LING 520 Introduction to Phonetics I
Fall 2008

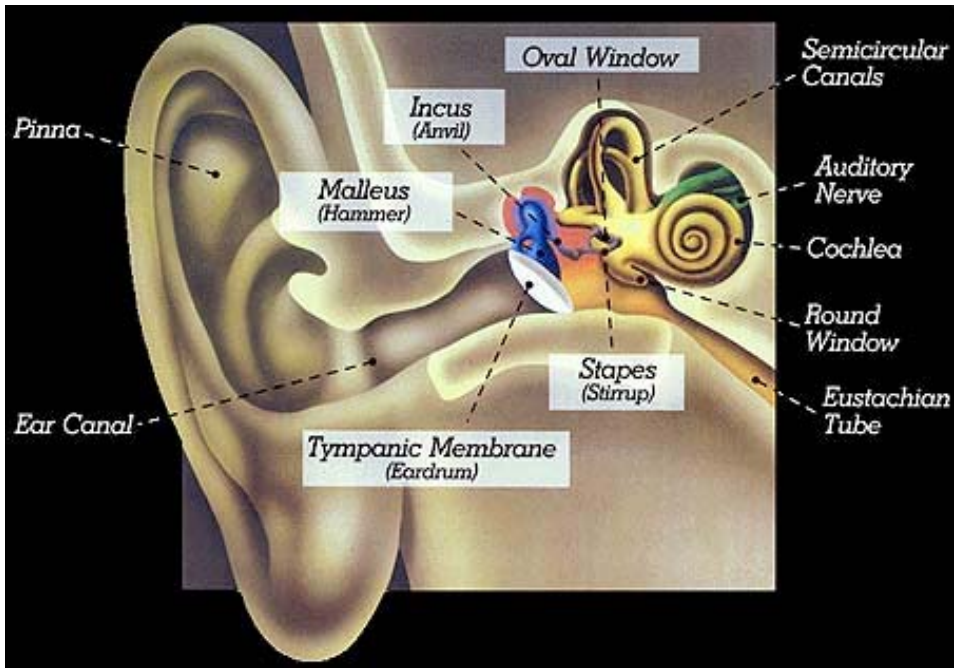
Week 9

Basic audition

Speech perception

Nov. 3, 2008

Auditory physiology



[From: www.kemt.fei.tuke.sk]

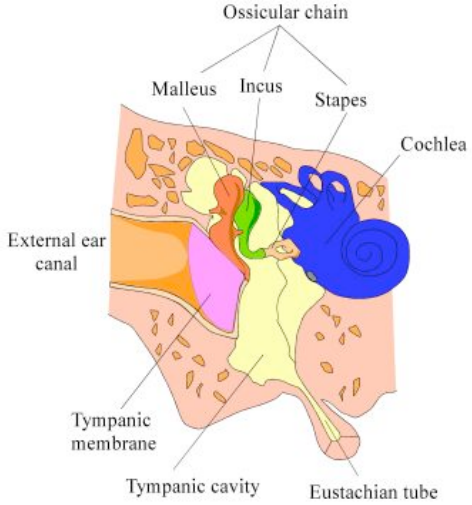
- Outer ear:

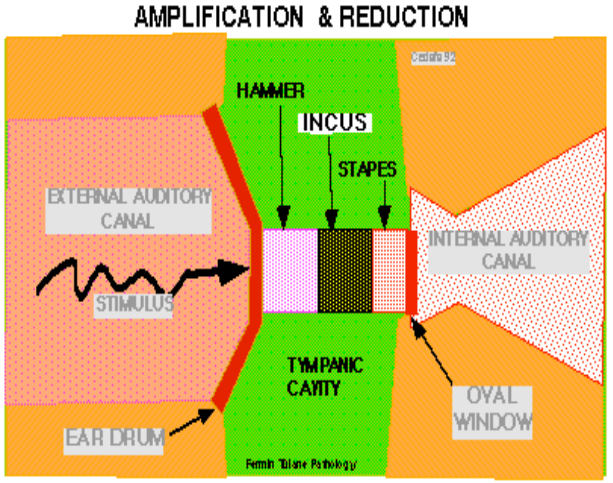
  The outer ear consists of the pinna and the auditory canal (external auditory meatus), which create a broad resonance resulting in an approximately 10 to 15 dB of amplification of the spectrum between 2.5 and 5 kHz (The auditory canal is a resonator, and can be compared to a uniform tube closed at one end and open at the other (~ 2.5 cm)).

