L2: Speech production and perception

Anatomy of the speech organs

Models of speech production

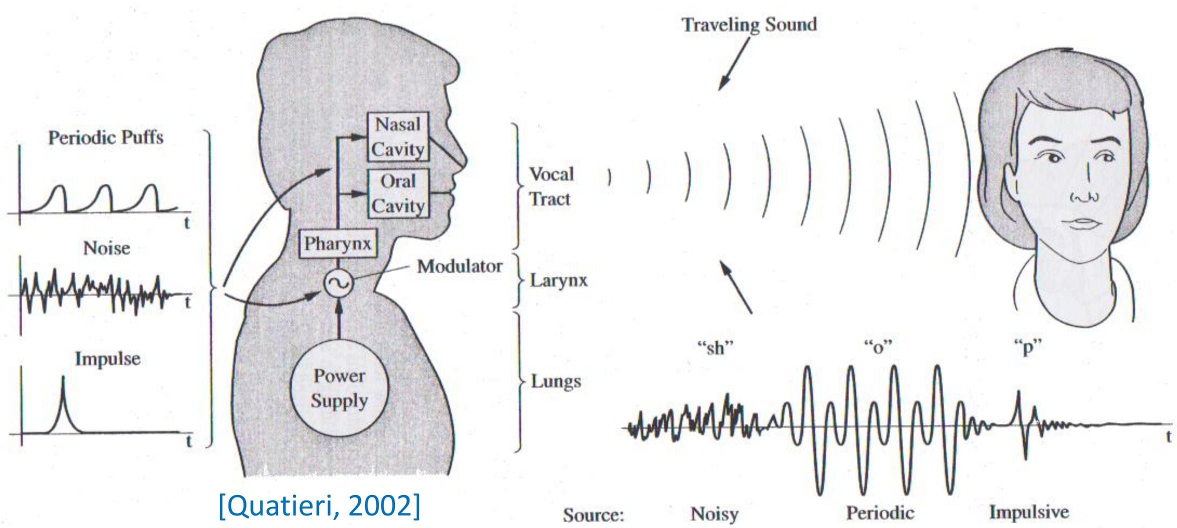
Anatomy of the ear

Auditory psychophysics

Anatomy of the speech organs

The speech organs can be broadly divided into three groups

- Lungs: serve as a "power supply" and provides airflow to the larynx
- Vocal chords (Larynx): modulate the airflow into either a periodic sequence of puffs or a noisy airflow source
  - A third type of source is impulsive
  - Exercise, say the word "shop" and determine where each sound occurs
- Vocal tract: converts modulated airflow into spectrally "colored" signal

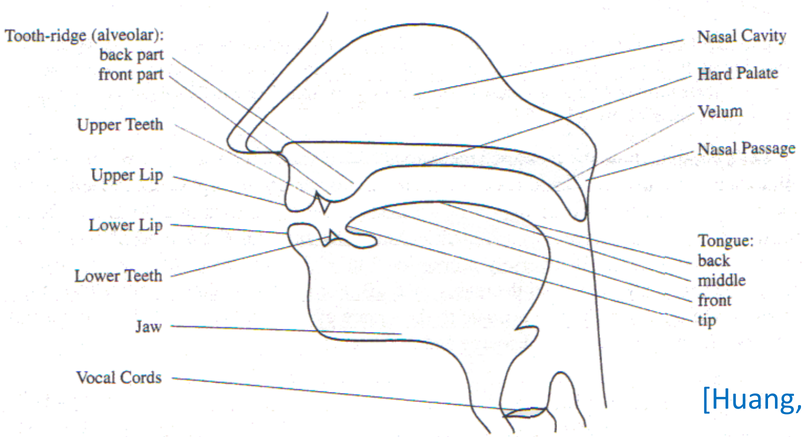


[Quatieri, 2002]

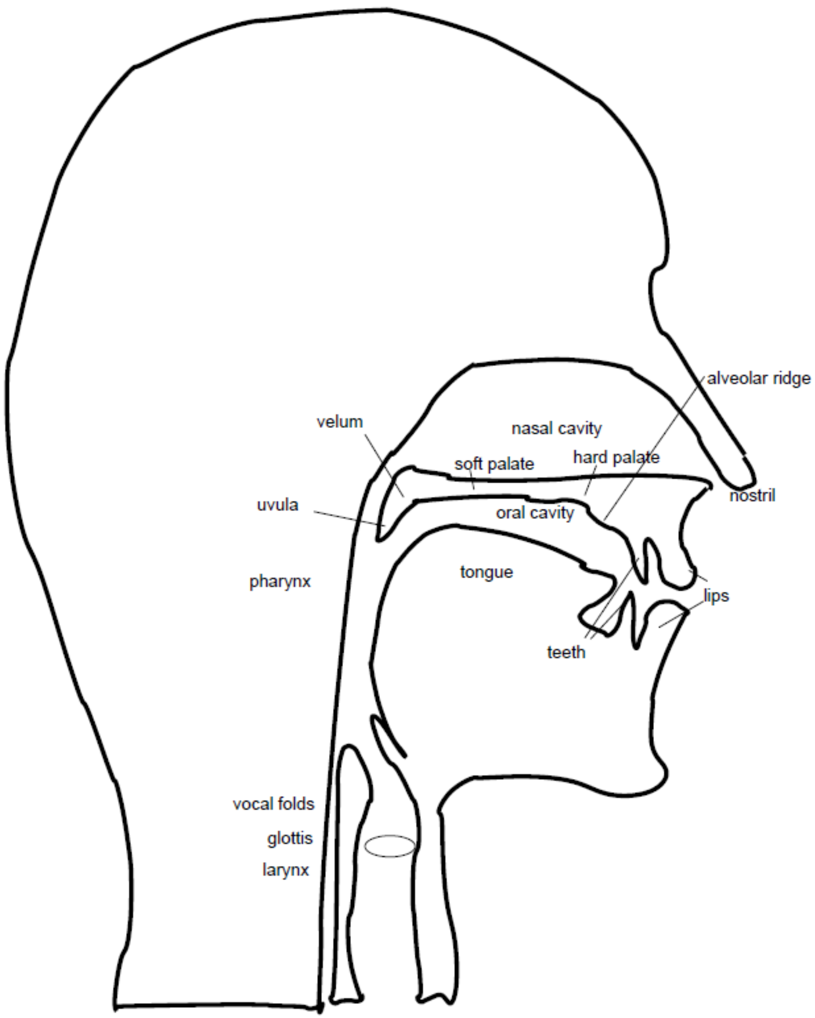
The vocal tract

The vocal tract can further be divided into

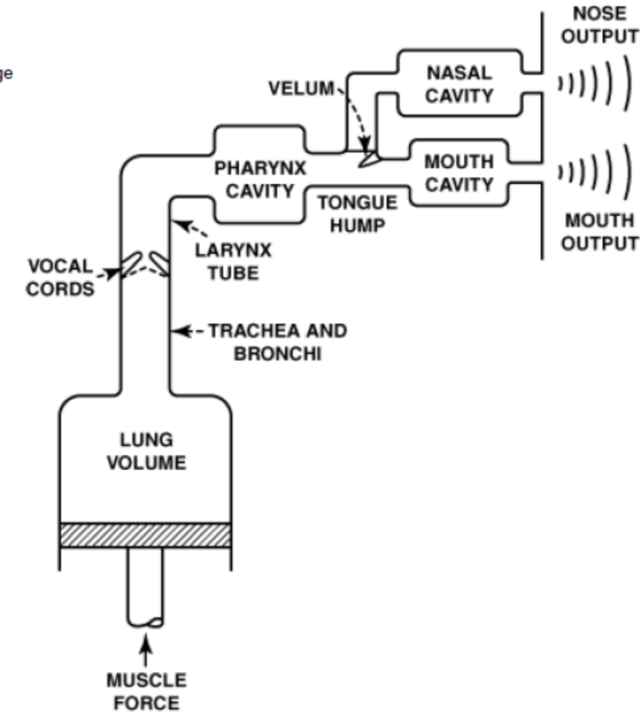
- Velum (soft palate): controls airflow through the nasal cavity. In its open position is used for “nasals” (i.e., [n], [m]).
- Hard palate: hard surface at the roof of the mouth. When tongue is pressed against it, leads to consonants
- Tongue: Away from the palate produces vowels; close to or pressing the palate leads to consonants
- Teeth: used to brace the tongue for certain consonants
- Lips: can be rounded or spread to shape consonant quality, or closed completely to produce certain consonants (i.e., [p], [b], [m])



