LING 520 Introduction to Phonetics I

Fall 2008

Week 5

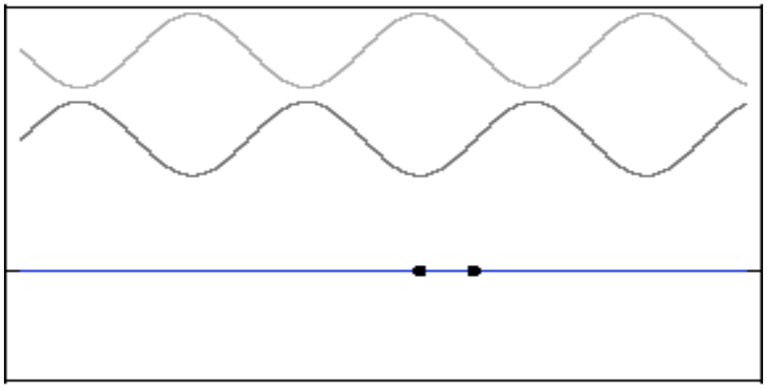
Acoustic theory of speech production

Acoustics of vowels

Oct. 6, 2008

Standing waves

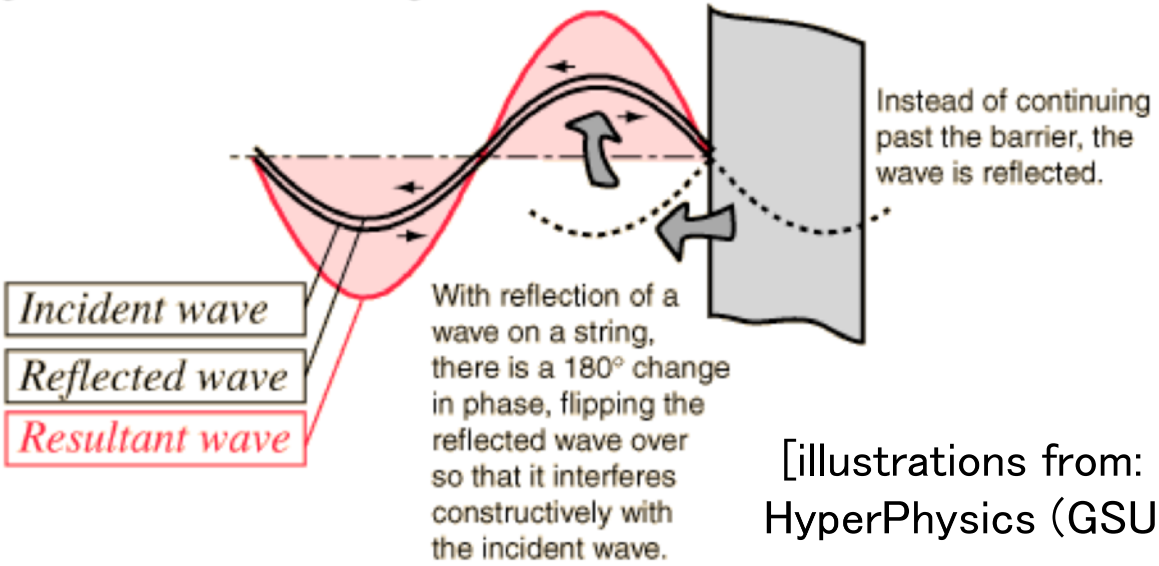
- What is a standing wave:
  - A stationary vibration pattern that results from the combination of reflection and interference such that the reflected waves interfere constructively with the incident waves.
  - It has nodes – points where the medium doesn't move, and antinodes – points where the motion is a maximum.
  - The medium appears to vibrate in segments or regions and the fact that these vibrations are made up of traveling waves is not apparent – hence the term "standing wave".



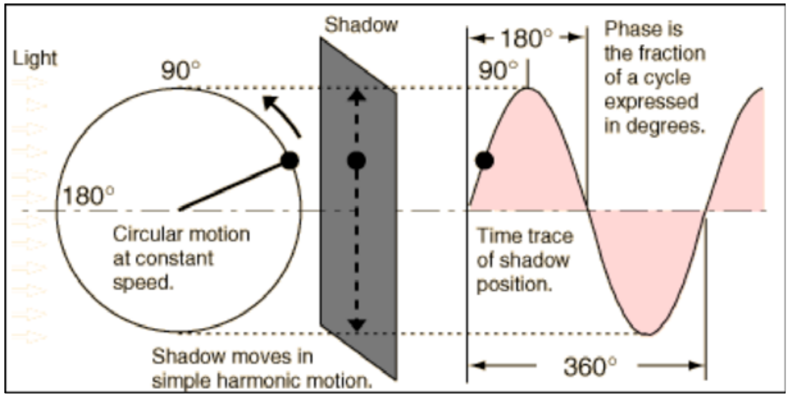
[Animation from: Daniel Russell website]

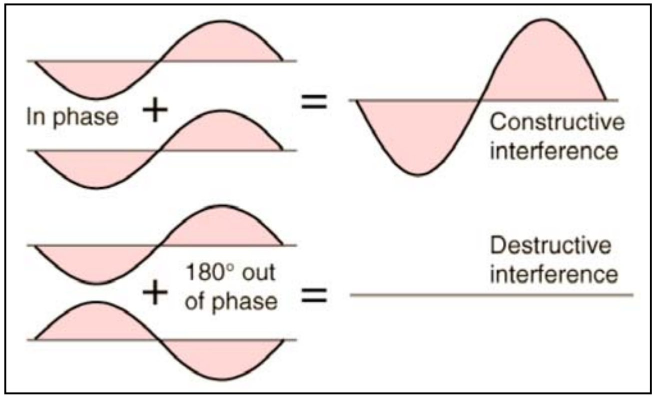
Standing waves

- Standing waves on a string that is fixed at both ends:


Incident wave
Reflected wave
Resultant wave
Instead of continuing past the barrier, the wave is reflected.
With reflection of a wave on a string, there is a 180° change in phase, flipping the reflected wave over so that it interferes constructively with the incident wave.

[illustrations from: C.R. Nave HyperPhysics (GSU) website)]

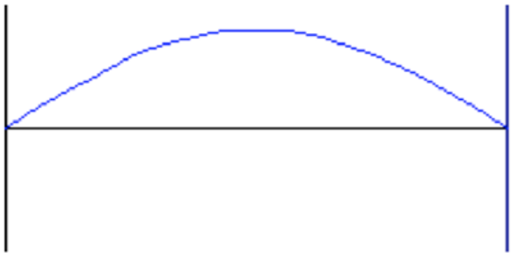

Light
90°
180°
Circular motion at constant speed.
Shadow
Shadow moves in simple harmonic motion.
180°
90°
Phase is the fraction of a cycle expressed in degrees.
Time trace of shadow position.
360°

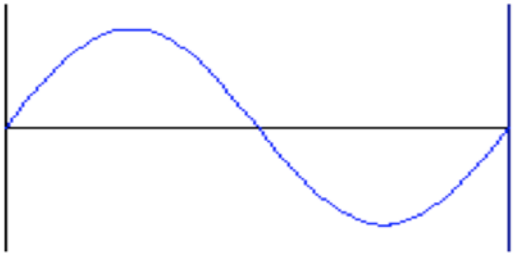


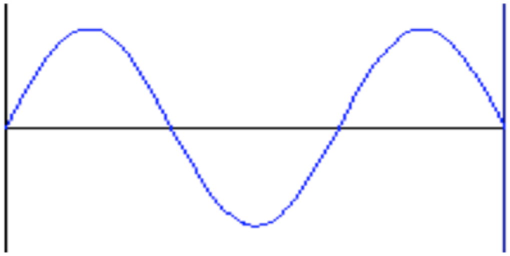
Standing waves

- Standing waves on a string that is fixed at both ends:

$$\lambda_n = \frac{2L}{n} \qquad f_n = \frac{v}{\lambda_n} = n\frac{v}{2L} \qquad n = 1, 2, 3\ldots$$

