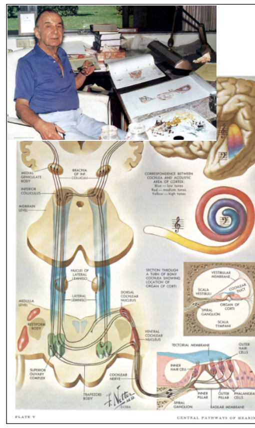
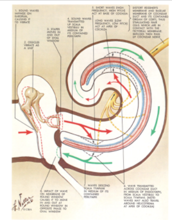
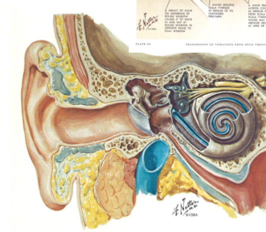
Computational Models of Mammalian Hearing: An “Auditory Image” Approach

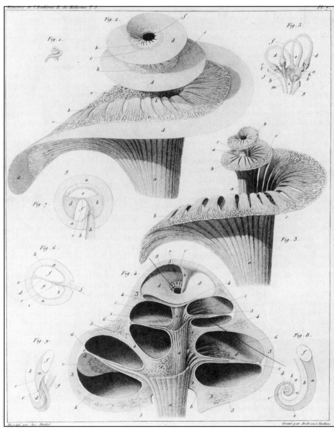
Dick Lyon
For Tom Dean’s Cortex Class
Stanford, April 14, 2010



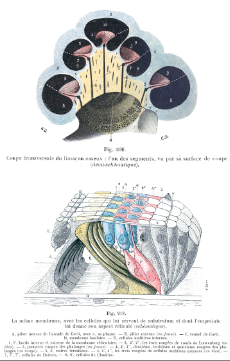
Frank Netter and his Ciba paintings



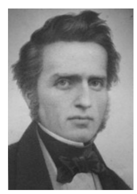




Breschet 1836, Testut 1897



August Seebeck 1841, 1844: time-domain pitch (from Schouten 1970 “The Residue Revisited”)



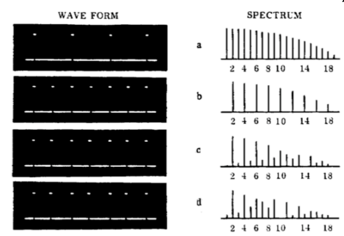
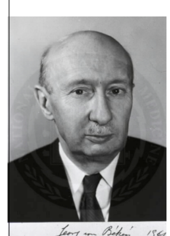


Fig. 2. Seebeck's experiment with four different pulse series.
The amplitudes in the spectra are those derived from the Fourier analysis of the four waveforms. The harmonics are indicated with their respective numbers. The indications of strength given below are in terms of perceived loudness.
(a) Spacing of holes 20°. Fundamental strong, harmonics weak.
(b) Spacing of holes 10°. The octave (now the fundamental) strong, higher even harmonics weak.
(c) Spacing alternately 9°.5 and 10°.5. The octave is stronger than the fundamental even though the latter is negligible in the Fourier spectrum.
(d) Spacing alternately 9° and 11°. The fundamental is somewhat louder than the octave, although still very small in the Fourier spectrum.

---

167 years of Fourier analysis model: “Ohm’s acoustic law” 1843, & Helmholtz

- Sensitivity to spectrum, ignoring phase or temporal structure
- Ohm and Helmholtz ignored the time-domain pitch evidence from Seebeck’s siren experiments
- This “frequency domain” approach continues to confuse the field to this day
- FFT and linearity are similarly harmful



My favorite quote

“… **dehydrated cats** and the application of **Fourier analysis** to hearing problems became more and more a handicap for research in hearing.”

— Georg von Békésy

“Some Biophysical Experiments from Fifty Years Ago”
Annual Review of Physiology
Vol. 36: 1-18 (1974)

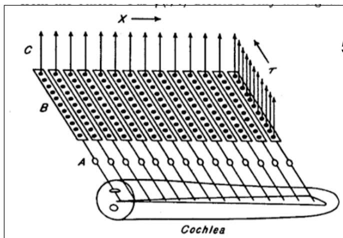


Fig. 2. – *Schematic diagram of overall analyzer.* At the bottom is the uncoiled cochlea. Its lengthwise dimension and the corresponding dimension in the neural tissue above it is designated the x-dimension. The cochlea performs a crude frequency analysis of the stimulus time function, distributing different 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.

