Digital Audio and Speech Processing

(Sayısal Ses ve Konuşma İşleme)

Prof. Dr. Nizamettin AYDIN

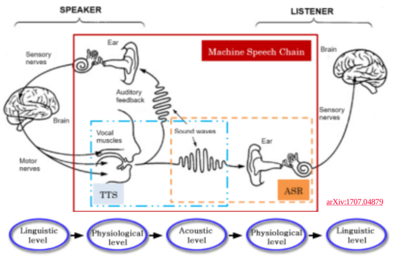
naydin@yildiz.edu.tr
nizamettinaydin@gmail.com
http://www3.yildiz.edu.tr/~naydin

Speech Perception

Speech Perception

- Understanding how we hear sounds and how we perceive speech leads to 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
  - speech and audio coding (MP3 audio, cellphone speech)
  - speech recognition
- Try to understand speech perception by looking at the physiological models of hearing

The Speech Chain



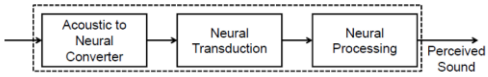
arXiv:1707.04879

- 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 Speech Chain

- The message to be conveyed by speech goes through five levels of representation between the speaker and the listener:
  - the linguistic level
    - where the basic sounds of the communication are chosen to express some thought of idea
  - the physiological level
    - where the vocal tract components produce the sounds associated with the linguistic units of the utterance
  - the acoustic level
    - where sound is released from the lips and nostrils and transmitted to both the speaker (sound feedback) and to the listener
  - the physiological level
    - where the sound is analyzed by the ear and the auditory nerves
  - the linguistic level
    - where the speech is perceived as a sequence of linguistic units and understood in terms of the ideas being communicated

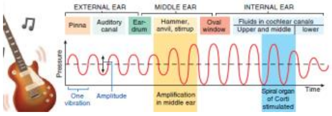
Auditory System



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

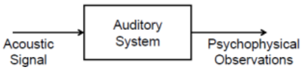
Auditory System

- Two major components:
  - The peripheral auditory organs (the ear)
    - Converts sounds pressure into mechanical vibration patterns, which then are transformed into neuron firings
  - The auditory nervous system (the brain)
    - Extracts perceptual information in various stages
- Route of sound waves through the ear (video)



- To excite the hair cells in the spiral organ of Corti in the inner ear, sound wave vibrations must pass through air, membranes, bone, and fluid.

The Black Box Model of the Auditory System



- Researchers have resorted to a black box behavioral model of hearing and perception
  - Model assumes that an acoustic signal enters the auditory system causing behavior that we record as psychophysical observations
  - Psychophysical methods and sound perception experiments determine how the brain processes signals with different loudness levels, different spectral characteristics, and different temporal properties
  - 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
  - We then determine how various attributes of sound (or speech) are processed by the auditory system

The Black Box Model Examples

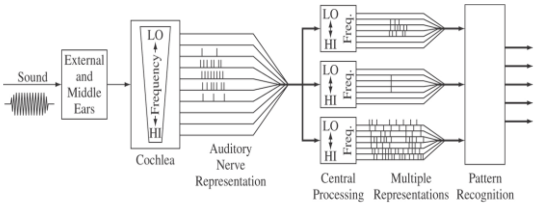
Physical Attribute	Psychophysical Observation
Intensity	Loudness
Frequency	Pitch

- Experiments with the black box model show:
  - Correspondences between sound intensity and loudness, and between frequency and pitch are complicated and far from linear
  - Attempts to extrapolate from psychophysical measurements to the processes of speech perception and language understanding are, at best, highly susceptible to misunderstanding of exactly what is going on in the brain

Why Do We Have Two Ears

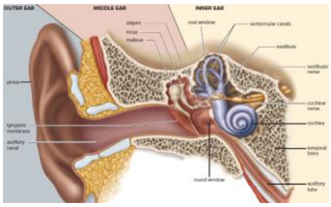
- Sound localization
  - Spatially locate sound sources in 3-dimensional sound fields, based on
    - two-ear processing,
    - loudness differences at the two ears,
    - delay to each ear
- Sound cancellation
  - Focus attention on a selected sound source in an array of sound sources
    - cocktail party effect', Binaural Masking Level Differences (BMLDs)
- Effect of listening over headphones
  - localize sounds inside the head (rather than spatially outside the head)