[Huang, Acero & Hon, 2001]



(a) mid-sagittal drawing of vocal organs



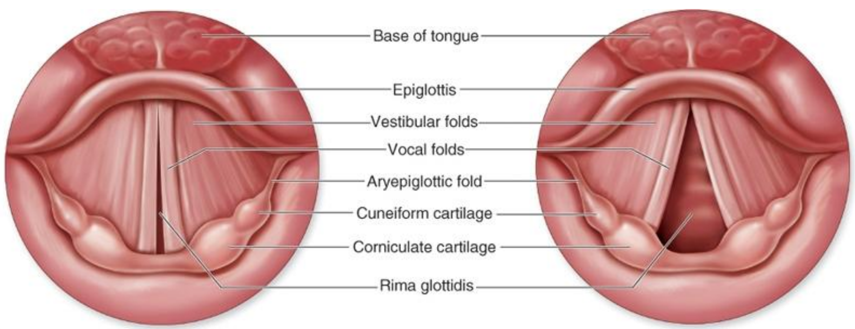
(b) Model of vocal organs with discrete components identified

[Taylor, 2009]

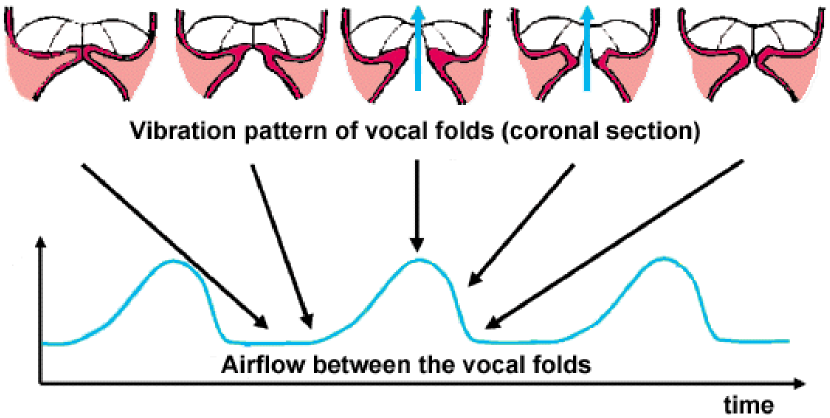
The vocal folds

Two masses of flesh, ligament and muscle across the larynx

- Fixed at the front of the larynx but free to move at the back and sides
- Can be in one of three primary states
  - Breathing: Glottis is wide, muscles are relaxed, and air flows with minimal obstruction
  - Voicing: vocal folds are tense and are brought up together. Pressure builds up behind, leading to an oscillatory opening of the folds (video)
  - Unvoiced: similar to breathing state, but folds are closer, which leads to turbulences (i.e. aspiration, as in the sound [h] in 'he') or whispering



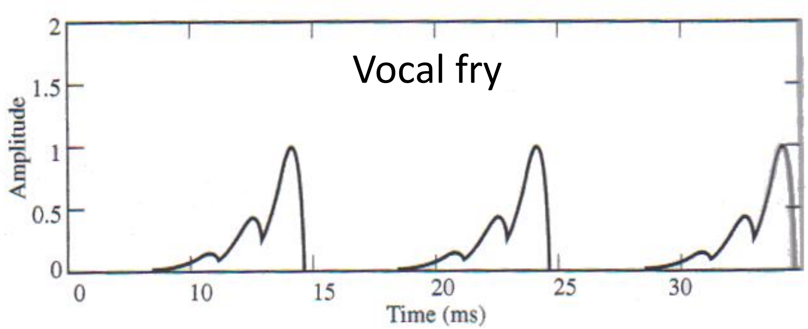
http://academic.kellogg.edu/herbrandsonc/bio201_mckinley/f25-5b_vocal_folds_lary_c.jpg

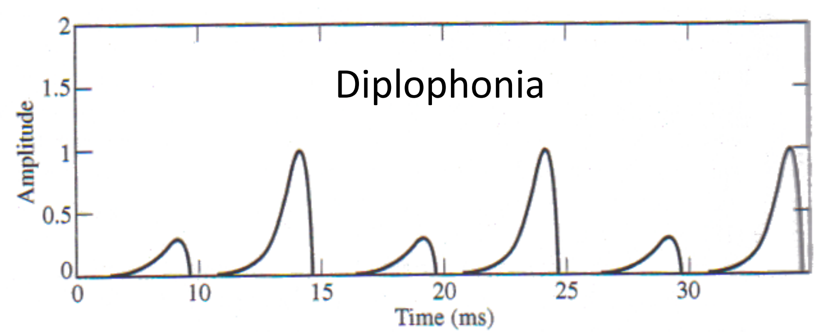


http://biorobotics.harvard.edu/research/heather2.gif

Other (minor) forms of voicing include

- Hoarse voice: voicing period (pitch) jitters, as what results from laryngitis or a cold
- Breathy voice: aspiration occurs simultaneously while voicing (audio)
- Creaky voice: vocal folds are very tense and only a portion oscillates. Result is a harsh sounding voice (audio)
- Vocal fry: folds are very relaxed, which leads to secondary glottal pulses (video)
- Diplophonic: secondary pulses occur, but during the closed phase





[Quatieri, 2002]

Models of speech production

Acoustic theory of speech production

– Speech occurs when a source signal passing through the glottis is modified by the vocal tract acting as a filter
– Models of this kind are generally known as *source-filter* models
– Using the theory of linear time invariant (LTI) systems, the overall process can be modeled in the z-domain (see lecture 4) as