59 years of auditory auto-correlogram: Licklider 1951 “duplex theory” unified the frequency and time views via 2D “auditory image”

ON THE IMPORTANCE OF TIME—A TEMPORAL REPRESENTATION OF SOUND

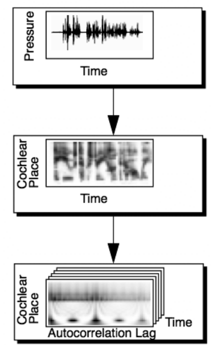


Fig. 1– Three stages of auditory processing are shown here. Sound enters the cochlea and is transduced into what we call a cochleagram (middle picture). A correlogram is then computed from the output of the cochlea by computing short time autocorrelations of each cochlear channel. One frame of the resulting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

Slaney & Lyon 1993

Correlograms are movies; this one is based on Roy Patterson’s “Stabilized Auditory Image” (SAI)

Lloyd A. Jeffress 1948: Binaural cross-correlator

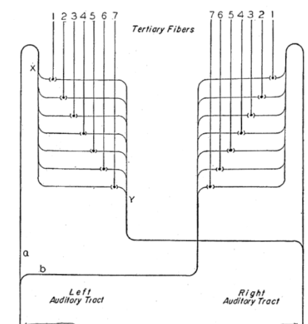
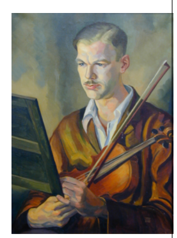
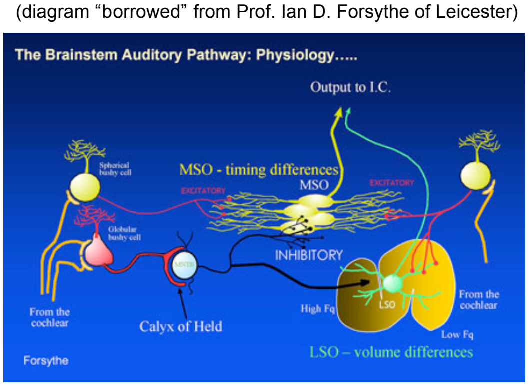


FIGURE 1. HYPOTHETICAL MID-BRAIN MECHANISM FOR THE LOCALIZATION OF LOW FREQUENCY TONES

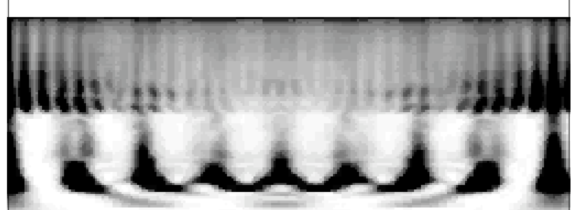


Neural correlators and such

(diagram “borrowed” from Prof. Ian D. Forsythe of Leicester)

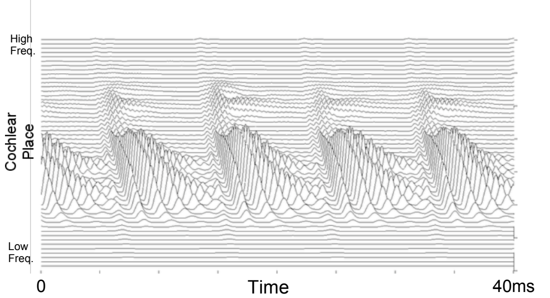


Stabilized Auditory Image – a very good way to visualize sounds such as speech…



Peripheral (Cochlea) Model

Example filterbank output for 40ms of a human vowel /a/



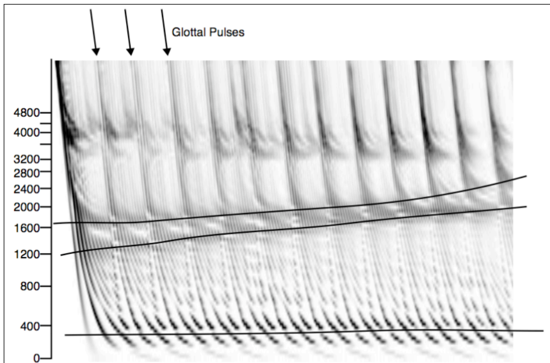
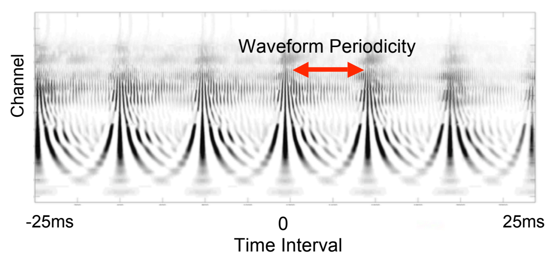


