



Speech Processing 15-492/18-492

Speech Recognition
Signal Processing

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?*

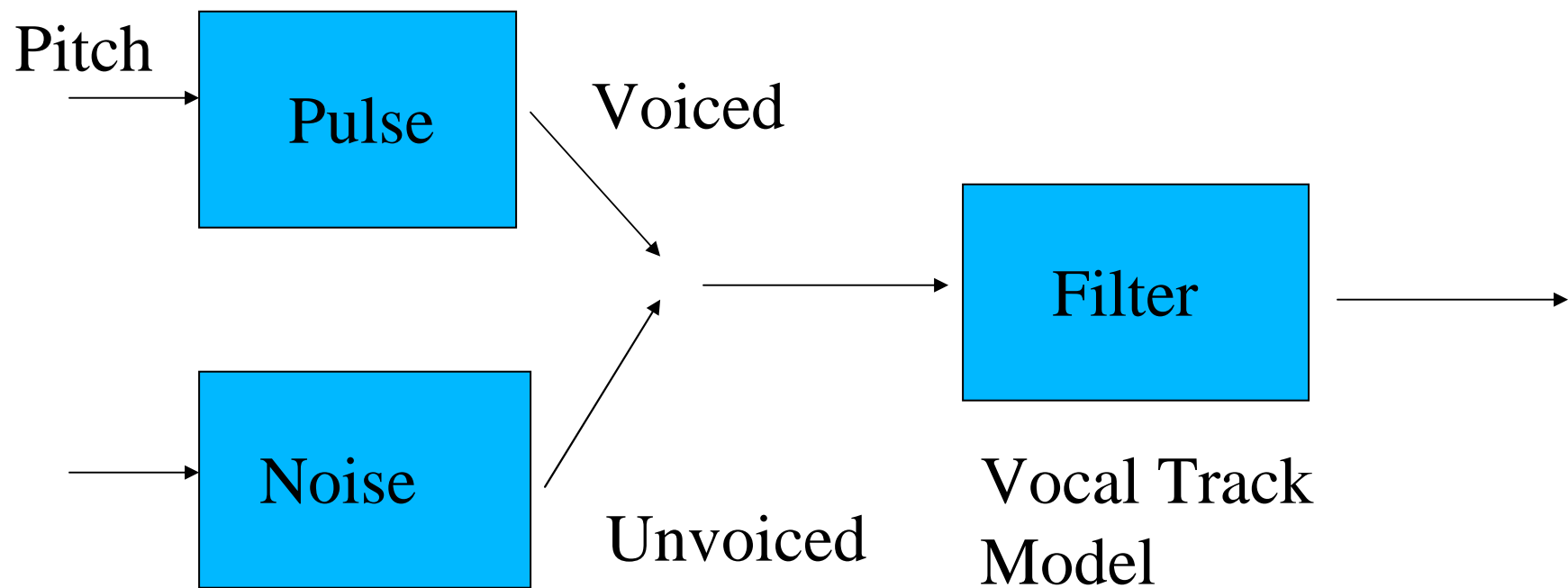
Goals of Signal Processing

- ◆ *Distinguish between phonetic types*
- ◆ *Be invariant to channel/room conditions*
- ◆ *Be invariant to speaker characteristics*
- ◆ *Computational efficiency*