	L = λ/2
	λ = 2L
	f = v/(2L) first harmonic

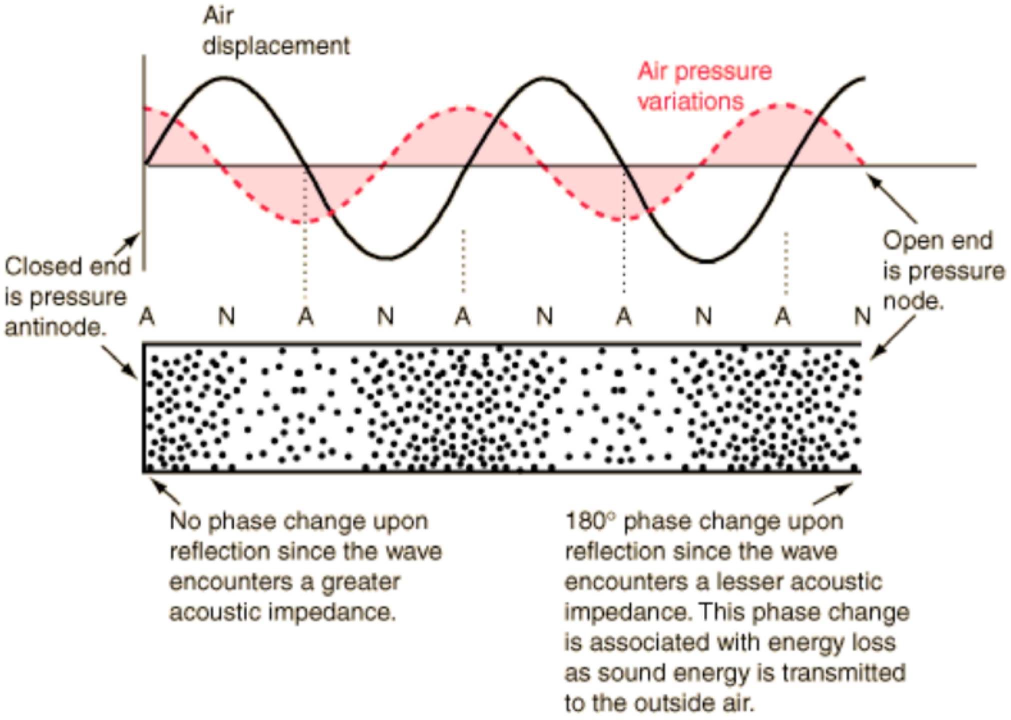
	L = λ
	λ = L
	f = v/(L) second harmonic

	L = 3λ/2
	λ = 2L/3
	f = 3v/(2L) third harmonic

Standing waves

- Standing waves in an air tube that is closed at one end and open at the other:
  - A node for air displacement is always an antinode for air pressure, and vice versa.

Production of a standing wave in an air column involves reflections from both the closed end and the open end of the column.



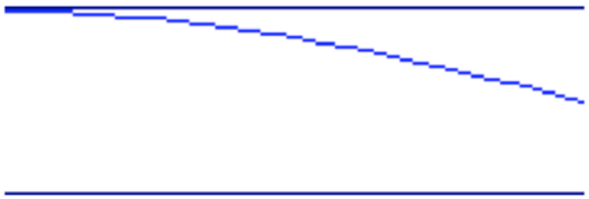
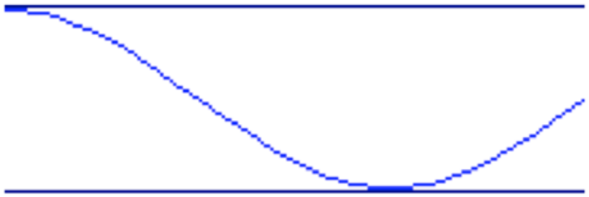
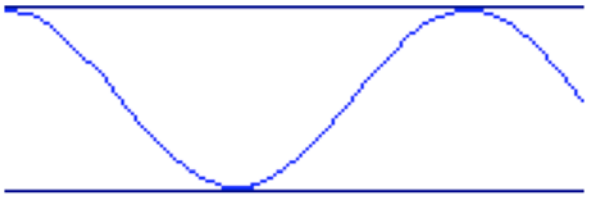
No phase change upon reflection since the wave encounters a greater acoustic impedance.

180° phase change upon reflection since the wave encounters a lesser acoustic impedance. This phase change is associated with energy loss as sound energy is transmitted to the outside air.

Standing waves

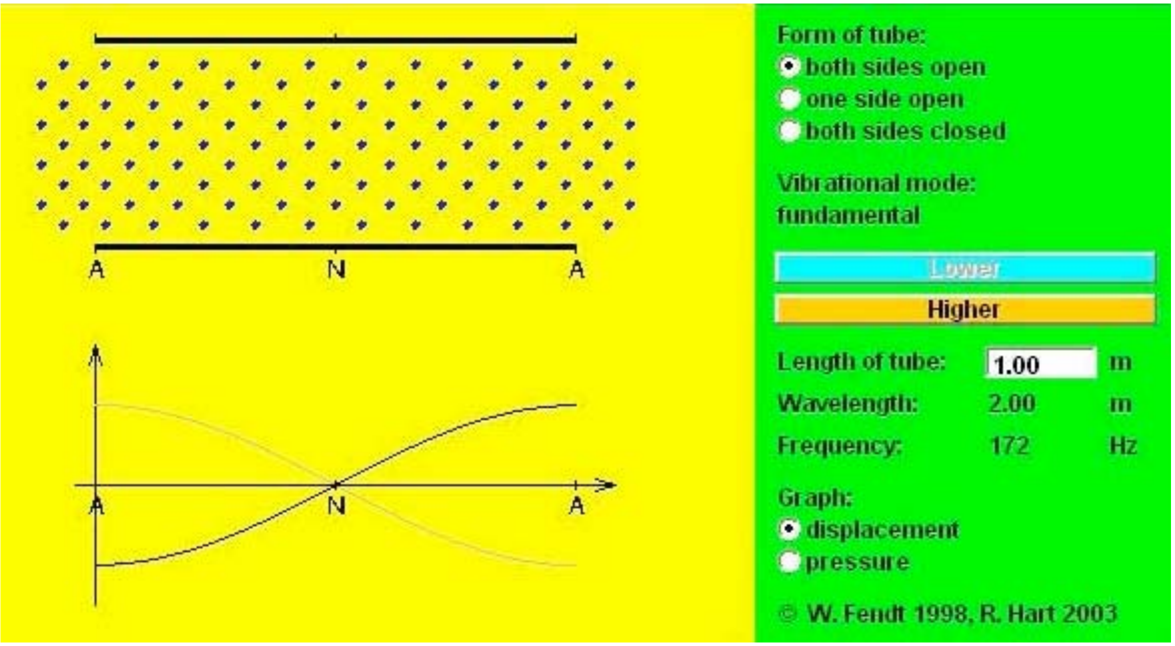
- Standing waves in an air tube that is closed at one end and open at the other:

$$\lambda_n = \frac{4L}{n} \qquad f_n = \frac{v}{\lambda_n} = n\frac{v}{4L} \qquad n = 1, 3, 5...$$

	$L = \lambda/4$ $\lambda = 4L$ $f = v/(4L)$ first harmonic
	$L = 3\lambda/4$ $\lambda = 4L/3$ $f = 3v/(4L)$ third harmonic
	$L = 5\lambda/4$ $\lambda = 4L/5$ $f = 5v/(4L)$ fifth harmonic

Standing waves

- **Demo:**



Form of tube:

● both sides open
○ one side open
○ both sides closed

Vibrational mode:
fundamental

Length of tube: 1.00 m
Wavelength: 2.00 m
Frequency: 172 Hz

Graph:

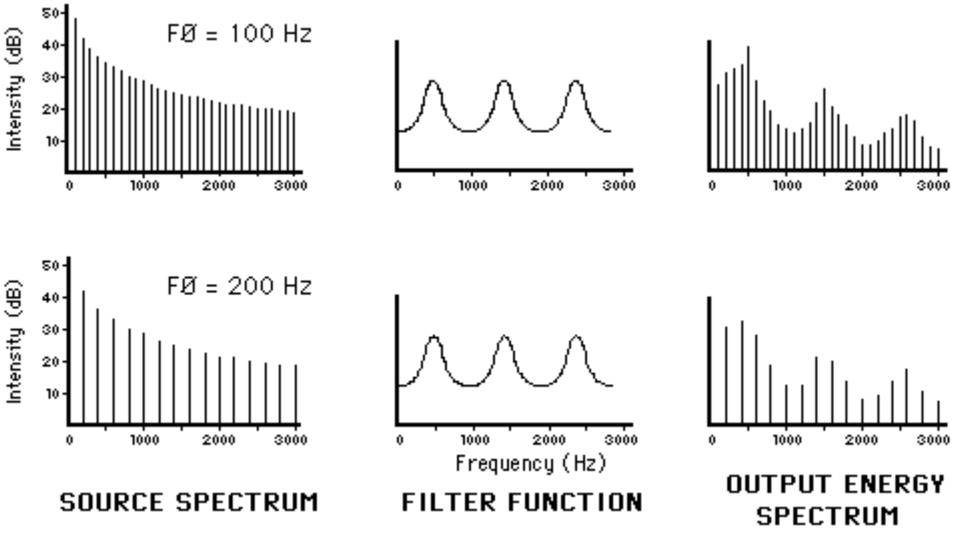
● displacement
○ pressure

© W. Fendt 1998, R. Hart 2003

[Applet from: Fredrick Olness website]

Standing waves

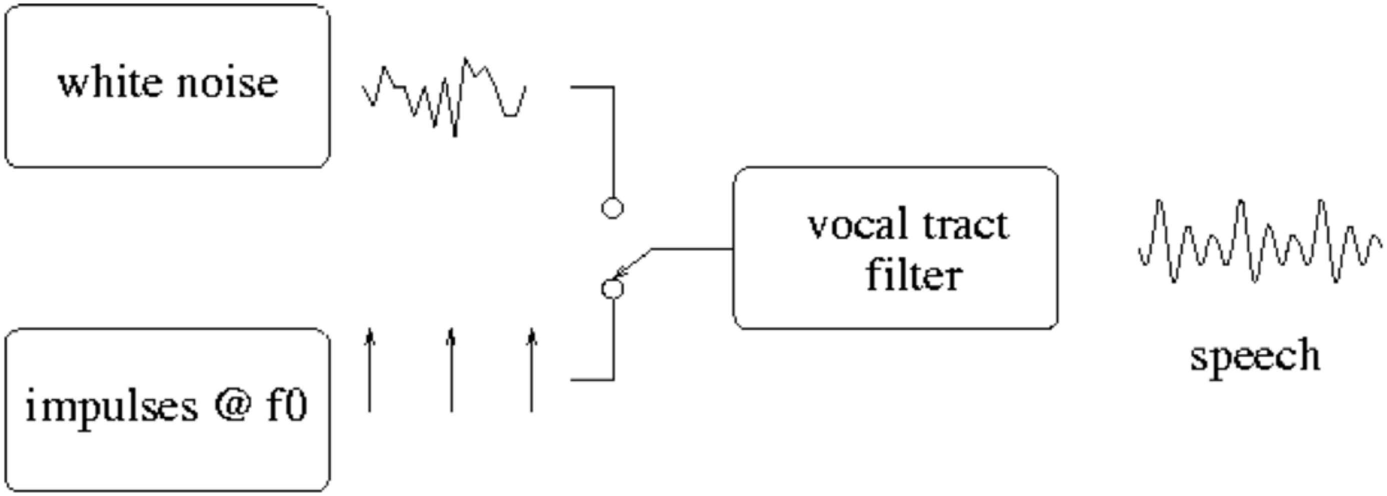
- **Standing waves as a property of a system**: filter, resonator.



[Picture from: Haskins Labs website]

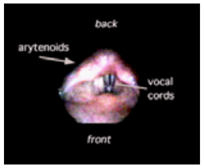
The source filter model of speech production

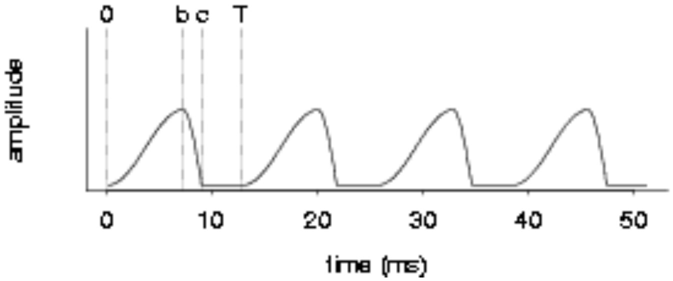
- The source filter model:
  - The production of speech consists of two kinds of operations: (1) the generation of sound sources, at the glottis or at some point along the length of the vocal tract, and (2) the filtering of these sources by the vocal tract.
  - There are two principle kinds of sources: a) turbulence noise (present in [s], [f], etc.), and b) vocal–fold vibration (present in vowels, nasals etc.).

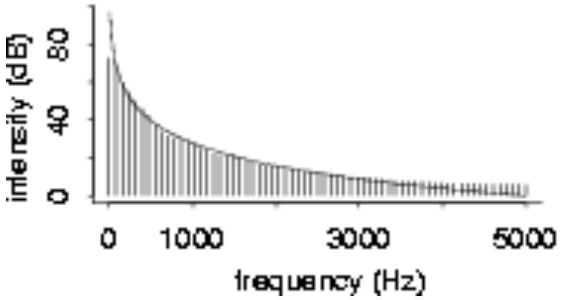


The source filter model of speech production

- The glottal source (impulses):
  - The vibration of the vocal folds are periodic, which determines the fundamental frequency.
  - The spectrum of the glottal source decreases in amplitude with increasing frequency at a rate of around –12dB per octave –– that is for each doubling in frequency, the amplitude of the spectrum decreases by around 12dB (later, radiation from the lips will raise the tilt by 6 dB/octave) .
  - Open Quotient (OQ): the ratio of the time in which the vocal folds are open and the whole pitch period duration.

