Speech perception

And acoustic phonetics

Overview

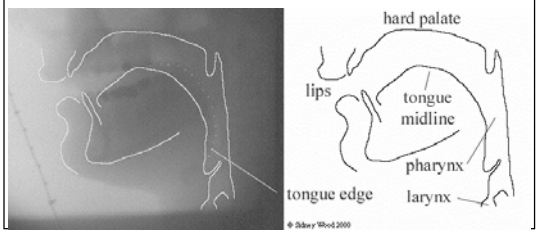
- Speech perception is relevant to many disorders & clinical groups, including:
  - Cleft palate
  - Articulation disorders
  - Phonological disorders
  - Hearing impairment
  - Cochlear implants
  - Dyslexia
  - Specific Language Impairment

Overview

- By the end of this section, you should understand:
  - Why one clinical treatment for dyslexia involves focusing on perception of stop consonants
  - Why individuals with sensorineural hearing loss have less problems hearing vowels than consonants
  - Why an individual with cleft palate cannot make the distinction between nasals & oral stops
  - Why cochlear implants, which only pass small amounts of the signal, can still be useful for speech.
  - Why someone with gross motor impairment (say, from a stroke), will be unable to produce some speech sounds.
  - Why second language learners often have particular difficulty with some sounds.

General overview

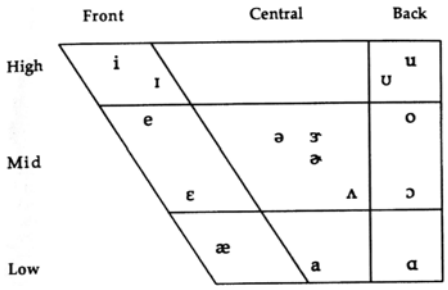
- Vowels vs. consonants
- Parts of system (midsagittal tracing)



Vowel types

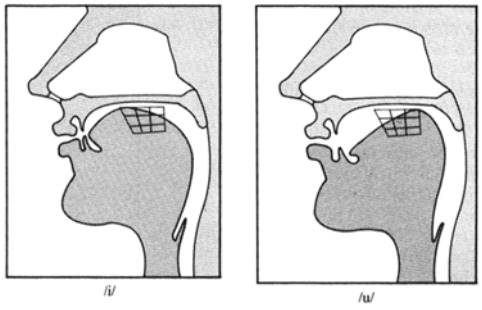
- Tongue height
- Tongue frontness/backness
- Rounding
- Tense/lax

Vowel quadrangle



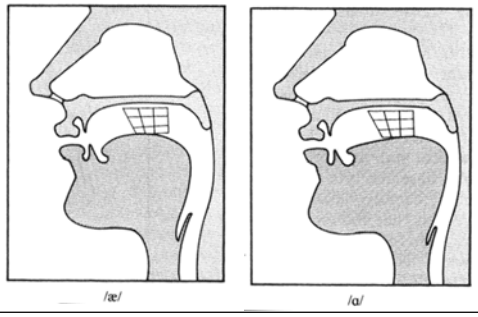
Vowel quadrangle, cont

/i/ vs. /u/



Source: I. Mackay, (1987) *Phonetics: The science of speech production*, 2nd ed.

/æ/ vs. /a/



Source: I. Mackay, (1987) *Phonetics: The science of speech production*, 2nd ed.

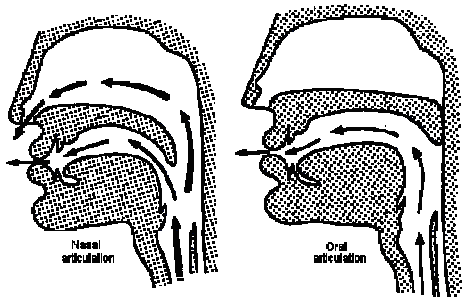
Consonants

- Manner of articulation
- Place of articulation
- Voicing

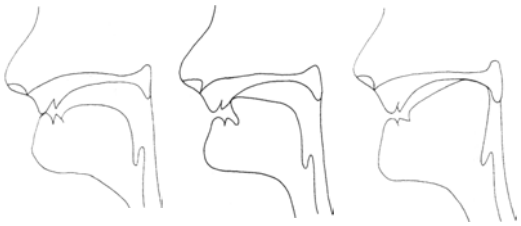
Manner of articulation

- Stop consonants
- Fricatives
- Affricates
- Nasals
- Glides
- Liquids

Nasals vs. Orals



Place of articulation



Place of articulation

- Bilabial
  - At lips
  - p, b, w, m
- Labidental
  - Lips & teeth
  - f, v
- Interdental
  - Between teet
  - th (soft & hard)
- Alveolar
  - Tongue behind teeth
  - t, d, s, z, n, l, r
- Palatal
  - Tongue against hard palate
  - sh, zh, ch, dj, y
- Velar
  - Tongue against back of mouth
  - k, g, ng

