

# CS 224S / LINGUIST 281 Speech Recognition and Synthesis

Dan Jurafsky

## Lecture 2: Acoustic Phonetics

1/17/06

CS 224S Winter 2006

1

## Today, Jan 12, Week 1

- **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

1/17/06

CS 224S Winter 2006

2

## Acoustic Phonetics

- **Sound Waves**

- <http://www.kettering.edu/~drussell/Demos/waves-intro/waves-intro.html>

1/17/06

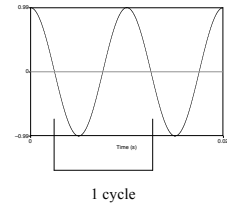
CS 224S Winter 2006

3

## Simple Period Waves (sine waves)

- Characterized by:

- period: T
- amplitude A
- phase  $\phi$
- Fundamental frequency in cycles per second, or Hz
  - $F_0 = 1/T$

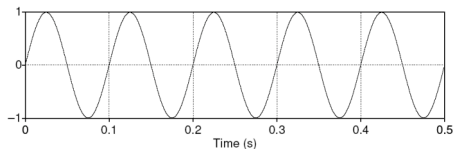


1/17/06

CS 224S Winter 2006

4

## Simple periodic waves



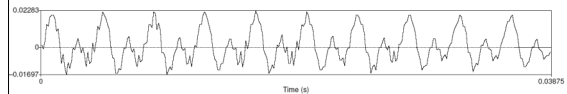
- Computing the frequency of a wave:
  - 5 cycles in .5 seconds = 10 cycles/second = 10 Hz
- Amplitude:
  - 1
- Equation:
  - $Y = A \sin(2\pi ft)$

1/17/06

CS 224S Winter 2006

5

## Speech sound waves



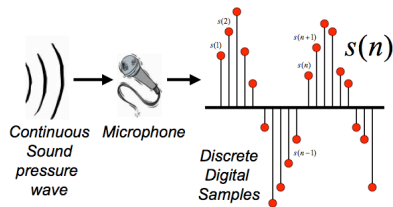
- A little piece from the waveform of the vowel [iy]
- Y axis:
  - Amplitude = amount of air pressure at that time point
    - Positive is compression
    - Zero is normal air pressure,
    - negative is rarefaction
- X axis: time.

1/17/06

CS 224S Winter 2006

6

## Digitizing Speech



Thanks to Bryan Pellow for this slide!

1/17/06

CS 224S Winter 2006

7

## Digitizing Speech

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

1/17/06

CS 224S Winter 2006

8

## Sampling

- Measuring amplitude of signal at time  $t$
- The sampling rate needs to have at least two samples for each cycle
  - Roughly speaking, one for the positive and one for the negative half of each cycle.
  - More than two sample 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**

1/17/06

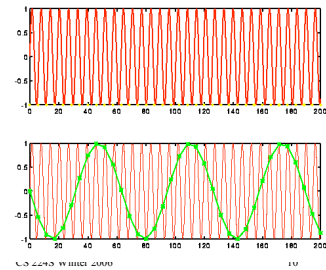
CS 224S Winter 2006

9

## Sampling

Original signal in red:

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



1/17/06

## Sampling

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

1/17/06

CS 224S Winter 2006

11

## 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
- Headers:
  - Raw (no header)
  - Microsoft wav
  - Sun .au

40 byte header

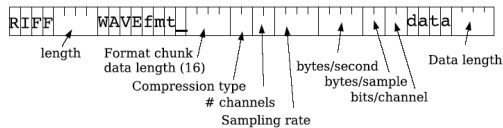


1/17/06

CS 224S Winter 2006

12

## WAV format



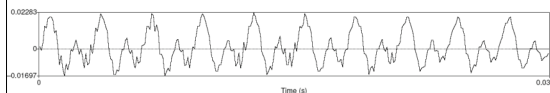
1/17/06

CS 224S Winter 2006

13

## Fundamental frequency

- Waveform of the vowel [iy]



