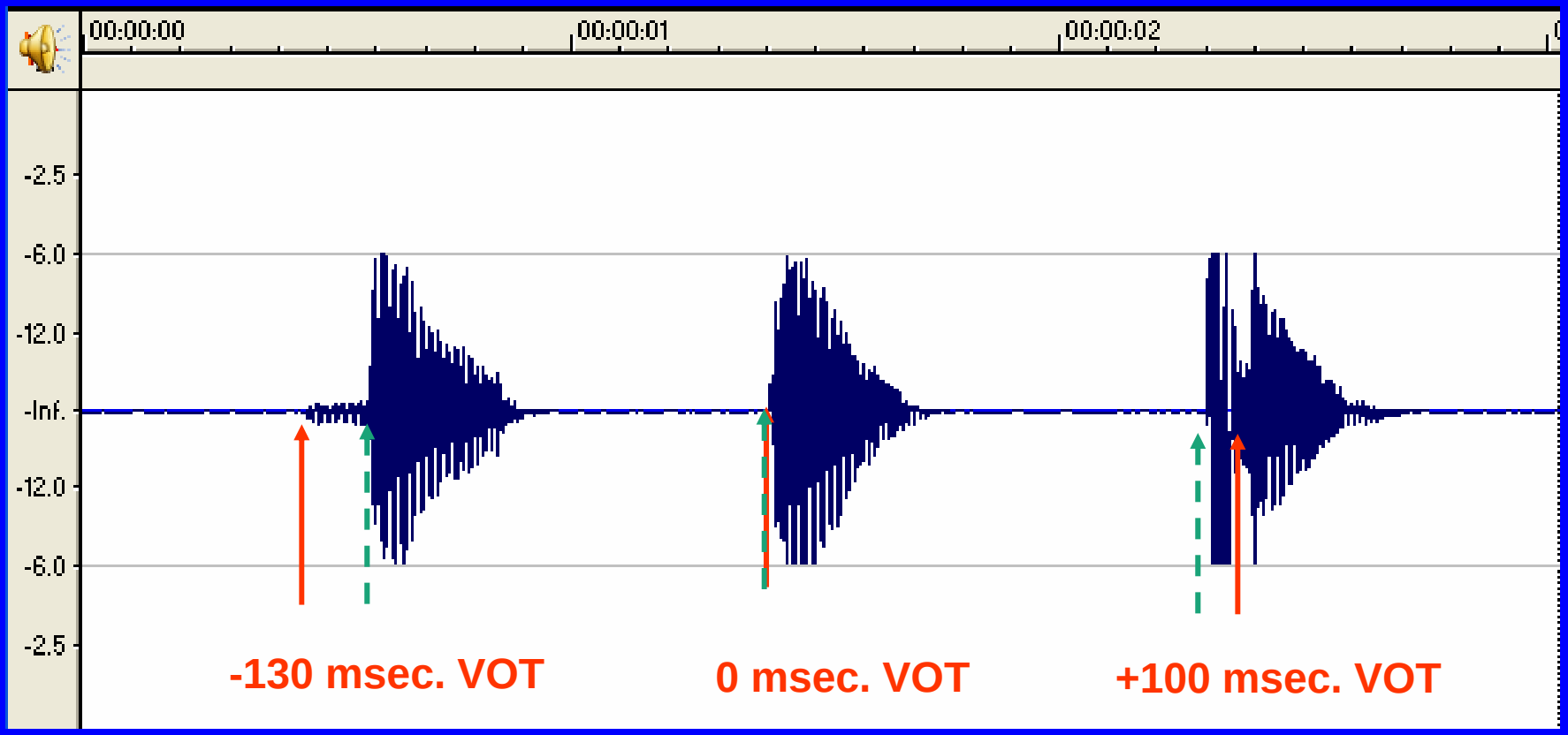
# Why speech perception is a challenge

- Something we do without effort
- Something machines do very poorly
- Characteristics:
  - Extremely rapid
  - No “white space”
  - “Lack of invariance”
    - Within a speaker
    - Across speakers

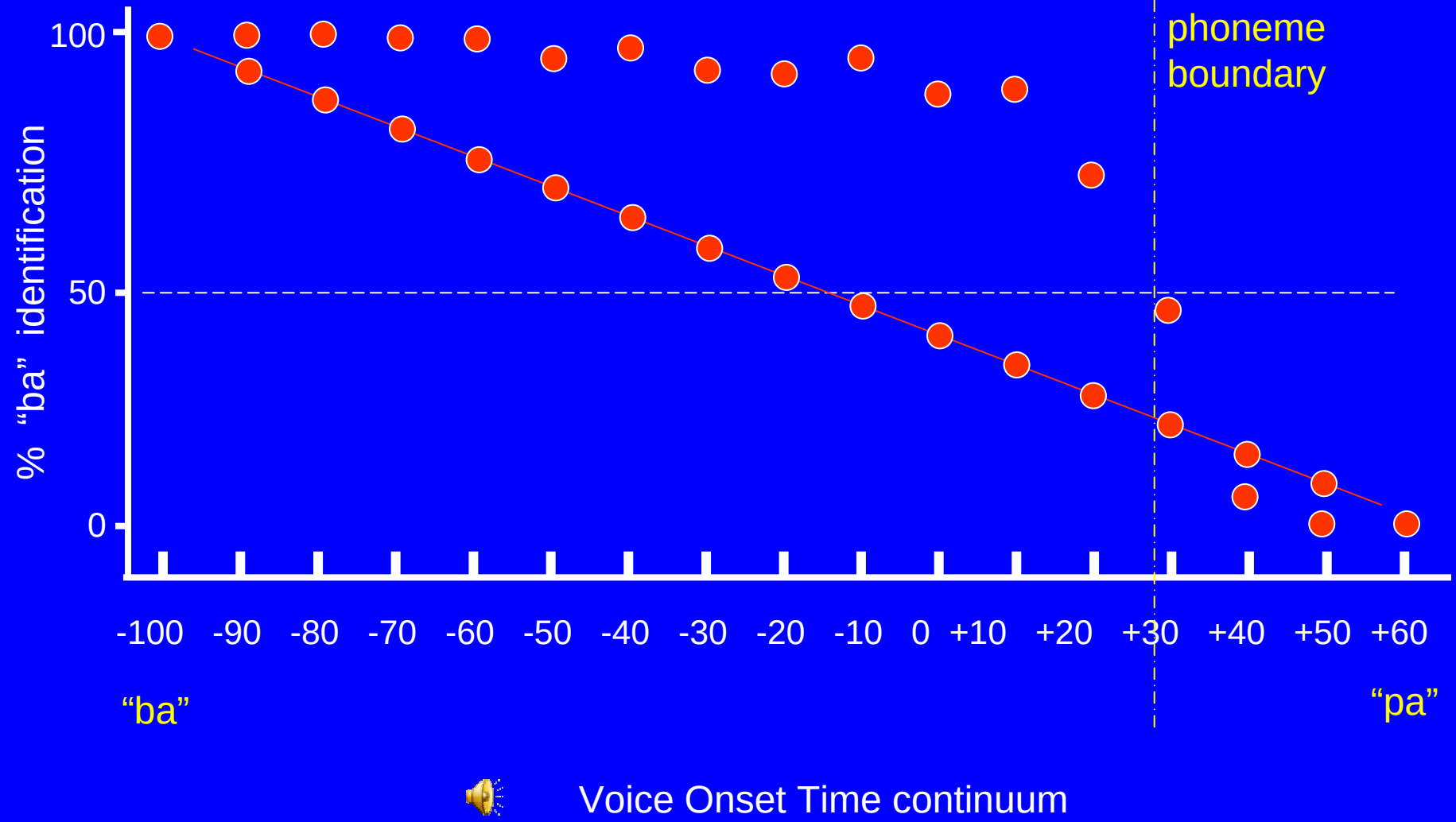
# “Speech is special” hypothesis

- Specialized neural hardware
- Innate categories




| | | | |
|---|---|---|---|
| English: | ba | ba | pa |
| Spanish: | ba | pa | - |