Overview of Auditory Mechanism



The Human Ear

- The ear can be divided into three main sections
  - Outer ear:
    - Encompasses the pinna (outer cartilage), auditory canal, and eardrum
    - Transforms sound pressure into vibrations

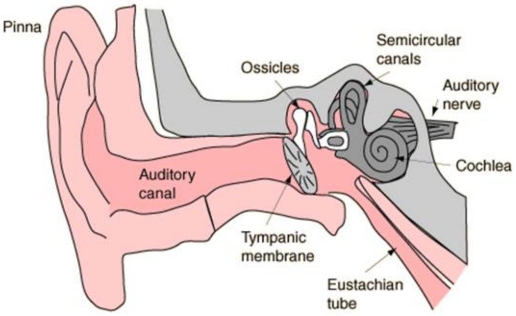


  - Middle ear:
    - Consists of three bones: malleus, incus and stapes
    - Transport eardrum vibrations to the inner ear
  - Inner ear:
    - Consists of the cochlea
    - Transforms vibrations into spike trains at the Basilar Membrane

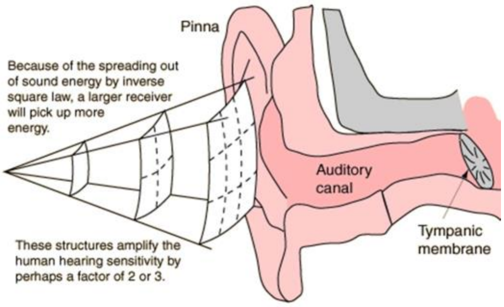
Auditory System

- The cochlea
  - A tube coiled in a snake-shaped spiral
  - Inside filled with gelatinous fluid
  - Running along its length is the Basilar Membrane
  - Along the BM are located approximately 10000 inner hair cells
- Signal transduction
  - Vibrations of the eardrum cause movement in the oval window
  - This causes a compression sound wave in the cochlear fluid
  - This causes vertical vibration of basilar membrane
  - This causes deflections in the inner hair cells, which then fire
- Frequency tuning
  - BM is stiff/thin at basal end (stapes), but compliant/massive at apex
  - Thus, traveling waves peak at different positions along BM
  - As a result, BM can be modeled as a filter bank (video1, video2)

Ear and Hearing



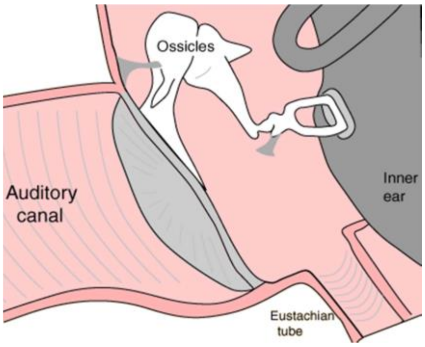
The Outer Ear



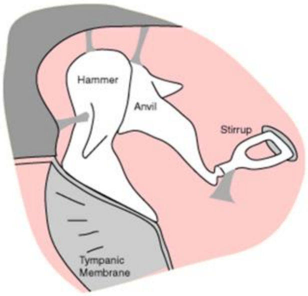
Because of the spreading out of sound energy by inverse square law, a larger receiver will pick up more energy.

These structures amplify the human hearing sensitivity by perhaps a factor of 2 or 3.

