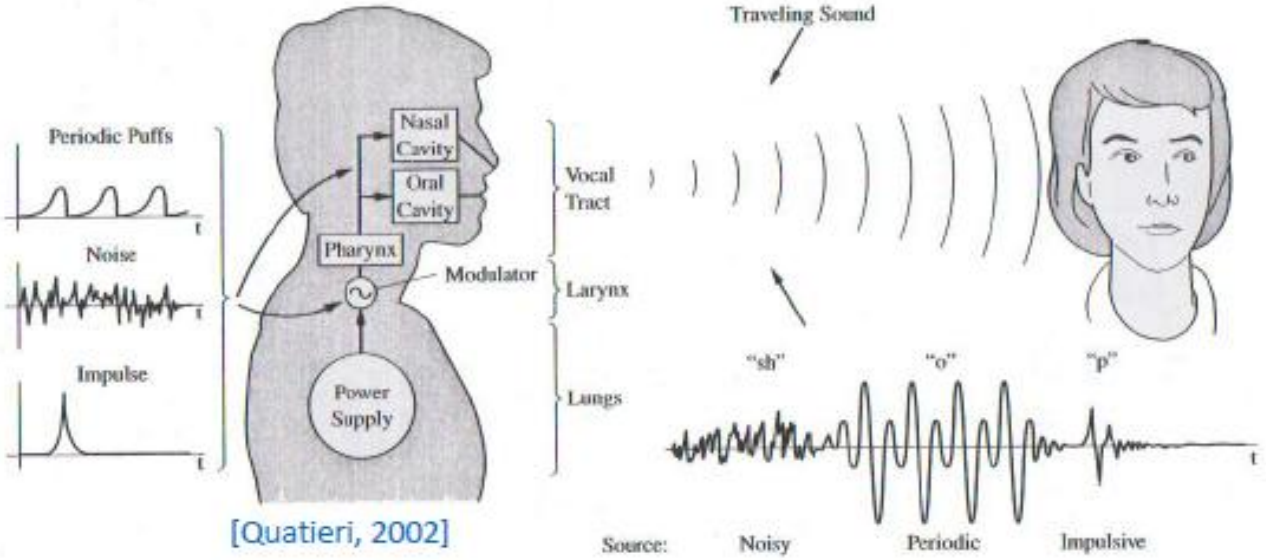
Speech production and perception

- **Anatomy of the speech organs**
- **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

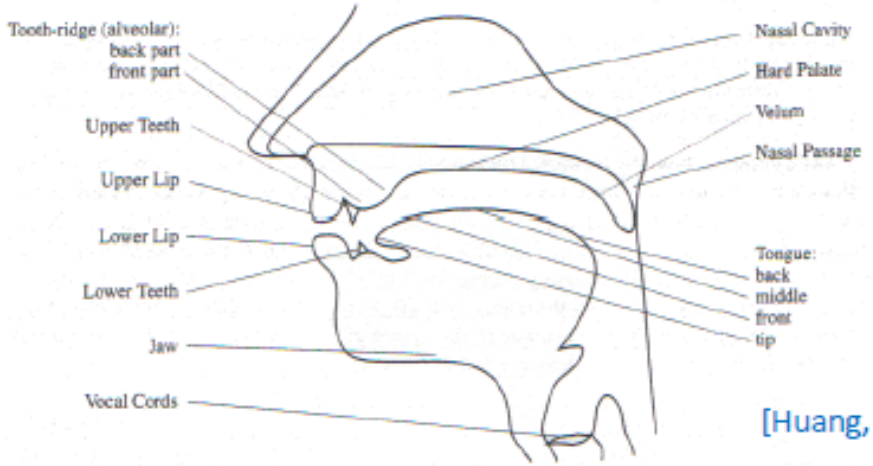


[Quatieri, 2002]

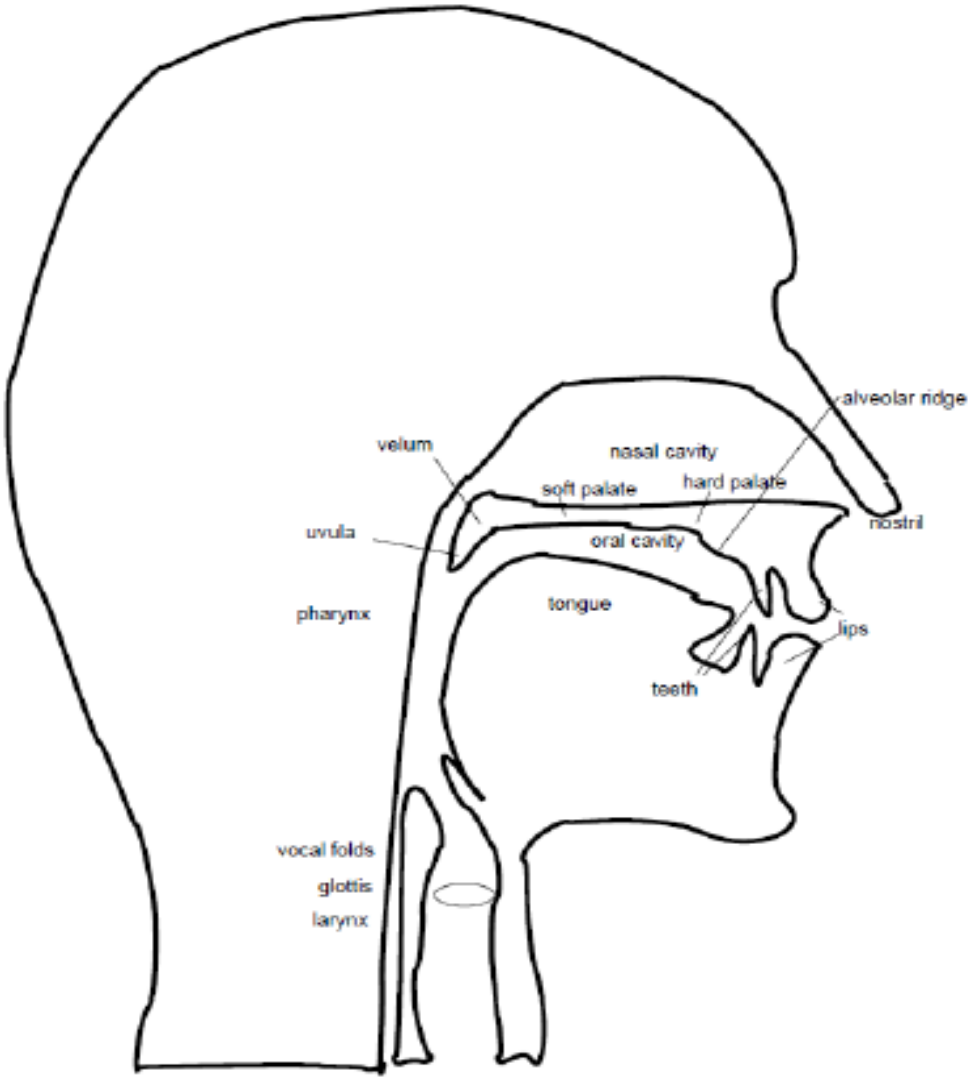
The vocal tract

The vocal tract can further be divided into

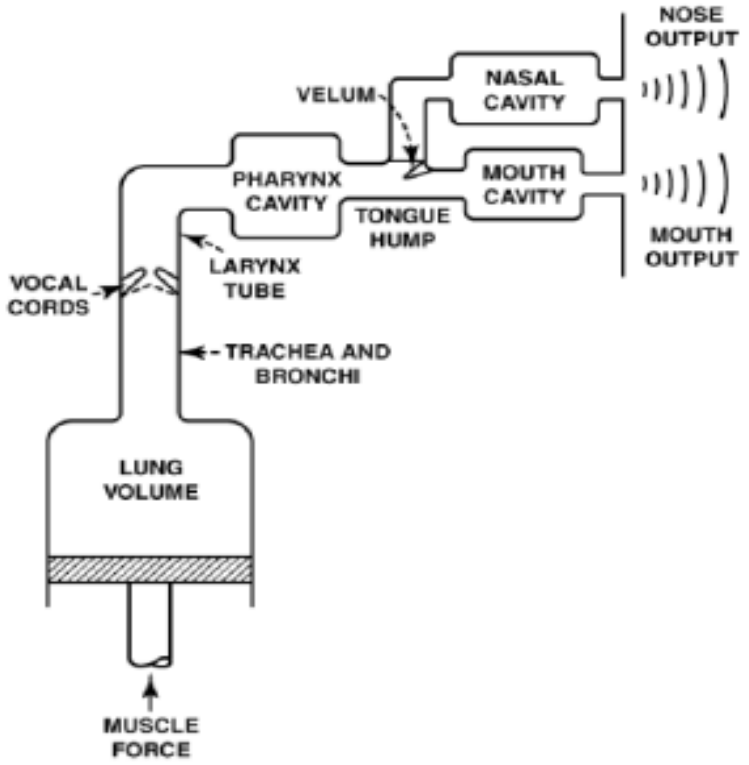
- Velum (soft palate): controls airflow through the nasal cavity. In its open position is used for "nasals" (i.e., [n], [m]).
- Hard palate: hard surface at the roof of the mouth. When tongue is pressed against it, leads to consonants
- Tongue: Away from the palate produces vowels; close to or pressing 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



