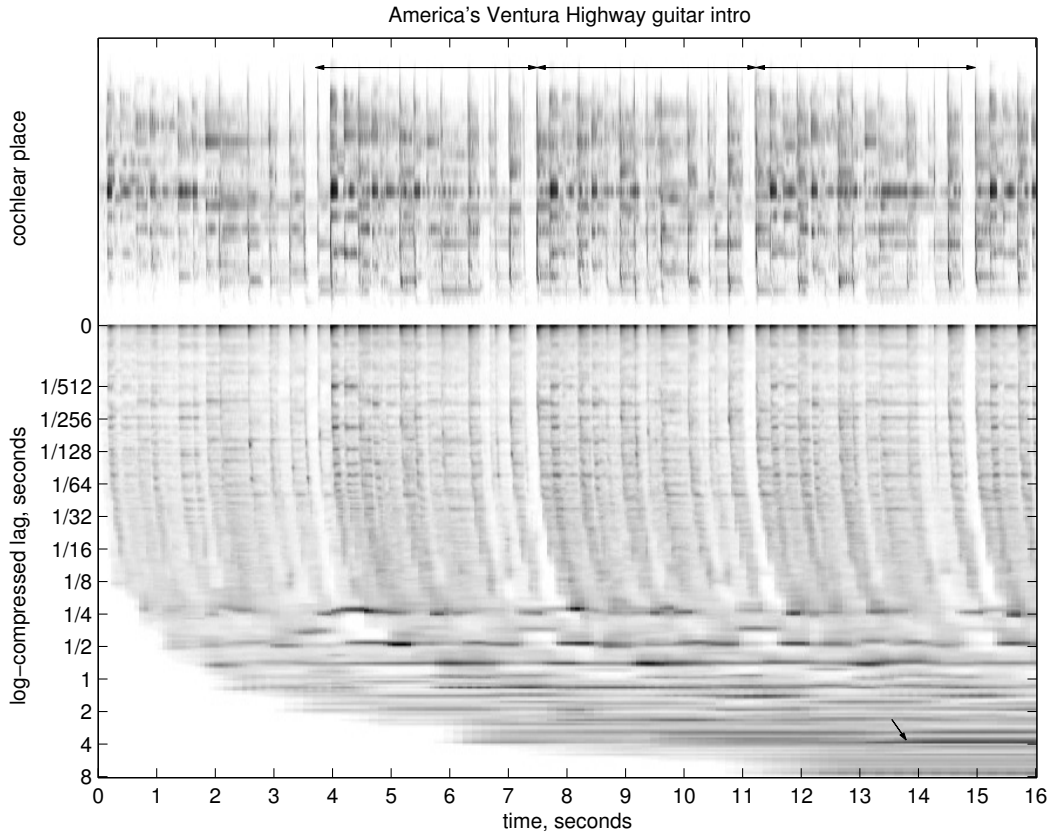


Figure 21.24: Combination cochleagram and log-lag pitchogram/rhythmogram of the popular opening guitar riff of America's "Ventura Highway." Prominent time intervals of one-quarter, one-half, and three-quarters seconds between notes show up, as does the phrase repetition near 4 s (indicated by arrows), but there is not much at 1 or 2 s. These time patterns summarize the rhythmic structure of the riff.

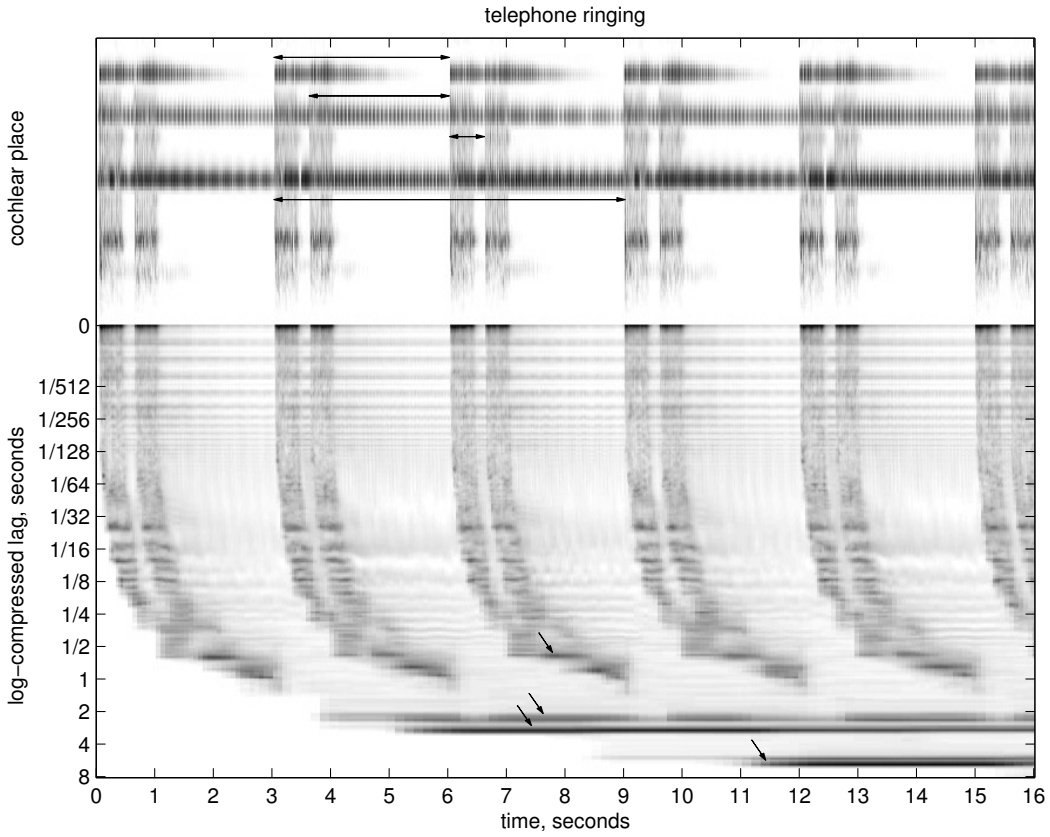


Figure 21.25: Combination cochleagram and log-lag pitchogram/rhythmogram of 16 s of the telephone ringing of Figure 21.22; two rings 0.7 s apart, every 3 s. The lag region that is not much used in music, between pitch periods and beat periods, is here filled with the pattern of the telephone's clapper intervals, at about 1/32 to 1/8 s. The rhythmic 0.7 s, 2.25 s, 3 s, and 6 s intervals are also prominent, as indicated by arrows.

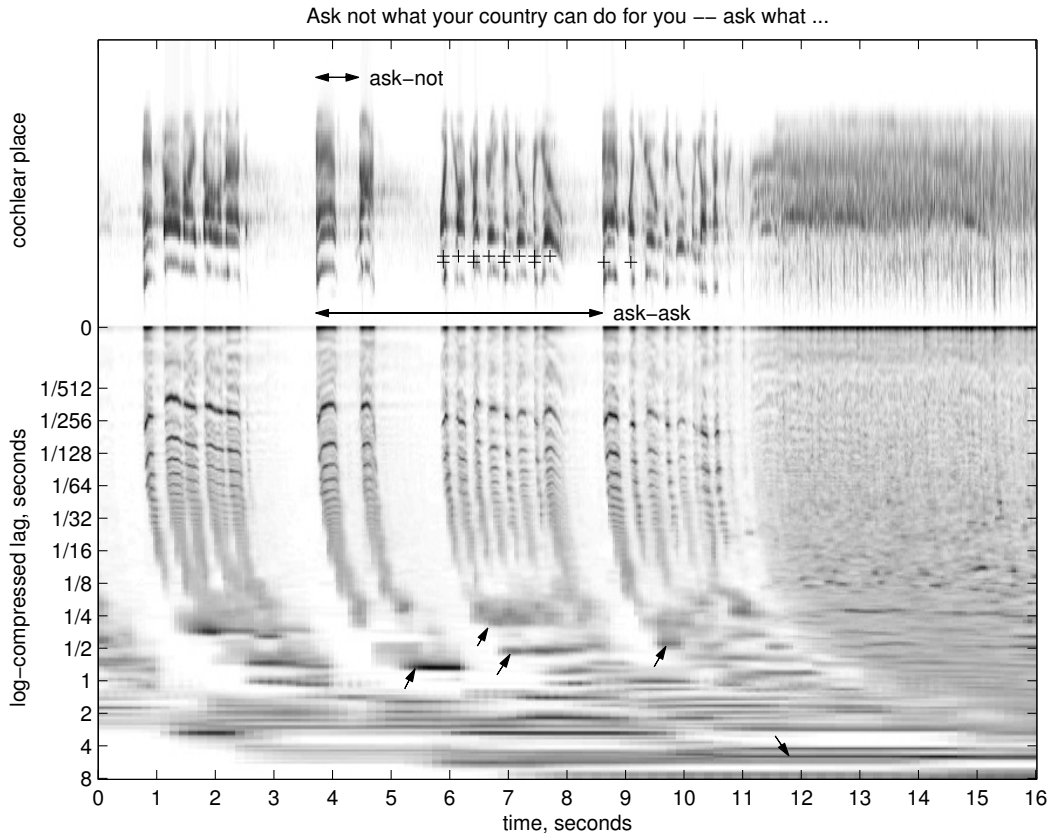


Figure 21.26: Combination cochleagram and log-lag pitchogram/rhythmogram of 16 s of John F. Kennedy's inaugural address: "And so my fellow Americans: Ask not what your country can do for you—ask what you can do for your country [applause]." English tends to have a regular cadence, but in oratory like this the rhythm is greatly altered. The interval between the syllables *ask–not* is lengthened to about 0.75 s, as marked on the plot, while the syllables in *what your country can do for you* come at about four per second, slightly syncopated, with a strong regularity at 0.5 s; these regular intervals are marked by crosses and arrows, as is the 0.4 s *ask–what* interval. The 5 s interval between the two occurrences of *ask* after a pause shows up relatively precisely.

Chapter 22

Binaural Spatial Hearing

A failure to distinguish between phase expressed as angle, ϕ, and phase expressed as time, t', has led to some confusion in the literature.

— "A place theory of sound localization," Lloyd A. Jeffress (1948)

To anyone who is familiar with our modern knowledge of room acoustics, one of the most puzzling questions must be how it is that sounds can be localized at all in a reverberant room, let alone heard in the rather precise positions that are often reported.

— "The precedence effect in sound localization," Wallach, Newman, and Rosenzweig (1949)

History Connection: Phase "Unwelcome"

The idea that the phases of the signals at the two ears could interact had been examined in the context of *binaural beats*, and had been discarded, with the conclusion that the effects must instead be due to intracranial sound propagation.

Sylvanus P. Thompson (1877) reported observing these beats between tones presented to the two ears. His hypothesis that "the tone-stimulus is transmitted along each auditory nerve to some common cerebral centre and that at this centre the beats arise" was ignored or rejected at the time, and was "unwelcome" even after Rayleigh's duplex theory. For example, Wilson and Myers (1908) critically assess "the influence of binaural phase differences":

> The importance of the transmission of stimuli by bone conduction from ear to ear is well seen in an experiment described by Thompson, in which two tones were generated in different rooms and were led by tubes, one to each ear of the observer. These tones were produced from two tuning-forks, having a pitch of 246 and 256 vibrations per second, respectively. Under these conditions, as is well known, beats are audible, just as if the two tones were presented to a single ear. Thompson concluded that under the conditions of binaural hearing above described, the tone-stimulus is transmitted along each auditory nerve to some common cerebral centre and that at this centre the beats arise. But this and the following interesting fact, also observed by Thompson, can be explained without recourse to such an unwelcome hypothesis, if we suppose that each tone is transmitted by bone conduction to the opposite ear and that the beats heard are due to the play of the two series of vibrations of different frequency on one and the same sense organ.

Wilson and Myers had discussed the possibility that the phase sensitivity implied that the auditory nerve carried the waveform directly; but then rejected that idea:

> In the case of vibrations of sound,—despite the fact that they are much slower than vibrations of light,—it is nevertheless just as difficult to suppose that such characters of the stimulus are actually communicable to the auditory nervous impulse. It is very hard to believe that every crest and every trough of each sound wave produce an exactly corresponding crest and trough in the impulses transmitted along each auditory nerve. Were this so, we could no longer regard the sensory nerve as an intermediary, knowing no more of the exact nature of the external stimulus than the telegraph wire knows of the mental processes of the operator who transmits the telegram. We could no longer regard the sensory nerve impulse as being determined solely by the method of response of the end organ with which it is connected.
>
> We hope to show that such a radical change in our views is unnecessary.

Eventually, researchers had to accept such a radical change, and admit that the nerves do convey waveform details to central sites for comparison. An accumulation of evidence, and Wever and Bray's 1930 *volley theory* that we discussed in Chapter 2, made the ideas less "unwelcome."

In the meantime, however, there were still plenty of efforts to explain binaural hearing without any central phase comparison, relying on level differences alone. One of these, by Henry J. Watt (1920), actually proposed an auditory-image-like model: a two-dimensional pattern of activity reflecting tonotopic organization along one dimension, and interaural intensity relationship along the other, narrower, dimension of neural tissue.

History Connection: From IPD to ITD

Rayleigh (Strutt, 1877) had made an observation that practically begged for someone to propose a cue based on the different times of arrival of sounds at the two ears:

> When one ear is stopped, mistakes are made between [tuning] forks right and left; but the direction of other sounds, such as those produced by clapping hands or by the voice, is often told much better than might have been expected.

The conceptual replacement of the phase differences of sine waves ("pure tones" as the acousticians always glorified them) with the time difference of more general or transient sounds started about 1908, but took a long time to catch on. Mallock (1908) noticed the strange apparent direction of the sound of a bullet, due to its bow shock wave (its miniature sonic boom), and after some experiments and analysis concluded (see Figure 22.1):

> A sound which is caused by the detached waves, such as those which accompany a bullet, can scarcely be said to have a pitch, but the wave-length is certainly small compared with the distance between the ears, and is indeed comparable with the dimensions of the bullet itself. It would seem, therefore, that the ears can determine the direction of a sound, not only by difference of phase, but by the actual difference in the times at which a single pulse reaches them.

