Speech production & perception

Professor Marie Roch

SAN DIEGO STATE UNIVERSITY

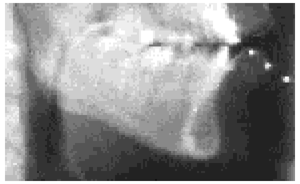
Phonetics & Phonology

- Phoneme
  - Description of a minimal unit of sound which can be used to distinguish one word for another.
  - We use symbols from the international phonetic alphabet to denote phonemes, typically between slashes
  - Example: "pet" /pɛt/ vs. "bet" /bɛt/

- Phone – A sound that corresponds to a phoneme.

Speech Production

Air, driven by our lungs, drives speech production.



Haskins - www.haskins.yale.edu/haskins /HEADS/production.html

The sound, or phone produced depends upon voicing & the configuration of our articulators.

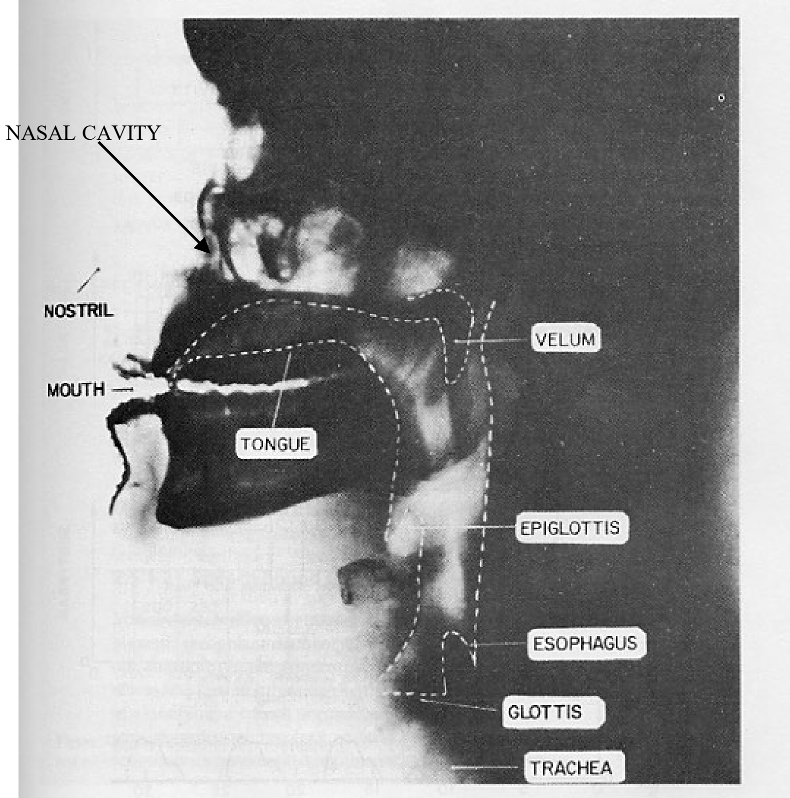


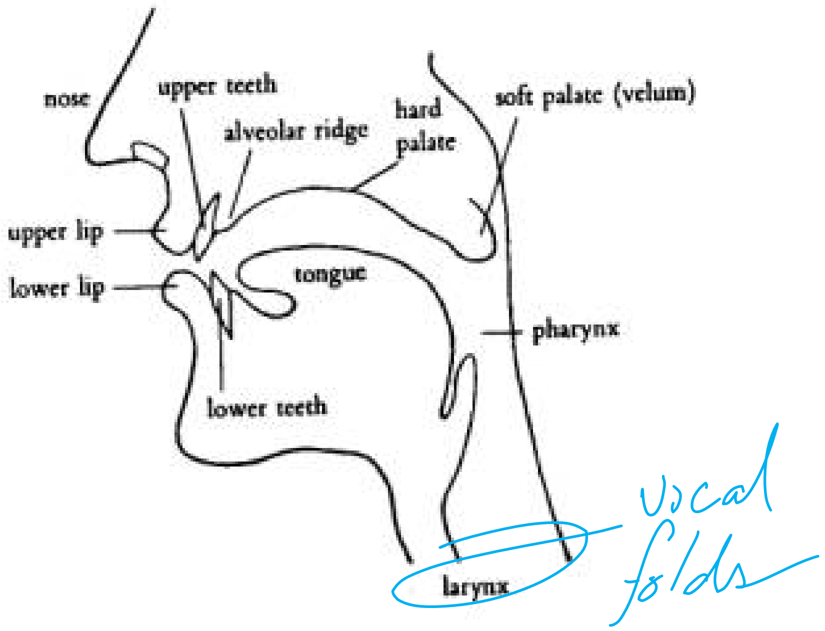
Figure 2.3 Mid-sagittal plane X-ray of the human vocal apparatus (after Flanagan et al. [2]).

Rabiner/Juang 1993

SAN DIEGO STATE UNIVERSITY

Articulators

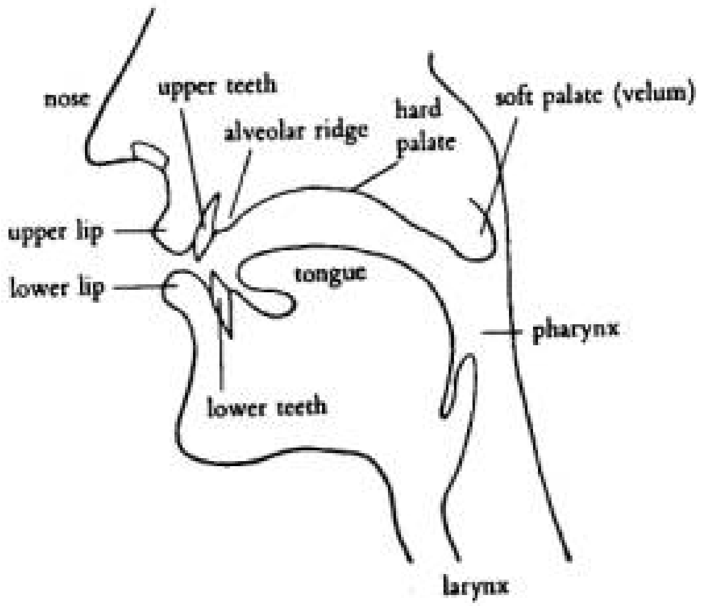
- Vocal folds (cords) - Responsible for voiced/unvoiced speech
- Velum (soft palate) – Serves as a valve to the nasal cavity and can be raised or lowered to allow air into th nasal cavity.



http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

Articulators

- Tongue – Flexible muscle, shape & position very important to phoneme production.
- Lips – Rounding can extend the length of the vocal tract. Closure can produce a stop, i.e. the /p/ in “apple.”

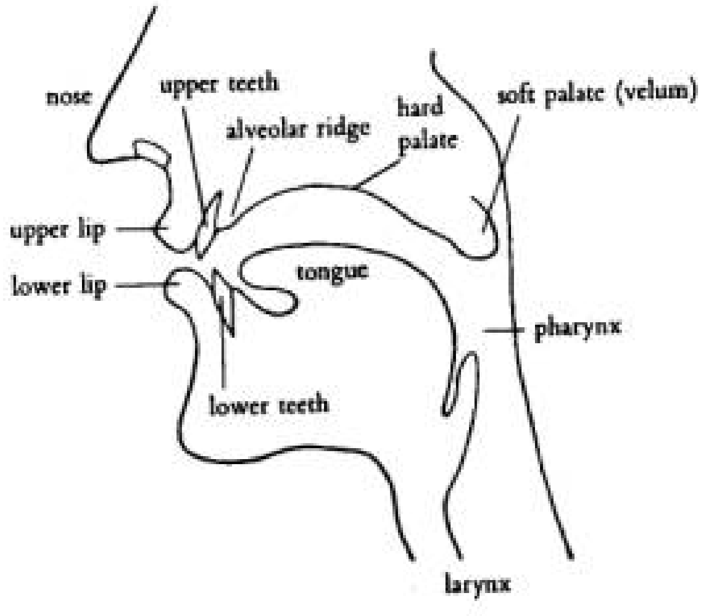


http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

Articulators

Targets: Tongue contacts these structures to change productions

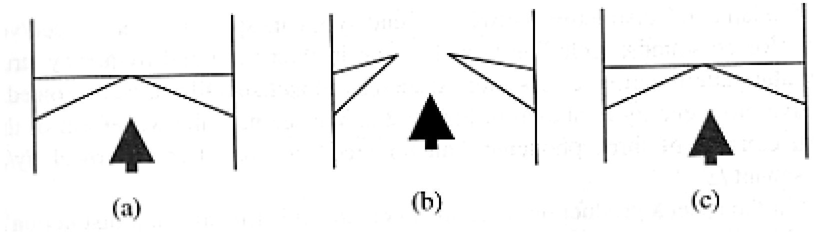
- Teeth – Target for the tongue for some consonants, i.e. /dh/ in “then.” (Teeth are actually moved by the jaw.)
- Alveolar ridge
- Hard palate – Hard part of the roof of your mouth.



