Speech Perception

Speech Perception

- Understanding ***how we hear sounds and how we perceive speech → better design and*** **implementation** of robust and efficient systems for analyzing and representing speech

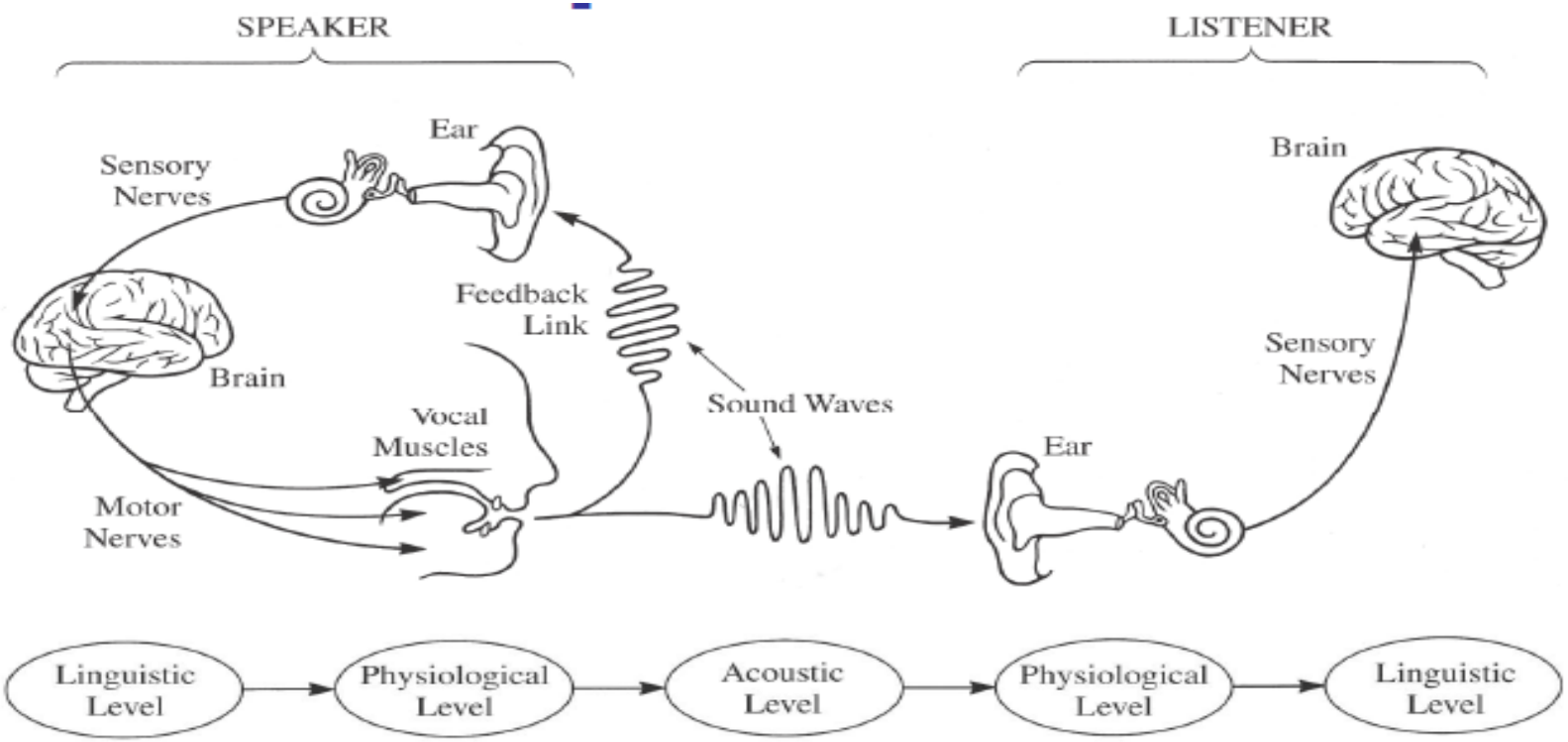
the better we understand signal processing in the human auditory system, the better we can (at least in theory) design practical speech processing systems like:

– speech coding

– speech recognition

- Try to understand speech perception by looking at the ***physiological models of hearing***

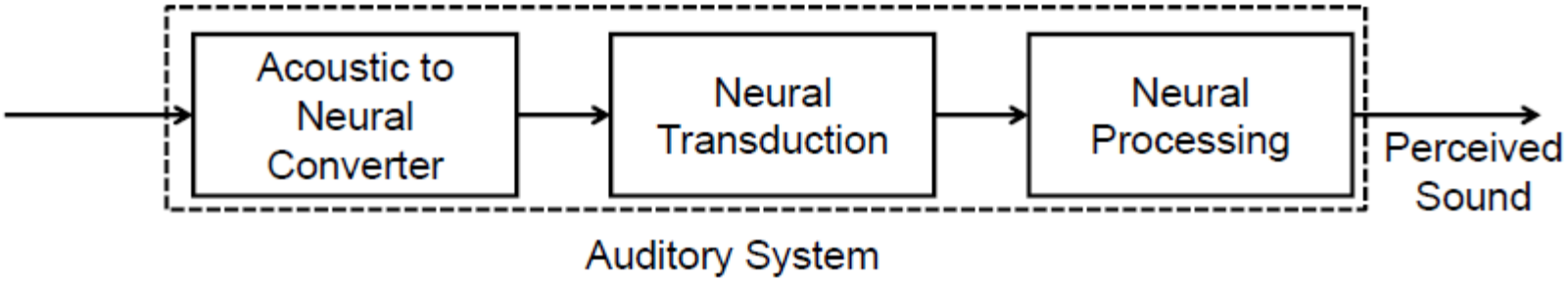
The Speech Chain



The Speech Chain comprises the processes of:

- ❑ Speech production,
- ❑ Auditory feedback to the speaker,
- ❑ Speech transmission (through air or over an electronic)
- ❑ Communication system (to the listener), and
- ❑ Speech perception and understanding by the listener.

The Auditory System



❑ The acoustic signal first converted to a neural representation by processing in the ear

- The conversion takes place in stages at the outer, middle and inner ear
- These processes can be measured and quantified

❑ The neural transduction step takes place between the output of the inner ear and the neural pathways to the brain

- Consists of a statistical process of nerve firings at the hair cells of the inner ear, which are transmitted along the auditory nerve to the brain
- Much remains to be learned about this process

❑ The nerve firing signals along the auditory nerve are processed by the brain to create the perceived sound corresponding to the spoken utterance

- **These processes not yet understood**

The Black Box Model of the Auditory System

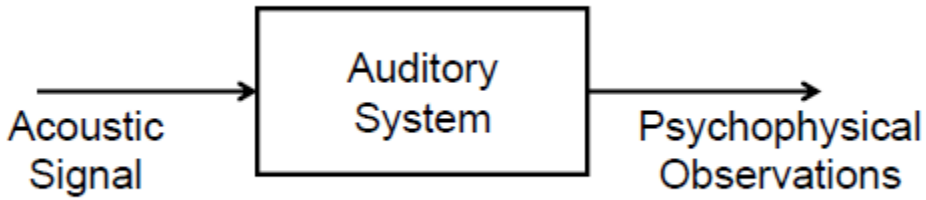
Researchers have resorted to a "black box" behavioral model of hearing and perception

I. Model assumes that an acoustic signal enters the auditory system causing behavior that we record as psychophysical observations

II. Psychophysical methods and sound perception experiments determine how the brain processes signals with different loudness levels, different spectral characteristics, and different temporal properties

III. Characteristics of the physical sound are varied in a systematic manner and the psychophysical observations of the human listener are recorded and correlated with the physical attributes of the incoming sound

IV. Then determine how various attributes of sound (or speech) are processed by the auditory system



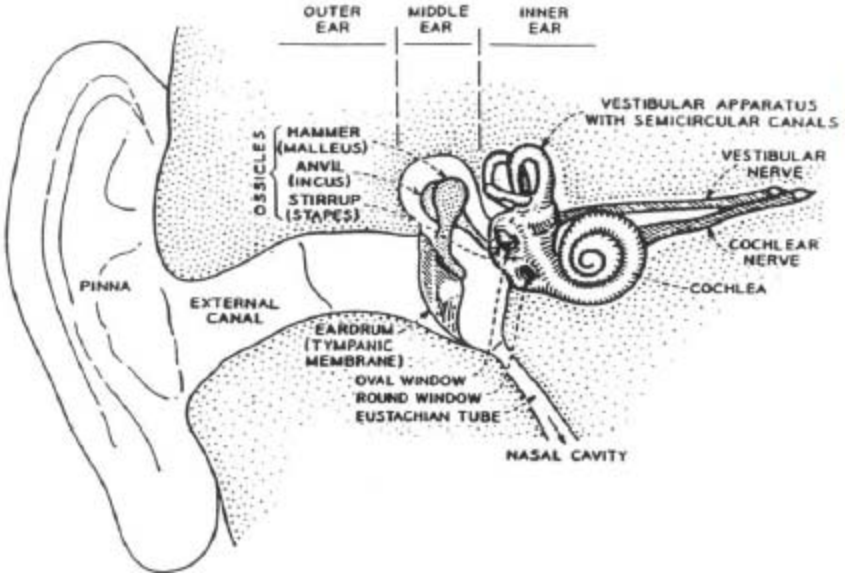
Why Do We Have Two Ears

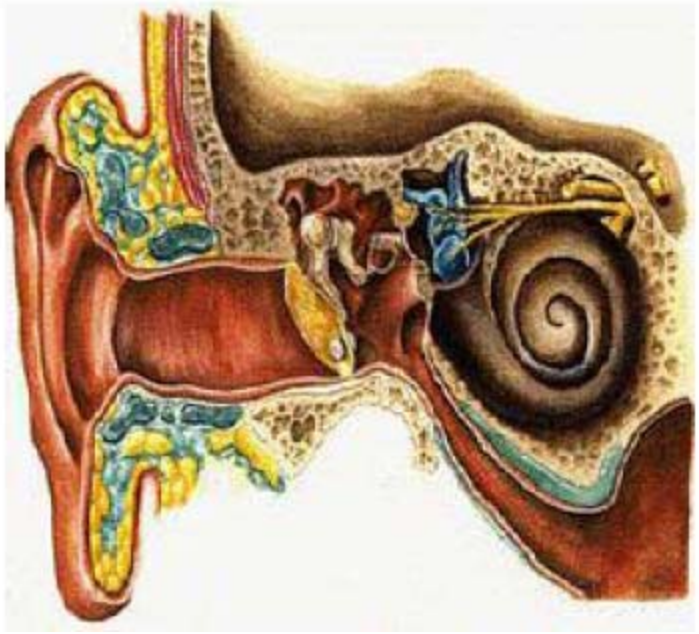
- ***Sound localization – spatially locate*** sound sources in 3-dimensional sound fields

- ***Sound cancellation – focus attention on*** a 'selected' sound source in an array of sound sources – 'cocktail party effect'

- Effect of ***listening over headphones* => ** localize sounds inside the head (rather than spatially outside the head)

The Human Ear



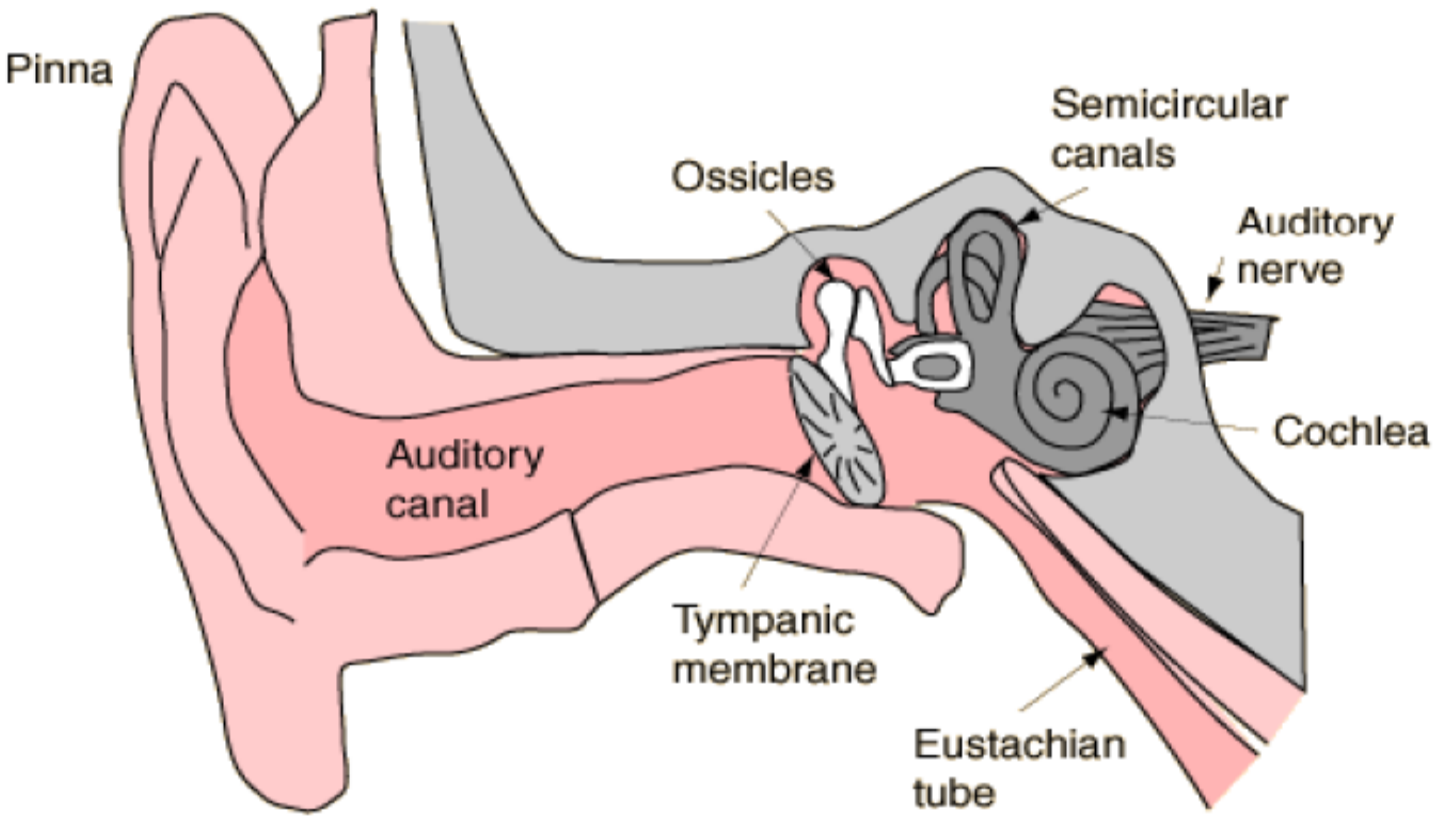


Outer ear: pinna and external canal

Middle ear: tympanic membrane or eardrum

Inner ear: cochlea, neural connections

Ear and Hearing



Human Ear

Outer ear: funnels sound into ear canal

Middle ear: sound impinges on tympanic membrane; this causes motion

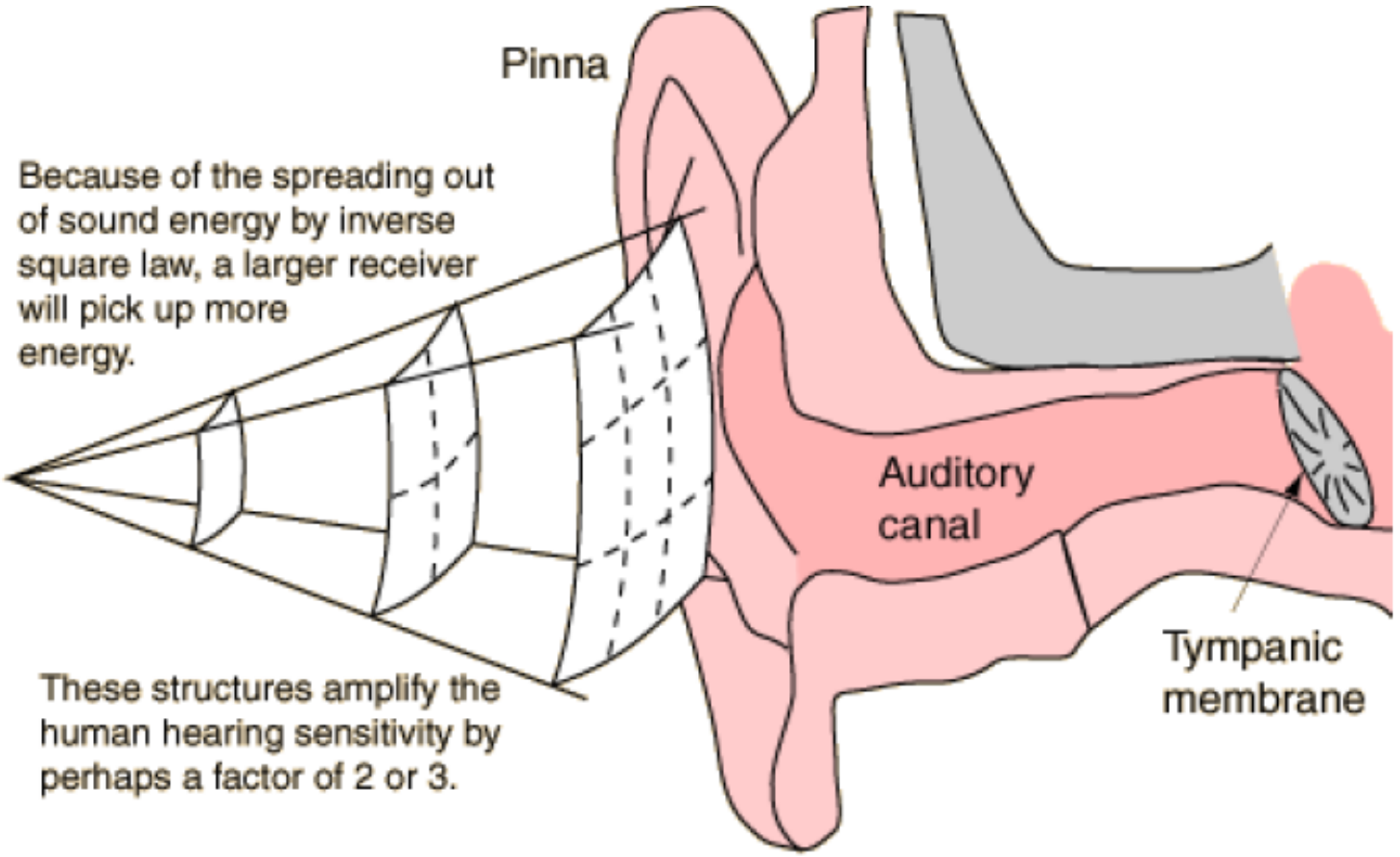
• Middle ear is a mechanical transducer, consisting of the hammer, anvil and stirrup; it converts acoustical sound wave to mechanical vibrations along the inner ear

Inner ear: the cochlea is a fluid-filled chamber partitioned by the basilar membrane