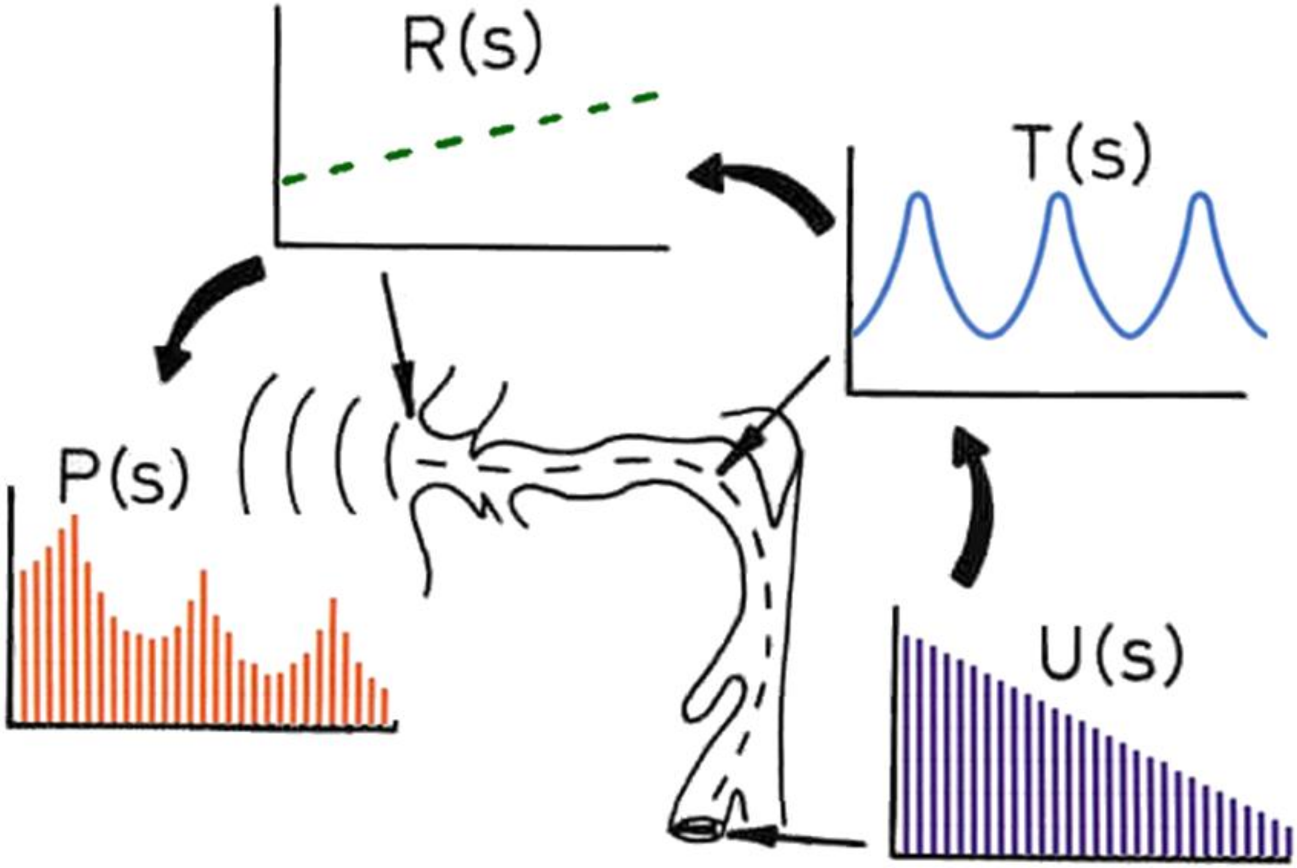
$$Y(z) = U(z)P(z)O(z)R(z)$$

  - where U(z) is the glottal source, and P(z), O(z), R(z) are the transfer functions at the pharynx, oral cavity and lips

– which can be simplified as

$$Y(z) = U(z)V(z)R(z)$$

  - by combining P(z) and O(z) into a single vocal-tract transfer function, which represents the filter component of the model



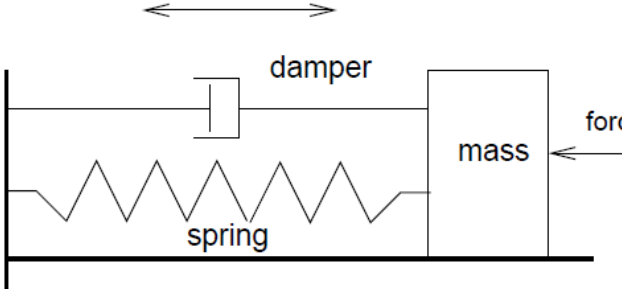
$$P(s) = U(s)\ T(s)\ R(s)$$

The physics of sounds

Resonant systems

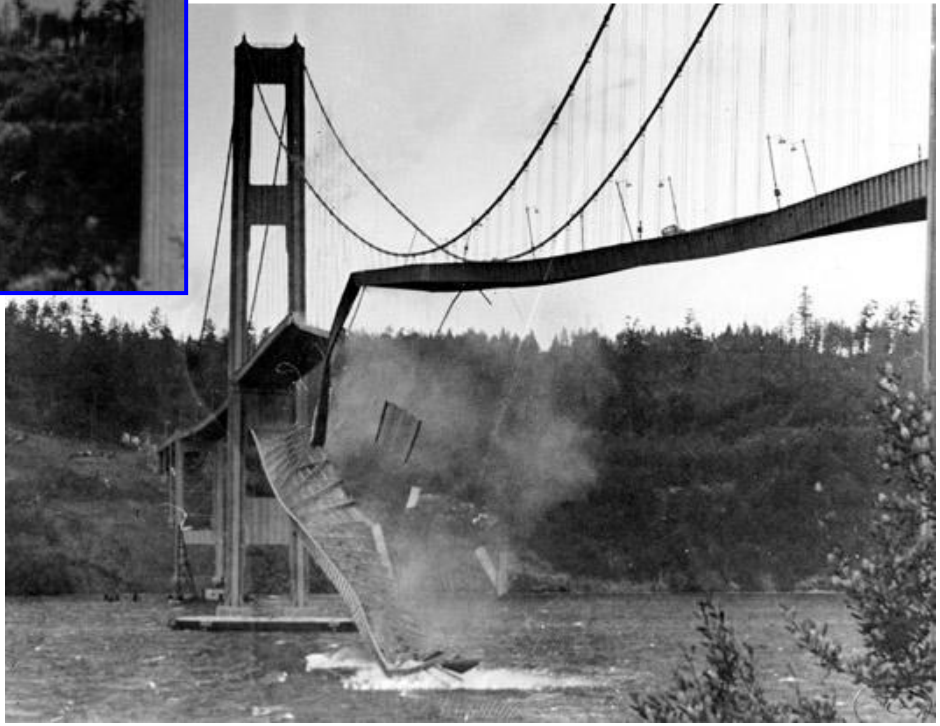
- Consider the mass spring shown below
  - If you displace the mass, the system will try to return to its rest position
  - In the process, it will lead to oscillations around the rest position
  - Due to frictions, the mass will eventually settle onto the rest position
- Now consider a periodic forcing function being applied
  - At a certain frequency f_R, the size of the oscillations will increase over time rather than decrease
  - Eventually, and in the absence of other factors, the system will break
  - Frequency f_R is known as the resonant frequency of the system

$$f_R = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$$



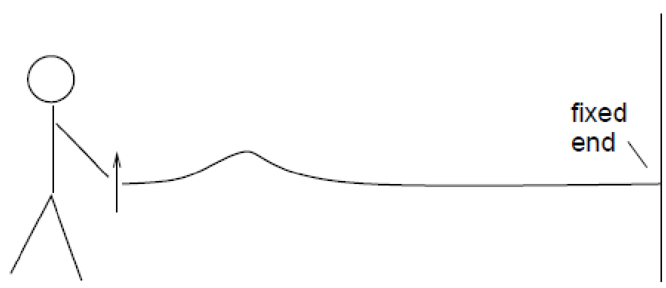
[Taylor, 2009]

The Tacoma Narrows Bridge

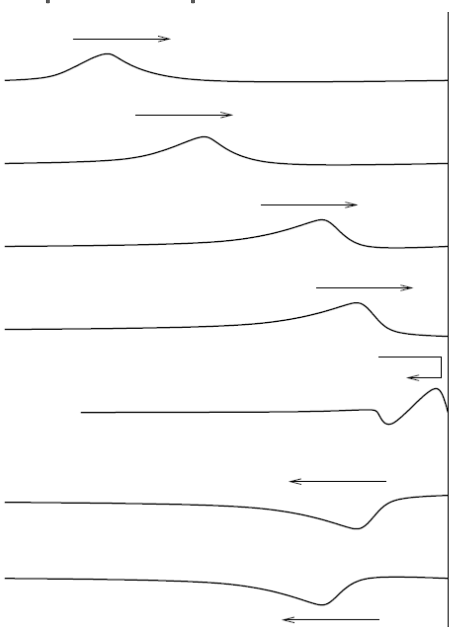


Travelling and standing waves

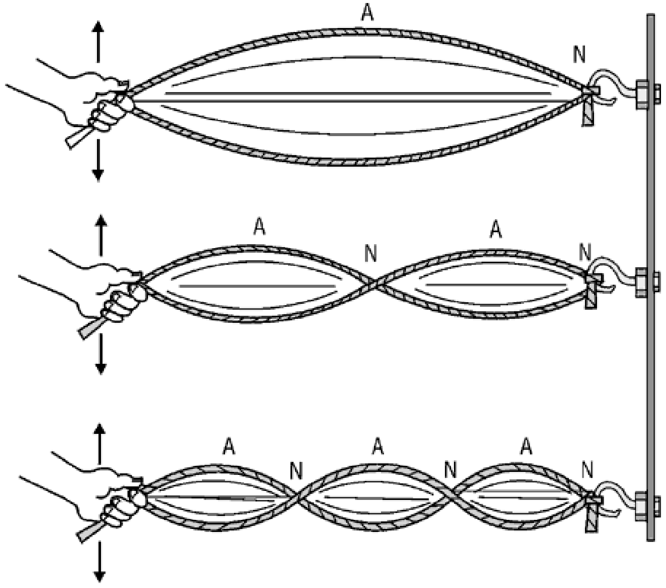
– Consider a person holding a rope fixed to a wall on the opposite end
  - The person gives a jerk to the rope, and as a result a wave forms and starts travelling down the rope
  - When the rope reaches the wall, it is reflected and begins to travel back towards the person
  - When it reaches the person, the wave is reflected back towards the wall
  - This process goes on until all energy in the rope dissipates