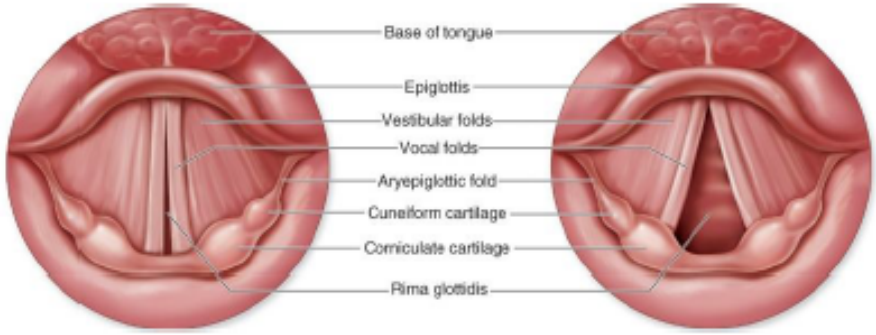
(b) Model of vocal organs with discrete components identified

[Taylor, 2009]

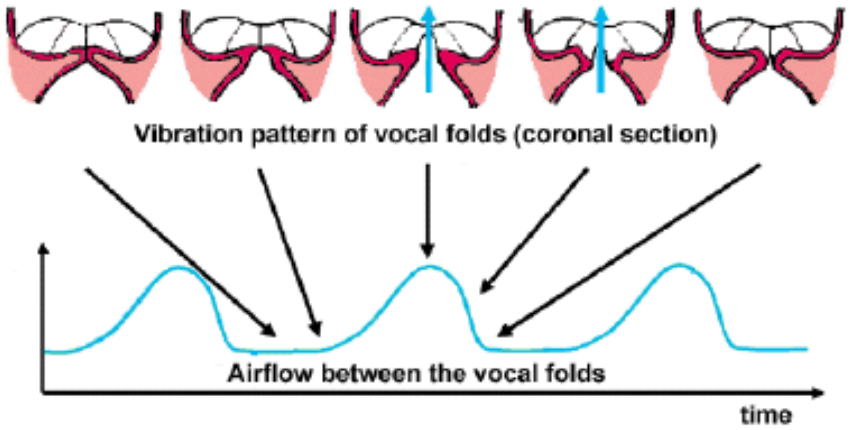
The vocal folds

Two masses of flesh, ligament and muscle across the larynx

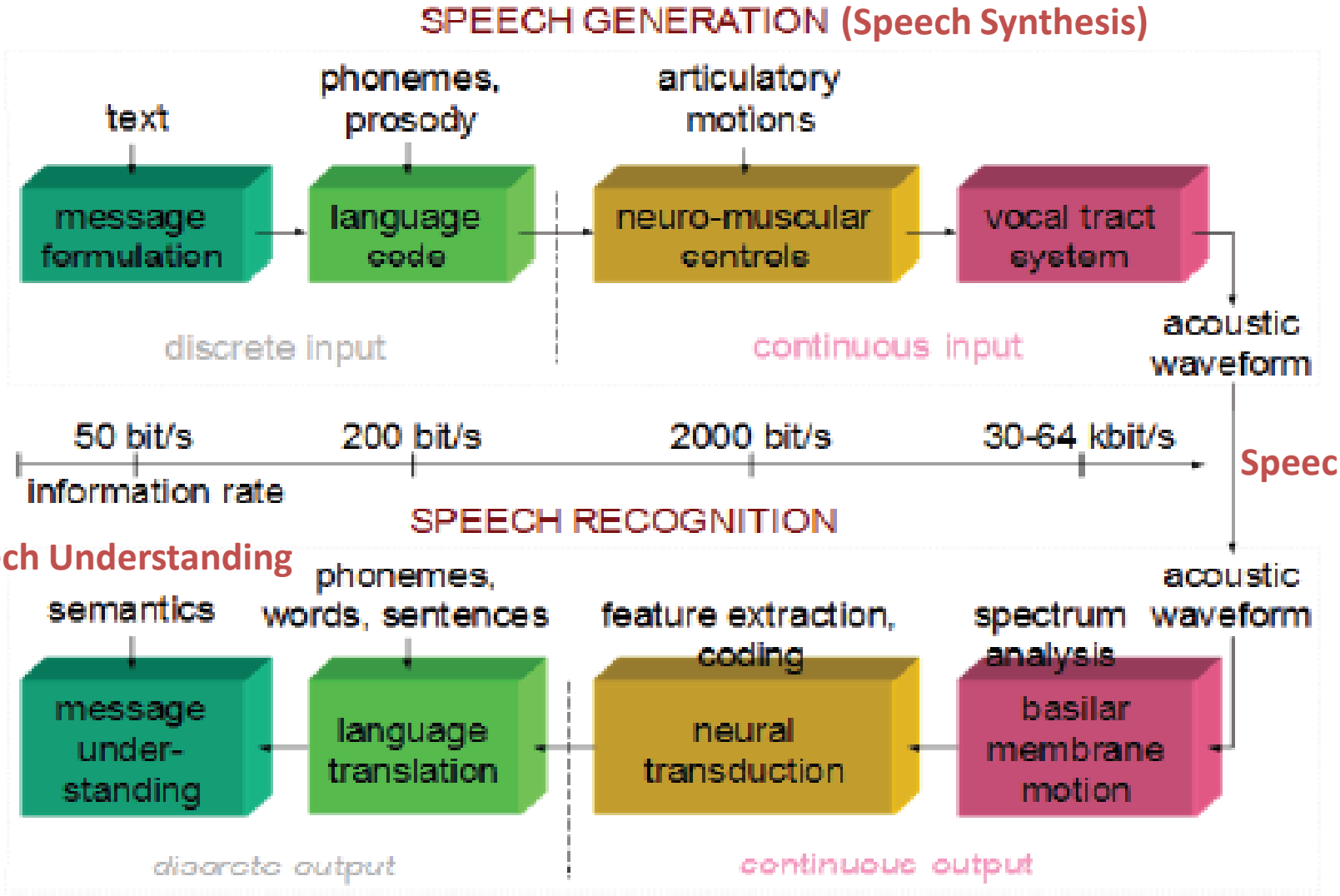
- Fixed at the front of the larynx but free to move at the back and sides
- Can be in one of three primary states
  - Breathing: Glottis is wide, muscles are relaxed, and air flows with minimal obstruction
  - Voicing: vocal folds are tense and are brought up together. Pressure builds up behind, leading to an oscillatory opening of the folds (video) **Link**
  - 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