He found a consistency of observations to within a few degrees of the calculated wavefront direction (corresponding to a time-difference error of a few tens of microseconds). Then Hornbostel and Wertheimer (1920) found time differences between clicks to be effective down to 30 microseconds, and even smaller "under favorable conditions." Otto Klemm (1920) published much more detailed experimental results, also in 1920, and found a time-difference threshold of about 20 microseconds in one subject, and even less than 10 microseconds in another! Several researchers published ITD thresholds in 1921: Aggazzotti (1921) found 70 microseconds, and Pérot (1921) found between 55 and 80 microseconds (and more at lower levels).

In spite of all these investigations, the paradigm shift to thinking of ITD as a genuine cue was slow to come. For example, Hartley and Fry (1922) analyzed the localization of complex tones, but interpreted it all in terms of the independent localization of sinusoidal components by their phase and amplitude differences.

While the ideas were still being debated, during World War I scientists on both sides were putting the "binaural sense" to use in military applications, for finding the directions to airplanes and submarines and tunnel diggers, using human listeners with binaural sound horns and time-delay compensators to steer their listening direction (Yerkes, 1920; Drysdale, 1920; Ferry, 1921). The 1920 report by Hornbostel and Wertheimer (1920) was based on their German wartime devices. They filed for a patent in 1915 on the "Richtungshörer" (directional listener) that used a wide spacing between listening horns to exaggerate the time difference cue (King and Wertheimer, 2007); similar to the one in Figure 22.2, it was popularly known for its inventors as the "Wertbostel."

In 1931, Erich von Hornbostel (1931) restated his "time theory," as part of a "discussion on audition," a discussion in which others argued as strongly against it. He pointed out that phase was a poor alternative to absolute time difference. He tried to push the field away from the overreliance on tones, saying, "Theory, and also experiment, must take into consideration the fact that noises are more important in life than musical sounds (complex and pure tones) which are of rare occurrence in Nature; the fact, therefore, that noises are better localized than tones is a useful one."

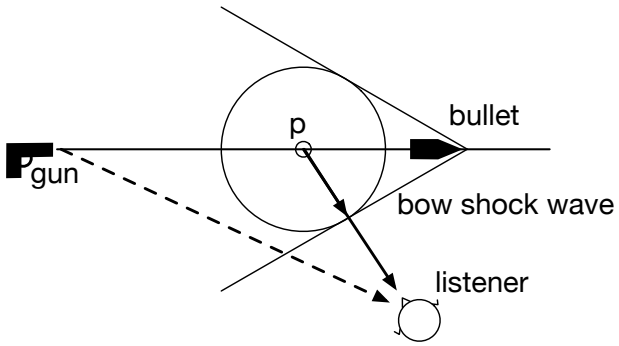


Figure 22.1: Mallock (1908) noticed that to a down-range listener, the crack of a supersonic bullet seemed to come from a direction different from the direction of the gun. The bullet's "sonic boom" or detached bow shock wave, moves at the speed of sound, while its apex at the bullet moves faster than sound, resulting in a wave angle as shown. When a two-eared listener is facing normal to the wave, it arrives at both ears at the same time, resulting in the apparent direction toward the point p.

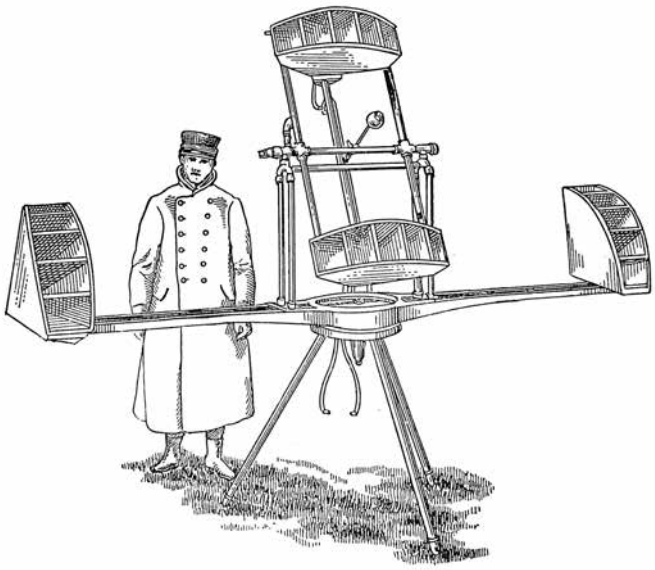


Figure 22.2: A World War I era acoustic goniometer for locating "invisible aeroplanes" (Ferry, 1921). Operators rotate pairs of pickup horns about vertical and horizontal axes, until a sound seems to be straight ahead. The device is clearly designed to exaggerate an interaural time difference cue.

History Connection: Getting Away from a Focus on Sinusoids

In 1936, while acknowledging that "phase difference is but a special case of time difference," Stevens and Newman (1936) concluded "that the localization of low tones is made on the basis of phase-differences at the two ears, and that the localization of high tones is made on the basis of intensitive differences. There is a band of intermediate frequencies near 3000 cycles in which neither phase nor intensity is very effective and in which localization is poorest."

The experiments of Scherer (1959) compared the ability to lateralize sinusoids, versus broadband signals, based on an ITD cue. Rayleigh had shown that the ITD cue for sinusoids, that is, a phase difference, can be totally ambiguous above about 640 Hz; but Scherer found that the ability to detect the insertion of a 20 μs delay is only reduced a little at 800 Hz, and falls gradually between about 800 and 1600 Hz. By 1600 Hz, with sine waves, his subjects had no ability to distinguish 0 from 20 μs ITD. But with noise filtered to a band around 1600 Hz, or even 3000 Hz, his subject were just as good at detecting the 20 μs ITD as at lower frequencies.

In light of Hornbostel's and Scherer's results, we know that the dip around 2–4 kHz that Stevens observed is purely an artifact of using tones, with their inherent cyclic time ambiguity, like those annoying beepers on carts in airports that you never notice coming up behind you because their beeps can't be localized. For more typical sounds, we localize quite well in that frequency range.

When Jeffress (1948) looked at the science, he concluded, "We may therefore reasonably assume that the basis for our ability to localize clicks and low frequency tones is the time difference." He wasn't denying that intensity difference is important, too, but made the point that time difference works across the frequency range, at least for sounds that contain time-localized events, as clicks do.

Even noises are not always so easy to localize; the normal or Gaussian amplitude distribution typical of some noises has few strong "outlier" features to drive robust localization. For sounds with "long tail" amplitude distributions or otherwise strongly fluctuating envelopes, frequent distinctive events in the tail of the amplitude distribution provide especially good points to localize based on time difference, using "envelope cues" and "onset cues" (Kollmeier et al., 2008). Onsets of pitch pulses in spoken vowels are such points. As McFadden and Pasanen (1976) say, "the auditory system obviously can be just as sensitive to this temporal difference at high frequencies as it is to cycle-by-cycle differences at low frequencies."

In spite of observations of precise neural synchrony to onsets and the ease of lateralizing sounds with transients, we still see papers stating that ITD works as a cue only at low frequencies, or that we're not very good at localizing in the 2–4 kHz region. These statements are correct, but only when applied to sinusoids or narrow-band signals; they can be compared to the view of *Flatlanders* (Abbott, 1884), people whose world is missing a dimension, limited to the infinitesimal slice of the sound space defined by sine waves.

Commenting on the brain area that extracts ITD cues, Karino et al. (2011) betray the usual preconception that ITD should be dominated by lower frequencies, as in Rayleigh's original conception: "Surprisingly, the tonotopic distribution of the afferent endings indicate that low characteristic frequencies are under-represented rather than over-represented in the MSO." Hopefully, we will get over being surprised that ITD is important at higher frequencies.

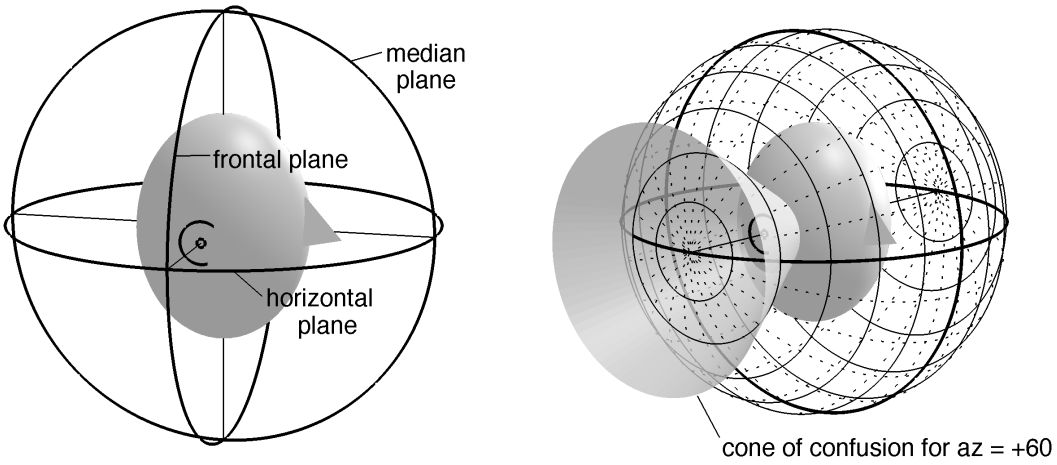


Figure 22.3: Sound directions are often described within three planes, depicted on the left by circles that determine the planes around an ellipsoidal head, aligned with the axes through the ears and the up–down and front–back directions. On the right, the *interaural-polar coordinate system* (Brown and Duda, 1998) is shown. In this system, elevation is like longitude, the angle about the polar axis between the ears, measured from the prime meridian in the horizontal plane; lines of constant elevation are shown dotted. Azimuth is like latitude, measured from the equator in the median plane; circles of constant azimuth are shown solid. One "cone of confusion" is also shown: the set of sound directions with a constant azimuth, or approximately a constant ITD.

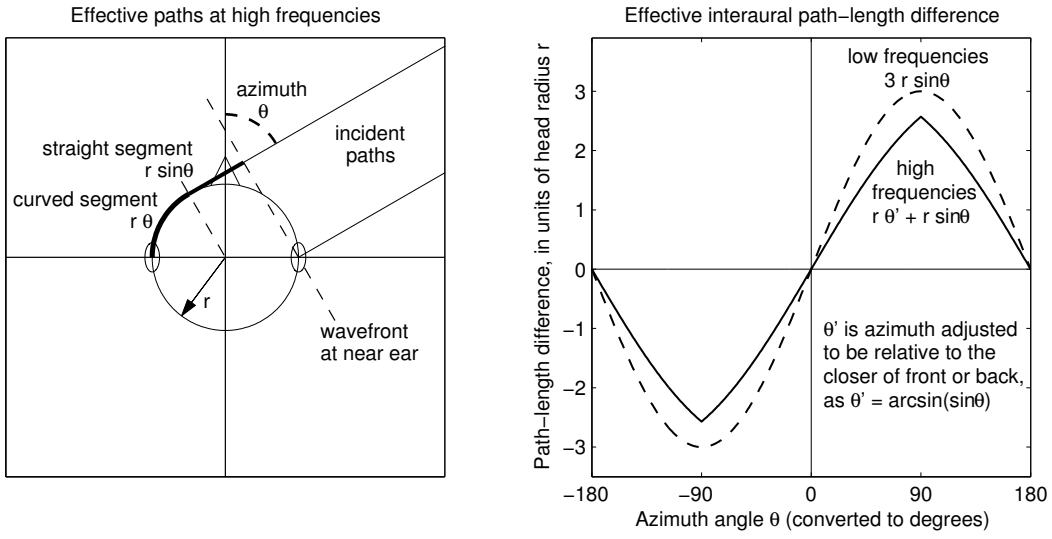


Figure 22.4: High-frequency sounds follow the shortest path around the head. In the horizontal plane, for sounds incident from an azimuth angle of θ, the extra path length to the far ear is $r\theta + r\sin\theta$ (using the azimuth angle modified as shown to be an angle of magnitude less than 90 degrees in the interaural-polar system). Due to diffraction effects, low-frequency sounds experience a somewhat larger time lag than this estimate would suggest; about 50% larger for small angles (Kuhn, 1977). The angle illustrated on the left is on the +60 degree cone of confusion shown in Figure 22.3.

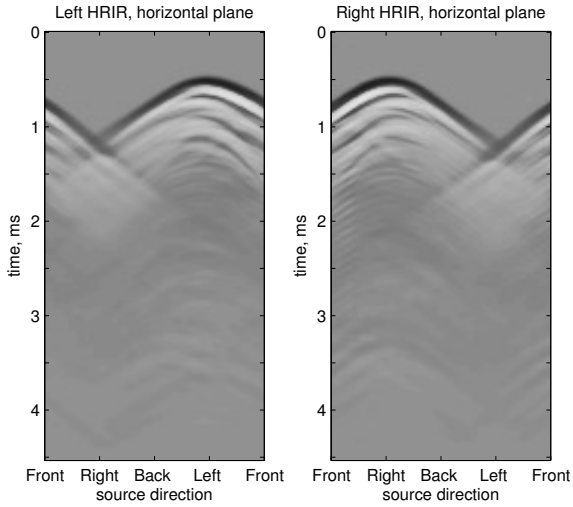


Figure 22.5: Head-related impulse responses of a dummy head, for sounds from various directions in the horizontal plane. The impulse responses are mapped to gray levels and displayed from top to bottom, from an arbitrary time origin shortly before the arrival of a sound impulse at the head; the x axis represents the azimuth angle of the sound source. Separate arrivals can be seen when the sound goes around the front and back of the head as in the left ear when the sound is from the right. For a sound from an intermediate azimuth (not a multiple of 90 degrees), differences in pinna echos, the ridges near the top of the plot, help distinguish front from back.