Auditory physiology

- **Middle ear:**

The auditory canal ends at the eardrum (the tympanic membrane). The sound pressure at the drum displaces the drum, which in turn causes displacement of three exceedingly small bones (ossicles, or ossicular chain). The mechanical vibrations of the auditory ossicles are transmitted to the oval window, a membrane that covers the opening to the inner-ear cochlea. This membrane is about one-fifteenth the area of the tympanic membrane. Because of this area ratio and the tight linkage of the ossicular bones, the middle ear amplifies sound approximately 20 dB.



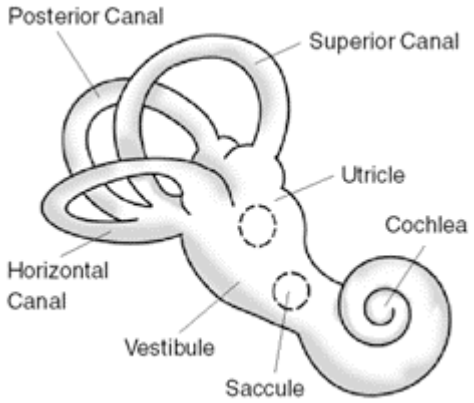


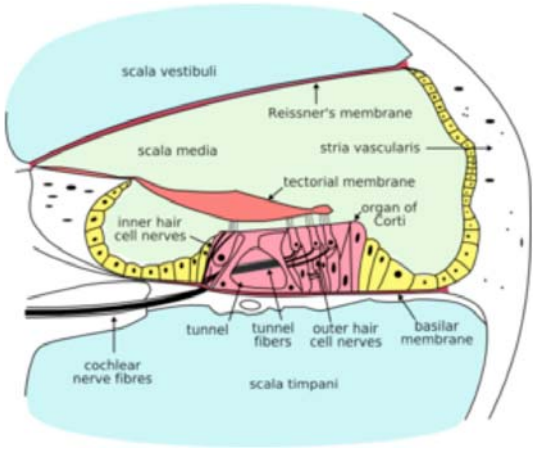
[From: www.som.tulane.edu; www.wadalab.mech.tohoku.ac.jp]

Auditory physiology

- **Inner ear:**

  The inner ear is a system of cavities in the bones of the skull which influence balance as well as hearing. The cavity that houses the sensory receptor for hearing is the cochlea, where the mechanical vibrations of the middle ear and oval window are transformed into nerve impulses. The cochlea is a fluid–filled coiled cavity. The vibration of the endplate of the stapes against the oval window results in pressure waves in the cochlear fluid, which in turn set the cochlear duct ––and the basilar membrane within the duct ––into vibration.



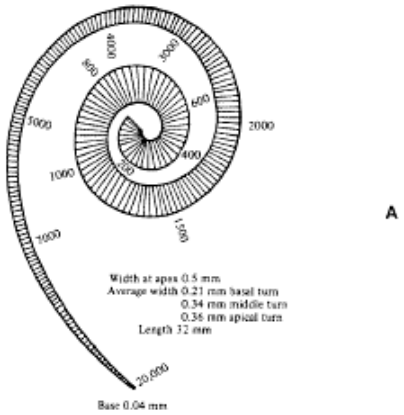


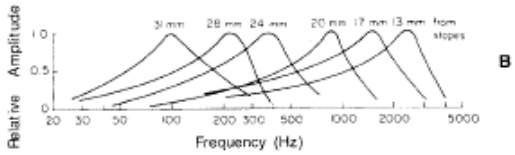
[From: en.wikipedia.org]

Auditory physiology

- **How the inner-ear recognizes sounds?**

  The basilar membrane is narrow and stiff at the basal (oval window) end, where it responds with greatest amplitude to high frequencies. At the apical end, where it is thicker and less stiff, the greatest amplitude of response is to low frequencies. Thus the basilar membrane is a spectrum analyzer, performing a kind of Fourier analysis on input complex waves albeit with limited power of resolution.





[From: *Principles of Experimental phonetics*]

For his seminal work in the biophysics of hearing, von Bèkèsy was awarded the Nobel Prize in Physiology or Medicine in 1961.

Auditory physiology

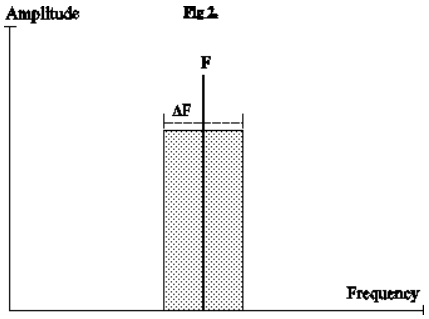
- Why the inner-ear is snail-shaped?