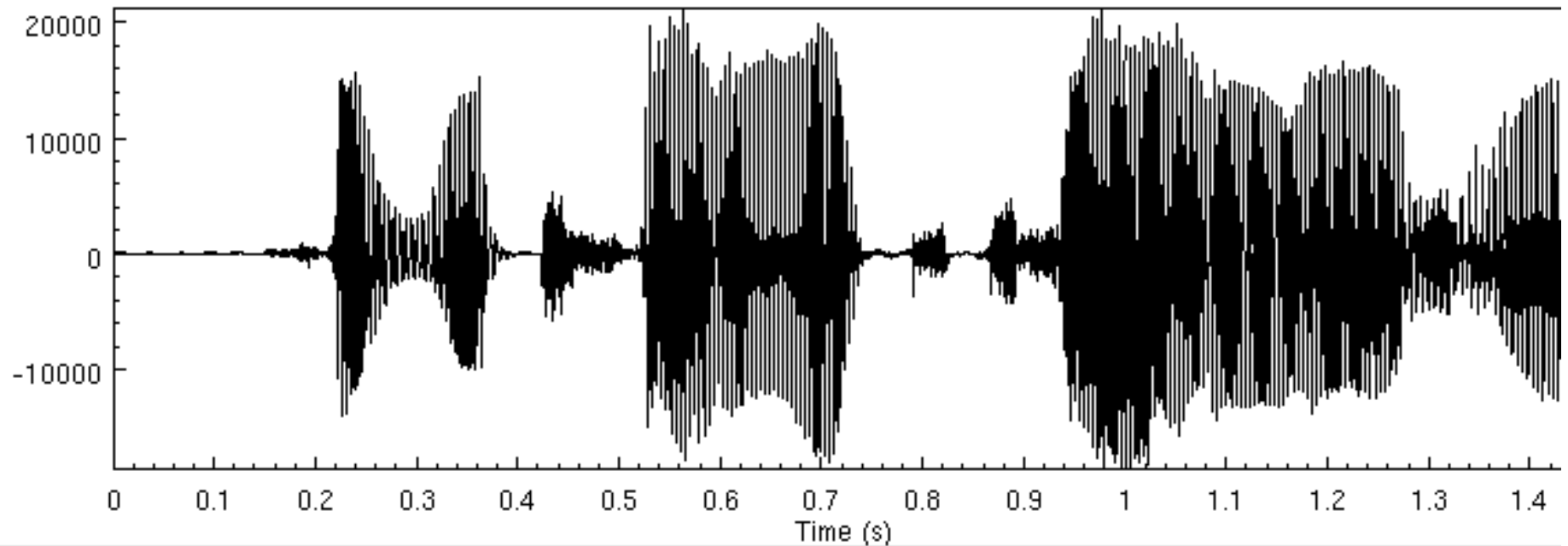
Time vs Frequency Domain

- ◆ *Human ear distinguishes frequencies*
- ◆ *Initial ASR used time domain features*
 - *Power*
 - *Zero crossings (sort of frequency)*