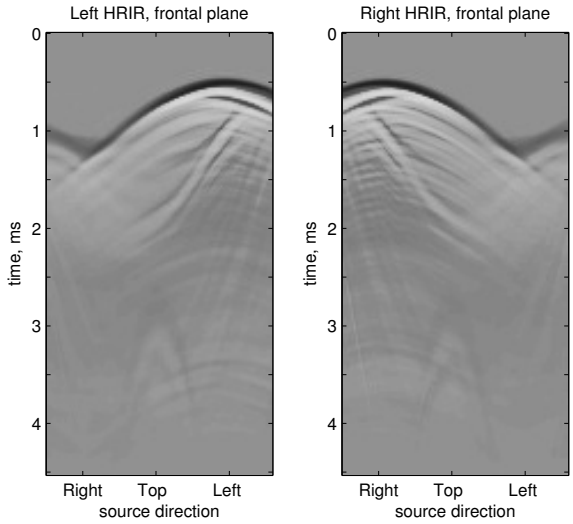


Figure 22.6: Head-related impulse responses of a dummy head, for sounds from various directions in the frontal plane. The x axis represents the angle of the sound source, from low on one side, over the top of the head, to low on the other side. The steeper patterns near the center, in the 1–2 ms range, represent shoulder-bounce arrivals, mostly at the ipsilateral ear, up to about 1 ms delayed from the main arrival (Brown and Duda, 1998).

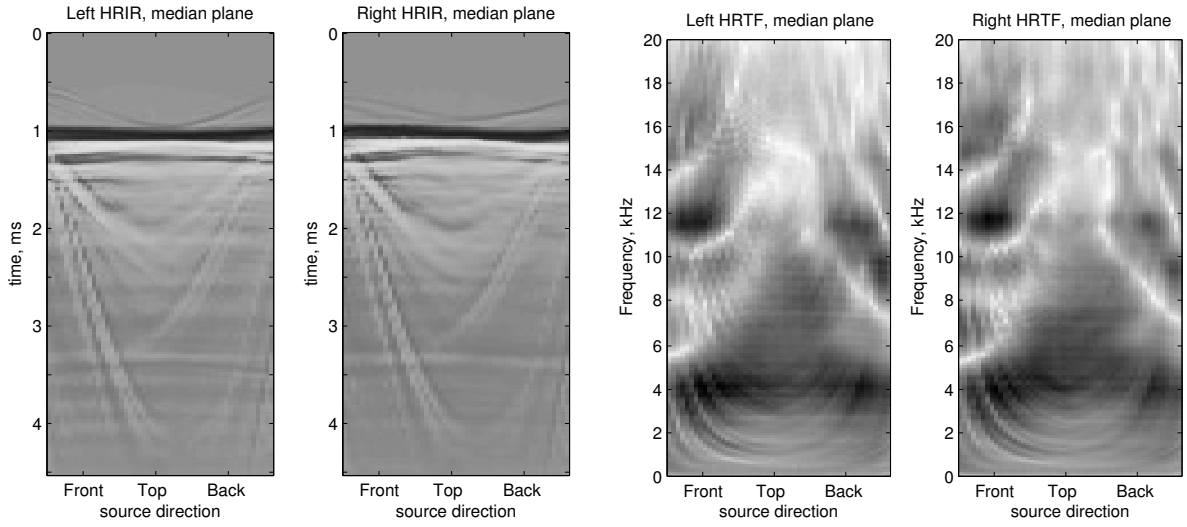


Figure 22.7: For sounds in the median plane, the responses of the two ears are essentially identical; the zero ITD and ILD cues indicate a sound straight ahead, or overhead, or straight behind, but there is no interaural cue for elevation angle. The HRIR (left panels) and HRTF (right panels) do show prominent elevation dependence, but the cue is essentially monaural, not based on a difference between the ears. The prominent spectral notches (white areas, since as always, we plot larger values as darker) above 5 kHz come from pinna diffraction, and the ripples below 4 kHz from torso (chest, shoulder, and back) echoes. Both provide useful cues to elevation (Algazi et al., 2001), and both might be detectable via either temporal or spectral patterns. These data are from a real person, not a dummy head, which is why the signals at the two ears are not quite identical.

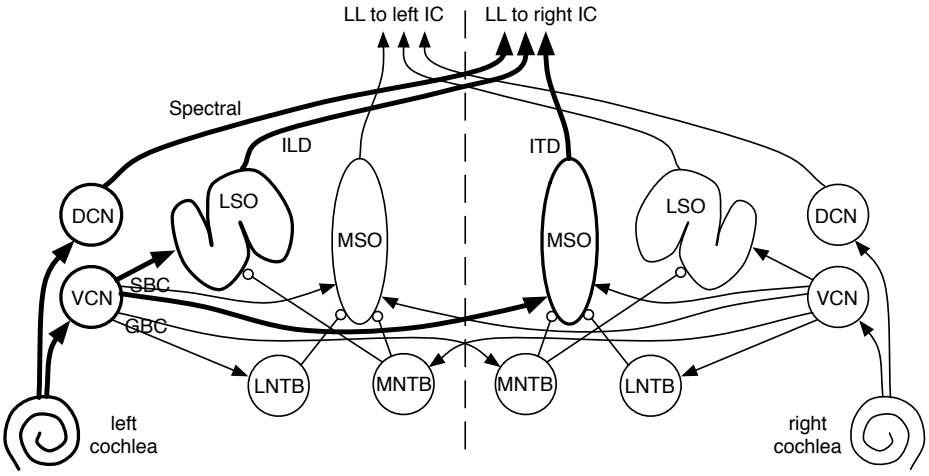


Figure 22.8: The ascending binaural circuits from the cochlea, through auditory nerve and ventral and dorsal cochlear nuclei (VCN and DCN), and through the olivary complex. Main excitatory pathways for a sound on the left are shown bold. A sound on the left primarily activates the left LSO and the right MSO, both of which project upward via the right lateral lemniscus (LL) and its nuclei to the right inferior colliculus (IC). Inhibitory connections are indicated with bubbles at their ends. The DCN is thought to provide spectral cues to IC, to help with vertical localization. Spectral, ILD, and ITD cues are probably integrated in IC. The division of VCN into AVCN and PVCN is not shown; the bushy cells (SBC and GBC) are mostly in AVCN.

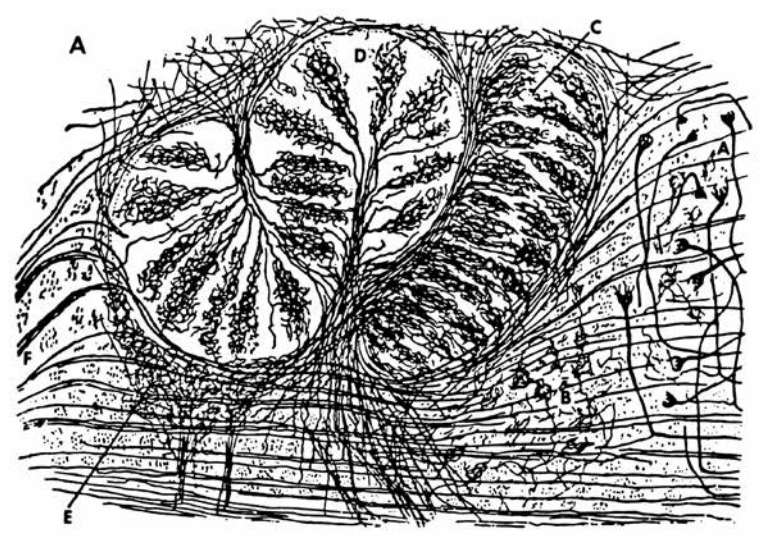


Figure 22.9: Superior olivary complex neurons as sketched by Cajal (1909), showing the S-shaped LSO and the sausage-shaped MSO, with trapezoid body neurons around them.

Chapter 23

The Auditory Brain

> ... how do we recognize what one person is saying when others are speaking at the same time (the "cocktail party problem")? On what logical basis could one design a machine ("filter") for carrying out such an operation? A few of the factors which give mental facility might be the following: (a) The voices come from different directions. (b) Lip-reading, gestures, and the like. (c) Different speaking voices, mean pitches, mean speeds, male and female, and so forth. (d) Accents differing. (e) Transition-probabilities (subject matter, voice dynamics, syntax ...).
>
> — "Some experiments on the recognition of speech, with one and with two ears," Cherry (1953)

> ... the majority of neurons in auditory thalamus and cortex coded well the presence of abstract entities in the sounds without containing much information about their spectro-temporal structure, suggesting that they are sensitive to abstract features in these sounds.
>
> — "Auditory abstraction from spectro-temporal features to coding auditory entities," Chechik and Nelken (2012)

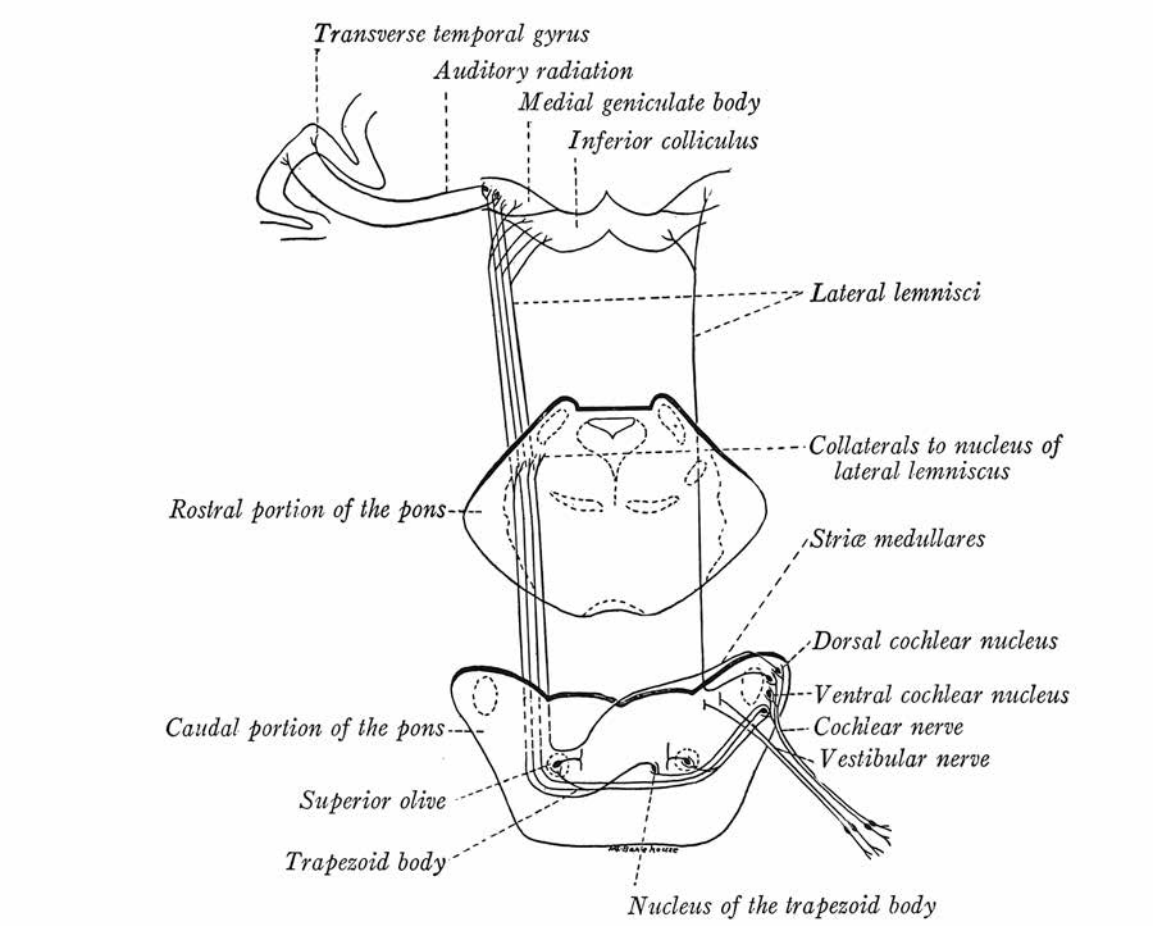


Fig. 233.—Diagram of the auditory pathway. (Based on the researches of Cajal and Kreidl.)

Figure 23.1: The afferent connections in the auditory nervous system, as rendered by Miss M. E. Bakehouse for Ranson (1920). The lower areas, labelled in the "caudal portion of the pons," are near the boundary between the medulla oblongata, the lower part of the brainstem, where the auditory nerve enters the brain at the cochlear nucleus, and the pons, the middle part of the brainstem. The main auditory area of the midbrain, the upper part of the brainstem, is the inferior colliculus (IC), and of the thalamus is the medial geniculate body (MGB). We have expanded our knowledge of the connections and functions and sub-areas since then, but the overall anatomy as known a century ago remains accurate. As the drawing shows, the main afferent projections go from the cochlea to contralateral brain areas—sounds from the left are processed on the right, and vice-versa.