# Categorical Perception

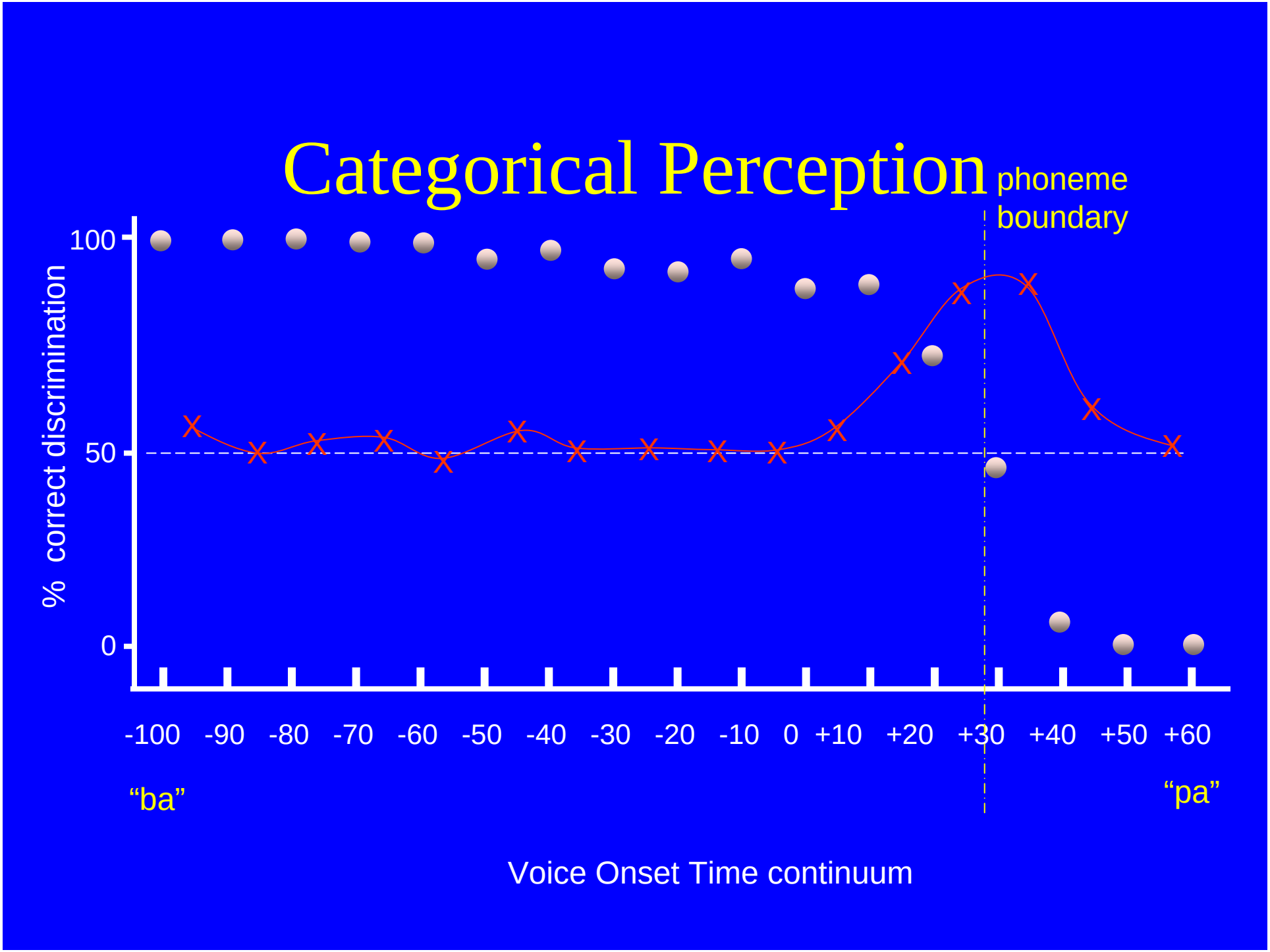


# Synthetic /pa-ba/ continuum

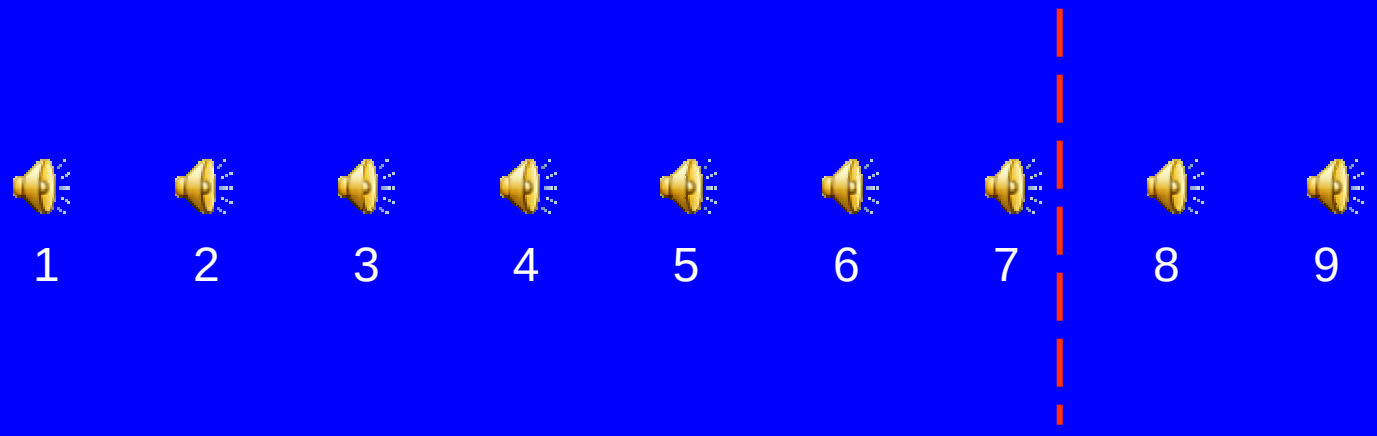
1 2 3 4 5 6 7 8 9

# AX discrimination task

- Hear 2 adjacent (i.e., very similar) stimuli
- Task:  Are they the Same?  Different?

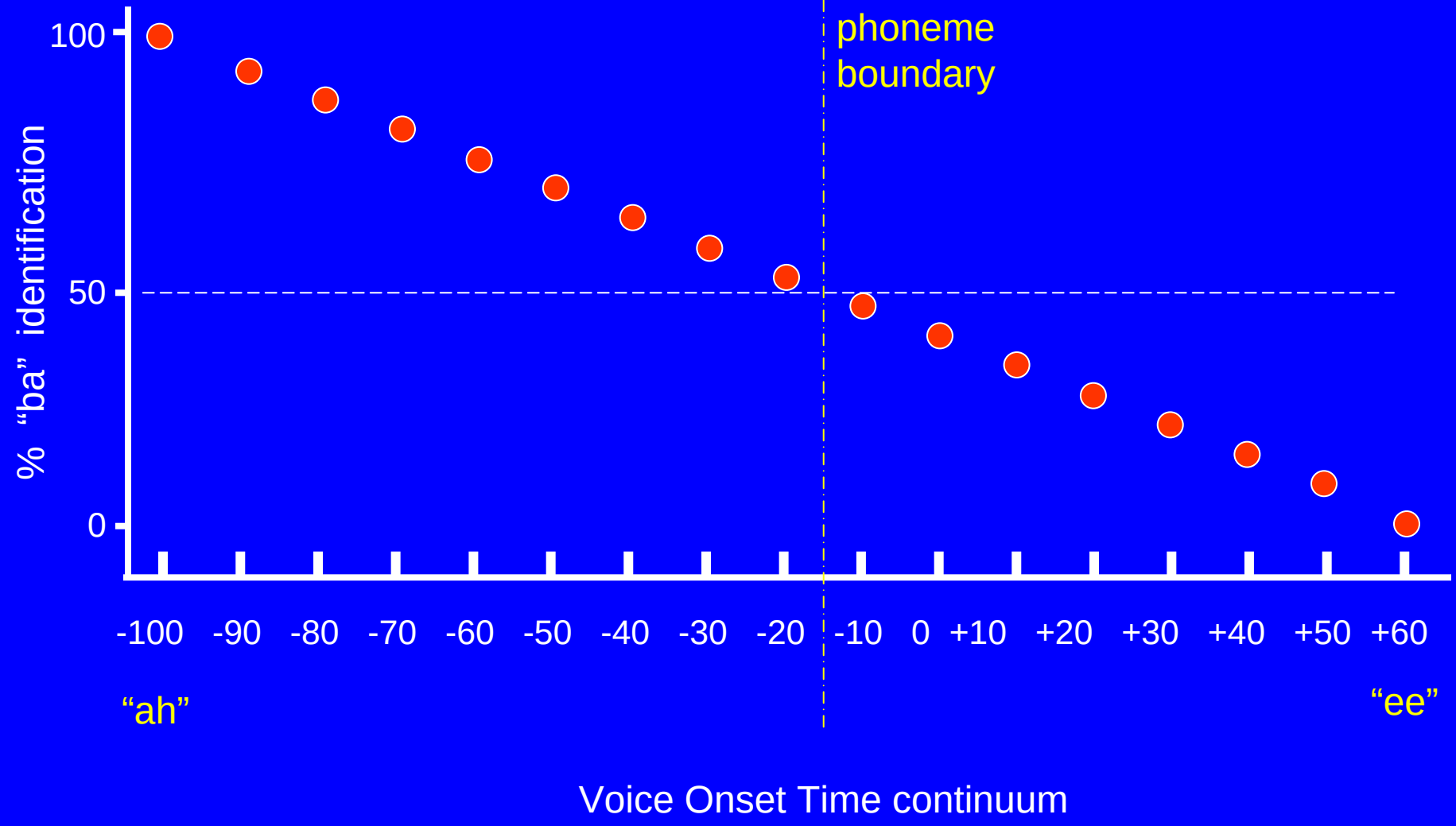

Categorical Perception
phoneme boundary
100
50
0
% correct discrimination
-100 -90 -80 -70 -60 -50 -40 -30 -20 -10 0 +10 +20 +30 +40 +50 +60
"ba"
"pa"
Voice Onset Time continuum

# Synthetic /pa-ba/ continuum




Phoneme boundary

# Continuous Perception



# Feature detectors?

If there are feature detectors, can we tire one of them out?

