CS 224S / LINGUIST 281
Speech Recognition and Synthesis

Dan Jurafsky

Lecture 2: Acoustic Phonetics

Outline for today

- Acoustic Phonetics
  - Waves, sound waves, and spectra
    - (**Informally!** We'll see it with more math when we do feature extraction)
  - Speech waveforms
  - F0, pitch, intensity
  - Spectra
    - Spectrograms
    - Formants
    - Reading spectrograms
  - Deriving schwa: why are formants where they are
  - PRAAT
  - Resources: dictionaries and phonetically-labeled corpora

Acoustic Phonetics

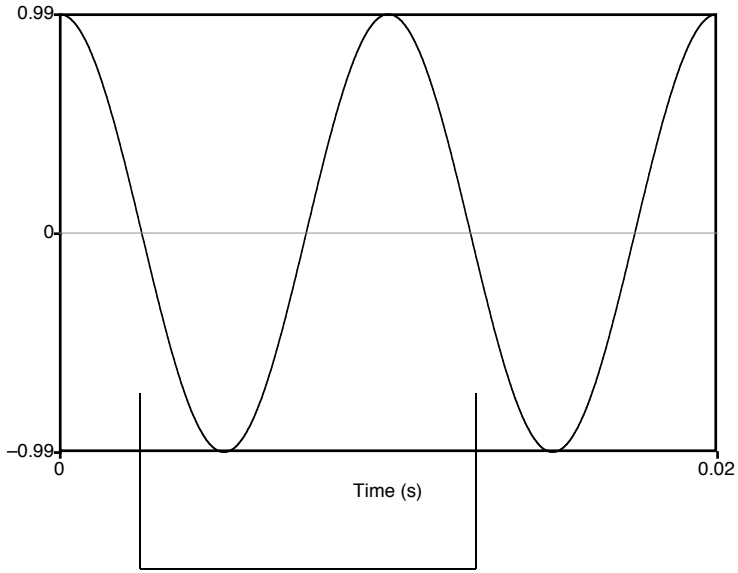
- Sound Waves
  - http://www.kettering.edu/~drussell/Demos/waves-intro/waves-intro.html

Simple Period Waves (sine waves)

- Characterized by:
  - period: T
  - amplitude A
  - phase ϕ
- Fundamental frequency

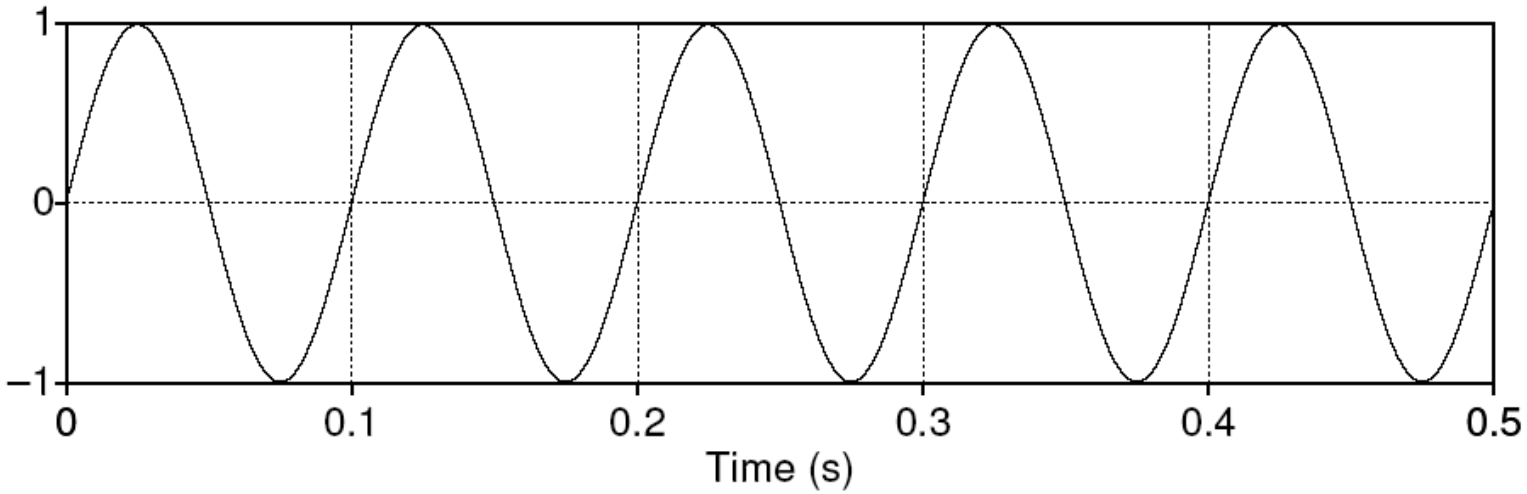
in cycles per second, or Hz

  - $F_0=1/T$



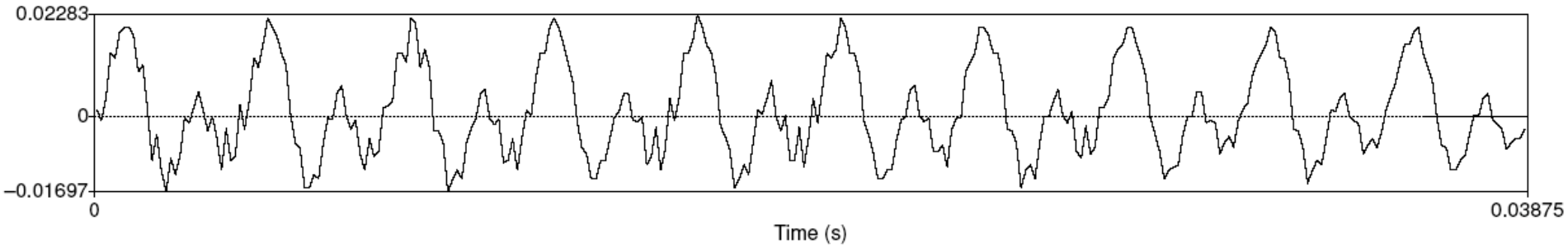
1 cycle

Simple periodic waves



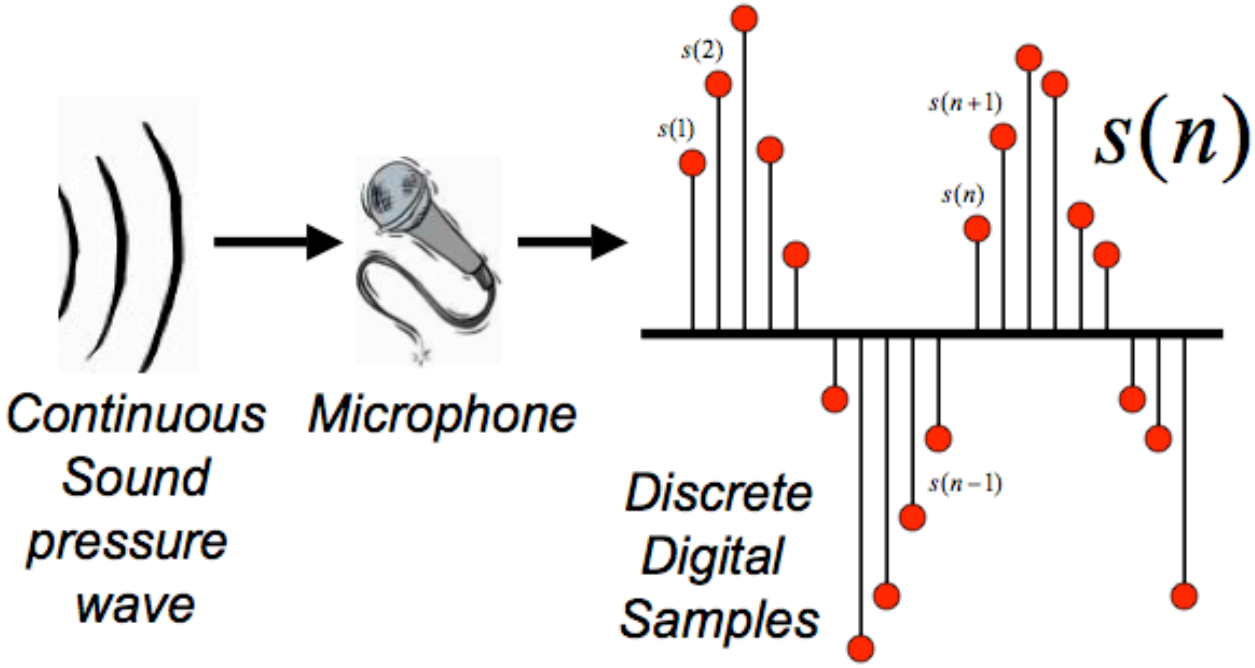
- Computing the frequency of a wave:
  - 5 cycles in .5 seconds = 10 cycles/second = 10 Hz
- Amplitude:
  - 1
- Equation:
  - $Y = A \sin(2\pi ft)$

Speech sound waves



- A little piece from the waveform of the vowel [iy]
- Y axis:
  - Amplitude = amount of air pressure at that time point
    - Positive is compression
    - Zero is normal air pressure,
    - negative is rarefaction
- X axis: time.

Digitizing Speech



Thanks to Bryan Pellom for this slide!