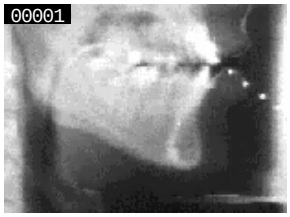
Voicing

- Source of sound, rather than location or type of constriction
  - voiceless sounds: the vocal folds are held wide open, and the air passes through the throat unimpeded.
  - voiced sounds: the vocal folds close together, blocking the air.

A clinical issue

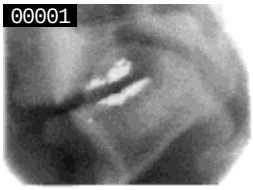
- Voiced stop consonants (b,d,g) are some of the shortest sounds in the language.
- One proposal: auditory processing deficits that prevent children from distinguishing among these fast sounds cause a variety of clinical disorders, esp. dyslexia

"Why did Ken set the soggy net on top of his deck"



Moviefrom K. Munhall, x-ray Film Database

“It’s 10 below outside”



Moviefrom K. Munhall, x-ray Film Database

“Try not to annoy her”

Moviefrom K. Munhall, x-ray Film Database

Vocal fold vibration

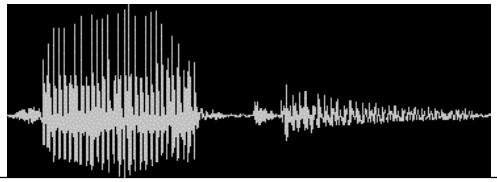
- Rate at which the vocal folds open & close is the fundamental frequency of the signal or F0.
- This is heard as a difference in pitch.
- Gender differences

Slow-motion of the vocal folds vibrating during speech

- Link

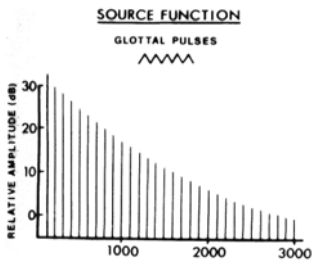
Speech waveform

- One way we can see speech is on a speech waveform
- Time is on the x-axis, & displacement of air on the y-axis. This is the syllable /adi/.
- Each vertical line is one opening/closing of the vocal folds.



Sound source

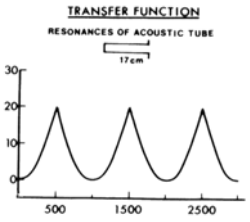
- Signal contains energy at each multiple of F0
  - These are called harmonics



Source: G J Borden & K S Harris (1984). *Speech science primer: Physiology, acoustics, and perception of speech*. 2nd ed,

Transfer function

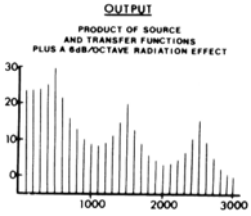
- The shape of the vocal tract determines what sounds are allowed to pass through.
- A wide open shape (such as for /^/) emphasizes frequencies at three evenly spaced points



Source: G J Borden & K S Harris (1984). *Speech science primer: Physiology, acoustics, and perception of speech*. 2nd ed,

Output function

- The combination of that vocal tract shape, and that glottal source, result in an output like this.
- This gets heard as the vowel /^/.



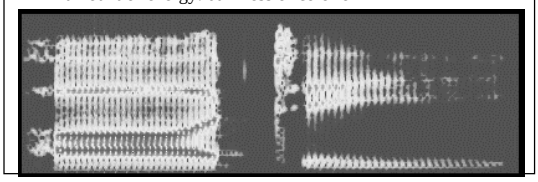
Source: G J Borden & K S Harris (1984). *Speech science primer: Physiology, acoustics, and perception of speech*. 2nd ed,

Resonances

- During speech, you move your tongue, changing the vocal tract shape.
- This results in different resonances.
- The band of resonant frequencies is called a formant.