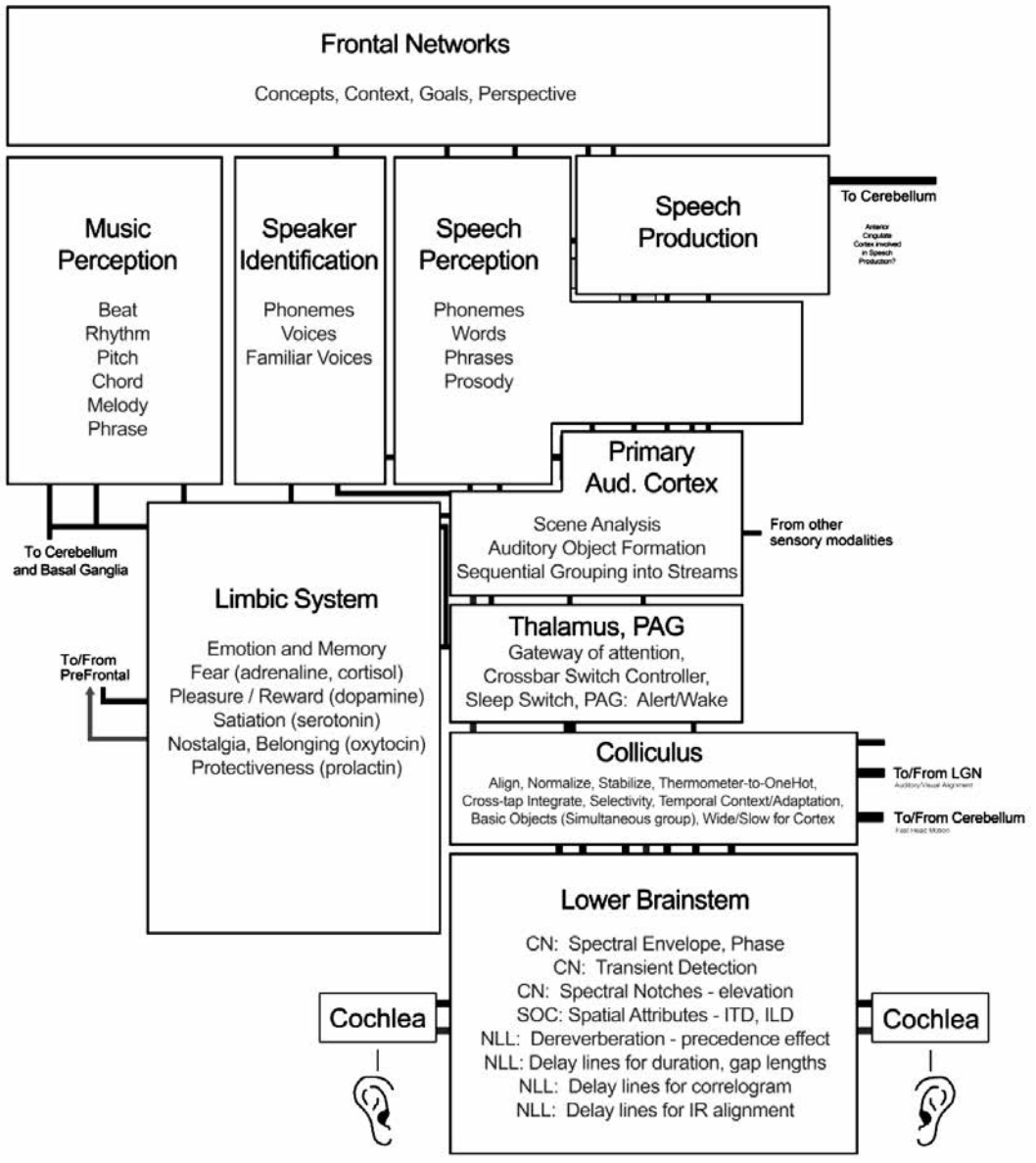


Figure 23.2: The brain function block diagram of Watts (2012) shows a hypothesized assignment of functions to structures. [Figure 1 (Watts, 2012) reproduced with permission of Springer.]

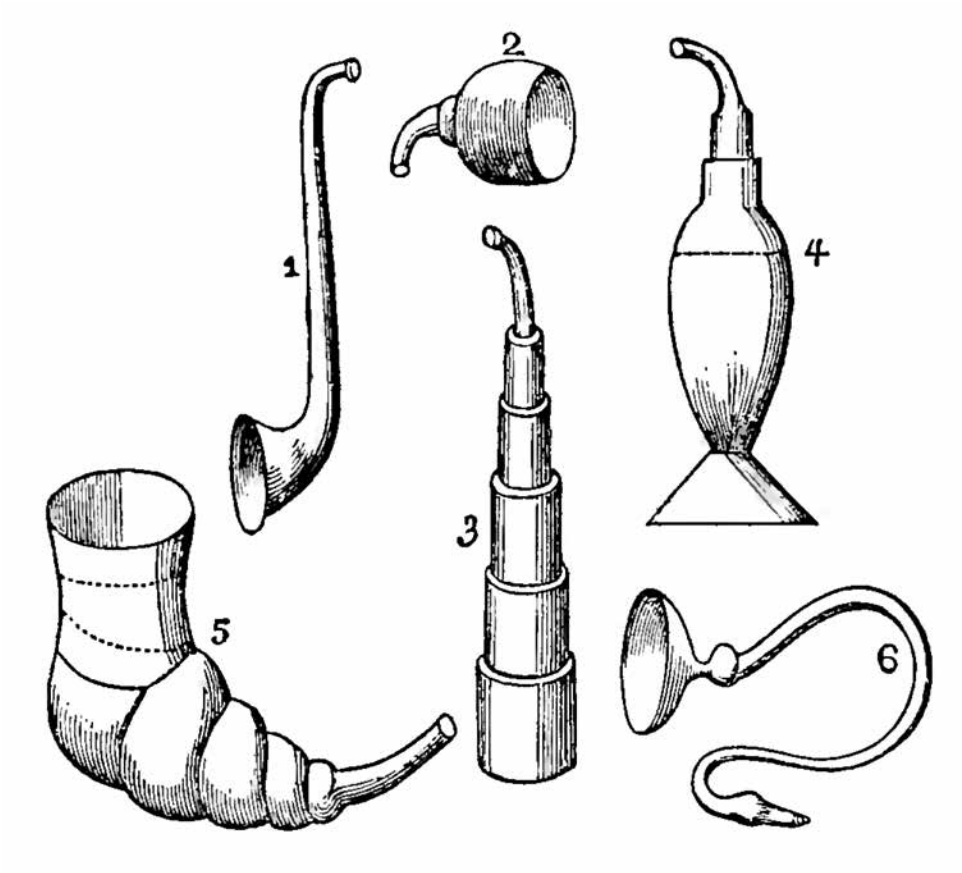
Part V

Learning and Applications

Part V Dedication: Max Mathews

This part is dedicated to the memory of Max Vernon Mathews (1926–2011), the father of computer music. Max had a decades-long focus on applications of computers to hearing and to music analysis, synthesis, and performance. His work on computer speech, music, and hearing started in the late 1950s at Bell Labs (Mathews, 1959, 1961, 1963). I had the opportunity to know Max at Stanford's CCRMA (Center for Computer Research in Music and Acoustics) where he worked for many years. When I taught my Human and Machine Hearing course at Stanford in 2010 (Psych 303, in affiliation with the Mind, Brain, and Computation center), Max came and audited the class once a week, climbing the stairs to the third floor with his hiking sticks. He invited me to his lab and explained the "coupled-form" filter that he was using for music synthesis; I subsequently adopted it as the basis for the digital implementation of my various cochlear filter models, so it figures prominently in earlier parts of the book.

In this part, we discuss the top two layers of our simple framework for machine hearing systems: types of systems that can be trained to address machine hearing applications, and ways that features can be extracted into a form suitable to be presented as inputs to such systems. We discuss several example applications, including ones on which we have published studies, and a survey of some others.



An early application of machine hearing concepts was in the improvement of hearing aids. These improved ear trumpets (Turnbull, 1887) are predecessors to more sophisticated hearing aids.

Chapter 24

Neural Networks for Machine Learning

In order for a digital neocortex to learn a new skill, it will still require many iterations of education, just as a biological neocortex does, but once a single neocortex somewhere and at some time learns something, it can share that knowledge with every other digital neocortex without delay. We can each have our own private neocortex extenders in the cloud, just as we have our own private stores of personal data today.

— *How to Create a Mind*, Ray Kurzweil (2012)

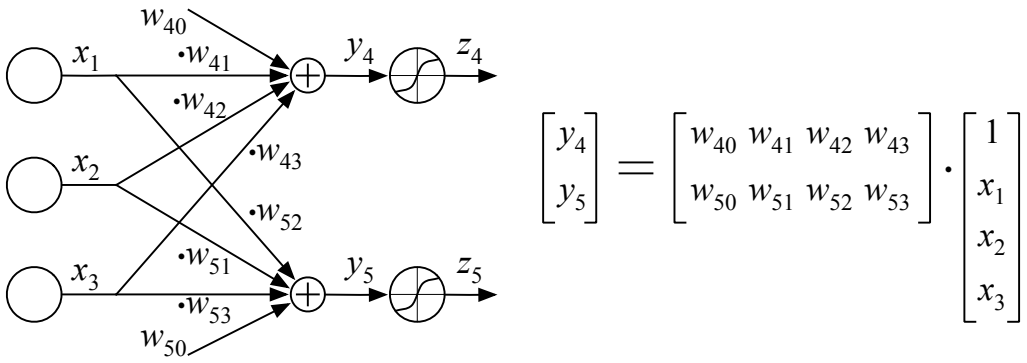


Figure 24.1: A three-input two-output single-layer perceptron with threshold adjustments, as a signal flow graph, and with the linear part as a matrix–vector multiplication. Each linear combination y_j is the dot product of the input vector with a row of the weight matrix: $y_j = \sum_i w_{ji} x_i$ (where $x_0 = 1$ to allow threshold adjustment through w_{j0}). The optional nonlinear part is shown as sigmoidal (s-shaped) nonlinearities. The empty circles on the left are input units, which may sometimes represent the outputs of another perceptron.

Example Problem for Neural Networks

Consider this example problem: we want to classify talkers as male or female, based on a single vowel utterance, using only measurements of the first and second formant frequencies (F_1 and F_2). For data, we use an online database of vowel data from North Texas talkers (Assmann and Katz, 2000).

With data from the ten adult male and ten adult female talkers, using only the averages between initial and final F_1 and F_2 values for 12 different vowels, we construct training and testing sets of two-dimensional features and two-class (one-bit) targets. The first five males and the first five females, repeating each vowel an average of 10 times each, make a training set of about 1200 points; the second half of the talkers make a similar-size testing set.

First, we train a simple perceptron: three trainable weights connect the F_1, F_2, and constant inputs to the decision output. The resulting decision boundary is necessarily a straight line in F_1–F_2 space, as shown in Figure 24.2

It is evident that there is a lot of confusion in the middle of the F_1–F_2 space. Males tend to have longer vocal tracts, and hence lower formant frequencies, than females, but some vowels also have lower F_1 and F_2 than other vowels, so the middle of the feature space has a confusing mixture of clumps of male and female sample points. If we added pitch (F_0) as a feature, the problem would be relatively easy, since pitch alone is enough to distinguish male from female with better than 95% accuracy. We use the F_1–F_2 example because the nonseparable nature of the problem helps to illustrate some of the issues in machine learning. Turner et al. (2009) provide a much more in-depth discussion of the relations between formant frequencies, vowel identity, and talker gender, vocal tract length, and pitch.

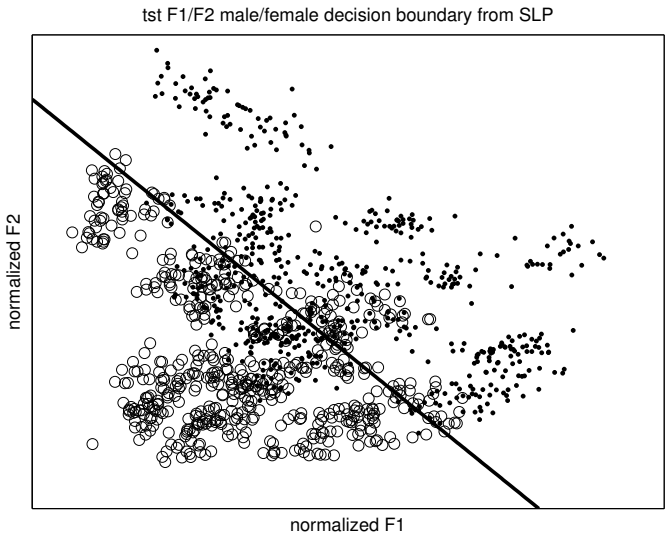


Figure 24.2: A feature-space map showing the decision boundary of a single-layer perceptron (SLP) classifying talker gender from F_1–F_2 data. Data points from the testing set are shown, with males as circles and females as dots (corresponding training points can be seen in Figure 24.6). Near the decision boundary, many classification errors are made: 267 errors of 1200 items in the testing set. This perceptron also makes 298 errors on the training set. The first two formant frequencies are apparently not quite sufficient to distinguish male from female talkers.

Quick Linear Training in MATLAB

In MATLAB, training a linear perceptron is a one-liner, if the training inputs and targets are already gathered up into matrices `x` and `targets`, with a column per training sample:

```
W = x \ targets;  % Least-squares training
y = W * x;  % y should now be close to targets
```

If there exists a matrix that will map all the inputs to all the targets exactly, this will find it, in which case `y` will be equal to `targets`. More generally, MATLAB's matrix division operator will find the matrix that maps `x` to a `y` near `targets`, minimizing the total squared error of `y` relative to `targets`. This formulation is easy to set up and solve as a least-squares problem in systems other than MATLAB, too, of course.

This kind of least-squares training is good for *regression* problems: learning a function that approximately maps input values to the training values. But perceptrons, and their training algorithm, were actually designed for *classification*, where the goal is to minimize a count of misclassifications, not a sum of squared errors. Whether for regression or classification, we typically use nonlinearities in our perceptrons, such that training is not quite this easy. That is, the *loss function* that we are minimizing may not be the total squared error of the linear part of the perceptron operation, so a different method for minimizing the loss needs to be found.