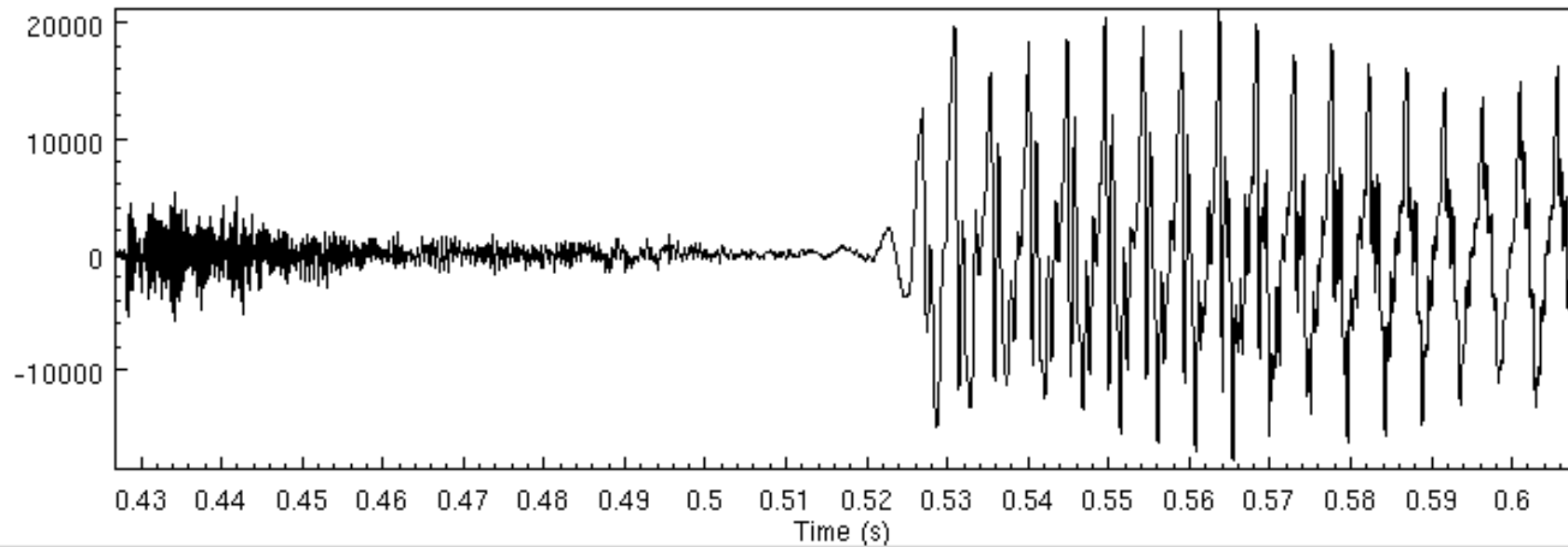
Source Filter Model



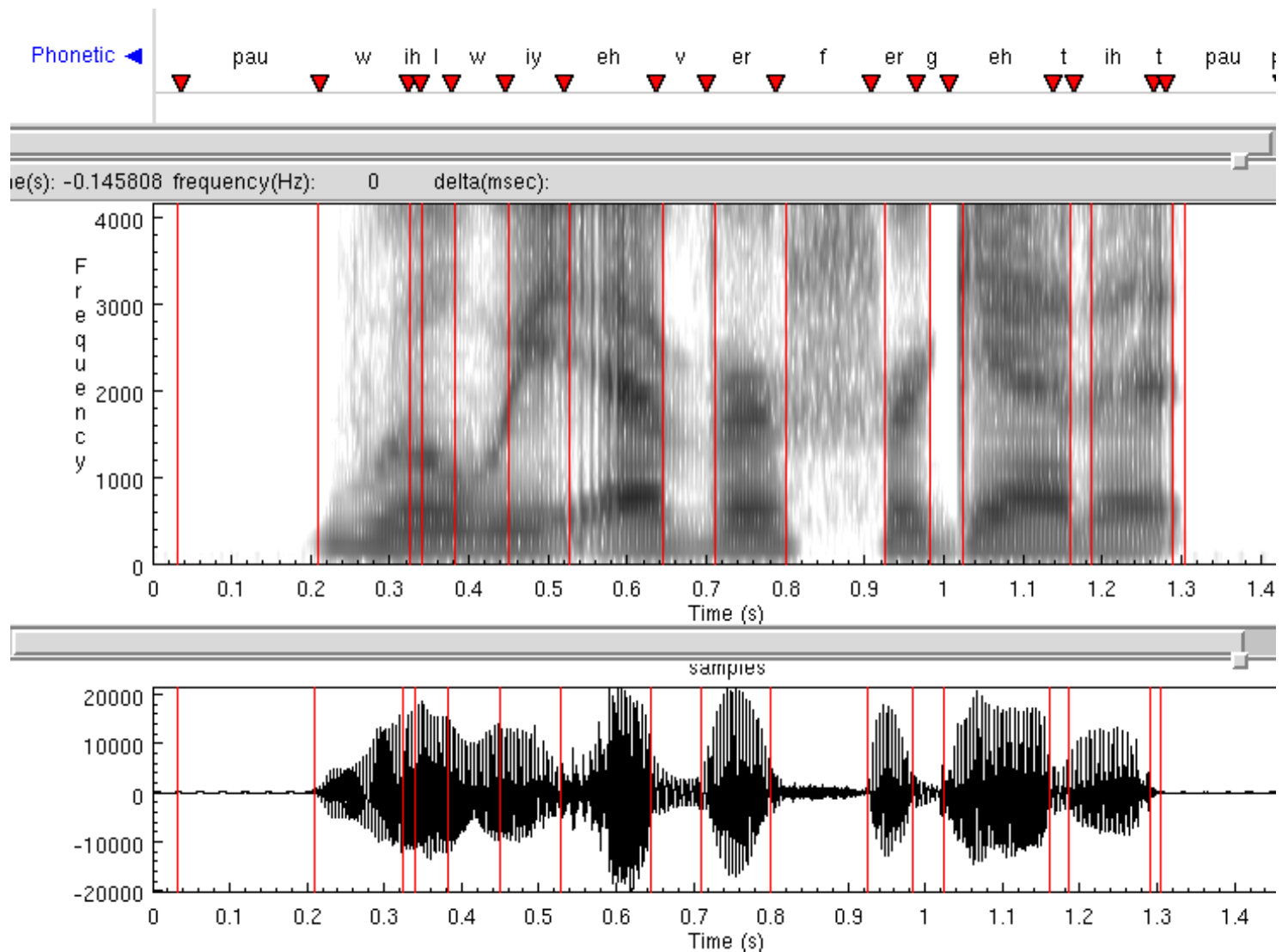
Time domain Signal



Waveform Representation

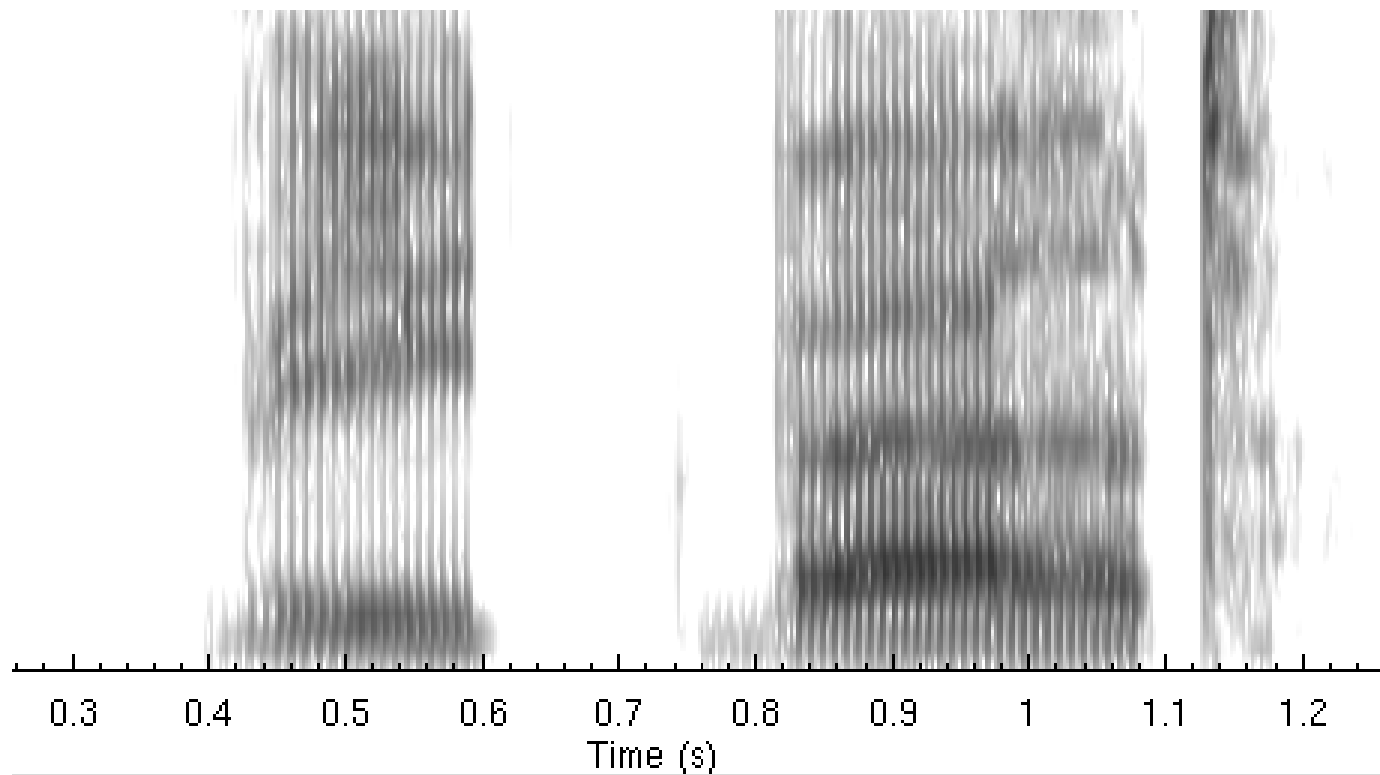


Speech Spectrogram



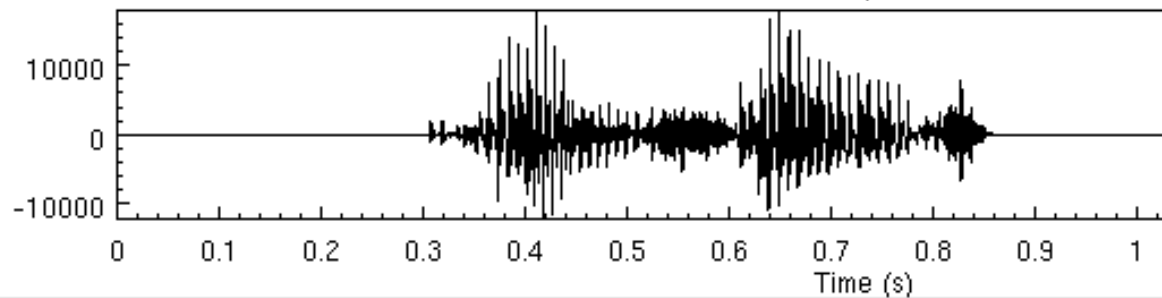