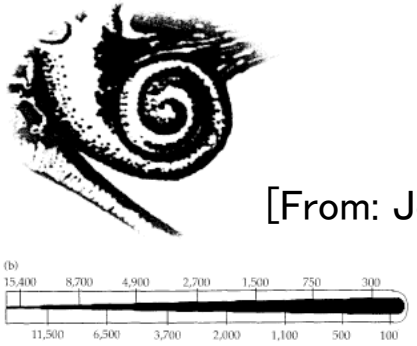
  Too boost sensitivity to low frequencies: Although the spiral shape of the cochlea had little impact on the average vibrational energy traveling along the tube, as the wave progresses, this energy increasingly accumulates near the outside edge of the spiral, rather than remaining evenly spread across it. Low frequencies travel the furthest into the spiral, so the effect is strongest for them.

http://focus.aps.org/story/v17/st8

Critical Bands

- The notion of critical bands, introduced by Fletcher (1940), explains the masking of a narrow band (sinusoidal) signal by a wideband noise source. Fletcher presented listeners with a pure tone signal plus a noise masker whose bandwidth (BW) was varied. (The noise was centered at the frequency of the pure tone.) He found that as the noise BW increased, so did the signal threshold, and what's interesting is that this finding held only up to a certain point: at some noise BW, the threshold function flattens off and further increases in noise BW do not affect the signal threshold.
- Physiologically, each critical band corresponds to a distance on the basilar membrane (about 1.3 mm). Since a larger portion of the basilar membrane responds to low frequencies than to higher frequencies, human listeners are more sensitive to differences in the lower than in the higher frequencies.





[From: Johnson 2003]

Pitch and Frequency

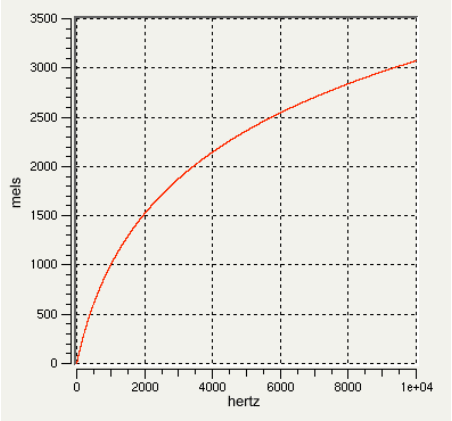
- The Bark scale:

  The Bark scale ranges from 1 to 24 Barks, corresponding to the first 24 critical bands of hearing.

$$\text{Criticalbandrate(bark)} = [26.81/(1 + 1960/f)] - 0.53$$

- The Mel scale:

  The Mel scale is based on experiments with pure tones in which listeners adjust the frequency of a test tone to be half as high (or twice as high) as that of a comparison tone (*1000 mel = pitch of 1000 Hz tone).* The Mel scale corresponds closely to the Hz scale up to ~500 Hz. At higher frequencies, the mel scale is more nearly logarithmic.