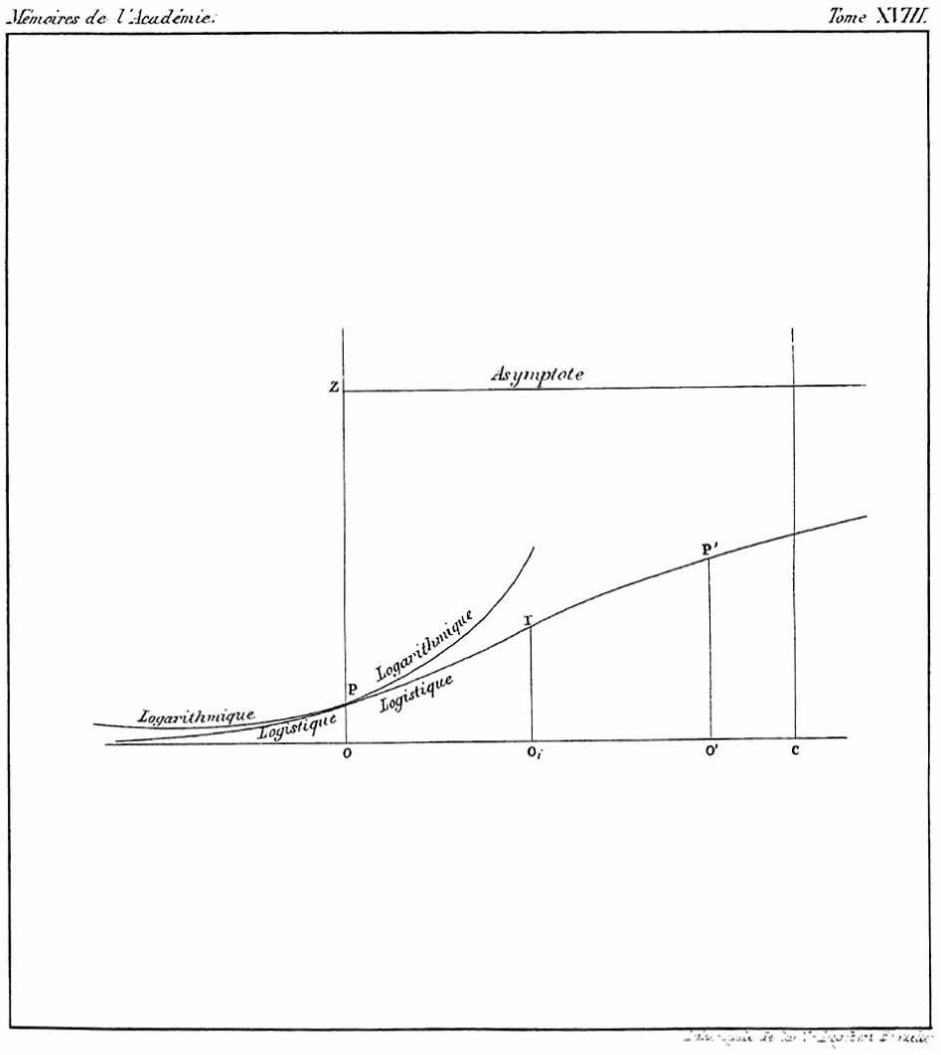


Figure 24.3: The logistic (*logistique*) function, which plays a big role in statistics and in artificial neural networks, was orignally derived as a population growth function (Verhulst, 1845), showing how exponential growth might be moderated as the carrying capacity of a region is approached.

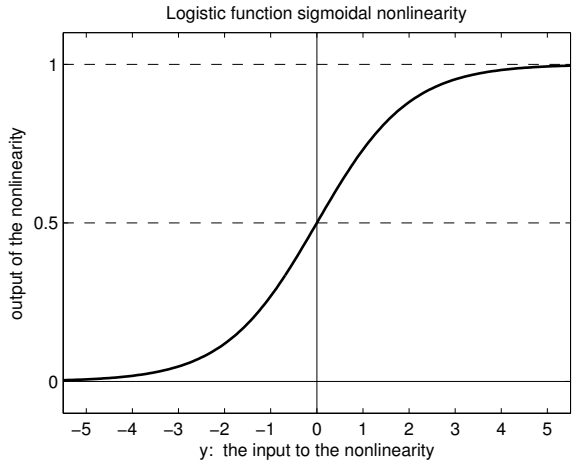


Figure 24.4: A popular nonlinearity at the output of a perceptron is the logistic function, which maps any value to an output between 0 and 1.

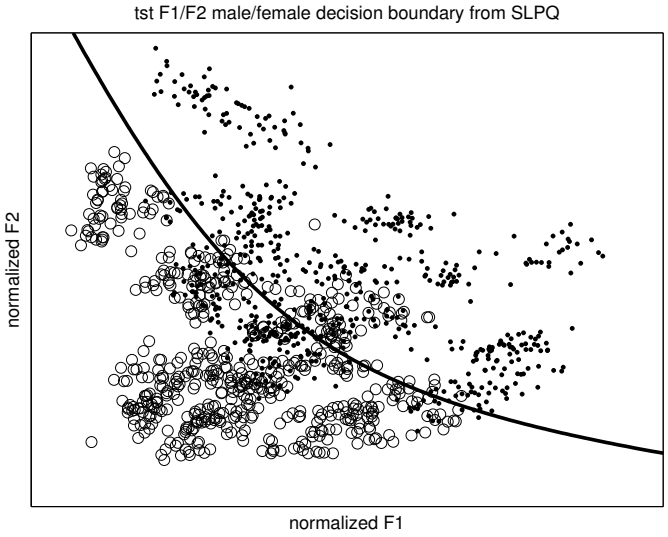


Figure 24.5: A feature-space map showing the decision boundary of a single-layer perceptron classifying talker gender from F_1–F_2 data, when the input feature vector is augmented with three quadratic dimensions (two squares and a cross term). The nonlinearites at the input help, but only a little; test-set errors are reduced to 247, from the 267 of Figure 24.2.

Multilayer Perceptron Examples

If we apply a much more powerful neural network, an MLP with two hidden layers, with six neurons per hidden layer, we produce a classifier that does a much better job of separating the training data into male and female regions of the feature space, as shown in Figure 24.6. This map illustrates the power of an MLP to learn complicated decision boundaries from training data. It also illustrates the problem of overfitting: when we test it on the test set, we find it gets even more errors than the simple linear perceptron did!

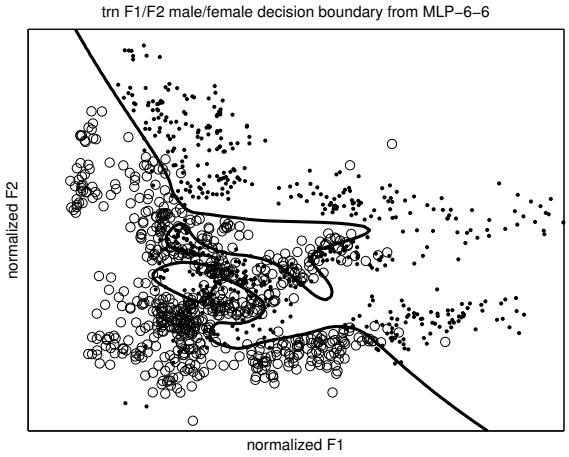


Figure 24.6: A map showing the decision boundary of a "too powerful" MLP, classifying talker gender from F_1–F_2 data. Data points from the training set are shown, using the same symbols as in Figure 24.2 (where the testing data points can be seen). The complicated decision boundary does a fairly good job of separating the training talkers into male and female, making only 186 errors on the training set. But it makes 290 errors on the testing set, which is worse than the simple linear perceptron.

A key problem in machine learning is to find a good compromise between the powerful capability of a trainable system to model the training data, and the need to generalize without over-fitting. One way to do that is to use a network with just enough trainable weights, or just enough modeling power. The network result shown in Figure 24.7 is an example of that approach: on our example problem, a net with only one hidden layer of 5 units makes fewer errors on the test set than do larger and smaller nets.

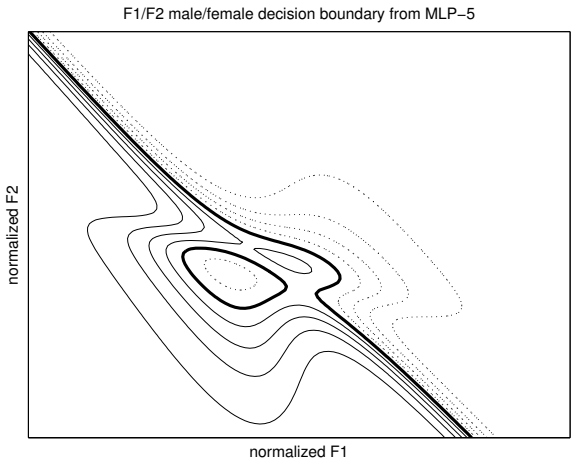
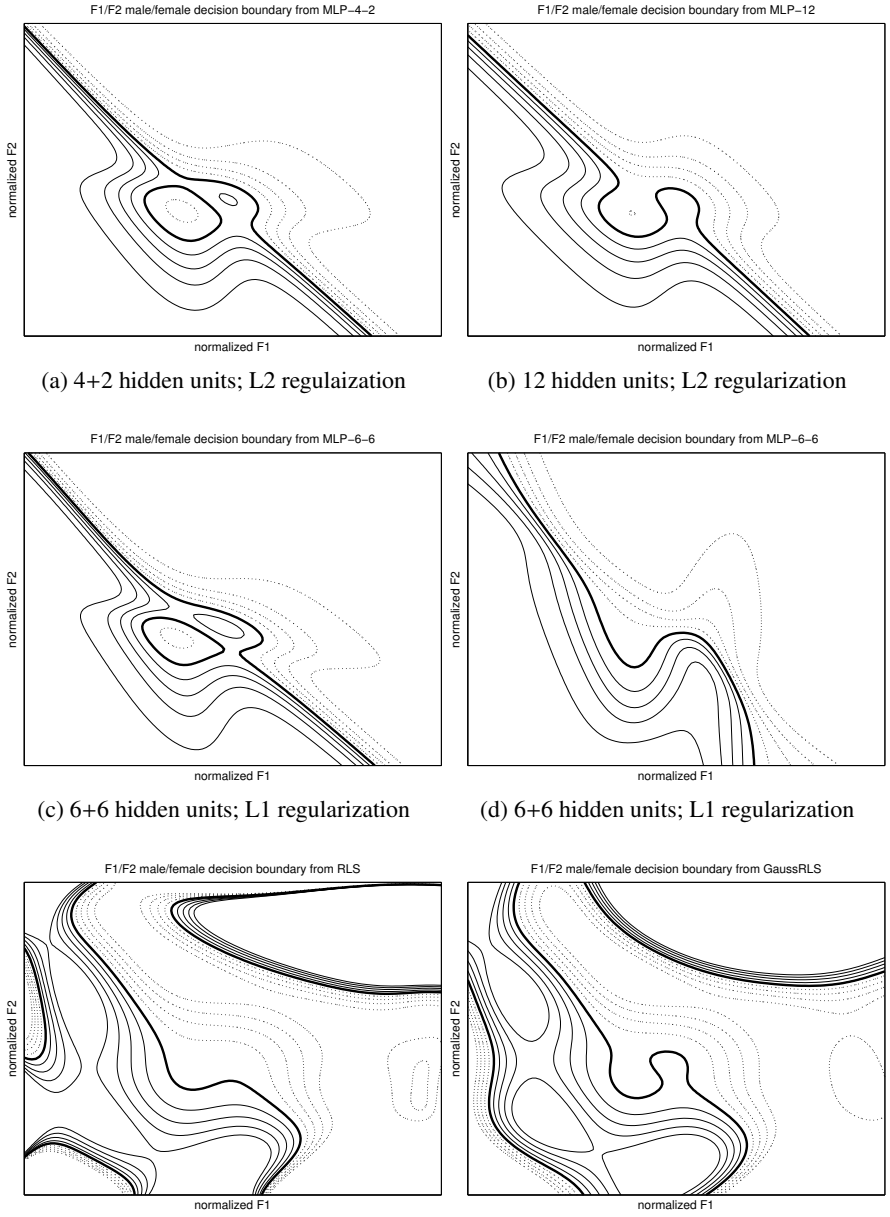


Figure 24.7: A map showing the decision boundary of a smaller MLP, classifying talker gender from F_1–F_2 data, along with contours of estimated class probability, in multiples of 10%. This network makes only 214 errors on the test set.



(a) 4+2 hidden units; L2 regulaization

(b) 12 hidden units; L2 regularization

(c) 6+6 hidden units; L1 regularization

(d) 6+6 hidden units; L1 regularization

(e) RLS with 8th-order polynomials

(f) RLS with Gaussian kernel

Figure 24.8: Decision boundaries of different neural nets and least-squares classifiers, all of which yield 206 to 218 errors on the testing set. Panels (a) and (b) are two-hidden-layer and one-hidden-layer nets trained with L2 regularization (weight decay). Panels (c) and (d) are identical two-hidden-layer structures trained with L1 regularization from different random starts; in both cases, enough weights go to zero to reduce them to effectively the structure in panel (a), with only 4 active units in the first hidden layer and 2 in the second. Panels (e) and (f) are examples of the modern regularized-least-squares method operating on a large nonlinearly-expanded input space, using polynomial expansion for (e), and Gaussians at the training points in (f). The RLS methods have formed additional decision regions to accommodate the training outlier points seen in Figure 24.6.