Digitizing Speech

- Analog-to-digital conversion
- Or A-D conversion.
- Two steps
  - Sampling
  - Quantization

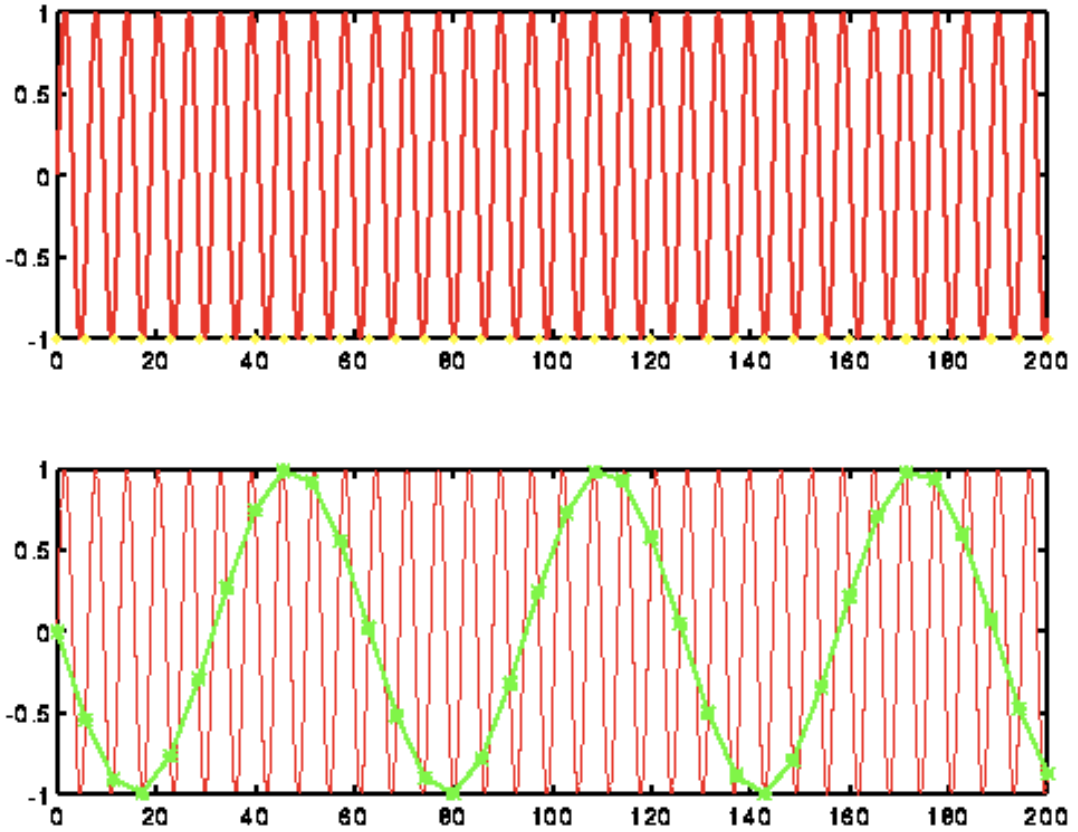
Sampling

- Measuring amplitude of a signal at time *t*
- The sample rate needs to have at least two samples for each cycle
  - One for the positiive, and one for the negative half of each cycle
  - More than two samples per cycle is ok
  - Less than two samples will cause frequencies to be missed
- So the maximum frequency that can be measured is one that is half the sampling rate.
- The maximum frequency for a given sampling rate called Nyquist frequency

Sampling

If measure at green dots, will see a lower frequency wave and miss the correct higher frequency one!

Original signal in red:

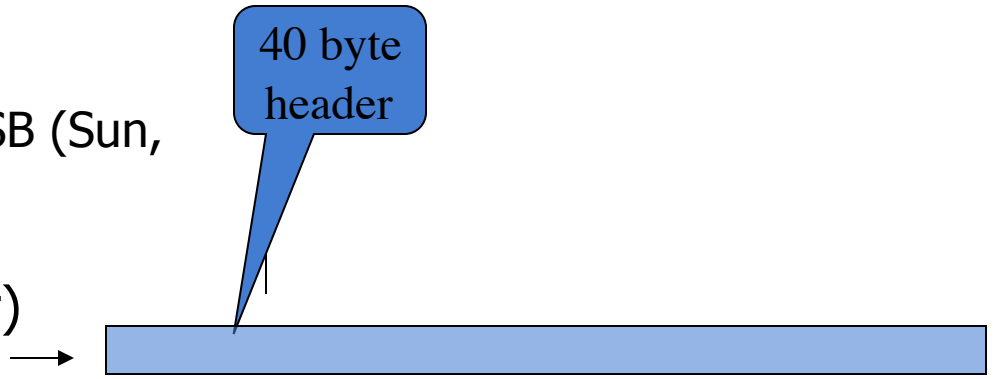


Sampling

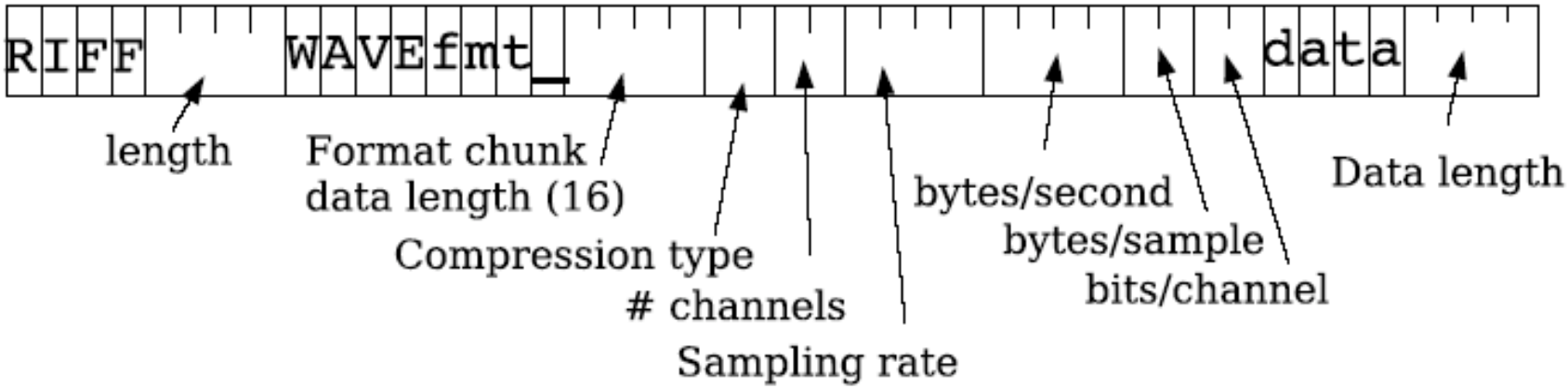
- In practice we use the following sample rates
  - 16,000 Hz (samples/sec), for microphones, "wideband"
  - 8,000 Hz (samples/sec) Telephone
- Why?
  - Need at least 2 samples per cycle
  - Max measurable frequency is half the sampling rate
  - Human speech < 10KHz, so need max 20K
  - Telephone is filtered at 4K, so 8K is enough.

Quantization

- Quantization
  - Representing real value of each amplitude as integer
  - 8-bit (-128 to 127) or 16-bit (-32768 to 32767)
- Formats:
  - 16 bit PCM
  - 8 bit mu-law; log compression
- Byte order
  - LSB (Intel) vs. MSB (Sun, Apple)
- Headers:
  - Raw (no header)
  - Microsoft wav
  - Sun .au

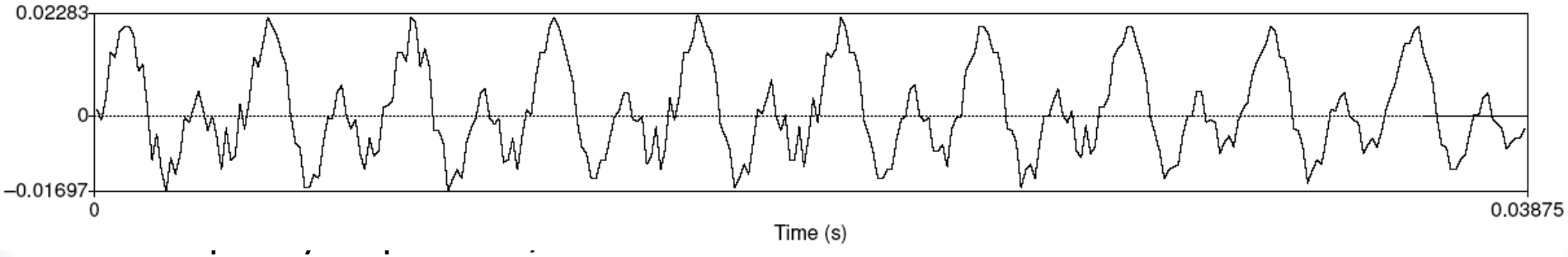


WAV format


RIFF
WAVEfmt_
data
length
Format chunk
data length (16)
Compression type
channels
Sampling rate
bytes/second
bytes/sample
bits/channel
Data length

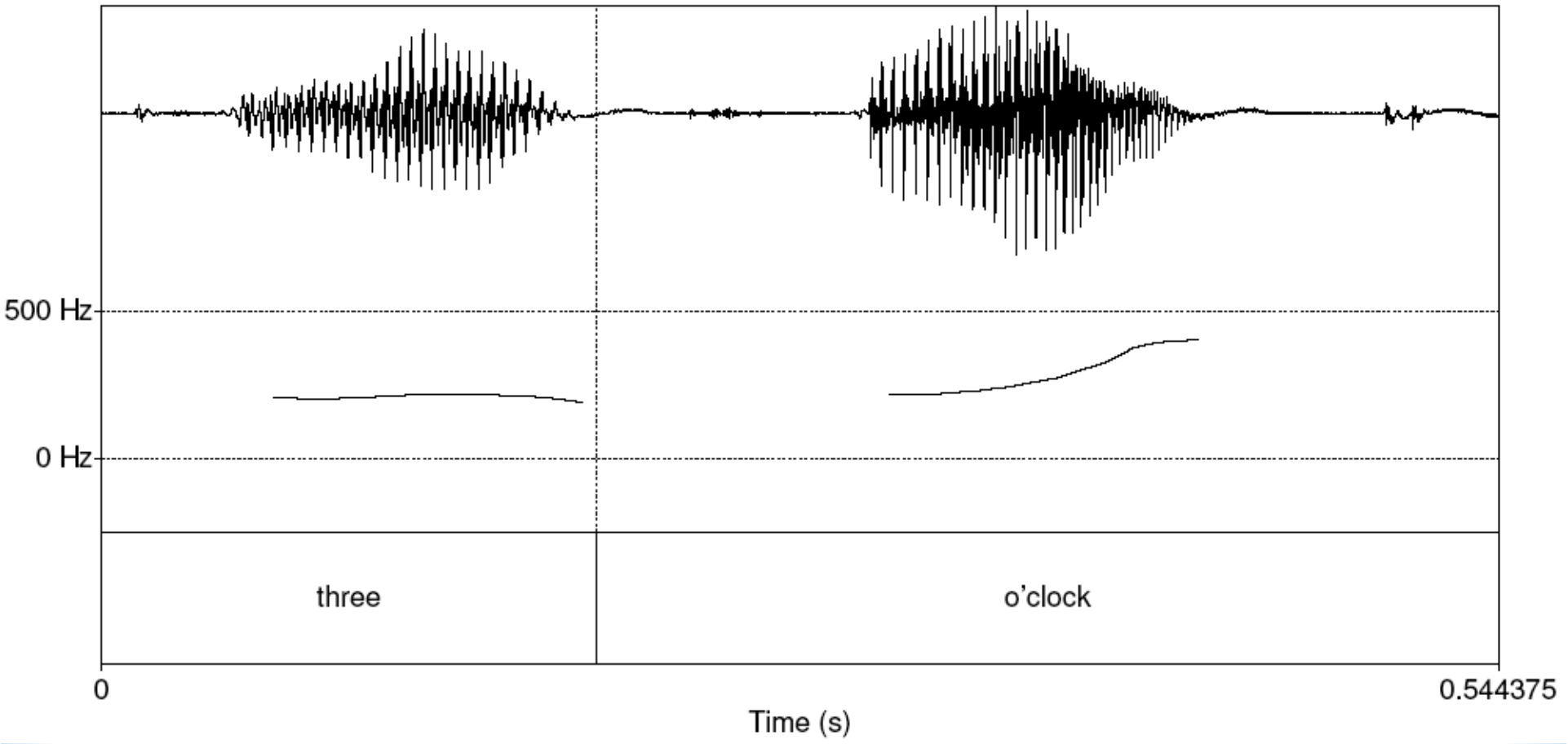
Fundamental frequency

- Waveform of the vowel [iy]