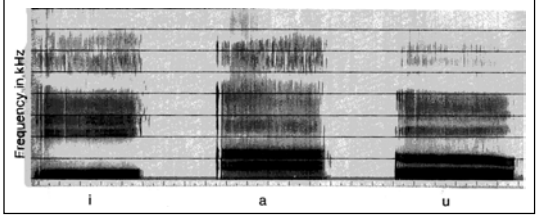
Speech Spectrogram

- Waveforms do not allow us to see formants.
- Spectrogram
  - time on the x-axis
  - frequency on the y-axis
  - amount of energy: darkness or color of ink

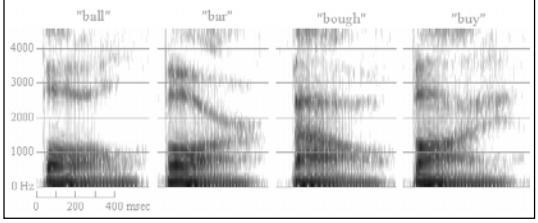


Formants

- First three formants are the most important cue to speech identity for vowels and some consonants (such as stops)



Formant transitions



Frequency range

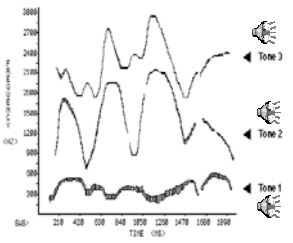
- Because first three formants are most important for distinguishing vowels & stop consonants, and they occur in 0-3000 Hz range, these sounds are more likely to be heard by someone with a hearing impairment.
- Voiceless fricatives tend to have energy in the 3000 - 8000 Hz range.

Synthetic speech

- We can measure what energy is in normal speech, and copy that to a computer
- We can then make slight changes to it and see how this affects perception

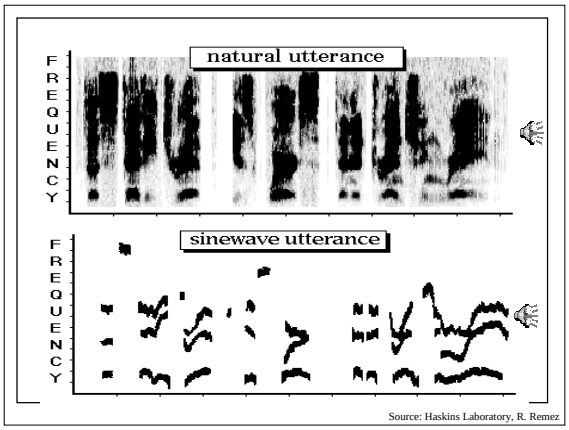
Sine wave analogs to speech

- Has a simple tone instead of each of the first three formants
- Doesn't sound like speech
- Can be heard as speech