http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

Voicing

- Voiced sounds occur when the vocal folds open & close at a regular interval:

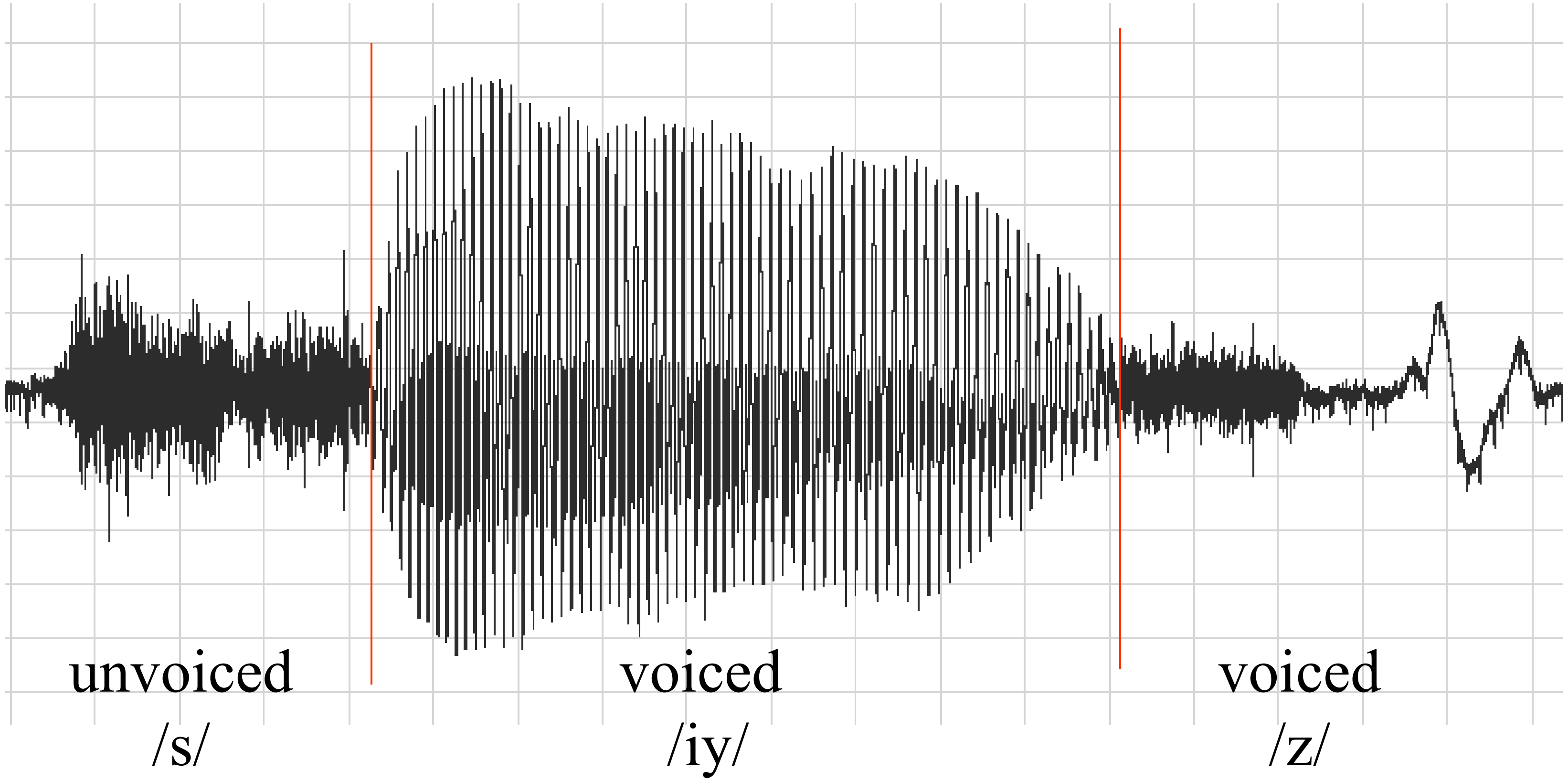


Huang et al., 2001, p 26

UCLA Phonetics Lab

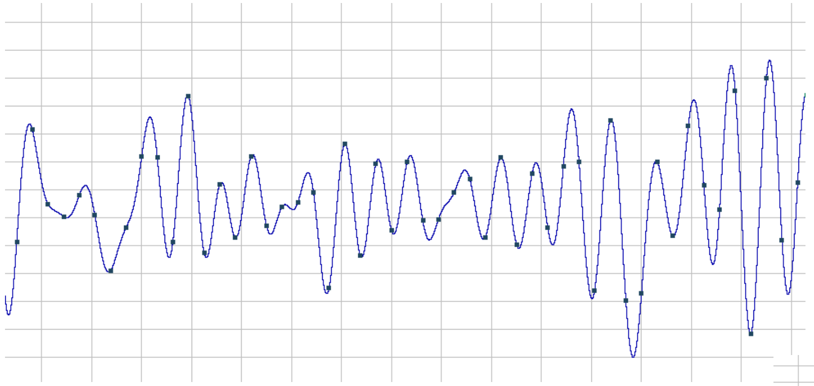
  – Subglottal pressure forces open the vocal folds
  – As the pressure differential drops, the folds close.

Voicing
"sees"



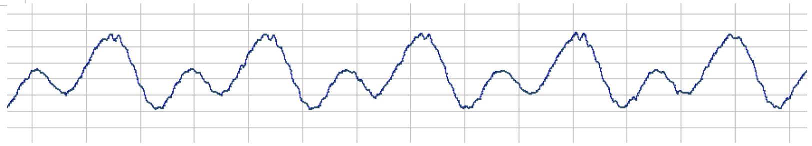
San Diego State University

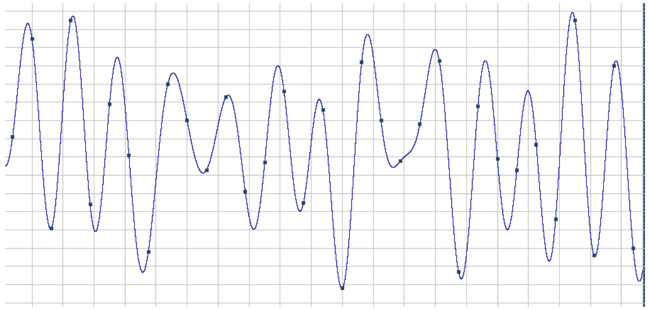
Zoomed time series of “sees” (different time scales)



unvoiced s /s/

voiced ee /iy/



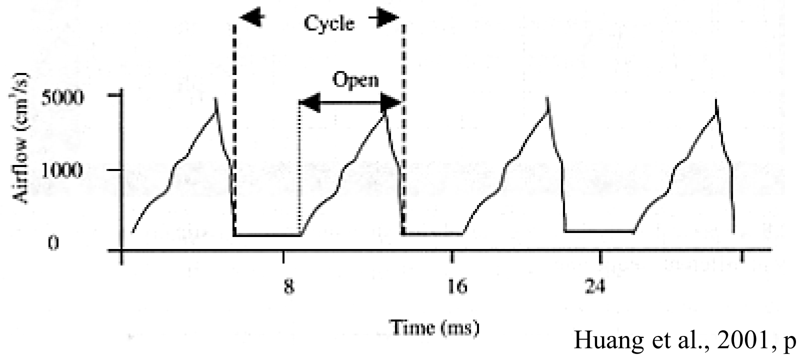


voiced s /z/
(constriction contributes to irregular pattern unlike the vowel)

SAN DIEGO STATE UNIVERSITY

F0 – Fundamental Frequency

- The *fundamental frequency*, or F0, is the number of times per second that the vocal folds open & close



Huang et al., 2001, p 27

Each cycle in the figure to the left is about 8.33 ms.