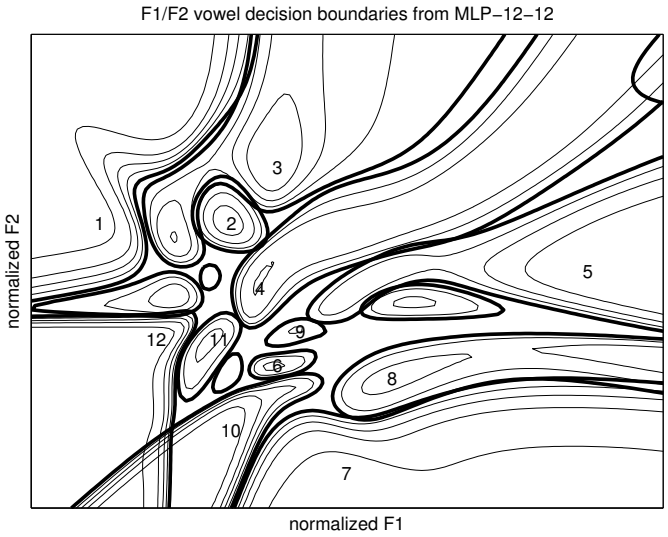
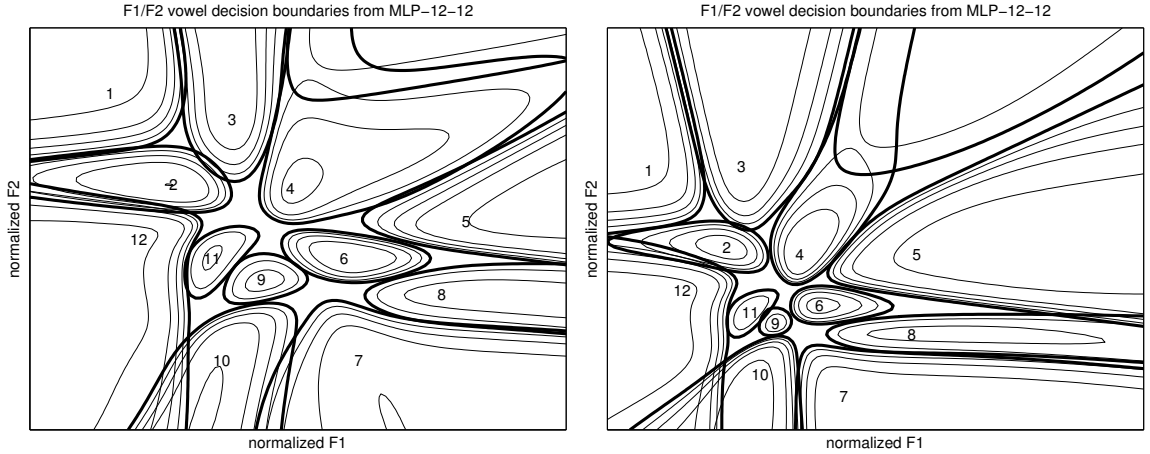


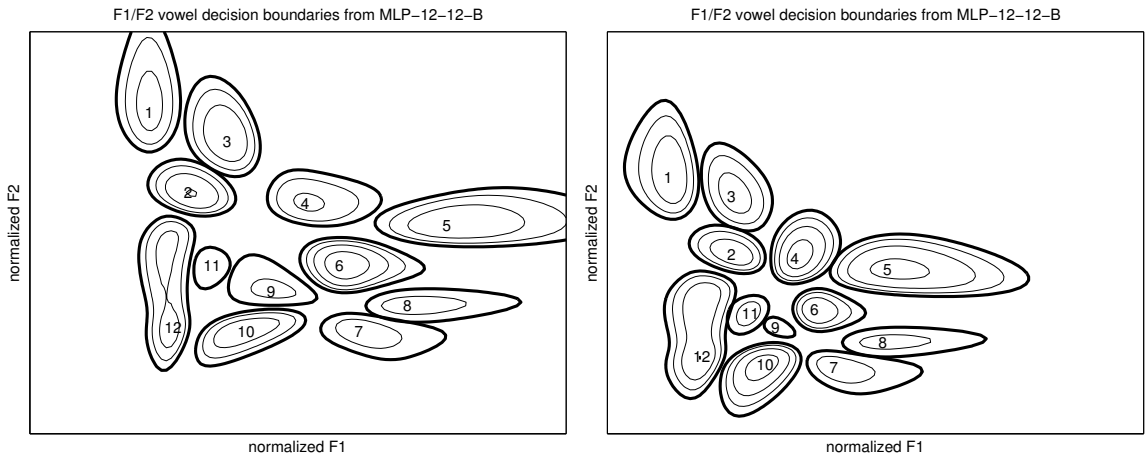
Figure 24.9: A map showing the decision boundaries of a 12-output MLP with 12+12 hidden units, classifying vowel identity from F_1–F_2 data for a mixed-gender training population. Vowels 2, 6, 9, and 12 have two regions each, but only one of each is labeled. Probability contours for 50%, 60%, 70%, 80% and 90% are shown, with the 50% contour darker. The probability estimates are not constrained to add to 1, and in regions with no training data they often add to more than 1, as the crossing contours show. The first-choice vowel classification accuracy with this net is about 50% on the testing set.



(a) Female vowel regions

(b) Male vowel regions

Figure 24.10: Decision boundaries of a neural network that classifies three-dimensional features (F_1, F_2, and talker gender) into 12 vowel classes; on the left, the gender input is low for female, and on the right it is high for male. Vowels are much more separable with the additional input information; the first-choice accuracy goes from about half to about two-thirds with this additional input. A continuous pitch input is similarly helpful, since it tends to correlate with the talker's vocal-tract length at least as well as gender does, but such a 3D feature space is harder to illustrate.



(a) Female vowel regions (b) Male vowel regions

Figure 24.11: Decision boundaries as in Figure 24.10, but for a net trained with additional synthetic training points, uniformly spread over the F_1–F_2 plane, with targets all low to represent a null class, or no vowel. Now the estimated vowel class probabilities usually add to less than one, especially in regions with no vowel training points. Vowels 7 and 8 are the ones corresponding to the English words *hawed* and *hod*, which are highly confusable, often not distinguished by American English speakers, even in North Texas, so their responses overlap and never get to probabilities as high as 0.7. Vowel 9, in *herd*, is also very fuzzy in F_1–F_2 space, as the *r* sound is mostly signaled by a low third formant; it is especially confusable with vowel 11, *hood*.

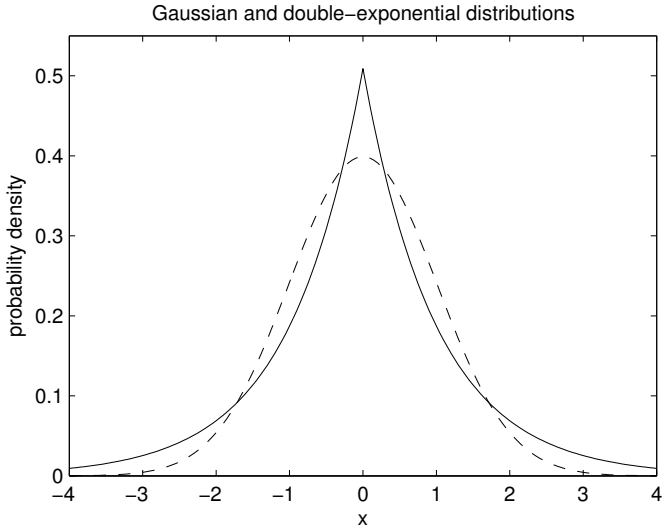


Figure 24.12: Gaussian (dashed curve) and double-exponential (solid curve) prior distributions on parameters, compared here with variances equal to 1, correspond to using L2 and L1 regularizers, respectively. The double-exponential distribution has many more values at zero, fewer small nonzero values, and more large values, compared to the Gaussian.

Chapter 25

Feature Spaces

> Rather than represent the entire transformation from the set of input variables to the set of output variables by a single neural network function, there is often great benefit in breaking down the mapping into an initial *pre-processing* stage, followed by the parameterized neural network model itself. ... The use of pre-processing can often greatly improve the performance of a pattern recognition system, and there are several reasons why this may be so.
>
> — *Neural Networks for Pattern Recognition*, C. M. Bishop (1995)

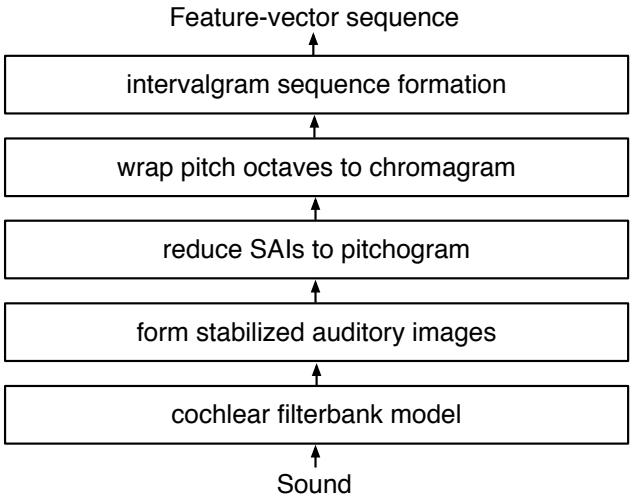


Figure 25.1: This multistage feature extraction pipeline for melody matching, described in detail in Chapter 27, maps into the first three levels of our *four-layer model* of Figure 1.5. Some of the stages here could alternatively be pushed into the uppermost layer, the machine learning system, with the possbiility that they would be end-to-end optimized for the application.

Chapter 26

Sound Search

This task aims at identifying the pictures relevant to a few word query, within a large picture collection. Solving such a problem is of particular interest from a user perspective since most people are used to efficiently access large textual corpora through text querying and would like to benefit from a similar interface to search collections of pictures.

— "A discriminative kernel-based model to rank images from text queries," Grangier and Bengio (2008)

Table 26.1: Parameters used for the SAI experiments

Parameter set	Smallest box	Total boxes	Means per box	VQ MP	Box cutting
Default "baseline"	32×16	49	256	VQ	Up
Codebook sizes	32×16	49	4, 16, 64, 256, 512, 1024, 2048, 3000, 4000 6000 8000	VQ	Up
Matching pursuit	32×16	49	4, 16, 64, 256, 1024, 2048, 3000	MP	Up
Box sizes (down)	16×8 32×16 64×32	1, 8, 33, 44, 66 8, 12, 20, 24 1, 2, 3, 4, 5, 6	256	VQ	Down
Box sizes (up)	16×8 32×16 64×32	32, 54, 72, 90, 108 5, 14, 28, 35, 42 2, 4, 6, 10, 12	256	VQ	Up

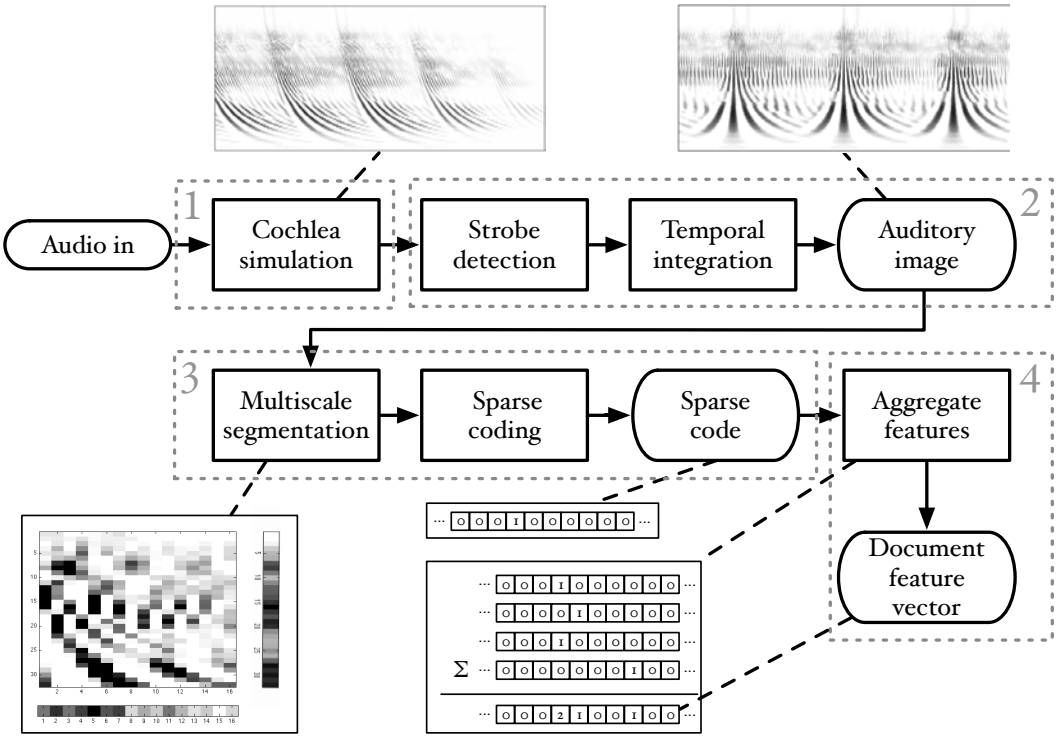


Figure 26.1: Generating sparse codes from an "audio document" using an auditory front end, in four steps: (1) cochlea simulation; (2) stabilized auditory image creation; (3) sparse coding of multiscale patches; (4) aggregation into a "bag of features" representation of the entire audio document. Steps 3 and 4 here correspond to the feature extraction layer in our four-layer system structure. From the point of view of the fourth layer, a PAMIR-based learning and retrieval system, this entire diagram represents a front end providing abstract sparse features for audio document characterization.

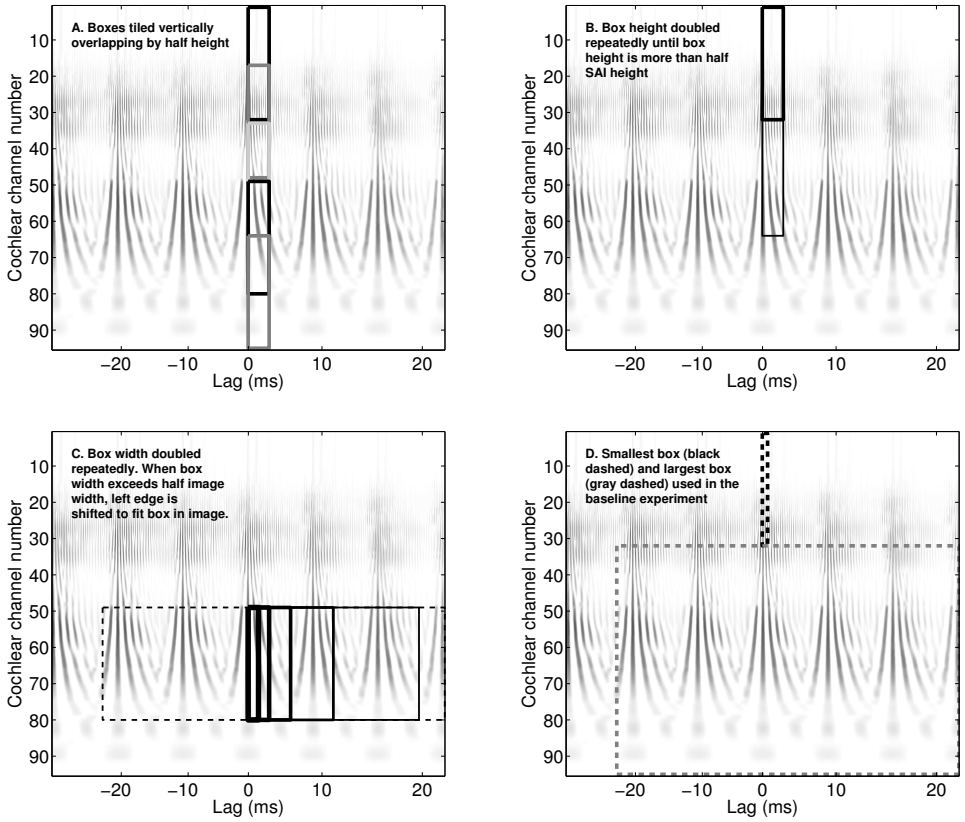


Figure 26.2: Defining a set of local rectangle regions in the SAI: rectangles are chosen to have different sizes, to capture multiscale patterns. In the default set of parameters we used, the smallest rectangle is 16 samples in the lag dimension and 32 channels high, and the largest is 1024 samples by 64 channels.