The Outer Ear

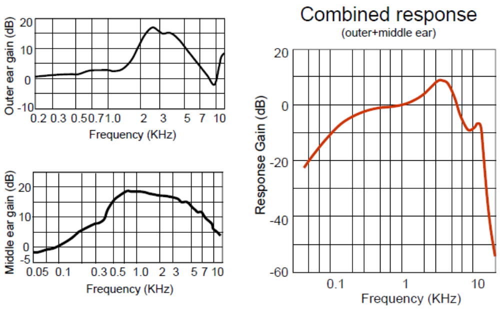


The Middle Ear

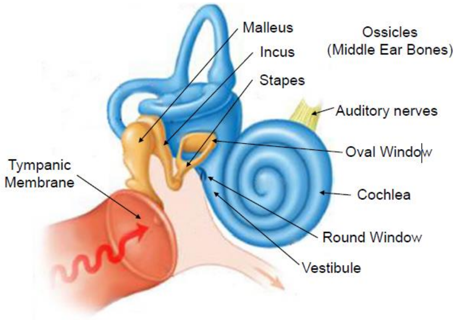


- The Hammer (Malleus), Anvil (Incus) and Stirrup (Stapes) are the three tiniest bones in the body.
  - Together they form the coupling between the vibration of the eardrum and the forces exerted on the oval window of the inner ear.
- These bones can be thought of as a compound lever which achieves a multiplication of force
  - by a factor of about three under optimum conditions.
- They also protect the ear against loud sounds by attenuating the sound.

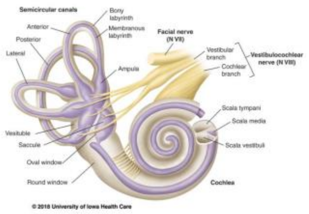
Transfer Functions at the Periphery



The Cochlea

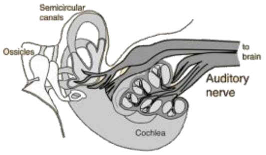


The Inner Ear



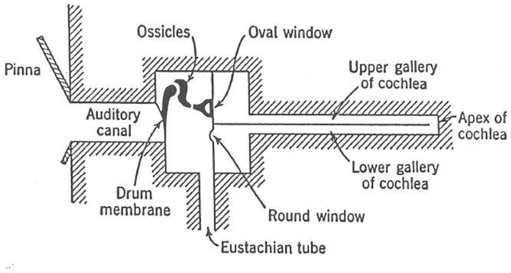
- The inner ear can be thought of as two organs:
  - The semicircular canals
    - which serve as the body's balance organ
  - The cochlea
    - which serves as the body's microphone,
      - converting sound pressure signals from the outer ear into electrical impulses which are passed on to the brain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

The Auditory Nerve

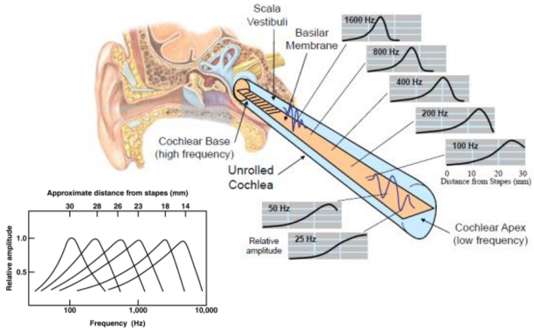


- Taking electrical impulses from the cochlea and the semicircular canals, the auditory nerve makes connections with both auditory areas of the brain.

Schematic Representation of the Ear

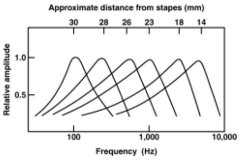


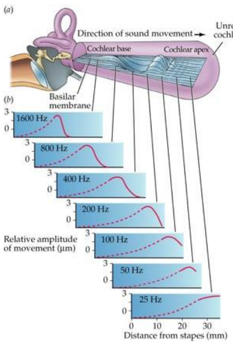
Stretched Cochlea & Basilar Membrane



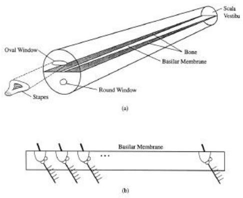
Stretched Cochlea & Basilar Membrane

- Frequency tuning





Basilar Membrane Mechanics



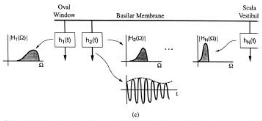
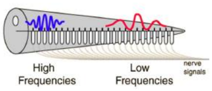


Figure 8.25 Schematic of front-end auditory processing and its model as a wavelet transform: (a) an uncoiled cochlea; (b) the transduction to neural firings of the deflection of hairs that protrude from the in hair cells along the basilar membrane; (c) a signal processing abstraction of the cochlear 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.