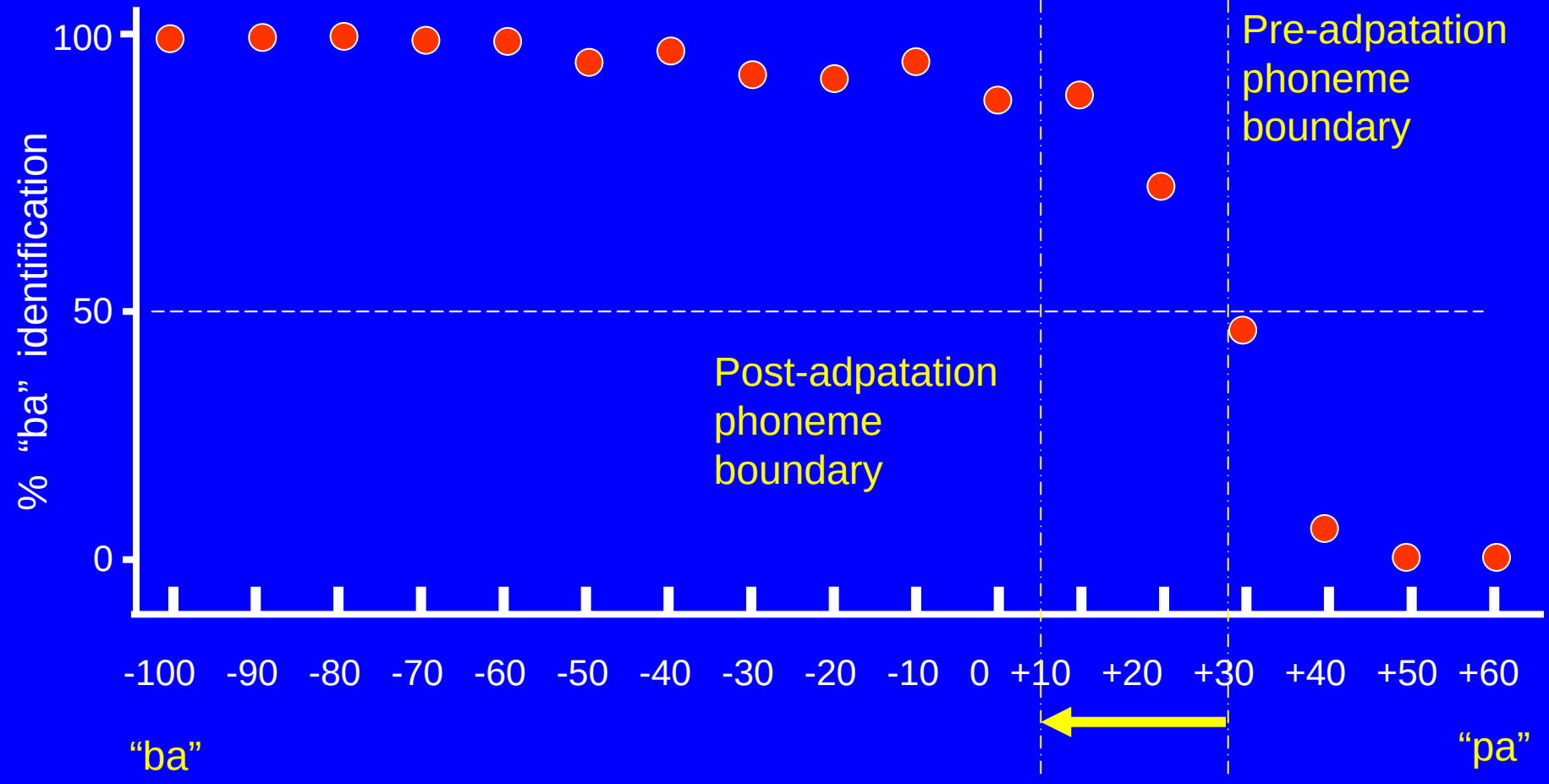
# Selective adaptation

1. Do phoneme identification test
   (e.g., “ba-pa” continuum)

2. Play a stimulus from one of the end-points many times (e.g., 100 times)

3. Repeat phoneme identification test

# Selective adaptation

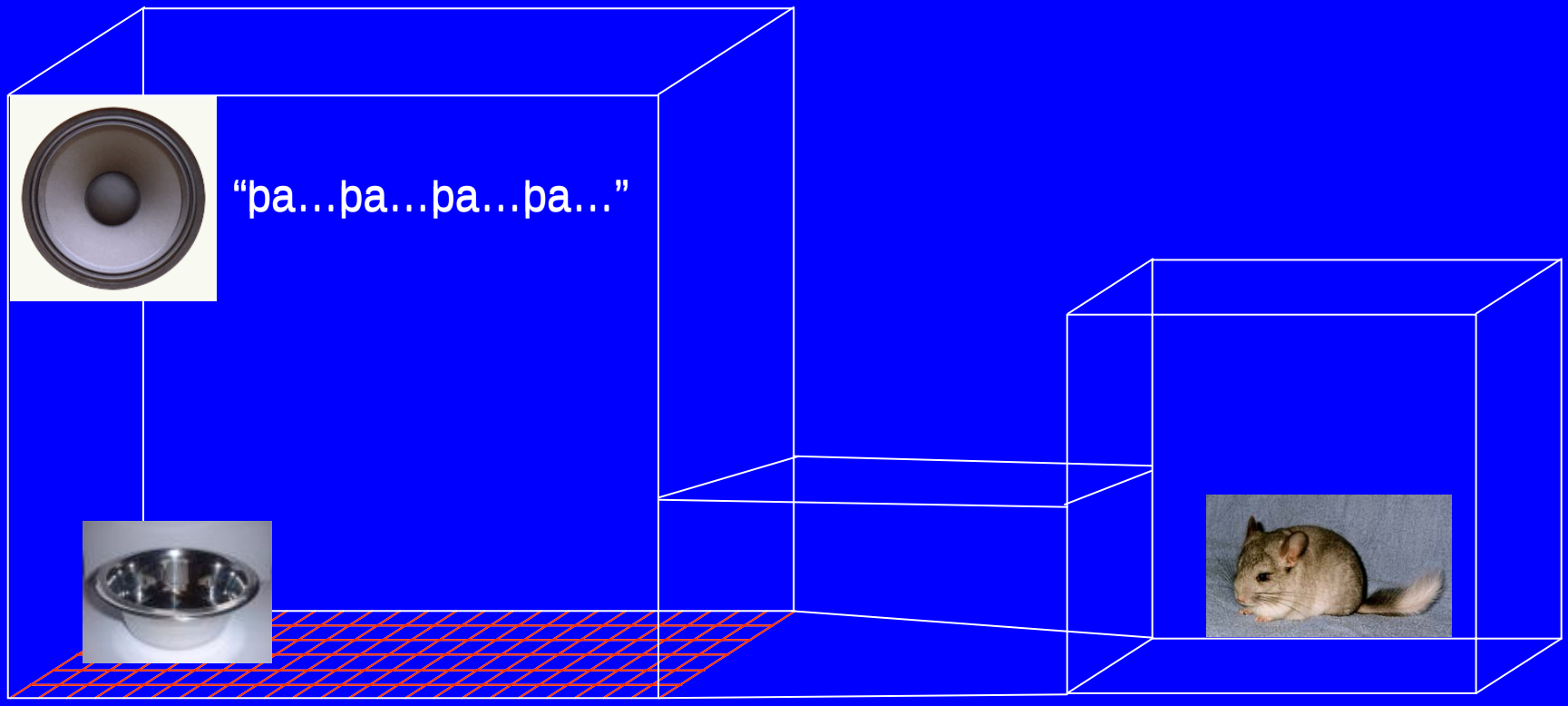



REPEAT -100 "ba"
100 times for one minute

# But are there really feature detectors?

- Chinchillas exhibit categorical perception as well

# Chinchilla experiment
## (Kuhl & Miller experiment)



- Train on end-point “ba” (good), “pa” (bad)
- Test on intermediate stimuli
- Results:
  - Chinchillas switched over from staying to running at about the same location as the English b/p phoneme boundary

# Categorical perception, Take 2

- Natural discontinuities in many sensory systems; many of these are common across mammalian species
- Some stimulus differences are hard; others are easy
- Language takes advantage of “natural boundaries”

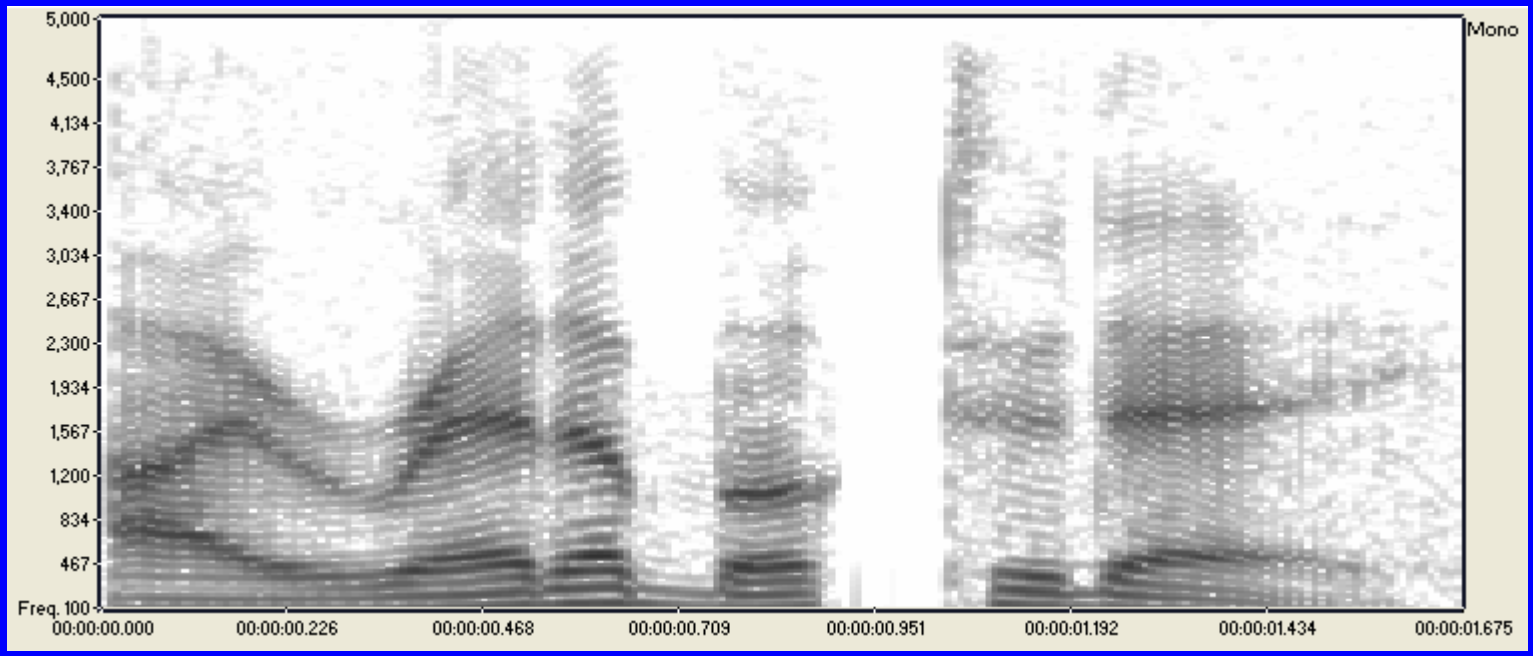
But other ways in which speech perception seems different

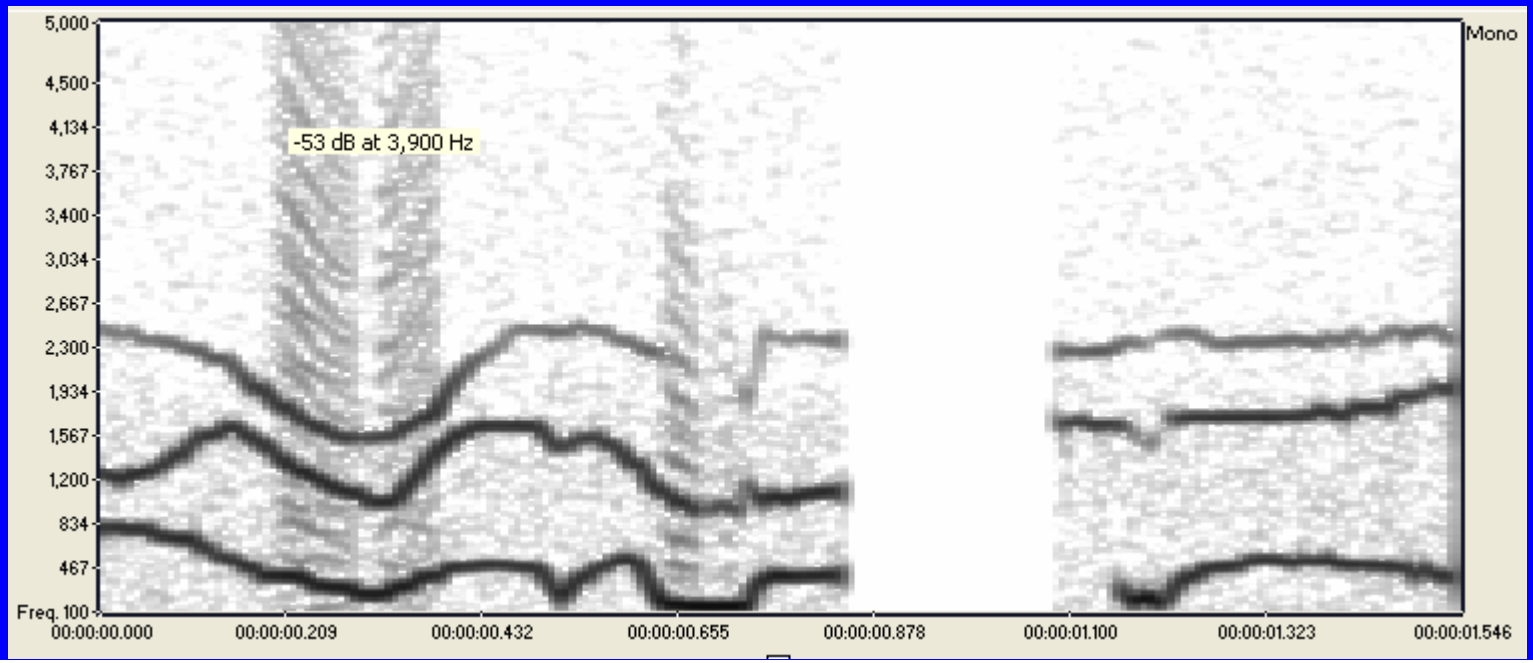
# Sine-wave analogs of speech

Sine wave: 

Original speech: 

I read a book today



# Other interesting effects. . .

- Speaker normalization
- Lack of invariance
- “top-down” effects – Phoneme Restoration

# Tentative conclusions

- The human articulatory system and auditory system is different from non-human primates in non-trivial ways that make speech possible
- There is a tight integration between perception & production… “our ears compensate for what our mouths do”
- There is a tight integration between the perception of speech and higher level cognition

# Phonetics & Phonology

- Phones
  - Sounds produced by human articulatory system
- Phonology
  - How sounds distinguish one word from another
- Phoneme
  - Smallest unit of sound that makes a difference for meaning; set of phones equivalent in determination of meaning

# Speech Sounds & Distinctive Features

- Trubetsky & Jakobson
  - Speech sounds are encoded in the brain in terms of more primitive specifications called *distinctive features*
- Phonological structure relates to movements of the vocal tract

# Sound waves

- Sound
  - wave-like transfer of energy through some medium (e.g. air, water)
  - results from disturbances in pressure of medium as a result of external force on some object (e.g. striking a tuning fork)

# A simple wave

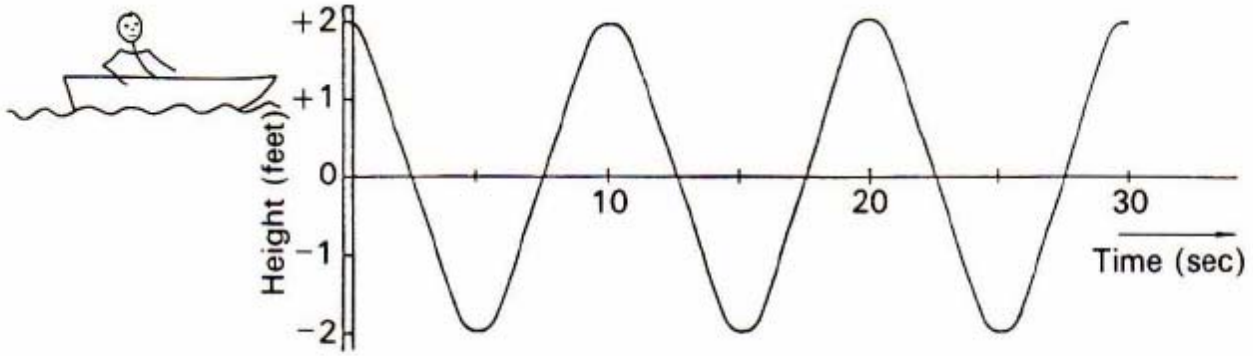



*Figure 3.2. Graph of water height showing the 10 second period of the waveform.*

# How waves are created

| | amplitude |
|---|---|
| (a) external force on object causes displacement in one direction (therefore maximal transfer of energy through medium) | maximum |
| (b) elastic recoil force (E) returns object to resting position; $E = \sigma/e$ ($\sigma$ = stress applied to object, e = amount of deformation of object) | 0 |
| (c) inertia carries object beyond resting position to other maximum amplitude | -maximum |
| (d) elastic recoil force returns object to resting position | 0 |

Etc.