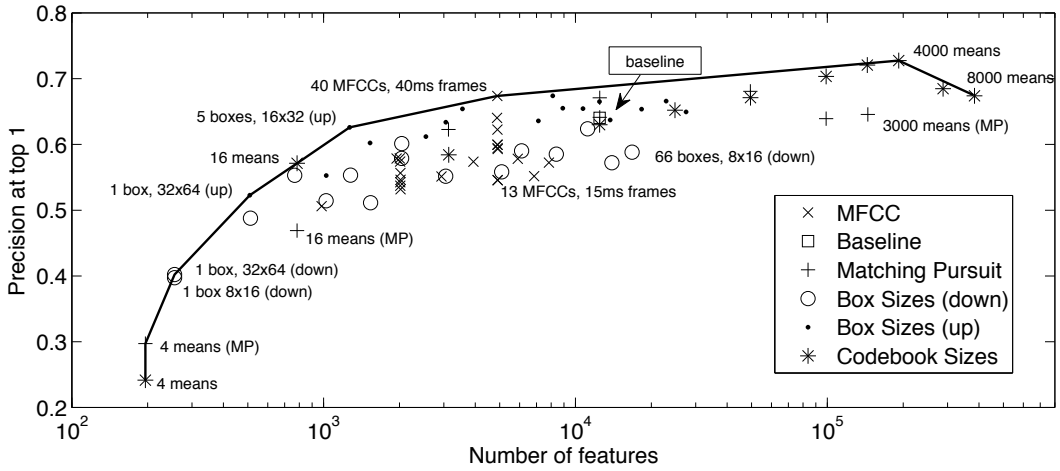


Figure 26.3: Ranking precision at the top-1 sound plotted against feature count, for all experiments. Selected experiment names are plotted on the figure near each point. The different experiment sets are denoted by different markers. The convex hull joining the best-performing points is plotted as a solid line.

top-*k*	SAI	MFCC	Percent error reduction
1	27	33	18%
2	39	44	12%
5	60	62	4%
10	72	74	3%
20	81	84	4%

Table 26.2: Comparison of error at top-*k* for best SAI and MFCC configurations (error defined as one minus precision).

Query	SAI file (labels)	MFCC file (labels)
tarzan	Tarzan-2 (tarzan, yell) tarzan2 (tarzan, yell) 203 (tarzan) wolf (mammal, wolves, wolf) morse (mors, code)	TARZAN (tarzan, yell) 175orgs (steam, whistle) mosquito-2 (mosquito) evil-witch-laugh (witch, laugh, evil) Man-Screams (horror, scream, man)
applause audience	27-Applause-from-audience 30-Applause-from-audience golf50 (golf) firecracker 53-ApplauseLargeAudienceSFX	26-Applause-from-audience phaser1 (trek, phaser, star) fanfare2 (fanfar, trumpet) 45-Crowd-Applause (crowd, applause) golf50
gulp	tite-flamn (hit, drum, roll) water-dripping (water, drip) Monster-growling (horror, monster, growl) Pouring (pour, soda)	GULPS (gulp, drink) drink (gulp, drink) california-myotis-search (blip) jaguar-1 (bigcat, jaguar, mammal)

Table 26.3: Top documents obtained for queries that performed very differently between the SAI and MFCC feature based systems.

Query, label		SAI + MFCC errors
clock-tick	cuckoo	8
door knock	door	8
evil laugh	laugh	7
laugh witch	laugh	7
bell-bicycle	bell	7
bee-insect	insect	7

Table 26.4: Error analysis. Queries that were repeatedly confused for another query. All pairs of true-label and confused labels with total count above seven are listed.

Chapter 27

Musical Melody Matching

I hope my critics will excuse me if I conclude from the opposite nature of their objections that I have struck out nearly the right path. As to my Theory of Consonance, I must claim it to be a mere systematisation of *observed facts* (with the exception of the functions of the *cochlea* of the ear, which is moreover an hypothesis that may be entirely dispensed with). But I consider it a mistake to make the Theory of Consonance the essential foundation of the Theory of Music, and, I had thought that this opinion was clearly enough expressed in my book. The essential basis of Music is *Melody*.

— *On the Sensations of Tone*, Hermann Ludwig F. Helmholtz (1870)

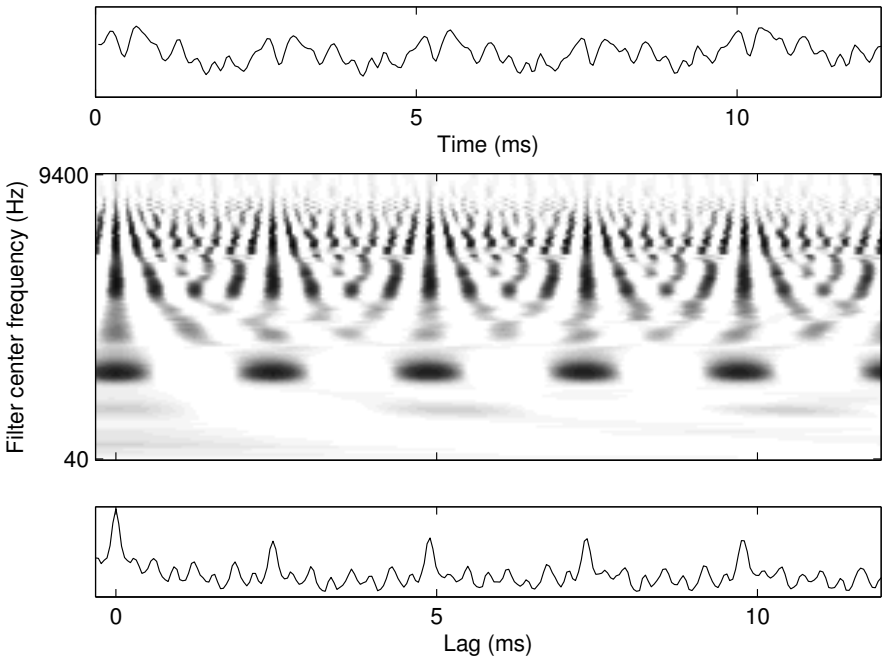


Figure 27.1: Waveform (top panel), stabilized auditory image (SAI, middle panel), and SAI temporal profile (bottom panel) for a human voice singing a note.

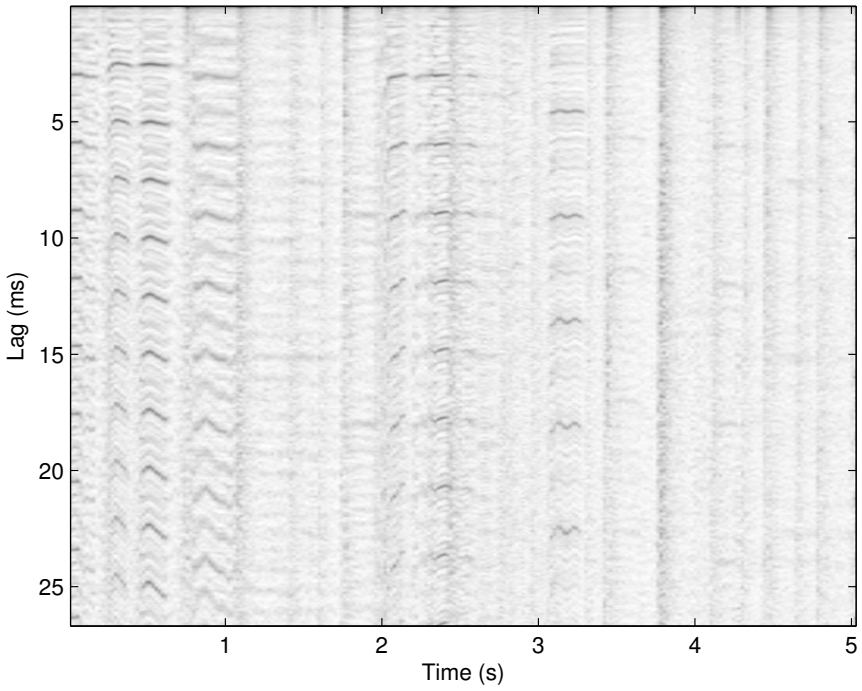


Figure 27.2: A pitchogram created by stacking a number of SAI temporal profiles in time. The lag dimension of the auditory image is now on the vertical axis. Dark ridges are associated with strong repetition rates in the signal.

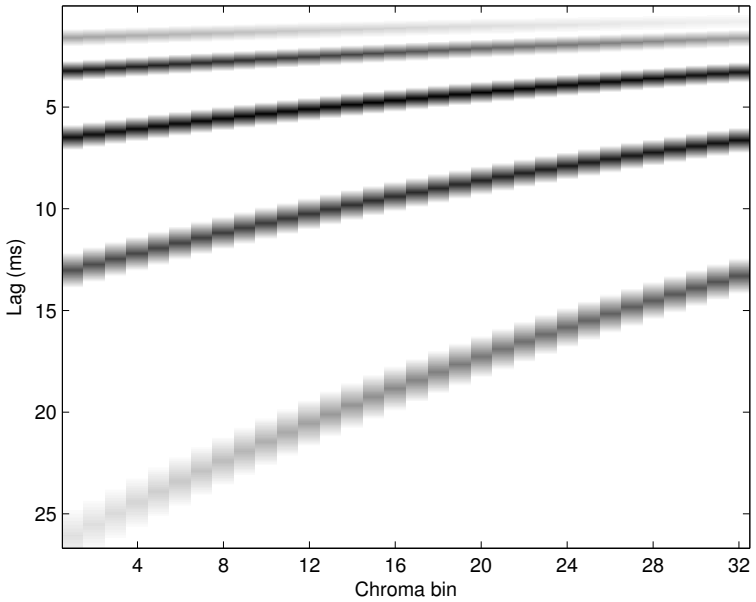


Figure 27.3: Weighting matrix to map from the time-lag axis of the SAI into 32 chroma bins

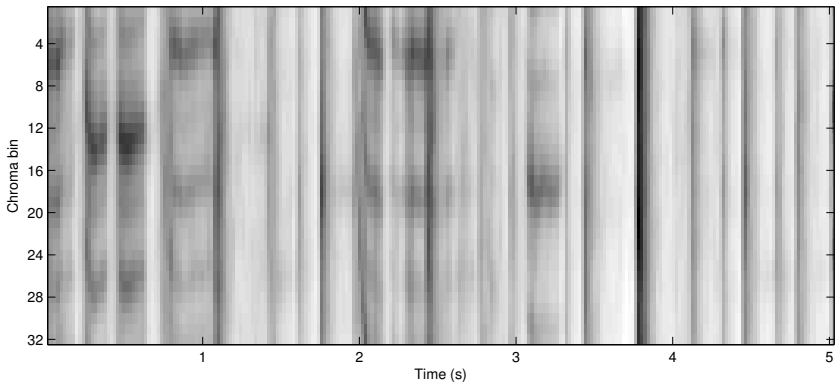


Figure 27.4: Chroma vectors generated from the pitchogram vectors shown in Figure 27.2.

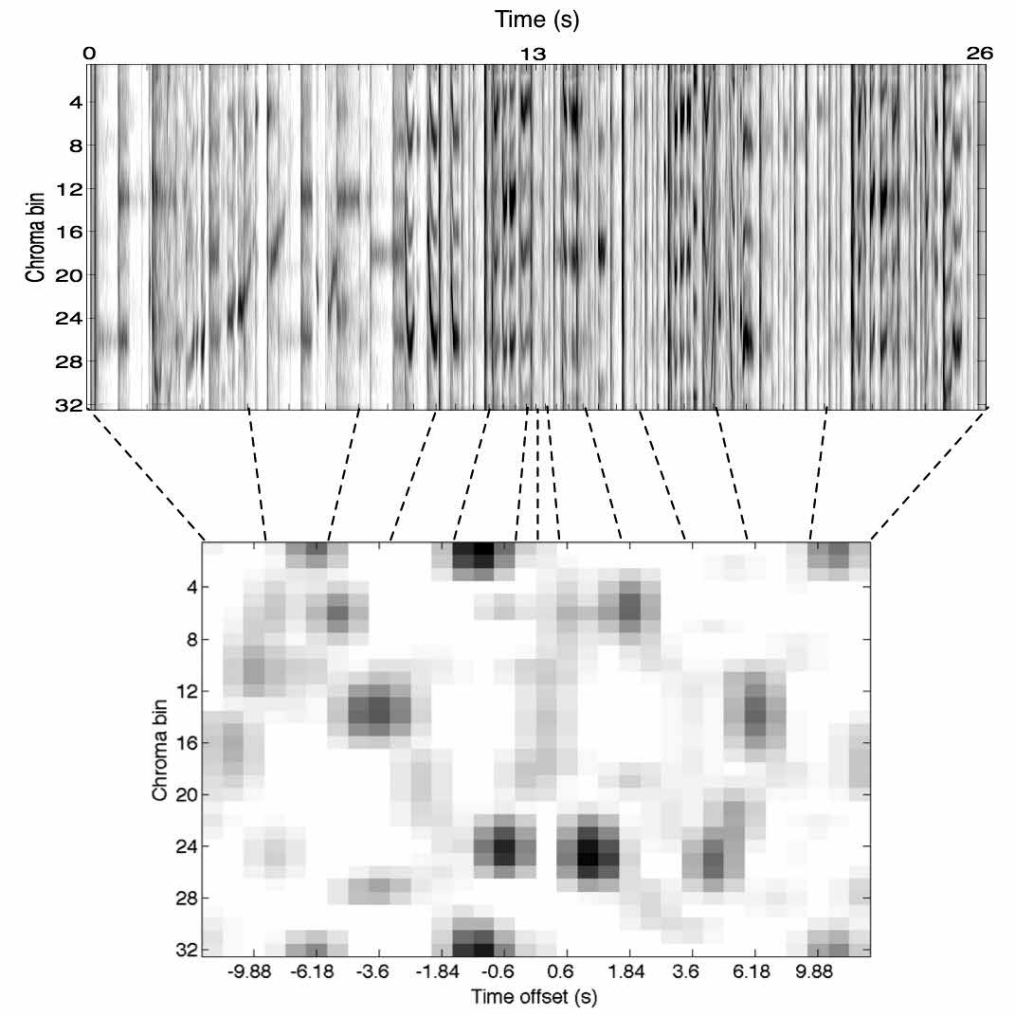


Figure 27.5: The intervalgram is generated from the chromagram using variable-width time bins and cross-correlation with a reference chroma vector to normalize chroma within the individual intervalgram.