[Taylor, 2009]



– Now consider the person performs a repetitive movement
  - The forward and backward wave will interfere, which may lead to full cancellation (if the two waves are in anti-phase)
  - At some frequencies, the reflected wave will reinforce the forward wave, and the rope settles into a fixed pattern
  - The resulting wave will appear not to be moving at all (a standing wave)
  - Thus the rope acts as a resonator: it amplifies some waves and attenuates others

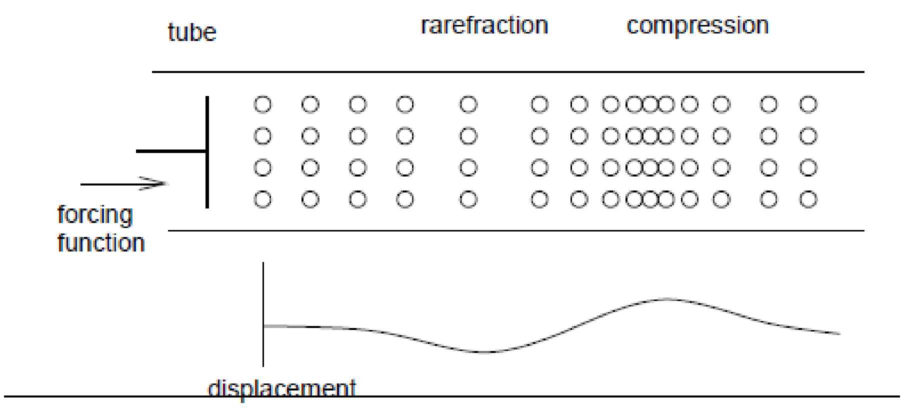


http://www.cccc.edu/instruction/slympany/ELN/236/Mod7/loet01-07-06new.gif

- What determines the behavior of the system?
  - The frequency of the oscillations is determined entirely by the hand
  - The rate of travel of the wave is determined entirely by the rope
  - Boundary conditions: whether the rope is fixed or free at each end.
- What is the relationship with the speech signal?
  - The hand acts as the source (the glottal pulses)
  - The rope acts as the filter (the vocal tract)

Acoustic waves

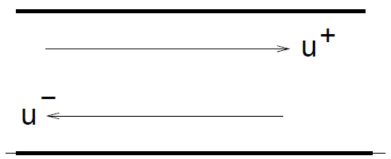
– Properties of sound waves traveling in a tube (i.e., vocal tract) are similar to those moving down a rope
  - The effect of a sound source causes air particles to move back and forth, so the wave spreads from the source
    – In some areas, particles come close together (compression) whereas in others they move further apart (rarefaction).
  - One difference is that sound waves are longitudinal whereas those in a rope are transverse, but otherwise the same mathematical model can be used for both systems



[Taylor, 2009]

Acoustic reflection and tube models

- As with the rope, boundary conditions in the tube will determine how acoustic waves are reflected at the end of the tube
  - At certain frequencies, determined by the length of the tube and the speed of sound, the backward and forward waves will reinforce each other and cause resonances
- We can model the volume velocity (e.g., particle velocity × area) at position *x* and time *t* as:



[Taylor, 2009]

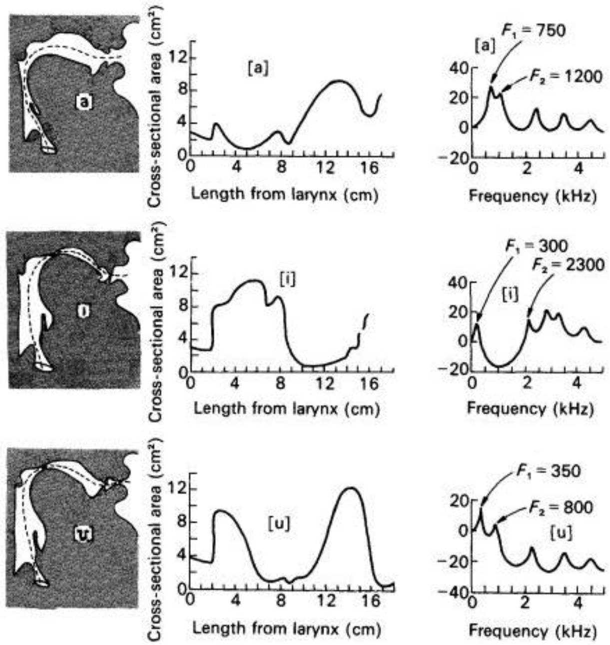
$$u(x,t) = u^+(t - x/c) - u^-(t + x/c)$$

  - where $u^+(t)$ and $u^-(t)$ are the forward- and backward-travelling waves, and c is the speed of sound
- And the pressure becomes

$$p(x,t) = \frac{\rho c}{A}\left(u^+(t - x/c) + u^-(t + x/c)\right)$$

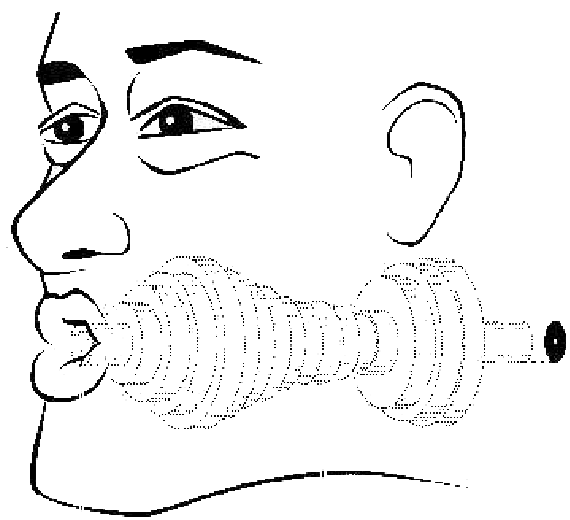
  - where $\rho c/A$ is the characteristic <u>impedance</u> of the tube
  - Notice how in this case the two waves add up as they meet at point x

– If the area of the tube remains constant, the wave propagates through the tube
– However, if the area changes, then the impedance changes, which causes reflection, which in turn leads to standing waves, which then cause resonances
– Thus, the impedance pattern on the tube determines the resonance properties of the model

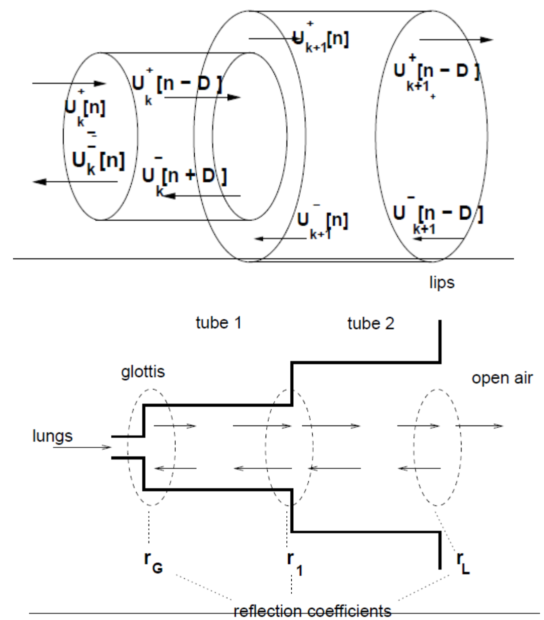


http://www.livingcontrolsystems.com/festschrift/nevin_files/image015.jpg

– The vocal tract can then be modeled as a series of short uniform tubes connected in series
  - By increasing the number of tubes, the vocal tract can be modeled with arbitrary accuracy
  - As we will see in linear predictive coding (lecture 7), a tube model with N sections leads to N/2 resonances, so in practice only a few tube sections are needed to model the main formants in the speech signal



http://www.gregandmel.net/burnett_thesis/2_3.ht7.jpg



[Taylor, 2009]

