



# Speech Processing 15-492/18-492

---

Computer Speech

# Analog to Digital

- ◆ *Speech (sound) is analog*
  - *Computers are digital*
    - ⊗ *We need to convert*
- ◆ *Sample from A-D converter*
  - *N times a second*
- ◆ *How many times a second?*

# Sample Frequency

## ◆ *Speech*

- *F0 (intonation contour) 80-300Hz*
- *F1/F2 250-3000Hz*
- *Fricatives, higher maybe 4KHz-8KHz*

## ◆ *We can hear higher frequencies*

- *Up to 20KHz (maybe)*

# What can you hear?

10Hz 100Hz 500Hz 1000Hz 2000Hz



4KHz 8KHz 10KHz 12KHz 14KHz



16KHz 18Khz 20KHz



# Human frequency perception

- ◆ *Highest perception 20Khz*
- ◆ *But it degrades with age.*
  - *The older you are the less high frequencies*
- ◆ *Starts degrading as late teenager!*
  
- ◆ *But is it important?*

# Sampling Frequency

- ◆ *How many samples a second*
  - *To capture an 8KHz signal?*
  - *To capture a 16KHz signal?*
- ◆ *At least 2 times the signal*
  - *Nyquist frequency (half the sample rate)*
- ◆ *So why is CD sampling rate 44.1KHz?*