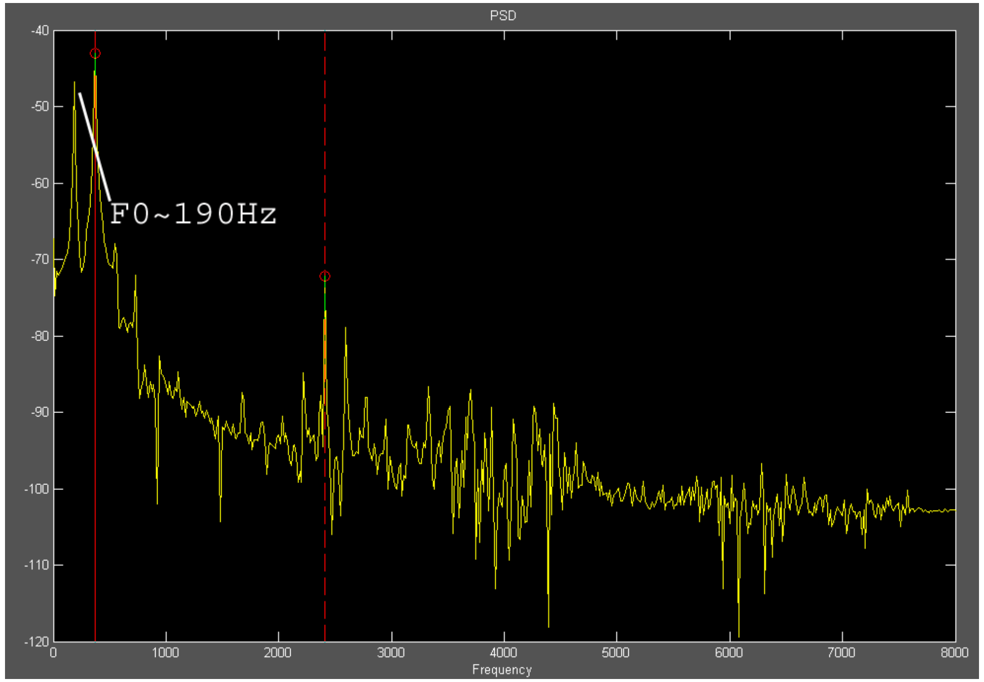
As $Frequency = \frac{cycles}{s}$,

$$\frac{1\ \text{cycle}}{8.33\ \text{ms}} \frac{1000\ \text{ms}}{1\ \text{s}} \approx 120Hz$$

and F0 is about 120 Hz

F0 and Harmonics

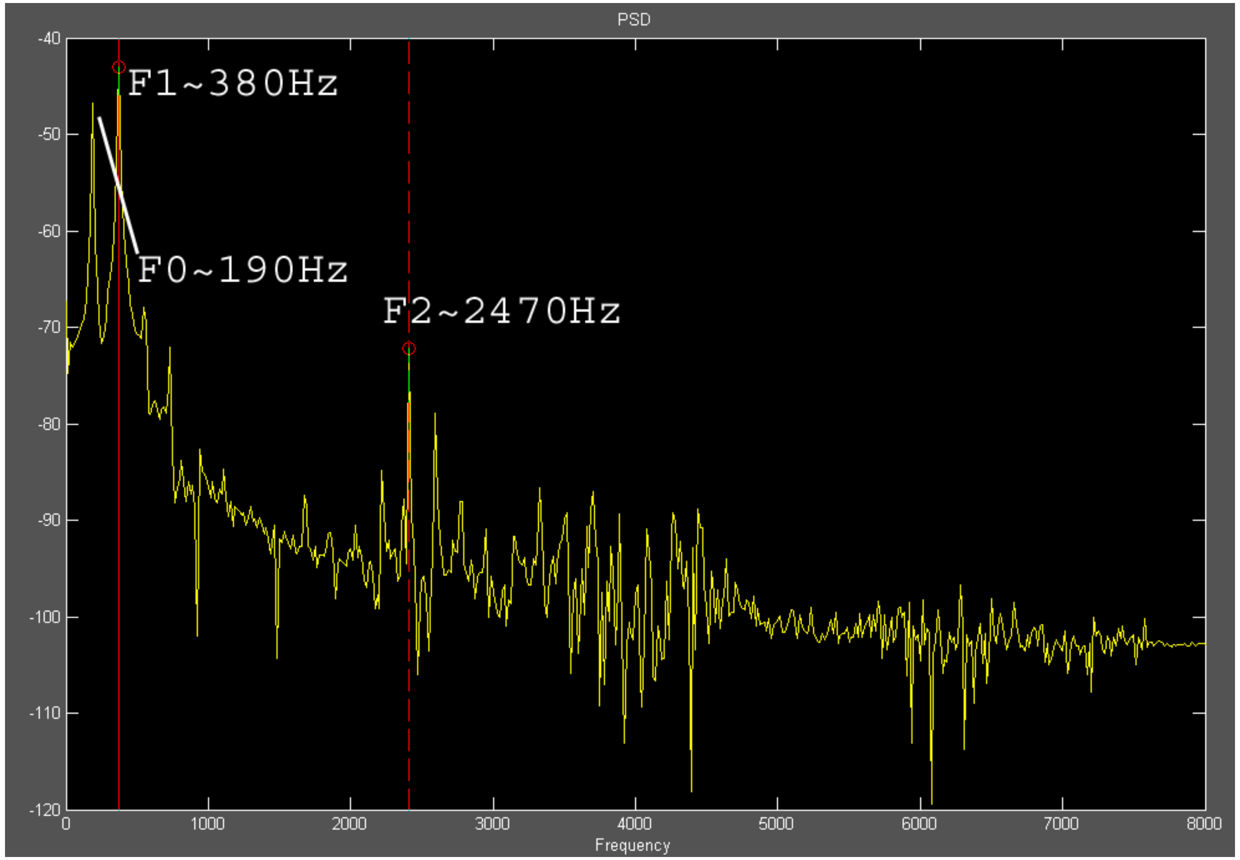
- F0 (if present), is not the only frequency.
- Harmonics are frequencies which occur at multiples of F0.
- Frequencies from a small portion of ee /iy/



SAN DIEGO STATE UNIVERSITY

Formants

- For any vocal tract shape, certain frequencies are reinforced.
- Harmonics (multiples) of F0 near resonances are reinforced.



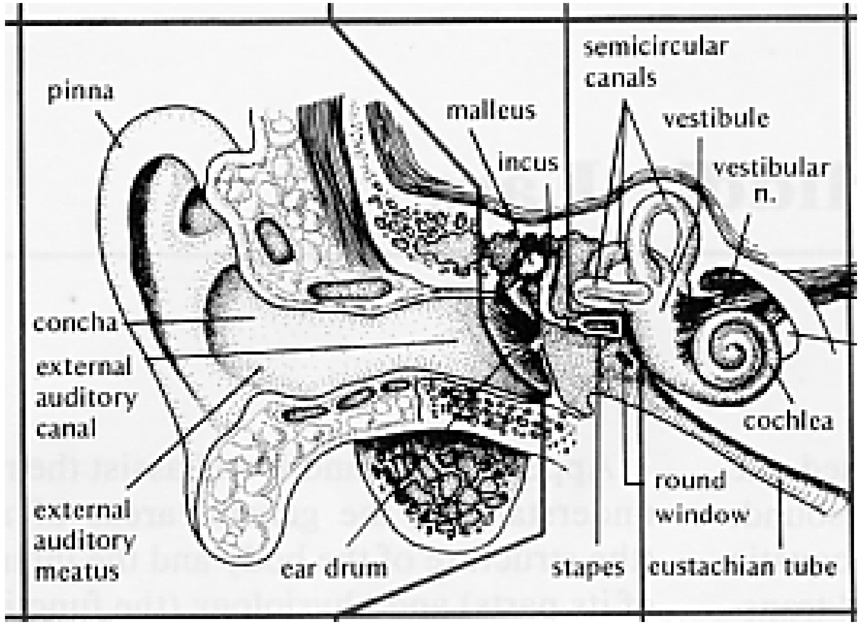
Formants

- These reinforced harmonics are called formants, and can play an important role in recognizing vowels.