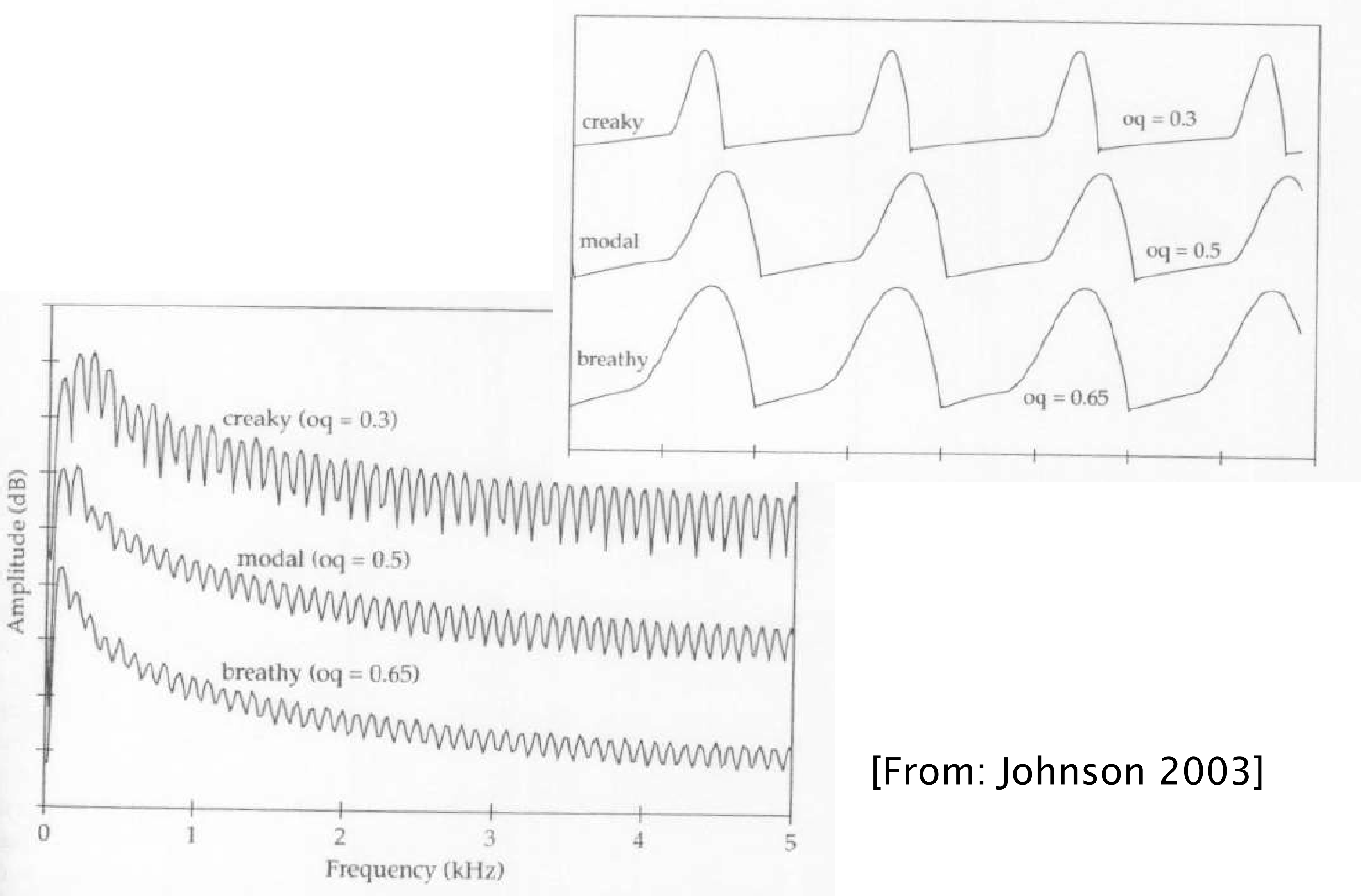






[Movie from: UCL phonetics website; figures from: Steve Cassidy website]

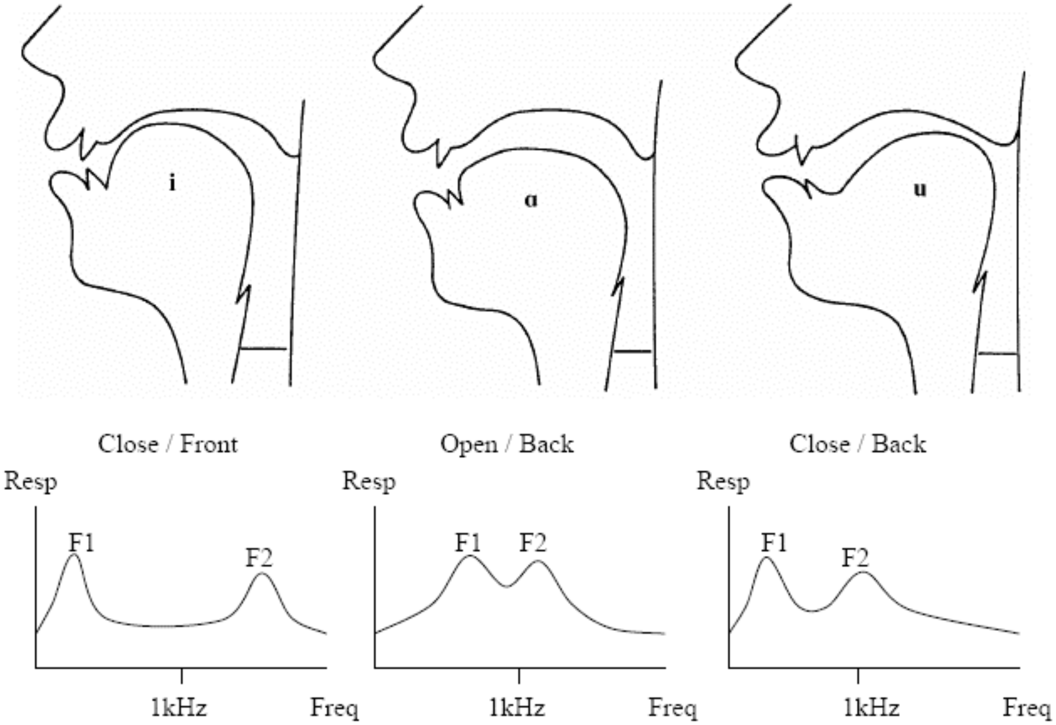
Open quotients and phonation types



[From: Johnson 2003]

The source filter model of speech production

- The vocal tract filter:



[From: UCL phonetics website]

- Formants are the resonance frequencies of the vocal tract.

Deriving schwa

- When the tongue is in the position of the neutral vowel, the schwa, the vocal tract (without the nasal cavity) can be approximated as a tube that is closed at one end (glottis) and open at the other (mouth).

$$\lambda_n = \frac{4L}{n} \qquad f_n = \frac{v}{\lambda_n} = n\frac{v}{4L} \qquad n = 1, 3, 5 \ldots$$

If L = 17.5 cm; v = 35,000 cm/s:

- $F_1 = v/\lambda_1 = v/(4L) =$ 500Hz
- $F_2 = v/\lambda_3 = v/(4/3L) =$ 1500Hz
- $F_3 = v/\lambda_5 = v/(4/5L) =$ 2500Hz

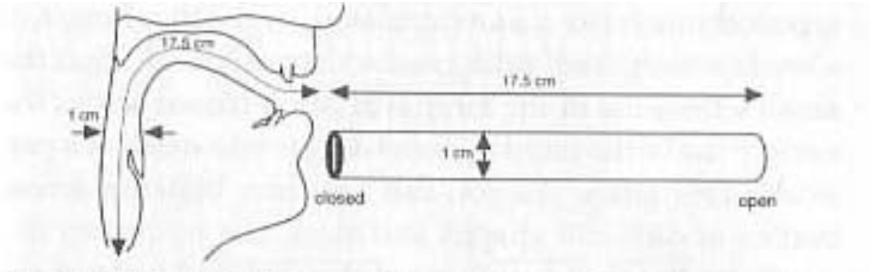
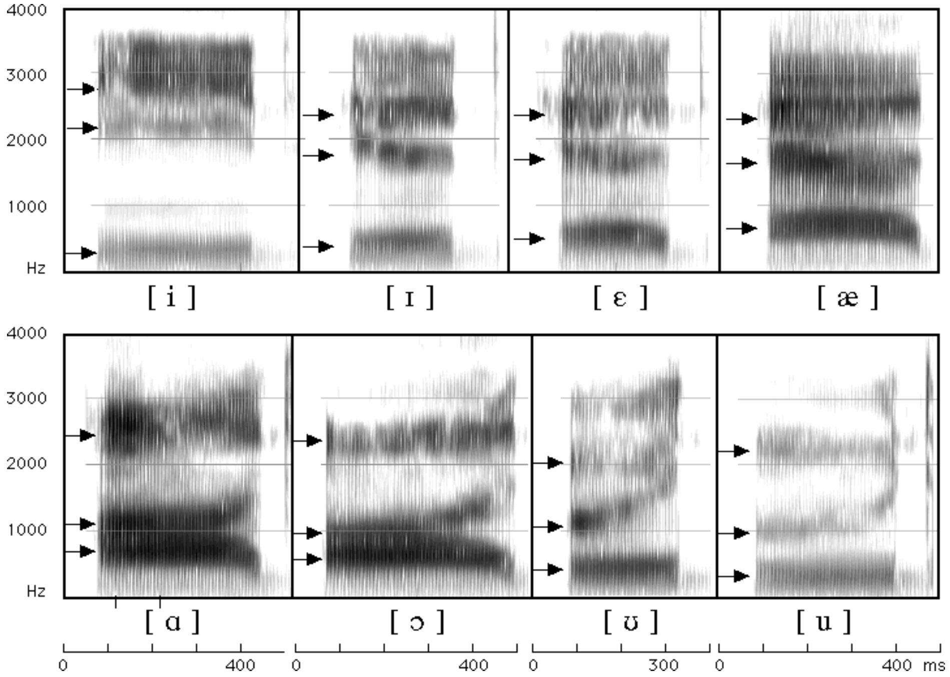


Fig. 8.2. A schematic diagram of a neutral vocal tract in the position for the vowel [ə] on the left, and a simplified version of that shape as a tube closed at one end on the right.