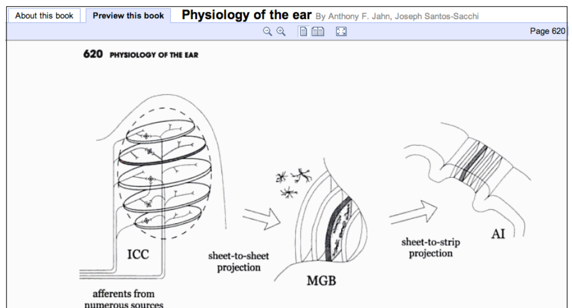
Fig. 10- Expanded cochleagram of the dipthong "rea" in greasy from the Sheffield 'timit.dip' utterance. The first three formant tracks are shown (the lowest formant is excited with two harmonics). The vertical lines, each of which represents a glottal pulse, are tilted slightly due to the natural delay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

Strobed Temporal Integration:

Align to trigger points, like an oscilloscope sync does, and average over time a bit.

Stabilized Auditory Image





620 PHYSIOLOGY OF THE EAR

Figure 26-3. Schematic depiction of the connectivity between the auditory midbrain and the auditory cortex. Note that the inferior colliculus receives input from many sources and that these inputs are neither completely overlapping in their spatial distribution nor completely segregated. Note that in the laminated, ventral division of the MGB, the principal cells have tufted dendrites and are oriented in the isofrequency plane; in the nonlaminated, medial division of the MGB, the cells are often larger and may have stellate dendritic arrangements (after Winer[84]). There is a sheet-to-sheet connectivity between the tonotopic arrays of the ICC and the ventral division of the MGB, but a convergent, sheet-to-strip connectivity between the thalamus and the cortex.

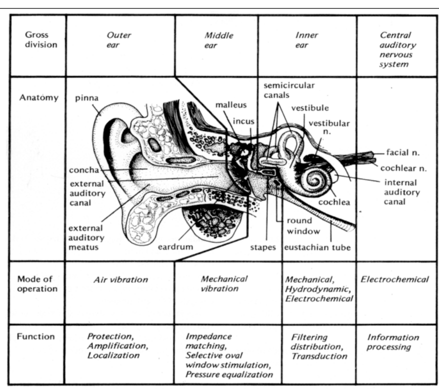
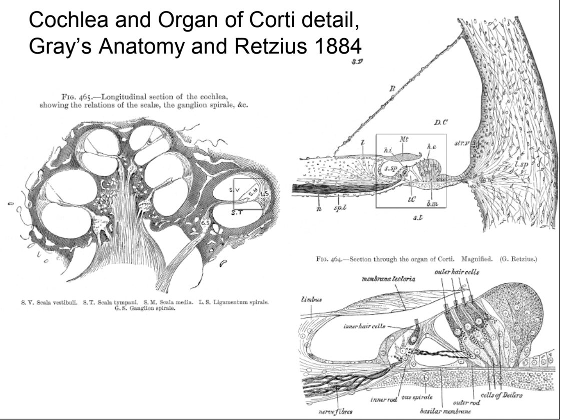


FIGURE 6.1 Cross section of human ear showing divisions of the outer, middle, and inner ears and central auditory nervous system. Below are listed the predominant modes of operation of each division and its suggested function. Adapted with permission from Ades and Engstrom (1974) and Dallos (1973).

Cochlea and Organ of Corti detail,
Gray's Anatomy and Retzius 1884



What does the cochlea tell the brain?

Geisler 1987 – waveforms, not just mag & phase of sinusoids

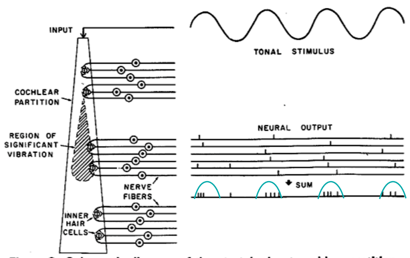
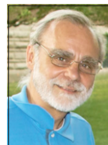


Figure 3. Schematic diagram of the stretched-out cochlear partition, as seen from above. The partition becomes wider and less stiff with distance from the input. The nerve fibers connected to the inner hair cells make up 90 to 95 percent of the auditory nerve's population.