- Frequency: repetitions/second of a wave
- 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**

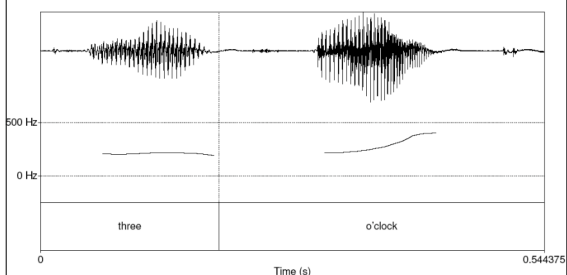
1/17/06

CS 224S Winter 2006

14



## Pitch track



1/17/06

CS 224S Winter 2006

15

## Amplitude

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

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

1/17/06

CS 224S Winter 2006

16

## 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.  $P_0$  is auditory threshold pressure =  $2 \times 10^{-5}$  pa

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

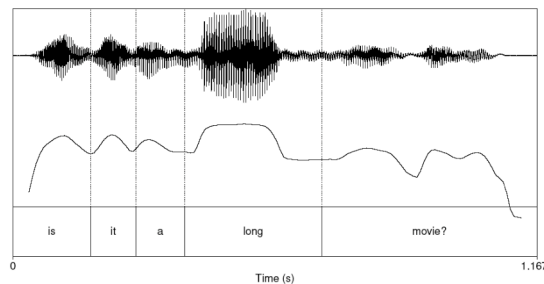
1/17/06

CS 224S Winter 2006

17



## Plot of Intensity



1/17/06

CS 224S Winter 2006

18

## 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 =  $1127 \ln(1 + f/700)$

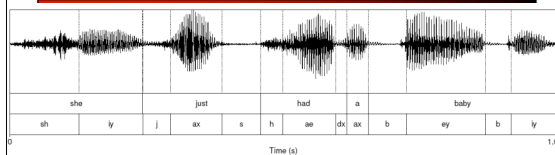
1/17/06

CS 224S Winter 2006

19



## She just had a baby



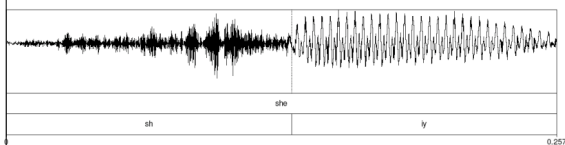
- Note that vowels all have regular amplitude peaks
- Stop consonant
  - Closure followed by release
  - Notice the silence followed by slight bursts of emphasis: very clear for [b] of "baby"
- Fricative: noisy. [sh] of "she" at beginning

1/17/06

CS 224S Winter 2006

20

## Fricative

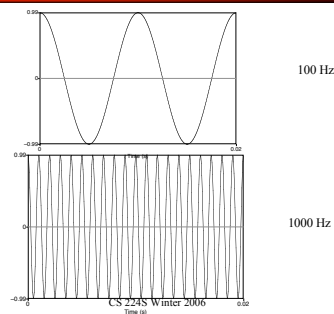


1/17/06

CS 224S Winter 2006

21

## Waves have different frequencies

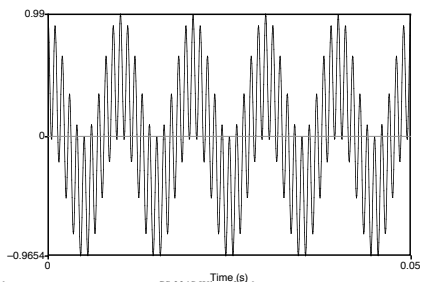


1/17/06

CS 224S Winter 2006

22

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



1/17/06

CS 224S Winter 2006

23

## Spectrum

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



1/17/06

CS 224S Winter 2006

24

## Spectra continued

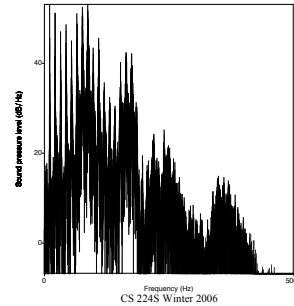
- **Fourier analysis: any wave can be represented as the (infinite) sum of sine waves of different frequencies (amplitude, phase)**