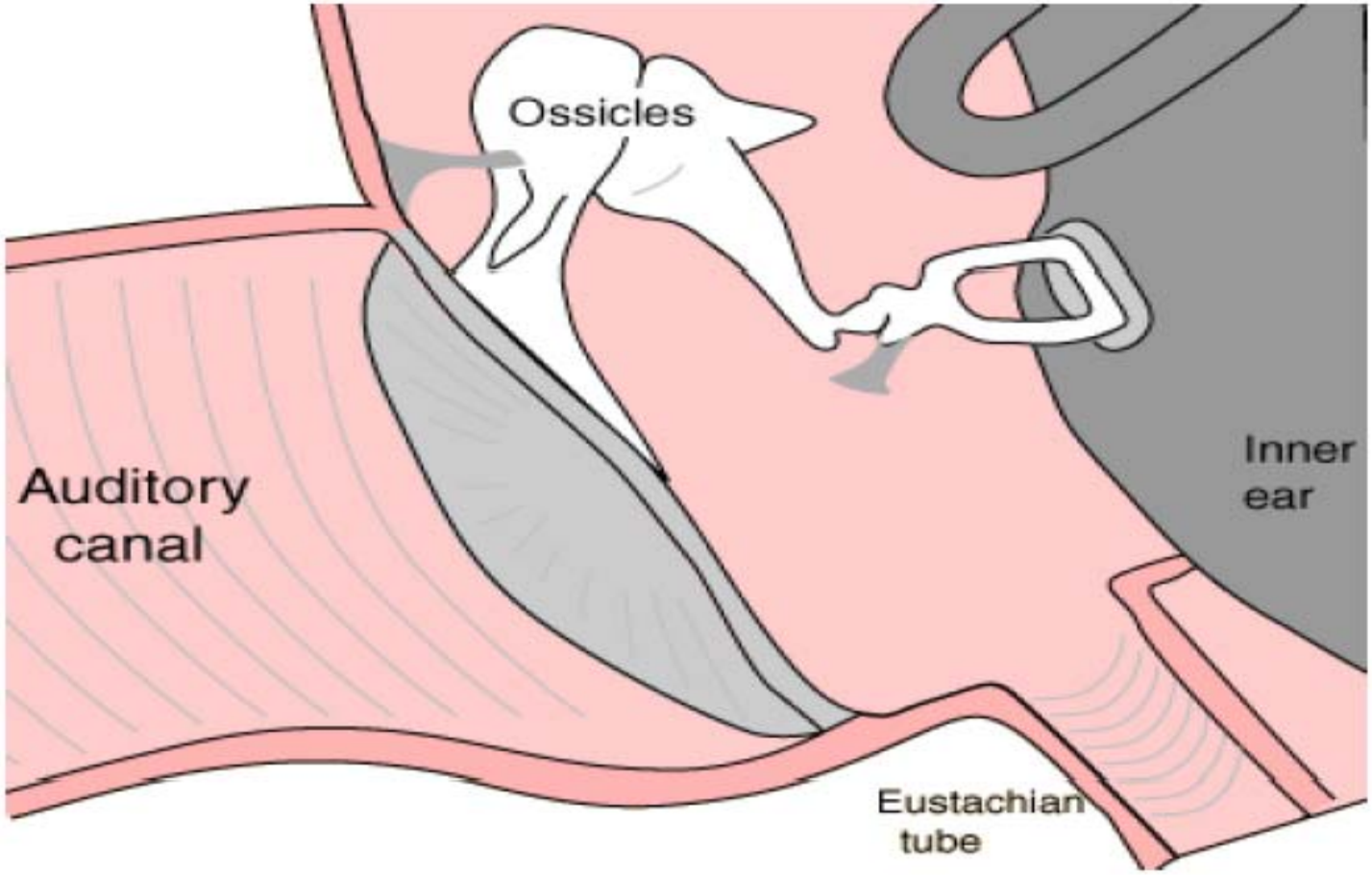
• The auditory nerve is connected to the basilar membrane via inner hair cells

• Mechanical vibrations at the entrance to the cochlea create standing waves (of fluid inside the cochlea) causing basilar membrane to vibrate at frequencies commensurate with the input acoustic wave frequencies (formants) and at a place along the basilar membrane that is associated with these frequencies

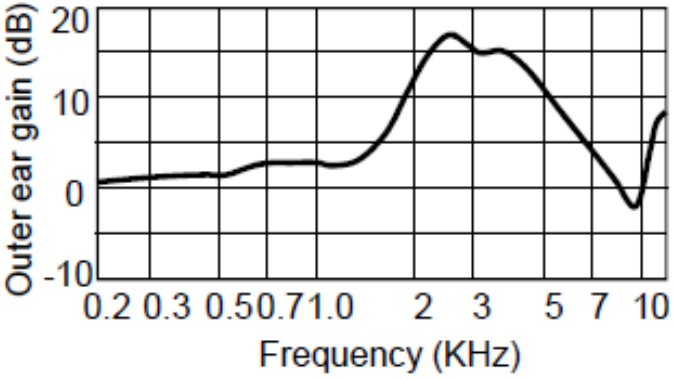
The Outer Ear

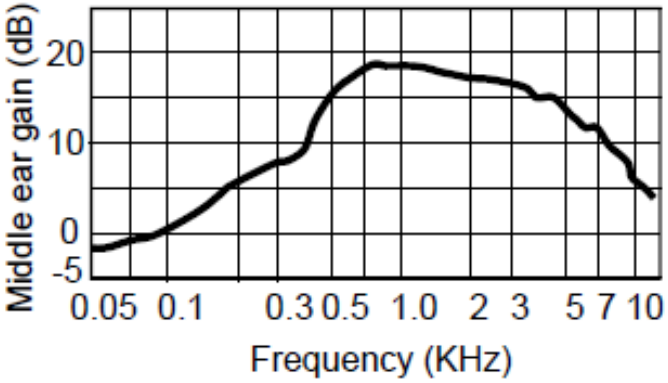


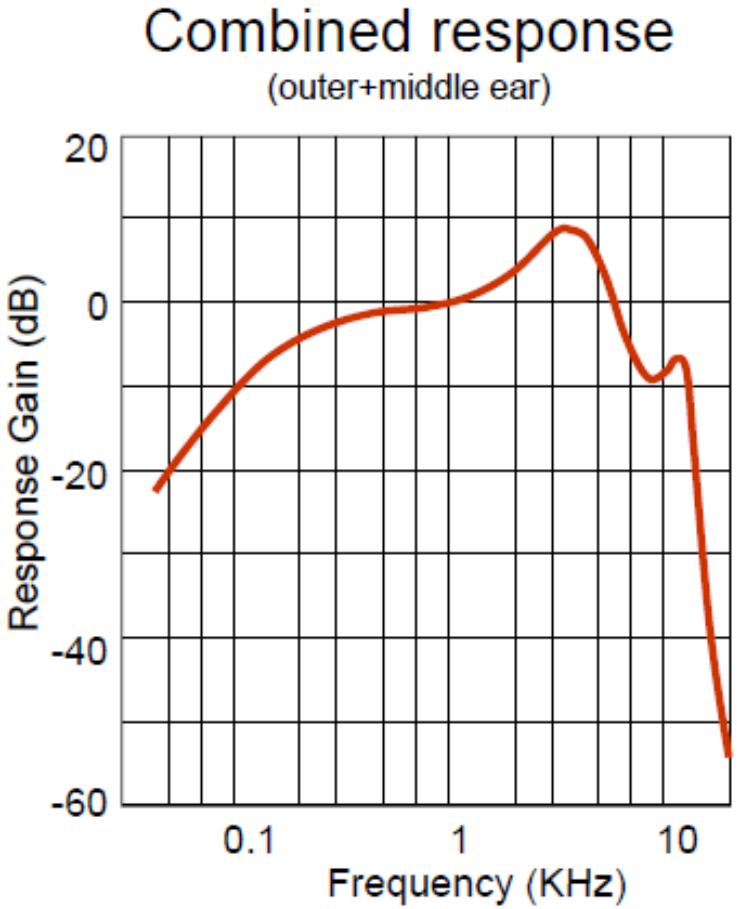
The Outer Ear



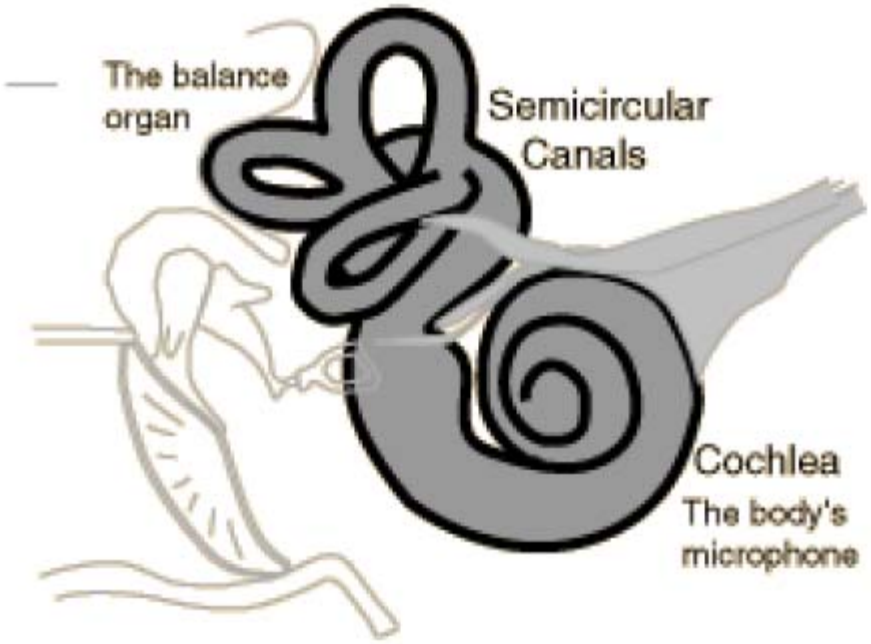
Transfer Functions at the Periphery







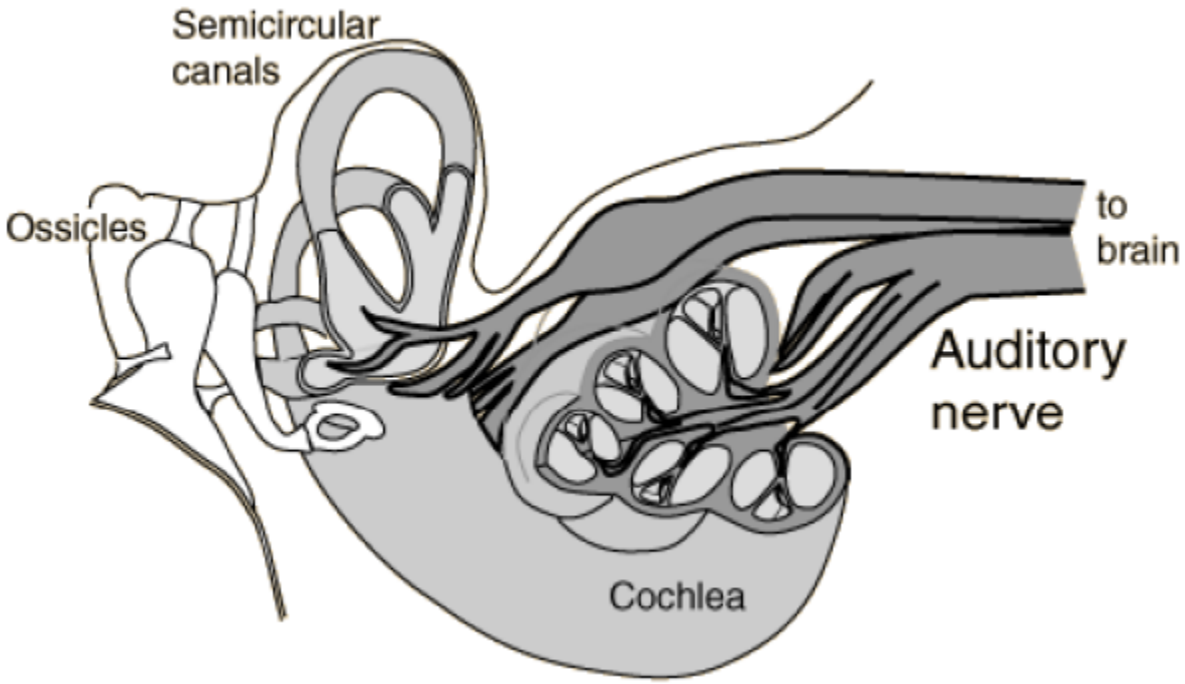
The Inner Ear



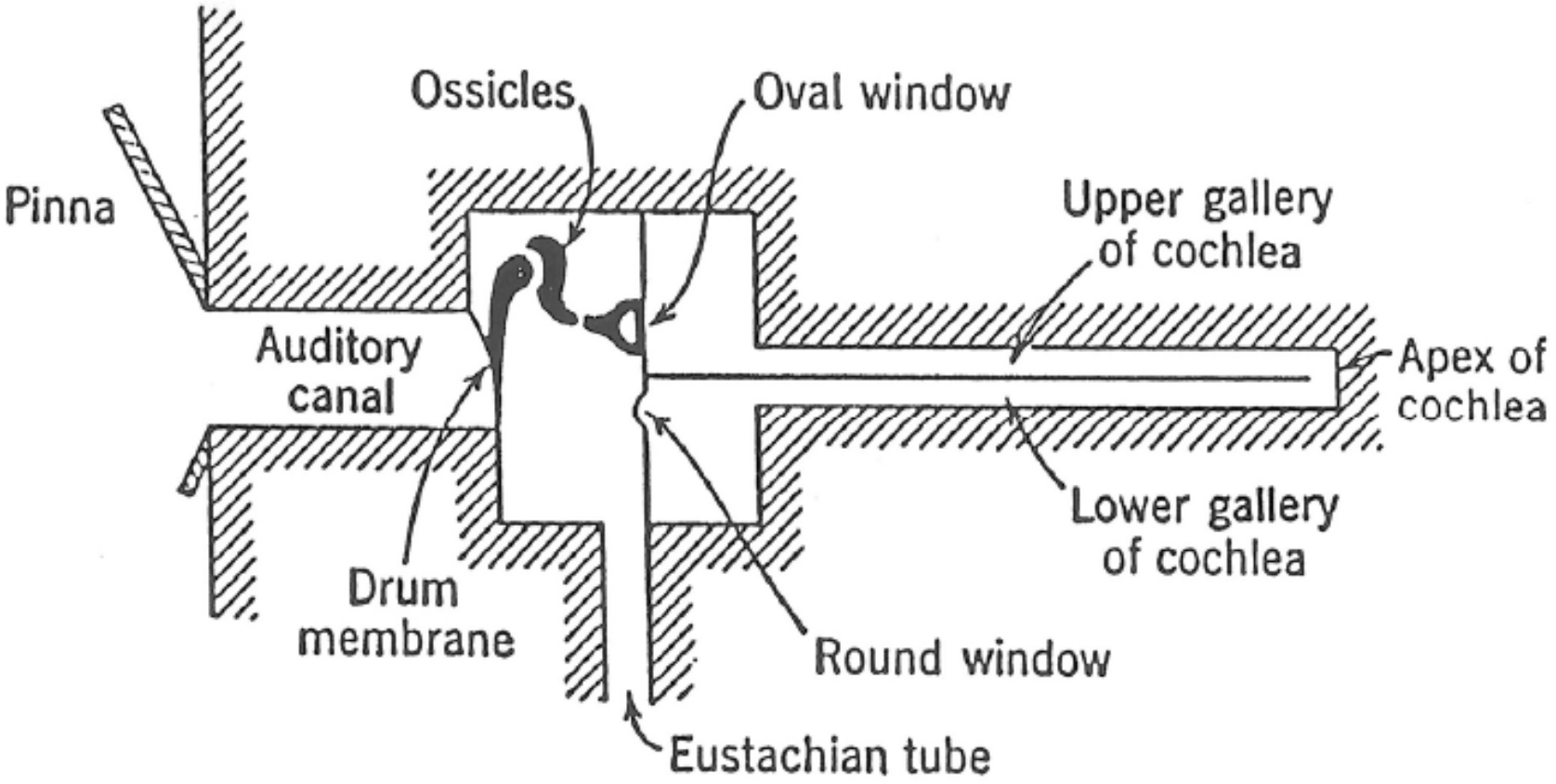
The inner ear can be thought of as two organs, namely the **semicircular canals** which serve as the body's balance organ and the **cochlea** which serves as the body's microphone, converting sound pressure signals from the outer ear into electrical impulses which are passed on to the brain via the auditory nerve.

The Auditory Nerve

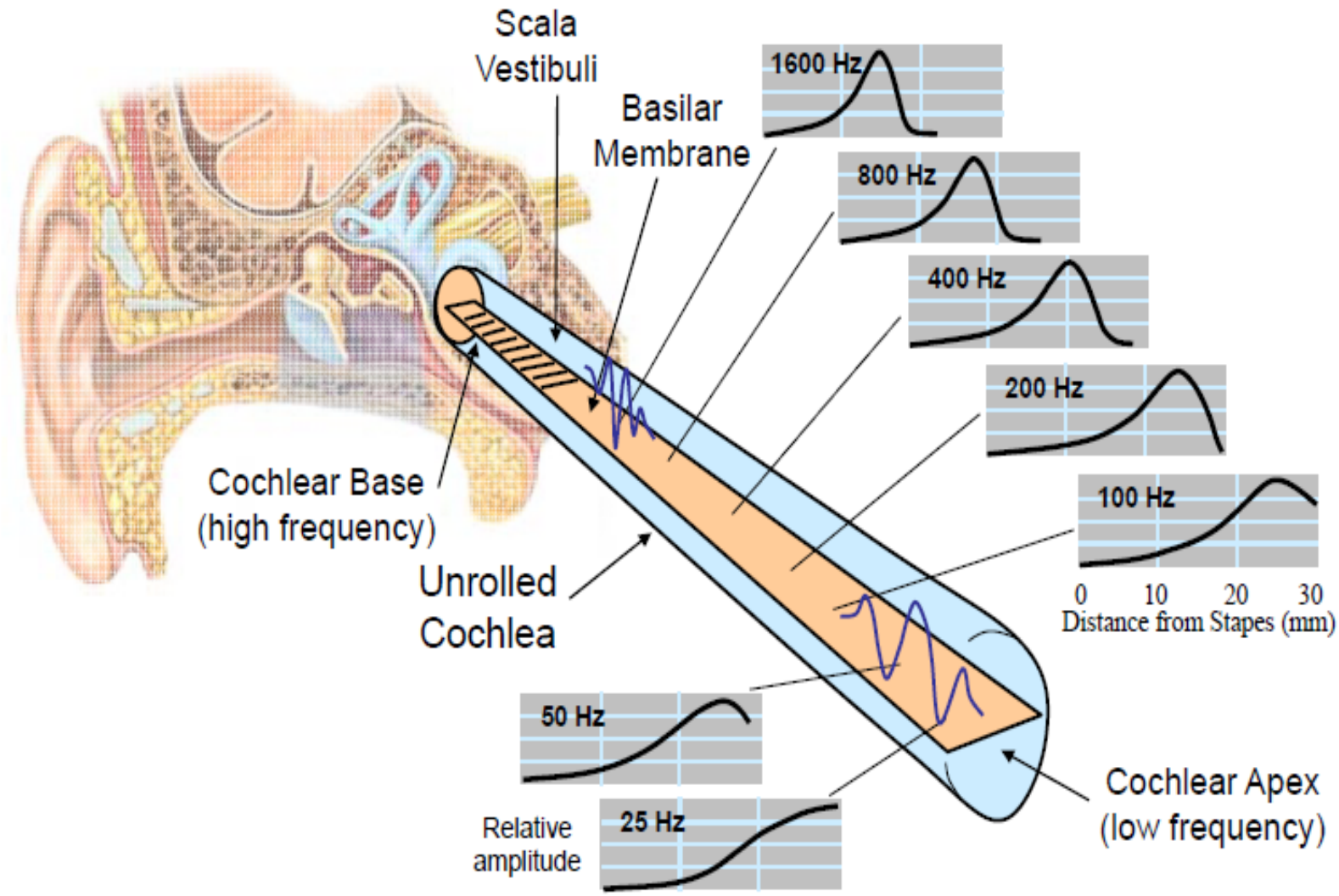
Taking electrical impulses from the cochlea and the semicircular canals, the auditory nerve makes connections with both auditory areas of the brain.