[Figure from: Ladefoged 1996]

Vowel formants

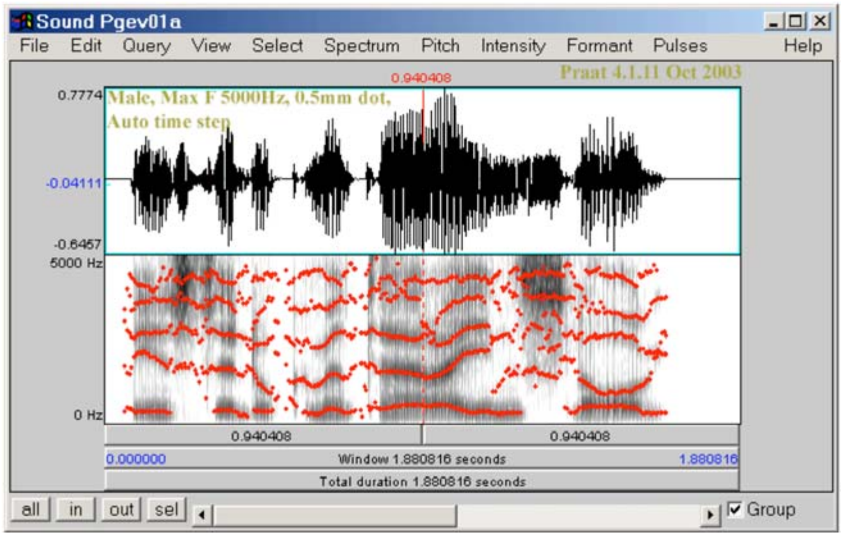
- Vowels are associated with a steady-state articulatory configuration and a steady-state acoustic pattern. Vowels often have been characterized with a very simple set pf acoustic descriptors, namely, the frequencies of the first three formants.
- American English: *heed, hid, head, had, hod, hawed, hood, who'd*



[From: Ladefoged, 2005]

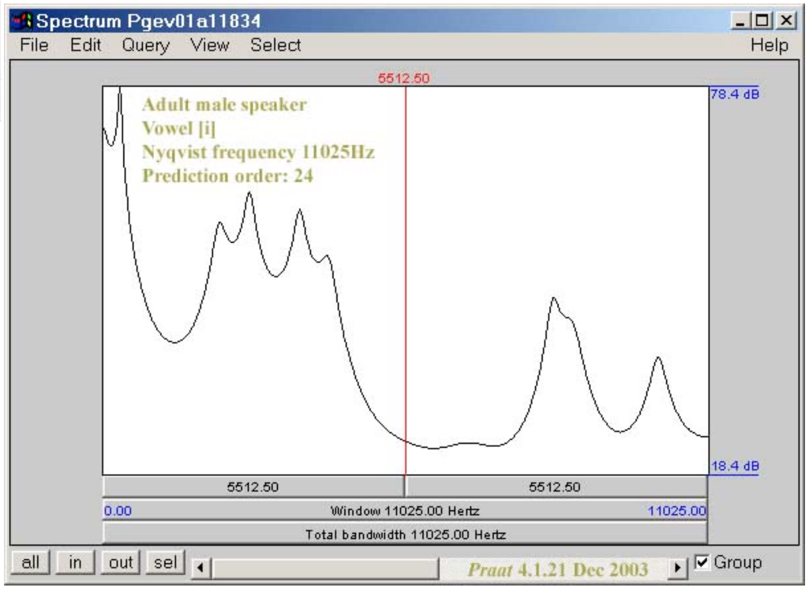
Measuring formants

- Spectrogram + formant tracking over a time span



Male, Max F 5000Hz, 0.5mm dot,
Auto time step

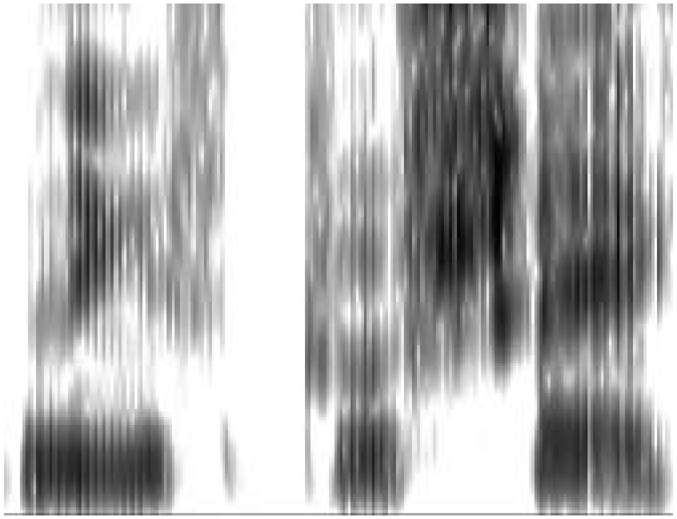
- LPC spectrum at one time point

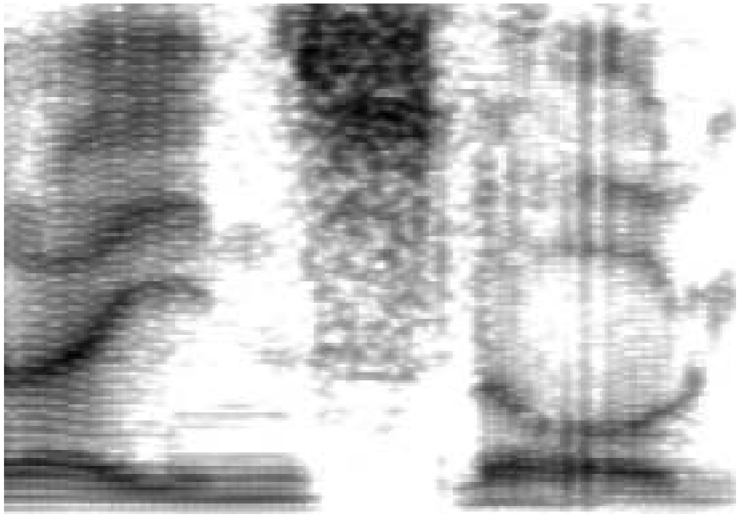


Adult male speaker
Vowel [i]
Nyqvist frequency 11025Hz
Prediction order: 24

Spectrograms

- Wideband and narrowband spectrograms
  - Traditionally, speech spectrograms are either *wideband* or *narrowband*, so called from the size of the electronic bandpass filter (330Hz or 45Hz respectively) that swept the frequency range on the original speech spectrographs.
  - The practical significance is that wideband spectrograms show formant structure while narrowband spectrograms reveal the harmonic structure.