- Above vowel has 10 reps in .03875 secs
- So freq is 10/.03875 = 258 Hz
- This is speed that vocal folds move, hence voicing
- Each peak corresponds to an opening of the vocal folds
- The frequency of the complex wave is called the fundamental frequency of the wave or F0

Pitch track



Amplitude

- We need a way to talk about the amplitude of a region of a signal over tune
- We can't just average all the values.
- Why not?
- So we often talk about RMS amplitude

$$A_{RMS} = \sqrt{\sum_{i=1}^{N} \frac{x[i]^2}{N}}$$

Power and Intensity

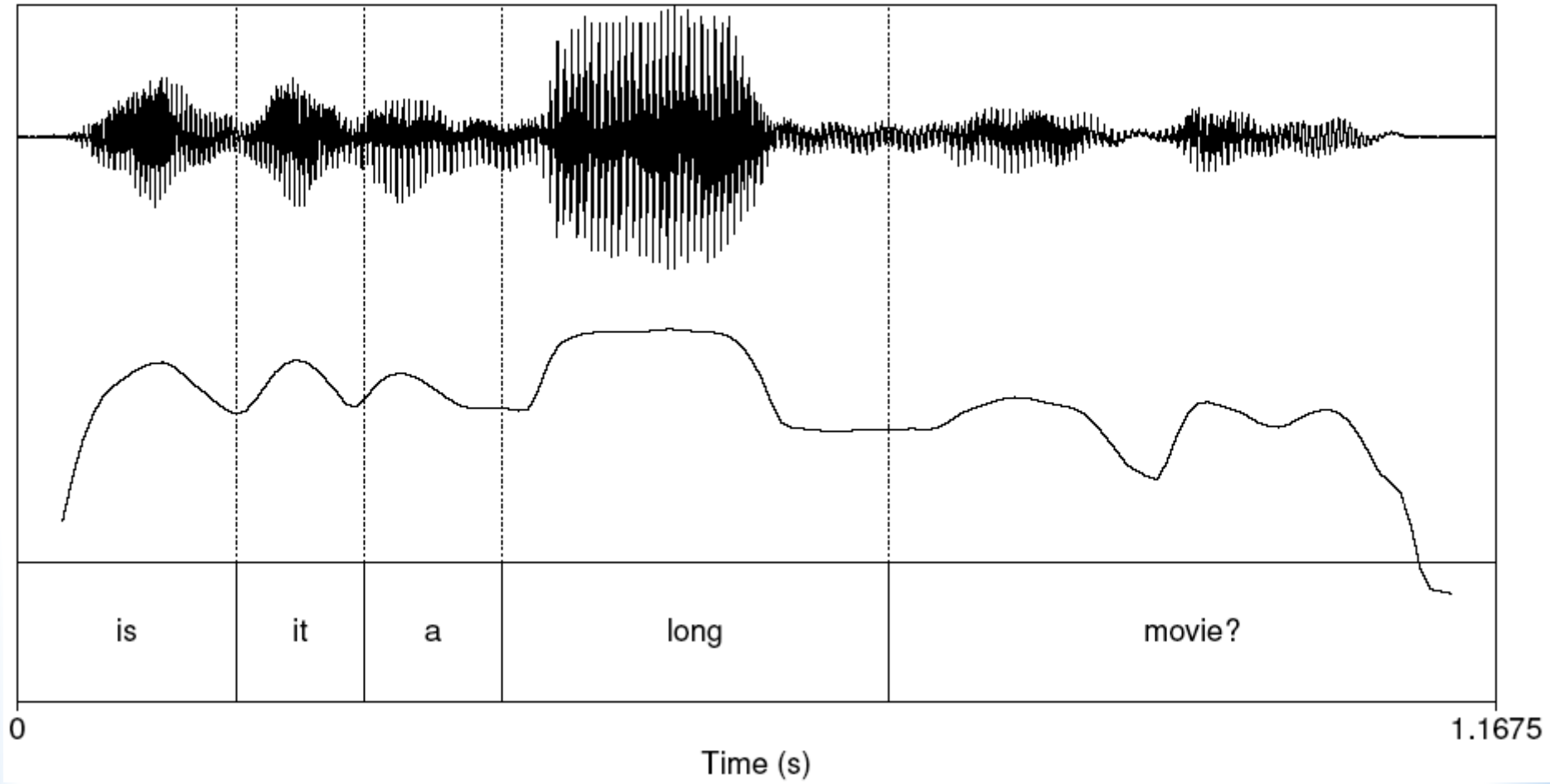
- Power: related to square of amplitude

$$Power = \frac{1}{N}\sum_{i=1}^{N} x[i]^2$$

- Intensity in air: power normalized to auditory threshold, given in dB. P0 is auditory threshold pressure = $2x10^{-5}$ pa

$$Intensity = 10\log_{10}\frac{1}{NP_0}\sum_{i=1}^{N} x[i]^2$$

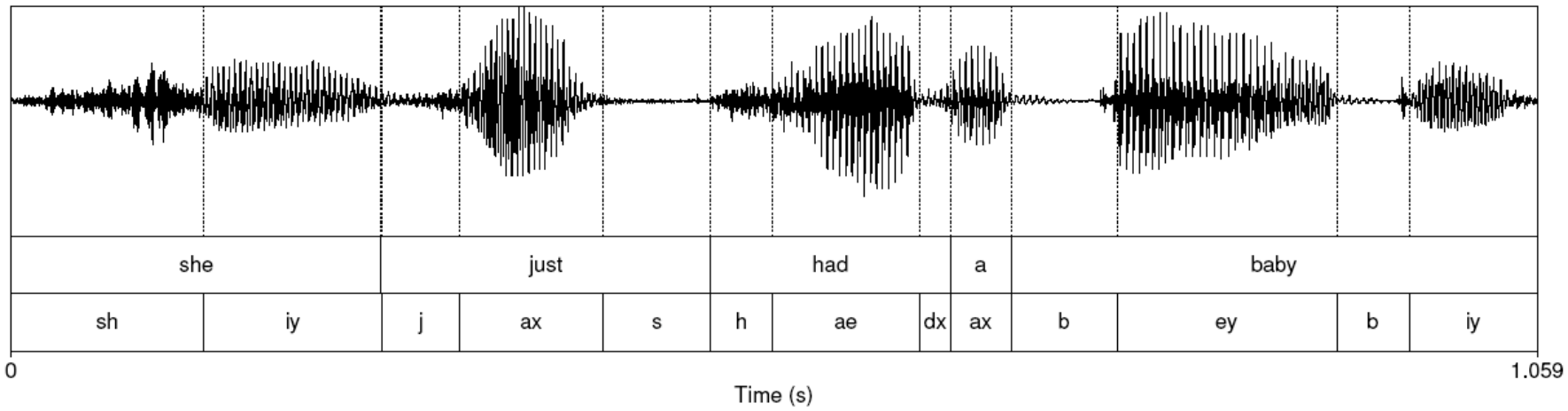
Plot of Intensity



Pitch and Loudness

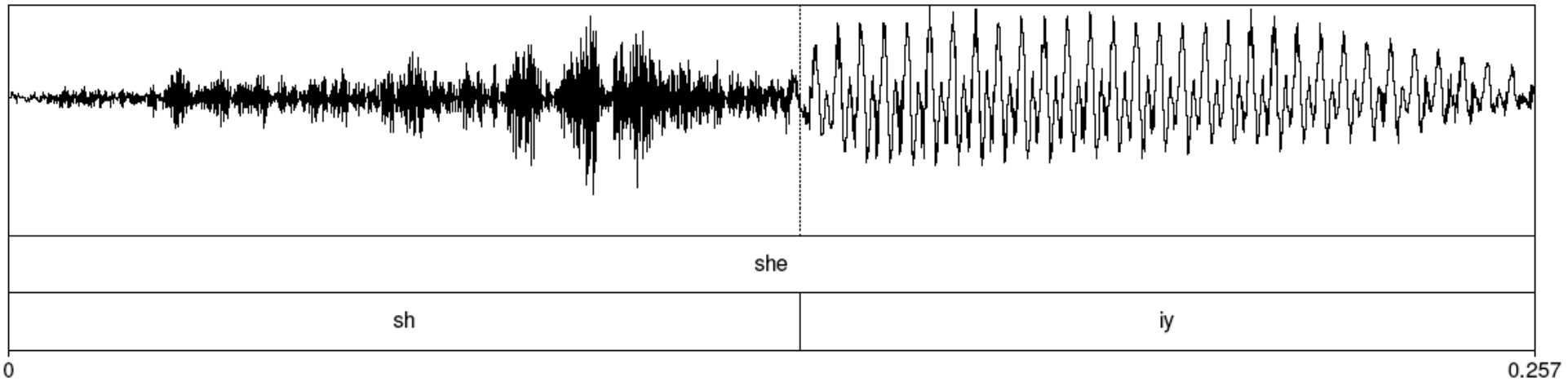
- Pitch is the mental sensation or perceptual correlated of F0
- Relationship between pitch and F0 is not linear;
  - human pitch perception is most accurate between 100Hz and 1000Hz.
    - Linear in this range
    - Logarithmic above 1000Hz
- Mel scale is one model of this F0-pitch mapping
  - A mel is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels
  - Frequency in mels = 1127 ln (1 + f/700)