# Human Speech

## ◆ *Human speech and sampling frequencies*

*32000Hz*



*22500Hz*



*16000Hz*



*11250Hz*



*8000Hz*



*6000Hz*



*4000Hz*



*2000Hz*

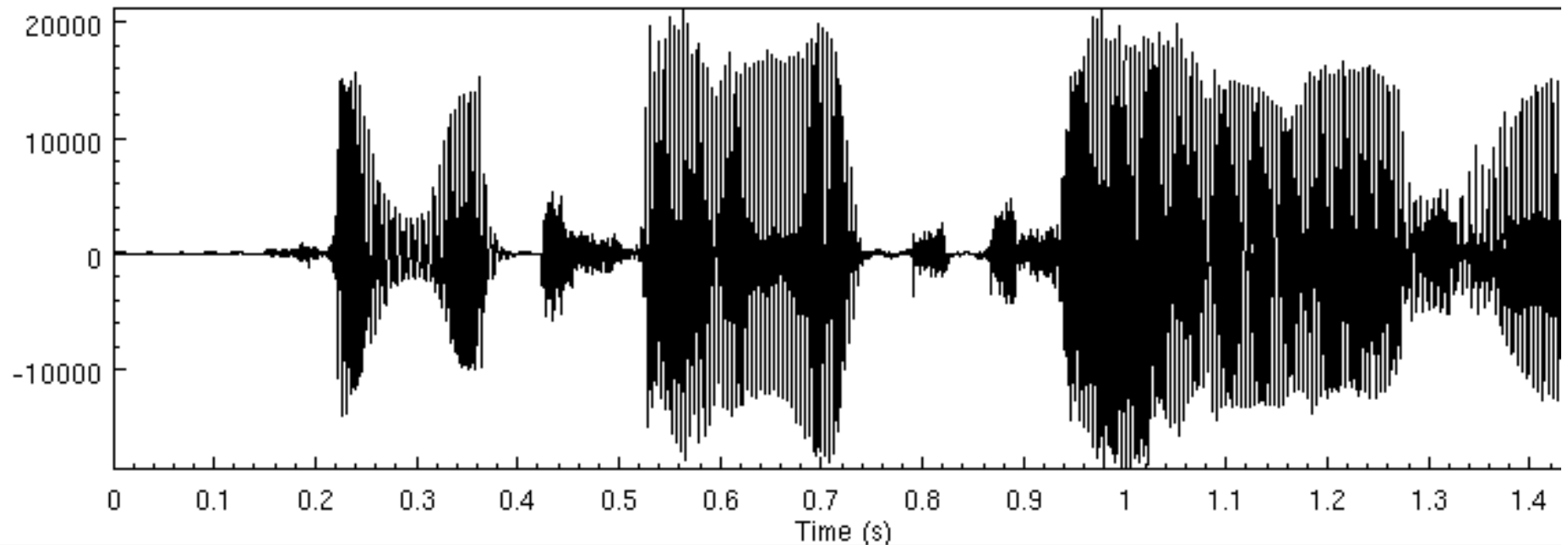


*1000Hz*

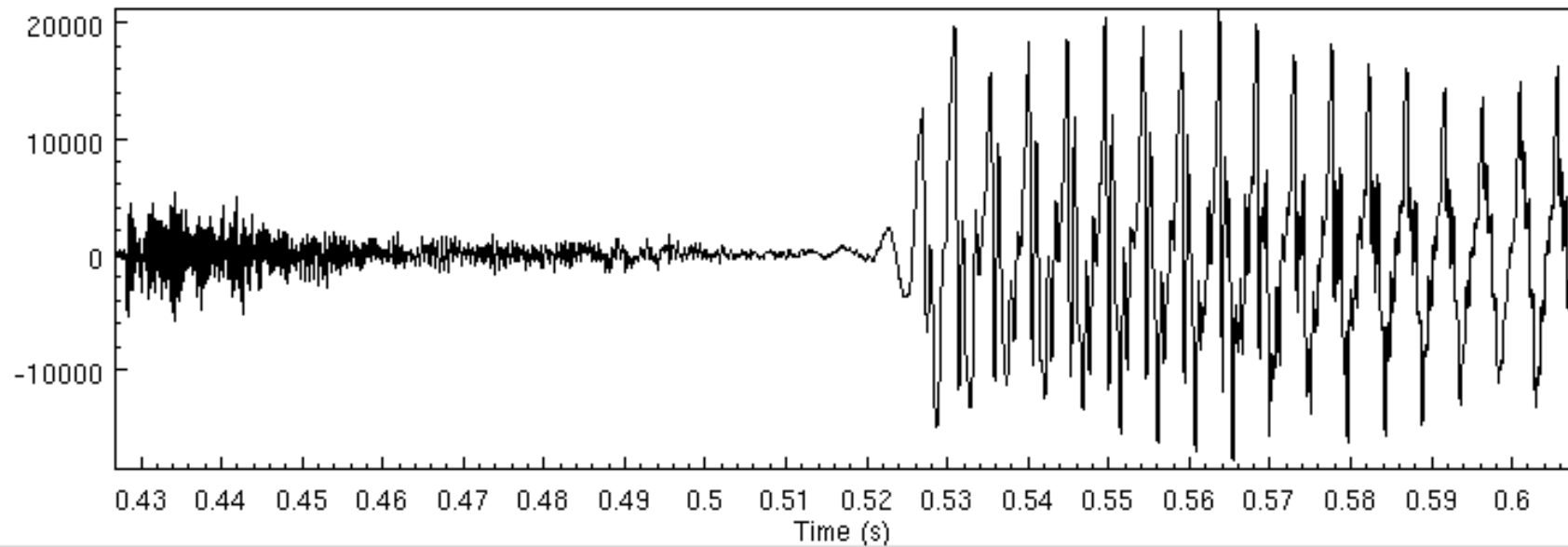


# Waveform Representation

- Sample magnitude at N Hz



# Waveform Representation



# Waveform Encoding

- ◆ *PCM (Pulse code modulation)*
  - *Simple +/-32768*
- ◆ *But human hearing is logarithmic*
  - *Changes are smaller amplitudes more important than changes at higher amplitudes*
  - *mulaw (alaw) encodings*
- ◆ *Human speech conventions*
  - *Wide band speech 16KHz*
  - *Narrow band speech 8KHz (telephone speech)*