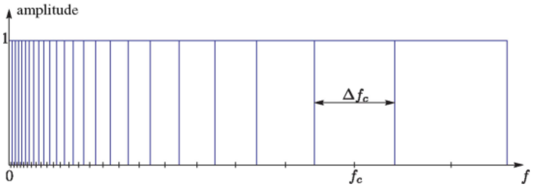
Basilar Membrane Mechanics

- 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 about 3000 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
- 30000 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 nonuniform 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
  - (Video1) (Video2)

Critical Bands



- Idealized basilar membrane filter bank
  - Center Frequency of Each Bandpass Filter: f_c
  - Bandwidth of Each Bandpass Filter:
    $\Delta f_c = 25 + 75[1 + 1.4(f_c/1000)^2]^{0.69}$
  - Real BM filters overlap significantly

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 cannot 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 cannot 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

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
  - The threshold of hearing (I_0) = 10^{-12} watts/m^2
  - The threshold of pain =10 watts/ m^2
- The intensity level of a sound (IL) is defined relative to I_0 as:

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

- 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:

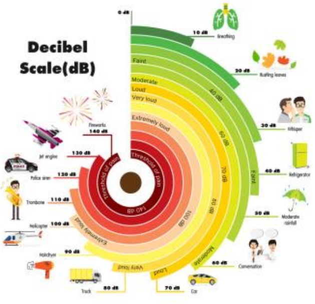
$$SPL = 10log_{10}\left(\frac{P^2}{P_0^2}\right) = 10log_{10}\left(\frac{P}{P_0}\right) \text{ dB}$$

where $P_0 = 2\times10^{-12}$ Newtons/m^2

Some Facts About Human Hearing

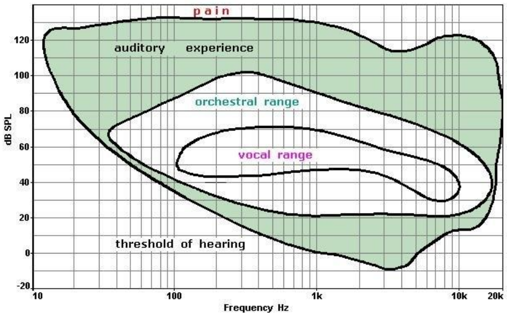
- The range of human hearing is incredible
  - Threshold of hearing
    - thermal limit of Brownian motion of air particles in the inner ear
  - Threshold of pain
    - intensities of from 10^{12} to 10^{16} greater than the threshold of hearing
- Human hearing perceives both sound frequency and sound 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)

Decibel Levels



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
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*

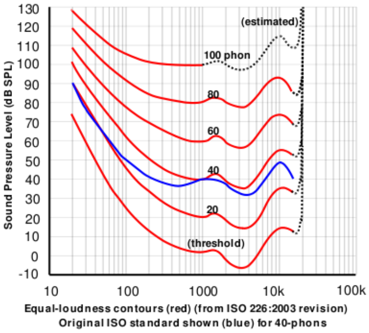
Range of Human Hearing



Hearing Thresholds

- Threshold of Audibility is the acoustic intensity level of a pure tone that can barely be heard at a particular frequency
  - 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
- Maximum sensitivity is at about 3000 Hz

Loudness Level



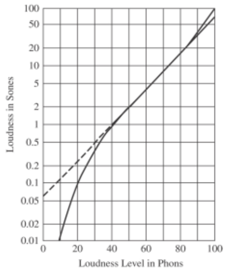
Equal-loudness contours (red) (from ISO 226:2003 revision)
Original ISO standard shown (blue) for 40-phons

- Loudness Level (LL) is equal to the *IL* of a 1000 Hz tone that is judged by the average observer to be equally loud as the tone