Spectrograms

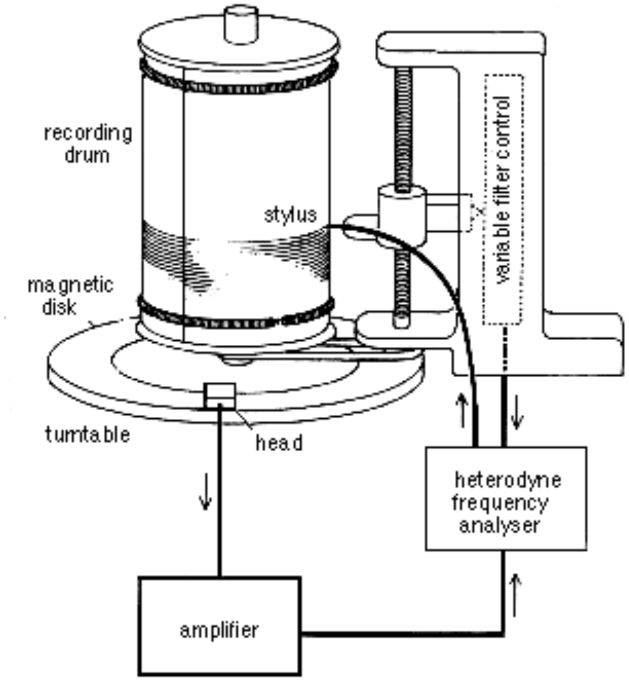
- Trade off between time and frequency resolution
  - The general rule is that the wider the passband the coarser the analysis in frequency but the finer the analysis in time, whereas the narrower the passband the finer the analysis in frequency but the coarser the analysis in time.

•DFT (to compute spectrograms):

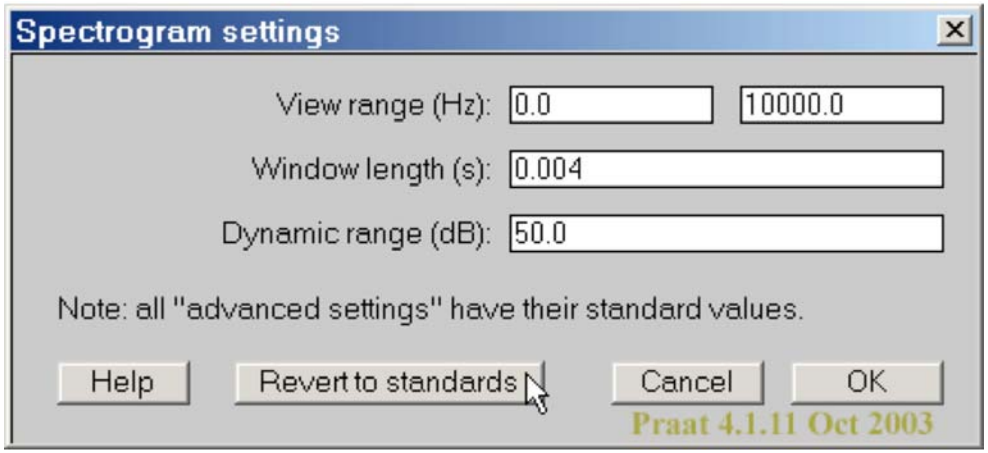
$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}n\cdot k}, \; k = 0,1,\ldots,N-1$$

•**FFT**: a computationally efficient algorithm for implementing DFT. It must use power of two points (zero padding).

•In DFT and FFT, time resolution and frequency resolution is a trade-off, due to the connection between window size (N) and frequency resolution (*2*pi/N*).



Spectral analysis in Praat



Spectrogram settings

View range (Hz): 0.0 10000.0

Window length (s): 0.004

Dynamic range (dB): 50.0

Note: all "advanced settings" have their standard values.

- *View Range*: the range of frequencies to display.
- *Window length:* the duration of the analysis window. The window length determines the *bandwidth* of the spectral analysis. To get a “wide-band” spectrogram (bandwidth 260 Hz), keep the standard window length of 5 ms; to get a “narrow-band” spectrogram (bandwidth 43 Hz), set it to 30 ms (0.03 seconds).
- *Dynamic range* (dB): All values that are more than *Dynamic range* dB below the maximum will be drawn in white. Values in-between have appropriate shades of grey.

Formant tracking and LPC

- Formant tracking from spectrograms (using eyes, image processing, etc.)
- The most popular formant tracking algorithms are, however, based on LPC.
- Linear Predictive Coding (LPC): Each sample of the speech signal is modeled as the weighted sum of past samples. The weights (or coefficients) are used to derive a linear prediction filter. The LP filter represents the vocal tract, the frequency response of the filter (peaks in the spectrum of the LP filter) provides estimates of the formants of the speech signal.

$$\hat{y}[n] = \sum_{k=1}^{p} a_k y[n-k].$$

$$y[n] = b_1 y[n-1] + b_2 y[n-2] + \cdots$$

Linear Predictive Coding (LPC)

- Predict y[n] from a linear combination of its past values.

$$\hat{y}[n] = \sum_{k=1}^{p} a_k y[n-k].$$

- Find the predictor coefficients by minimizing the mean squared error E:

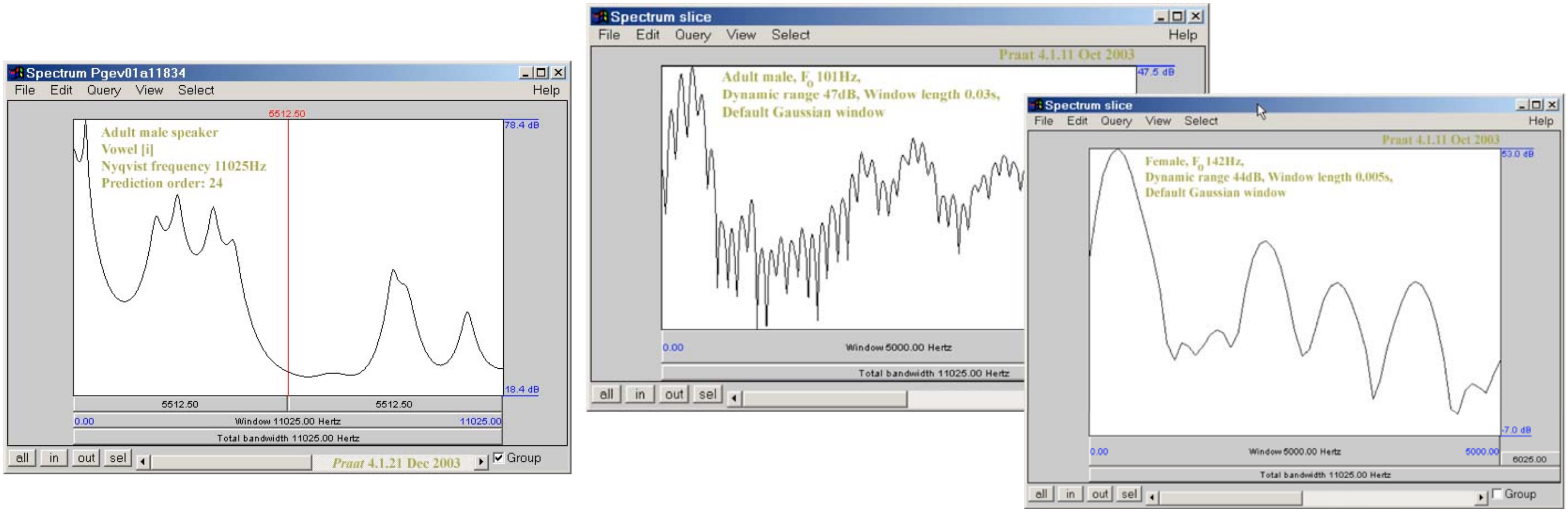
$$\varepsilon[n] \triangleq y[n] - \hat{y}[n] = y[n] - \sum_{k=1}^{p} a_k y[n-k].$$

$$E = \frac{1}{N}\sum_{n=0}^{N-1} \varepsilon[n]^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left(y[n] - \sum_{k=1}^{p} a_k y[n-k]\right)^2 .$$

- The linear predictor is an all-pole IIR filter. The all-pole IIR filter represents the vocal tract, therefore, the frequency response of the filter shows the resonances that shape the formants of the speech wave.
- In order for LPC to work, the vocal tract (the filter) must not have zeros. In mathematical terms, side branches introduces zeros. So LPC doesn’t work well for nasals.

LPC spectrum and FFT spectrum

- LPC spectrum: Linear prediction estimates the vocal tract filter and shows the resonances that shape the formants of the speech wave. There is no distinction between wideband and narrowband analysis in LPC. The LPC slice shows the envelope of the vocal tract resonances and looks superficially like a wideband FFT.
- FFT spectrum: Wideband FFT slices show the formant structure at the selected location, narrowband FFT slices show more detailed spectral structure.