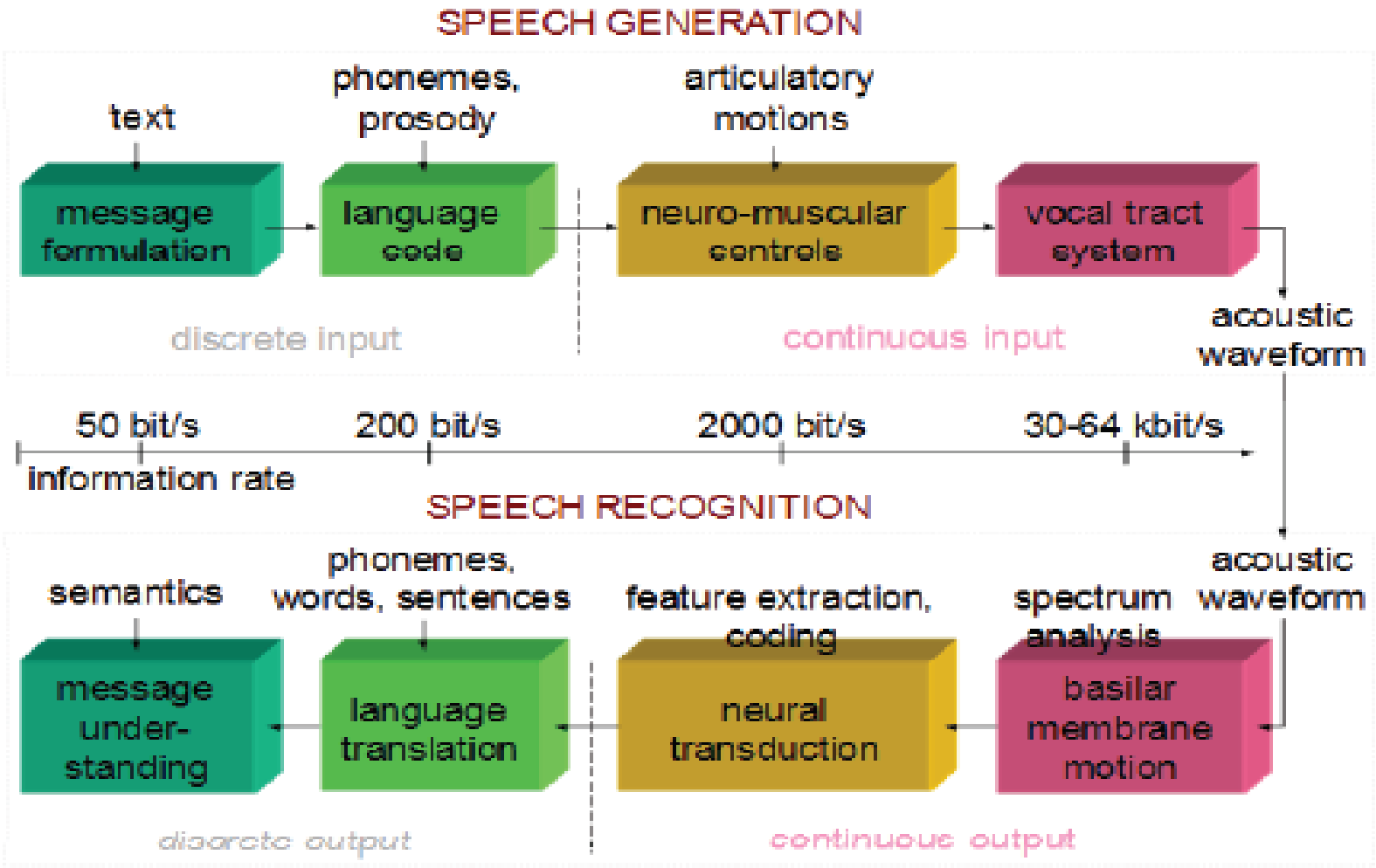
http://academic.kellogg.edu/herbrandsonc/bio201_mckinley/f25-5b_vocal_folds_lary_c.jpg



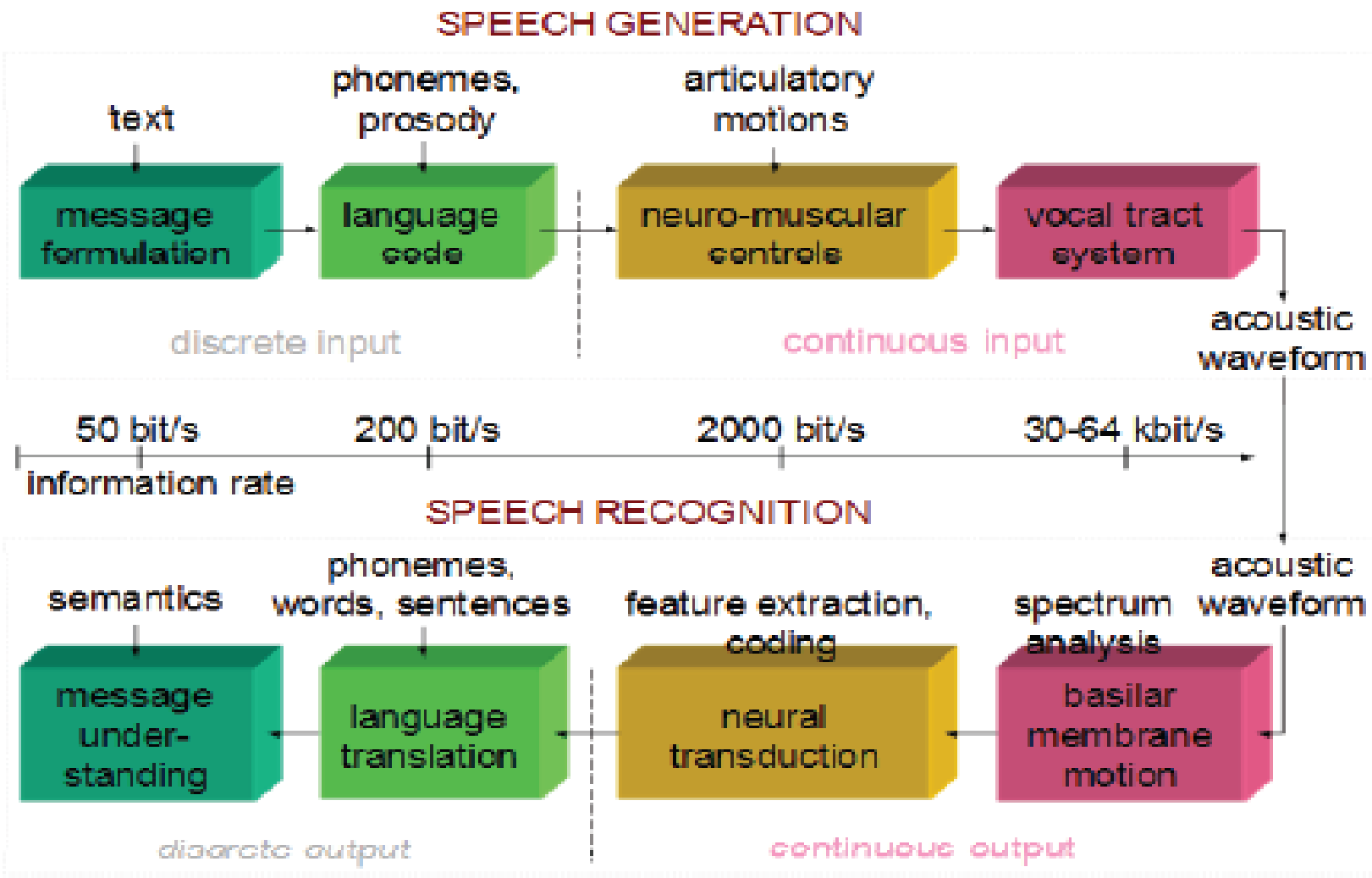
http://biorobotics.harvard.edu/research/heather2.gif



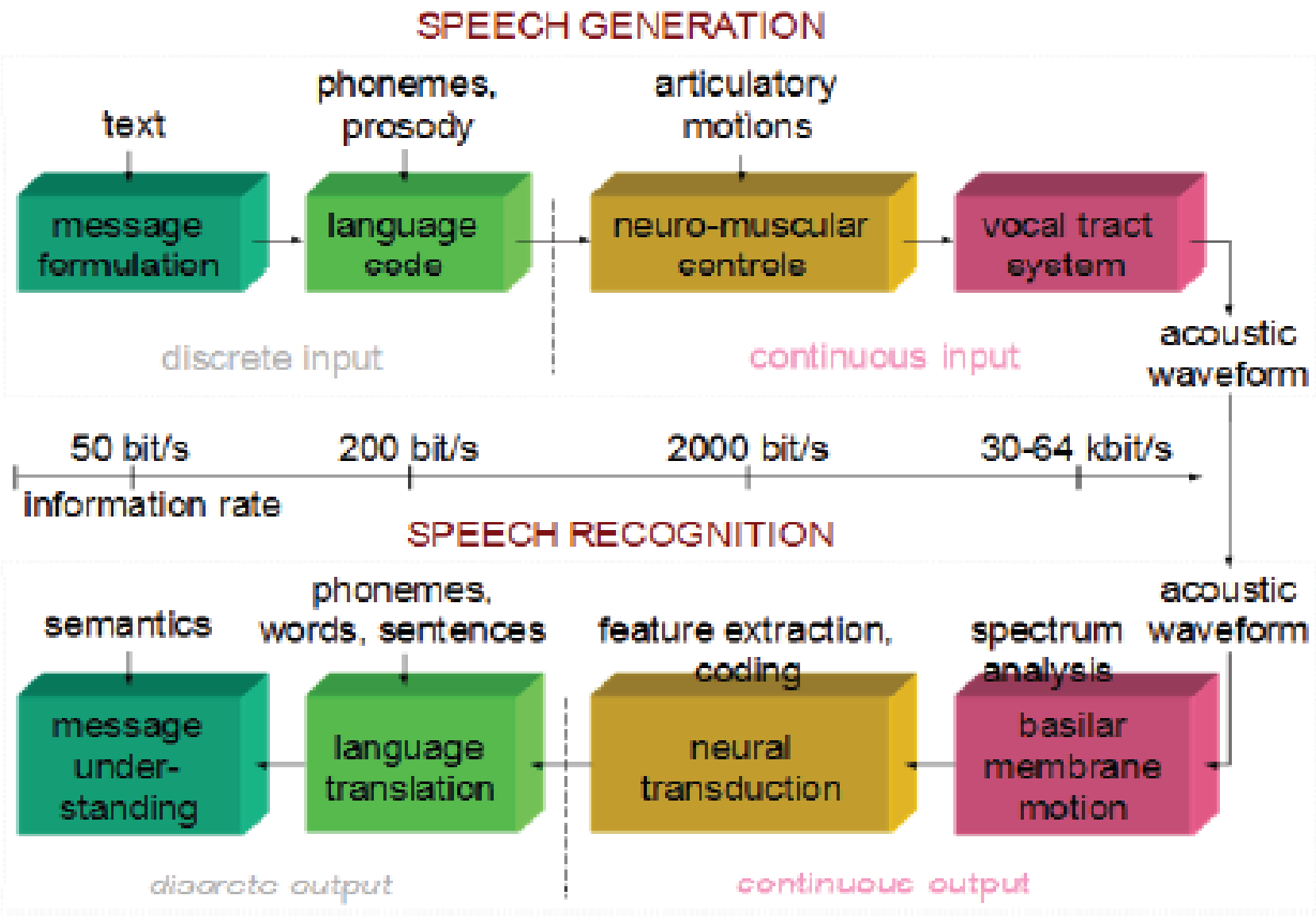
The speech production process begins when the talker formulates a message in his/her mind to transmit to the listener via speech