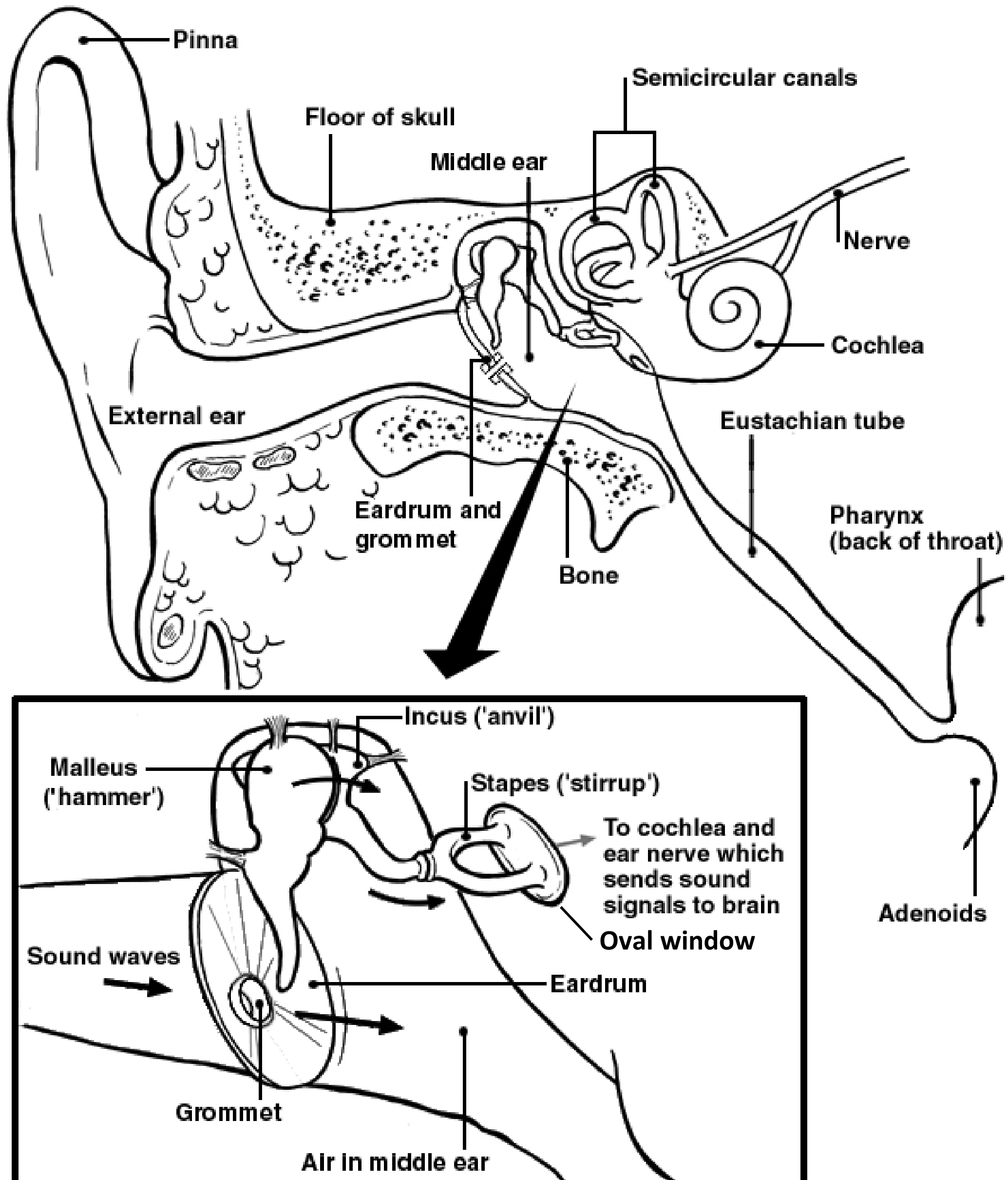
Anatomy of the ear

There are two major components in auditory system

- 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
- We will focus on the peripheral auditory organ

The ear can be further divided into

- 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



http://www.bissy.scot.nhs.uk/master_code/pilsinl/042.gif

The cochlea

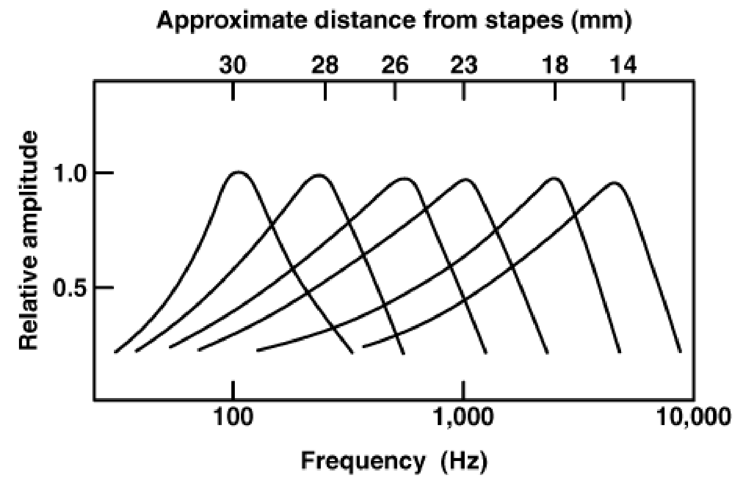
- A tube coiled in a snake-shaped spiral
- Inside filled with gelatinous fluid
- Running along its length is the basilar membrane
- Along the BM are located approx. 10,000 inner hair cells

Signal transduction

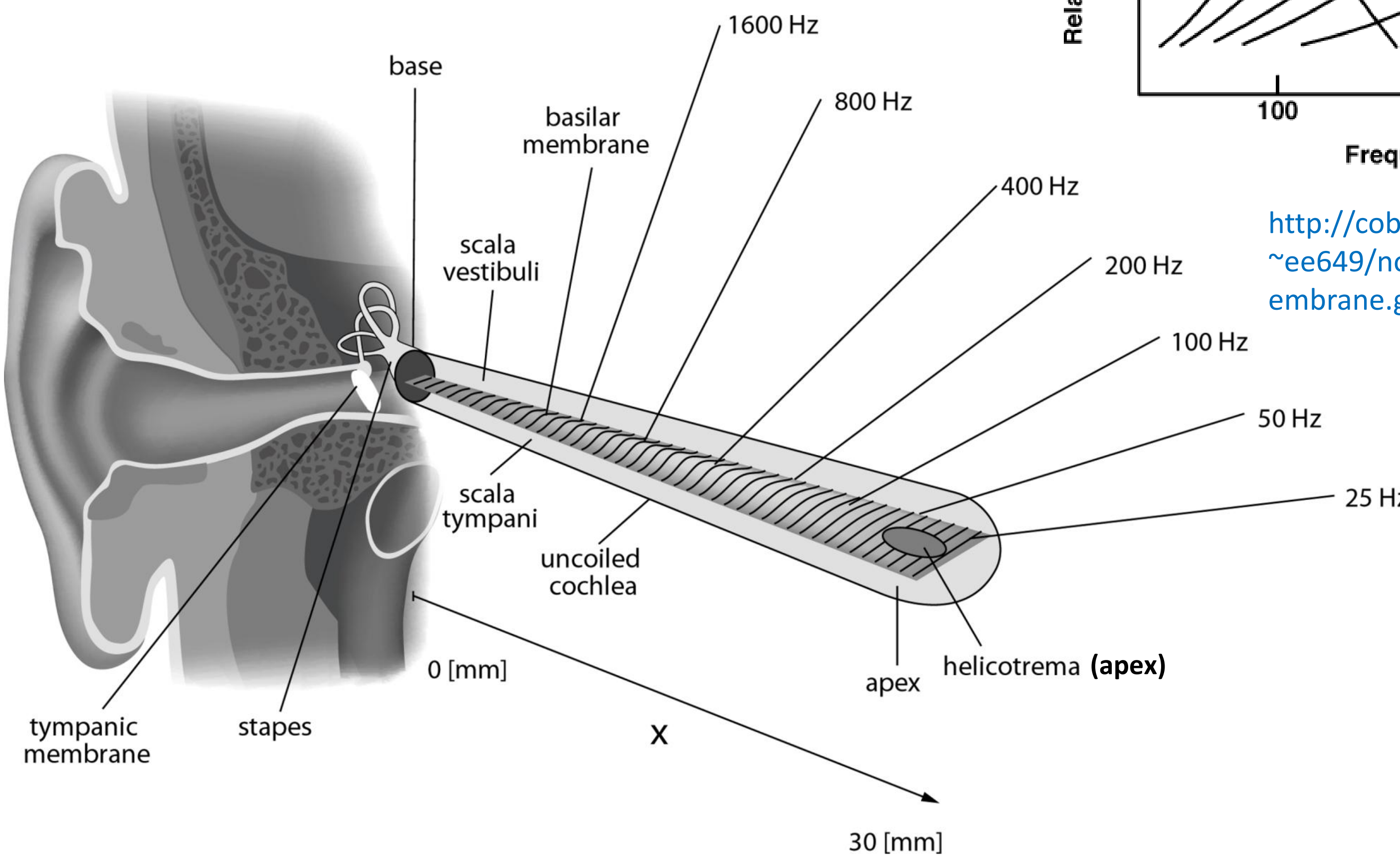
- Vibrations of the eardrum cause movement in the oval window
- This causes a compression sound wave in the cochlear fluid
- This causes vertical vibration of basilar membrane
- This causes deflections in the inner hair cells, which then fire