/iy/ vs /ae/

- “beat” /b iy t/ and “bat” /b ae t/

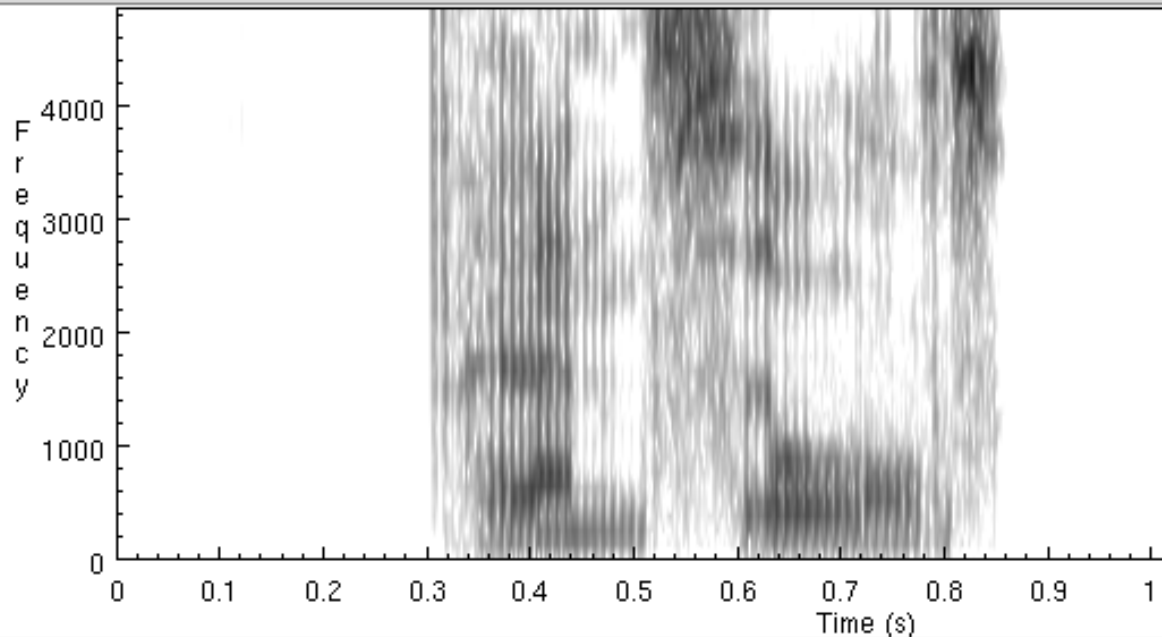


Frequency Domain

- “pencils” /p eh n s ih l z/

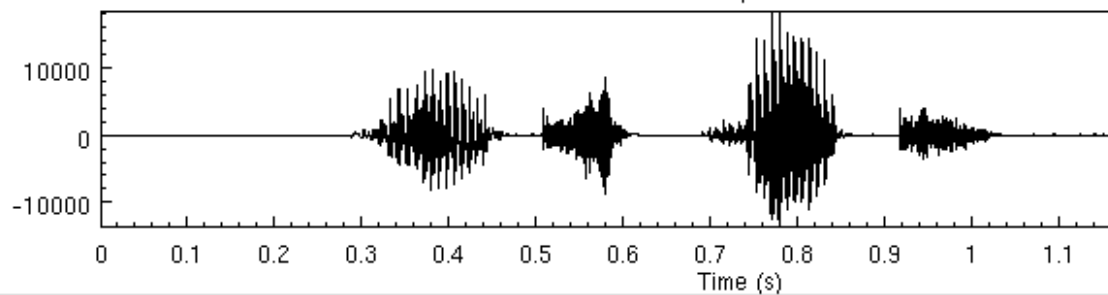


s): -0.192673 frequency(Hz): -460 delta(msec):

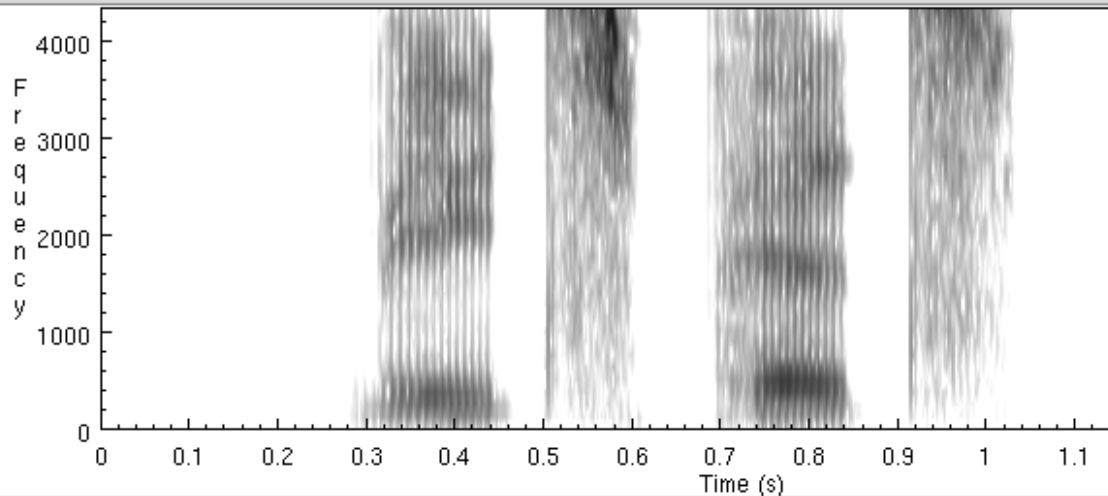


Frequency Domain

- “beats pits” / b i y t s p i h t s /



Frequency Domain



Speech Analysis

Standard Parameterization

- ◆ *Split waveform into “frames”*
 - *Advance every 10ms*
 - *Size around 25ms (overlapping frames)*
 - *Window them*
 - *Perform FFT/Mel Cepstral analysis*
 - *Find Deltas (difference from previous)*
 - *Find Delta Deltas (difference in delta)*

Summary

- ◆ *Time domain vs Frequency domain*
- ◆ *Parameterization of speech*
 - *Frequency domain*
 - *Short term FFTs*
 - *FFT vs MEL Cepstrum*