- Note that F0 ***is not a formant!***

The Human Ear

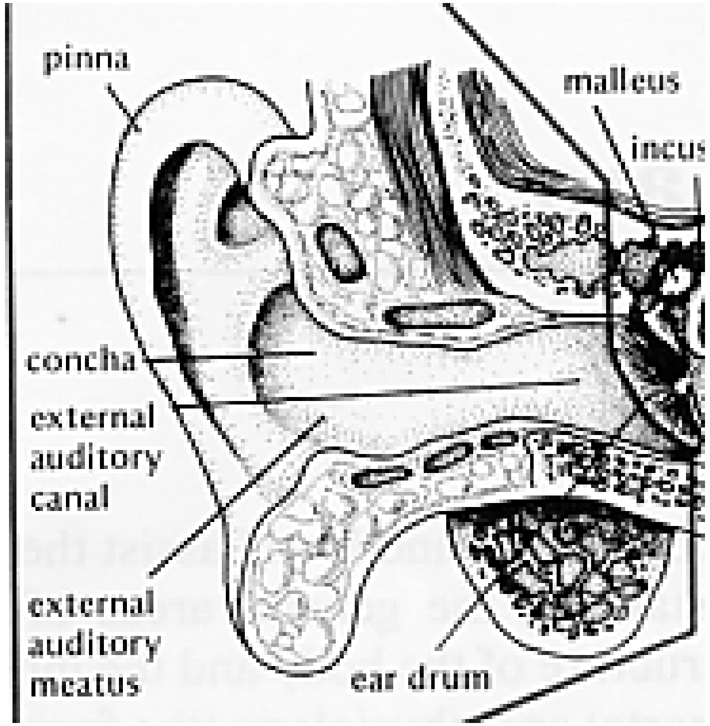
- Outer
- Middle
- Inner



Yost, 1994

The outer ear

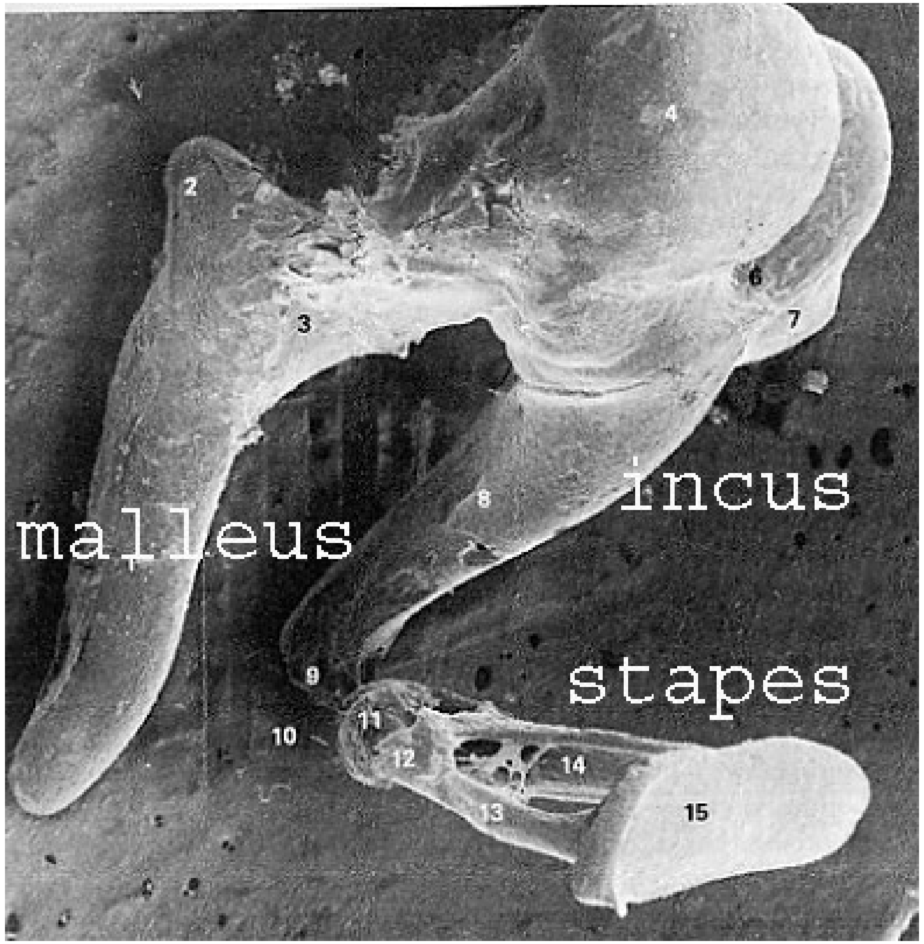
- Pinna - protect & filter
- Ear canal & concha - amplify frequencies between 1.5-7kHz.
- tympanic membrane (ear drum)



Yost, 1994

The middle ear

- Outer ear's tympanic membrane connected to the inner ear's oval window by ossicles
  - malleus
  - incus
  - stapes



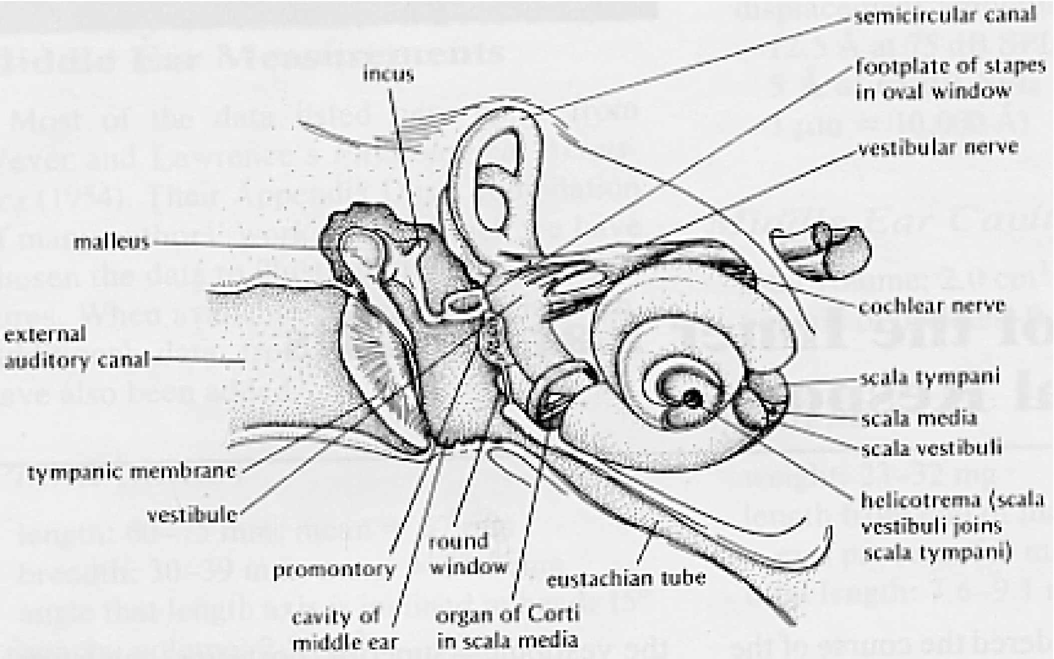
Yost 1994

Middle ear contd.

- Ossicle functioning
  - mechanical transfer of energy
  - compression to prevent overload
  - stapes connected to the inner ear's oval window
- Eustachian tube
  - Connects to nasal cavity
  - Normally closed
  - When open, permits pressure equalization between outer/middle ear.