Frequency tuning

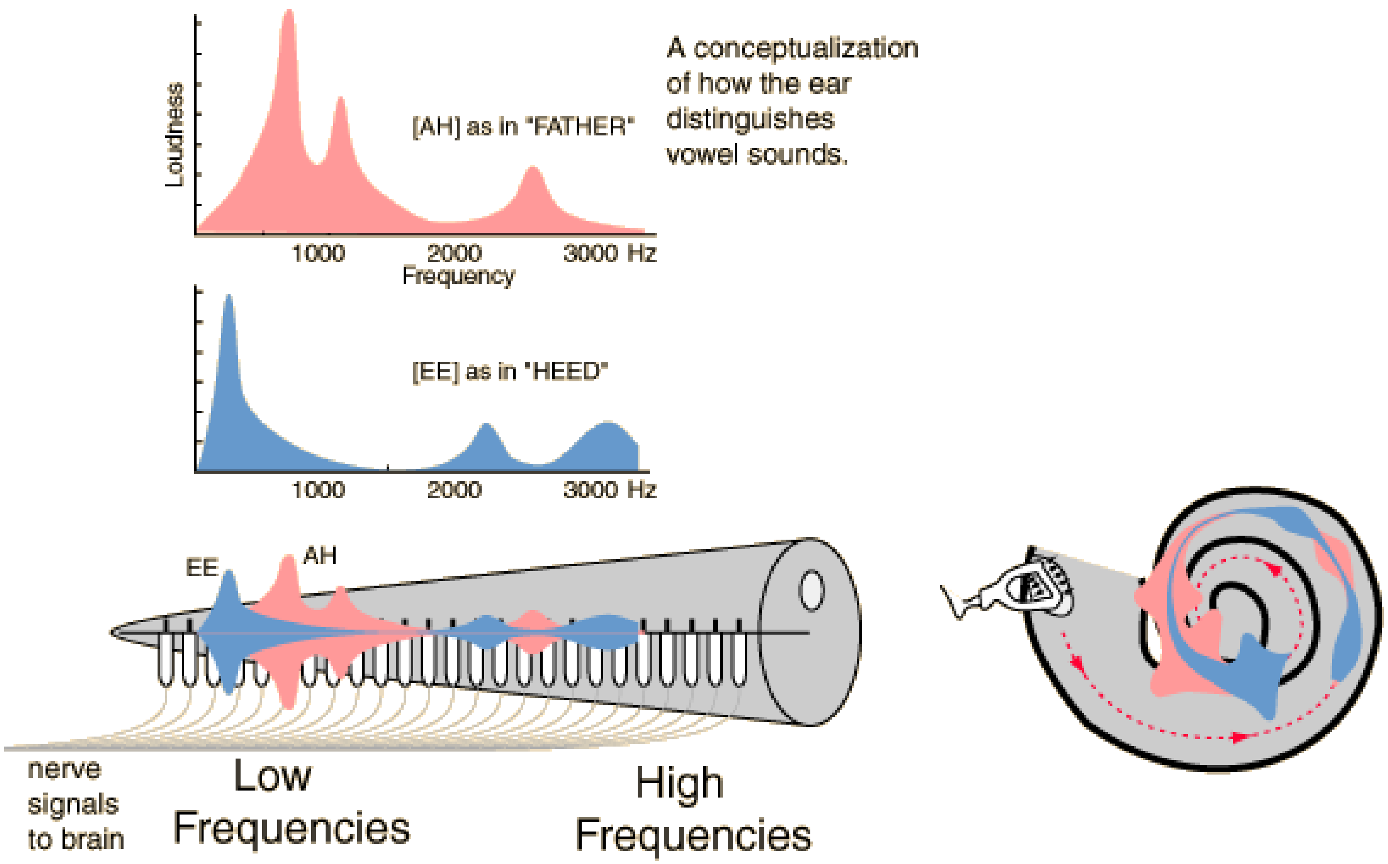
- BM is stiff/thin at basal end (stapes), but compliant/massive at apex
- Thus, traveling waves peak at different positions along BM
- As a result, BM can be modeled as a filter bank (video)



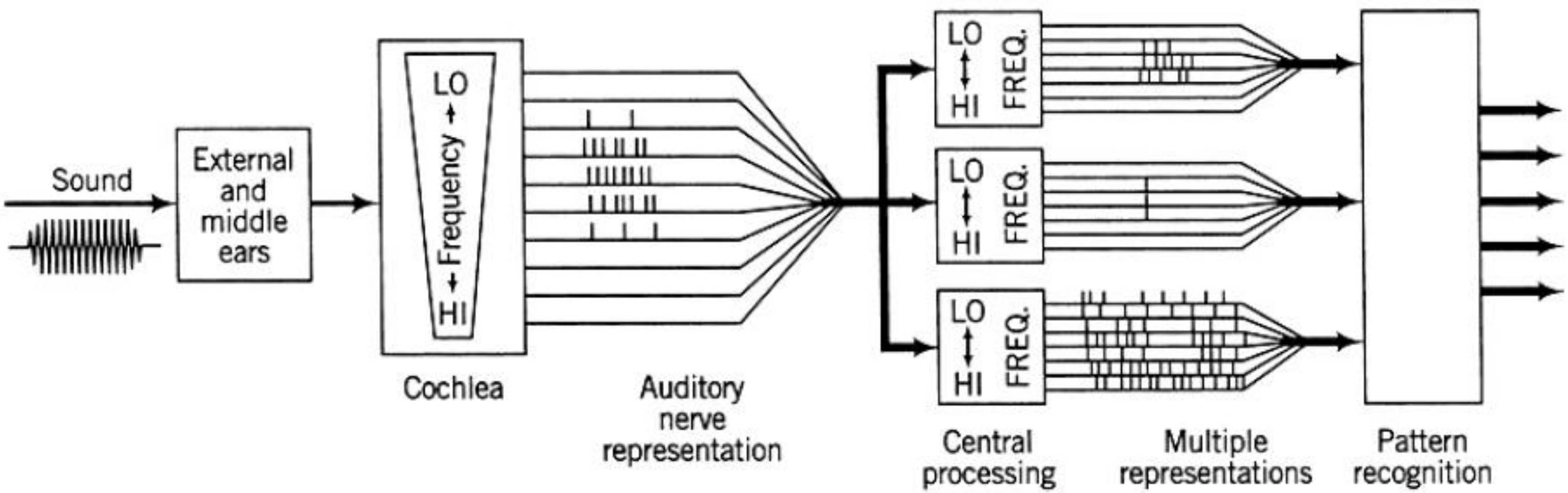
http://cobweb.ecn.purdue.edu/~ee649/notes/figures/basilar_membrane.gif