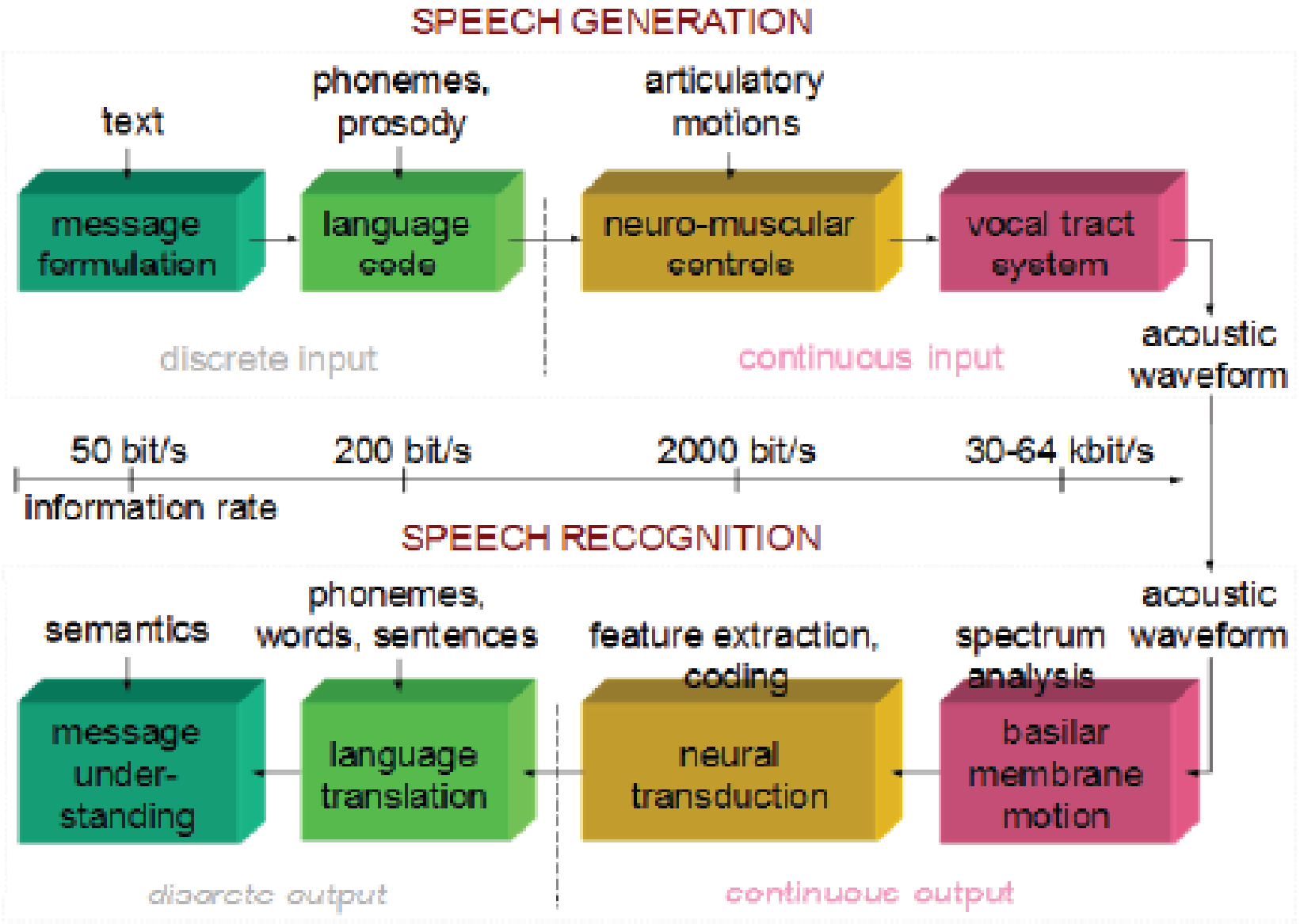
The next step in the process is the conversion of the message into a language code. This corresponds to converting the message into a set of phoneme sequences corresponding to the sounds that make up the words, along with prosody (syntax) markers denoting *duration* of sounds, *loudness* of sounds, and *pitch* associated with the sounds.



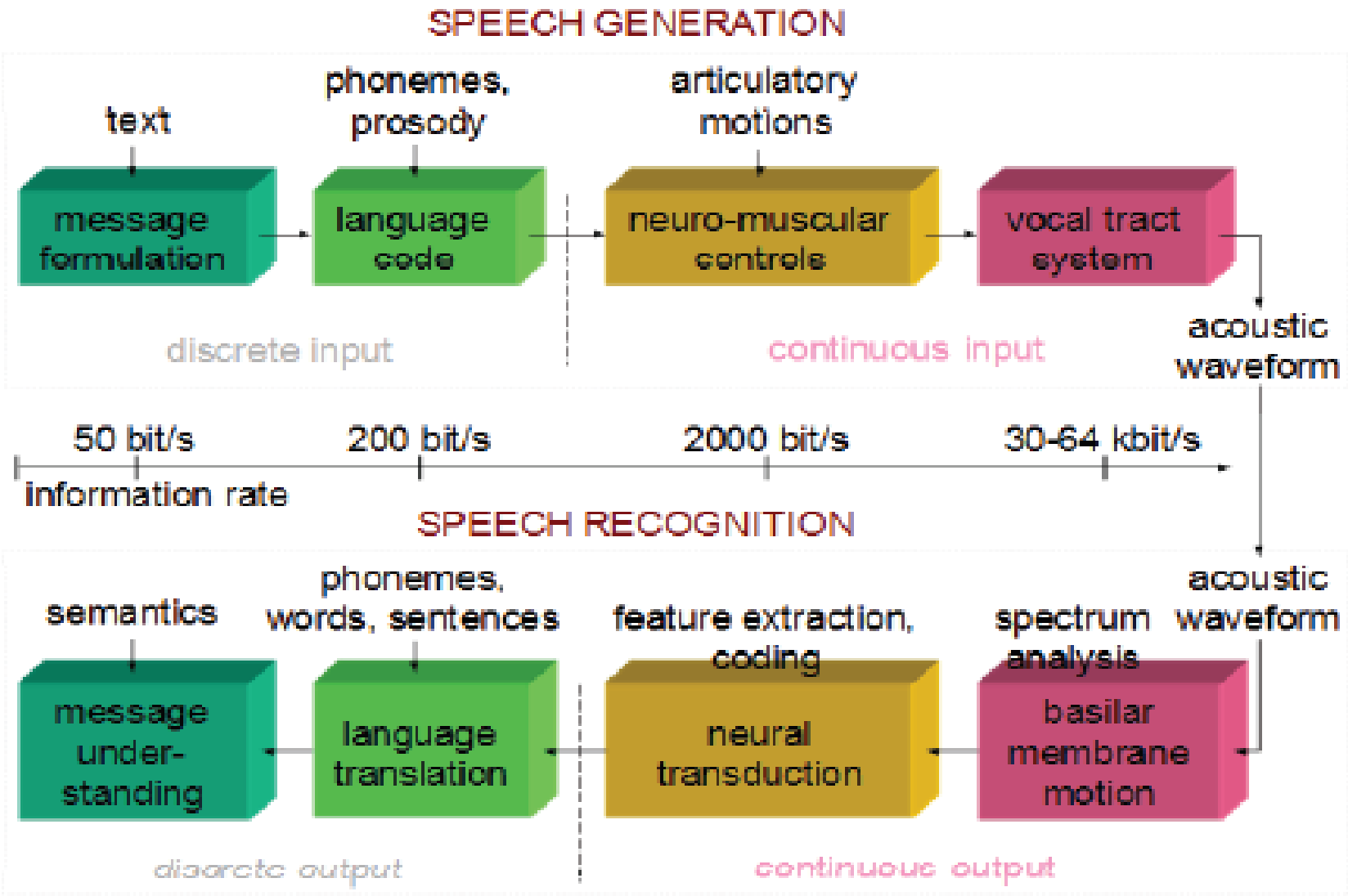
Once the language code is chosen the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output.