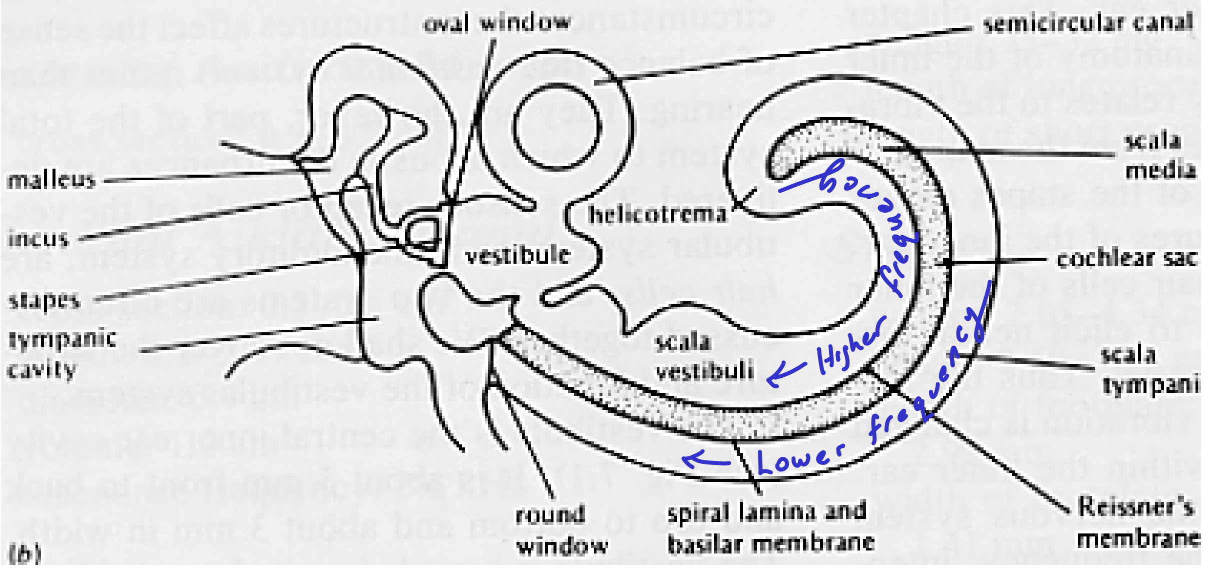
The inner ear

- Vestibule
- Semicircular canals
  - sense of balance
- Cochlea
  - coiled ≈ 2 and ¾ turns.
  - mechanical → neural impulses



Yost, 1994

Cochlea (simplified view)



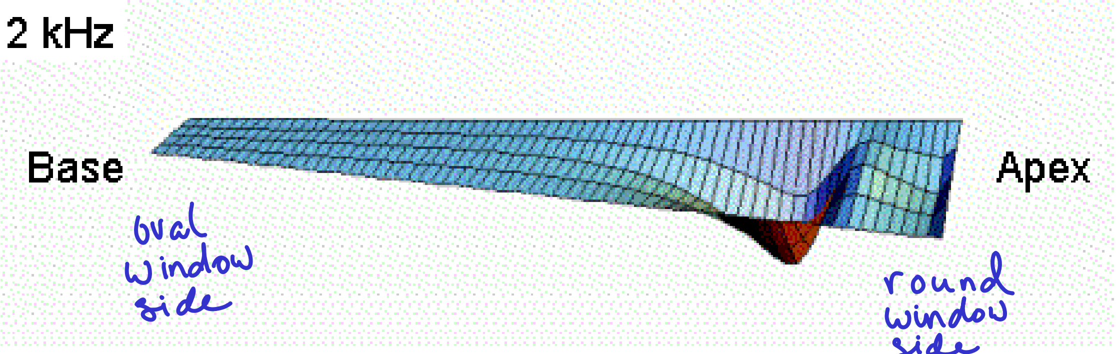
Yost, 1994

- filled with fluid
- scala vestibuli and scala tympani joined at apex (helicotrema)

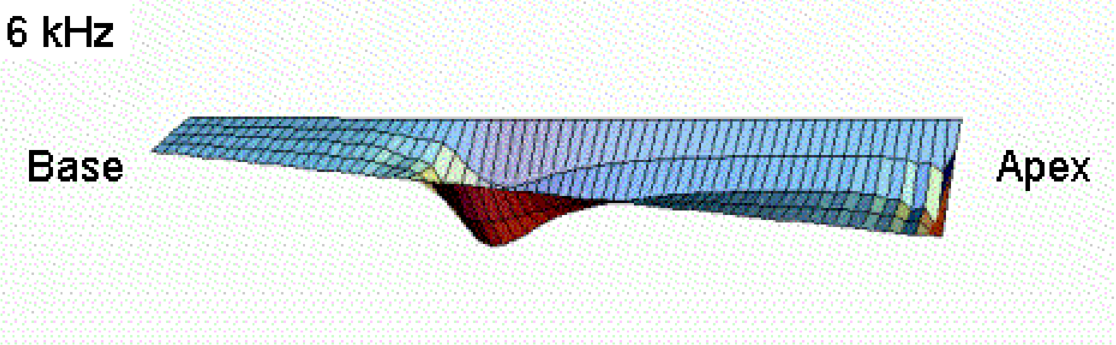
- traveling waves vibrate the basilar membrane moving hair cells which fire neurons

Deformation of basilar membrane

- Point of maximum deformation is frequency dependent
- The cochlea acts as a spectrum analyzer.



finite element model animations from WADA laboratory, Japan

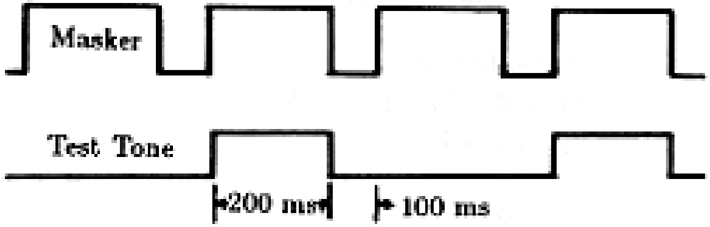


Masking

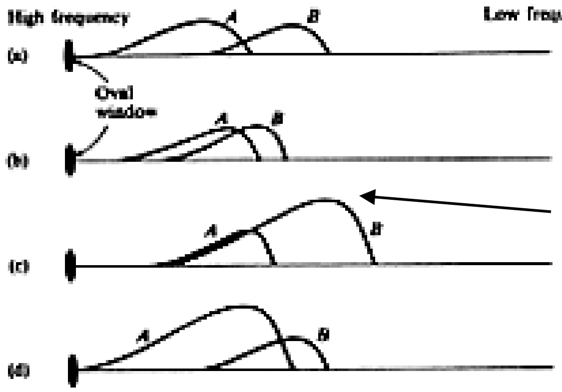
- Simultaneous tones close in frequency:
  - Louder tone can "hide" the softer ones.
  - Lower frequency tones are better maskers.
- When a short tone follows a sound closely (20-30 ms), the tone may be hidden (forward masking).

Masking Demonstration

- Low vs. high frequency masker
  - Masker/Test 1200/2000Hz then 2000/1200 Hz.
  - Ten repetitions, volume of test tone decreases each time.
- Basilar membrane response
  - Lower pitch masks more effectively than lower pitch tone.



Houtsma et al., *Auditory Demonstrations,*1987 p 29

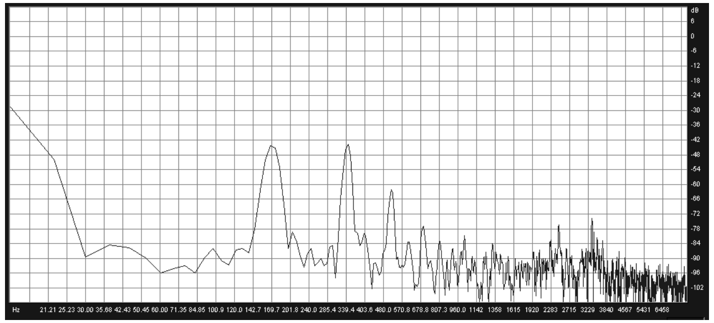


Lower pitch tone hides higher pitch one.