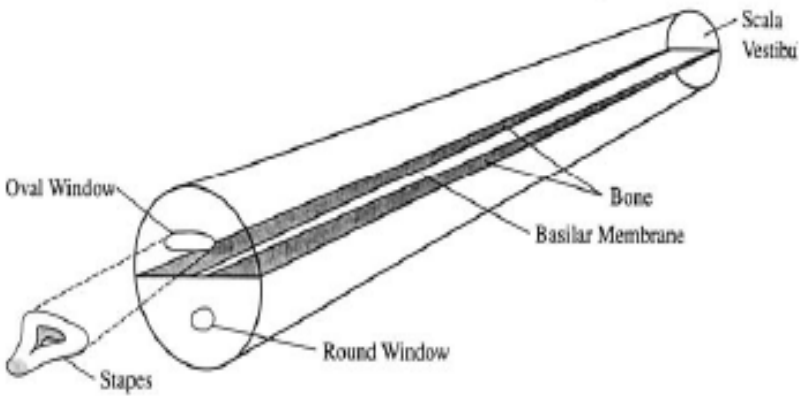
Schematic Representation of the Ear



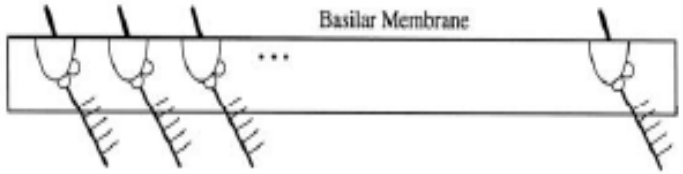
Stretched Cochlea & Basilar Membrane



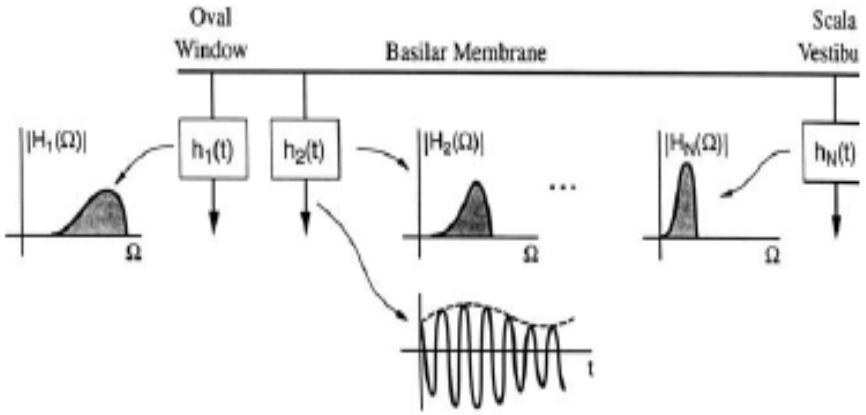
Basilar Membrane Mechanics



(a)

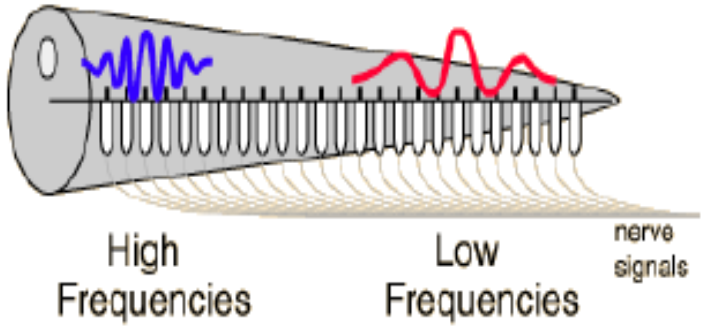


(b)



(c)

Figure 8.25 Schematic of front-end auditory processing and its model as a wavelet transform: (a) uncoiled cochlea; (b) the transduction to neural firings of the deflection of hairs that protrude from the in hair cells along the basilar membrane; (c) a signal processing abstraction of the cochlear filters along basilar membrane. The filter tuning curves, i.e., frequency responses, are roughly constant-Q with bandwi decreasing logarithmically from the oval window to the scala vestibuli.



Basilar Membrane Mechanics

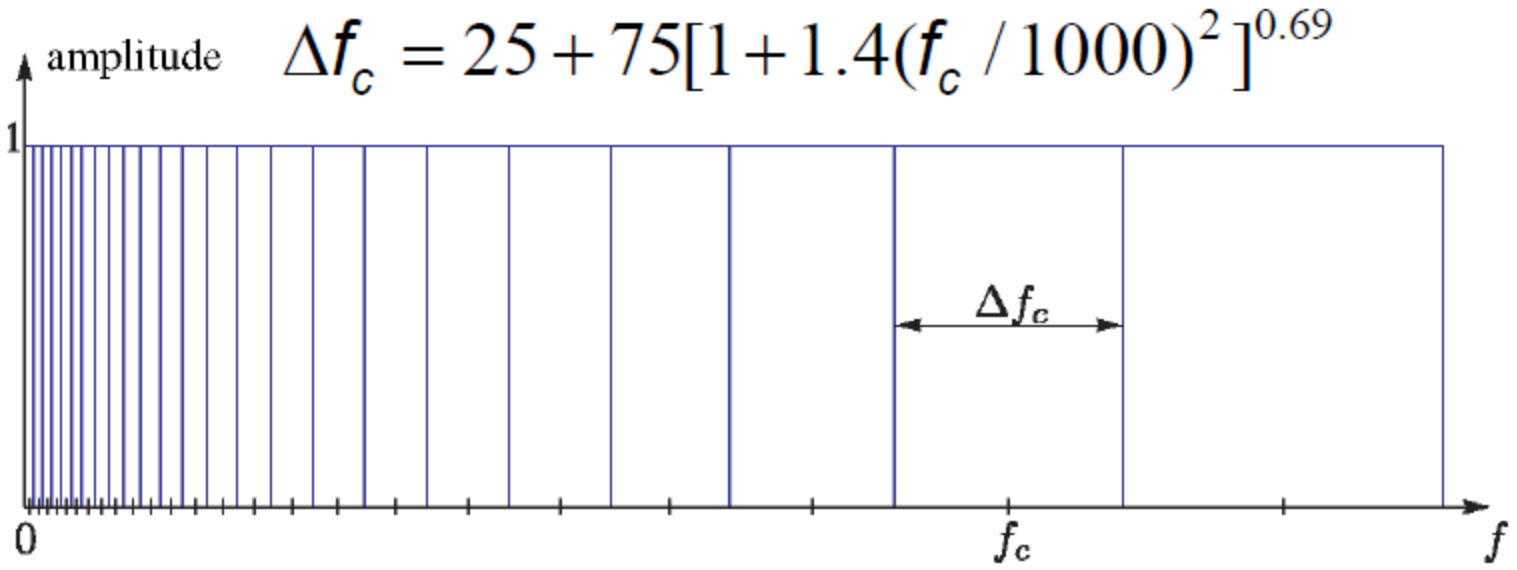
- Characterized by a set of ***frequency responses at different points*** along the membrane
- Mechanical realization of a ***bank of filters***
- Filters are roughly ***constant Q (center frequency/bandwidth) with*** logarithmically decreasing bandwidth
- Distributed along the Basilar Membrane is a set of sensors called ***Inner Hair Cells (IHC) which act as mechanical motion-to-neural*** activity converters
- Mechanical motion along the BM is sensed by local IHC causing ***firing activity at nerve fibers that innervate bottom of each IHC***
- Each IHC connected to about 10 ***nerve fibers, each of different*** diameter => thin fibers fire at high motion levels, thick fibers fire at lower motion levels
- 30,000 nerve fibers link IHC to ***auditory nerve***
- Electrical pulses run along auditory nerve, ultimately reach higher levels of auditory processing in brain, perceived as ***sound***

Basilar Membrane Motion

The ear is excited by the input acoustic wave which has the spectral properties of the speech being produced

- Different regions of the BM respond maximally to different input frequencies => frequency tuning occurs along BM
- The BM acts like a bank of non-uniform cochlear filters
- Roughly logarithmic increase in BW of filters (<800 Hz has equal BW) => constant Q filters with BW decreasing as we move away from cochlear opening
- Peak frequency at which maximum response occurs along the BM is called the characteristic frequency

Critical Bands



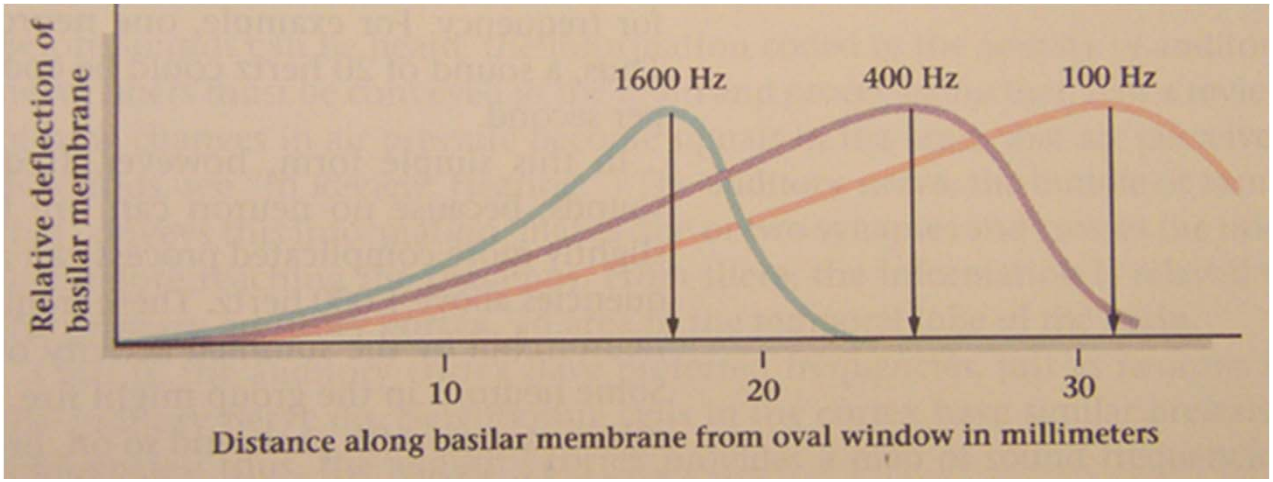
Real BM filters overlap significantly

Frequency (temporal) Theory

- The basis for the *temporal theory* of pitch perception is the timing of neural firings, which occur in response to vibrations on the basilar membrane.
- Periodic stimulation of membrane matches frequency of sound
  - one electrical impulse at every peak
  - maps time differences of pulses to pitch

Place Theory

- Waves move down basilar membrane
  - →stimulation increases, peaks, and quickly tapers
  - → location of peak depends on frequency of the sound, lower frequencies being further away



The Perception of Sound

Key questions about sound perception:

- What is the `resolving power' of the hearing mechanism

- How good an estimate of the fundamental frequency of a sound do we need so that the perception mechanism basically can't tell the difference

- How good an estimate of the resonances or formants (both center frequency and bandwidth) of a sound do we need so that when we synthesize the sound, the listener can't tell the difference

- How good an estimate of the intensity of a sound do we need so that when we synthesize it, the level appears to be correct

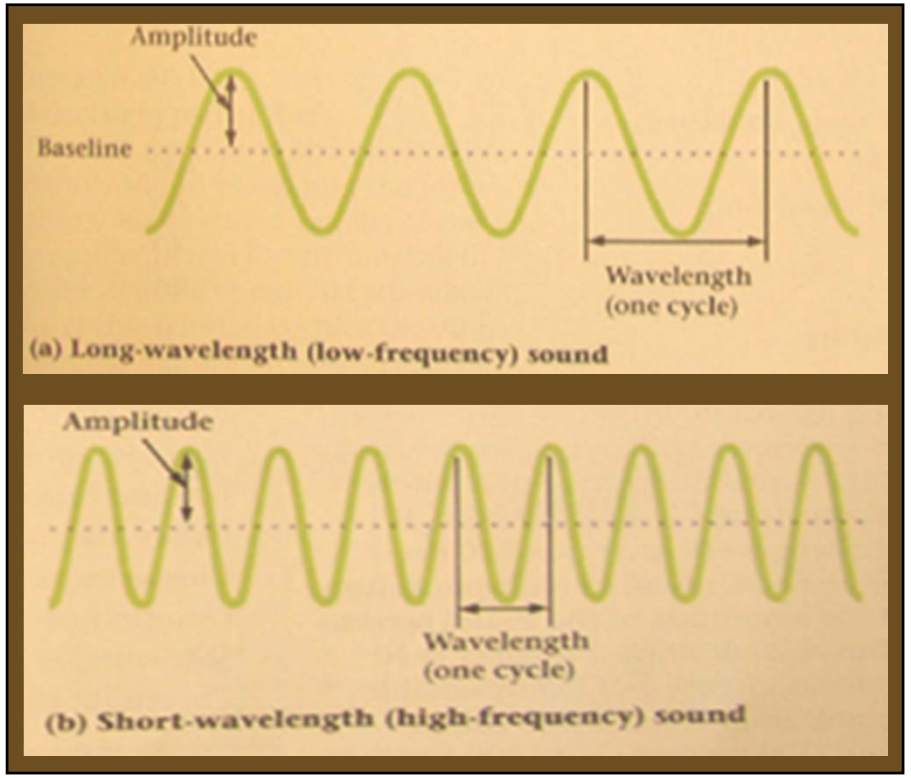
Parameter Discrimination

JND – Just Noticeable Difference
Similar names: differential limen (DL), …

Parameter	JND/DL
Fundamental Frequency	0.3-0.5%
Formant Frequency	3-5%
Formant bandwidth	20-40%
Overall Intensity	1.5 dB

Physical Dimensions of Sound

- Amplitude
  - height of a cycle
  - relates to loudness
- Wavelength (*w*)
  - distance between peaks
- Frequency (λ)
  - cycles per second
  - relates to pitch
  - λ *w* = velocity
- Most sounds mix many frequencies & amplitudes



Sound is repetitive changes in air pressure over time

Auditory Perception

Auditory perception is a branch of ***psychophysics.***

Psychophysics studies relationships between ***perception*** and ***physical properties*** of stimuli.

Physical dimensions: Aspects of a physical stimulus that can be measured with an instrument (e.g., a light meter, a sound level meter, a spectrum analyzer, a fundamental frequency meter, etc.)

Perceptual dimensions: These are the ***mental experiences*** that occur inside the mind of the observer. These experiences are actively created by the sensory system and brain based on an analysis of the physical properties of the stimulus. Perceptual dimensions can be measured, but not with a meter. Measuring perceptual dimensions requires an observer (e.g., a listener).

Visual Psychophysics:

Perceptual Dimensions	Physical Properties of Light
Hue	Wavelength
Brightness	Luminance
Shape	Contour/Contrast

Auditory Psychophysics:

Perceptual Dimensions	Physical Properties of Sound
Pitch	Fundamental Frequency
Loudness	Intensity
Timbre (sound quality)	Spectrum Envelope/Amp Env

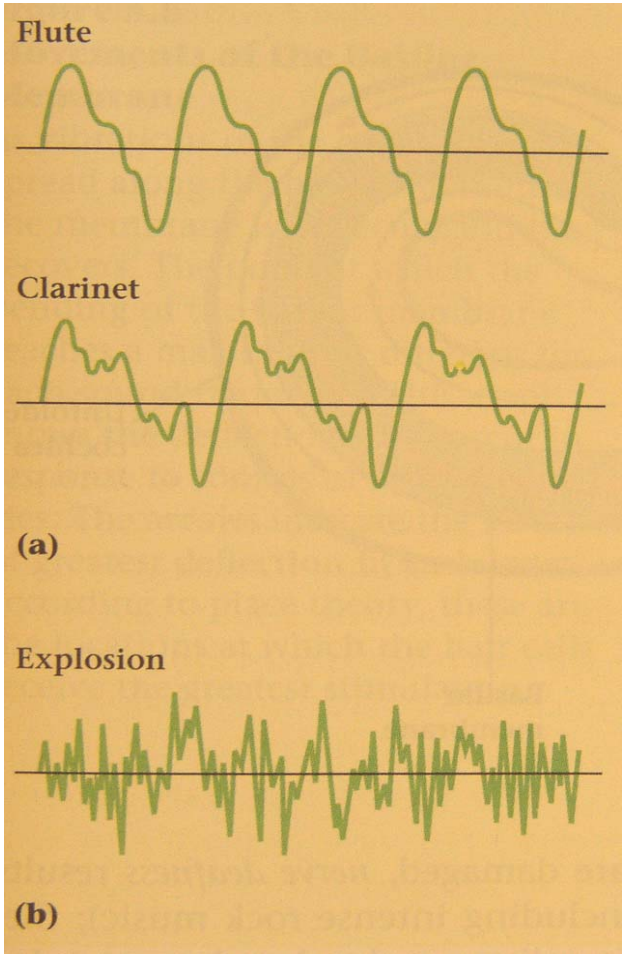
The terms ***pitch***, ***loudness***, and ***timbre*** refer not to the physical characteristics of sound, but to the mental experiences that occur in the minds of listeners.

Perceptual Dimensions

- Pitch
  - higher frequencies perceived as higher pitch
  - humans hear sounds in 20 Hz to 20,000 Hz range
- Loudness
  - higher amplitude results in louder sounds
  - measured in decibels (db), 0 db represents hearing threshold

Perceptual Dimensions (cont.)

- Timbre
  - complex patterns added to the lowest, or *fundamental*, frequency of a sound, referred to as *spectrum envelope*
  - spectrum envelopes enable us to distinguish musical instruments
- Multiples of fundamental frequency give music
- Multiples of unrelated frequencies give noise



The Range of Human Hearing

❑ Human hearing perceives both ***sound frequency*** and ***sound direction***

❑ ***Threshold of hearing*** — Thermal limit of Brownian motion of air particles in the inner ear

The acoustic intensity level of a pure tone that can barely be heard at a particular frequency is called ***Threshold of Audibility***

- *threshold of audibility ≈ 0 dB at 1000 Hz*
- *threshold of feeling ≈ 120 dB*
- *threshold of pain ≈ 140 dB*
- *immediate damage ≈ 160 dB*

Thresholds vary with frequency and from person-to-person

❑ Masking is the phenomenon whereby one loud sound makes another softer sound inaudible

- masking is most effective for frequencies around the masker frequency

Sound Intensity

- Intensity of a sound is a physical quantity that can be measured and quantified
- Acoustic Intensity (I) defined as the average flow of energy (power) through a unit area, measured in watts/square meter
- Range of intensities between 10^{-12} watts/square meter to 10 watts/square meter; this corresponds to the range from the threshold of hearing to the threshold of pain

Threshold of hearing defined to be:

$I_0 = 10^{-12}$ watts/m^2

The intensity level of a sound, IL is defined relative to I_0 as:

$$IL = 10 \log_{10}\left(\frac{I}{I_0}\right) \text{ in dB}$$

For a pure sinusoidal sound wave of amplitude P, the intensity is proportional to P^2 and the sound pressure level (SPL) is defined as:

$$SPL = 10 \log_{10}\left(\frac{P^2}{P_0^2}\right) = 20 \log_{10}\left(\frac{P}{P_0}\right) \text{ dB}$$

where $P_0 = 2 x 10^{-5}$ Newtons/m^2

Some Facts About Human Hearing

- the ***range of human hearing*** is incredible
  - ***threshold of hearing*** — thermal limit of Brownian motion of air particles in the inner ear
  - ***threshold of pain*** — intensities of from 10**12 to 10**16 greater than the threshold of hearing
- human hearing perceives both ***sound frequency*** and ***sound direction***
  - can detect weak spectral components in strong broadband noise
- ***masking*** is the phenomenon whereby one loud sound makes another softer sound inaudible
  - masking is most effective for frequencies around the masker frequency
  - masking is used to hide quantizer noise by methods of spectral shaping (similar grossly to Dolby noise reduction methods)

Sound Intensity

Intensity of a sound is a physical quantity that can be measured and quantified

Acoustic Intensity (*I) defined as the average flow of energy (power)* through a unit area, measured in watts/square meter

Threshold of hearing defined to be: $I_0 = 10^{-12}$ *watts/m*^{2}

The intensity level of a sound I_L is defined relative to I_0

$$I_L = 10 \log_{10} I/I_0$$

Why the decibel?