# Some sound wave terminology

- amplitude (intensity) = size of wave
  - generally measured in decibels (ratio between two intensities)
  - doubling/halving the ratio of signal intensities adds/subtracts 3 dB.
    - E.g. if sound X is 6 dB and sound Y is 3 dB, then X is twice as loud as Y
- Damping = decrease in amplitude over time (e.g. caused by friction from air molecules).

# Amplitudes of various sounds

| | |
|---|---|
| 0 dB | threshold of hearing at 1000 Hz |
| 30 dB | soft whisper |
| 80 dB | conversation |
| 100 dB | shouting |
| 130 dB | loud thunder |
| 150 dB | pain threshold (most people) |

# More wave terminology

- period = time it takes to complete a wave
- frequency = inverse of period (1/P) (# of waves/cycles that can occur in a given time period)
  - 1 Hertz (Hz) = 1 cycle/second
- periodic (e.g. voiced sounds) vs. aperiodic waves (e.g. fricatives)

# Frequencies of various sounds

| | |
|---|---|
| 20 Hz | lowest human(?) audibility |
| 27.5 Hz | lowest note on piano |
| 104.8 Hz | lowest note on clarinet |
| 261.6 Hz | middle C on piano |
| 440 Hz | standard tuning pitch (A above middle C) |
| 1,000 Hz | upper range of soprano |
| 4,180 Hz | highest note on piano |
| 10,000 Hz | harmonics of musical instruments |
| 12,000 Hz | limit of hearing for older persons |
| 16,000-20,000 Hz | limit of human hearing |
| 20,000-100,000 Hz | hearing range of bats |

# Different types of periodic sound waves

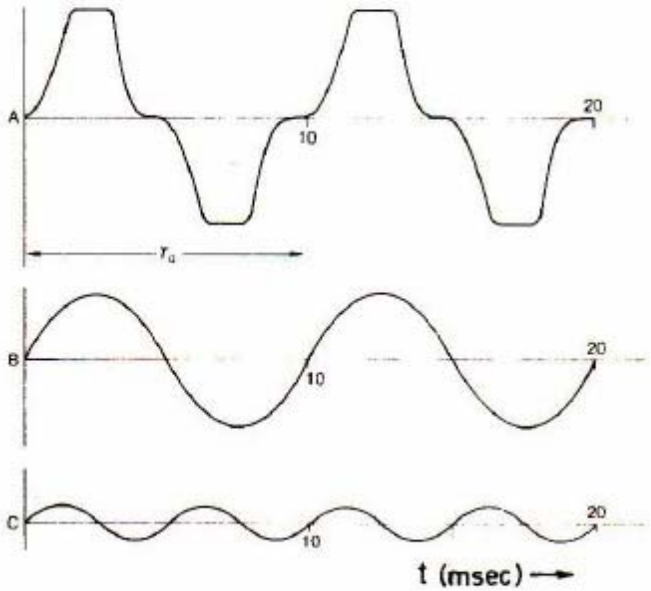
- simple
  - consist of wave of one frequency and one amplitude
    - e.g. wave produced by tuning fork
- complex
  - consist of one or more simple waves of different frequencies and amplitudes
    - e.g. chord, **vowel**

# Fourier analysis

- A complex periodic wave can be analyzed as a sum of simple waves that differ in frequency and/or amplitude.

| | |
|---|---|
| lowest frequency wave | (f0, fundamental frequency, first harmonic) |
| + wave of frequency (f0)(2) | second harmonic |
| + wave of frequency (f0)(3) | third harmonic |

Etc.




← f0 = first harmonic = fundamental

← second harmonic

*Figure 3.6. (A) A complex waveform. (B) and (C) Its first two sinusoidal Fourier components.*

# Fast Fourier transform (FFT)

- provides the frequency/amplitude components of the simple waves that comprise a complex wave
- output of a fast Fourier transform is a spectrum, graph of amplitude vs. frequency

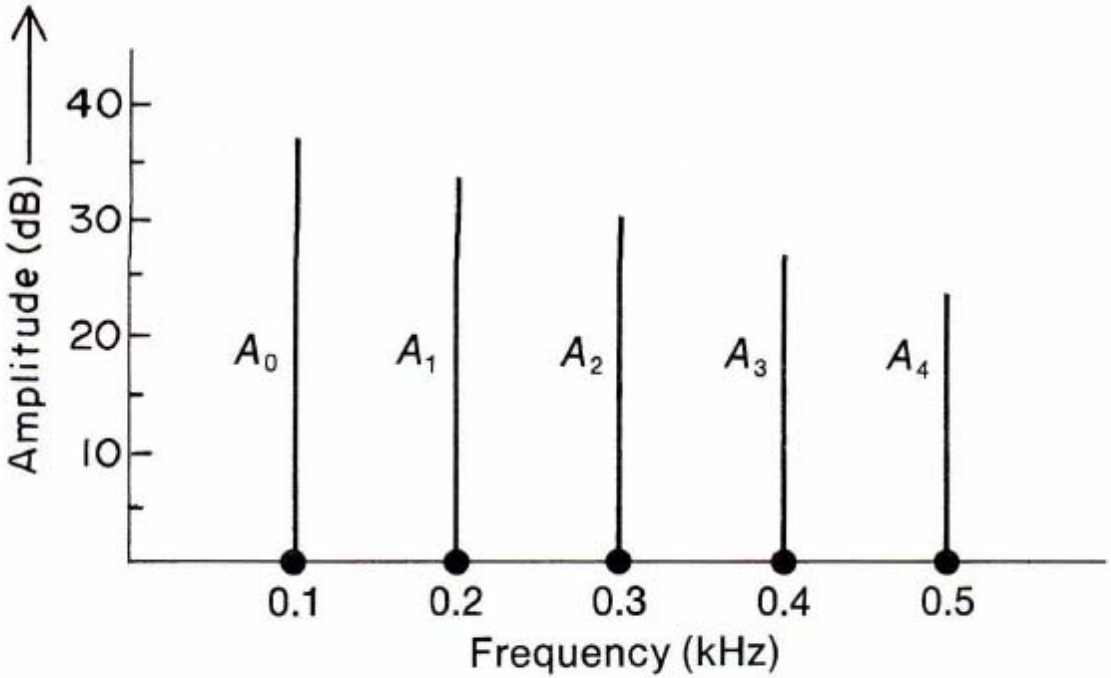



*Figure 3.7. Graph of a spectrum.*

# Resonance

- Resonance = transfer of energy/vibration from one medium to another
  - e.g. transfer of vibrating energy from vocal cords to vocal tract
- Different objects prefer to vibrate at various characteristic frequencies (resonant frequencies).
  - Certain harmonics of sound waves that travel through or within these objects are amplified, others are damped.

# Production of voiced sounds

- The vocal cords are a sound source.
  - interrupt air flow from lungs, producing a complex, periodic wave
  - f0 for adult male speakers: apx 80-200 Hz
  - f0 for adult female speakers: apx 150-220 Hz (can go up to 400 Hz)
  - f0 for children: up to 500 Hz
- The vocal tract acts as a resonator (or filter: 'source-filter theory')

# Simplified model of vowel production

- an open tube (= lips) with a vibrating membrane (= vocal cords) at one end.

# Resonant frequencies

- Resonant frequencies, or formant frequencies, or formants, of tube depend on its length of tube: expressible by formula:
  - $F_n = (2n-1)c/4l$
    - F = formant frequency
    - n = which formant
    - c = speed of sound (1086 feet/sec or 35,000 cm/sec) (or speed of propagation of wave through medium)
    - l = length of tube
      - 17.5 cm—average adult male vocal tract
      - 14.9 cm—average adult female vocal tract length (80-90% of male)

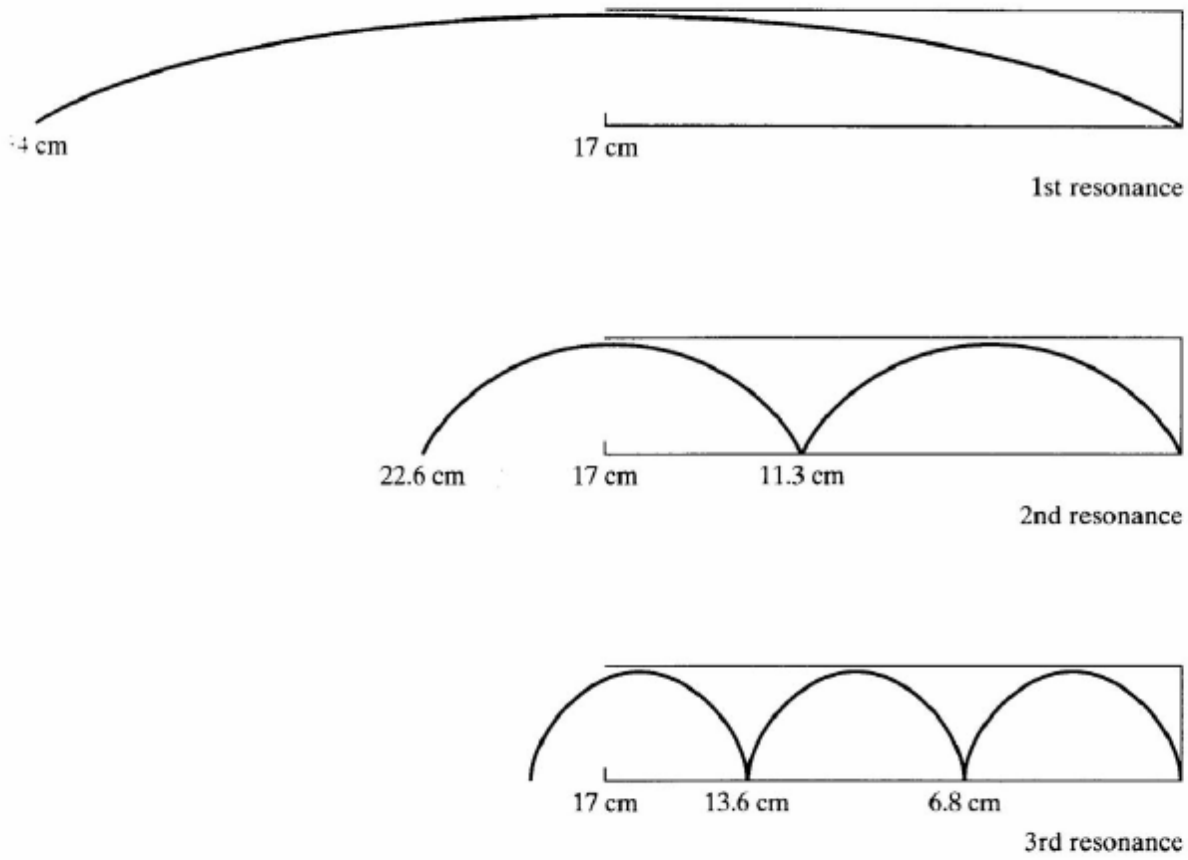



**Figure 4.19** Standing wave patterns in 17-cm tube of uniform cross section approximating vocal tract.

# Calculating F1 and F2

- average adult male vocal tract
  - F1 = (2x1 -1)35000/4x17.5 = 500 Hz
  - F2 = (2x2 -1)35000/4x17.5 = 1500 Hz
- average adult female vocal tract
  - F1 = (2x1 -1)35000/4x14.9 = 570 Hz
  - F2 = (2x2 -1)35000/4x14.9 = 1711 Hz
- note: vocal tract length is inversely proportional to formant frequency

# More on vocal tract as resonator

- Certain harmonic frequencies are damped, others amplified, by different vocal tract shapes, depending on location of main constriction.
- Several bands of frequencies may be amplified around the resonant frequency
  - e.g. resonant frequency of 1000 Hz may amplify frequencies at 990-1010 or 800-1200 Hz