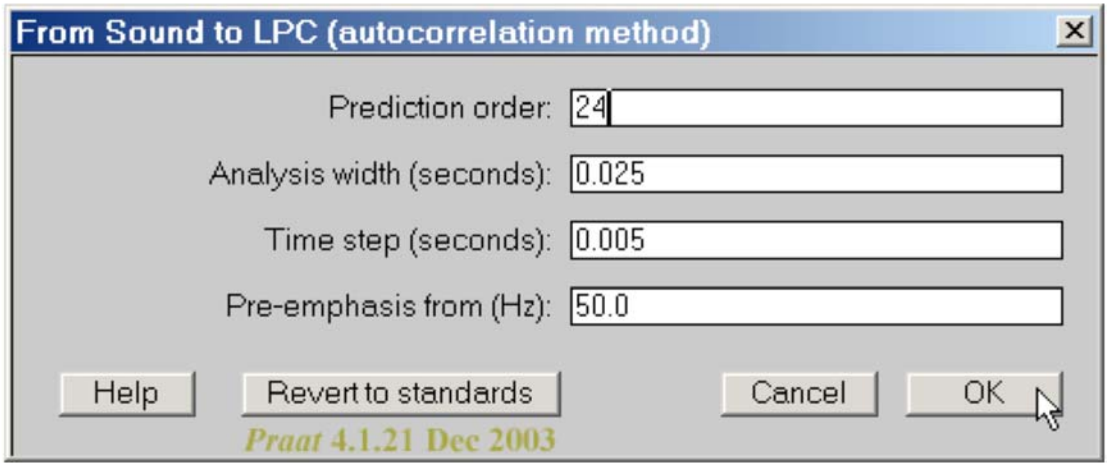


Adult male speaker
Vowel [i]
Nyqvist frequency 11025Hz
Prediction order: 24

Adult male, F_0 101Hz,
Dynamic range 47dB, Window length 0.03s,
Default Gaussian window

Female, F_0 142Hz,
Dynamic range 44dB, Window length 0.005s,
Default Gaussian window

LPC analysis in Praat



From Sound to LPC (autocorrelation method)

Prediction order: 24

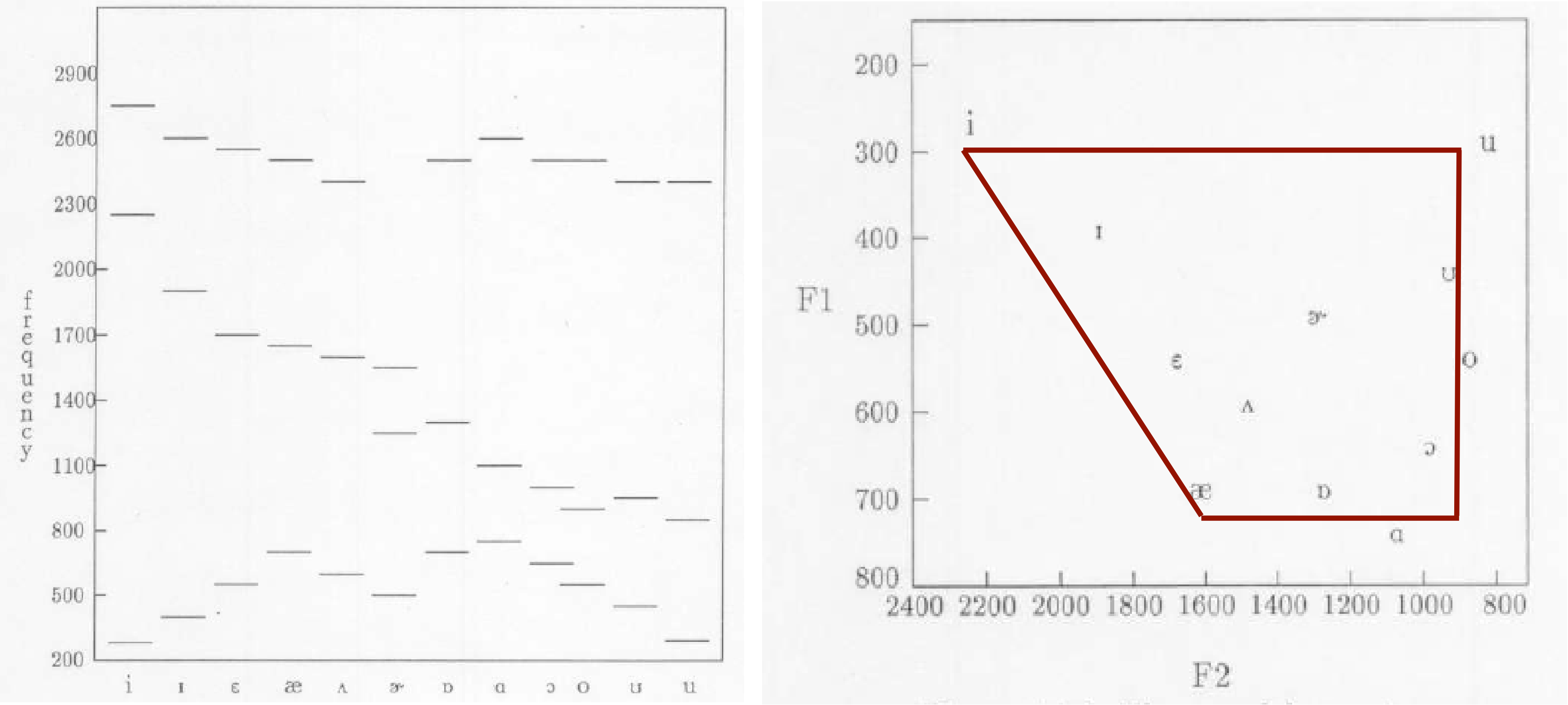
Analysis width (seconds): 0.025

Time step (seconds): 0.005

Pre-emphasis from (Hz): 50.0

- *Prediction order*: the number of filter coefficients; this is a critical setting; the absolute minimum setting is twice the number of formants in the frequency range of the signal (but a better setting is twice the number of formants plus at least 2);
- *Analysis width*: the amount of data from the signal that is needed for the computation; a good setting includes at least two glottal periods;
- *Time step*: the time interval between computed analysis frames;
- *Pre-emphasis from*: the effect is to high pass filter the signal by +6dB/octave, enhancing energy at higher frequencies.