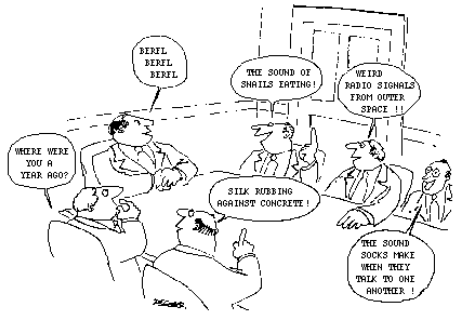
Complete Sine wave

Source: Haskins Laboratory, R. Remez



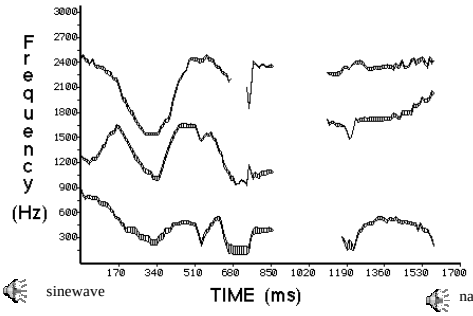
Source: Haskins Laboratory, R. Remez

How people normally hear this

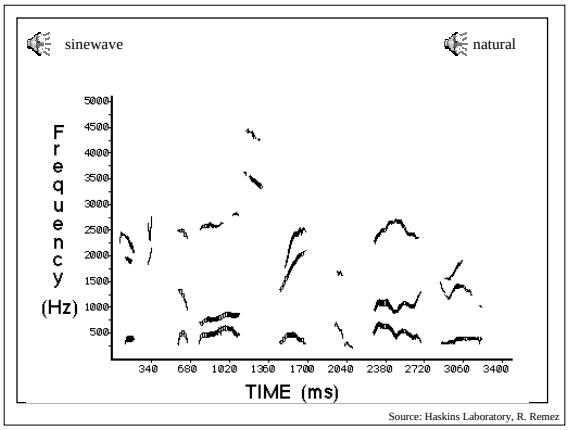


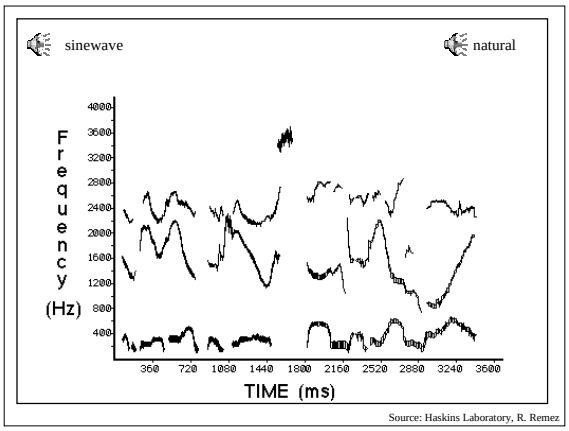
Source: Haskins Laboratory, R. Remez

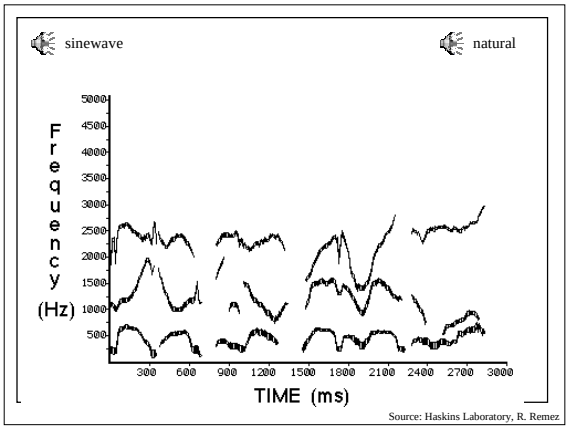
Some more examples….

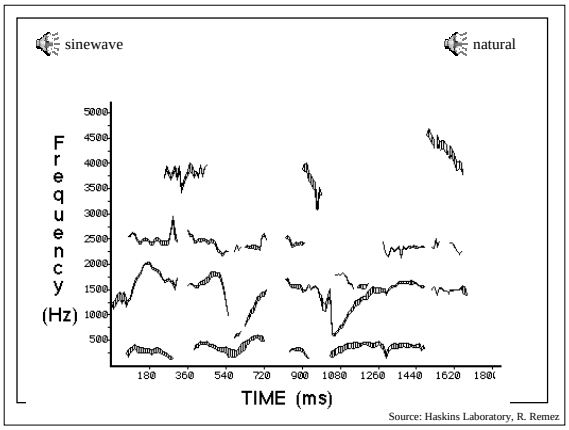


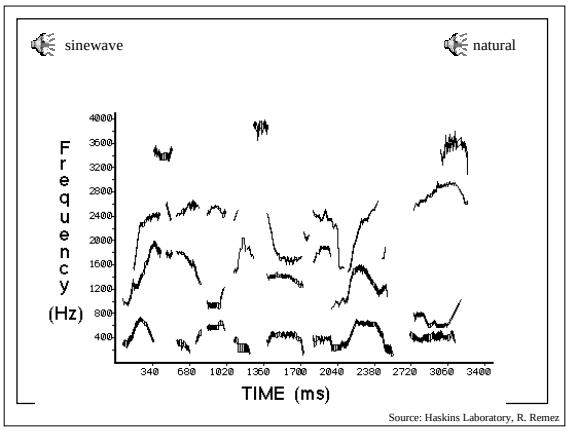
Source: Haskins Laboratory, R. Remez


sinewave
natural
Frequency (Hz)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
340 680 1020 1360 1700 2040 2380 2720 3060 3400
TIME (ms)
Source: Haskins Laboratory, R. Remez


sinewave
natural
Frequency (Hz)
4000
3600
3200
2800
2400
2000
1600
1200
800
400
360 720 1080 1440 1800 2160 2520 2880 3240 3600
TIME (ms)
Source: Haskins Laboratory, R. Remez

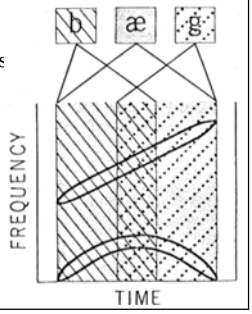

sinewave
natural
Frequency (Hz)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
300 600 900 1200 1500 1800 2100 2400 2700 3000
TIME (ms)
Source: Haskins Laboratory, R. Remez





Issues in speech perception

- Lack of invariance
  - Variability across genders & talkers
  - Contextual variability
  - Coarticulation



Source: Liberman, A.