- Ears judge loudness on a logarithmic vice linear scale
- Alexander Graham Bell
- deci = $\frac{1}{10}$
- 1 bel = 10 decibel

$$\text{"bel"} \equiv \log \frac{I}{I_{ref}}$$

$$IL(\text{in dB}) = 10 \log \left(\frac{\langle I \rangle}{I_{ref}} \right)$$

Reference Level Conventions

$$I_{ref} = \frac{p_{ref}^2}{\rho_o c}$$

Location	**Reference Intensity**	**Reference Pressure**
Air	1×10^{-12} W/m^2	20 μPa
Water	6.67×10^{-19} W/m^2	1 uPa

Sound Pressure Level

Mean Squared Quantities:
Power, Energy, **Intensity**

$$IL = 10 \log \left(\frac{\langle I \rangle}{I_{ref}} \right)$$

"Intensity Level"

Root Mean Squared Quantities:
Voltage, Current, **Pressure**

$$SPL = 20 \log \left(\frac{\sqrt{\langle p^2 \rangle}}{p_{ref}} \right) = 20 \log \frac{p_{rms}}{p_{ref}}$$

"Sound Pressure Level"

Decibels and Percentages

% change= $100\left(1-10^{\frac{-dB}{10}}\right)$

% change= $100\left(10^{\frac{+dB}{10}}-1\right)$

$$dB = M\log_b\left(1-\frac{\%change}{100}\right)$$ Below Reference

$$dB = M\log_b\left(\frac{\%change}{100}+1\right)$$ Above Reference

1. An anechoic chamber absorbs 99% of the power and reflect only 1%. What percent of the initial sound pressure level(SPL) is reflected

W_i =100 W

Reflected power=1W

$$10\log\frac{100}{1}=20dB=$$ 99% power absorbed

% change= $$100\left(1-10^{\frac{-20}{20}}\right)=90\%$$

So reflected SPL=100-90=10%

2. A sound system gain is increased by 15dB. What is the % power increase

% change= $$100\left(10^{\frac{15}{10}}-1\right)=3062\%$$

3. An acoustic signal is reflected off of a surface that is 80% absorptive, the reflected signal will be drop by how many dB?

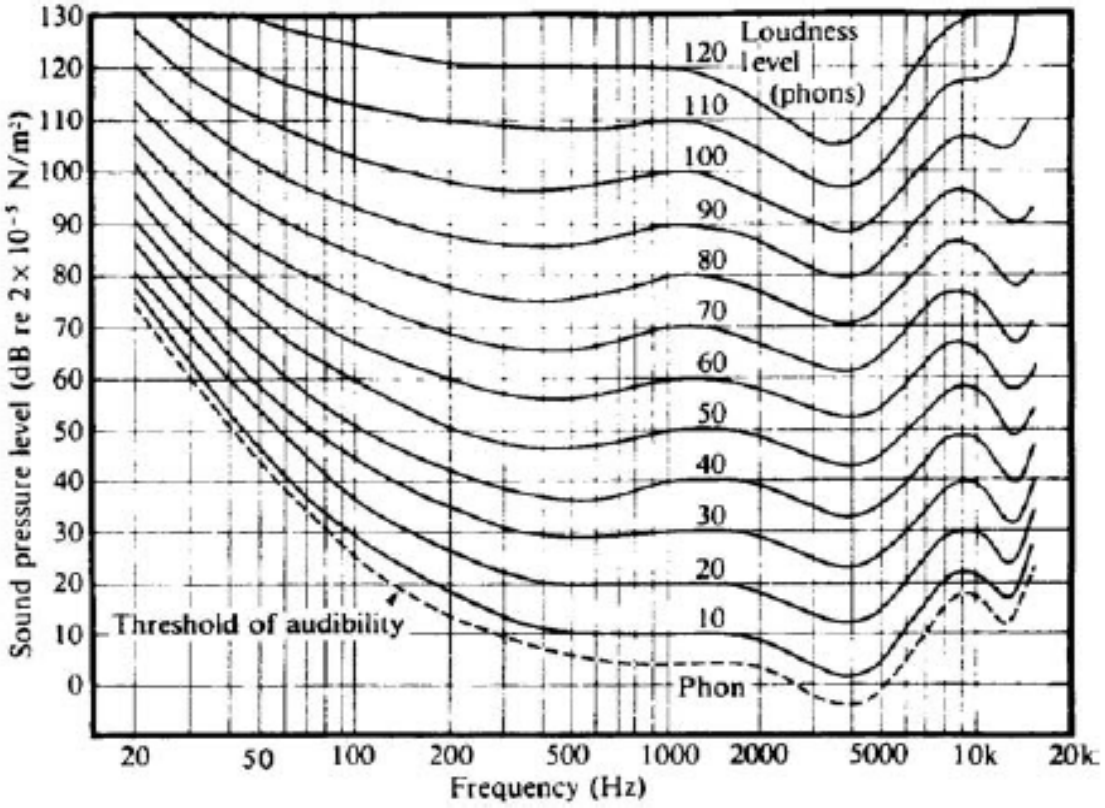
$$dB=10\log_{10}\left(1-\frac{-80}{100}\right)=6.99dB$$

4. Input voltage of a loudspeaker is raised by 30% what will be the amount of increase in L_p

Loudness

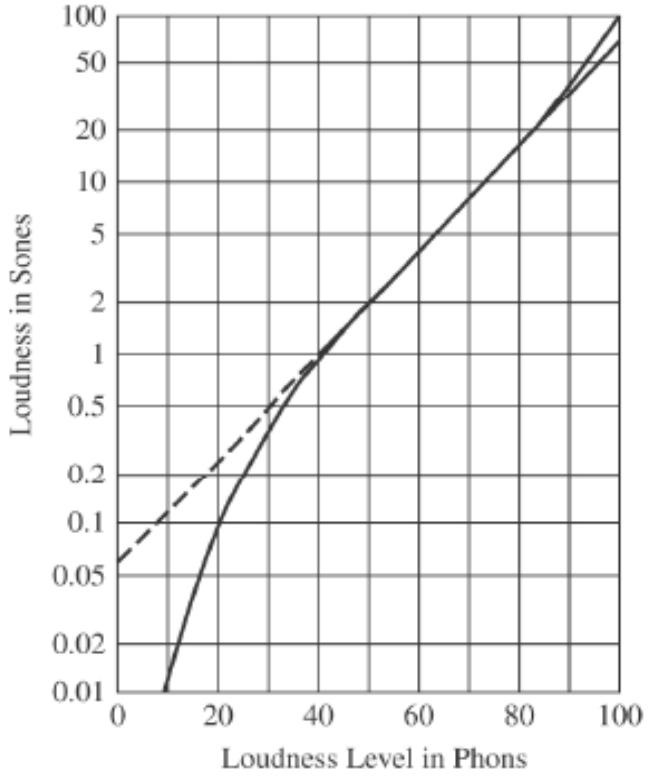
Loudness is a perceptual quality that is related to the physical property of sound pressure level.

Loudness is quantified by relating the actual sound pressure level of a pure tone (in dB relative to a standard reference level) to the perceived loudness of the same tone (in a unit called phons) over the range of human hearing (20 Hz–20 kHz)



Loudness

- ***Loudness (L)*** (in sones) is a scale that doubles whenever the *perceived* loudness doubles



$$\log L = 0.033\,(LL - 40)$$
$$= 0.033LL - 1.32$$

• for a frequency of 1000 Hz, the loudness level, LL, in phons is, by definition, numerically equal to the intensity level IL in decibels, so that the equation may be rewritten as

$$LL = 10\log(I / I_0)$$

or since $I_0 = 10^{-12}$ watts/m^2

$$LL = 10\log I + 120$$

Substitution of this value of *LL* in the equation gives

$$\log L = 0.033(10\log I + 120) - 1.32$$
$$= 0.33\log I + 2.64$$

which reduces to

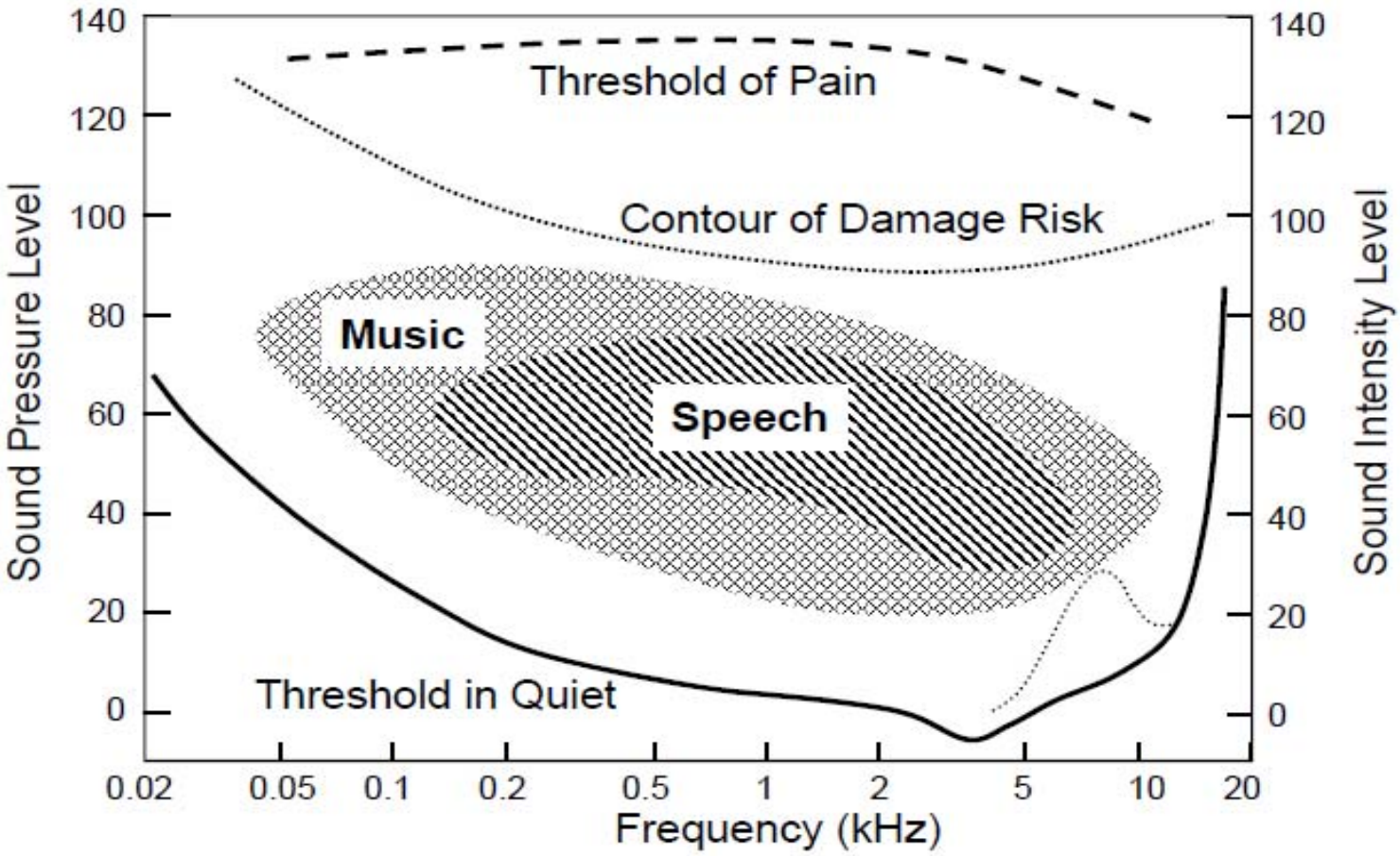
$$L = 445I^{0.33}$$

Sound Pressure Levels (dB)

SPL (dB)	Sound Source
160	Jet Engine — close up
150	Firecracker; Artillery Fire
140	Rock Singer Screaming into Microphone: Jet Takeoff
130	***Threshold of Pain***; .22 Caliber Rifle
120	Planes on Airport Runway; Rock Concert; Thunder
110	Power Tools; Shouting in Ear
100	Subway Trains; Garbage Truck
90	Heavy Truck Traffic; Lawn Mower
80	Home Stereo — 1 foot; Blow Dryer

SPL (dB)	Sound Source
70	Busy Street; Noisy Restaurant
60	Conversational Speech — 1 foot
50	Average Office Noise; Light Traffic; Rainfall
40	Quiet Conversation; Refrigerator; Library
30	Quiet Office; Whisper
20	Quiet Living Room; Rustling Leaves
10	Quiet Recording Studio; Breathing
0	***Threshold of Hearing***

Range of Human Hearing



Perception of Frequency

Pure tone

Pitch is a perceived quantity while Frequency is a physical one (cycle per second or Hertz)

Mel is a scale that doubles whenever the perceived pitch doubles;

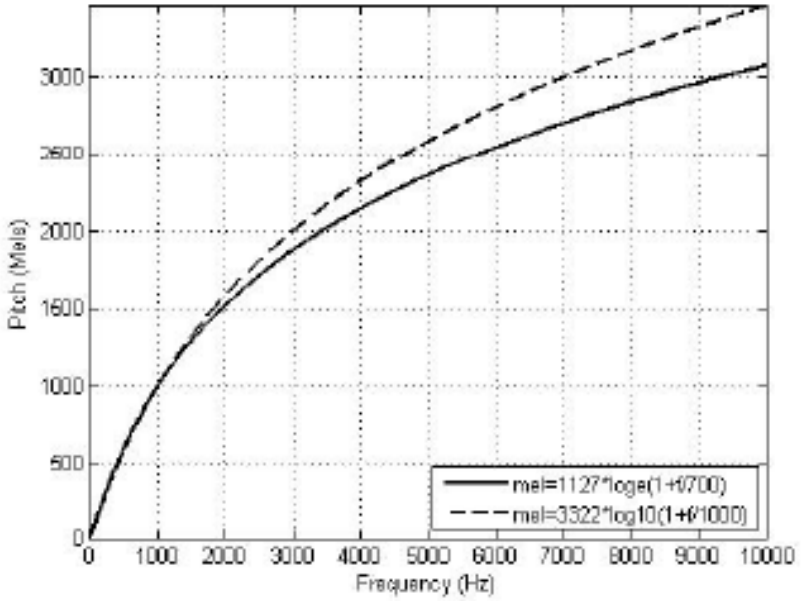
start with 1000 Hz = 1000 mel, increase frequency of tone until listener perceives twice the pitch (or decrease until half the pitch) and so on to find mel-Hz relationship

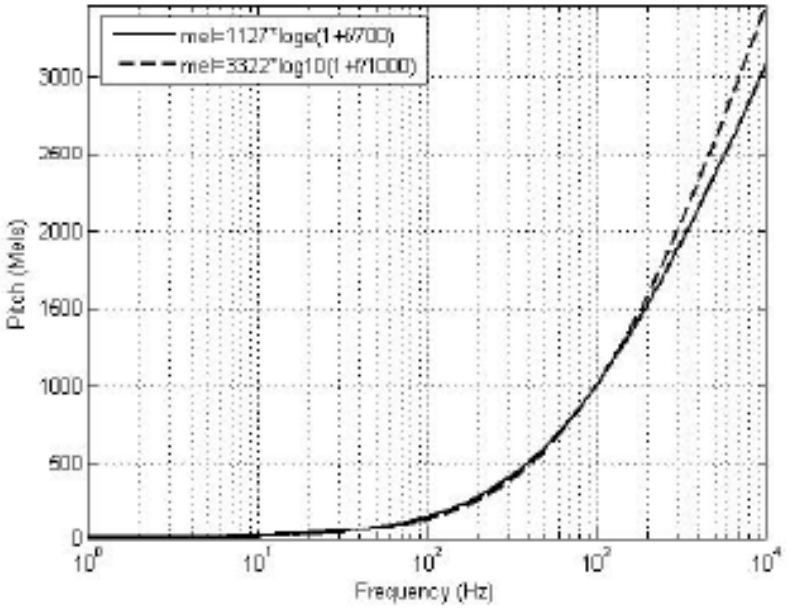
The relationship between pitch and frequency is non-linear

Complex sound such as speech

Pitch is related to fundamental frequency but not the same as fundamental frequency; the relationship is more complex than pure tones

Pitch-The Mel Scale





$$\text{Pitch } (mels) = 3322 \log_{10}(1 + f / 1000)$$

Alternatively, we can approximate curve as:

$$\text{Pitch } (mels) = 1127 \log_{e}(1 + f / 700)$$

Critical Bands

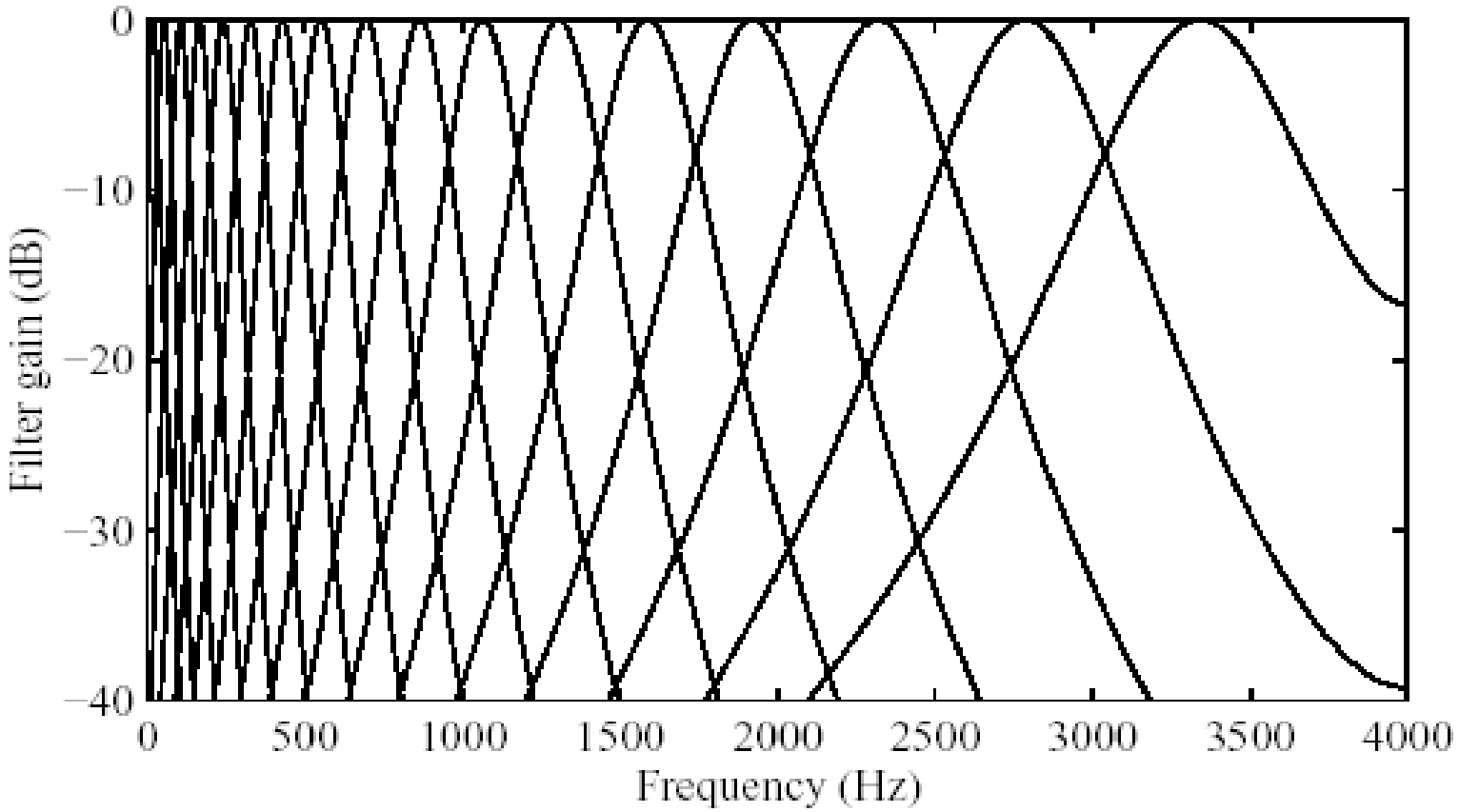
- The ear cannot distinguish sounds within the same band that occur simultaneously.
- Each band is called a critical band
- The auditory system can be roughly modeled as a filterbank, consisting of 25 overlapping bandpass filters, from 0 to 20 KHz
- The bandwidth of each critical band is about 100 Hz for signals below 500 Hz, and increases non-linearly after 500 Hz up to 5000 Hz

1 bark = width of 1 critical band

$$\text{Bark} = \begin{cases} f/100, & f \le 500\text{Hz} \\ 9 + 4\log_2(f/1000), & f > 500\text{Hz} \end{cases}$$

Critical Bands in Masking

Critical bands: The widths of the masking bands for different masking tones are different, increasing with the frequency of the masking tone.