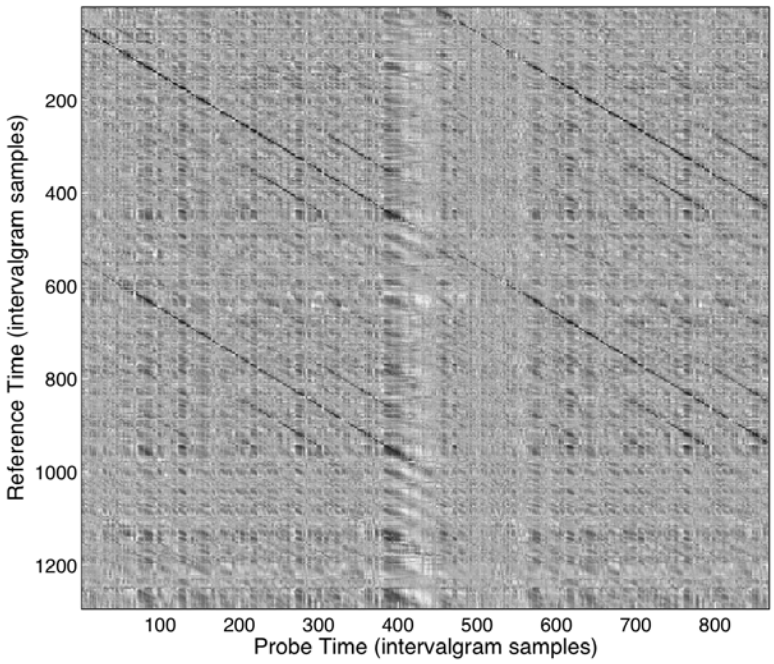


Figure 27.6: Example distance matrix for a pair of songs that share an underlying melody. The darker pixels show the regions where the intervalgrams match closely.

Parameter	Value
Chromagram step size (ms)	20
Chroma bins per octave	32
Total intervalgram width (s)	26.44
Intervalgram step size (ms)	240
Reference chroma vector width (chroma vectors)	4

Table 27.1: Parameters of the best intervalgram for cover song matching

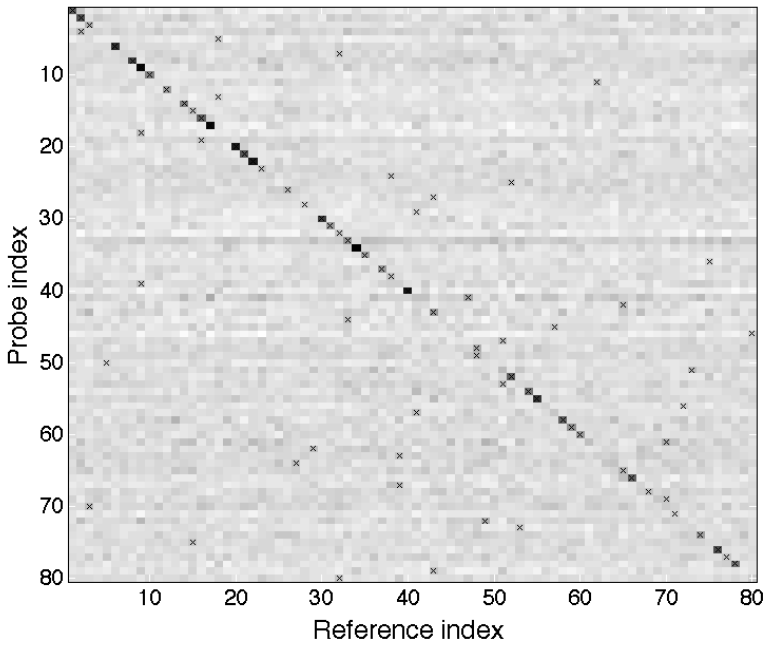


Figure 27.7: Scores matrix for comparing all probes and references in the dataset. Darker pixels denote higher scores, indicating a more likely match. Black crosses denote the best-matching reference for each probe.

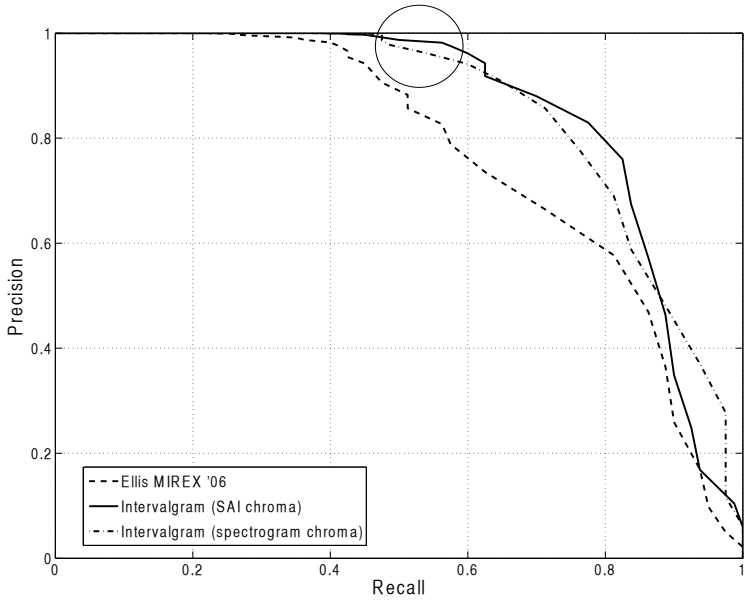
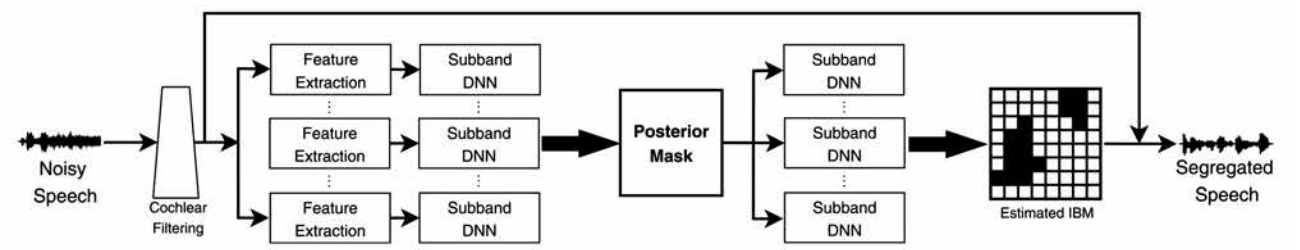


Figure 27.8: Precision–recall curves for the intervalgram-based melody-matching system described in this chapter, and the LabROSA MIREX'06 entry (Ellis and Cotton, 2007). Precision is one minus the probability of falsely matching a song as a cover, while recall is the probability of correctly identifying a cover song. In the high-precision region, near 50% recall and above 95% precision, shown circled, the SAI-based features lead to about half as many false matches as the spectrogram-based features.

Chapter 28

Other Applications

> Computational modeling of the auditory periphery has become an integral part of hearing and speech research in recent years. This reflects the importance of computers and computational models as a research tool for experimenting flexibly in the domain of complex auditory phenomena. Both our general understanding and the fragmental knowledge of details known from hearing research can be reconstructed and tested in the form of functional models.
>
> — "Auditory models for speech processing," Matti Karjalainen (1987)



Schematic diagram of the current speech-segregation system. DNN = deep neural network, IBM = ideal binary mask.

Figure 28.1: The hearing aid architecture of Healy et al. (2013), using a cochlear filterbank and binary masking, has been shown to yield a net intelligibility improvement for hearing-impaired subjects in noisy situations. [Figure 1 (Healy et al., 2013) reproduced with permission of AIP Publishing.]

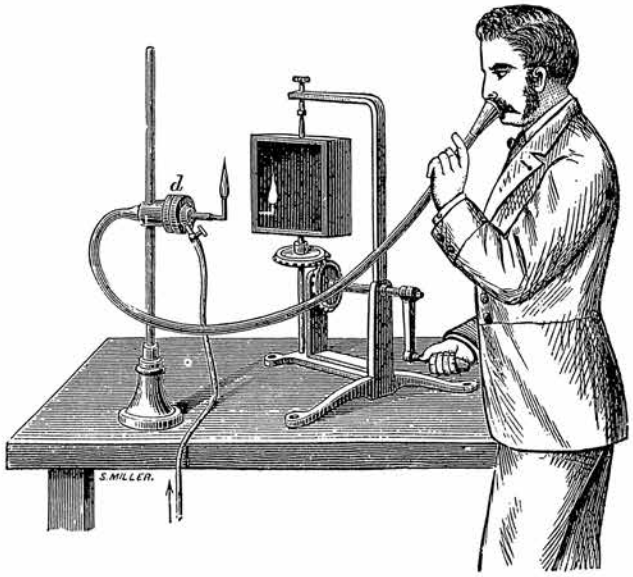


FIG. 432.—König's apparatus for illustrating the quality of vowel tones by a manometric flame.

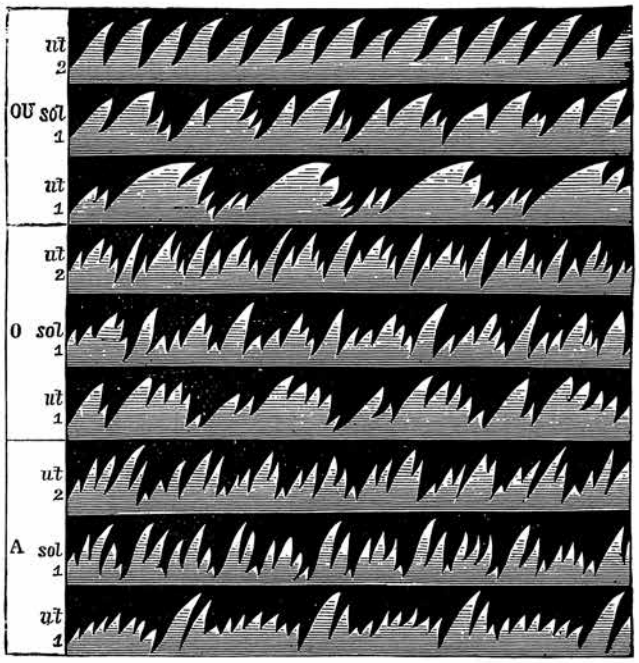


FIG. 433.—Flame pictures of the vowels *OU*, *O*, and *A*.—König.

Figure 28.2: Rudolf König's manometric flame apparatus was among the devices that Alexander Graham Bell used to visualize sound waveforms. The rotating quad mirror converted fast sound-induced modulations of the flame into spatio-temporal patterns—but not stabilized. These images from McKendrick (1889) show how the experimenter might use the apparatus, and the flame patterns that would appear in the mirror in response to several steady vowels, each at several different pitches.

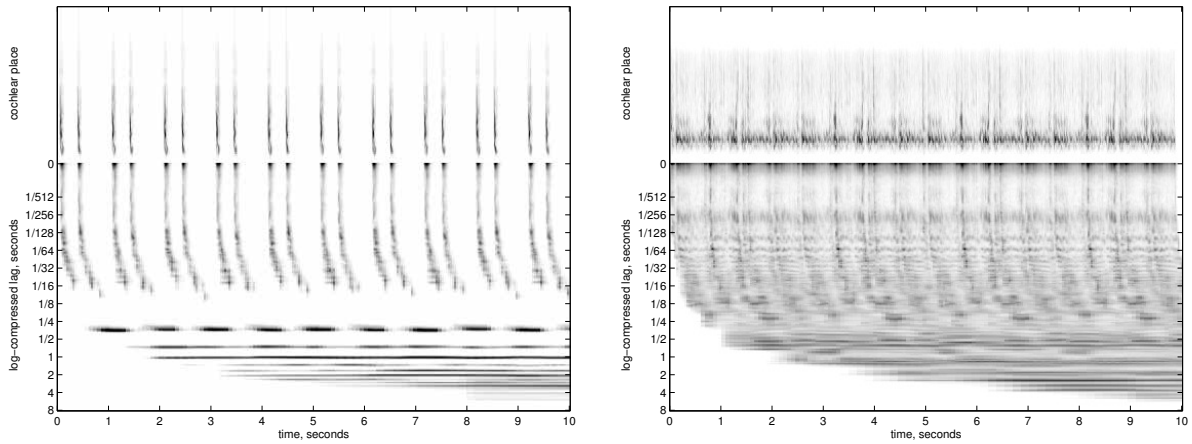


Figure 28.3: Phonocardiograms visualized with the cochleagram/log-lag pitchogram representation of Chapter 21, showing a normal clean low-frequency "lub-dub" heart sound on the left, and a heart with *patent ductus arteriosus* (PDA) on the right. The heart with PDA has a continuous murmur, or "machinery murmur"—essentially a modulated noise from blood squirting under pressure through an opening that should not be there. Other heart problems have more subtle sonic signatures.