1/17/06

CS 224S Winter 2006

25

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

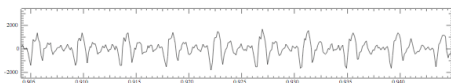


1/17/06

CS 224S Winter 2006

26

## Part of [æ] 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

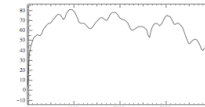
1/17/06

CS 224S Winter 2006

27

## Back to spectrum

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



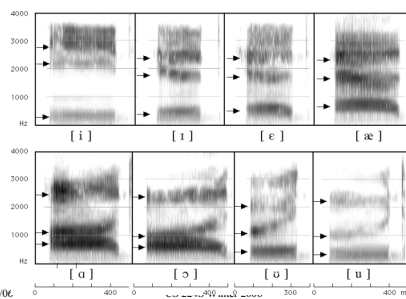
- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

1/17/06

CS 224S Winter 2006

28

## Seeing formants: the spectrogram



1/17/06

29

## Formants

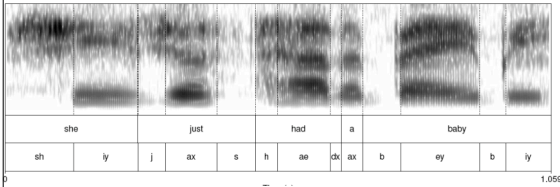
- Vowels largely distinguished by 2 characteristic pitches.
- One of them (the higher of the two) goes downward throughout the series iy ih eh ae aa ao u
- The other goes up for the first four vowels and then down for the next four.
- These are called "formants" of the vowels, lower is 1st formant, higher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

1/17/06

CS 224S Winter 2006

30

## Spectrogram: spectrum + time dimension



1/17/06

CS 224S Winter 2006

31

## 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.

1/17/06

CS 224S Winter 2006

32

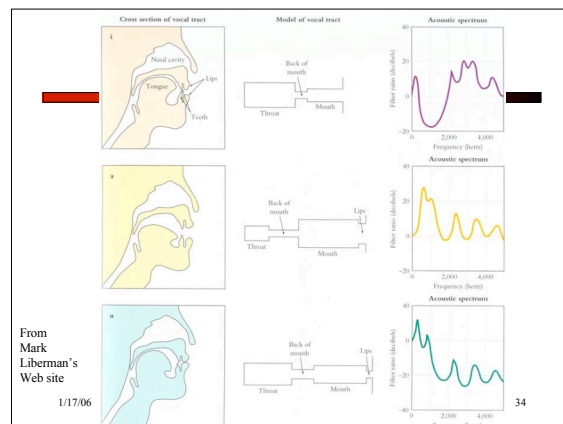
## Again: why is a speech sound wave composed of these peaks?

- 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

1/17/06

CS 224S Winter 2006

33



From Mark Liberman's Web site

1/17/06

34

## How formants are produced

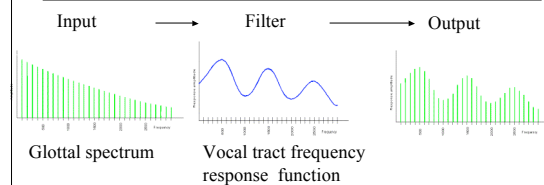
- **Q:** Why do vowels have different pitches if the vocal cords are same rate?
- **A:** This is a confusion of frequencies of **SOURCE** and frequencies of **FILTER**!