She just had a baby

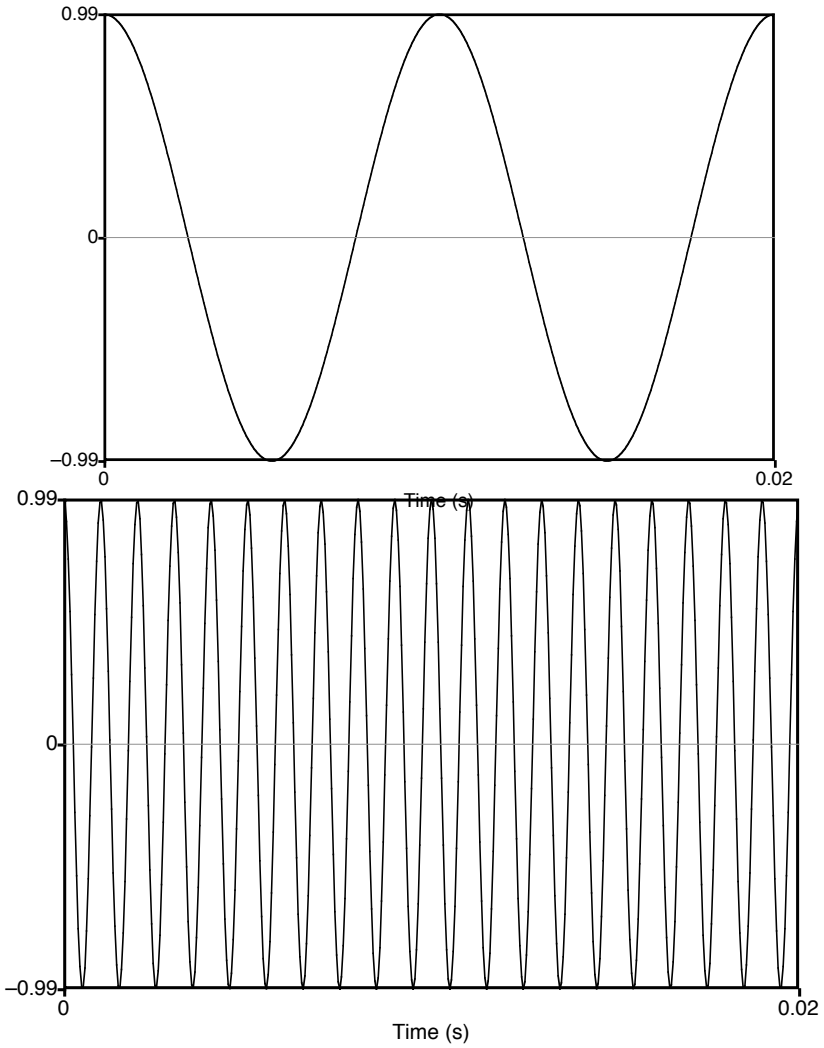


- Note that vowels all have regular amplitude peaks
- Stop consonant
  - Closure followed by release
  - Notice the silence followed by slight bursts of emphasis: very clear for [b] of "baby"
- Fricative: noisy. [sh] of "she" at beginning