$$m = 1127.01048 \log_e(1 + f/700)$$

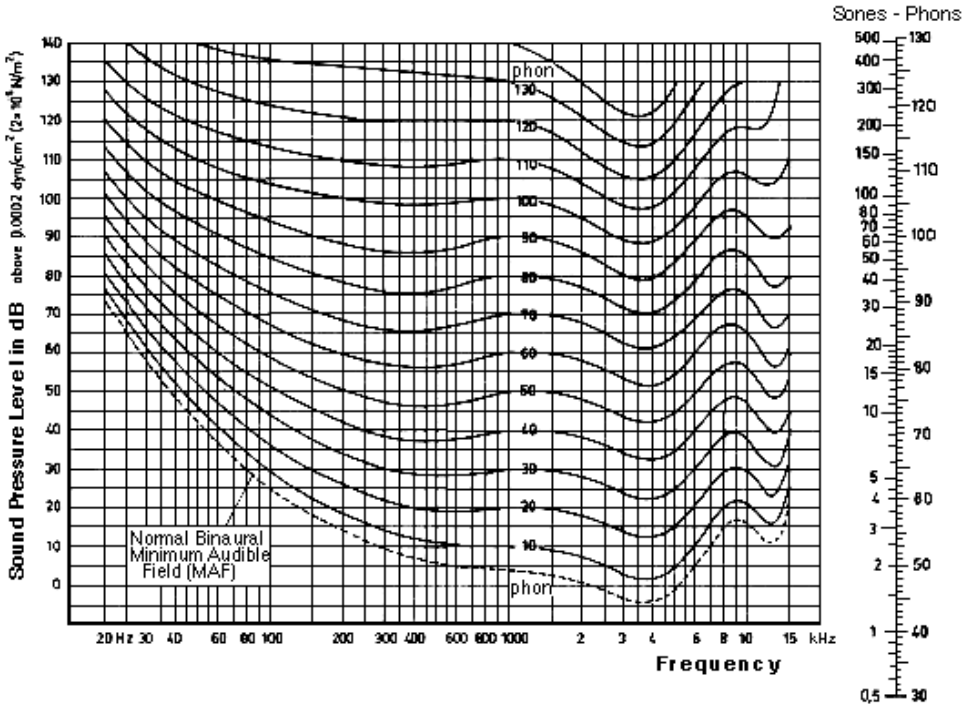
Loudness and Intensity

- **Decibel:** Unit of measurement of relative intensity of a sound, compared to an arbitrary reference point. The decibel scale is not sensitive to the effects of frequency on the sensation of loudness. For example, a 300 Hz and a 3000 Hz tone at 50 dB differ substantially in loudness.
- **Phon:** The phon scale is determined by having listeners adjust the intensity of a tone at a different frequency until it has the same loudness as a tone of a 1000Hz. Sounds judged to have equal loudness in this way are assigned the same "phon" value (e.g., all tones judged as having the same loudness as a 20 dB 1000 Hz tone have a loudness of 20 phons). It cannot be used to measure relationships between sounds of different loudness. For instance, 40 phons is not twice as loud as 20 phons.
- **Sone**: For the purpose of measuring sounds of different loudness, the sone scale of subjective loudness was invented. The sone scale is determined by having listeners adjust the loudness of a tone until it is twice as loud, or half as loud, as another tone. 1 sone = loudness of a 40 dB 1000 Hz tone. 2 sones = sound judged to be 2x as loud as this.

Loudness and Intensity

- An increase of 10 phons is sufficient to produce the impression that a sine tone is twice as loud. One sone is 40 phons *at any frequency*. Two sones are twice as loud, e.g. 40 + 10 phons = 50 phons. Four sones are twice as loud again, e.g. 50 + 10 phons = 60 phons. The relationship between phons and sones is expressed by the equation:

Phon = 40 + 10 log2 (Sone)



Just Noticeable Differences (JNDs)

- **JNDs:** Threshold of difference in frequency, intensity, or duration can the human auditory system detect.

- **Intensity JNDs:**

  1. **Threshold of audibility:**

     Pure tone of 1000 Hz: 0 dB ($2 * 10^{-5}$ newtons/m^2)

     Wideband noise at amplitudes in the speech range: 0.3–1.0 dB

  2. The intensity jnd is about 1 dB for soft sounds around 30-40 dB at low and midrange freqencies. It may drop to 1/3 to 1/2 a decibel for loud sounds.

- **Pitch JNDs:**

  1. Pure tones (normal listening levels):
     ~ 1 Hz for frequencies up to 1 kHz;
     ~ 2 Hz at about 2 kHz and ~ 4 Hz at about 4 k Hz
     Increases rapidly above 5 kHz

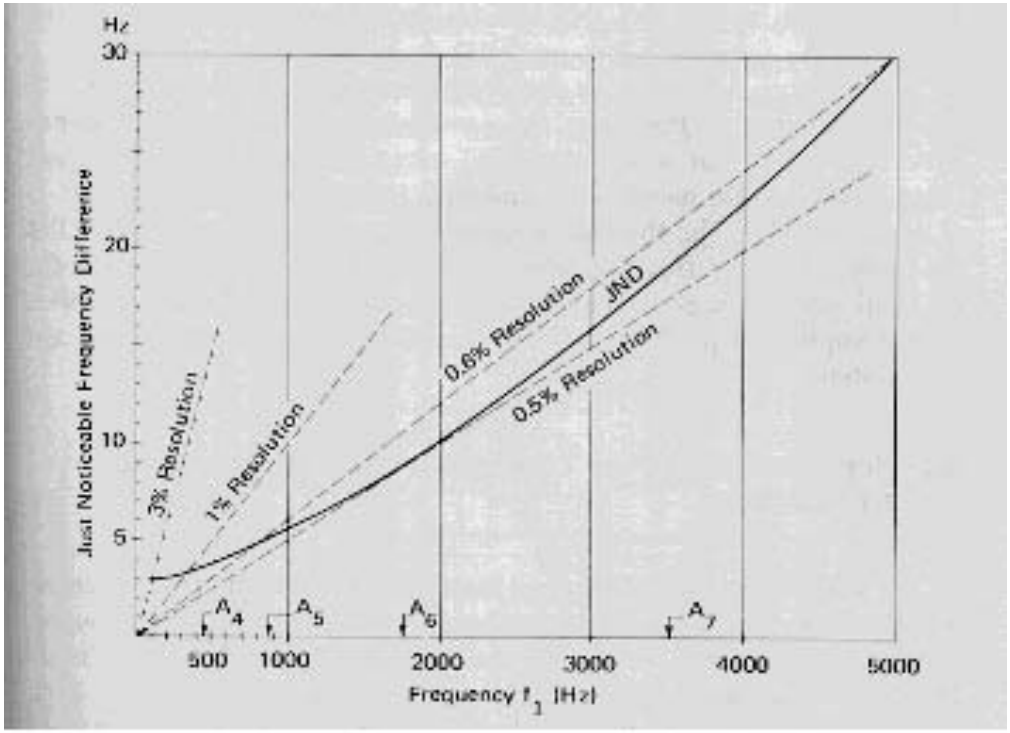
  2. F2: 20–100 Hz, depending on the F1–F2 or F2–F3 distance