1/17/06

CS 224S Winter 2006

35

## Source-filter model of speech production



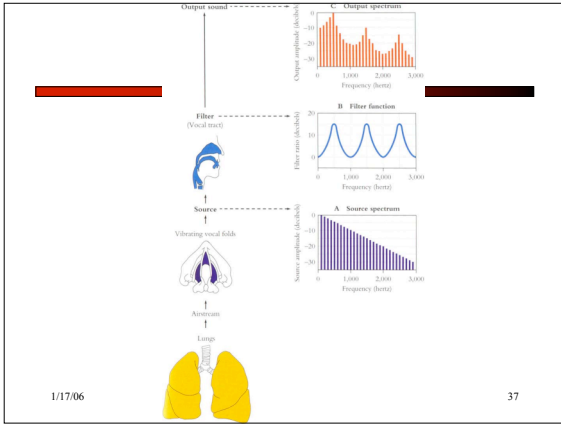
Source and filter are independent, so:  
 Different vowels can have same pitch  
 The same vowel can have different pitch

1/17/06

CS 224S Winter 2006

36

Figures and text from Ratrete Wayland slide from his website



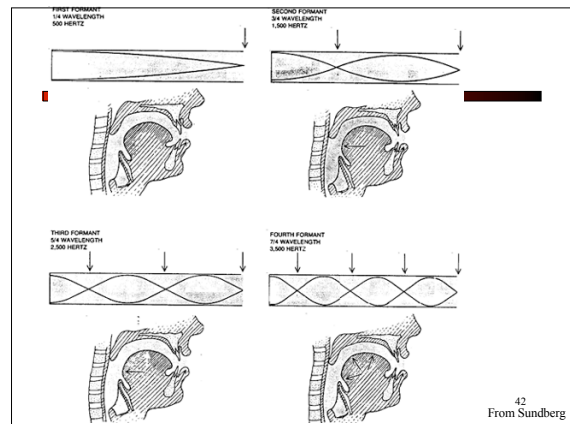
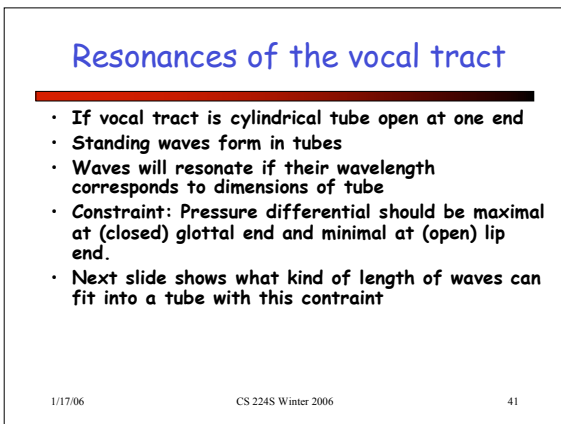
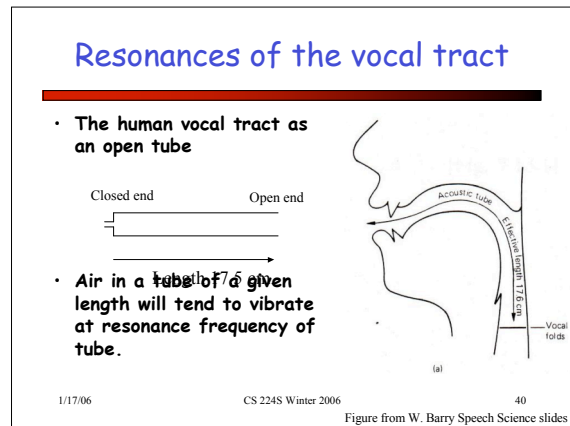
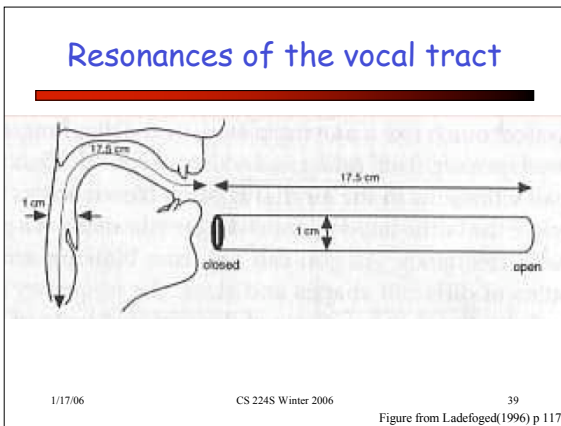
## Deriving schwa: how shape of mouth (filter function) creates peaks!