Jont Allen on function:

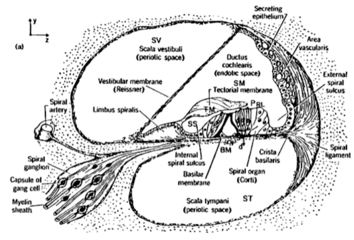
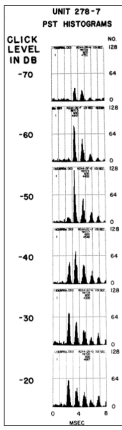


Figure 1: *Cross section through the cochlear duct showing all the major structures of the cochlea. The three chambers are filled with fluid. Reissner's membrane is an electrical barrier and is not believed to play a mechanical role.*

Function of the Inner Ear

The goal of cochlear modeling is to refine our understanding of how auditory signal are processed. Science is at its best as an interplay between experiment and theory. The two main roles of the cochlea are to separate the input acoustic signal into overlapping frequency bands, and to compress the large acoustic intensity range into the much smaller mechanical and electrical dynamic range of the inner hair cell. This is a basic problem of information processing by the ear. The eye plays a similar role as a peripheral organ.



Nelson Kiang 1965 – cat auditory nerve shows phase detail – as expected

Figure 5.1 Responses of a unit with a CF of 0.65 kc to clicks as a function of click level. Stimuli: 10/sec clicks. Reference level was 4 V into dynamic ear-phone 1.

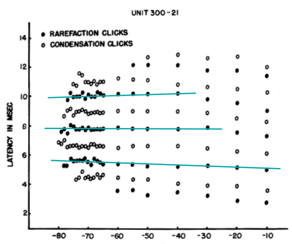
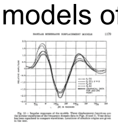


Figure 5.11 Latencies of the peaks in the PST histograms shown in Figures 5.7 and 5.8 plotted against click level. The latency shifts at the higher click levels are the greatest that were encountered in this study.

Jim Flanagan 1962 filter models of basilar membrane



order-3 "gamma-tone" filter

a reasonable fit to Békésy's results is

$$F_1(s) = c_1\beta_1^4 \left(\frac{2000\pi\beta_1}{\beta_1 + 2000\pi}\right)^{0.8} \left(\frac{s+\varepsilon_1}{s+\beta_1}\right) \left[\frac{1}{(s+\alpha_1)^2+\beta_1^2}\right]^2 e^{\frac{-3\pi s}{4\beta_1}}$$

The membrane response at any point is therefore approximated in terms of the poles and zeros of the rational function part of $F_1(s)$. As

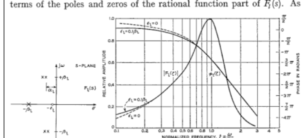


Fig. 4.17a. Pole-zero diagram for the approximating function $F_1(s)$ (after Flanagan, 1962a)

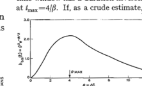
Fig. 4.17b. Amplitude and phase response of the basilar membrane model $F_1(s)$. Frequency is normalized in terms of the characteristic frequency β_1

126 Techniques for Speech Analysis

good approximation to the displacement impulse response of the basilar membrane, at a point maximally responsive to radian frequency β, is

$$p(t) = (\beta t)^3 e^{-\beta t/2} \sin \beta t = h_{\beta m}(t) \sin \beta t. \quad (5.9)$$

The time window for the basilar membrane, according to this modeling, is the "surge" function plotted in Fig. 5.6. One notices that the time window has a duration inversely related to β. It has its maximum at $t_{max} = 4/\beta$. If, as a crude estimate, $2t_{max}$ is taken as the effective duration D of the window, then for several membrane places:



$\beta/2\pi$ (cps)	$D=2t_{max}$ (msec)
100	12.0
1000	1.2
5000	0.3

Fig. 5.6. The effective time window for short-time frequency analysis by the basilar membrane in the human ear. The weighting function is deduced from the ear model discussed in Chapter IV

For most speech signals, therefore, the mechanical analysis of the ear apparently provides better temporal resolution than spectral resolution. Generally, the only harmonic component resolved mechanically is the fundamental frequency of voiced segments. This result is borne out by observations on the models described in Chapter IV.