http://upload.wikimedia.org/wikipedia/commons/6/65/Uncoiled_cochlea_with_basilar_membrane.png



http://hyperphysics.phy-astr.gsu.edu/hbase/sound/cochimp.html

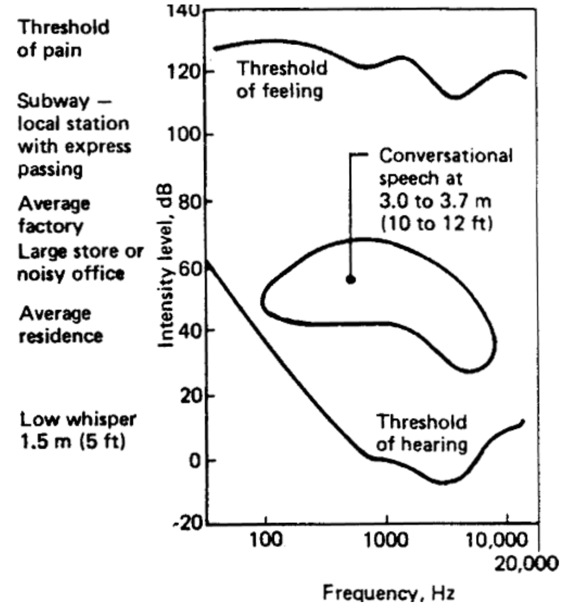


[Rabiner & Schafer, 2007]

Auditory psychophysics

Psychoacoustics is concerned with quantitative modeling of human auditory perception

– How does the ear respond to different intensities and frequencies?
– How well does it focus on a sound of interest in the presence of interfering sounds?



http://msis.jsc.nasa.gov/images/Section04/Image126.gif

Thresholds

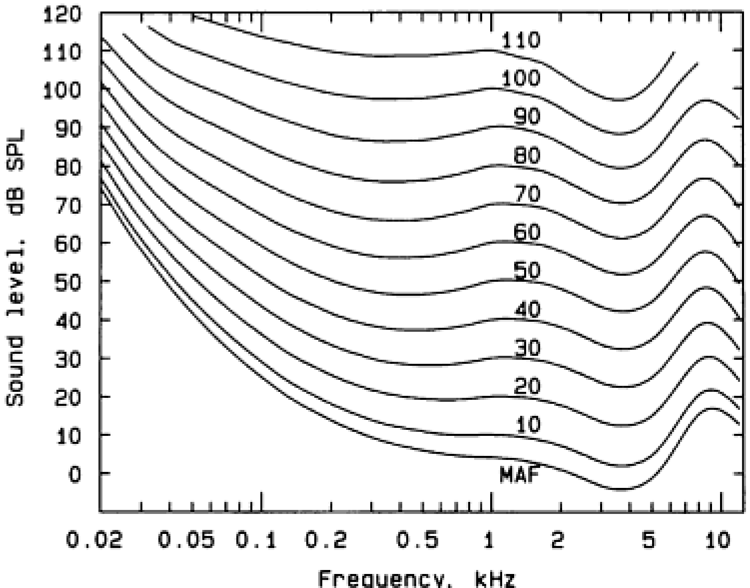
– The ear is capable of hearing sounds in the range of 16Hz to 18kHz
– Intensity is measured in terms of sound pressure levels (SPL) in units of decibels (dB)
– Hearing threshold: Minimum intensity at which a sound is perceived
  - Sounds below 1kHz or above 5kHz have increasingly higher thresholds
  - Threshold is nearly constant across most speech frequencies (700Hz-7kHz)

SPL and loudness

- As with other sensory systems (seeing, smelling), auditory sensations increase logarithmically with the intensity of the stimulus
- The relation between sound pressure *p*, sound intensity *I* and loudness S follows Steven's power law

$$S \propto p^{0.6} \propto I^{0.3}$$

  - where the unit of S is the *sone*, and the proportionality constant implied by the equation is frequency dependent
- The ear is most sensitive to tones around 4kHz
  - Each loudness contour corresponds to a unit of a phons (the SPL in dB of a 1kHz tone)



http://replaygain.hydrogenaudio.org/pics/equal_loudness.gif

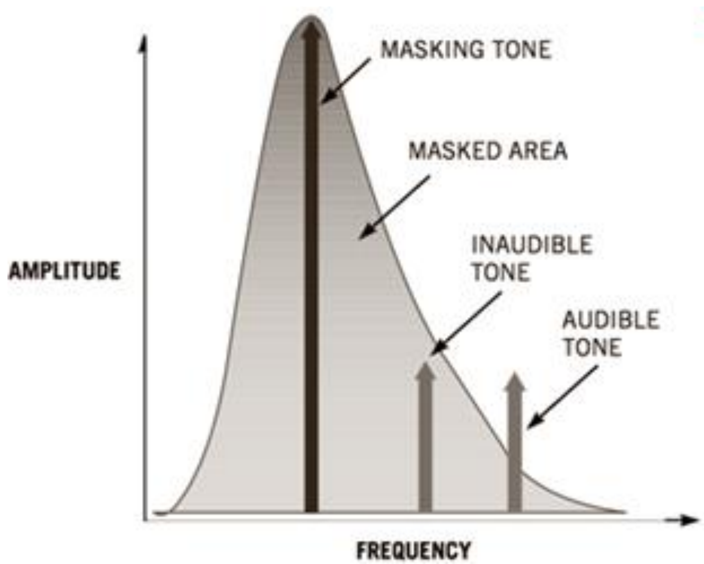
Masking

- A phenomenon whereby the perception of a sound is obscured by the presence of another (i.e., the latter raises the threshold of the former)
- Masking is the major non-linear phenomenon that prevents treating the perception of speech sounds as a summation of responses

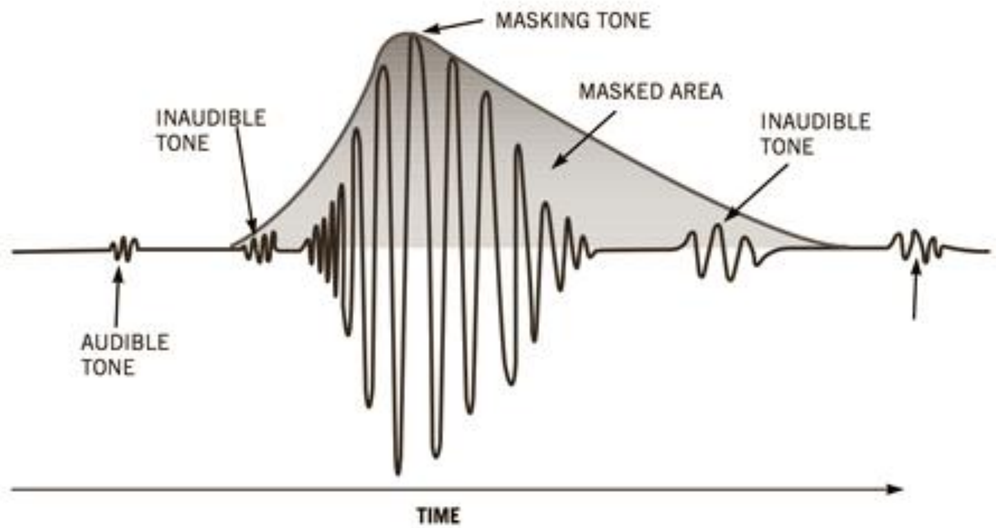
Two types of masking phenomena

- Frequency masking
  - A lower frequency sound generally masks a higher frequency one
  - Leads to the concept of critical bands (next)
- Temporal masking
  - Sounds delayed wrt one another can cause masking of either sound
  - Pre-masking tends to last 5ms; post-masking can last up to 50-300ms