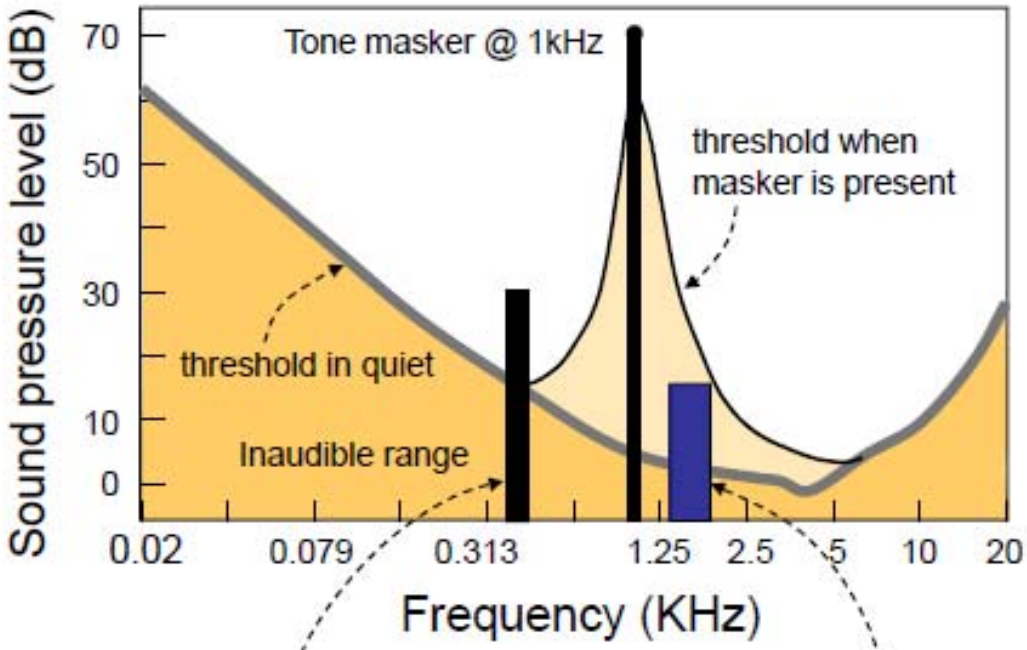
Cochlea-Overlapping Bandpass Filter

A distance of 1 critical band is commonly referred to as “one bark” in the literature.

Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

Masking as defined by the American Standards Association (ASA) is the amount (or the process) by which the threshold of audibility for one sound is raised by the presence of another (masking) sound (B.C.J. Moore 1982, p. 74)

Auditory Masking

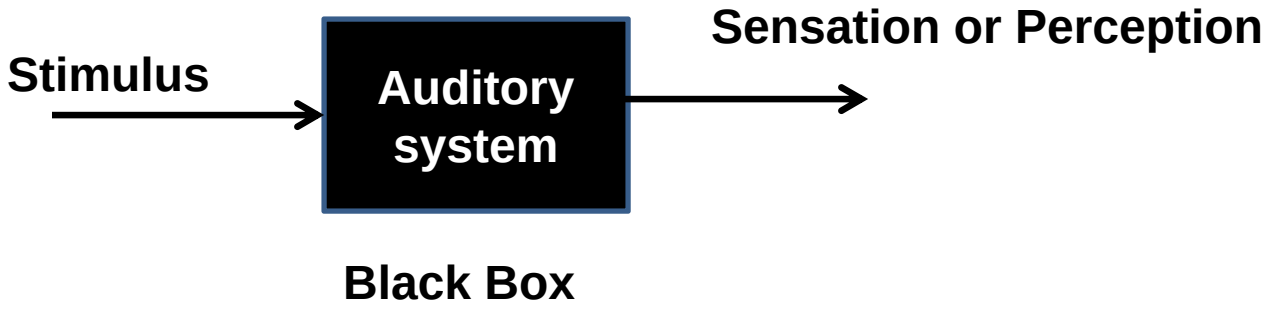


Signal perceptible even in the presence of the tone masker

Signal not perceptible due to the presence of the tone masker

Different Views of Auditory Perception

Functional: based on studies of psychophysics – relates stimulus (*physics)* to perception (*psychology): e.g. frequency in Hz. vs. Mel/Bark scale.*



Structural: based on studies of physiology/anatomy – how various body parts work with emphasis on the process; e.g. neural processing of a sound

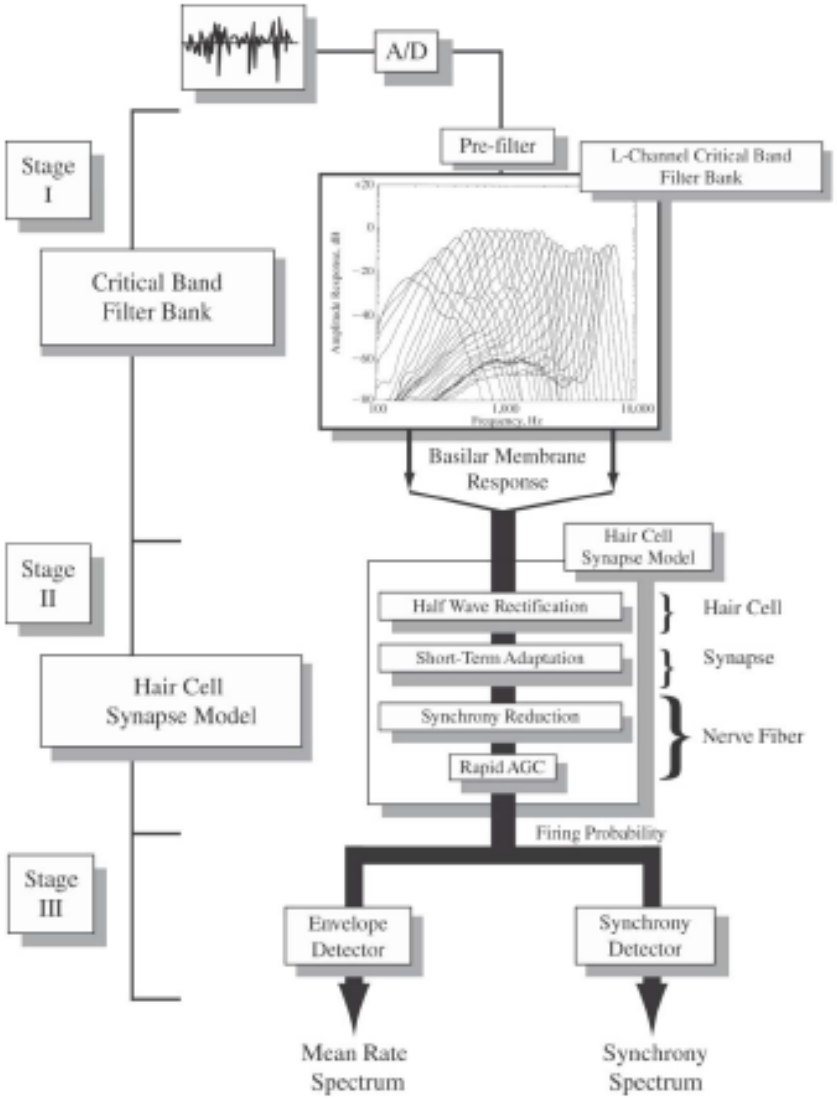
Perceptual effects included in most auditory models

- Spectral analysis on a non-linear frequency scale (usually mel or Bark scale)
- Spectral amplitude compression (dynamic range compression)
- Loudness compression via some logarithmic process
- Decreased sensitivity at lower (and higher) frequencies based on results from equal loudness contours
- Utilization of temporal features based on long spectral integration intervals (syllabic rate processing)
- Auditory masking by tones or noise within a critical frequency band of the tone (or noise)

Auditory Models

- Perceptual Linear Prediction
- Seneff Auditory Model
- Lyon’s Cochlear Model
- Gamma tone Filter Bank Model for Inner Ear
- Inner Hair Cell Model

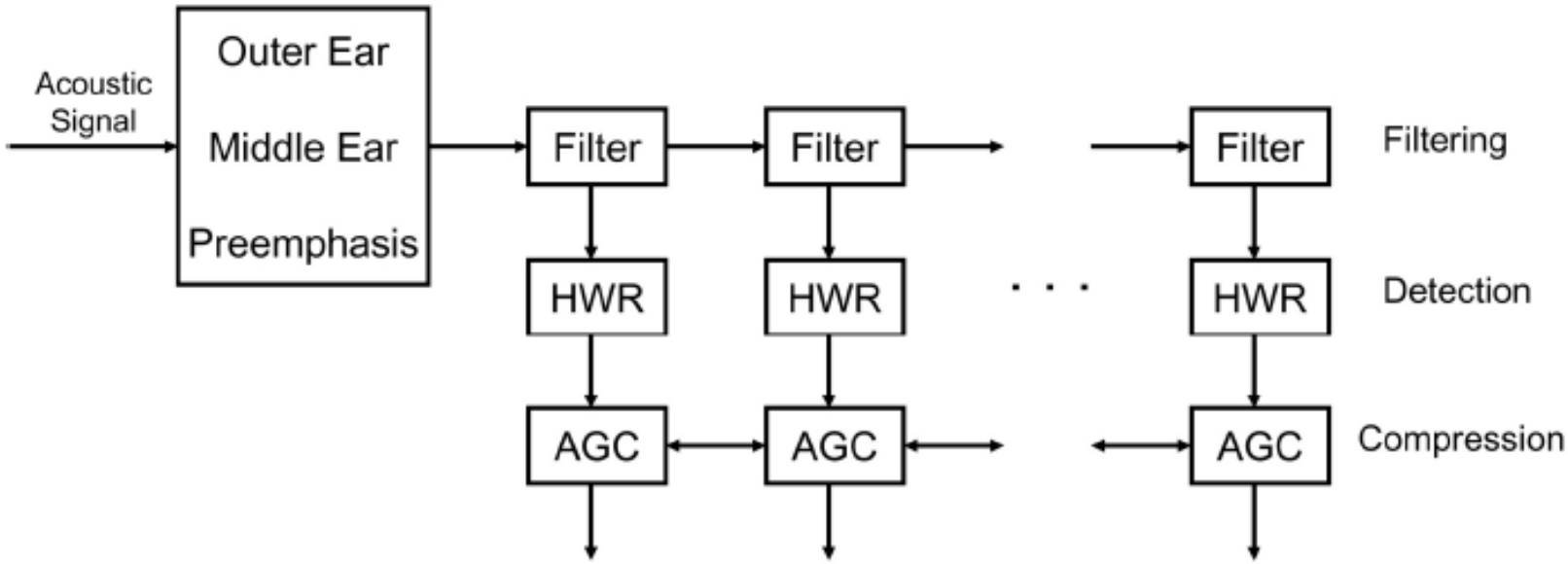
Seneff Auditory Model



Seneff Auditory Model

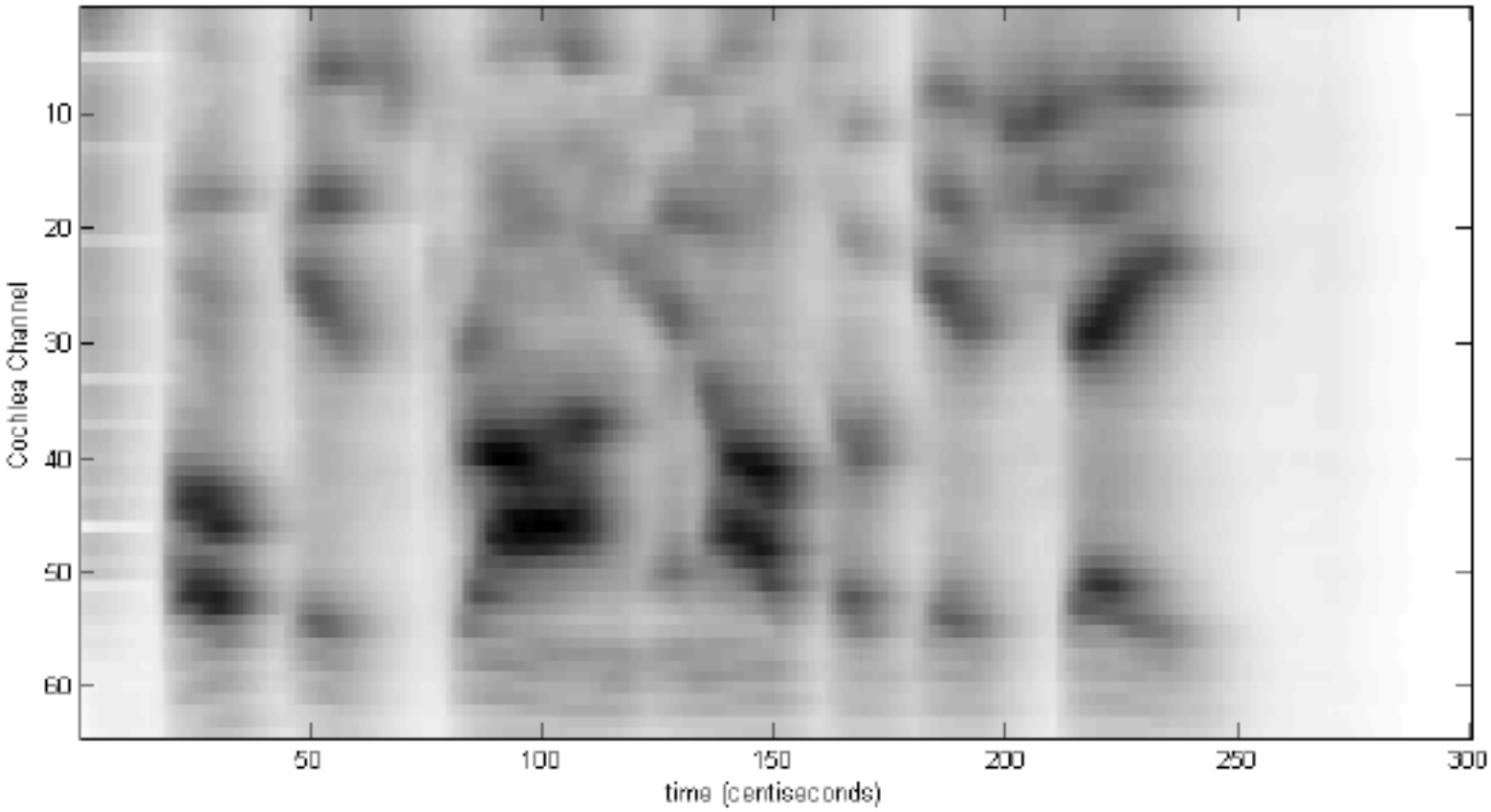
- This model tried to capture essential features of the response of the cochlea and the attached hair cells in response to speech sound pressure waves
- Three stages of processing:
  - stage 1 pre-filters the speech to eliminate very low and very high frequency components, and then uses a 40-channel critical band filter bank distributed on a Bark scale
  - stage 2 is a hair cell synapse models which models the (probabilistic) behavior of the combination of inner hair cells, synapses, and nerve fibers via the processes of half wave rectification, short-term adaptation, and synchrony reduction and rapid automatic gain control at the nerve fiber; outputs are the probabilities of firing, over time, for a set of similar fibers acting as a group
  - stage 3 utilizes the firing probability signals to extract information relevant to perception; i.e., formant frequencies and enhanced sharpness of onset and offset of speech segments; an Envelope Detector estimates the Mean Rate Spectrum (transitions from one phonetic segment to the next) and a Synchrony Detector implements a phase-locking property of nerve fibers, thereby enhancing spectral peaks at formants and enabling tracking of dynamic spectral changes

Lyon's Cochlear Model



- Pre-processing stage (simulating effects of outer and middle ears as a simple pre-emphasis network)
  - three full stages of processing for modeling the cochlea as a non-linear filter bank
  - first stage is a bank of 86 cochlea filters, space non0uniformly according to mel or Bark scale, and highly overlapped in frequency
  - second stage uses a half wave rectifier non-linearity to convert basilar membrane signals to Inner Hair Cell receptor potentials or Auditory Nerve firing rates
  - third stage consists of inter-connected AGC circuits which continuously adapt in response to activity levels at the outputs of the HWRs of the second stage to compress the wide range of sound levels into a limited dynamic range of basilar membrand motion, IHC receptor potential and AN firing rates

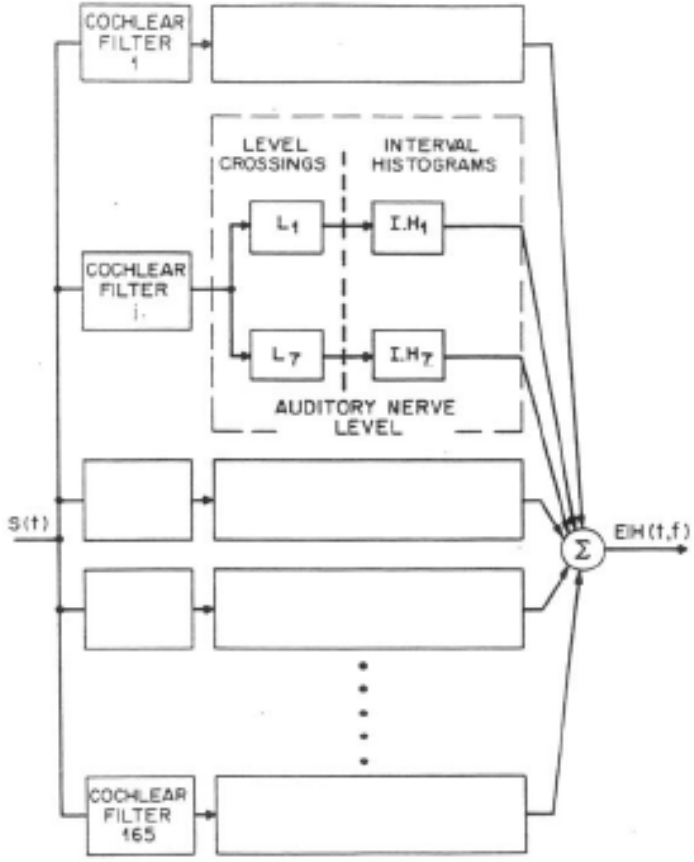
Lyon's Cochleargram



Cochleagram is a plot of model intensity as a function of place (warped frequency) and time; i.e., a type of auditory model spectrogram.