Bringing mathematical physics to cochlear modeling:

"The Cochlear Compromise," Zweig, Lipes, & Pierce 1976

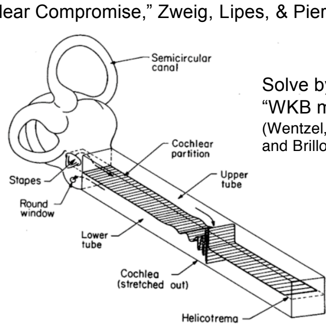
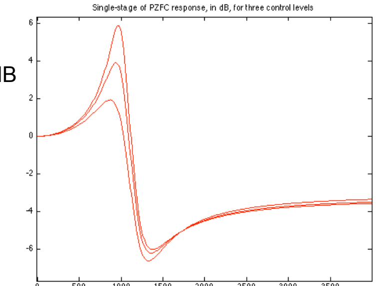


Figure 1. Schematic sketch of the cochlea, stretched out from its actual spiral form, catching a mid-frequency traveling wave at a particular instant. Adapted from Zweig et al. [2].

Solve by "WKB method" (Wentzel, Kramers, and Brillouin)

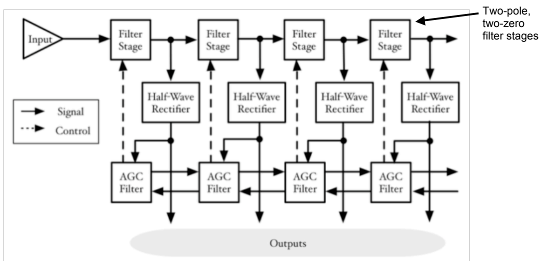
PZFC stage frequency response: **pole** pair makes a bump (variable via pole Q),and **zero** pair makes a dip

dB

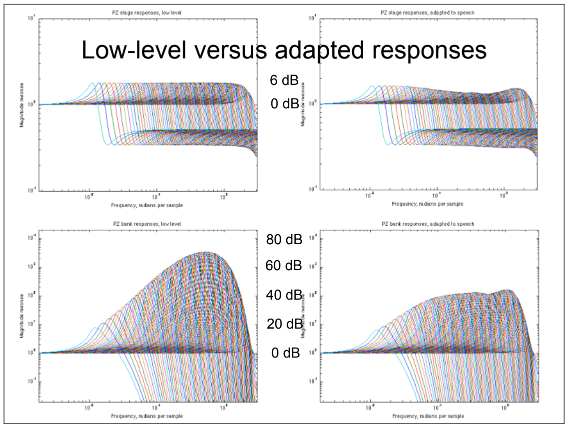


PZFC Cochlea Model

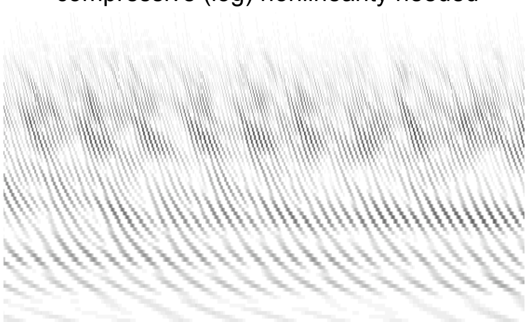
Pole–Zero Filter Cascade: Cascade model of cochlear filtering