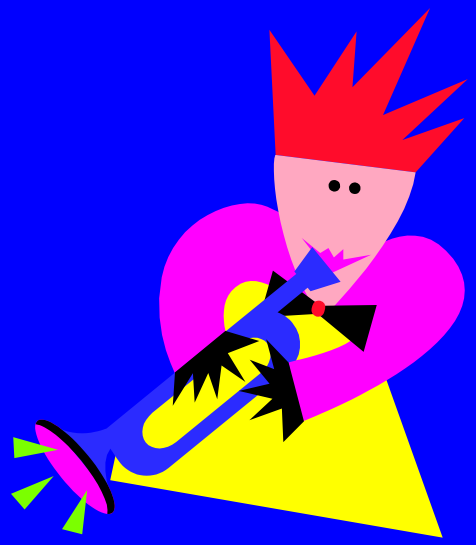
# Effect of source

- Shape of wave entering cavity can vary in frequency, amplitude
  - length of vocal cords can vary, causing the cords to vibrate at different rates, affecting f0 and harmonics
  - loudness of sound can vary, depending on effort by lungs

# Effect of resonator (filter)

- Resonating cavity differences in size, shape, material may affect
  - which frequencies are amplified
  - width of resonant band
  - numbers of resonant bands

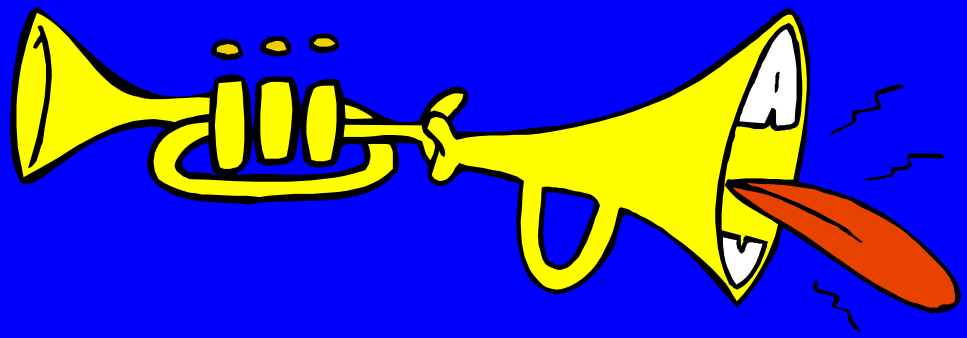
# Mechanics of Speaking



- Vibration is a function of the vibration of the lips and the resonant frequencies of the trumpet’s tube

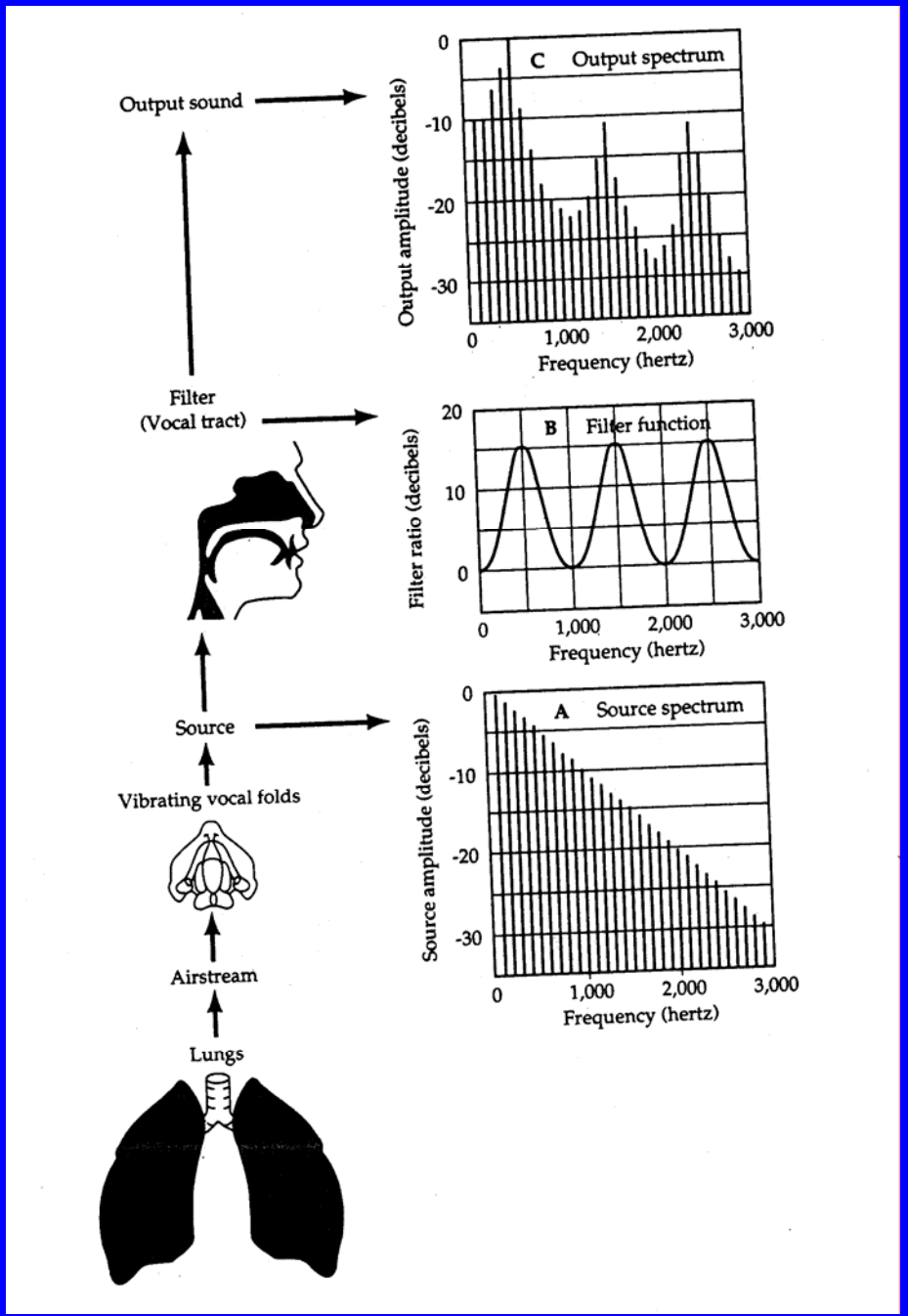
# Imaginary Trumpet

- Made of rubber
- Has 2 tubes
- Stretching changes resonant frequencies
- Choice of tube offers possibility of different tone qualities

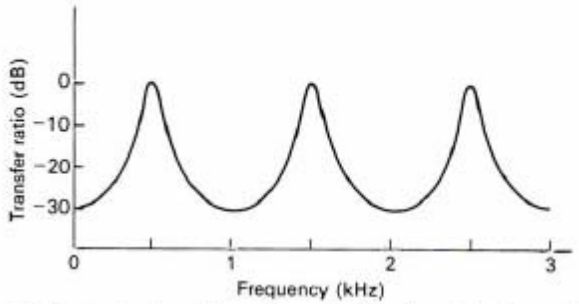


# Analogy

- Trumpet Player's Lips
- Trumpet Tube
  - Tube 1
  - Tube 2
- Vocal Cords in Larynx
- Throat & its Branches
  - Nasal Cavity
  - Oral Cavity


Output sound
Filter
(Vocal tract)
Source
Vibrating vocal folds
Airstream
Lungs
C Output spectrum
Output amplitude (decibels)
0
-10
-20
-30
0
1,000
2,000
3,000
Frequency (hertz)
B Filter function
Filter ratio (decibels)
20
10
0
A Source spectrum
Source amplitude (decibels)

passed through




yields

*Figure 4.3. Transfer function of the supralaryngeal vocal tract for the vowel [ə]. For our purposes, the term "transfer function" is equivalent to "filter function." Note the locations of the formant frequencies at 0.5, 1.5, and 2.5 kHz.*

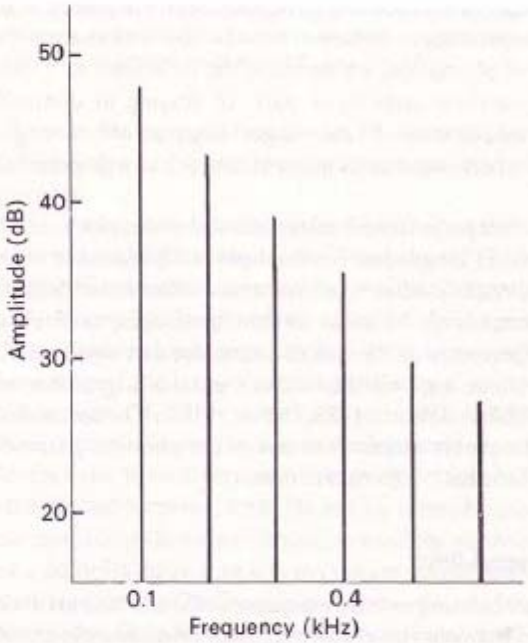



*Figure 4.2. Spectrum of typical glottal air flow.*

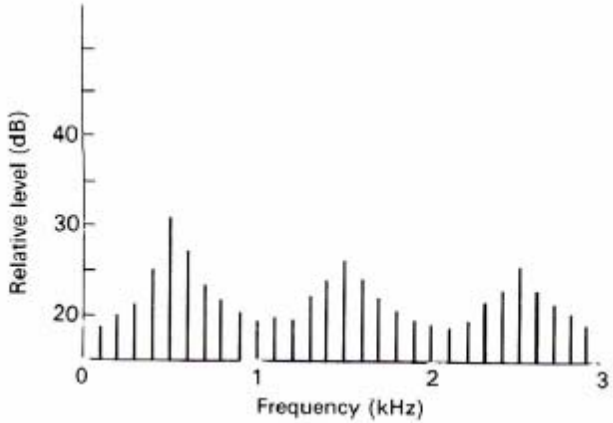



*Figure 4.4. The spectrum that would result if the transfer function plotted in Figure 4.3 were "excited" by the glottal source plotted in Figure 4.2. The sound is the vowel [ə].*