Fricative



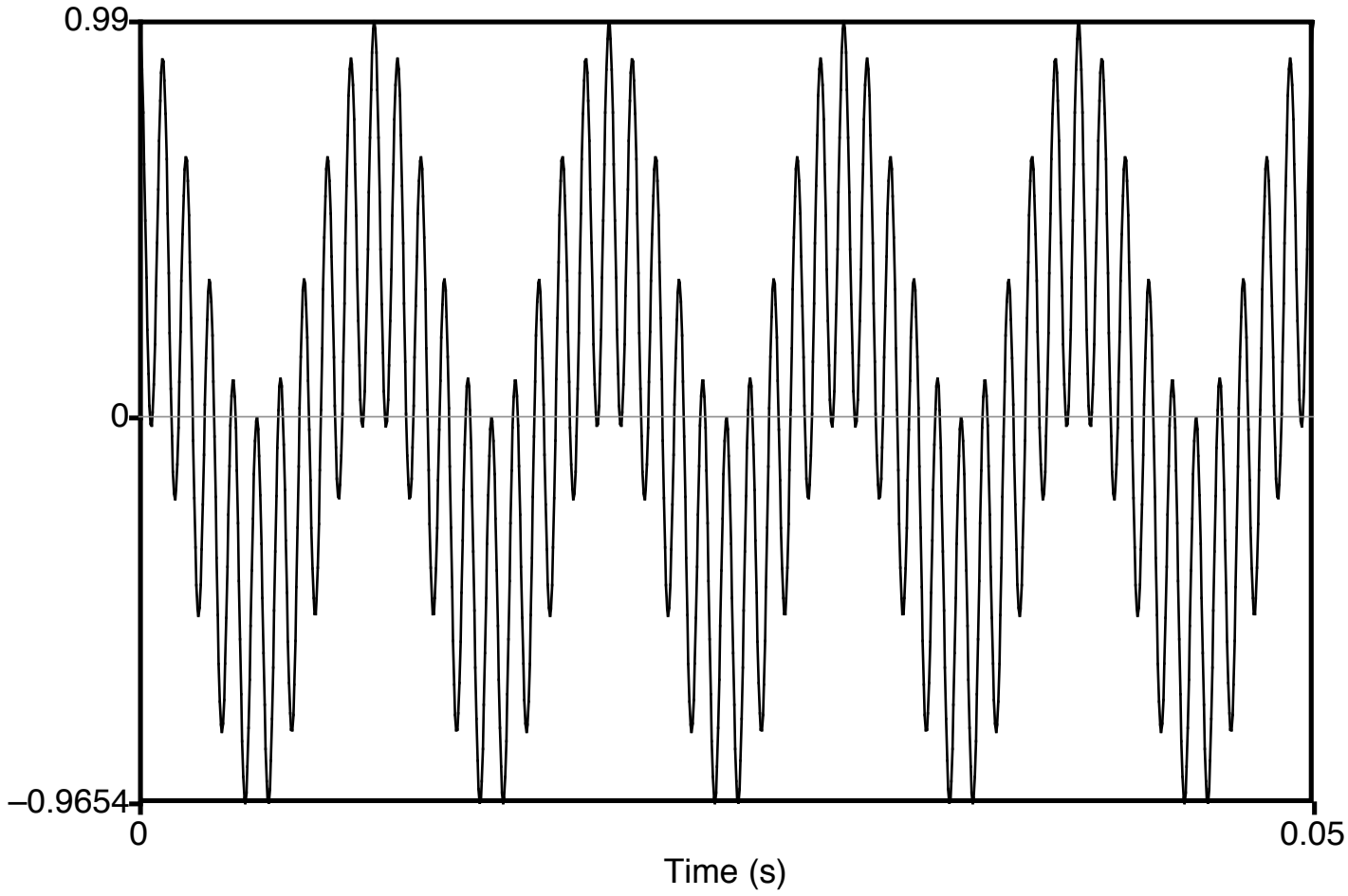
Waves have different frequencies



100 Hz

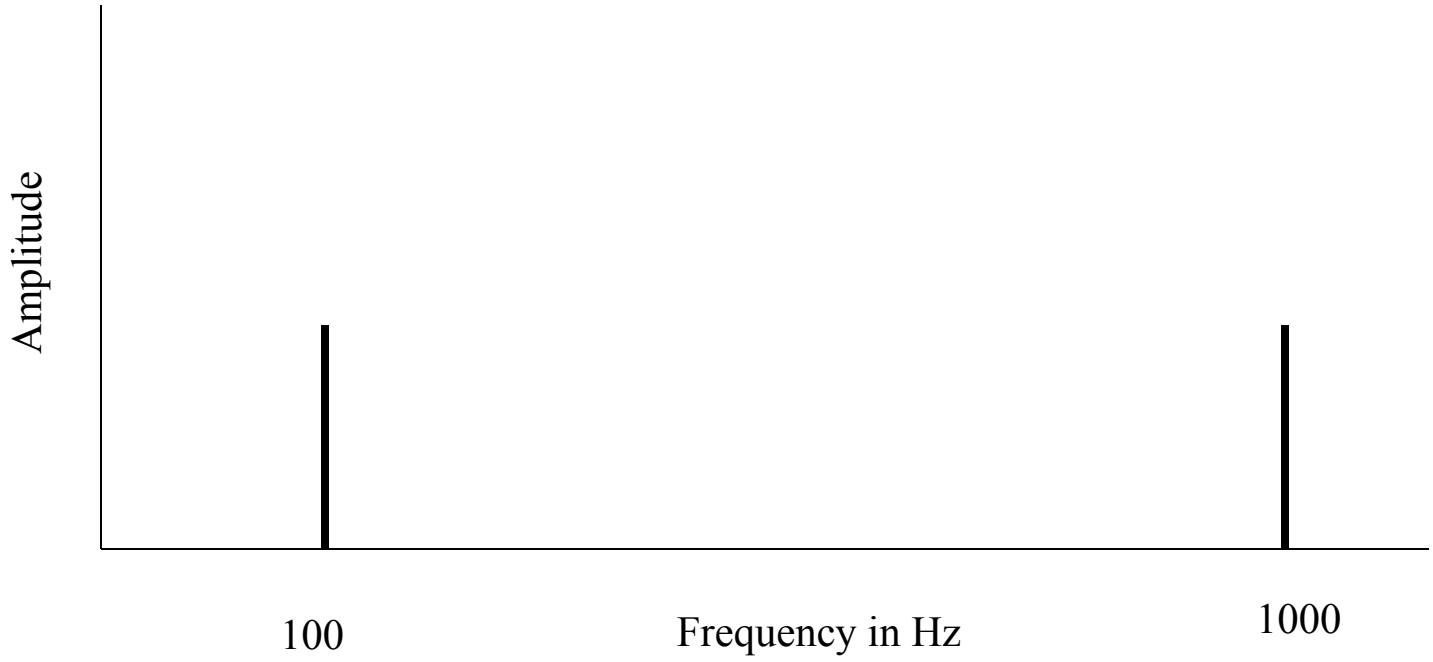
1000 Hz

Complex waves: Adding a 100 Hz and 1000 Hz wave together



Spectrum

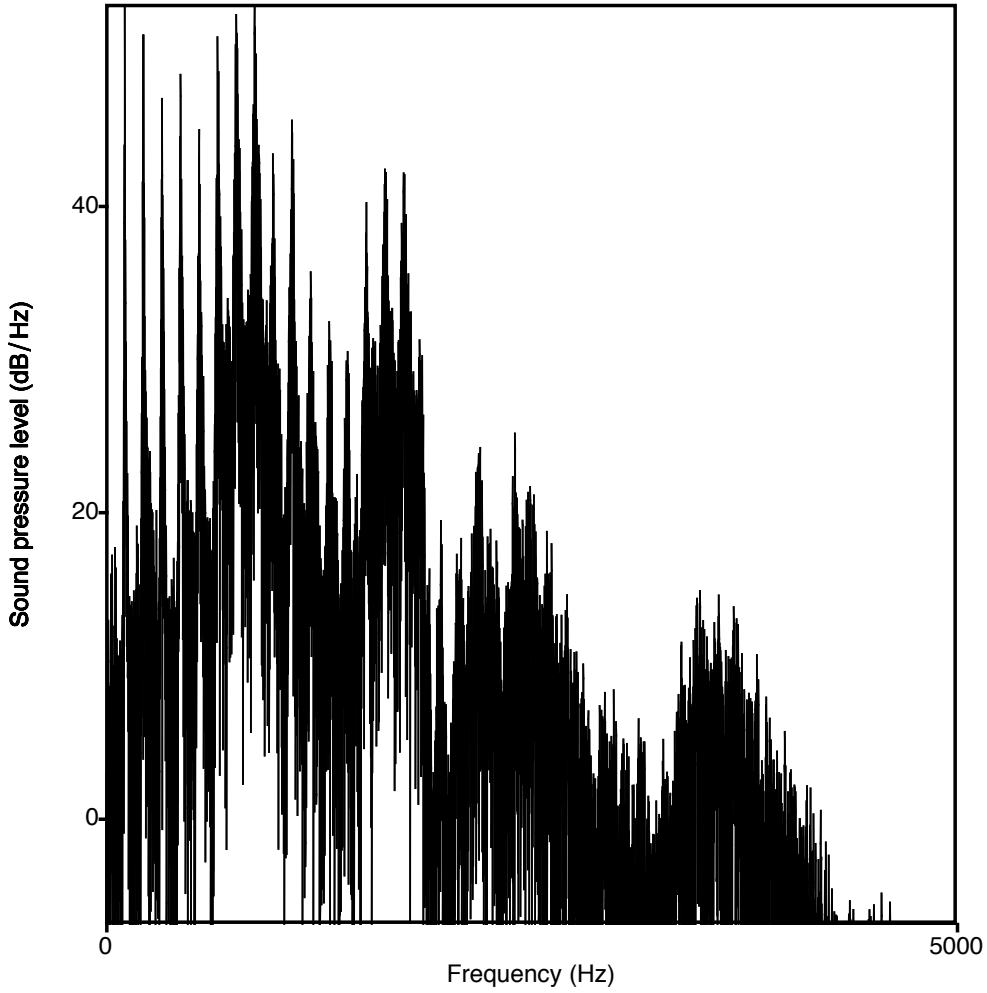
Frequency components (100 and 1000 Hz) on x-axis



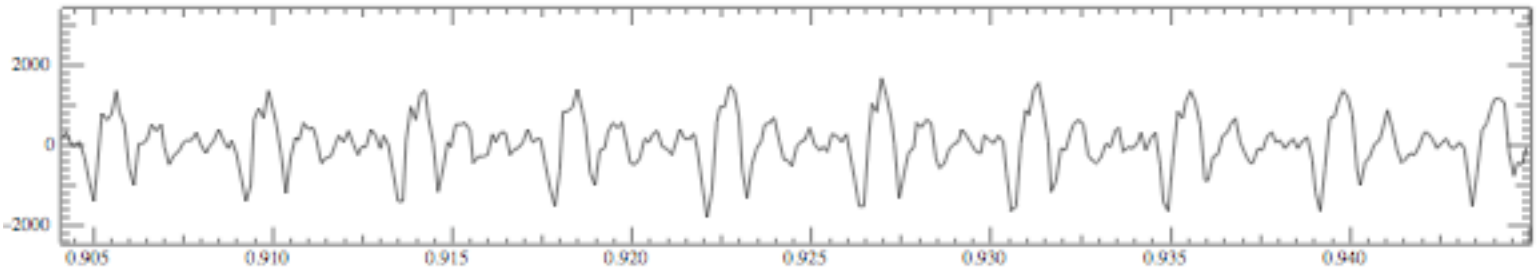
Spectra continued

- Fourier analysis: any wave can be represented as the (infinite) sum of sine waves of different frequencies (amplitude, phase)

Spectrum of one instant in an actual soundwave: many components across frequency range



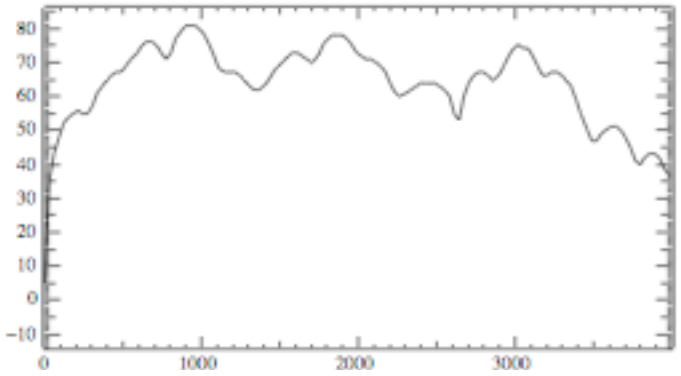
Part of [ae] waveform from "had"



- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

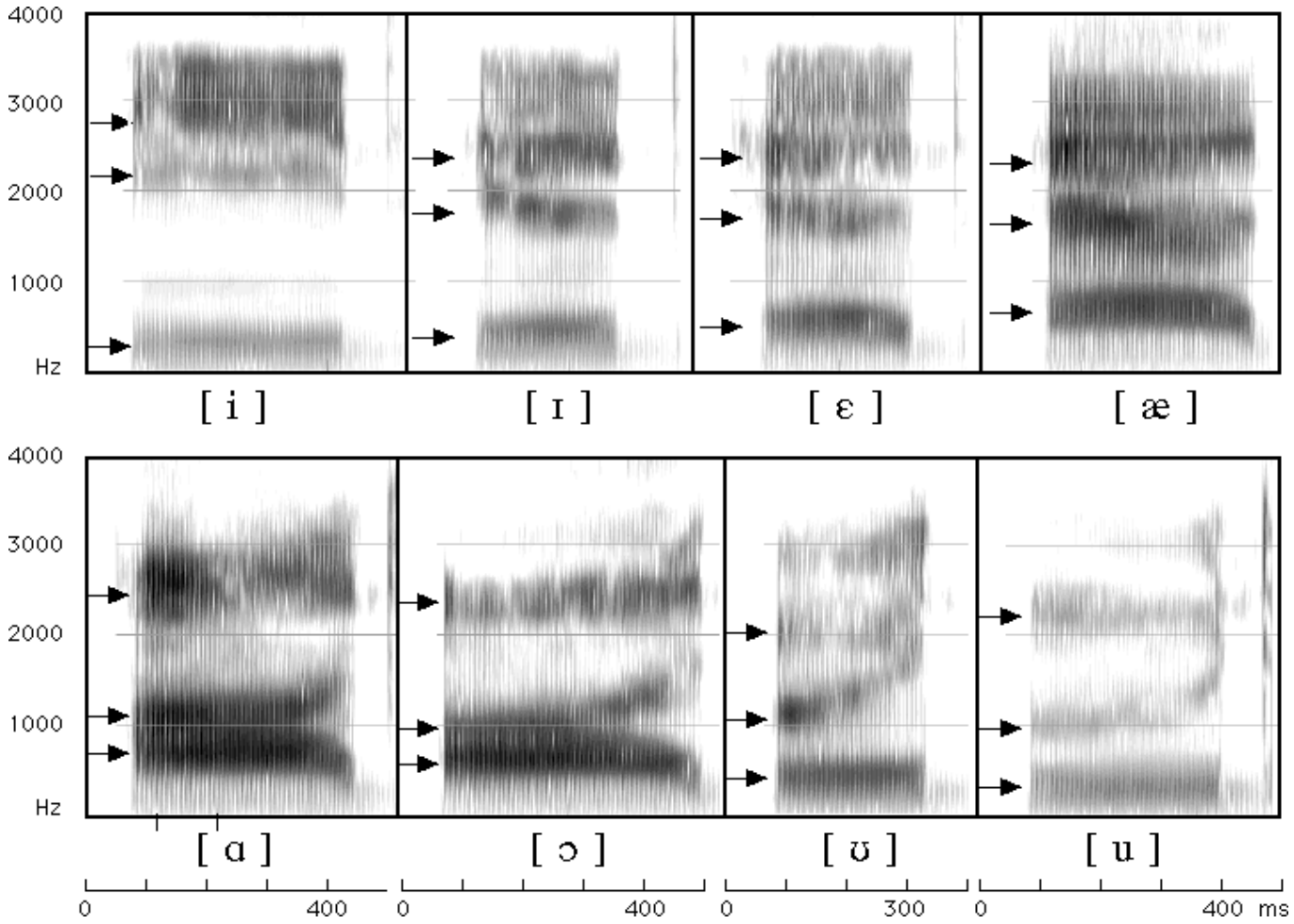
Back to spectrum

- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.



- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

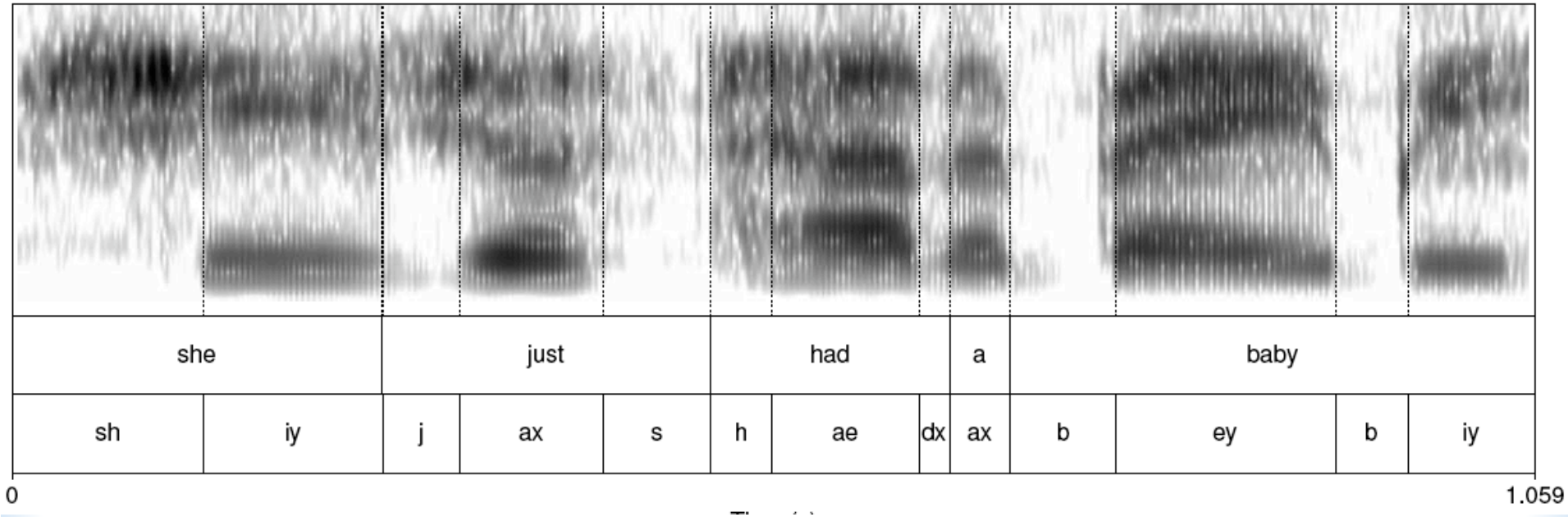
Seeing formants: the spectrogram



Formants

- Vowels largely distinguished by 2 characteristic pitches.
- One of them (the higher of the two) goes downward throughout the series iy ih eh ae aa ao ou u
- The other goes up for the first four vowels and then down for the next four.
- These are called "formants" of the vowels, lower is 1st formant, higher is 2nd formant.