# Speech Compression

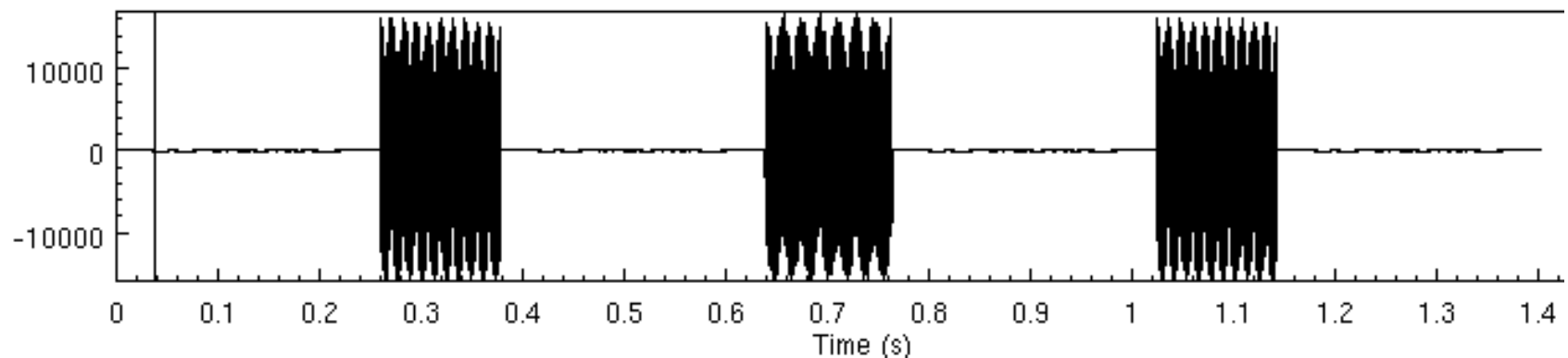
- ◆ *Bandwidth is money (or time)*
- ◆ *Telephone Speech*
  - *64KBs (8KHz/8bit ulaw/alaw)*
- ◆ *Wide band:*
  - *256KBz (16KHz/16bit)*
- ◆ *CDs*
  - *1.4MBs (44.1KHz 16bit stereo)*
- ◆ *Mp3s (music)*
  - *128KBs (expands to 44.1KHz stereo)*
- ◆ *Cell phone*
  - *9.8KBs (or even 4.8KBs)*

# Time vs Frequency Domain

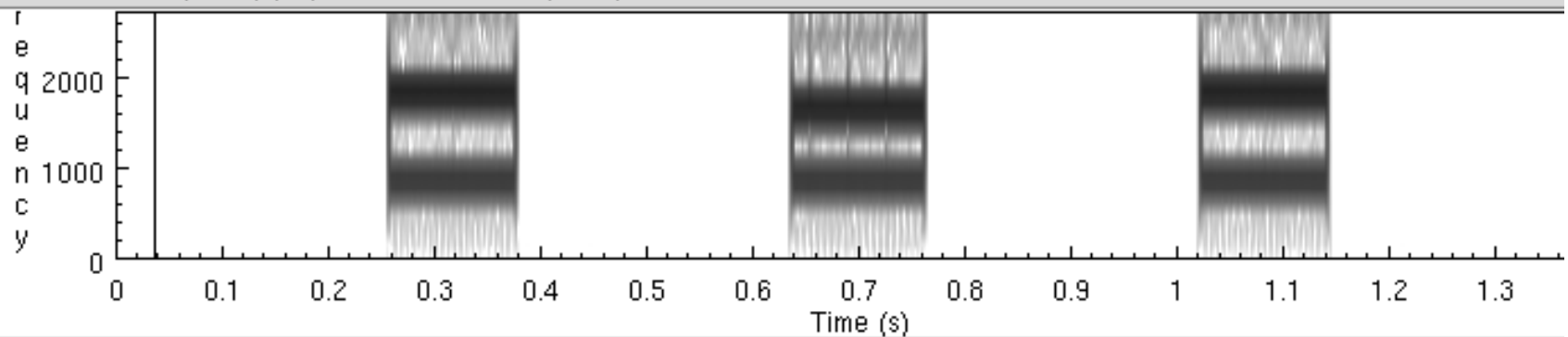
- ◆ *All signals can be constructed*
  - *From sum of sine waves*
- ◆ *We can convert any signal into a set of sine waves*
- ◆ *Fourier Transform*
  - *Conversion of time signal to frequency spectrum*
- ◆ *Fast Fourier Transform*
  - *An efficient computer algorithm to do it*