The neuromuscular commands must simultaneously control all aspects of articulatory motion including control of the **lips, jaw, tongue and velum.**



A neural transduction process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process.



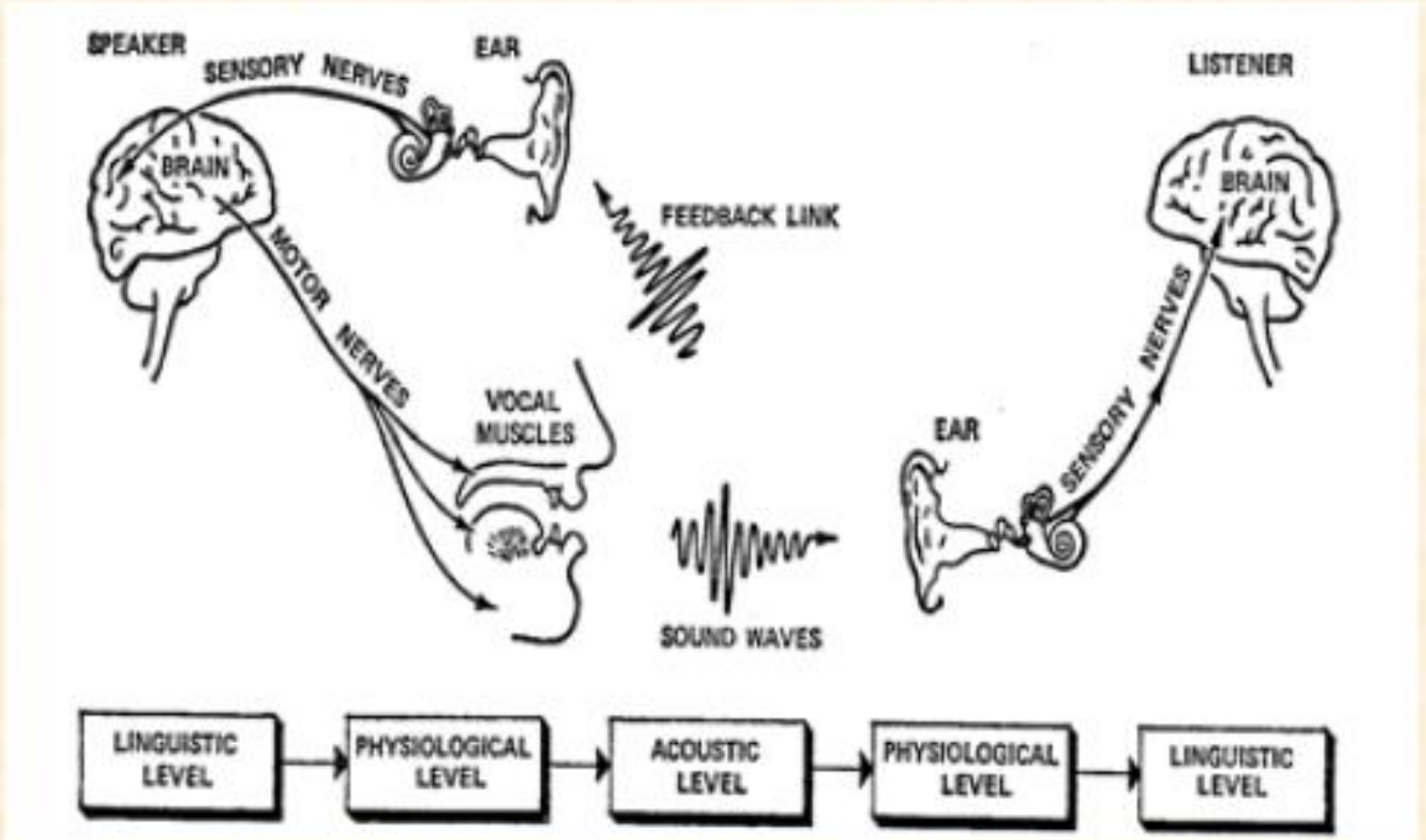
The neural activity along the auditory nerve is converted into a language code at higher centres of processing within the brain, and finally message comprehension (understanding of meaning) is achieved.

The mechanism of Speech Production

- In order to apply DSP techniques to speech processing problems it is important to understand the fundamentals of the speech production process.

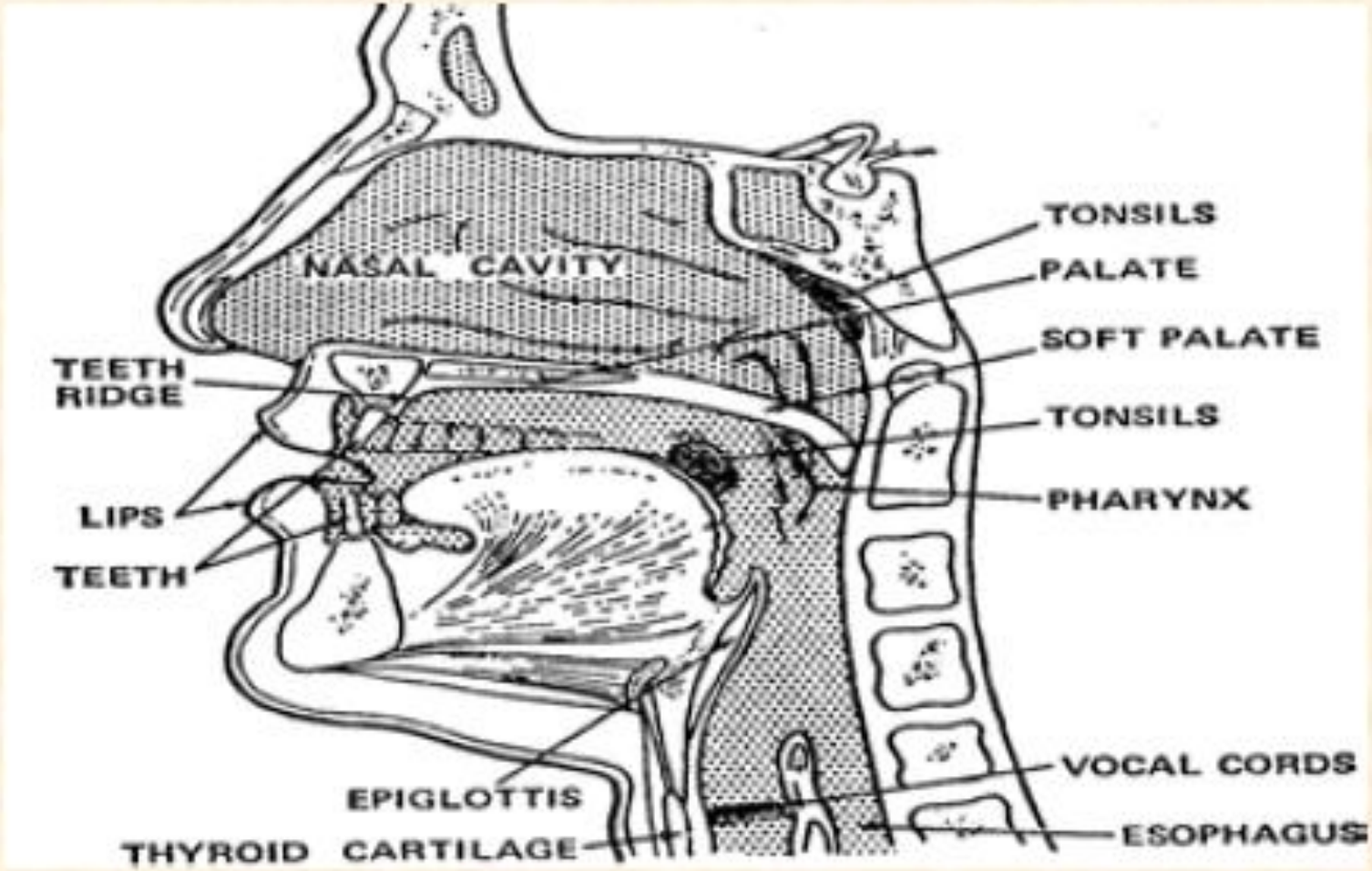
- Speech signals are composed of a sequence of sounds and the sequence of sounds are produced as a result of acoustical excitation of the vocal tract when air is expelled from the lungs