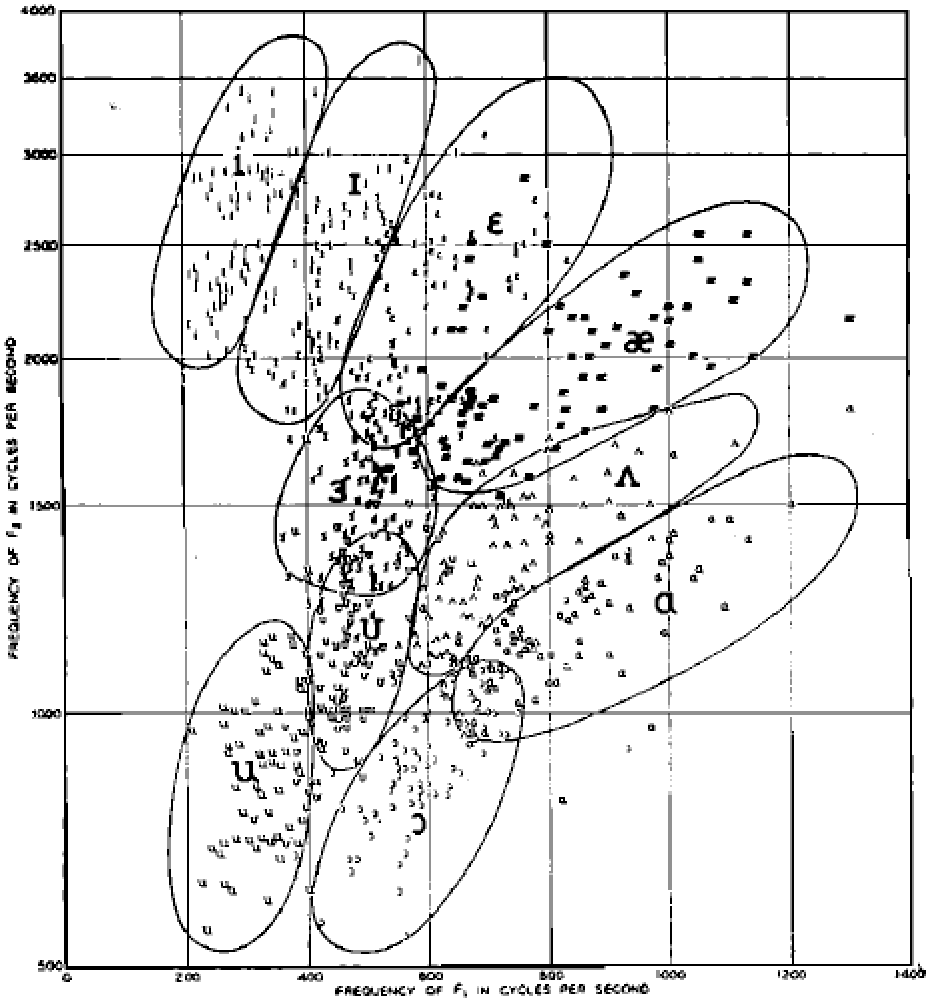
Acoustic vowel space



[From: *Acoustics of American English Speech*, 1993]

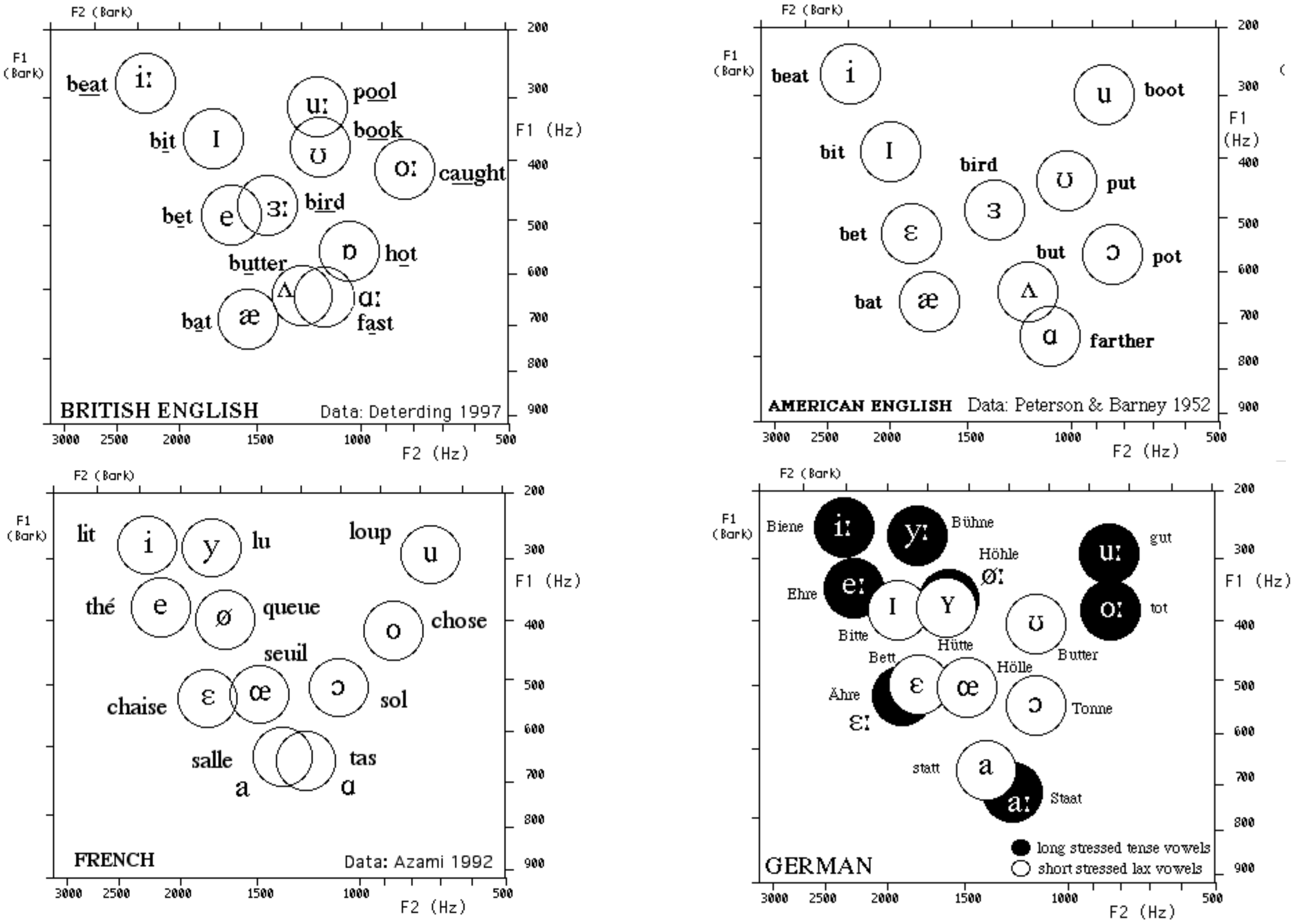
- **First formant: vowel height; Second formant: backness**

Acoustic vowel space



[From: Peterson and Barney 1952]

Acoustic vowel space



[From: www.helsinki.fi/hum/hyfl/projektit/vokaalikartat_eng.html]

Diphthongs

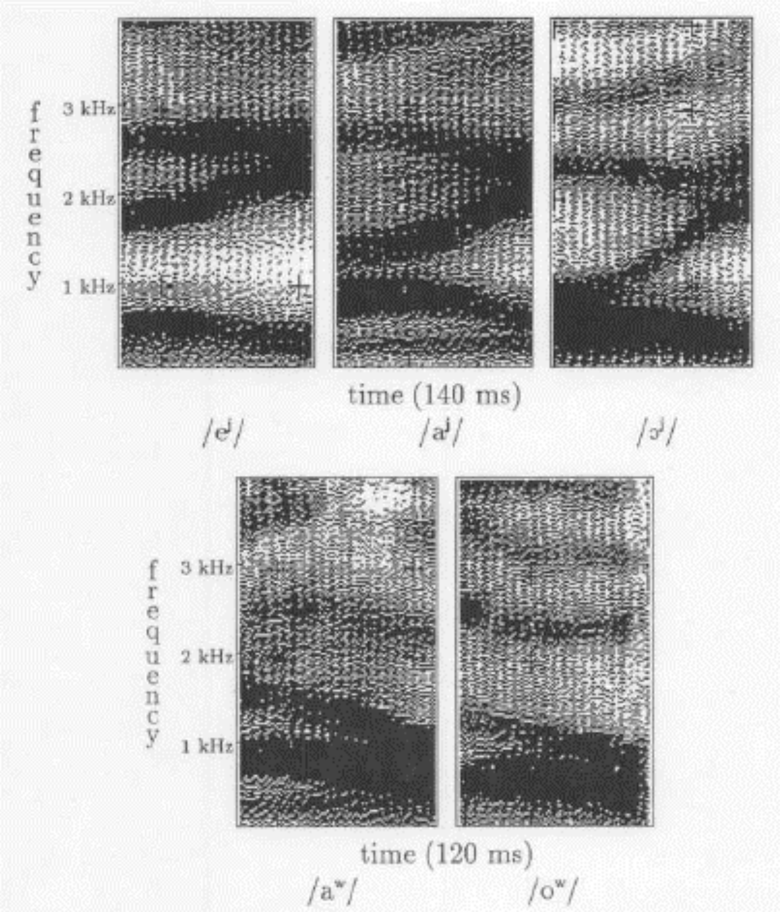


Figure 5.1: English diphthongs

[From: *Acoustics of American English Speech*, 1993]