Loudness

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

$$logL = 0.033(LL - 40)$$
$$= 0.033LL - 1.32$$



- 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

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

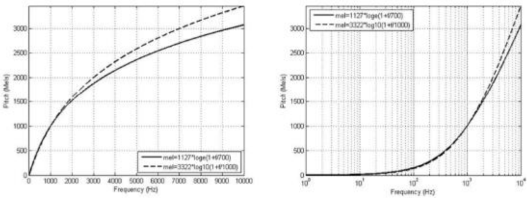
  which reduces to

  $L = 445I^{0.33}$

Pitch

- pitch and fundamental frequency are not the same thing
- We are quite sensitive to changes in pitch
  - F < 500 Hz, $\Delta F \approx 3$ Hz
  - F > 500 Hz, $\Delta F/F \approx 0.003$
- Relationship between pitch and fundamental frequency is not simple, even for pure tones
  - The tone that has a pitch half as great as the pitch of a 200 Hz tone has a frequency of about 100 Hz
  - The tone that has a pitch half as great as the pitch of a 5000 Hz tone has a frequency of less than 2000 Hz
- The pitch of complex sounds is an even more complex and interesting phenomenon

Pitch - The Mel Scale



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

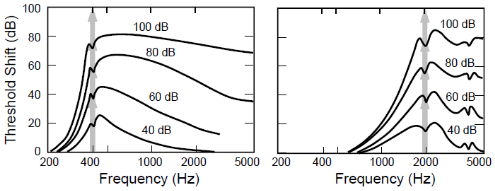
- Alternatively, we can approximate curve as:

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

Perception of Frequency

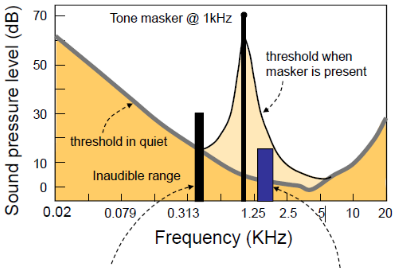
- Pure tone
  - Pitch is a perceived quantity
  - Frequency is a physical quantity (cycle per second or Hertz)
  - Mel is a scale that doubles whenever the perceived pitch doubles;
    - start with 1000 Hz = 1000 mels, increase frequency of tone until listener perceives twice the pitch (or decrease until 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 period is related to time.

Pure Tone Masking



- Masking is the effect whereby some sounds are made less distinct or even inaudible by the presence of other sounds
- Plots above show shift of threshold over non-masking thresholds as a function of the level of the tone masker

Auditory Masking

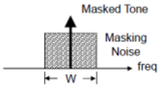


Signal perceptible even in the presence of the tone masker

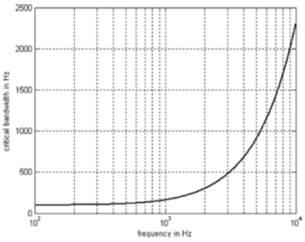
Signal not perceptible due to the presence of the tone masker

Masking & Critical Bandwidth

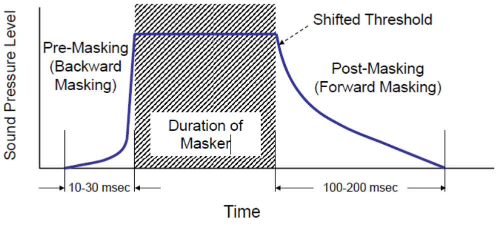
- Critical Bandwidth is the bandwidth of masking noise beyond which further increase in bandwidth has little or no effect on the amount of masking of a pure tone at the center of the band



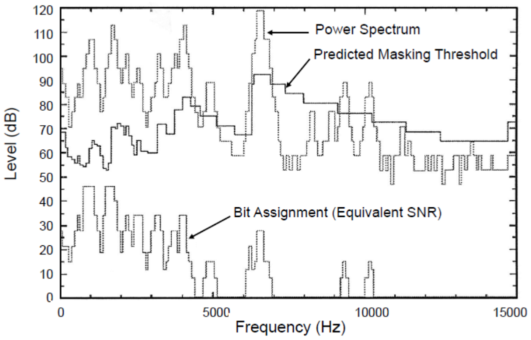
The noise spectrum used is essentially rectangular, thus the notion of equivalent rectangular bandwidth (ERB)