Spectrogram: spectrum + time dimension

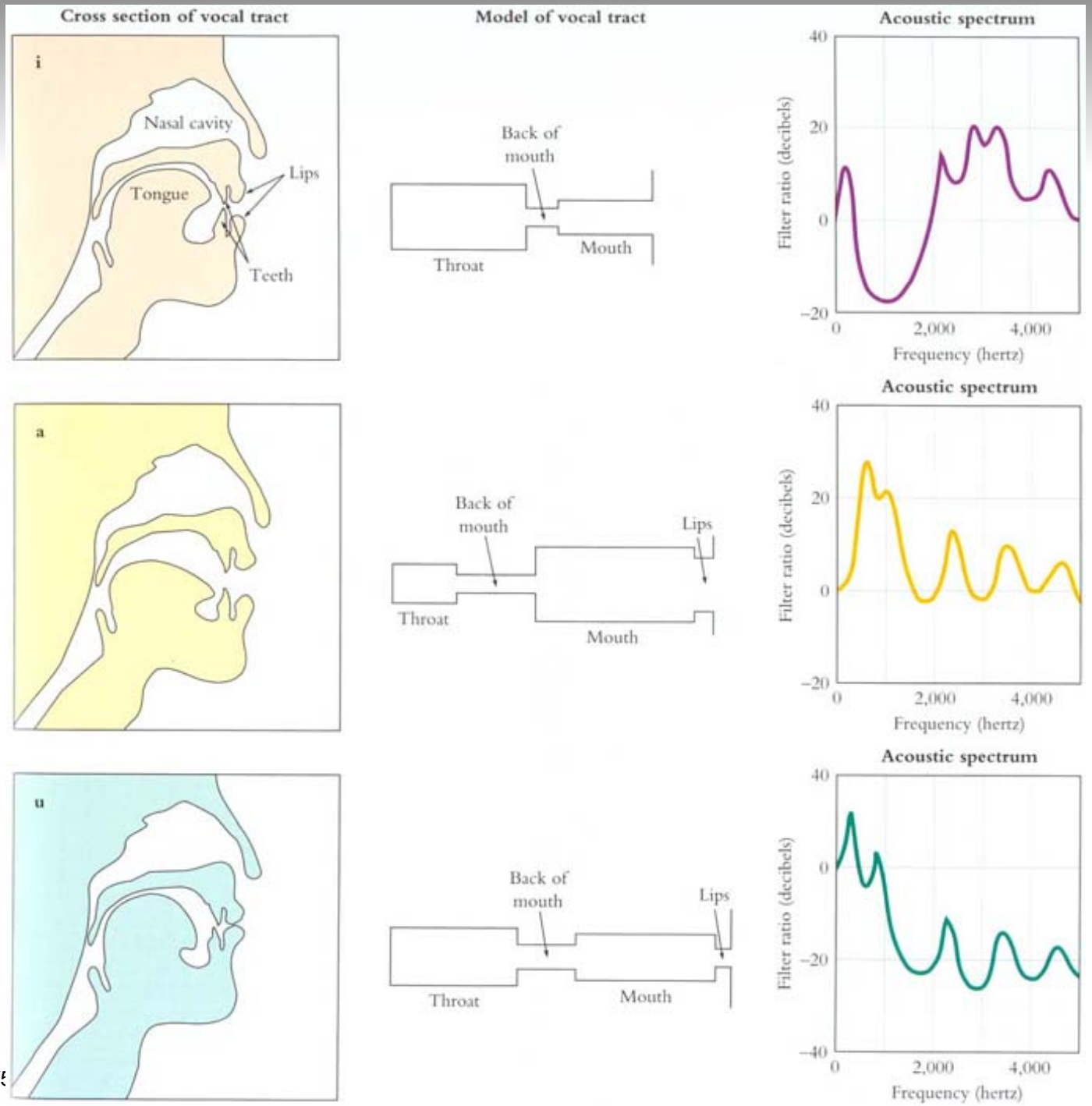


Different vowels have different formants

- Vocal tract as "amplifier"; amplifies different frequencies
- Formants are result of different shapes of vocal tract.
- Any body of air will vibrate in a way that depends on its size and shape.
- Air in vocal tract is set in vibration by action of vocal cords.
- Every time the vocal cords open and close, pulse of air from the lungs, acting like sharp taps on air in vocal tract,
- Setting resonating cavities into vibration so produce a number of different frequencies.

Again: why is a speech sound wave composed of these peaks?

- Articulatory facts:
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others



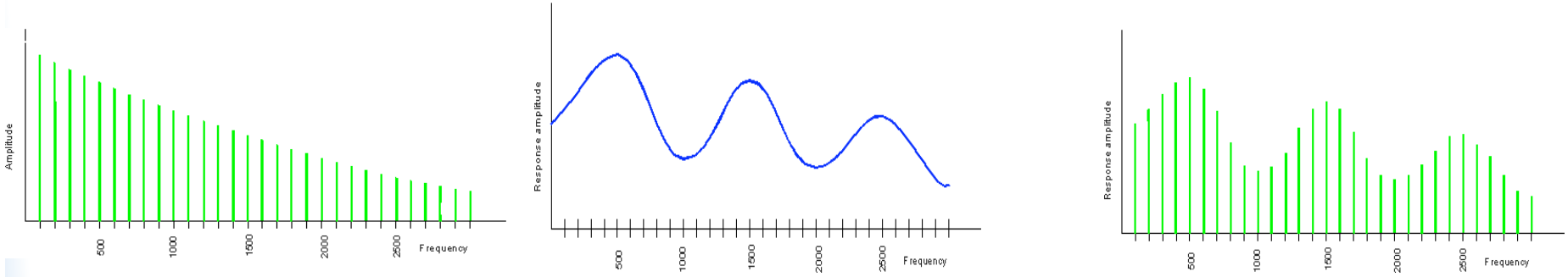
From Mark Liberman's Web site

How formants are produced

- Q: Why do different vowels have different pitches if the vocal cords are vibrating at the same rate?
- A: This is a confusion of frequencies of SOURCE and frequencies of FILTER!

Source-filter model of speech production

Input → Filter → Output



Glottal spectrum

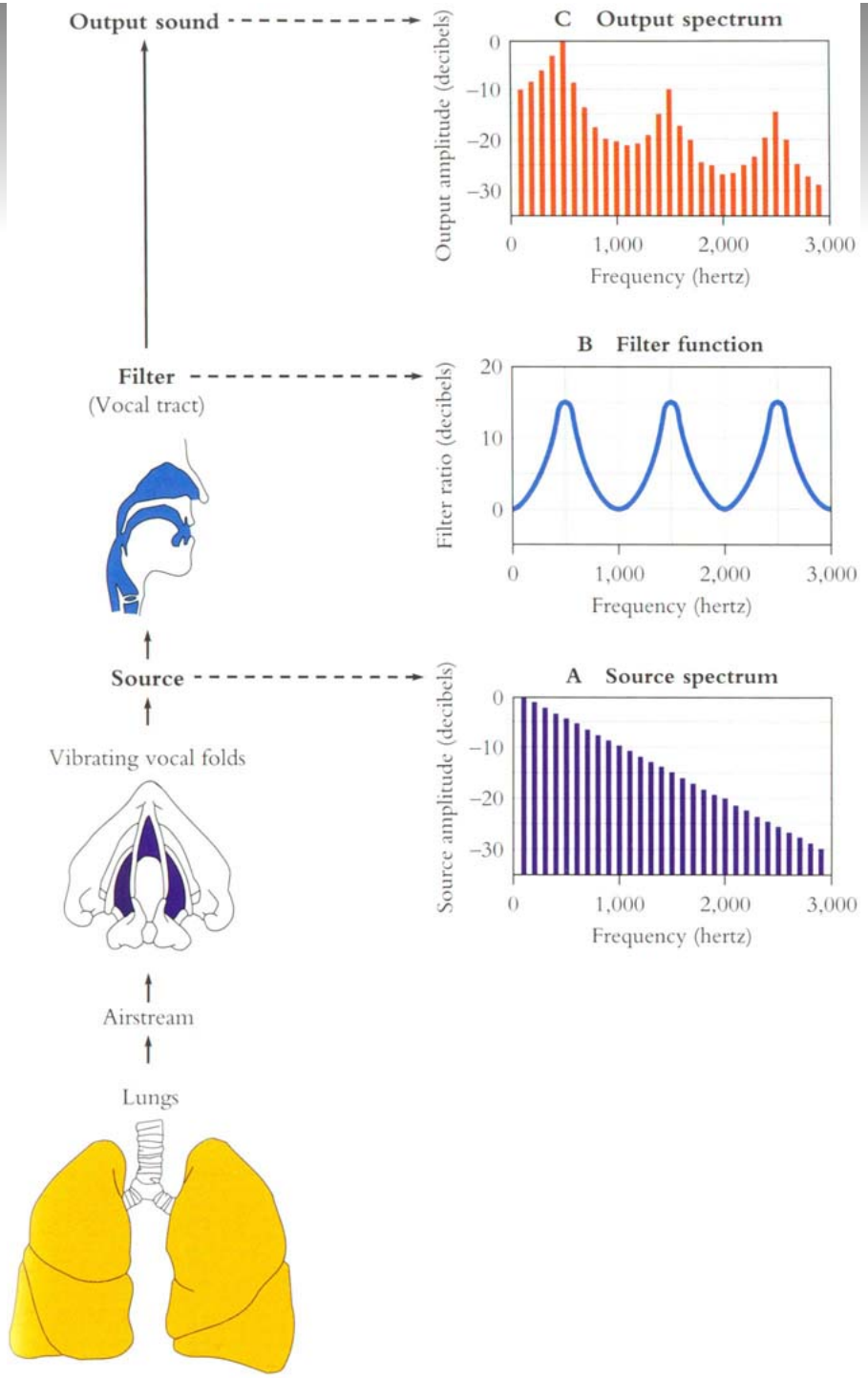
Vocal tract frequency response function

Source and filter are independent, so:

Different vowels can have same pitch

The same vowel can have different pitch

Figures and text from Ratree Wayland slide from his website


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

Deriving schwa: how shape of mouth (filter function) creates peaks!