Just Noticeable Differences (JNDs)

- **Pitch JNDs:**

  1. Figure below shows the average JND in frequency for pure tones of constant intensity (80 dB) whose frequency was slowly and continuously modulated up and down.

  2. The JND is about 5 Hz for a pure tone of 1000 Hz and about 20 Hz for a pure tone of 2000 Hz.



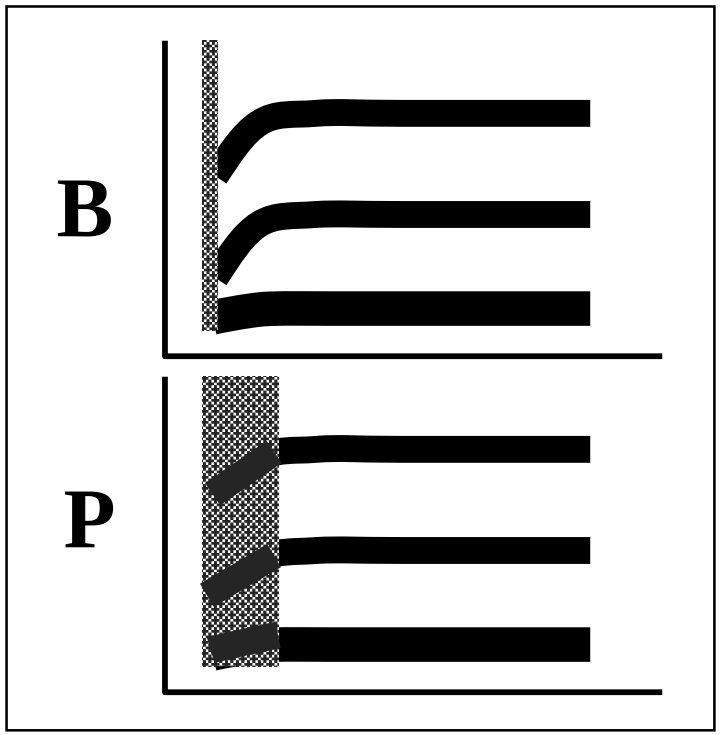
[Roederer, 1973]

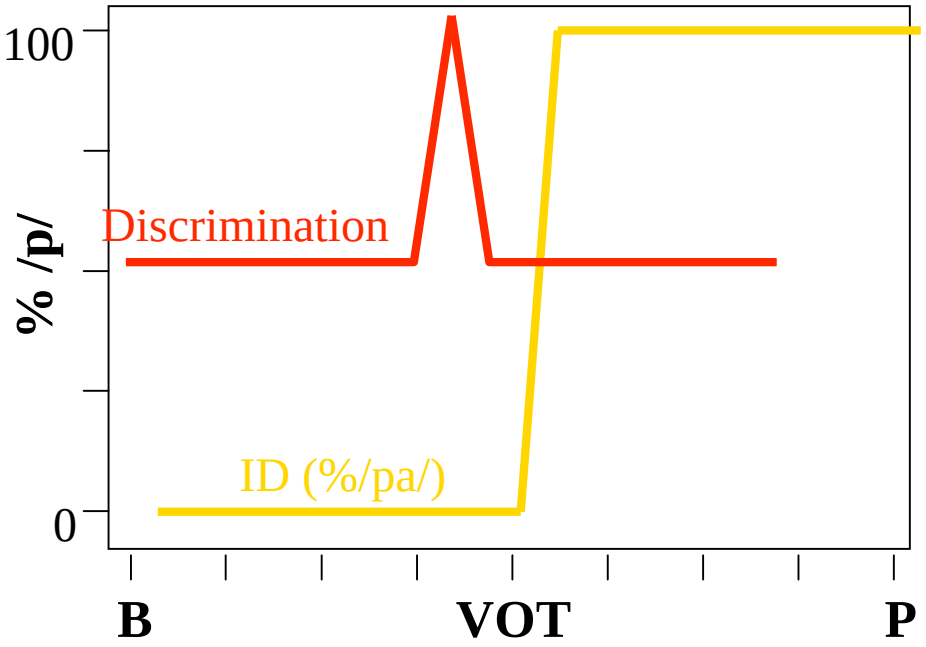
Specialization of speech perception?