# Sound Source

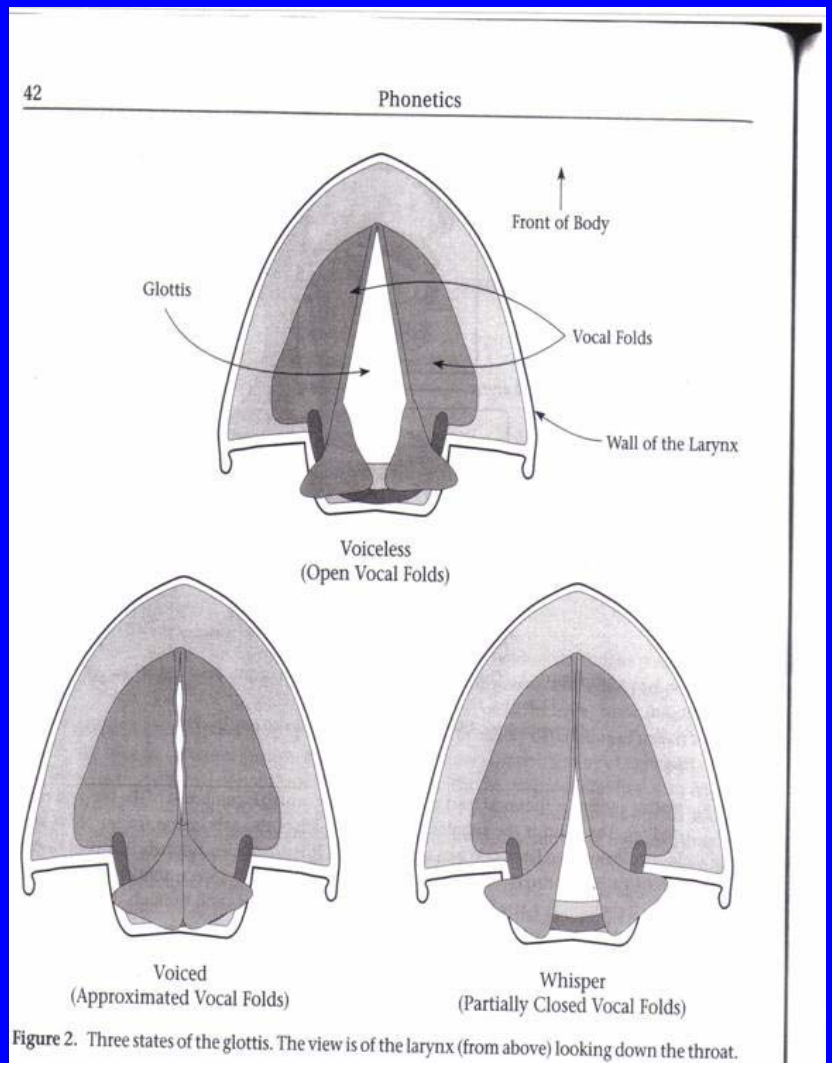



**Figure 2.** Three states of the glottis. The view is of the larynx (from above) looking down the throat.

# Resonant Filter

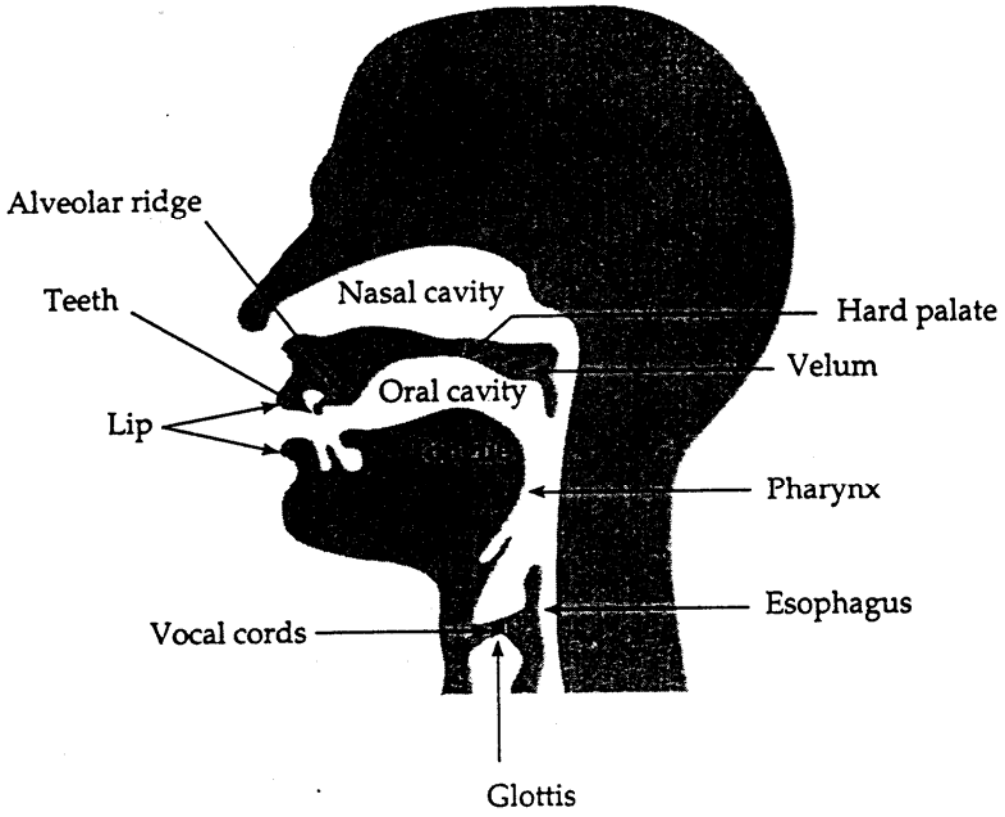



**Figure 9.1** **Positions of Articulation in the Mouth.**
*Source*: Clark, H. H., & Clark, E. V. (1977).

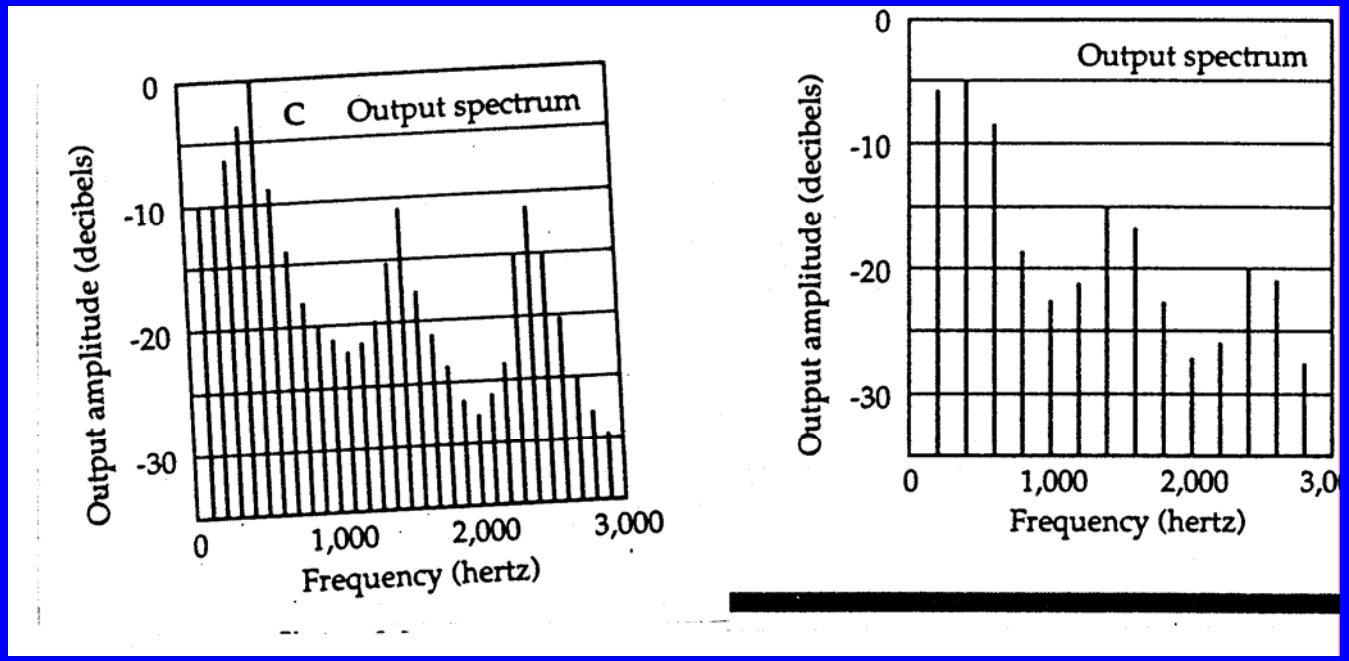
# Speech production


KNS_X-ray_Film.mov

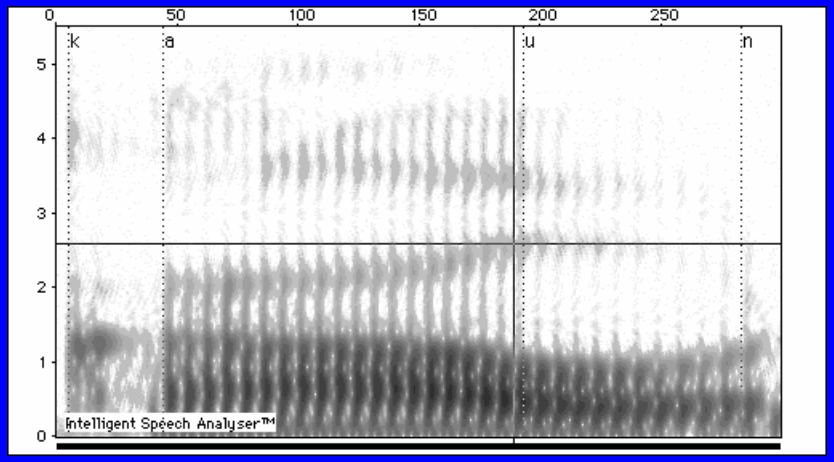
# Spectra and spectrograms

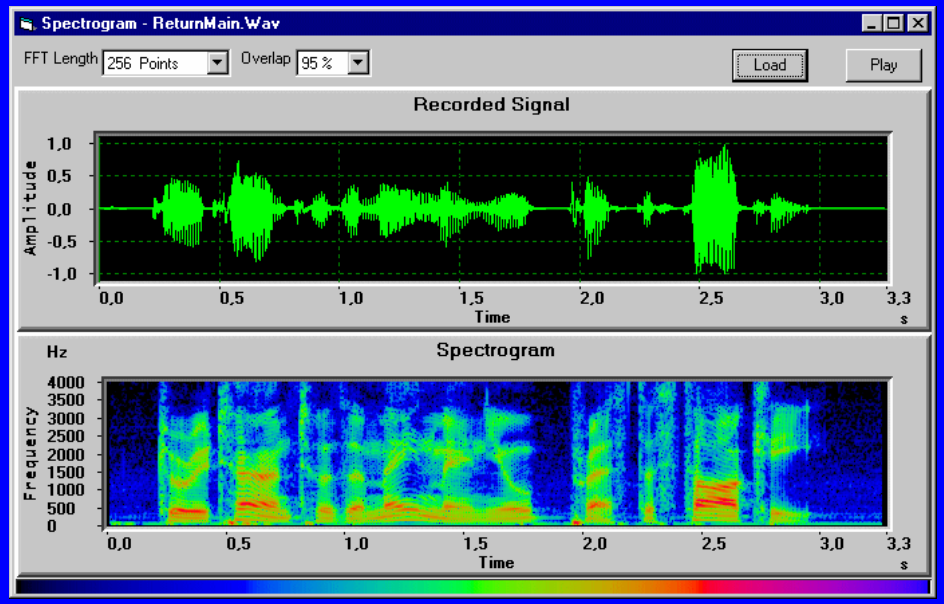
- Graphs of sound waves commonly encountered in speech analysis
  - waveform: amplitude x time
  - spectrum (or spectral slice): energy x frequency (at "one point" in time)
  - spectrogram: frequency x time (x amplitude indicated by darkness)

# Spectrum(s)


C Output spectrum
Output amplitude (decibels)
0
-10
-20
-30
0
1,000
2,000
3,000
Frequency (hertz)
Output spectrum
Output amplitude (decibels)
0
-10
-20
-30
0
1,000
2,000
Frequency (hertz)

# Spectogram(s)



# Potential points of confusion

- Two (basically) independent parameters:
  - shape of vocal tract (F1 etc.)
  - length of vocal cords (f0)
- f0 independent of F1
- formant ≠ harmonic

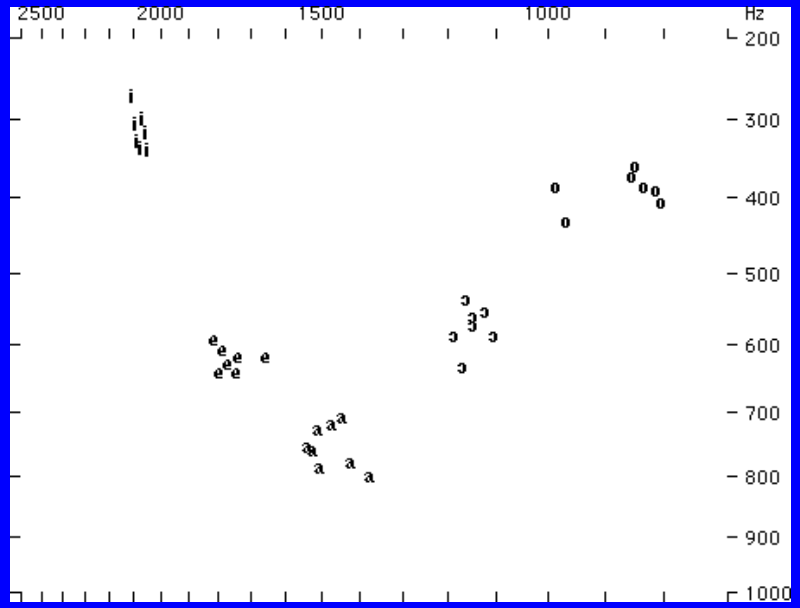
# Directing the Flow

- To disconnect nasal cavity
  - Raise velum (soft palate)
- To cut off air from oral cavity
  - Close lips
  - Raise tongue