The Speech Chain



Taken from 'The Speech Chain: The Physics and Biology of Spoken Language' by P B Denes & E N Pinson, New York: Anchor Press, 1973

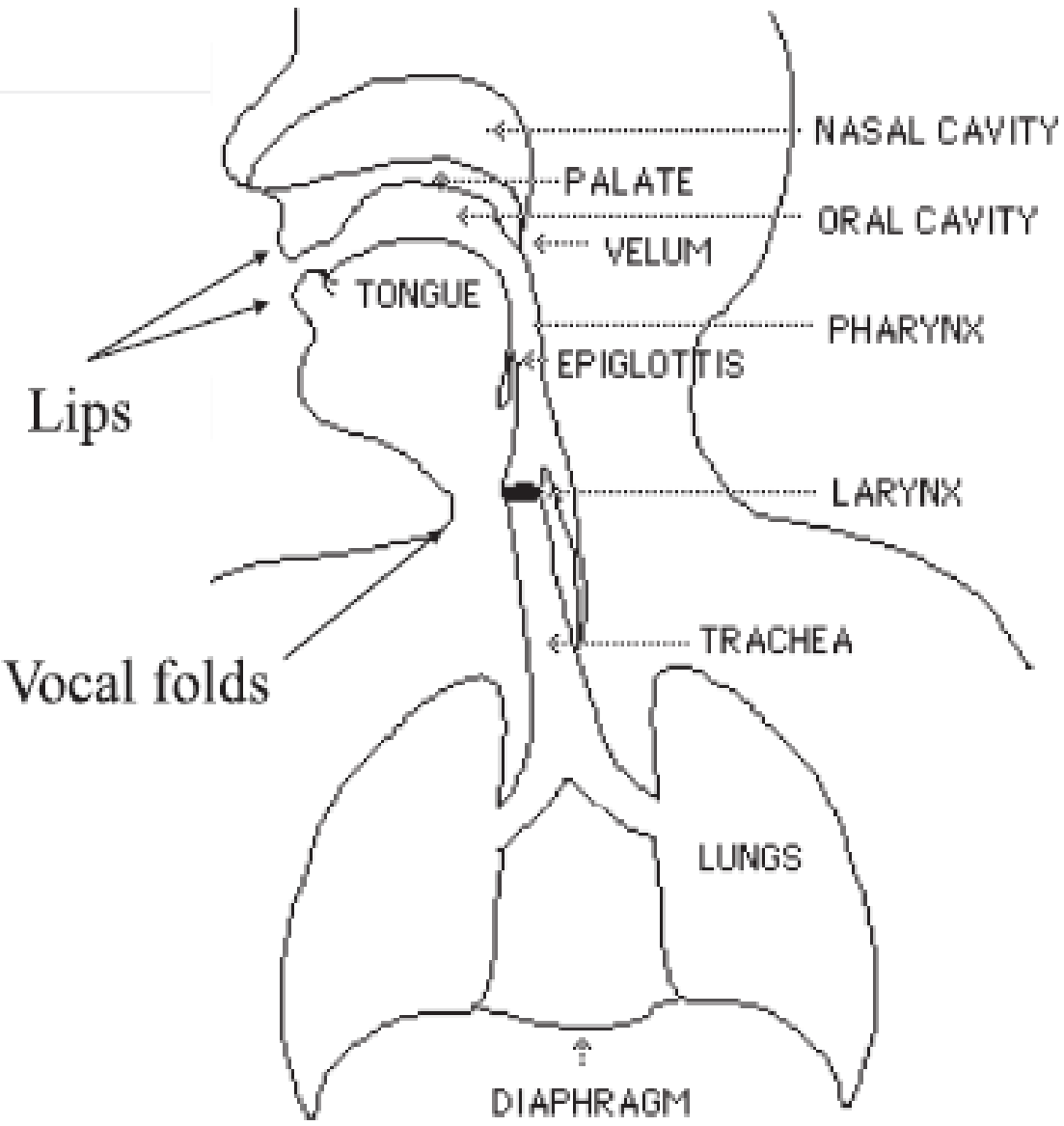
The Human Vocal Tract



Taken from 'The Speech Chain: The Physics and Biology of Spoken Language' by P B Denes & E N Pinson, New York: Anchor Press, 1973

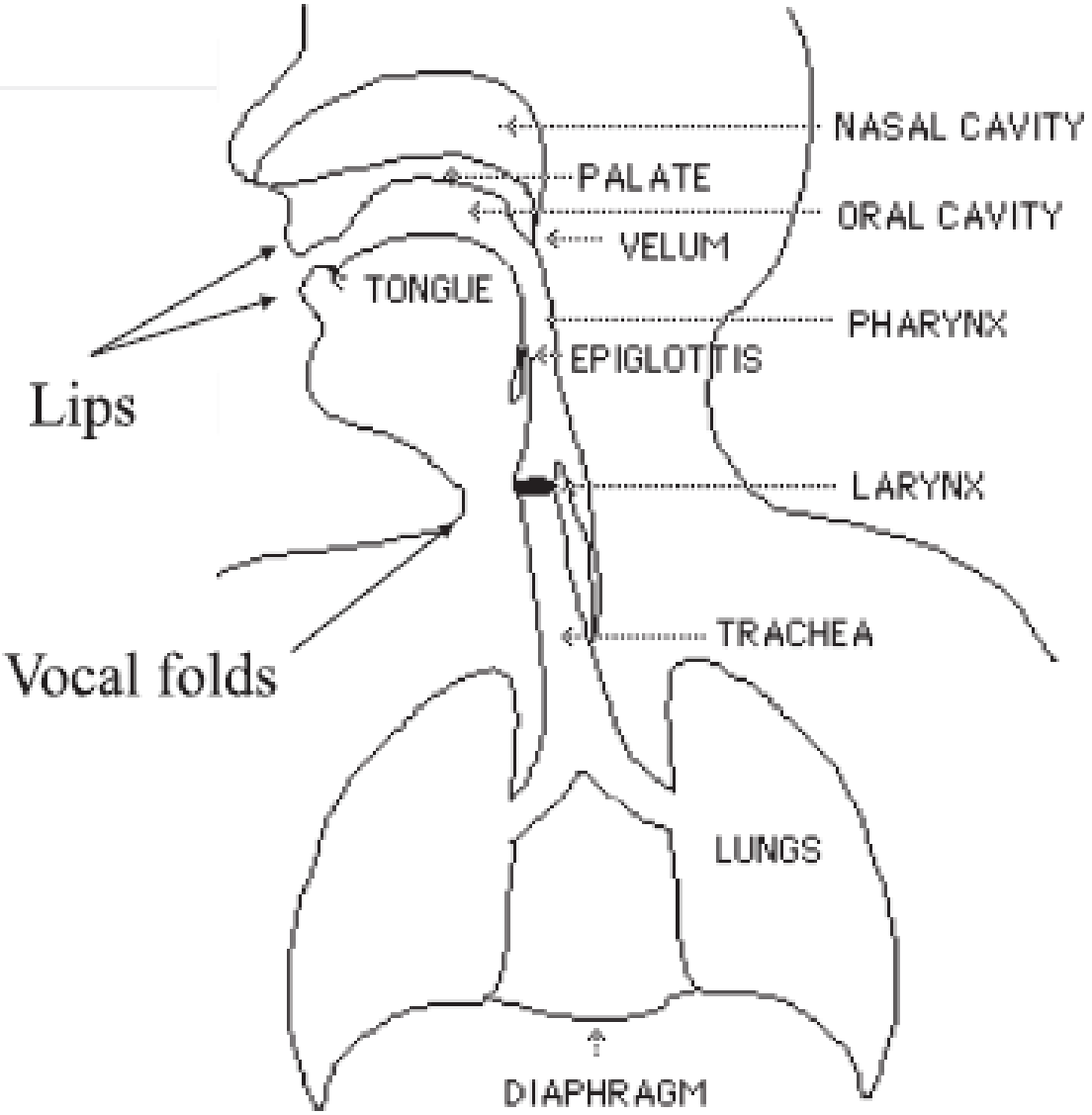
Speech Production Mechanism

- Vocal tract begins at the opening between the vocal cords and ends at the lips
- In the average male, the total length of the vocal tract is about 17 cm.
- The cross-ectional area of the vocal, determined by the positions of the tongue, lips, jaw and velum varies from zero (complete closure) to about 20 cm^2.