Temporal Masking

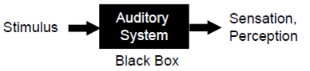


Exploiting Masking in Coding

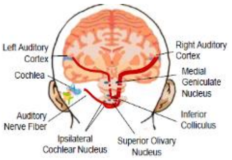


Different Views of Auditory Perception

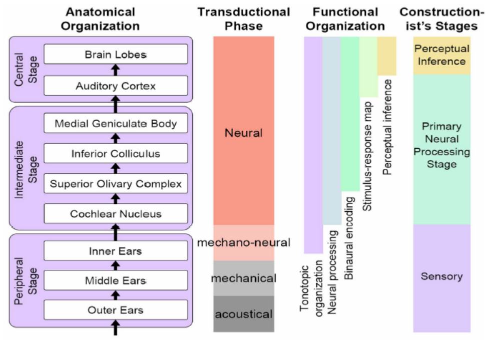
- Functional:
  - based on studies of psychophysics – relates stimulus (physics) to perception (psychology): e.g. frequency in Hz. vs. Mel/Bark scale.


Stimulus
Auditory System
Sensation, Perception
Black Box

- Structural:
  - based on studies of physiology/anatomy
    - how various body parts work with emphasis on the process; e.g. neural processing of a sound
- Auditory System:
  - Periphery:
    - outer, middle, and inner ear
  - Intermediate:
    - CN, SON, IC, and MGN
  - Central:
    - auditory cortex, higherprocessing units



Anatomical & Functional Organizations


Anatomical Organization
Transductional Phase
Functional Organization
Construction-ist's Stages

Central Stage
Brain Lobes
Auditory Cortex

Intermediate Stage
Medial Geniculate Body
Inferior Colliculus
Superior Olivary Complex
Cochlear Nucleus

Peripheral Stage
Inner Ears
Middle Ears
Outer Ears

Neural
mechano-neural
mechanical
acoustical

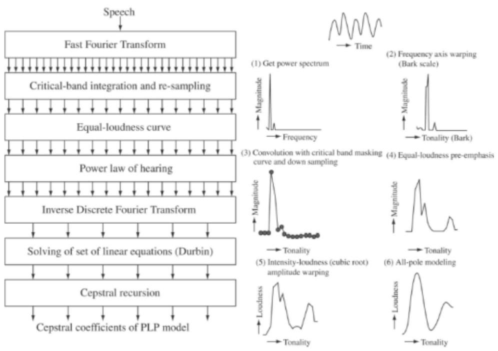
Tonotopic organization
Neural processing
Binaural encoding
Stimulus-response map
Perceptual inference

Perceptual Inference
Primary Neural Processing Stage
Sensory

Auditory Models

- 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)

Perceptual Linear Prediction

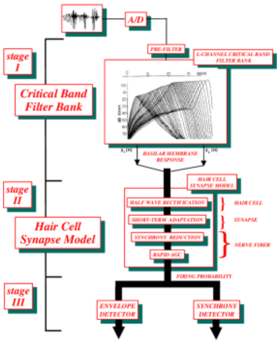

Speech
Fast Fourier Transform
Critical-band integration and re-sampling
Equal-loudness curve
Power law of hearing
Inverse Discrete Fourier Transform
Solving of set of linear equations (Durbin)
Cepstral recursion
Cepstral coefficients of PLP model

Time
(1) Get power spectrum
Magnitude
Frequency
(2) Frequency axis warping (Bark scale)
Magnitude
Tonality (Bark)
(3) Convolution with critical band masking curve and down sampling
Magnitude
Tonality
(4) Equal-loudness pre-emphasis
Magnitude
Tonality
(5) Intensity-loudness (cubic root) amplitude warping
Loudness
Tonality
(6) All-pole modeling
Loudness
Tonality