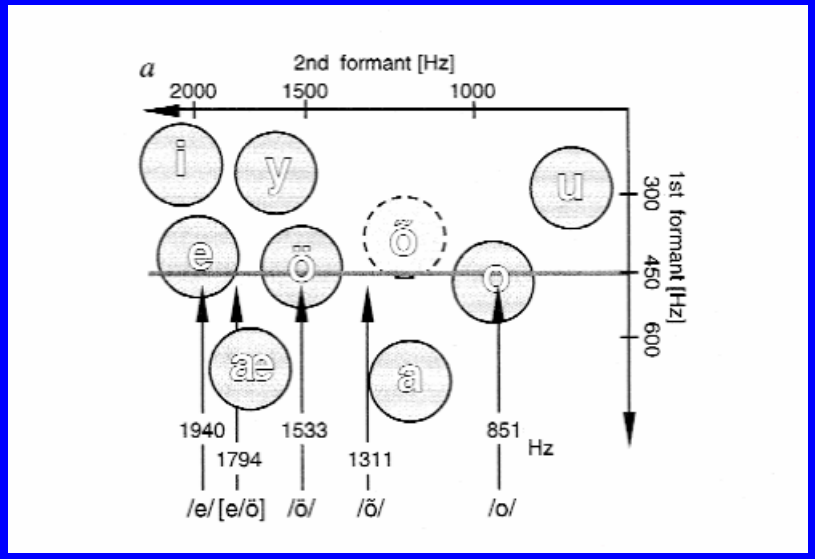
AEMS.mov

# Vowels

- Tongue Height
- Tongue Position
- Defined by first two
  Formants
  (peaks in spectra)
  (dark bands in spectro-gram)
- Synthesized by first
  3 formants



# Vowels


a
2nd formant [Hz]
2000
1500
1000
i
y
u
e
ö
ö
o
æ
ä
1st formant [Hz]
300
450
600
1940
1794
1533
1311
851
Hz
/e/
[e/ö]
/ö/
/ö/
/o/

# Studying Speech Perception w/ERPs

- Is there an ERP component that could serve as an objective record of perceptual discriminations people make when comprehending speech?

# Mismatch Negativity (MMN)

- Frontocentral negative ERP component
- Peaks 100-250 ms post-stimulus onset
- Change in repetitive aspect of on-going auditory stimulation

# Mismatch Negativity



MMN as a Function of Frequency Change

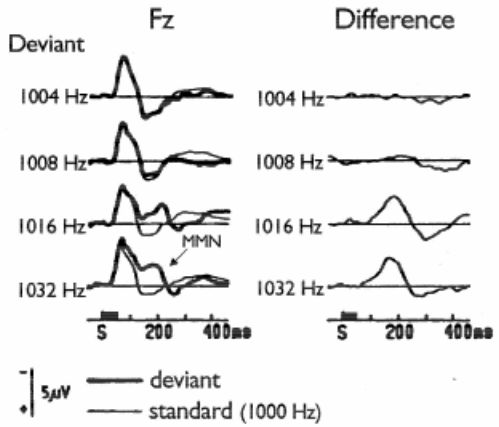


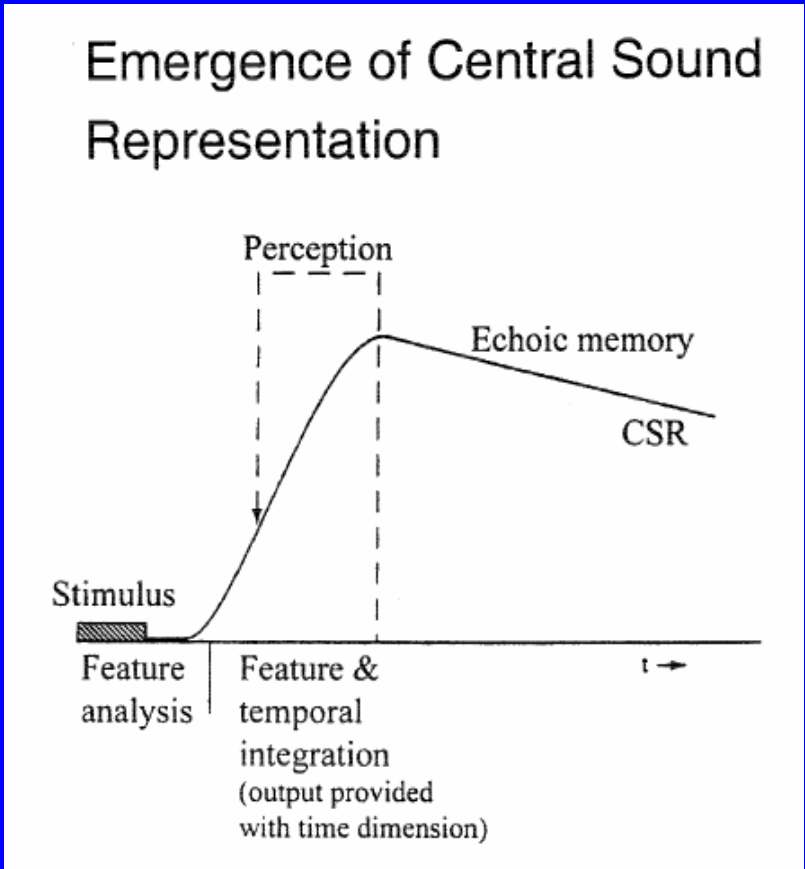

**Figure 1.** Left: Frontal (Fz) event-related potentials (ERPs; averaged across subjects) to 1000-Hz standard (thin line) and to deviant (thick line) stimuli of different frequencies, as indicated on the left side. Right: The difference waves obtained by subtracting the standard-stimulus ERP from that to the deviant stimulus separately for the different deviant stimuli. MMN = mismatch negativity. Adapted from Sams et al. (1985). 

# Reflects Automatic Processing

- Occurs with or without attention to auditory stimuli
- Sleep
  - Stage 2
  - REM
- Coma
- MMN signal more “pure” without attention
  - Without: subtraction yields only MMN
  - With: subtraction yields MMN, N2, P3

# MMN & Echoic Memory

- Seems to reflect unified sound percepts (not acoustic features)
  - Simple tones
  - Complex stimuli (phonemes)
  - “complex spectrotemporal pattern”
- Does not arise unless “standard” is repeated a few times
- Does not arise if 5-10 s intervenes between stimuli
  - Matches estimated length of echoic memory



# How different is different?

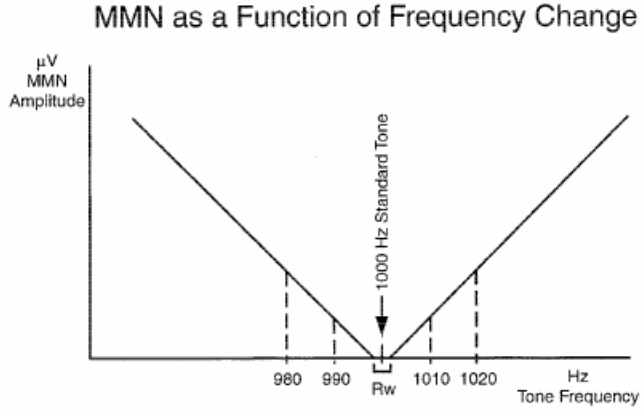



**Figure 3.** An illustration of the concept of the representational width (Rw) along a sensory dimension, here for tone frequency. Rw is the range around the standard stimulus for the deviant-stimulus levels that are too near to the standard-stimulus level to elicit a mismatch negativity (MMN) and to cause behavioral change discrimination. Outside this range, MMN is elicited at an amplitude that is larger for wider deviations in either direction, and behavioral change detection usually occurs. From Näätänen and Alho (1997). Copyright by S. Karger AG, Basel.

# MMN as a Function of Discrimination Performance

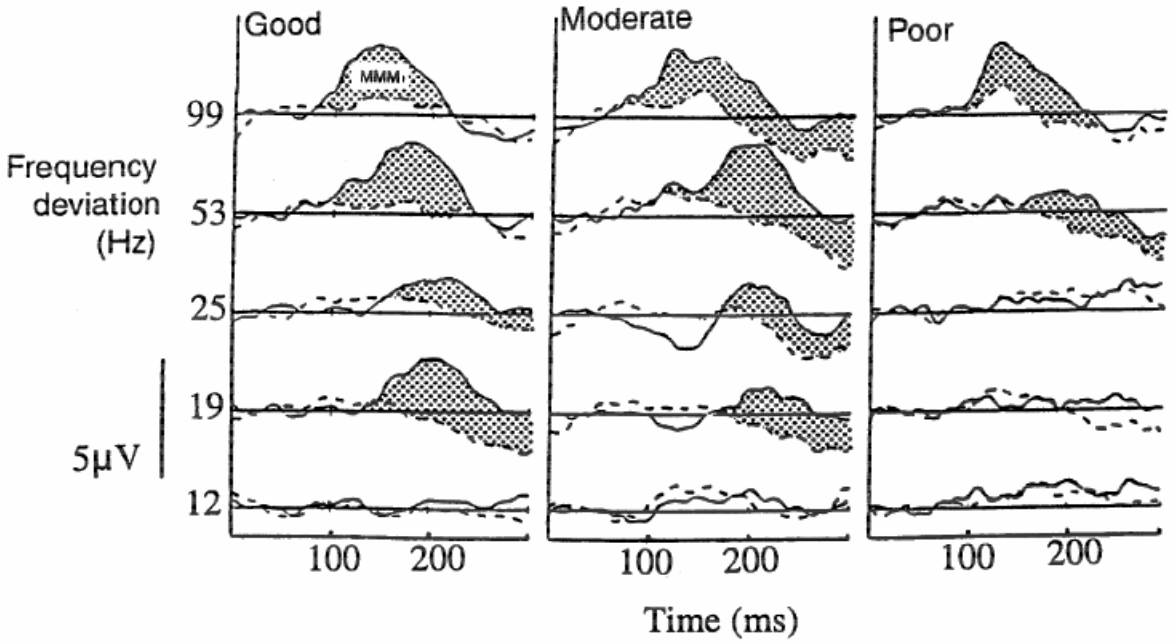



**Figure 4.** Grand-average event-related potentials (ERPs) for subgroups of subjects who were "good," "moderate," or "poor" discriminators (in a separate behavioral pitch-discrimination session). ERPs to standard tones of 698 Hz (dashed lines) and to infrequent deviant tones (solid lines) which were 12, 19, 25, 53, or 99 Hz higher in frequency than the standard tones are overlaid. Note the between-group differences in the mismatch negativity (MMN; shaded areas) amplitudes for different frequency deviations. Adapted from Lang et al. (1990). Copyright 1990 Tilburg University Press.