Words taken out of context

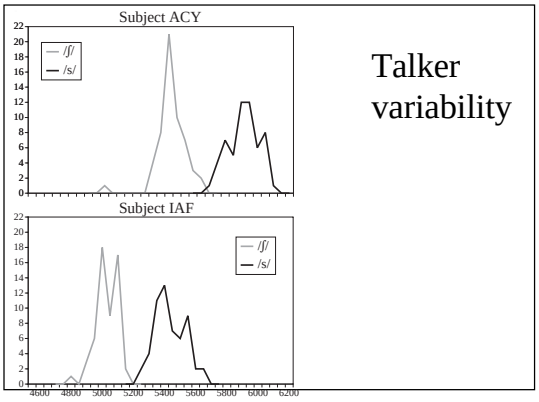
- Try to identify these words; each is repeated three times.
- There are 34 items.

Words taken out of context

- 1. Like
- 2. At
- 3. Home
- 4. Box
- 5. For
- 6. Get
- 7. Phone
- 8. Put
- 9. Hand
- 10. Box
- 11. Tape
- 12. Don't
- 13. Nice
- 14. Stay
- 15. Down
- 16. There
- 17. See
- 18. Box
- 19. Toys
- 20. Books
- 21. Doll
- 22. Comb
- 23. Ball
- 24. Have
- 25. Door
- 26. Can
- 27. Go
- 28. Go
- 29. Shoes
- 30. Books
- 31. Can
- 32. Sit
- 33. Floor
- 34. Play

Talker normalization

- Different individuals produce the same sound in different ways.
- Because of this, different phoneme categories overlap.
- We need to interpret speech in reference to the talker.



Talker variability

Adjusting for variability

- Mullennix, Pisoni, and Martin
  - Identification was more accurate and naming was faster for a single-talker condition
- Magnuson et al.
  - Same results when voices are spouse & children
- Sommers, Nygaard, and Pisoni
  - Similar decrements for rate variability

- Adjusting for variation requires cognitive resources, which may be why it is particularly problematic for older individuals & those with hearing impairments

Phoneme restoration

- Richard Warren

The state governors met with their respective legislatures convening in the capital city.

- A cough replaced the first /s/ in legislatures.
- He asked Ss where the cough occurred.

Phoneme restoration, cont.

- Another example: Warren presented a sentence like

  It was found that the #eel was on the _____,

  – # was the noise.
  – The last word of the sentence could be “axle”, “table”, “shoe” or “orange”
  – People heard the word as whichever was most appropriate: wheel, meal, heel, or peel.

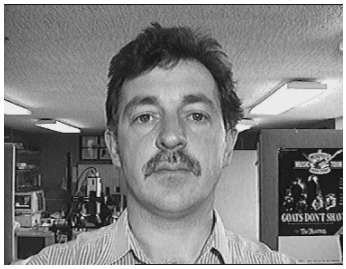
Mispronunciation detection

- People in seldom caught mispronunciations that differed by only a single feature.
- For mispronunciations that differed in several features, it depended on WHERE it occurred.

What do these findings mean?

- Speech perception is not based only on the signal – it is also influenced by your prior knowledge of the language.
- Thus, speech involves top-down processing as well as bottom-up processing.
- Poor cognitive processing will limit speech perception!

McGurk effect



Second example

Source: www.media.uio.no/personer/arntm/McGurk_english.html

Third example

- Link

Source: Lawrence D. Rosenblum www.psych.ucr.edu/avspeech/lab

McGurk & MacDonald study

- Combined an auditory “ba” with a visual “ga”
- People heard a fusion of the two signals, the syllable “da”.

McGurk effect in infants

- Saw & heard a talker saying “va va va.”
- After they’d gotten bored with (habituated to) that, one of three things happened:
  - It stayed the same (infants should remain bored)
  - It changed; the face said “va va va” but the voice said “ba ba ba” (adults hear this as “va”)
  - It changed; the face said “va va va” but the voice said “da da da” (adults hear this as “da”)
- Infants dishabituated to the last, but not the first two -- so they perceive these like adults