Ensemble Interval Histogram (EIH)

• model of cochlear and hair cell transduction => filter bank that models frequency selectivity at points along the BM, and nonlinear processor for converting filter bank output to neural firing patterns along the auditory nerve



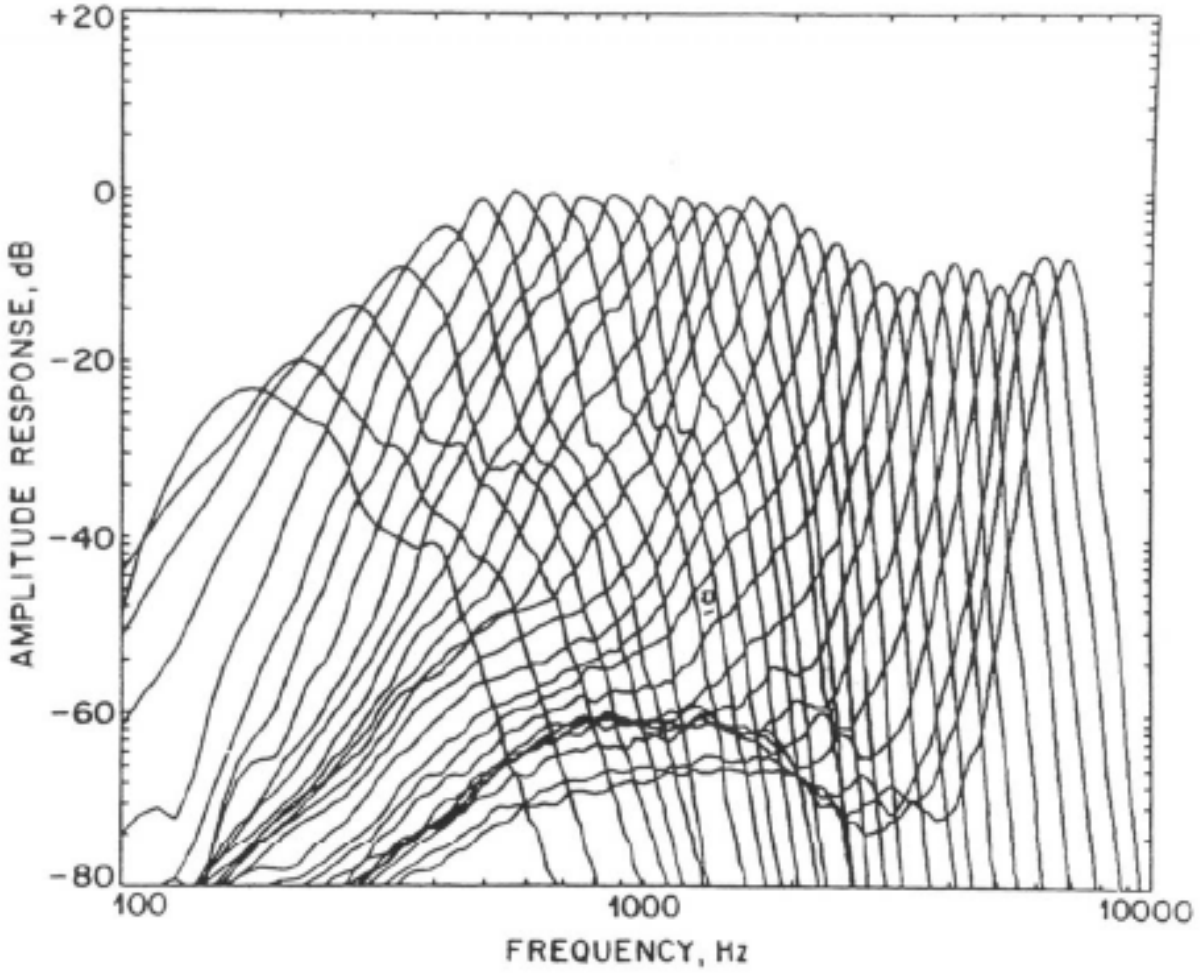
• 165 channels, equally spaced on a log frequency scale between 150 and 7000 Hz

• cochlear filter designs match neural tuning curves for cats => minimum phase filters

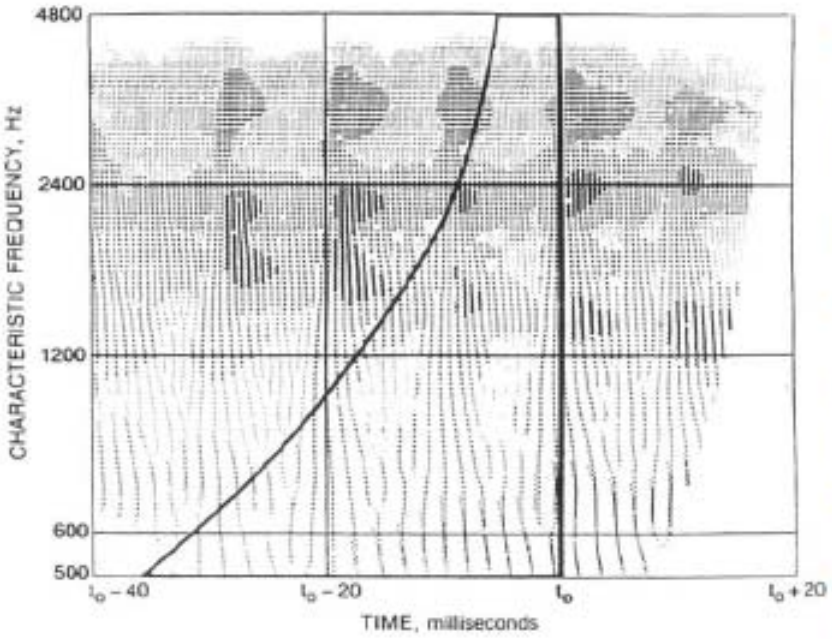
• array of level crossing detectors that model motion-to-neural activity transduction of the IHCs

  • detection levels are pseudo-randomly distributed to match variability of fiber diameters

Cochlear Filter Designs



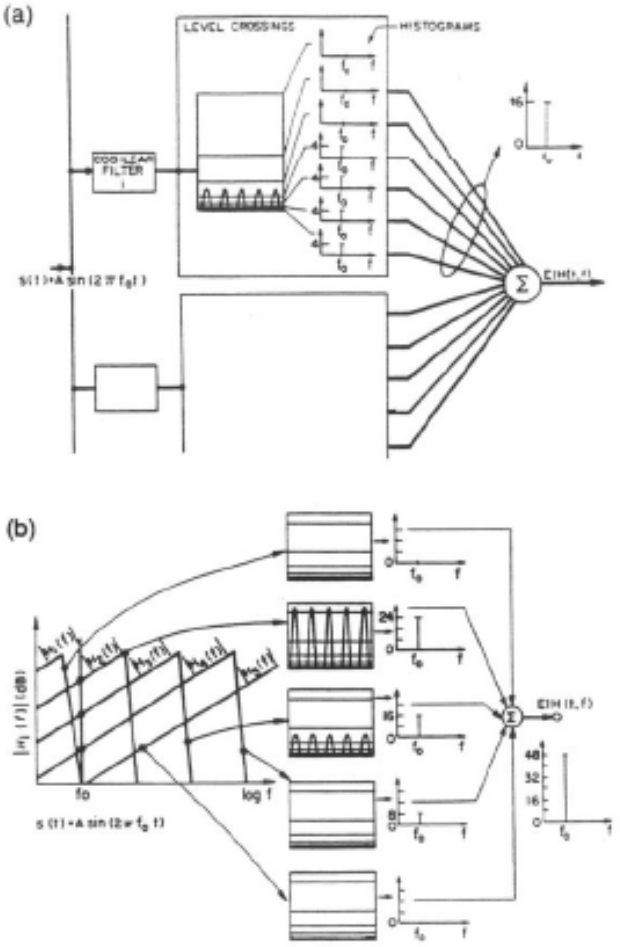
EIH Responses



- plot shows simulated auditory nerve activity for first 60 msec of /o/ in both time and frequency of IHC channels
- log frequency scale
- level crossing occurrence marked by single dot; each level crossing detector is a separate trace
- for filter output low level—1 or fewer levels will be crossed
- for filter output high level—many levels crossed => darker region

Overall EIH

- EIH is a measure of spatial extent of coherent neural activity across auditory nerve
- it provides estimate of short term PDF of reciprocal of intervals between successive firings in a characteristic frequency-time zone
- EIH preserves signal energy since threshold crossings are functions of amplitude
  - as A increases, more levels are activated



response to pure sinusoid

Why Auditory Models

- Match human speech perception
  - Non-linear frequency scale – mel, Bark scale
  - Spectral amplitude (dynamic range) compression – loudness (log compression)
  - Equal loudness curve – decreased sensitivity at lower frequencies
  - Long spectral integration – “temporal” features

What Do We Learn From Auditory Models

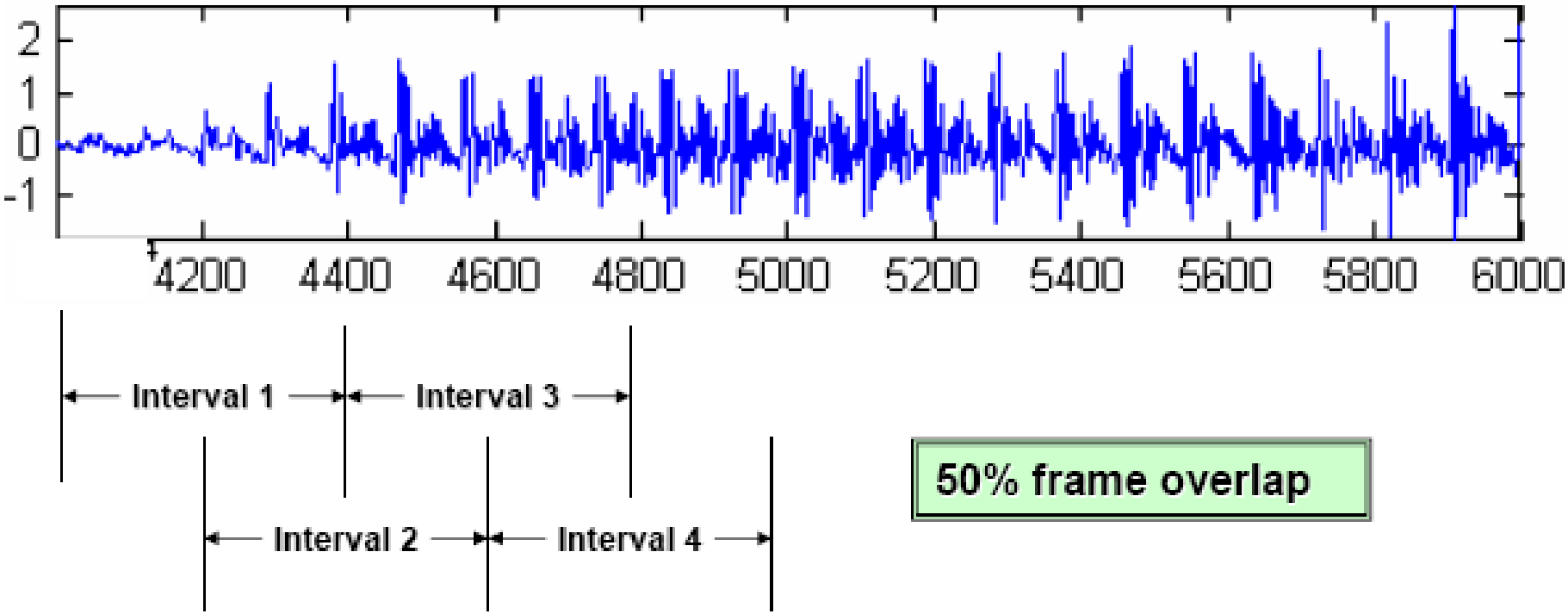
- Need both short (20 msec for phonemes) and long (200 msec for syllables) segments of speech
- Temporal structure of speech is important
- Spectral structure of sounds (formants) is important
- Dynamic (delta) features are important

Time Domain Methods in Speech Processing

Fundemental Assumptions

- Properties of Speech Signal change relatively slowly with time (5-10 sounds per second)
- Uncertainty in short/Long time measurements and estimates
  - Over very short (5-20ms) intervals
    - Uncertainty due to small amount of data, varying pitch and amplitude
  - Over medium Length intervals (20-100ms)
    - Uncertainty due to changes in sound quality, transition between sounds, rapid transients in speech
  - Overlong Intervals (100-500ms)
    - Uncertainty due to large amount of sound changes

Frame-by-Frame Processing



50% frame overlap

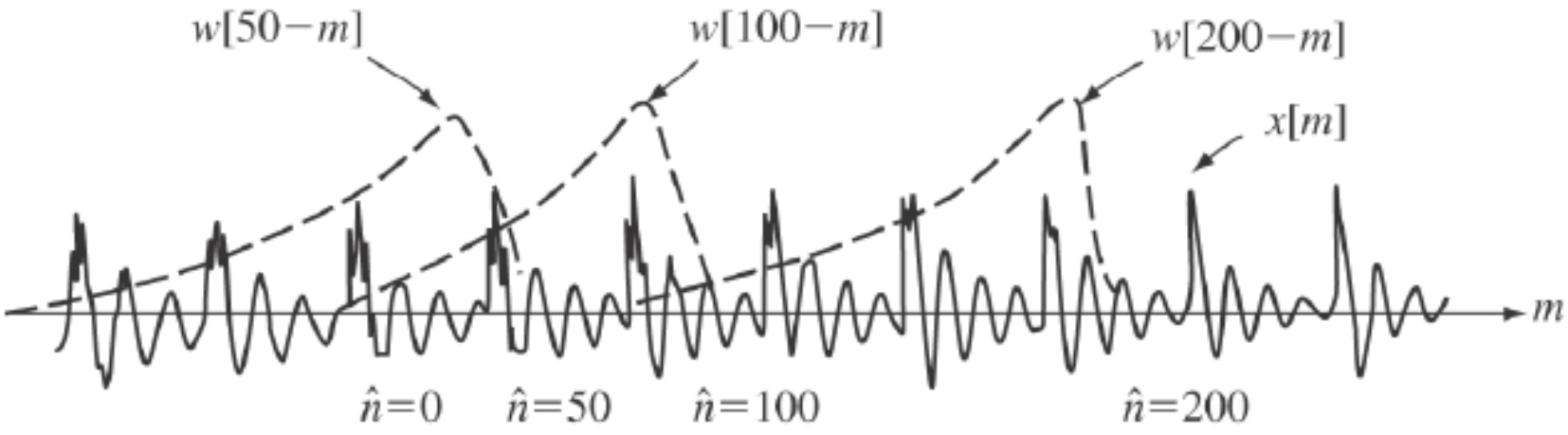
- speech is processed frame-by-frame in overlapping intervals until entire region of speech is covered by at least one such frame
- results of analysis of individual frames used to drive model parameters in some manner

Definition of STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

both $\hat{n}$ and $\hat{\omega}$ are variables

- $w(\hat{n}-m)$ is a real window which determines the portion of $x(\hat{n})$ that is used in the computation of $X_{\hat{n}}(e^{j\hat{\omega}})$



Time-domain processing

- **Time-domain parameters**
  - Short-time energy
  - Short-time average magnitude
  - Short-time zero crossing rate
  - Short-time autocorrelation
  - Short-time average magnitude difference

Short-Time Energy

$$E = \sum_{m=-\infty}^{\infty} x^2[m]$$

-- this is the long term definition of signal energy

-- there is little or no utility of this definition for time-varying signals

$$E_{\hat{n}} = \sum_{m=\hat{n}-N+1}^{\hat{n}} x^2[m] = x^2[\hat{n}-N+1] + \ldots + x^2[\hat{n}]$$

-- short-time energy in vicinity of time $\hat{n}$

$$T(x) = x^2$$

$$\tilde{w}[n] = 1 \qquad 0 \le n \le N-1$$

$$= 0 \qquad \text{otherwise}$$

Computation of Short-Time Energy

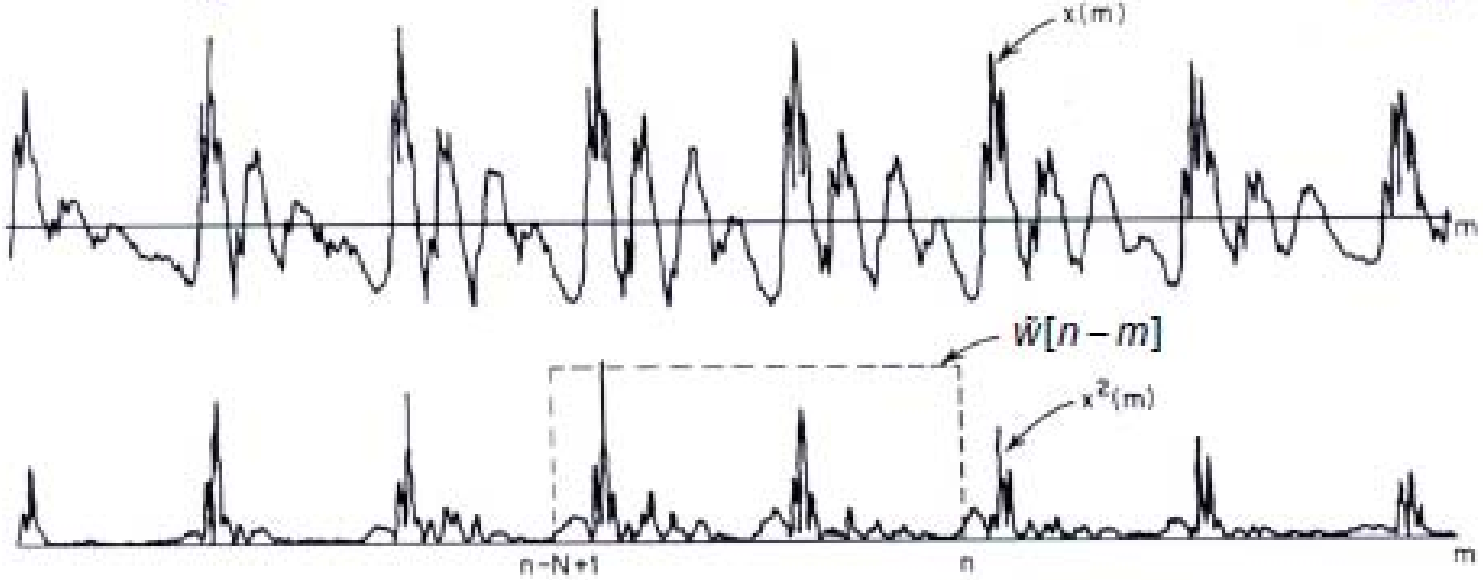


Fig. 4.2 Illustration of the computation of short-time energy.

- **window jumps/slides across sequence of squared values**, selecting interval for processing

- what happens to E_n as sequence jumps by 2,4,8,...,*L* samples (E_n is a lowpass function—so it can be decimated without lost of information; why is E_n lowpass?)

- effects of decimation depend on *L*; if *L* is small, then E_n is a lot more variable than if *L* is large (window bandwidth changes with *L*!)