Frequency masking



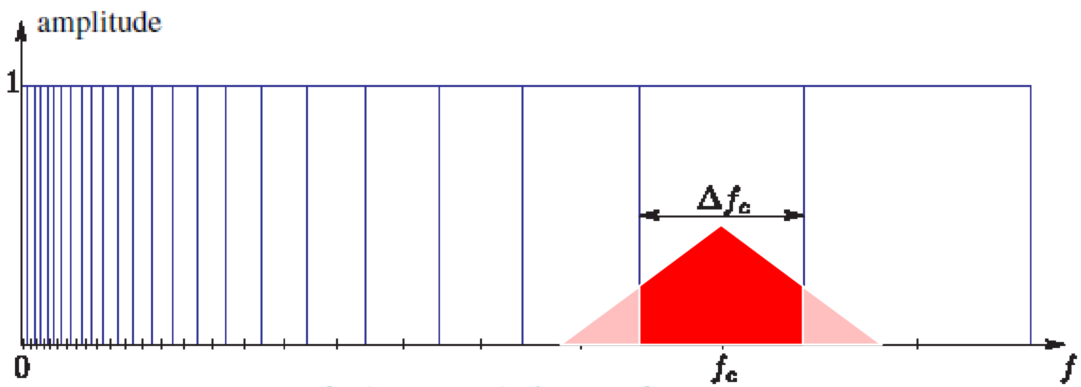
Temporal masking



http://homepage.mac.com/marc.heijligers/audio/ipod/compression/encoding/encoding.html

Critical bands

- For a given frequency, the critical band is the smallest band of frequencies around it which activate the same part of the BM
  - Critical bandwidths correspond to about 1.5 mm spacing along the BM
  - This suggests that a set of 24 bandpass filters (with increasing bandwidth with frequency) would model the BM well
- If a signal and masker are presented simultaneously, only the masker frequencies within the CB contribute to masking of the signal
  - The amount of masking is equal to the total energy of the masker within the CB of the probe



[Rabiner & Schafer, 2007]

How can you test a critical band experimentally?

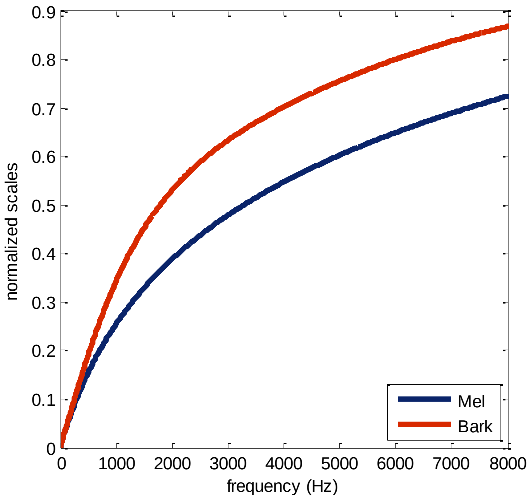
- Take a band-limited noise signal with a center frequency of 2 kHz, and play it alongside a sinusoidal 2 kHz tone
- Make the tone very quiet relative to the noise
  - You will not be able to detect the tone because the noise signal will mask it
  - Now, turn up the level of the tone until you can hear it and write down its level
- Increase the bandwidth of the noise (w/o turning up its level) and repeat
  - You'll find that your threshold for detecting the tone will be higher
  - In other words, if the bandwidth of the masking signal is increased, you have to turn up the tone more in order to be able to hear it
- Increase the bandwidth and do the experiment over and over
  - As you increase the bandwidth of the masker, the detection threshold of the tone will increase up to a certain bandwidth. Then it won't increase any more!
  - This means that, for a given frequency, once you get far enough away in frequency, the noise does not contribute to the masking of the tone
- The bandwidth at which the threshold for the detection of the tone stops increasing is the critical bandwidth

http://www.tonmeister.ca/main/textbook/node331.html

Two perceptual scales have been derived from critical bands

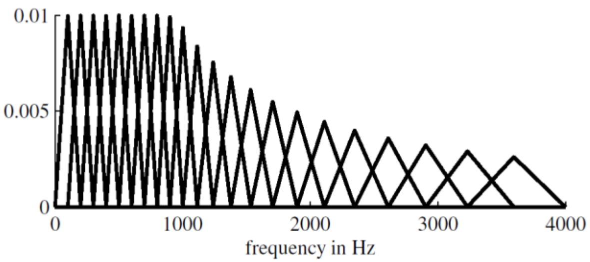
- Bark scale
  - Relates acoustic frequency to perceptual frequency resolution
  - One Bark equals one critical band

$$z = 13tan^{-1}\left(0.76\frac{f}{kHz}\right) + 3.5tan^{-1}\left(\frac{f}{7.5kHz}\right)$$



- Mel scale (more on lecture 9)
  - Linear mapping up to 1 kHz, then logarithmic at higher frequencies

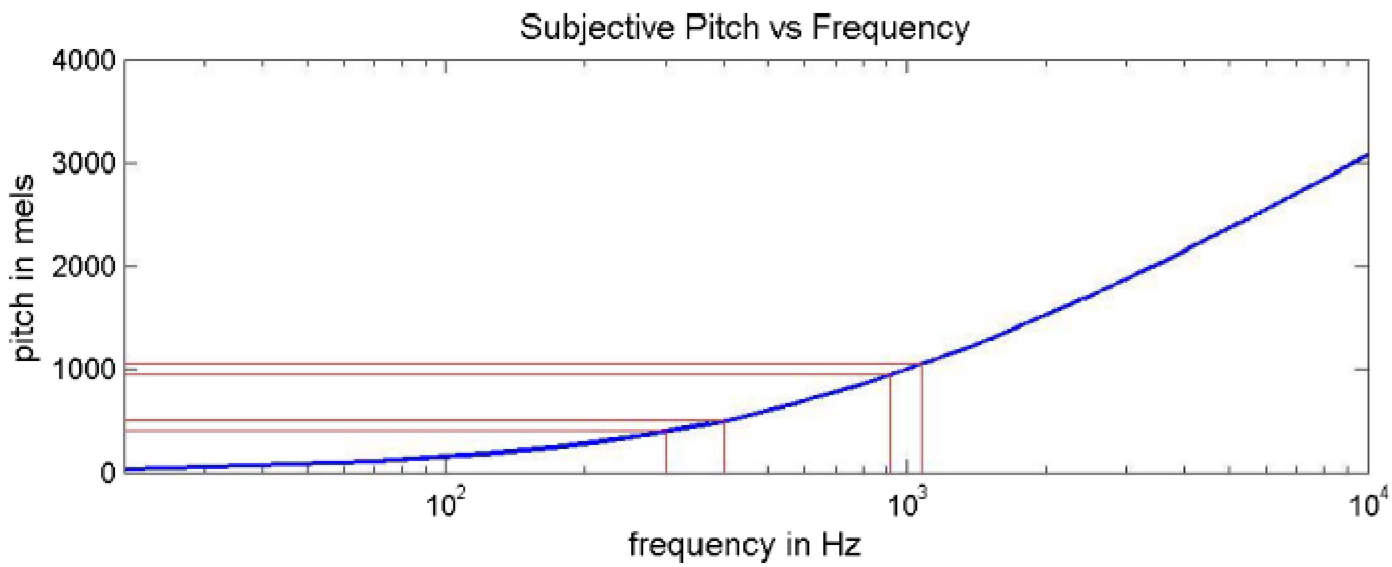
$$m = 2595log_{10}(1 + f/700)$$



[Rabiner & Schafer, 2007]

Pitch perception

– Like loudness, pitch is a subjective attribute, in this case related to the fundamental frequency (F0) of a periodic signal
– The relationship between pitch and F0 is non linear and can be described by the Mel scale



[Rabiner & Schafer, 2007]