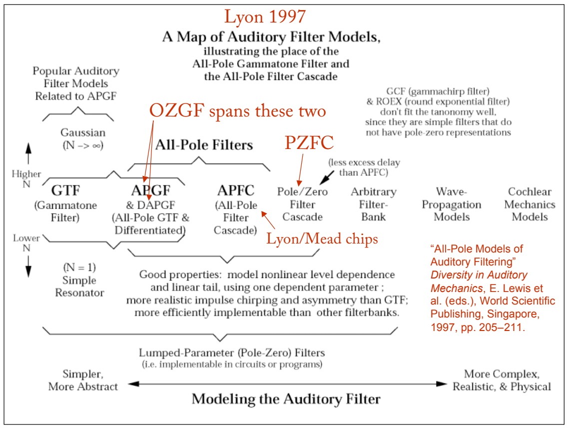


Low-level versus adapted responses



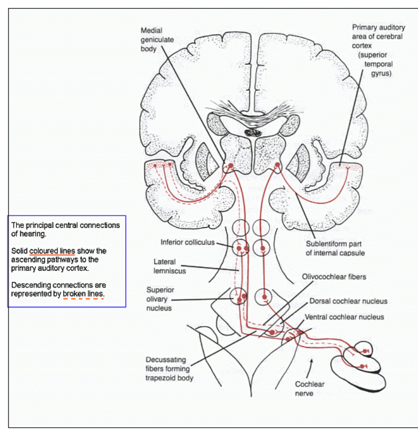
Response to sounds – no further compressive (log) nonlinearity needed





Important nonlinearities (linear filters are not enough)

1. Automatic (adaptive) Gain Control (AGC) via filter parameter variation at all time scales
2. Half-wave detection nonlinearity (needed to unify all types of pitch via autocorrelation)
3. Distortion, especially third-order compressive nonlinearity in travelling wave or at filter outputs
4. Coincidence detection (AND gate, or multiplier) for monaural and binaural correlation analysis



Afferent and efferent auditory pathways

(Stuart Rosen et al.)

Iso-frequency unit: afferent and efferent innervation of inner and outer hair cells in the cochlea

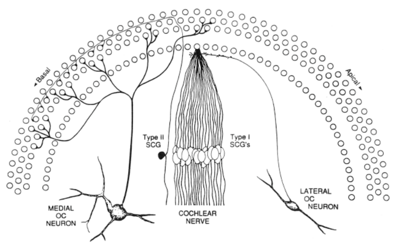


FIGURE 7.12. A hypothetical isofrequency unit of afferent and efferent innervation from the middle of the cochlea. The unit is comprised of two types of afferent and two types of efferent fibers tuned to the same frequency. Note the basal displacement and partial spatial overlap of MOC efferent and type-II afferent innervation patterns on the outer hair cells relative to the corresponding inner hair cell and lateral efferent neuron. (See text for further details.)

Geisler 1974 crossed olivo-cochlear bundle (COCB) stimulation reduces gain from SPL to neural signals

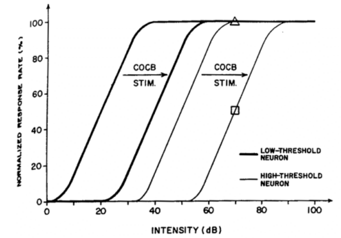


FIG. 1. Idealized curves of discharge rate as a function of intensity for two auditory nerve neurons discharging to best frequency stimuli. The two neurons represent the most and least sensitive neurons having that best frequency. With maximum COCB stimulation, each curve is shifted right by 20 dB, to cover a higher range of intensities. Ordinate: discharge rate expressed as a percentage of each neuron's dynamic range (maximum discharge rate less spontaneous rate).

Sharp Tuning Curves from Unsharp Filters and Adaptive (AGC) Nonlinearity

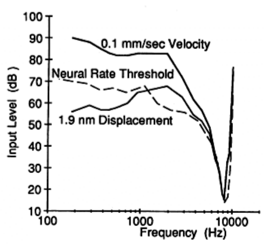


Figure 5

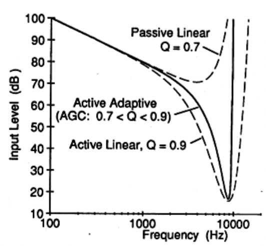


Figure 6

Why all this complicated stuff?

- Goal is for ***simple*** and ***effective*** structures that preserve all psychophysically relevant features of sounds, including mixtures.
- Pole-zero cascade with coupled AGC via variable pole Q is simple and effective; for binaural, too.
- Follow by temporal periodicity detection (autocorr or S.A.I.) and then do CASA, streaming, or whatever…

4-part Machine Hearing System Structure

- Peripheral model (PZFC with HWR/AGC)
- "Auditory image" generation
  - Patterson's SAI variants or correlograms
  - Binaural cross-correlogram
  - Chromagram for melody
  - …
- Feature extraction (cortical simple and complex…)
  - High-dim sparse features
  - Low-dim dense features
  - "Key point" features, multi-scale features, etc.
- Machine-learning back end trained for application