# How different is different?

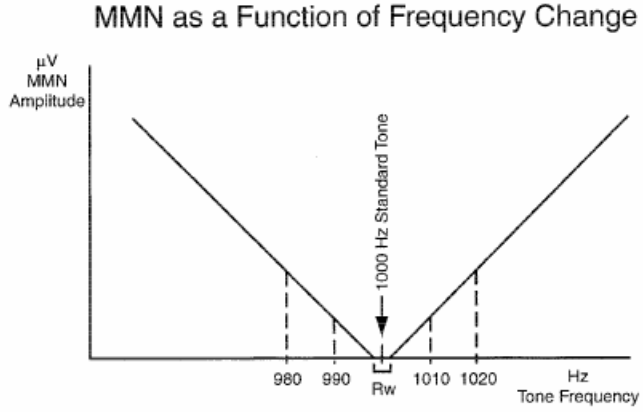
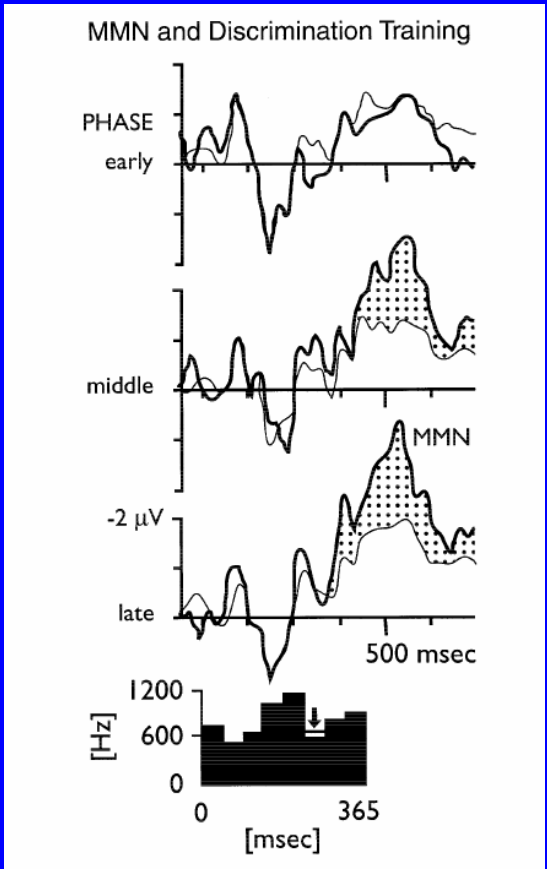



**Figure 3.** An illustration of the concept of the representational width (Rw) along a sensory dimension, here for tone frequency. Rw is the range around the standard stimulus for the deviant-stimulus levels that are too near to the standard-stimulus level to elicit a mismatch negativity (MMN) and to cause behavioral change discrimination. Outside this range, MMN is elicited at an amplitude that is larger for wider deviations in either direction, and behavioral change detection usually occurs. From Näätänen and Alho (1997). Copyright by S. Karger AG, Basel.

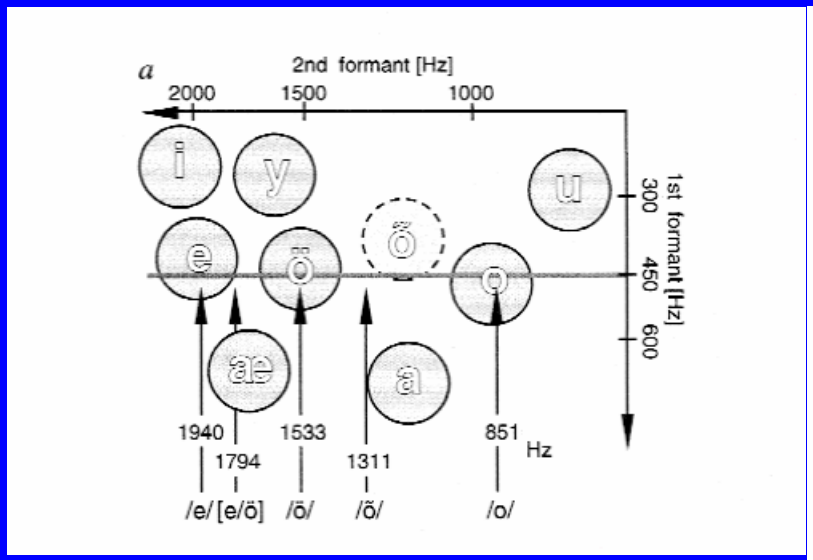
- How does discrimination ability change as a function of experience?

# MMN and Experience




- Subjects read a book while tones played in background
- Tested periodically in their ability to discriminate between tones
  - Increased over course of study
- MMN increases in amplitude as function of experience

# Language Experience: Infants



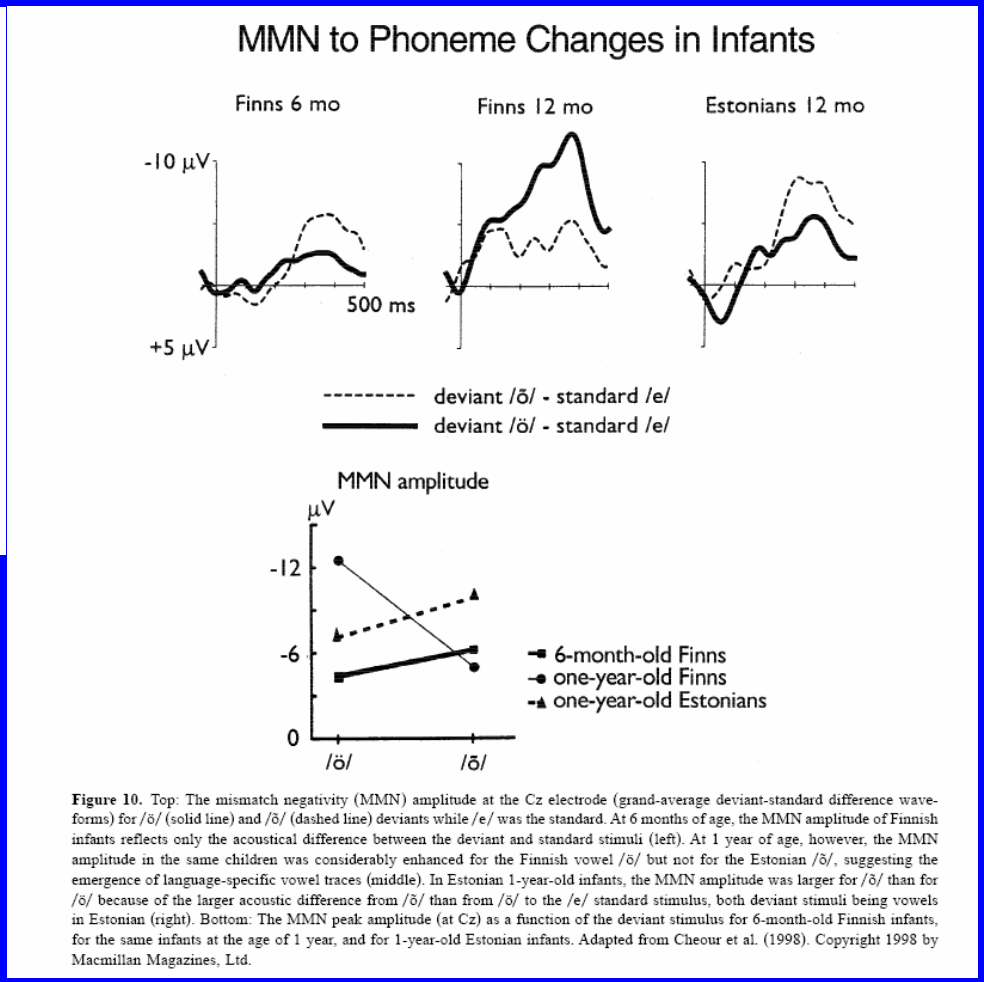
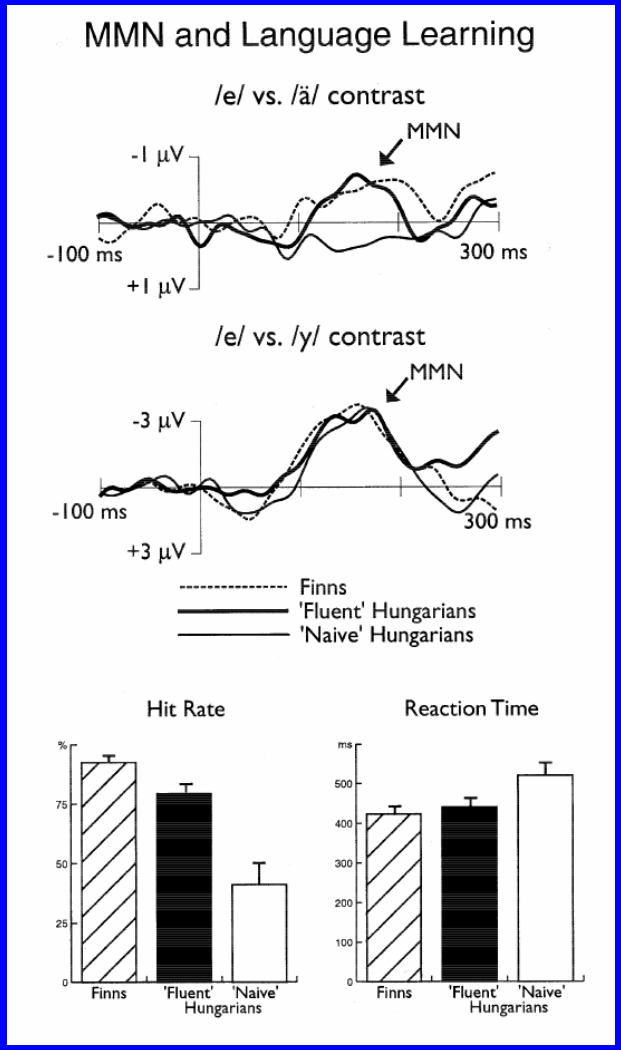




**Figure 10.** Top: The mismatch negativity (MMN) amplitude at the Cz electrode (grand-average deviant-standard difference waveforms) for /õ/ (solid line) and /ð/ (dashed line) deviants while /e/ was the standard. At 6 months of age, the MMN amplitude of Finnish infants reflects only the acoustical difference between the deviant and standard stimuli (left). At 1 year of age, however, the MMN amplitude in the same children was considerably enhanced for the Finnish vowel /ö/ but not for the Estonian /õ/, suggesting the emergence of language-specific vowel traces (middle). In Estonian 1-year-old infants, the MMN amplitude was larger for /ð/ than for /ö/ because of the larger acoustic difference from /ð/ than from /õ/ to the /e/ standard stimulus, both deviant stimuli being vowels in Estonian (right). Bottom: The MMN peak amplitude (at Cz) as a function of the deviant stimulus for 6-month-old Finnish infants, for the same infants at the age of 1 year, and for 1-year-old Estonian infants. Adapted from Cheour et al. (1998). Copyright 1998 by Macmillan Magazines, Ltd.

# Language Experience: Adults

- /a/ embedded in /e/
  - Relevant in Finnish
  - Not relevant in Hungarian
- /y/ embedded in /e/
  - Relevant in Finnish
  - Relevant in Hungarian



Speech Stimuli
(Left Temporal)

## MMN Generator Loci Shown by PET

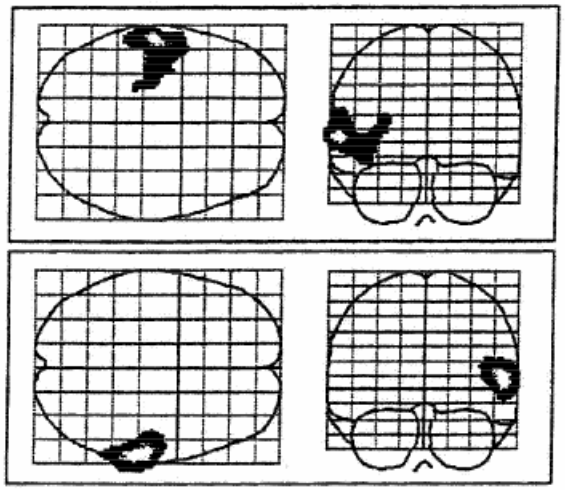


**Figure 9.** The functional lateralization of the mismatch negativity (MMN) shown with positron emission tomography (PET). The upper panel (on the left, seen from above; on the right, seen from back) shows the activation resulting from subtracting the activity in a condition in which only /e/ phonemes were presented from the activity in a condition in which /e/ phonemes were presented as standards and /o/ phonemes as deviants. The differential neural activity reflecting the MMN response is seen in the left superior and medial temporal gyri of the auditory cortex, indicating that change detection for speech sounds is left lateralized. The lower panel (on the left, seen from above; on the right, seen from back) shows the activation resulting from subtracting the activity in a condition in which only A-major chords were presented from that in a condition in which A-major chords were presented as standards and A-minor chords as deviants. Neural activity reflecting the MMN response is now seen in the right superior temporal gyrus of the auditory cortex, indicating that change detection for musical information is right lateralized. Adapted from Tervaniemi et al. (2000).

Musical Stimuli
(Right Temporal)