- **Categorical perception**
- **Duplex perception**
- **Trading relations and integration of cues**
- **The McGurk effect: cross-modal cue integration**

  [Following slides follow Goldinger et al., "Speech perception and spoken word recognition: research and theory", in *Principles of Experimental Phonetics*, Norman Lass (ed.), 1996.]

Categorical perception





[graphs from: McMurray et al. slides]

- A change in some variable along a continuum is perceived, not as gradual but as instances of discrete categories;
- Discrimination between stimuli is much more accurate between categories than within them.

Categorical perception

- Categorical perception was one of the most important evidence for a speech mode of perception, motor theory of speech perception (Liberman et al.).
- But: Are Nonspeech stimuli continuously perceived? Is synthetic speech real speech? How about speech perception in nonhumans?

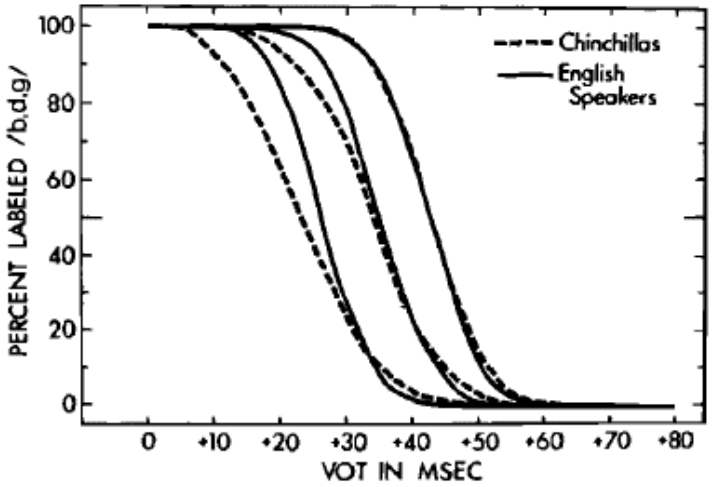
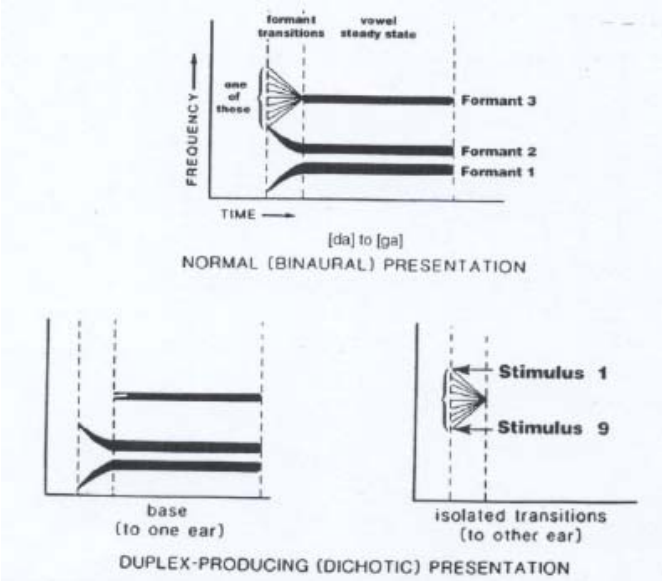


FIG. 10. A composite of the mean identification functions for chinchilla and human subjects that were obtained with bilabial, alveolar, and velar synthetic stimuli. No significant differences between species on the absolute values of the phonetic boundaries were obtained, but chinchillas produced identification functions that were slightly, but significantly, less steep.