Speech Production Mechanism

- The nasal tract begins at the velum and ends at the nostrilss
- When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.



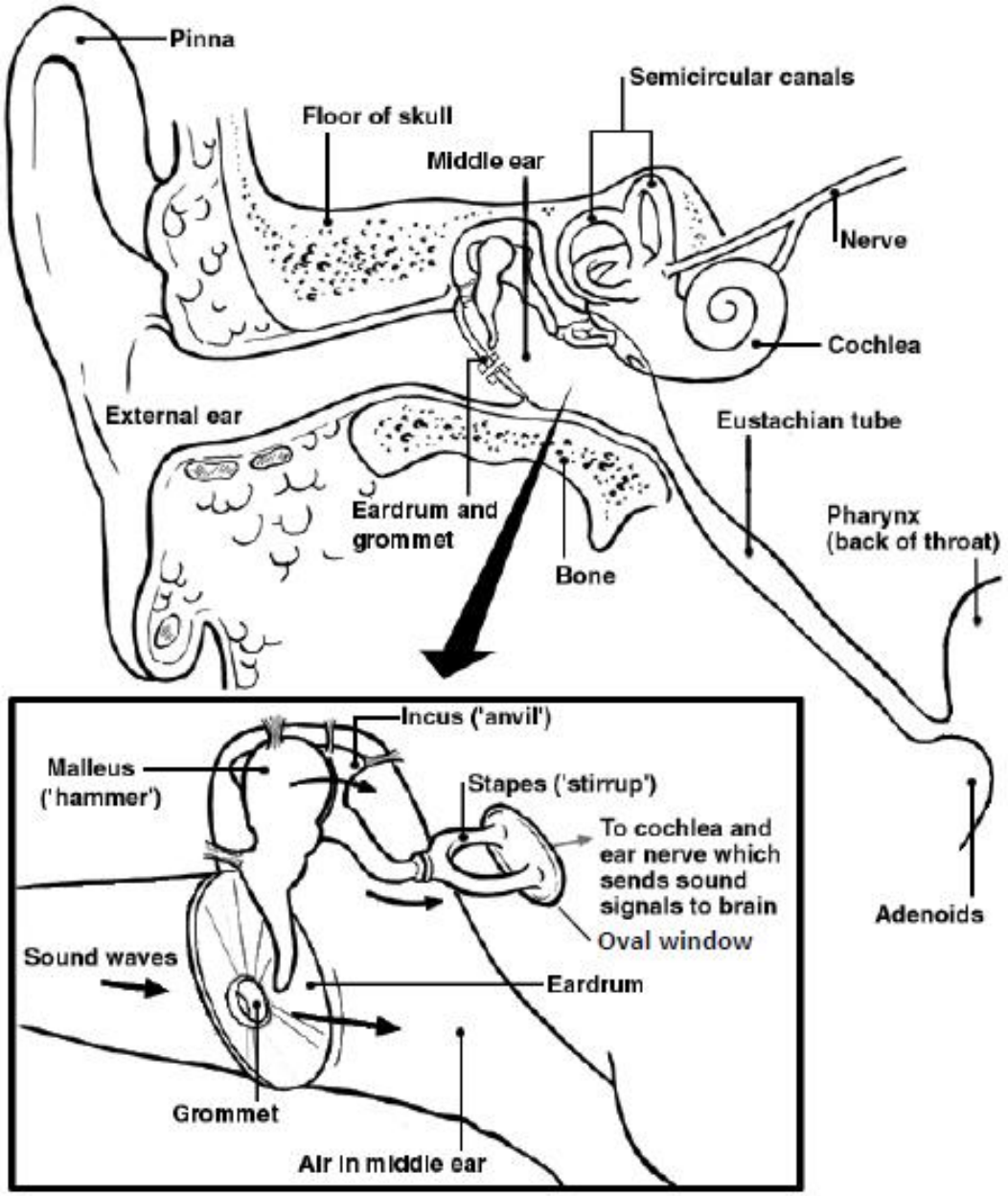
Anatomy of the ear

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

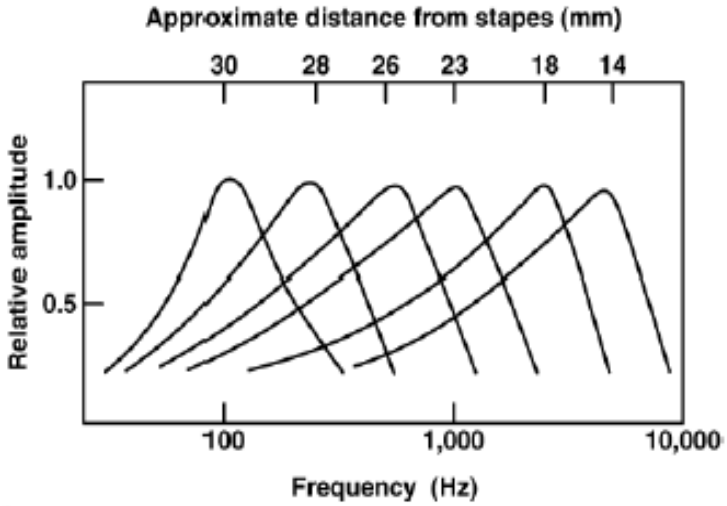
- A tube coiled in a snake-shaped spiral
- Inside filled with gelatinous fluid
- Running along its length is the basilar membrane
- Along the BM are located approx. 10,000 inner hair cells

Signal transduction

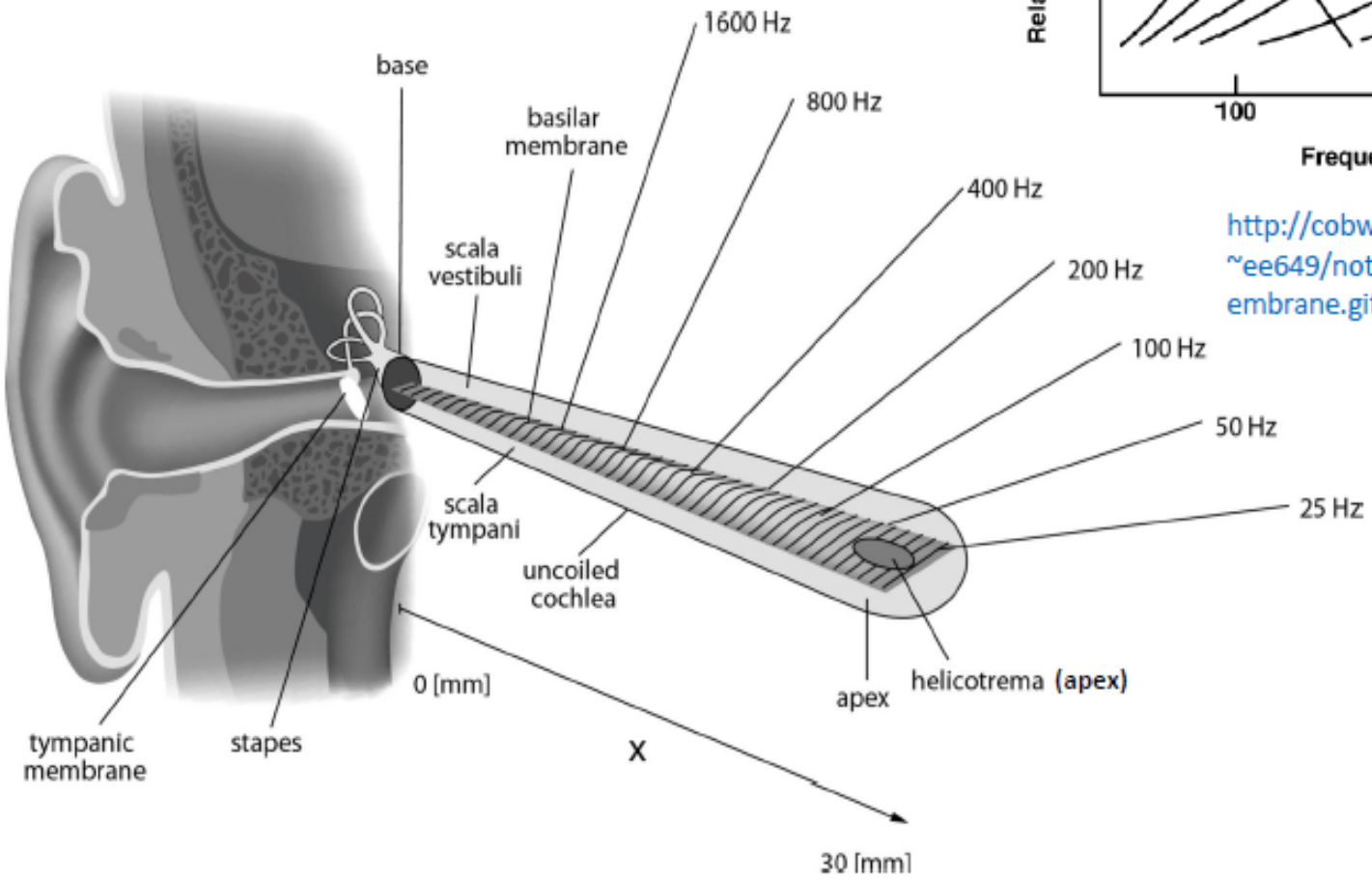
- Vibrations of the eardrum cause movement in the oval window
- This causes a compression sound wave in the cochlear fluid
- This causes vertical vibration of basilar membrane
- This causes deflections in the inner hair cells, which then fire