Simplified response of the basilar membrane (from Rossing, 1982).

Spectral shape and Timbre

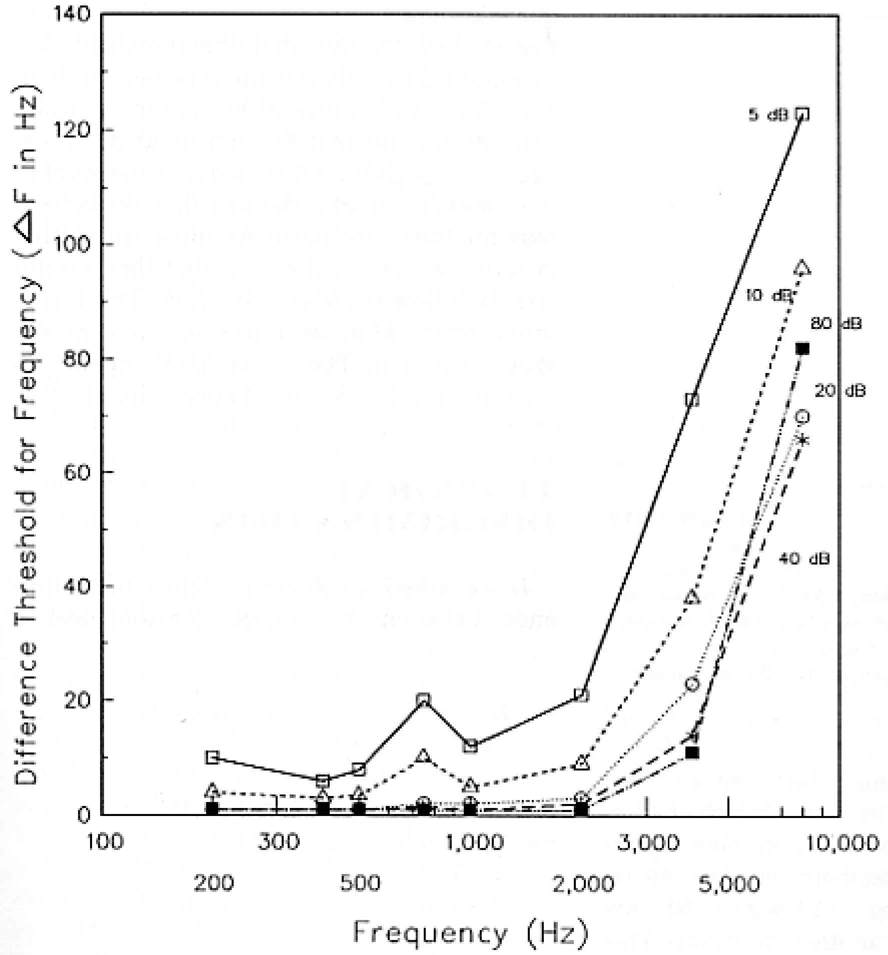
- Spectral shape is the shape of the frequency domain:



- Timbre is our perception of the frequencies, i.e. a sound is “rich” or “tinny.”

Frequency discrimination

- 0-4000 Hz – Good frequency resolution
- > 4000 Hz – Requires greater separation of frequency to distinguish

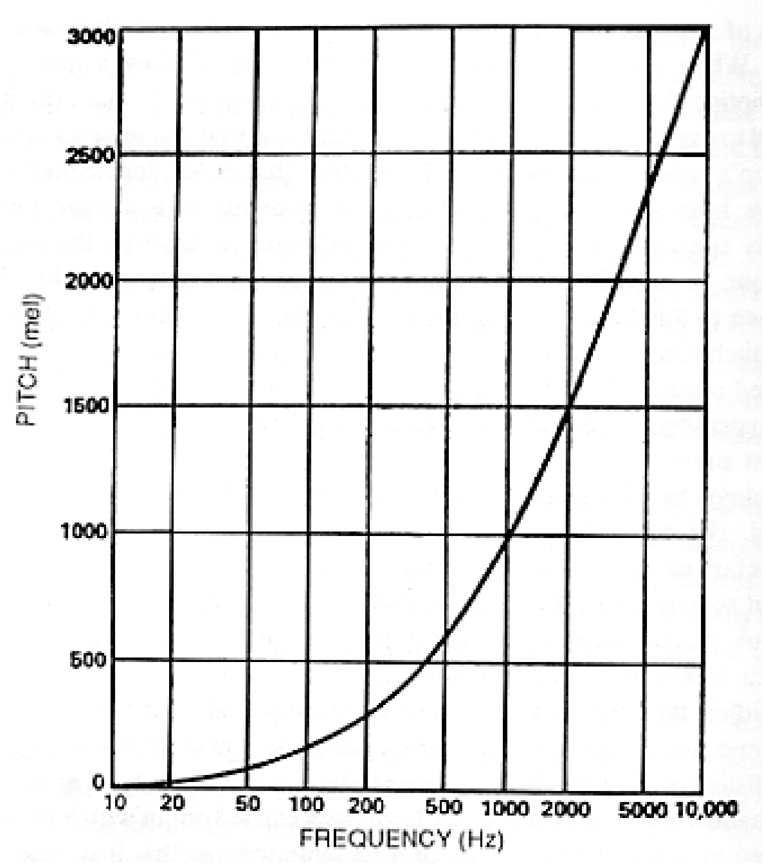


Yost, 1994

SAN DIEGO STATE UNIVERSITY

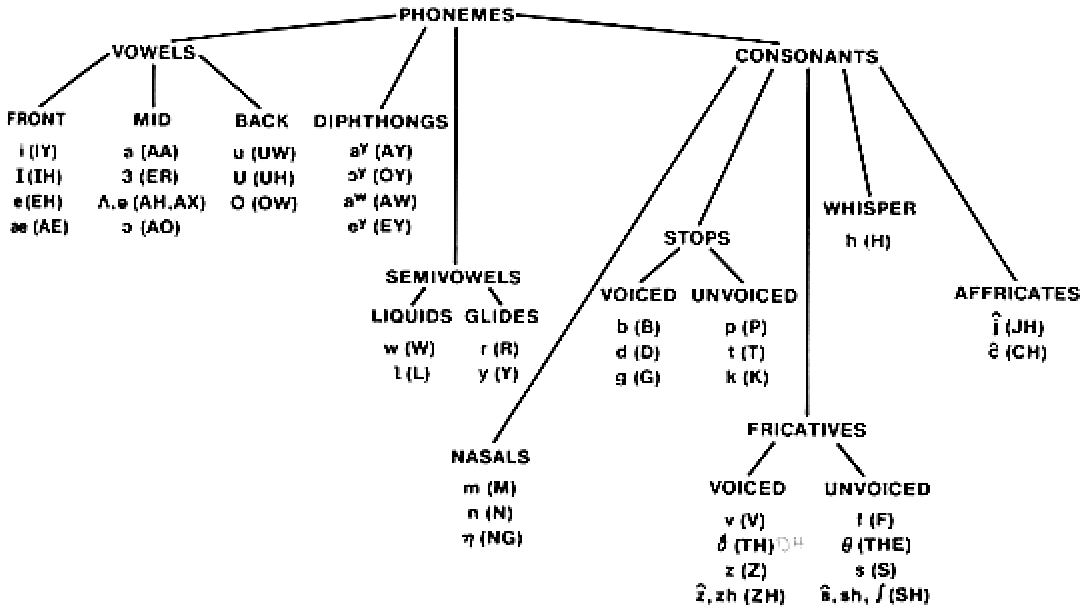
Mel Scale

- Subjective scale
- 2N mel seems twice as high pitched as N mel.



Sundberg, 1991

Classes of phonemes



Rabiner & Juang, p. 25

Phones are described with the international phonentic alphabet, or combinations of letters calls ARPABET. This figure contains IPA and an ARPABET variant.

Note that experts sometimes disagree on some of the classifications, e.g. OW.