- Reminder of basic facts about sound waves
- $f = c/\lambda$
- c = speed of sound (approx 35,000 cm/sec)
- A sound with λ=10 meters has low frequency f = 35 Hz (35,000/1000)
- A sound with λ=2 centimeters has high frequency f = 17,500 Hz (35,000/2)

Resonances of the vocal tract

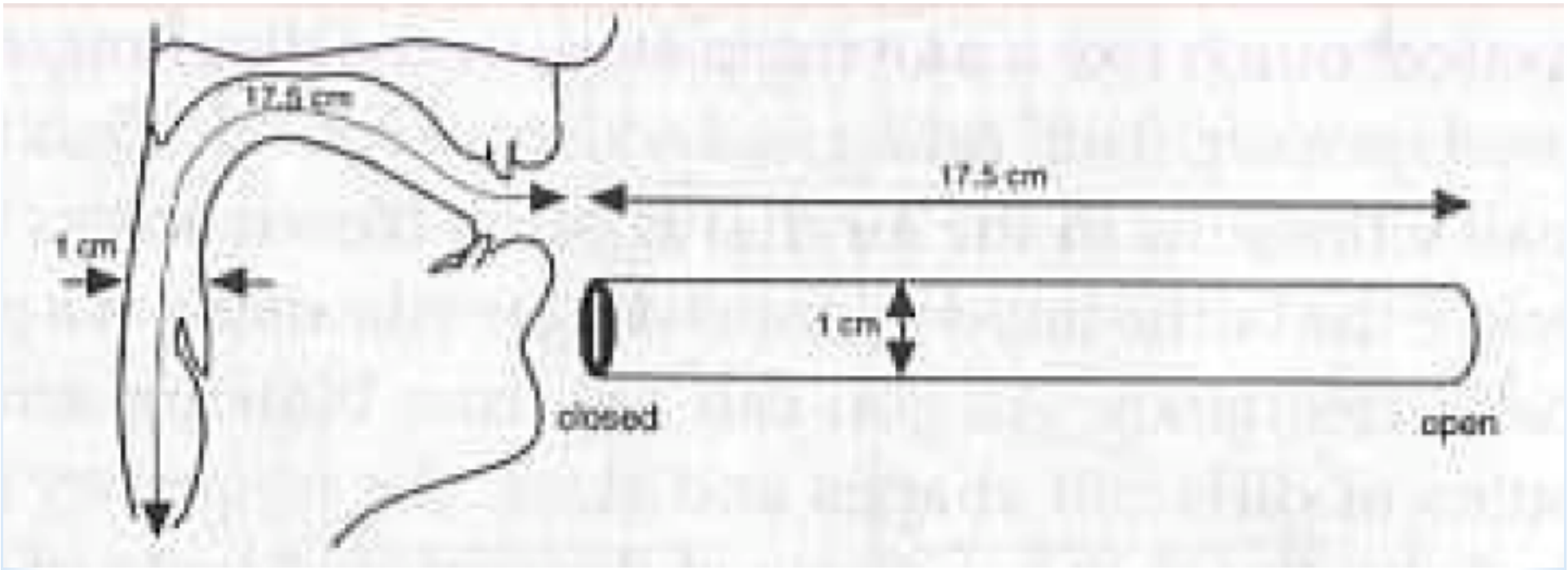
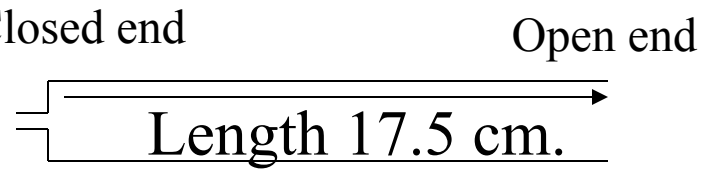


Figure from Ladefoged(1996) p 117

Resonances of the vocal tract

- The human vocal tract as an open tube



- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.

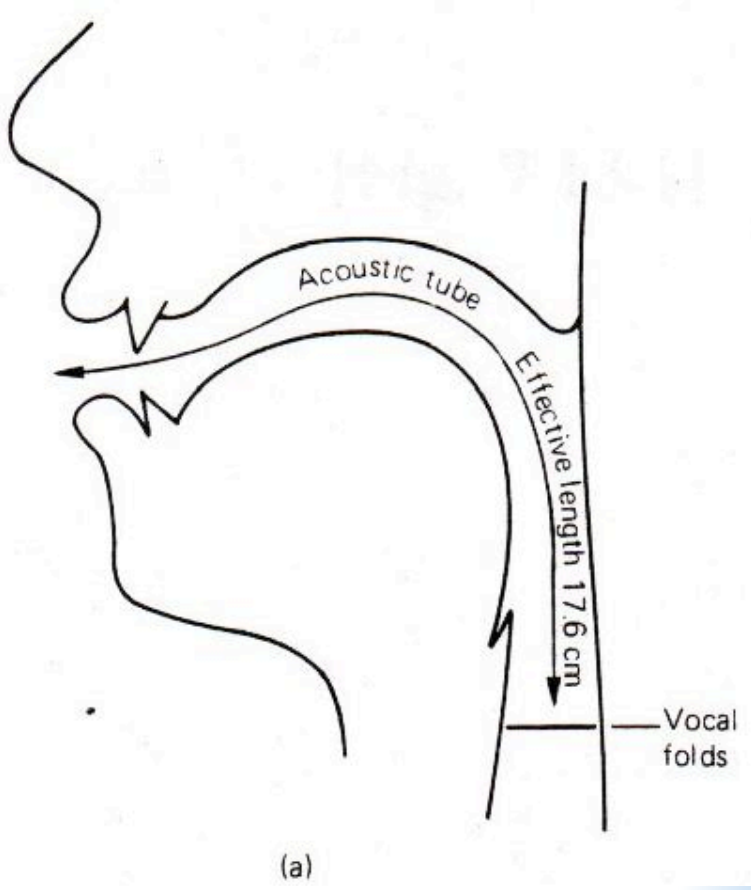
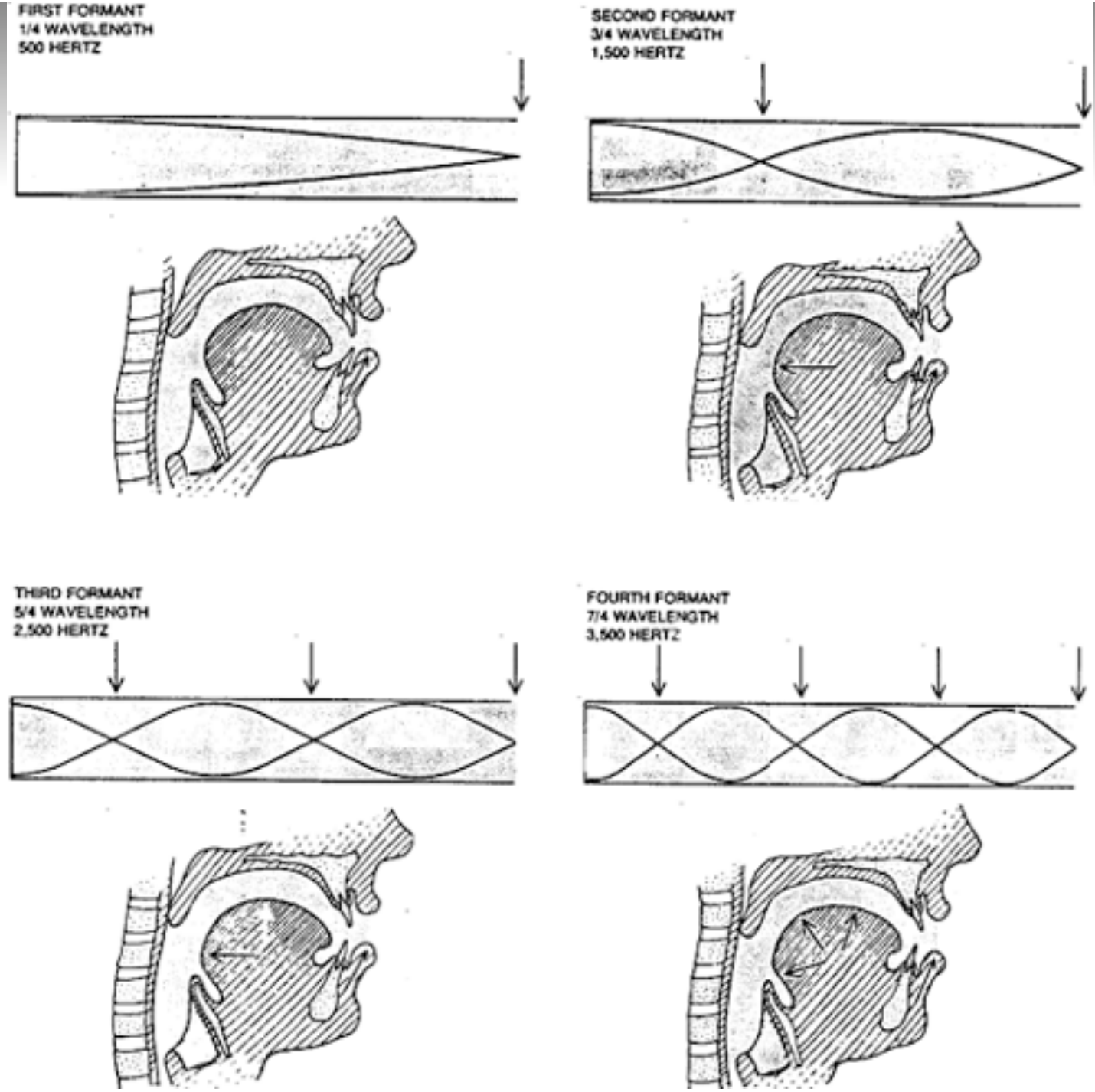


Figure from W. Barry Speech Science slides

Resonances of the vocal tract

- If vocal tract is cylindrical tube open at one end
- Standing waves form in tubes
- Waves will resonate if their wavelength corresponds to dimensions of tube
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.
- Next slide shows what kind of length of waves can fit into a tube with this contraint

