Digitization of Speech

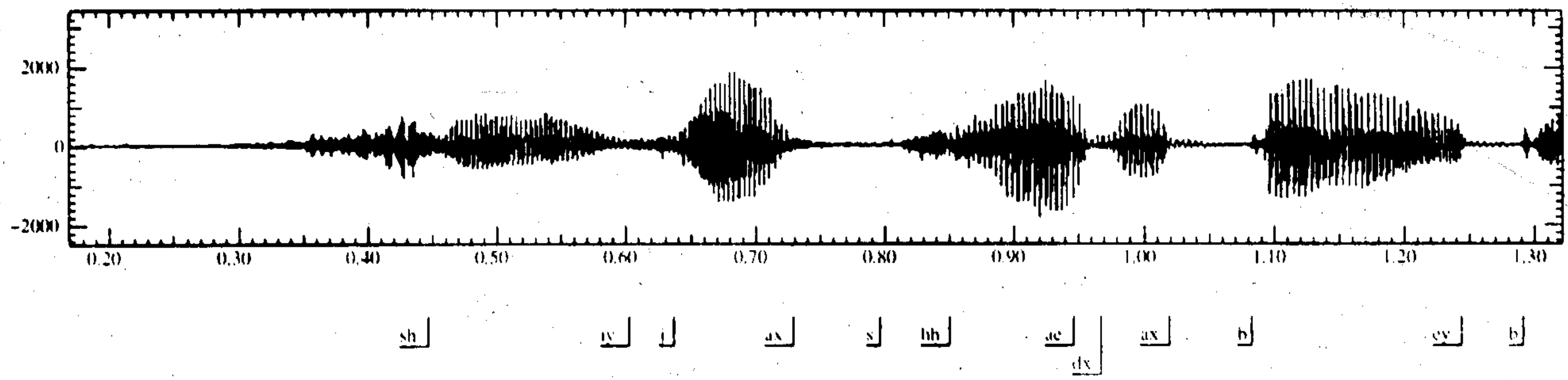
Gerald Penn

CSC 401
University of Toronto

http://www.cs.toronto.edu/∼gpenn/csc401

The physical speech signal (1)

Jurafsky and Martin



She just had a baby (Switchboard Corpus). The *x*-axis is time; the *y*-axis is amplitude.

How to Digitize Speech

Speech is a longitudinal pressure wave (although we often represent it transversally).

Speech recognizers must first:

1. Sample this. Sampling rate is typically between 6 and 40kHz.
    - Often 16 kHz per channel
    - Telephone speech: 8 kHz
    - "CD-quality:": 44.1 kHz
    - The human ear can distinguish pressure waves between 20 Hz and 20 kHz as sound, but *Nyquist's Theorem* says that the sampling frequency must be twice that of the maximum frequency that we wish to faithfully preserve.

How to Digitize Speech

2. Quantize the samples. Place bins at intervals along the y-axis, and indicate in which bin the pressure is measured at each sample time step.
   - This technique is called *pulse code modulation*
   - The number of bins determines the *sample size* — often 16 bits.
   - But long-term characteristics of speech do not yield a uniform distribution across y-bins unless we distort them — bigger bins near peaks of signal, smaller, better resolved bins near x-axis.

"Companding"

Distortion of y-bins to improve fidelity of signal relative to a fixed signal size.

Two companding methods are common in telephony: A-law (European digital), and μ-law (North America and Japan).

- A-law: $w(s) = \begin{cases} s & \text{if } \mid s \mid < \kappa A \\ \log s & \text{o.w.} \end{cases}$
- μ-law: $w(s) = sgn(s) A \frac{\log(1+\mu/A|s|)}{\log(1+\mu/A)}$

where:

- A is the maximum amplitude of the signal being quantized,
- κ is a compression parameter (in European telephony, 1/8756), and
- μ is determined by the sample size (in North America, $\mu = 255$ because the sample size is 8 bits).