# Spectrogram vs Time domain

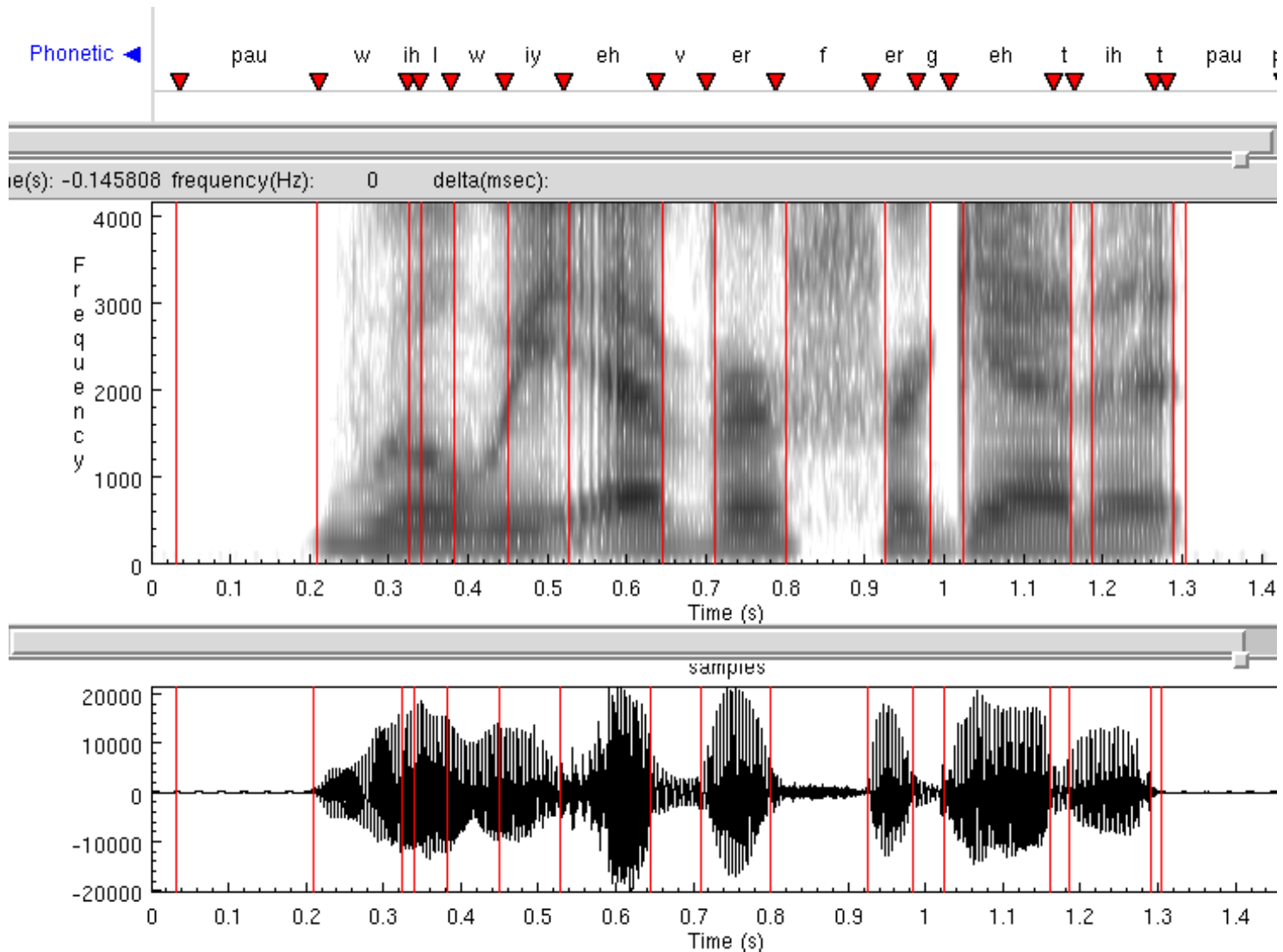
- Three telephone tones



Time: 0.0377524 frequency(Hz): 0 delta(msec):