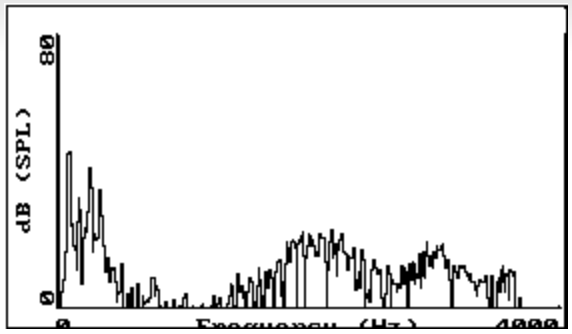


From Sundberg

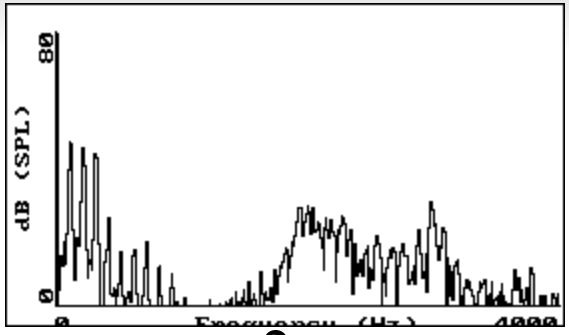
Computing the 3 formants of schwa

- Let the length of the tube be L
- $F_1 = c/\lambda_1 = c/(4L) = 35{,}000/4*17.5 =$ 500Hz
- $F_2 = c/\lambda_2 = c/(4/3L) = 3c/4L = 3*35{,}000/4*17.5$ = 1500Hz
- $F_1 = c/\lambda_2 = c/(4/5L) = 5c/4L = 5*35{,}000/4*17.5$ = 2500Hz
- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz
- These vowel resonances are called formants

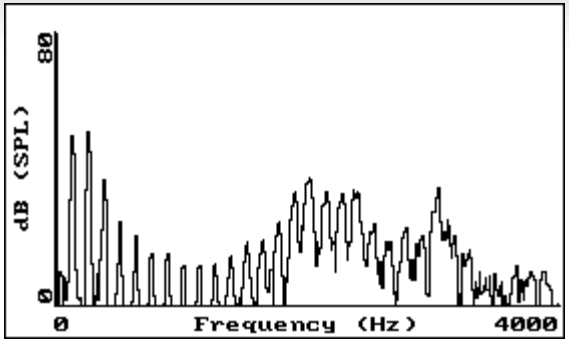
Vowel [i] sung at successively higher pitch.



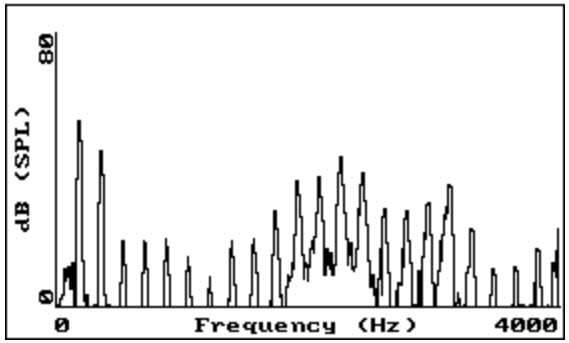
1



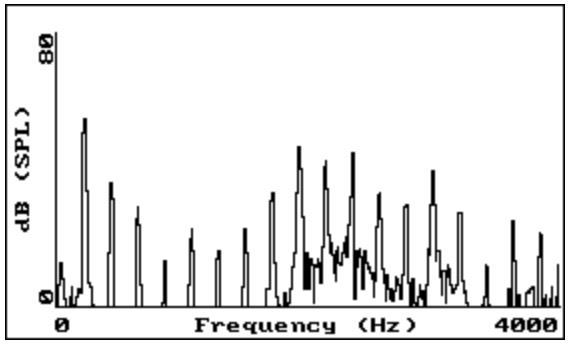
2



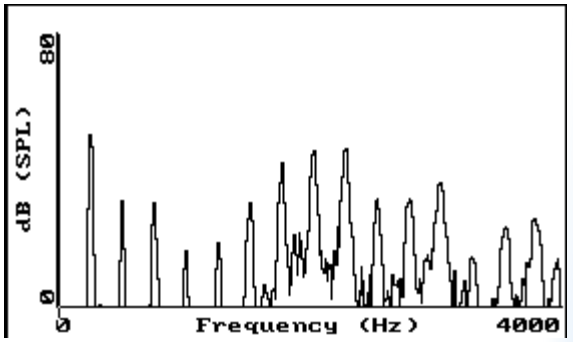
3



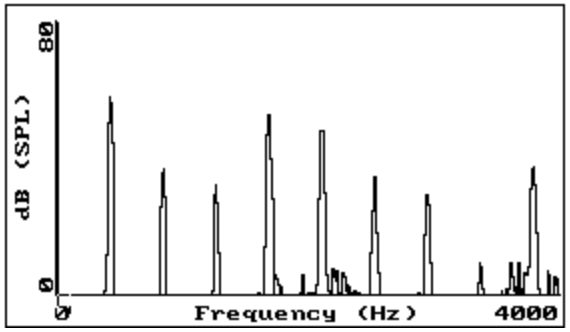
4



5



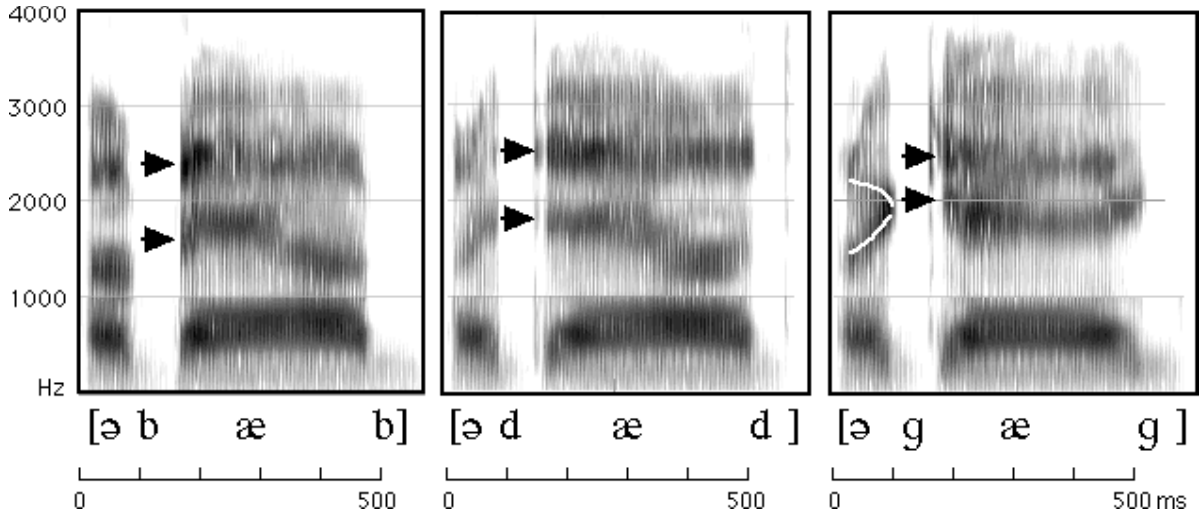
6



7

Figures from Ratree Wayland slides from his website

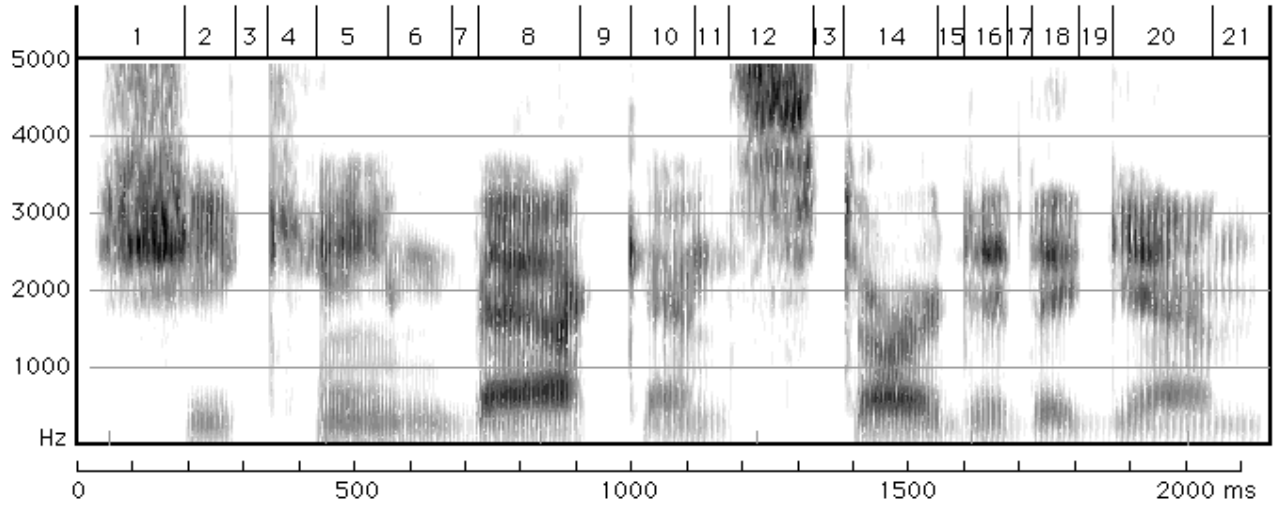
How to read spectrograms



- bab**: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"**
- dad**: first formant increases, but F2 and F3 slight fall**
- gag**: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials**

From Ladefoged "A Course in Phonetics"

She came back and started again



1. lots of high-freq energy
3. closure for k
4. burst of aspiration for k
5. ey vowel;faint 1100 Hz formant is nasalization
6. bilabial nasal
7. short b closure, voicing barely visible.
8. ae; note upward transitions after bilabial stop at beginning
9. note F2 and F3 coming together for "k"

From Ladefoged "A Course in Phonetics"

Homework 1

- http://www.stanford.edu/class/cs224s/hw1.html
- You'll need to download PRAAT; details are in the homework.

Phonetic Resources

- Phonetic dictionaries
  - CMU dict
  - CELEX
- Phonetically transcribed corpora
  - TIMIT
  - Switchboard

TIMIT

- Read speech corpus, time aligned

she	had	your	dark	suit	in
sh iy	hv ae dcl	jh axr	dcl d aa r kcl	s ux q	en

x q	in	greasy	wash	water
	en	gcl g r iy s ix	w aa sh	q w aa dx axr q

Switchboard

- Spontaneous speech corpus
- Telephone conversations between strangers
- "They're kind of in between right now"
-

0.470 dh er	0.640 k aa	0.720 n ax	0.900 v ih m	0.953 b ix	

53	1.279 t w iy n	1.410 r ay	1.630 n aw

Summary

- Acoustic Phonetics
  - Waves, sound waves, and spectra
  - Speech waveforms
  - F0, pitch, intensity
  - Spectra
    - Spectrograms
    - Formants
    - Reading spectrograms
  - Deriving schwa: why are formants where they are
  - PRAAT
  - Resources: dictionaries and phonetically-labeled corpora.

Examples from Ladefoged

pad

bad

spat