[From: Kuhl & Miller, 1978]

Duplex perception

- Discovered by Rand (1974): A listener is presented with two simultaneous, dichotic stimuli. One ear hears an isolated third-formant transition that sounds like a nonspeech chirp. At the same time the other ear receives a base syllable. This base syllable consists of the first two formants, complete with formant transitions, and the third formant without a transition.
- The listener's percept is duplex, that is, the completed syllable is perceived and the nonspeech chirp is heard at the same time.



[From: Mann and Liberman, 1983]

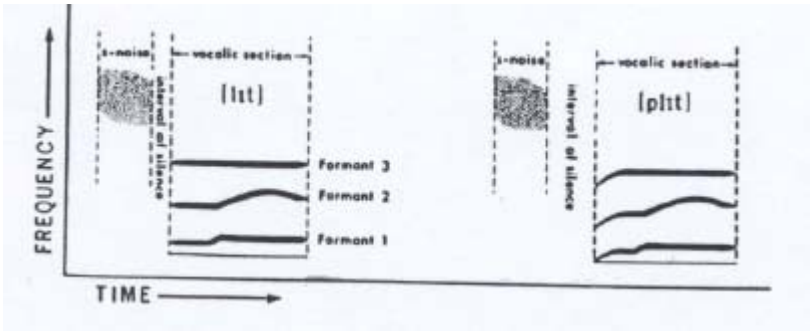
Duplex perception

- Cited as strong evidence for a dissociation of phonetic perception from general auditory perception.
  - "Duplex perception phenomena provide evidence for the distinction between auditory and phonetic modes of perception. They show that, in the duplex situation, the auditory mode can gain access to the input from the individual ears, whereas the phonetic mode operates on the combined input from both ears" (Repp 1982)
  - The "phonological fusion" discovered by Day (1968) - two dichotic utterances such as "banket" and "lanket" yield the percept "blanket" - is yet another example of the abstract, nonauditory level of integration that categorizes the phonetic mode.

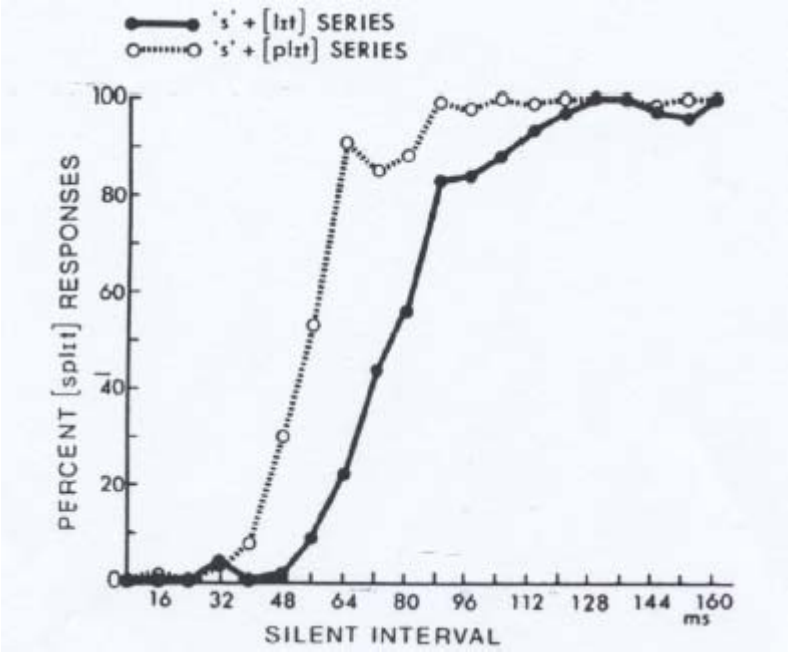
- Challenge:
  - Our perceptual and cognitive systems are attuned to perceive meaning from any simulation. Duplex perception would occur whenever two acoustic fragments, when integrated, specify a natural event (speech or non-speech) and when one of the fragments has any unnatural quality (Fowler and Rosenblum 1990).

Trading relations and integration of cues

- The speech signal is replete with cues to phonetic contrasts, and several different cues may indicate a single contrast.
- This makes it possible that when the utility of one cue is reduced, another cue becomes primary.



[From: Fitch et al. 1980]



Trading relations and integration of cues

- Cited as evidence of a speech mode of perception for two main reasons (Repp 1982):
  - It is difficult to imagine that such cues (temporal and spectral) would be integrated into a single percept unless some speech specific system were mediating perception.
  - Trading relations may occur because listeners perceive speech in terms of the underlying articulation and resolve inconsistencies in the acoustic formation by perceiving the most plausible articulatory act.
- Challenge:
  - Fuzzy logical model of perception can account for trading relations while making no assumptions of specialized processing of speech.
  - "once the prototypical patterns are known in any perceptual domain, trading relations follow as the inevitable product of a general pattern matching operation. Thus, speech perception is the application of general perceptual principles to very special patterns." (Repp 1983)

The McGurk effect

- A subject is presented with a video display of a talker articulating dimple CV syllables and heats spoken syllables synchronized with the visual display.
- The listener typically reports hearing neither the spoken syllable not the lip-read syllable, but something in between.