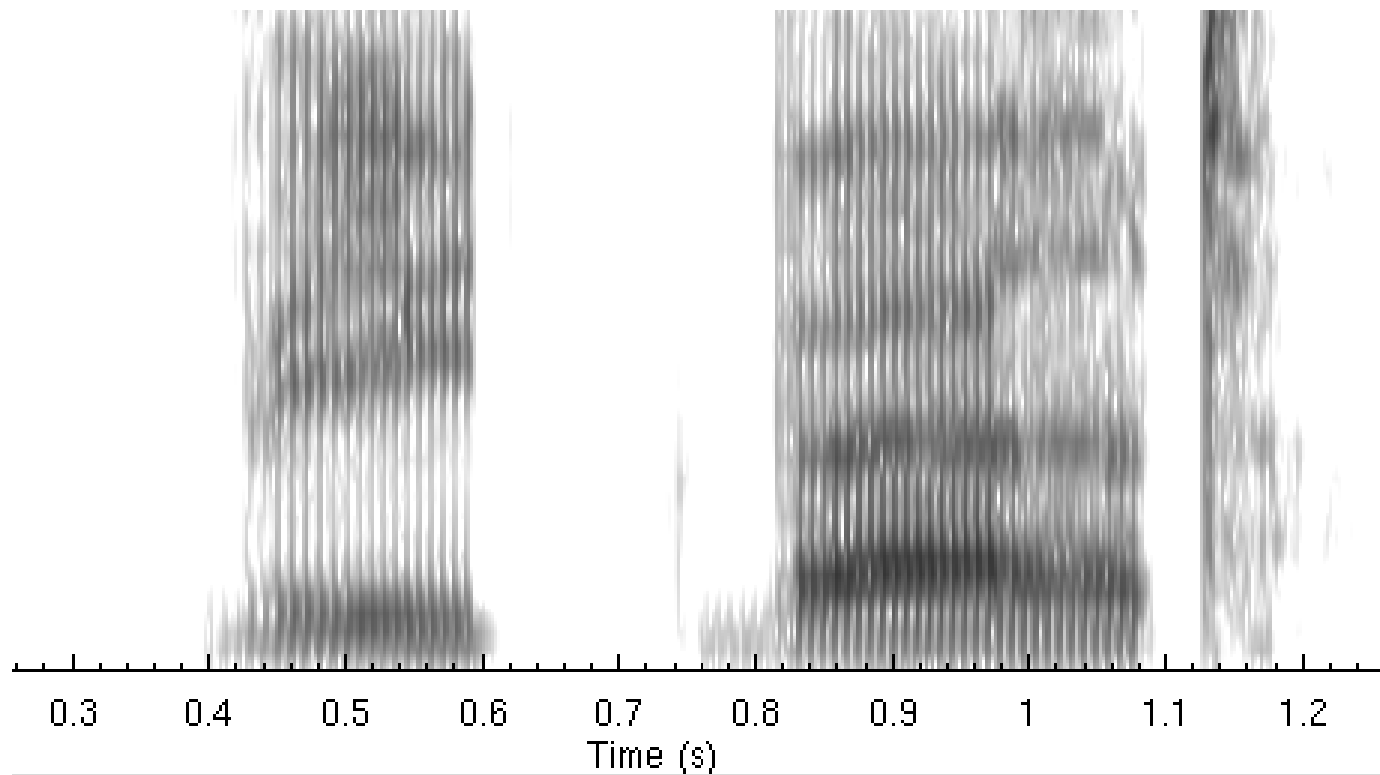


# Speech Spectrogram



# /iy/ vs /ae/

- “beat” /b iy t/ and “bat” /b ae t/



# Microphones

- ◆ *Head mounted microphone:*
  - *Close–talking, noise cancelling*
- ◆ *Far field microphone*
  - *Speaker will move giving different acoustics*
- ◆ *Array microphone*
  - *“follows” where speaker is*

# Background noise

- ◆ *Quiet offices*
  - *Consistent “white” noise (computer fan/AC)*
- ◆ *Outside*
  - *Wind, traffic*
- ◆ *Human babble*
  - *Hardest time of noise to deal with*



# Summary

## ◆ *Computer speech*

- *Digitized by sampling 8KHz to 44KHz*
- *Telephone speech is 8KHz*
- *Wide band is 16KHz (or more)*

## ◆ *Time vs Frequency domain*

- *More distinctions in the frequency domain*
- *FFT to convert to frequency from time*
- *Easier to “see” difference in speech*