Short-Time Energy Properties

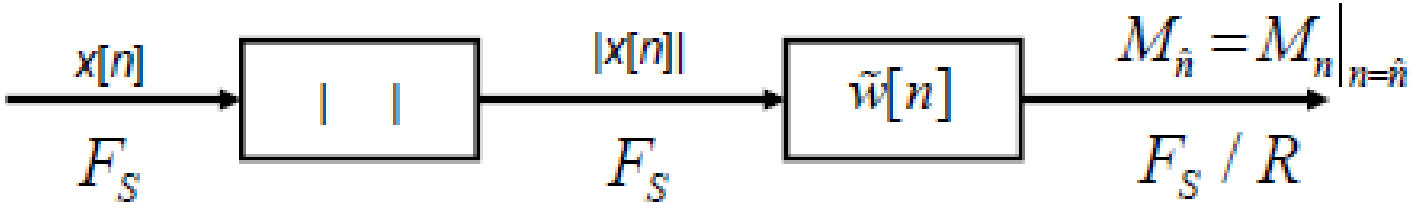
- depends on choice of $h[n]$, or equivalently, window $\tilde{w}[n]$
  - if $w[n]$ duration very long and constant amplitude ($\tilde{w}[n]=1$, $n=0,1,\ldots,L-1$), E_n would not change much over time, and would not reflect the short-time amplitudes of the sounds of the speech
  - very long duration windows correspond to **narrowband lowpass filters**
  - want E_n to change at a rate comparable to the changing sounds of the speech => this is the essential ***conflict*** in all speech processing, namely we need short duration window to be responsive to rapid sound changes, but short windows will not provide sufficient averaging to give smooth and reliable energy function

Short-Time Magnitude

- short-time energy is very sensitive to large signal levels due to $x^2[n]$ terms
  - consider a new definition of 'pseudo-energy' based on average signal magnitude (rather than energy)

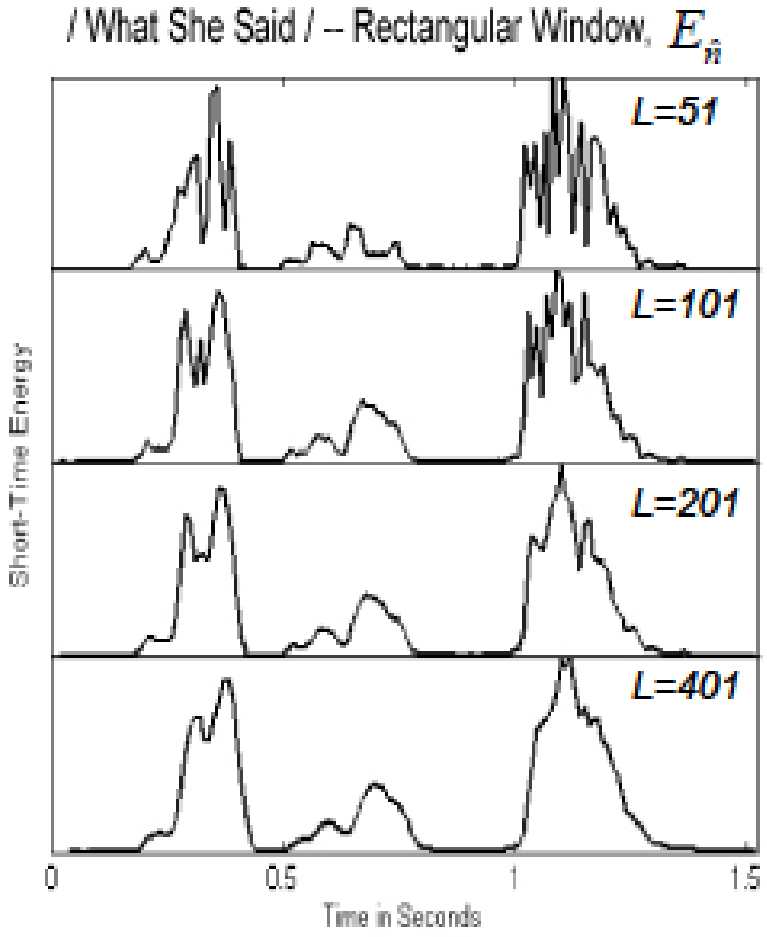
$$M_{\hat{n}} = \sum_{m=-\infty}^{\infty} |x[m]| \, \tilde{w}[\hat{n} - m]$$

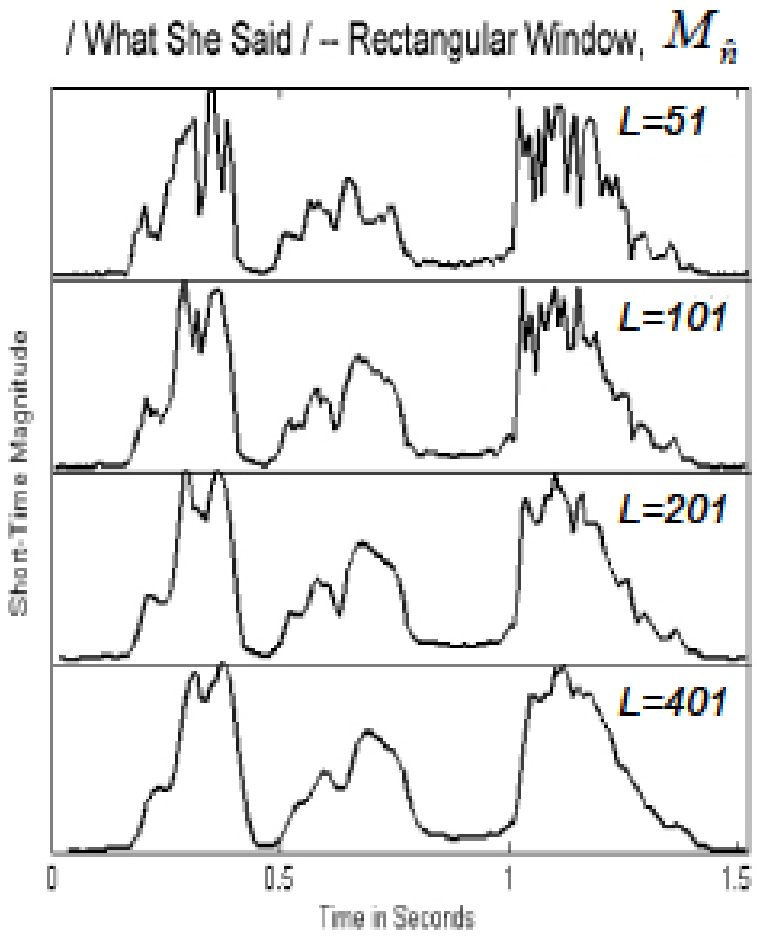
  - weighted sum of magnitudes, rather than weighted sum of squares



- computation avoids multiplications of signal with itself (the squared term)

Short Time Energy and Magnitude—Rectangular Window

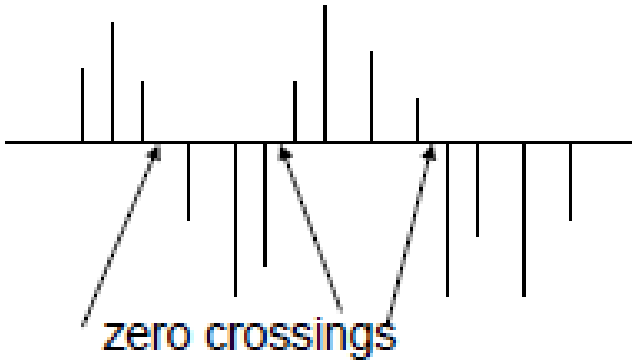




Zero Crossing

- Number of times unvoiced speech crosses the zero line is significantly higher than that of voiced speech.
- Gender of speaker can also have an effect on zero crossing.
- Small pitch weighting can be used to weight the decision threshold.

Short-Time Average ZC Rate



zero crossing => successive samples
have different algebraic signs

- zero crossing rate is a simple measure of the 'frequency content' of a signal—especially true for narrowband signals (e.g., sinusoids)
- sinusoid at frequency F_0 with sampling rate F_S has F_S/F_0 samples per cycle with two zero crossings per cycle, giving an average zero crossing rate of

$z_1=(2)$ crossings/cycle x (F_0/F_S) cycles/sample

$z_1=2F_0/F_S$ crossings/sample (i.e., z_1 proportional to F_0)

$z_M=M\,(2F_0/F_S)$ crossings/(M samples)

Sinusoid Zero Crossing Rates

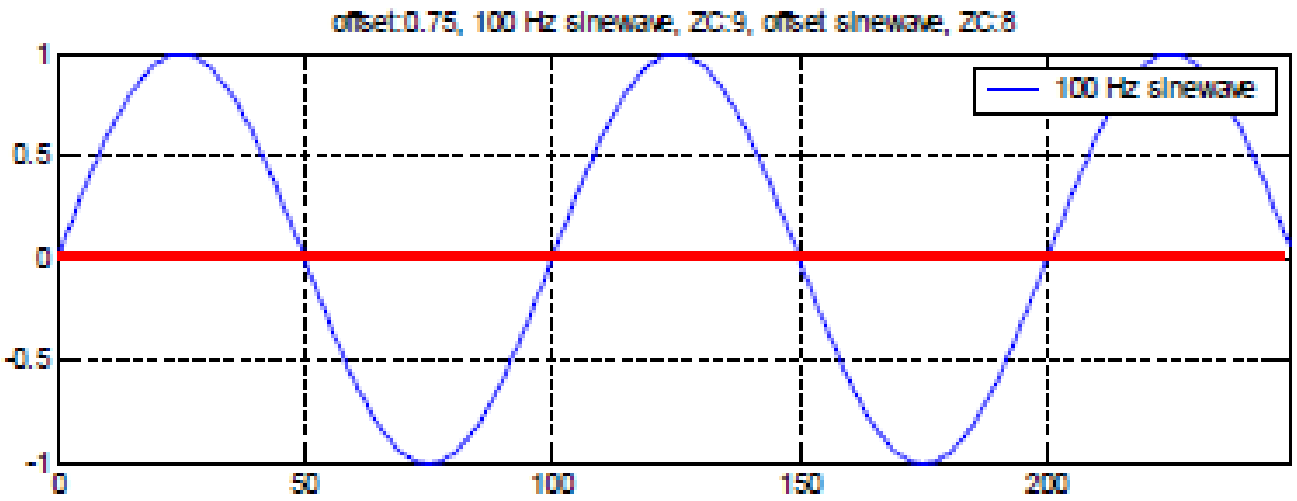
Assume the sampling rate is $F_S = 10{,}000$ Hz

1. $F_0 = 100$ Hz sinusoid has $F_S / F_0 = 10{,}000 / 100 = 100$ samples/cycle;
   or $z_1 = 2/100$ crossings/sample, or $z_{100} = 2/100 * 100 =$
   2 crossings/10 msec interval

2. $F_0 = 1000$ Hz sinusoid has $F_S / F_0 = 10{,}000 / 1000 = 10$ samples/cycle;
   or $z_1 = 2/10$ crossings/sample, or $z_{100} = 2/10 * 100 =$
   20 crossings/10 msec interval

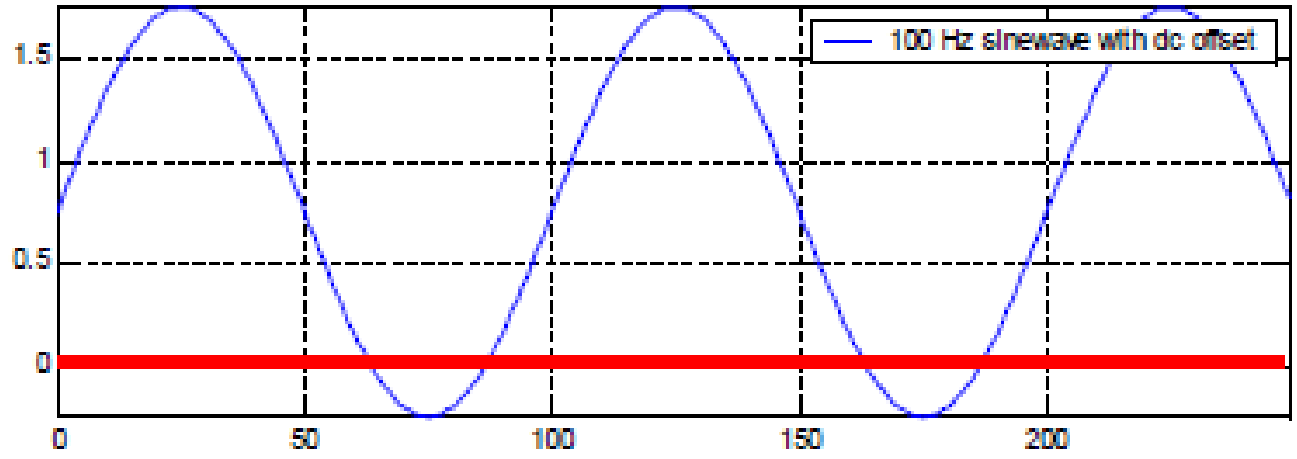
3. $F_0 = 5000$ Hz sinusoid has $F_S / F_0 = 10{,}000 / 5000 = 2$ samples/cycle;
   or $z_1 = 2/2$ crossings/sample, or $z_{100} = 2/2 * 100 =$
   100 crossings/10 msec interval

Zero Crossing for Sinusoids



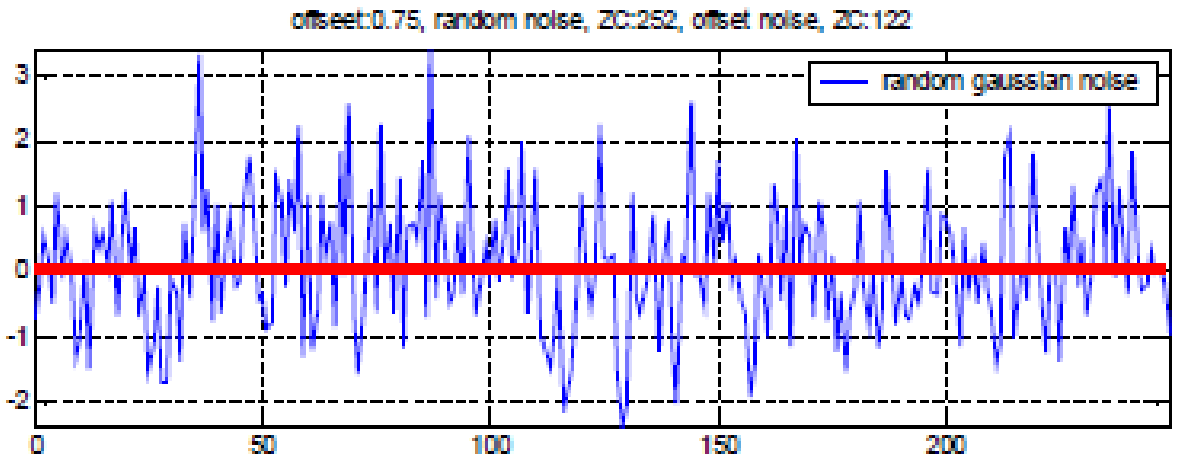
ZC=9

Offset=0.75



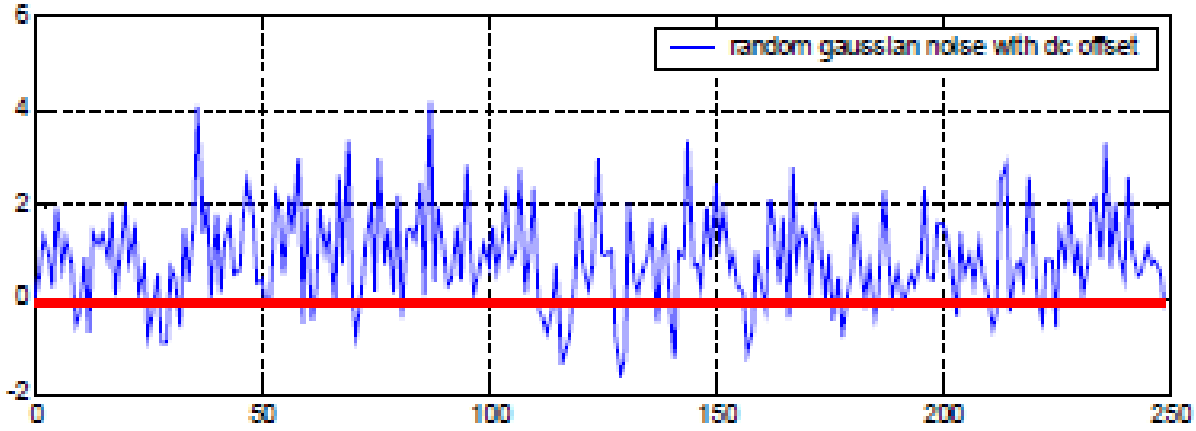
ZC=8

Zero Crossings for Noise



ZC=252

Offset=0.75



ZC=122

ZC Rate Definitions

$$Z_{\hat{n}} = \frac{1}{2L_{\text{eff}}} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\,\text{sgn}(x[m]) - \text{sgn}(x[m-1])\,|\,\tilde{w}[\hat{n}-m]$$

$$\text{sgn}(x[n]) = 1 \qquad x[n] \geq 0$$

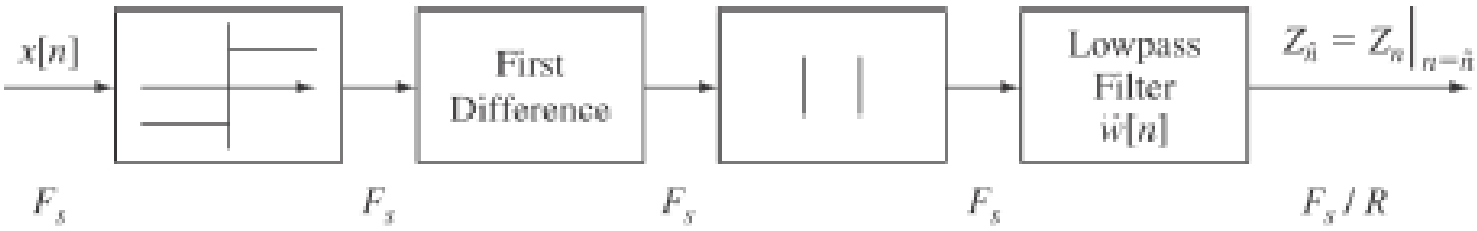
$$= -1 \qquad x[n] < 0$$

- simple rectangular window:

$$\tilde{w}[n] = 1 \qquad 0 \leq n \leq L-1$$

$$= 0 \qquad \text{otherwise}$$

$$L_{\text{eff}} = L$$



Same form for $Z_{\hat{n}}$ as for $E_{\hat{n}}$ or $M_{\hat{n}}$

ZC Normalization

- The formal definition of z_n is:

$$Z_{\hat{n}} = z_1 = \frac{1}{2L} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\operatorname{sgn}(x[m]) - \operatorname{sgn}(x[m-1])|$$

  is interpreted as the number of zero crossings per sample.

- For most practical applications, we need the rate of zero crossings per fixed interval of M samples, which is

$$z_M = z_1 \cdot M = \text{rate of zero crossings per } M \text{ sample interval}$$

  Thus, for an interval of τ sec., corresponding to M samples we get

$$z_M = z_1 \cdot M; \quad M = \tau F_S = \tau / T$$

ZC Normalization

- For a 1000 Hz sinewave as input, using a 40 msec window length (L), with various values of sampling rate (F_S), we get the following:

F_S	L	z_1	M	z_M
8000	320	1 / 4	80	20
10000	400	1 / 5	100	20
16000	640	1 / 8	160	20

- Thus we see that the normalized (per interval) zero crossing rate, z_M, is independent of the sampling rate and can be used as a measure of the dominant energy in a band.

Autocorrelation Technique

- Autocorrelation is a cross-correlation of a signal with itself.

$$\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau)$$

  - The maximum of similarity occurs for time shifting of zero.
  - An other maximum should occur in theory when the time-shifting of the signal corresponds to the fundamental period.

Autocorrelation function

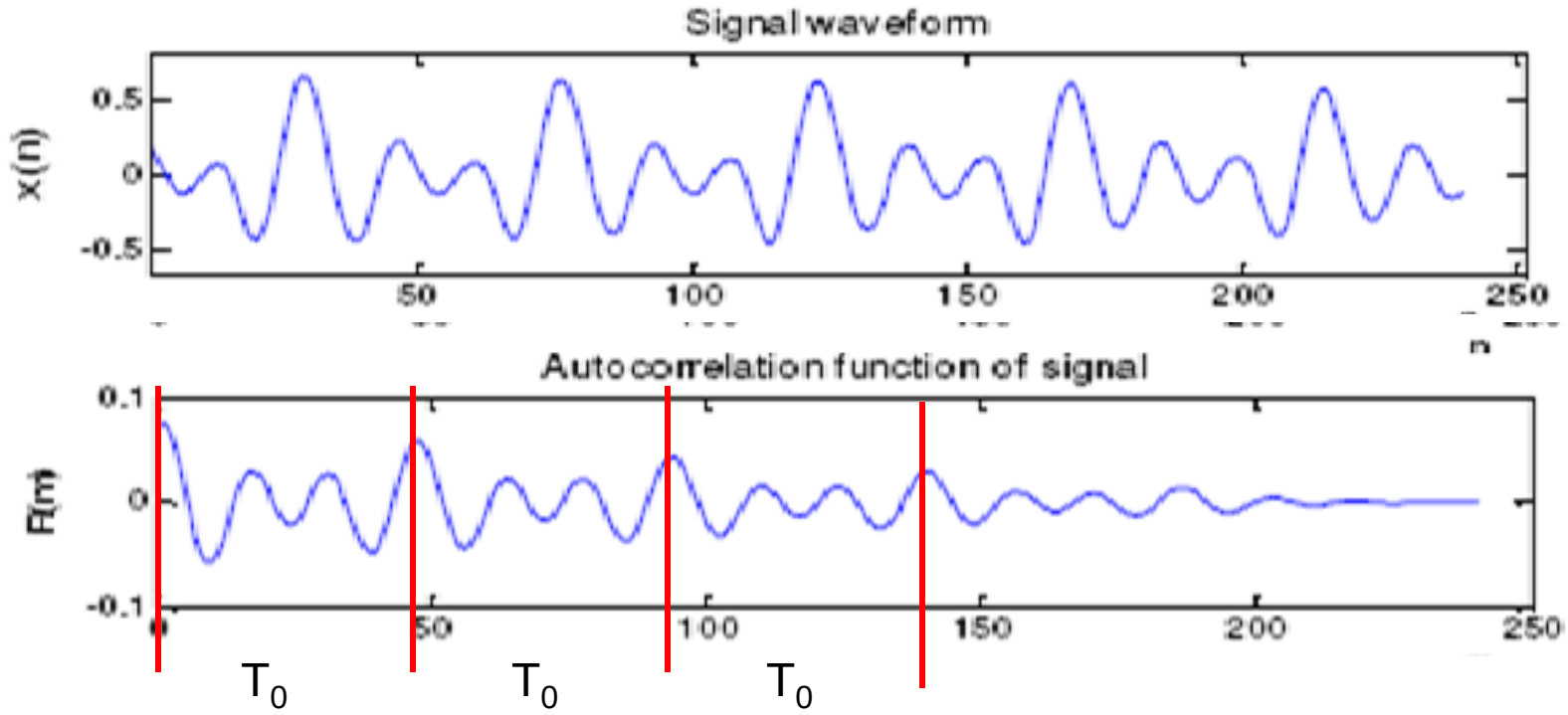
By definition, auto-correlation is

$$R[k] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n] \cdot x[n+k], \quad 0 \le k \le K_0$$