[From: http://www.brl.ntt.co.jp/IllusionForum/basics/auditory/mcgurk-e.html]

The McGurk effect

- The McGurk illusion has been interpreted as particularly strong evidence for a specialized speech perceptual system that makes reference to articulatory gestures.
  - "Why does integration occur? One answer is that both sources of information, the optical and the acoustic, provide information about the same event of talking, and they do so by providing information about the talker's phonetic gestures." (Fowler and Rosenblum 1991)
  - Infants prefer to watch a display of an articulating face if the accompanying spoken syllables match the articulation rather than incongruent audiovisual displays (Kuhl and Melzoff 1982).
  - Selective adaptation: Present subjects with an auditory syllable /be/ and visual /ge/, producing the percept of /de/. However, the perceived audiovisual syllable had the same effect as a purely auditory /be/ on a /be/–/de/ series; subject's phonetic perception of the stimulus as /de/ was not reflected in their adaptation data. (Roberts and Summerfield, 1981)
- Challenge:
  - The pathway from audition to phonetic perception is composed of processing stages. The integration of information from vision and audition occurs somewhat late in the speech perception process.

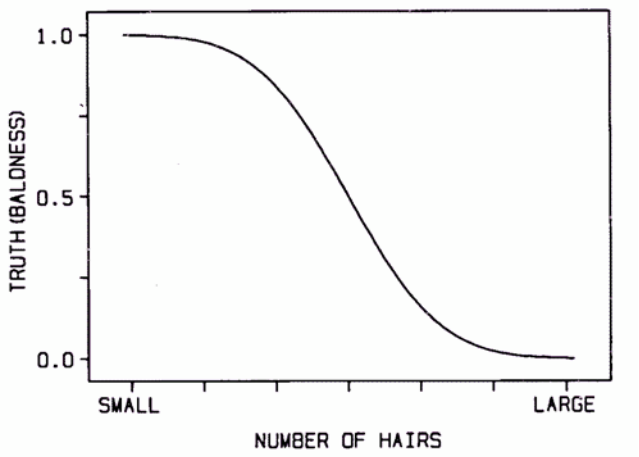
Motor theory of speech perception

- Proposed by Liberman et al.
- Listeners interpret speech sounds in terms of
  - motoric gestures they would make them with (1967)
  - intended gestures of the speaker (1985)
- Gestural unit: 'phonetic category'
- The motor theory can account for the invariance problem; that is, the ways that phonemes are produced and perceived have more in common than the ways they are acoustically represented and perceived.
- Are there many-to-one mappings from production to perception?

Fuzzy logical model of perception

- Proposed by Massaro et al.
- It was developed to account for feature integration in speech perception, regardless of the nature of the relevant features.
- FLMP assumes three operations in phoneme identification: *feature evaluation, prototype matching, and pattern classification*.
- *Features* are assigned continuous, 'fuzzy' values ranging from 0 to 1, indicating the degree of certainty that the feature appears in the signal.
- *The prototype matching* operation specifies the degree of correspondence between ideal phonemes and the input sets of features.
- *Pattern classification* determines the best match between the candidate phonemes and the input by using goodness of fit algorithms.
- It claims to be a universal principle of perceptual cognitive performance.

Fuzzy logical model of perception



- Evaluation:
  /ba/ - Rising F2-F3 and Closed Lips
  /da/ - Level F2-F3 and Open Lips
- Integration:
  s(/ba/) = (1 - a)(1 - v)
  s(/da/) = av
- Decision:

$$P(/da/) = \frac{av}{av + (1-a)(1-v)}$$

Feature Evaluation 特徴評価

$\forall x$ and $\forall y, t(x,y)$
where x is relevant alternative
and y is relevant source of information.
$0 \le t(x,y) \le 1$ represents the degree
to which source y supports alternative x.

Feature Integration 特徴統合

$\forall x$ and $\forall y, t(x,y)$

$$t(x) = \prod_i t(x, y_i)$$

where $t(x)$ = total support for alternative x

Decision 決定

$$P(x) = \frac{t(x)}{\sum_i t(x_i)}$$