Vowels

/ARPABET, IPA/
/iy, i/ f*ee*l, el*i*te, /ih,ɪ/ f*i*ll, /ae, æ/ g*a*s, /aa, ɑ/ f*a*ther,
/ah, ʌ/ c*u*t, /ao, ɑ/ d*o*g, /ax, ɜ/ c*o*mply, /eh, e/ p*e*t,
/er, ɝ/ t*u*rn, /uh, ʊ/ g*oo*d, /uw, u/ t*oo*l

- Phonemes whose phones are characterized by:
  - voicing
  - lack of major constrictions of the air
  - pharyngeal cavity produces F1, oral cavity F2
  - rounding the lips increases the oral cavity length, lowering F2

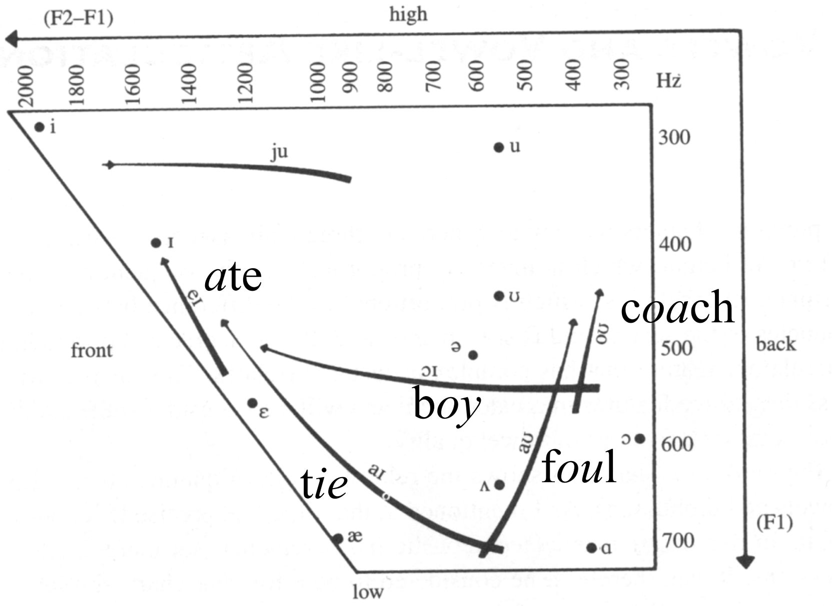
Diphthongs (vowels)

/ARPABET, IPA/

/ay, aɪ/ t*i*e, /ey, eɪ/ *a*te, /oy, ɔɪ/ c*oi*n, /aw, aʊ/, f*ou*l, /ow, oʊ/ c*oa*ch, /ow, oʊ/ t*o*ne

- Articulators start to form one vowel & move into another:

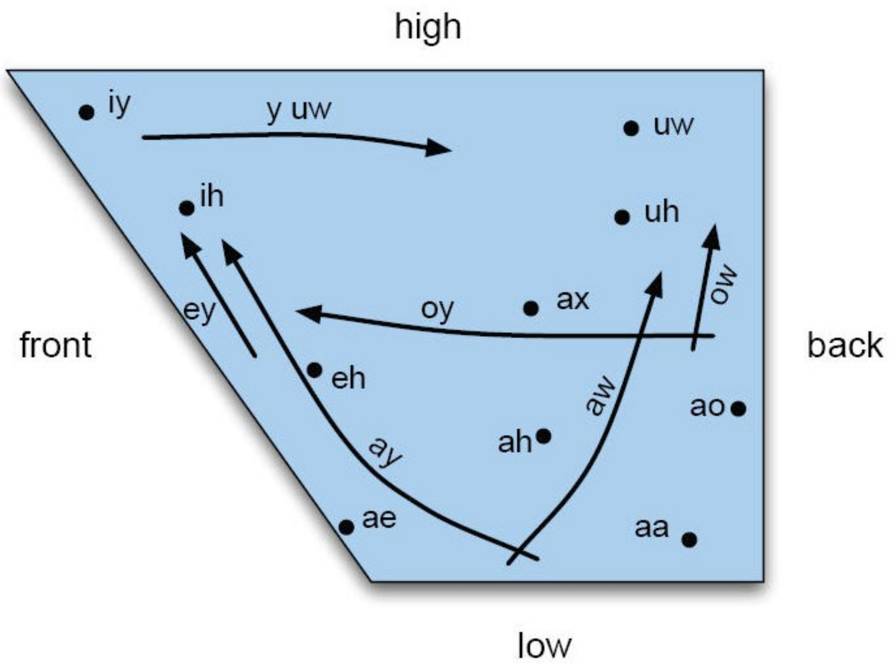
diphthong	from	to
/ay/ t*i*e	/aa/ f*a*ther	/iy/ *e*ve
/ey/ *a*te	/eh/ t*e*n	/iy/ *e*ve
/oy/ c*oi*n	/ao/ d*o*g	/iy/ *e*ve
/aw/ f*ou*l	/aa/ f*a*ther	/uw/ t*oo*l
/ow/ c*oa*ch		



Ladefoged, 2001, p. 200

Major articulators for vowels

- Tongue height
  - high (i.e. /iy, iː/ *e*ve)
  - versus low (i.e. /ae, æ/ *a*t)
- Tongue position
  - front (i.e. /iy, iː/ *e*ve)
  - back (i.e. /uh, ʊ/ b*oo*k)
- Lip rounding
  - flat (i.e. /iy, iː/ s*ee*)
  - rounded (i.e. /uw, u/ bl*ue*)



Jurafsky & Martin 2009, p. 223

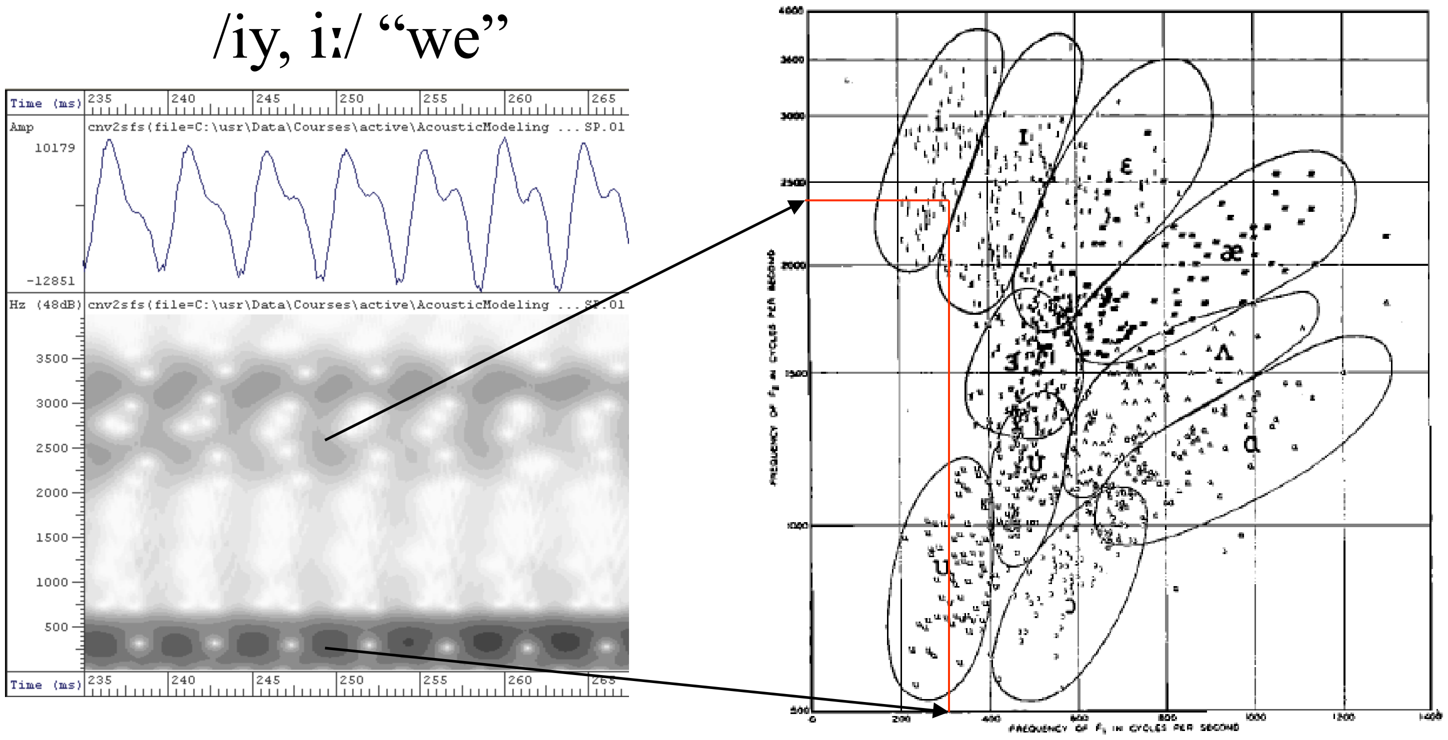
Vowels

- Vowels can typically be characterized by F1 & F2

/iy, iː/ “we”

F2~2400

F1~350



Peterson and Barney, 1952, p. 182

FIG. 8. Frequency of second formant *versus* frequency of first formant for ten vowels by 76 speakers.