Properties of Autocorrelations is

1. R[k] = R[-k]
2. R[k]is maximum at k = 0

$$R[k] = \frac{1}{N} \sum_{n=0}^{N-1-k} x[n] \cdot x[n+k], \quad 0 \le k \le K_0$$



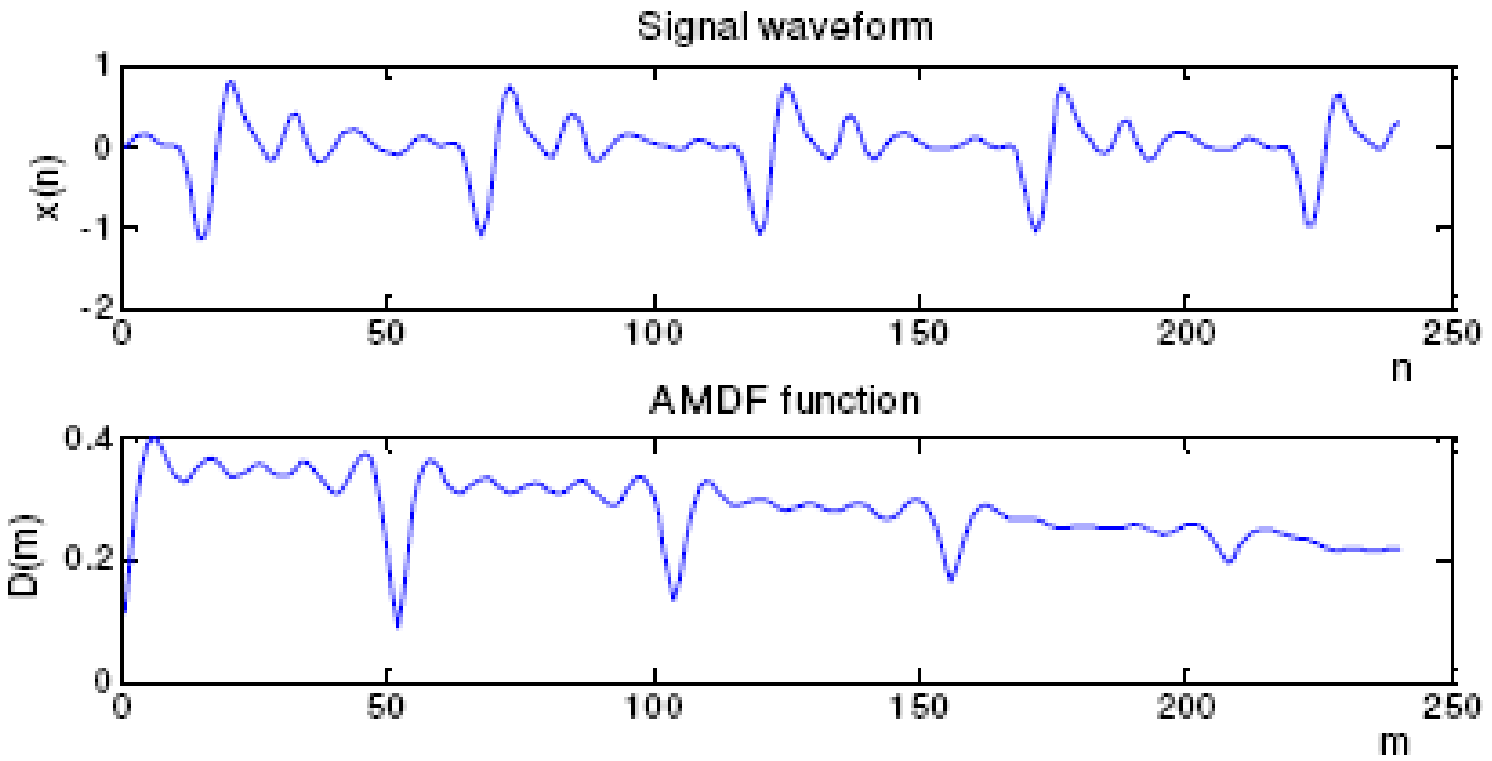
When a segment of a signal is correlated with itself, the distance (*Lag_time_in_samples*) between the positions of the maximum and the second maximum is defined as the *fundamental period* (pitch) of the signal.

Average Magnitude Difference Function(AMDF)

- It is an alternate to Autocorrelation function.
- It compute the difference between the signal and a time-shifted version of itself.

$$D_x[k] = \frac{1}{N} \sum_{n=0}^{N-1-k} \left| x(k) - x(n+k) \right|, \qquad 0 \le k \le K_0$$

- While autocorrelation have peaks at maximum similarity, there will be valleys in the average magnitude difference function.


Signal waveform
x(n)
1
0
-1
-2
0
50
100
150
200
250
n
AMDF function
D(m)
0.4
0.2
0
0
50
100
150
200
250
m

Speech/Non-speech Detection

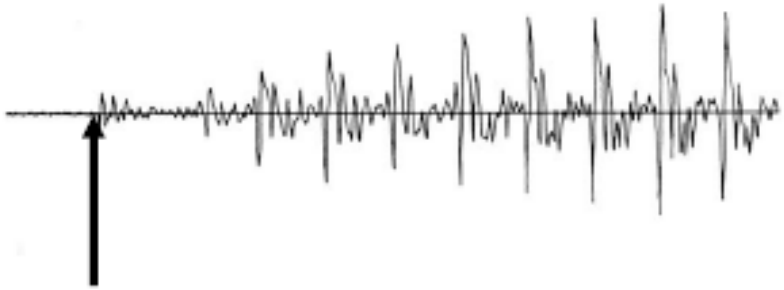
Ideal Speech/Non-Speech Detection

Beginning of
speech interval

Ending of speech
interval

Speech Detection Issues

- key problem in speech processing is locating accurately the beginning and end of a speech utterance in noise/background signal



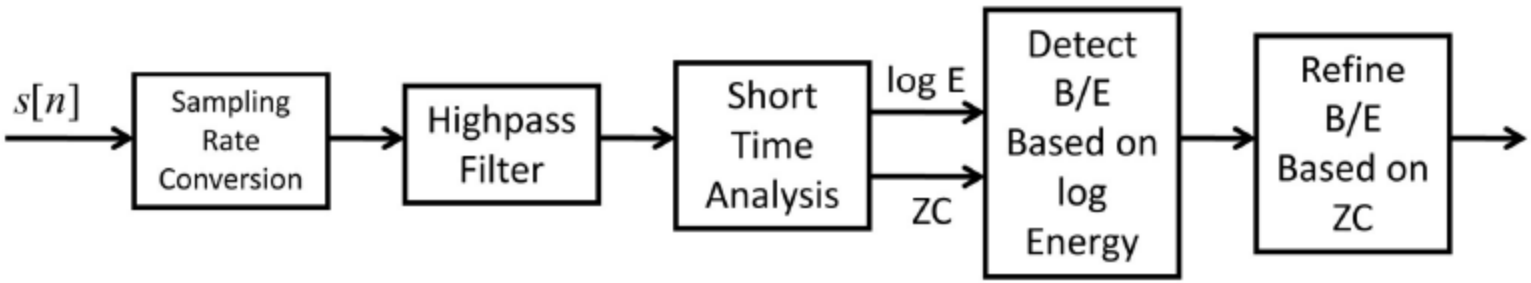
- need endpoint detection to enable:
  - computation reduction (don't have to process background signal)
  - better recognition performance (can't mistake background for speech)
- non-trivial problem except for high SNR recordings

Problems for Reliable Speech Detection

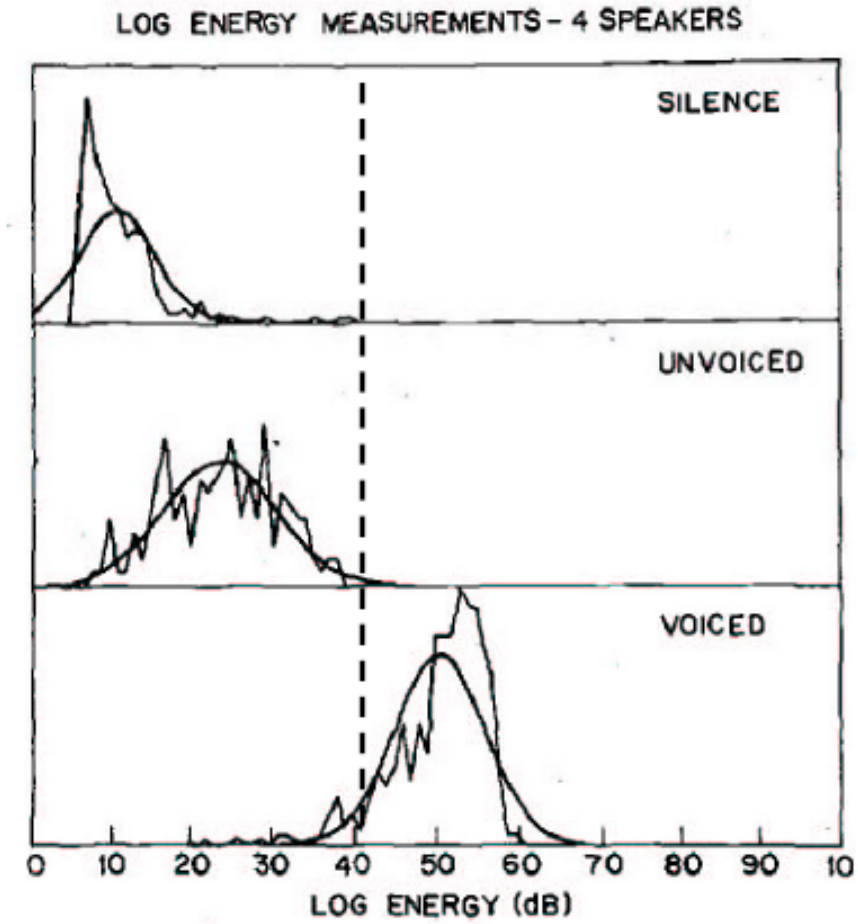
- weak fricatives (/f/, /th/, /h/) at beginning or end of utterance
- weak plosive bursts for /p/, /t/, or /k/
- nasals at end of utterance (often devoiced and reduced levels)
- voiced fricatives which become devoiced at end of utterance
- trailing off of vowel sounds at end of utterance

the good news is that highly reliable endpoint detection is not required for most practical applications; also we will see how some applications can process background signal/silence in the same way that speech is processed, so endpoint detection becomes a moot issue

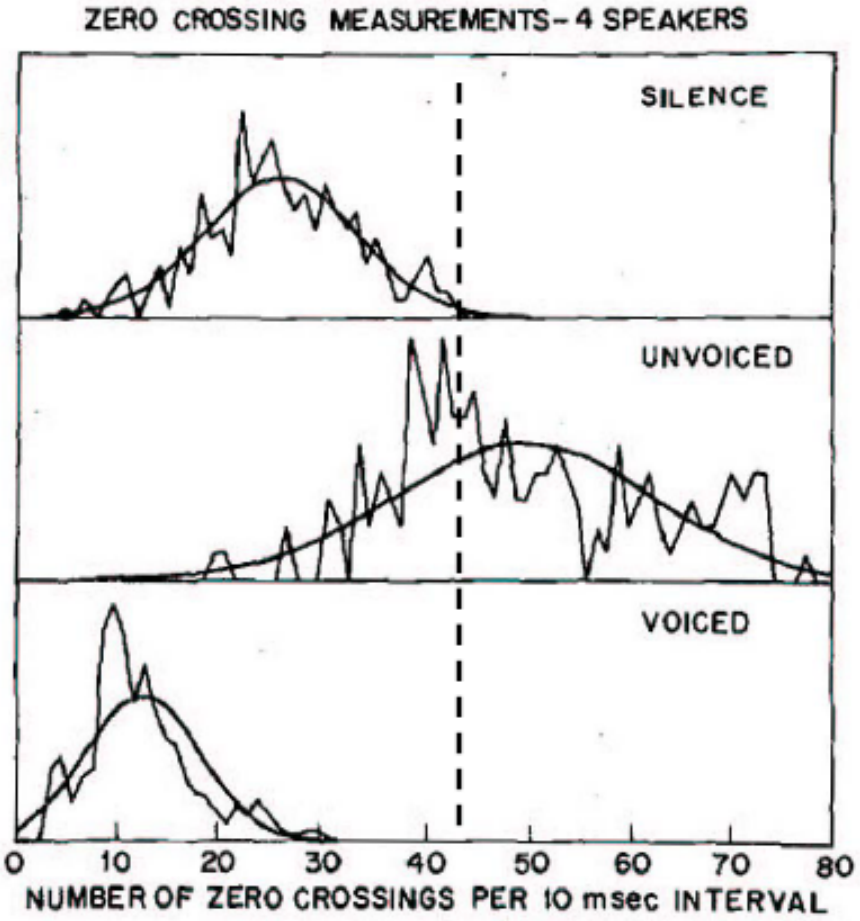
Speech/Non-Speech Detection



Speech/Non-Speech Detection



Log energy separates Voiced from Unvoiced and Silence



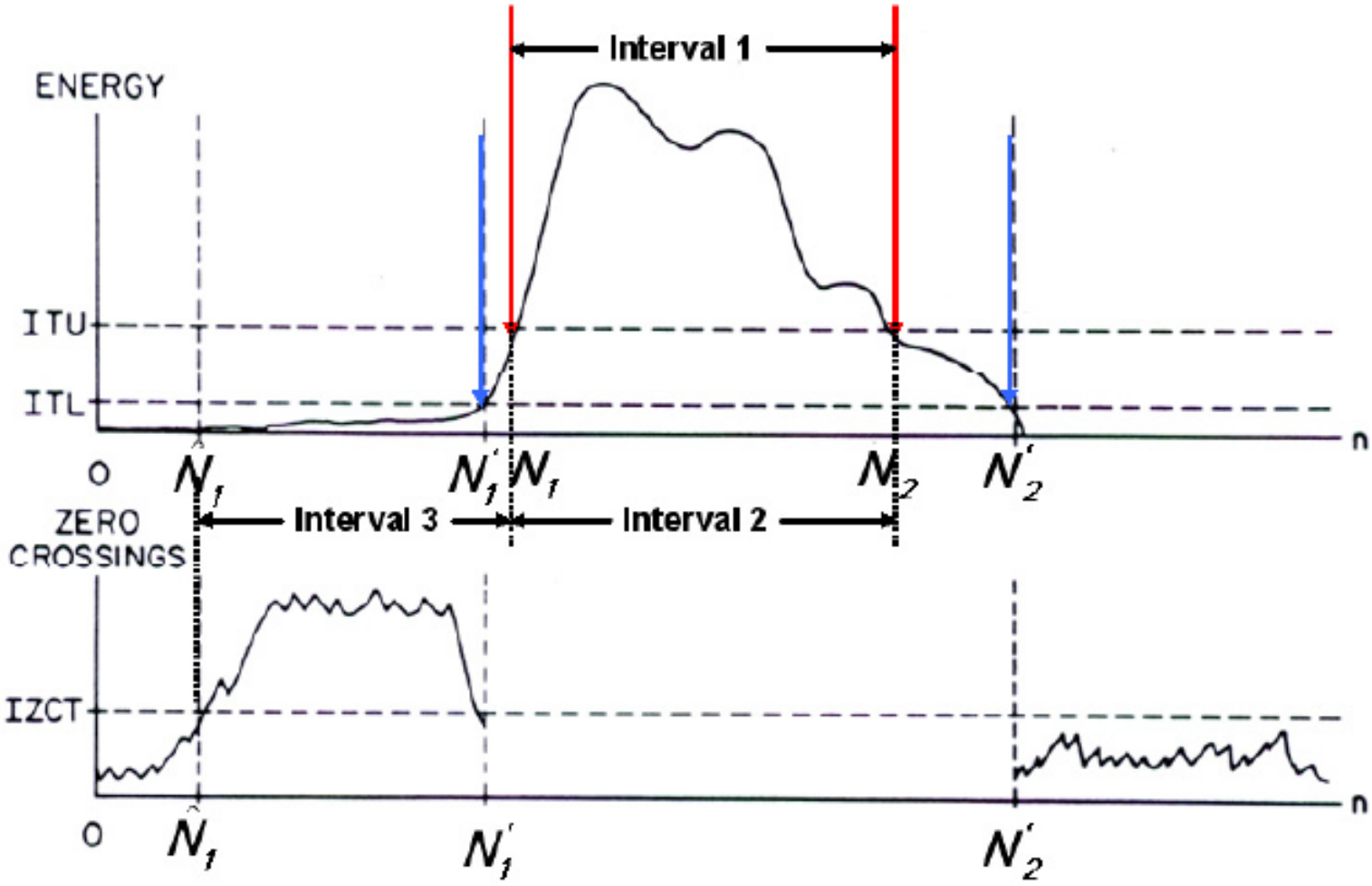
Zero crossings separate Unvoiced from Silence and Voiced

Rule-Based Short-Time Measurements of Speech

Algorithm for endpoint detection:

1. compute mean and σ of log E_n and Z_{100} for first 100 msec of signal (assuming no speech in this interval and assuming F_S=*10,000 Hz*).
2. determine maximum value of log E_n for entire recording => normalization.
3. compute log E_n thresholds based on results of steps 1 and 2—e.g., take some percentage of the peaks over the entire interval. Use threshold for zero crossings based on ZC distribution for unvoiced speech.
4. find an interval of log E_n that exceeds a high threshold ITU.
5. find a putative starting point (N_1) where log E_n crosses ITL from above; find a putative ending point (N_2) where log E_n crosses ITL from above.
6. move backwards from N_1 by comparing Z_{100} to IZCT, and find the first point where Z_{100} exceeds IZCT; similarly move forward from N_2 by comparing Z_{100} to IZCT and finding last point where Z_{100} exceeds IZCT.

Endpoint Detection Algorithm



Speech Parameters

$X = [x_1, x_2, x_3, x_4, x_5]$

$x_1 = \log E_S$ -- short-time log energy of the signal

$x_2 = Z_{100}$ -- short-time zero crossing rate of the signal

for a 100-sample frame

$x_3 = C_1$ -- short-time autocorrelation coefficient at unit sample delay

$x_4 = \alpha_1$ -- first predictor coefficient of a p^{th} order linear predictor

$x_5 = E_p$ -- normalized energy of the prediction error of a p^{th} order linear predictor

Manual Training

- Using a designated training set of sentences, each 10 msec interval is classified manually (based on waveform displays and plots of parameter values) as either:
  - Voiced speech – clear periodicity seen in waveform
  - Unvoiced speech – clear indication of frication or whisper
  - Background signal – lack of voicing or unvoicing traits
  - Unclassified – unclear as to whether low level voiced, low level unvoiced, or background signal (usually at speech beginnings and endings); not used as part of the training set
- Each classified frame is used to train a single Gaussian model, for each speech parameter and for each pattern class; i.e., the mean and variance of each speech parameter is measured for each of the 3 classes

Frequency-domain Processing

- **Spectrogram – short-time Fourier analysis**
  - two-dimensional waveform (amplitude/time) is converted into a three-dimensional pattern (amplitude/frequency/time)
- **Wideband spectrogram:**
  - analyzed on 15ms sections of waveform with a step of 1ms
  - voiced regions with vertical striations due to the periodicity of the time waveform (each vertical line represents a pulse of vocal folds) while unvoiced regions are solid/random, or ‘snowy’
- **Narrowband spectrogram:**
  - analyzed on 50ms sections of waveform with a step of 1ms
  - pitch for voiced intervals in horizontal lines

Frequency-domain Processing