Perceptual Linear Prediction

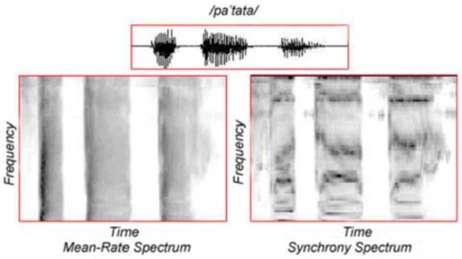
- Included perceptual effects in PLP:
  - Critical band spectral analysis using a Bark frequency scale with variable bandwidth trapezoidal shaped filters
  - Asymmetric auditory filters with a 25 dB/Bark slope at the high frequency cutoff and a 10 dB/Bark slope at the low frequency cutoff
  - Use of the equal loudness contour to approximate unequal sensitivity of human hearing to different frequency components of the signal
  - Use of the non-linear relationship between sound intensity and perceived loudness using a cubic root compression method on the spectral levels
  - A method of broader than critical band integration of frequency bands based on an autoregressive, all-pole model utilizing a fifthorder analysis

Seneff Auditory Model



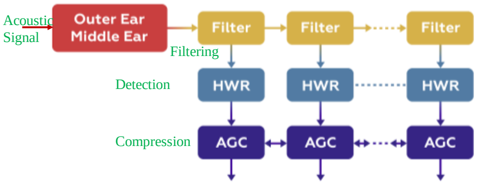
- This model tried to capture essential features of the response of the cochlea and the attached hair cells in response to speech sound pressure waves
- Three stages of processing:
  - Stage 1 pre-filters the speech to eliminate very low and very high frequency components, and then uses a 40-channel 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

Seneff Auditory Model



- Segmentation into well defined onsets and offsets (for each stop consonant in the utterance) is seen in the Mean-Rate Spectrum;
- Speech resonances clearly seen in the Synchrony Spectrum.

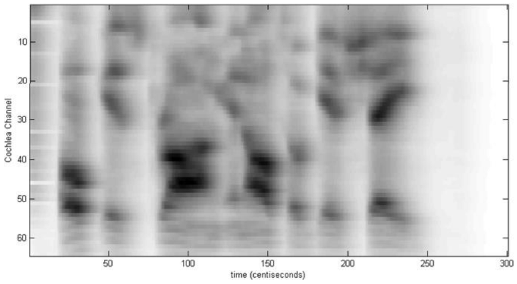
Lyon’s Cochlear Model



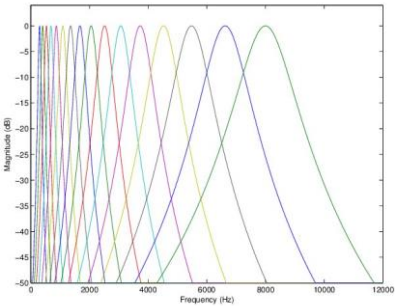
- Pre-processing stage (simulating effects of outer and middle ears as a simple pre-emphasis network)
  - Three full stages of processing for modeling the cochlea as a non-linear filter bank
  - First stage is a bank of 86 cochlea filters, space non0uniformly according to mel or Bark scale, and highly 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.

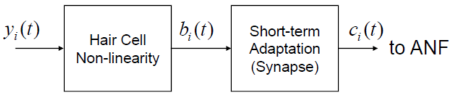


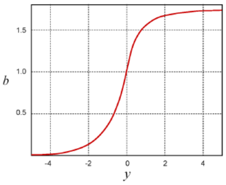
Gammatone Filter Bank Model for Inner Ear



- A popular auditory filter model:
- The magnitude responses (in dB) of 16 gammatone filters in the frequency range 300-8000 Hz are represented on a linear frequency scale.