SAN DIEGO STATE UNIVERSITY

Consonants

- Manner of articulation describes the major distinction between different consonant classes.

- Many consonants come in pairs, where the only difference between them is whether or not they are voiced.
  Example: /s/ vs. /z/

Note: Many IPA consonants are the same as for ARPABET.
Only one symbol is shown when they are identical.

Consonants: Approximants

- Voiced with less obstruction of the vocal tract than normal consonants:
  - Liquids (/l/ edib*le*, /r/ fa*r*) are very vowel-like and can even take the place of a vowel in a syllable. e.g., funnel /fʌnl/.
  - Glides (/y, j/ *y*ak, /w/ *w*alrus or *o*nce) are shortened & unstressed versions of the vowels /iy, iː/ *e*ve & /uw, u/ m*oo.*
- This manner is sometimes called semivowels

Consonants: Nasals

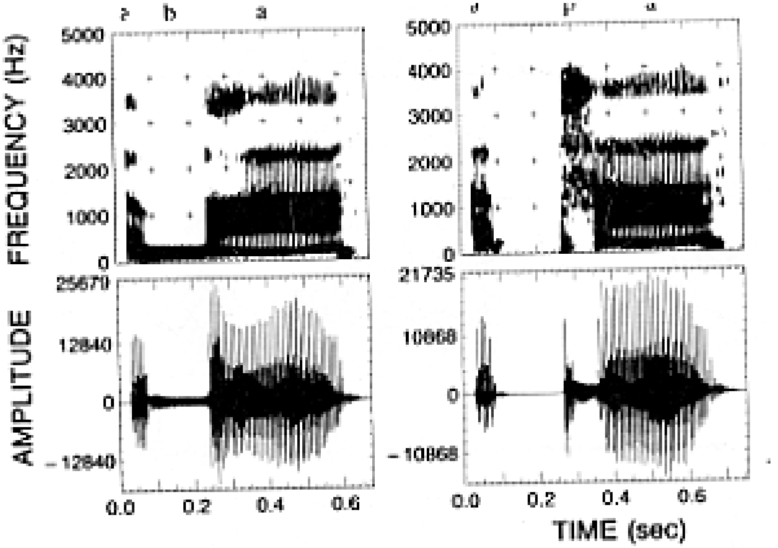
- Nasals, /m/ *m*ouse, /n/ *n*ose, /ng, ŋ/ thi*ng*, are characterized by:
  - Constriction of oral cavity making it difficult for air to pass through it.
  - Lowering of the velum, permitting air to move through the nasal passage.
- Semivowels, nasals, & vowels form the category of *sonorants*.

Consonants: Plosives (Stops)

- Complete blockage of the oral cavity
- Voiced & unvoiced pairs: /b/-/p/, /d/-/t/, /k/-/g/, /g/
- Easy to recognize in a spectrogram from the lack of energy right before the plosive.

/əbɔ/ vs. /əpɔ/

"uh-bah" vs. "uh-pah"



Rabiner & Juang, p. 38

Consonants: Fricatives

- Nearly complete closure of the vocal tract creates turbulent, noise like sound.
- Can be voiced or unvoiced:
  - /v/-/f/ *v*oiced, *f*ree
  - /dh, ð/ - /th, θ/ *th*en, ma*th*
  - /z/-/s/ mi*zz*en, *s*igh
  - /zh, ʒ/-/sh, ʃ/ *Zs*a-*Zs*a, *sh*eepi*sh*

Consonants: Affricates

- Combination: stop followed by a fricative
- voiced: /d/ + /zh, ʒ/ = /jh, dʒ/ a*g*ile
- unvoiced: /t/ + /sh, ʃ/ = /ch, tʃ/ *ch*eese

Distinctions between consonants

- We indicated that many consonants belong to the same classes which are determined by the *manner of articulation*
- What makes consonants within a class unique?

Place of articulation

- The distinction is caused by where the manner of articulation occurs.

Table 2.10 The consonants of English arranged by place (columns) and manner (rows).

	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar	Glottal
Plosive	*p b*			*t d*		*k g*	*?*
Nasal	*m*			*n*		*ng*	
Fricative		*f v*	*th dh*	*s z*	*sh zh*		*h*
Retroflex sonorant				*r*			
Lateral sonorant				*l*			
Glide	*w*				*y*		

Huang et al., 2001, p 47

Other languages

- Other subsets of the phonemes
  e.g. Spanish, French
- Use of pitch to distinguish phones
  e.g. Mandarin Chinese
- Use of vowel length
  e.g. Japanese

Allophones & Coarticulation

- Allophone – Phone which is recognizable even though it is atypical.
- Coarticulation
  - Surrounding phonemes affect production.
  - Try "pin" versus "spin" (The plosive /p/ is stronger in pin)
  - As speech rate increases, these effects will be more prominent.

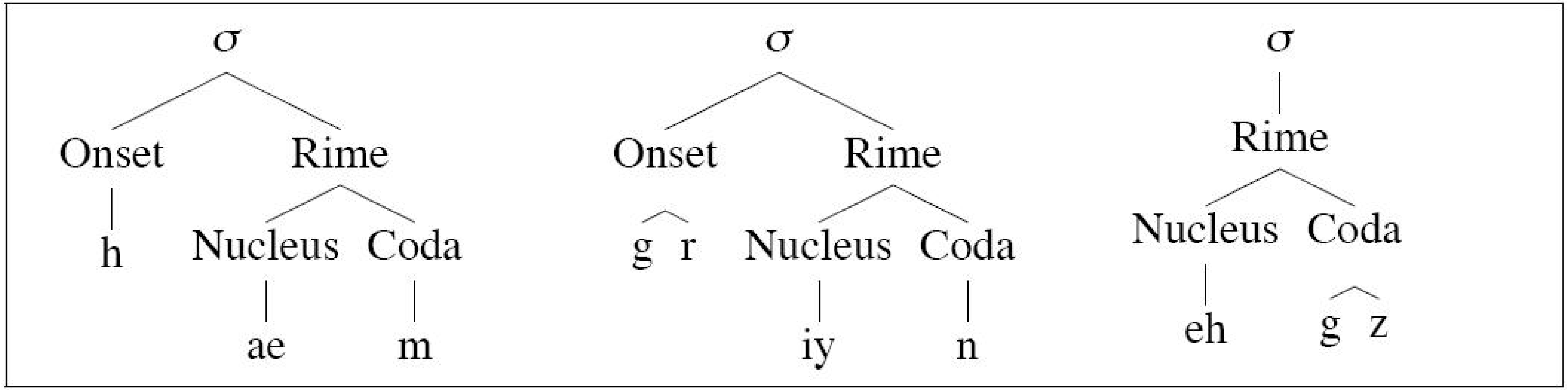
Insertions and Deletions

- We sometimes insert (epenthis) sounds:
  strength: [strɛŋkθ]

- Similarly, we can drop sounds
  e.g. alveolar stops between consonant pairs
  “last game” becomes [læs geɪm]

Syllables

ham green eggs



Jurafsky & Martin 2009, p. 223

Syllables

- Linguists consider *phonotactics*, rules about syllable construction

- In practice, not a serious issue for speech recognition systems as cross syllable boundaries are usually modeled.