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

1/17/06

CS 224S Winter 2006

38



## Computing the 3 formants of schwa

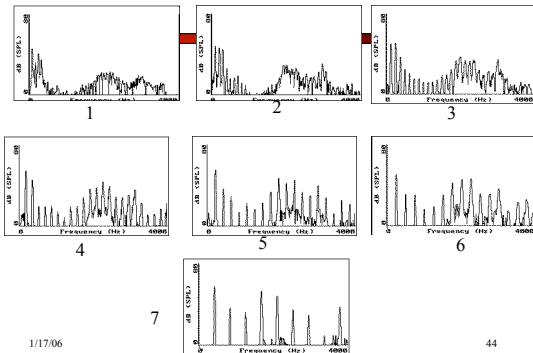
- Let the length of the tube be  $L$
- $F_1 = c/\lambda_1 = c/(4L) = 35,000/4*17.5 = 500\text{Hz}$
- $F_2 = c/\lambda_2 = c/(4/3L) = 3c/4L = 3*35,000/4*17.5 = 1500\text{Hz}$
- $F_3 = c/\lambda_3 = c/(4/5L) = 5c/4L = 5*35,000/4*17.5 = 2500\text{Hz}$
- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz
- These vowel resonances are called **formants**

1/17/06

CS 224S Winter 2006

43

Vowel [i] sung at successively higher pitch.

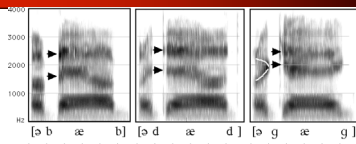


1/17/06

44

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

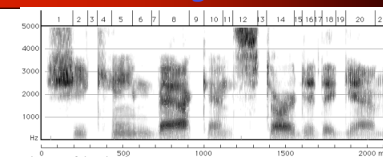
1/17/06

CS 224S Winter 2006

45

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"

1/17/06

CS 224S Winter 2006

46

From Ladefoged "A Course in Phonetics"

## Homework 1

- <http://www.stanford.edu/class/linguist236/homework1.html>
- You'll need to download PRAAT; details are in the homework.

1/17/06

CS 224S Winter 2006

47

## Phonetic Resources

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

1/17/06

CS 224S Winter 2006

48

## TIMIT

- Read speech corpus, time aligned

|       |            |         |                |        |
|-------|------------|---------|----------------|--------|
| she   | had        | your    | dark           | suit   |
| sh iy | h v ae dcl | j h axr | dcl d aa r kcl | s ux q |

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

1/17/06

CS 224S Winter 2006

49

## Switchboard

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

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

|          |       |       |
|----------|-------|-------|
| 1.279    | 1.410 | 1.630 |
| t w iy n | r ay  | n aw  |

50

## Summary

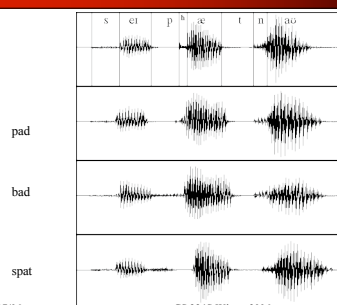
- Acoustic Phonetics
  - Waves, sound waves, and spectra
  - Speech waveforms
  - F0, pitch, intensity
  - Spectra
    - Spectrograms
    - Formants
    - Reading spectrograms
  - Deriving schwa: why are formants where they are
  - PRAAT
  - Resources: dictionaries and phonetically-labeled corpora.

1/17/06

CS 224S Winter 2006

51

## Examples from Ladefoged



1/17/06

CS 224S Winter 2006

52