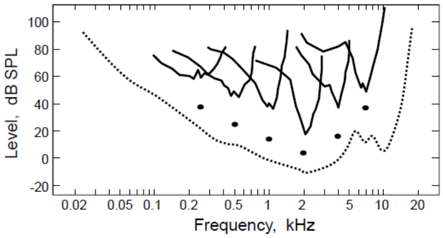
Inner Hair Cell Model





$$\frac{dc_i(t)}{dt} = \begin{cases} \alpha\left[b_i(t) - c_i(t)\right] - \beta c_i(t), & b_i(t) > c_i(t) \\ -\beta c_i(t), & b_i(t) \le c_i(t) \end{cases}$$

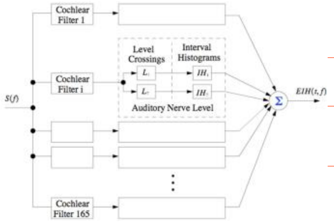
Psychophysical Tuning Curves (PTC)



- Each of the psychophysical tuning curves (PTCs) describes the simultaneous masking of a low intensity signal by sinusoidal maskers with variable intensity and frequency.
- PTCs are similar to the tuning curves of the auditory nerve fibers (ANF).

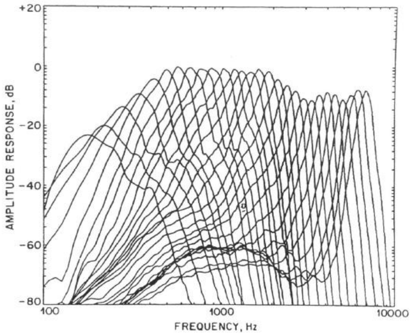
Ensemble Interval Histogram (EIH)

- 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



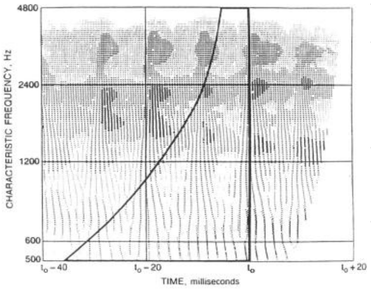
- 165 channels, equally spaced on a log frequency scale between 150 and 7000 Hz
  - cochlear filter designs match neural tuning curves for cats
    - minimum phase filters
  - array of level crossing detectors that model motion-to-neural activity transduction of the IHCs
  - detection levels are pseudo-randomly distributed to match variability of fiber diameters

Cochlear Filter Designs



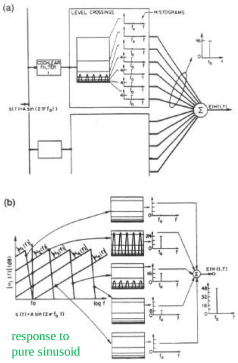
EIH Responses

- 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

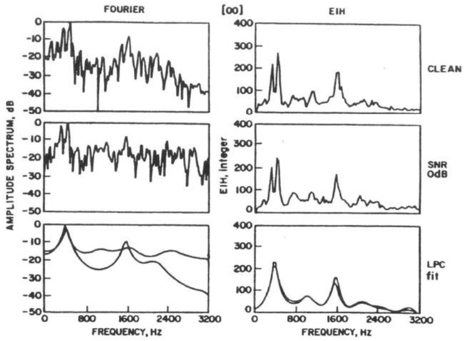
Overall EIH



response to pure sinusoid

- 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

EIH Robustness to Noise



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

Summary of Auditory Processing

- Human hearing ranges
- Speech communication model
  - from production to perception
- Black box models of hearing/perception
- The human ear
  - outer, middle, inner
- Mechanics of the basilar membrane
- The ear as a frequency analyzer
- The Ensemble Interval Histogram (EIH) model

Lecture Summary

- The ear acts as a sound canal, transducer, spectrum analyzer
- The cochlea acts like a multi-channel, logarithmically spaced, constant Q filter bank
- Frequency and place along the basilar membrane are represented by inner hair cell transduction to events (ensemble intervals) that are processed by the brain
  - This makes sound highly robust to noise and echo
- Hearing has an enormous range from threshold of audibility to threshold of pain
  - Perceptual attributes scale differently from physical attributes
    - e.g., loudness, pitch
- Masking enables tones or noise to hide tones or noise
  - this is the basis for perceptual coding (MP3)
- Perception and intelligibility are tough concepts to quantify
  - but they are key to understanding performance of speech processing systems