Frequency tuning

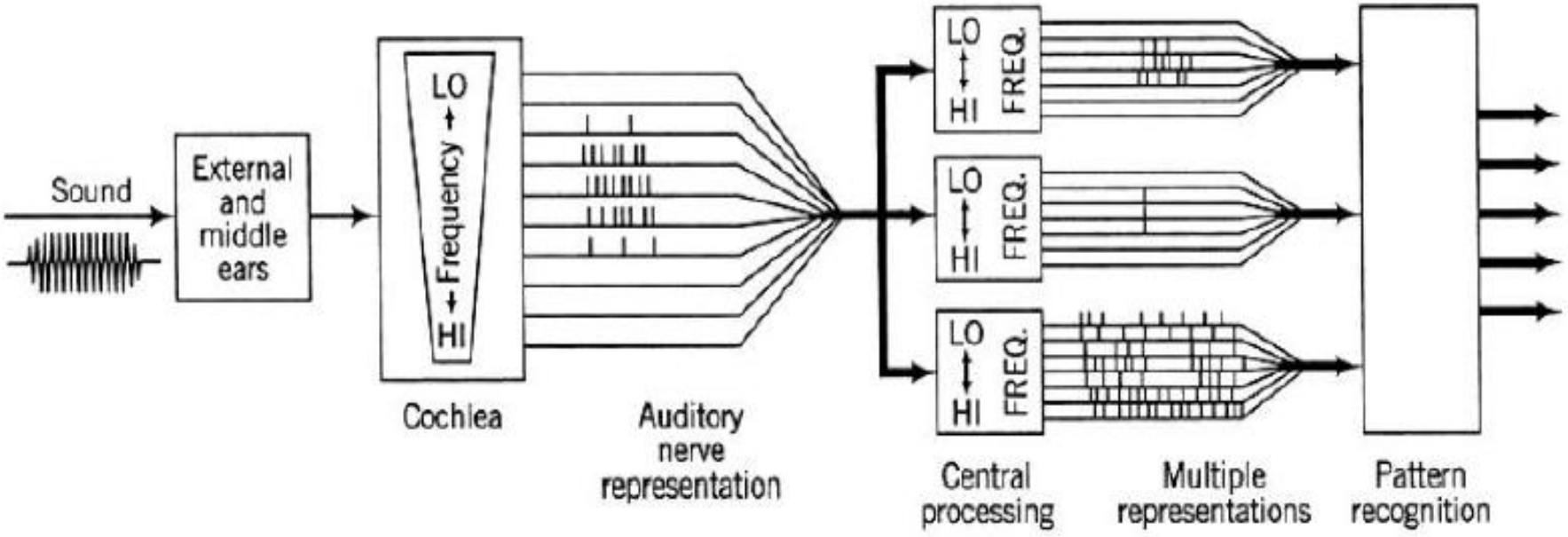
- BM is stiff/thin at basal end (stapes), but compliant/massive at apex
- Thus, traveling waves peak at different positions along BM
- As a result, BM can be modeled as a filter bank (video) **Link**



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

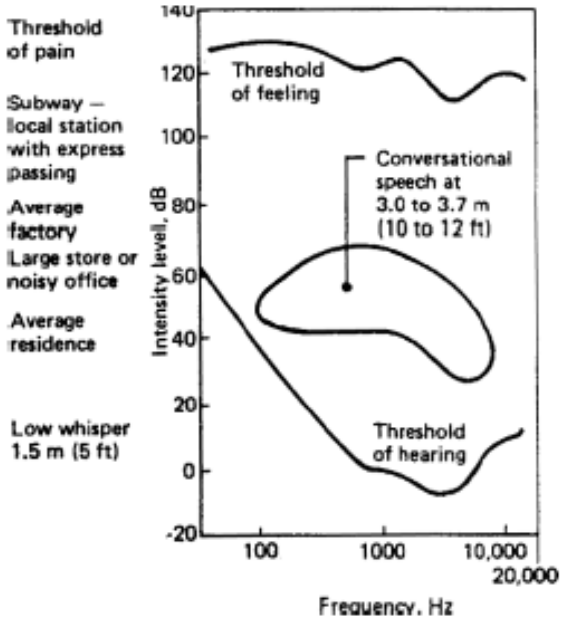


[Rabiner & Schafer, 2007]

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)

Masking

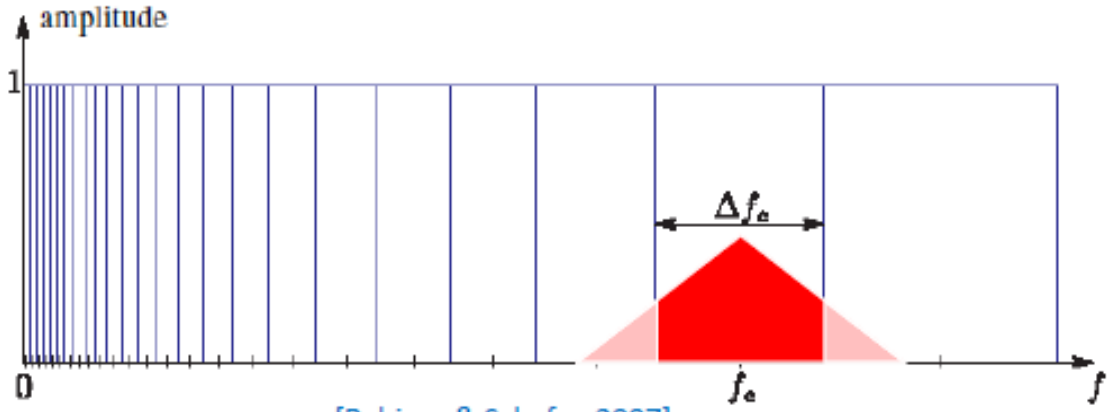
- A phenomenon whereby the perception of a sound is obscured by the presence of another (i.e., the latter raises the threshold of the former)
- Masking is the major non-linear phenomenon that prevents treating the perception of speech sounds as a summation of responses

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

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?

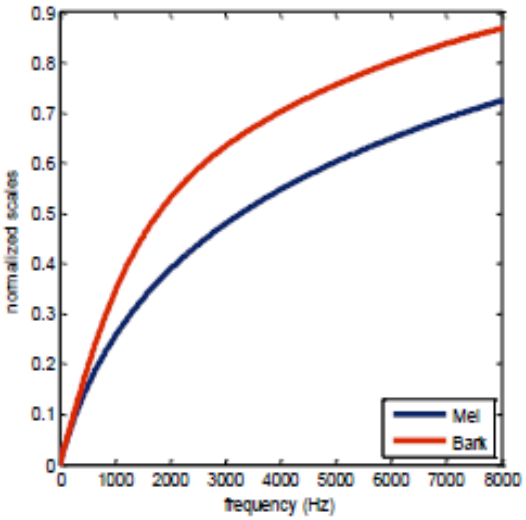
- Take a band-limited noise signal with a center frequency of 2 kHz, and play it alongside a sinusoidal 2 kHz tone
- Make the tone very quiet relative to the noise
  - You will not be able to detect the tone because the noise signal will mask it
  - Now, turn up the level of the tone until you can hear 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

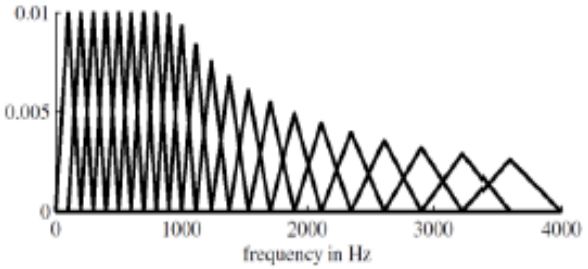
- Bark scale
  - Relates acoustic frequency to perceptual frequency resolution
  - One Bark equals one critical band

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



- Mel scale (more **Later on**)
  - Linear mapping up to $1\ kHz$, then logarithmic at higher frequencies

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



[Rabiner & Schafer, 2007]