Speech Recognition
Lecture 1: Introduction

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.com

Logistics

- Prerequisites: basics in analysis of algorithms and probability. No specific knowledge about signal processing.
- Workload: 2-3 homework assignments, 1 project (your choice).
- Textbooks: no single textbook covering the material presented in this course. Lecture slides available electronically.

Objectives

- Computer science view of automatic speech recognition (ASR) (no signal processing).
- Essential algorithms for large-vocabulary speech recognition.
- But, emphasis on general algorithms:
  - automata and transducer algorithms.
  - statistical learning algorithms.

Topics

- introduction, formulation, components, features.
- weighted transducer software library.
- weighted automata algorithms.
- statistical language modeling software library.
- ngram models.
- maximum entropy models.
- pronunciation models, decision trees, context-dependent models.

Topics

- search algorithms, transducer optimizations, Viterbi decoder.
- search algorithms, N-best algorithms, lattice generation, rescoring.
- structured prediction algorithms.
- adaptation.
- active learning.
- semi-supervised learning.

This Lecture

- Speech recognition problem
- Statistical formulation
- Acoustic features

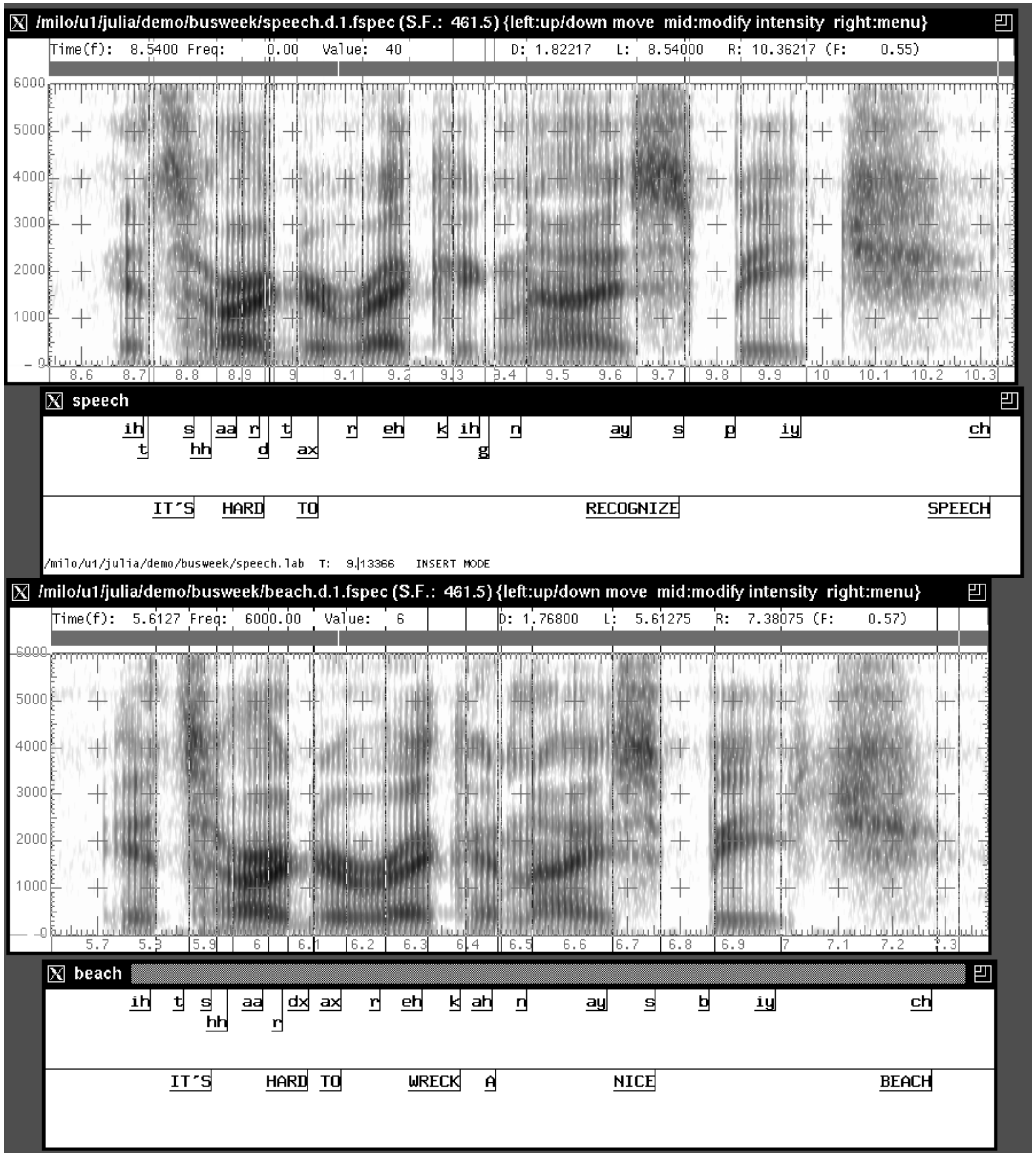
Speech Recognition Problem

- Definition: find accurate written transcription of spoken utterances.
  - transcriptions may be in words, phonemes, syllables, or other units.
- Accuracy: typically measured in terms of the edit-distance between reference transcription and sequence output by the model.

Other Related Problems

- Speaker verification.
- Speaker identification.
- Spoken-dialog systems.
- Detection of voice features, e.g., gender, age, dialect, emotion, height, weight!
- Speech synthesis.

Speech Spectogram



Speech Recognition Is Difficult

- Highly variable: the same words pronounced by the same person in the same conditions typically lead to different waveforms.
  - source variation: speaking rate, volume, accent, dialect, pitch, coarticulation.
  - channel variation: microphone (type, position), noise (background, distortion).
- Key problem: robustness to such variations.

ASR Characteristics

- Vocabulary size: small (digit recognition, 10), medium (Resource Management, 1000), large (Broadcast News, 100,000), very large (+1M).
- Speaker-dependent or speaker-independent.
- Domain-specific or unconstrained, e.g., travel reservation, modern spoken-dialog systems.
- Isolated (pause between units) or continuous.
- Read or spontaneous, e.g., dictation, news broadcast, conversational speech.

Example - Broadcast News

History

See (Juang and Rabiner, 1995)

- 1922: Radio Rex, toy, single-word recognizer (*rex*).
- 1939: voder and vocoder (mechanical synthesizer), Dudley (Bell Labs).
- 1952: isolated digit recognition, single speaker (Bell Labs).
- 1950s: 10 syllables of single speaker, Olson and Belar, (RCA Labs).
- 1950s: speaker-independent 10-vowel recognizer (MIT).

History

- 1960s: Linear Predictive Coding (LPC), Atal and Itakura.
- 1969: John Pierce's negative comments about ASR (Bell Labs).
- 1970s: Advanced Research Projects Agency (ARPA) funds speech understanding program. CMU's Harpy system based on automata had reasonable accuracy for 1,000 words.

History

- 1980s: n-gram models. ARPA Resource Management, Wall Street Journal, and ATIS tasks. Delta/delta-delta cepstra, mel cepstra.
- mid-1980s: Hidden Markov models (HMMs) become the preferred technique for speech recognition.
- 1990s: Discriminative training, vocal tract normalization, speaker adaptation. Very large-vocabulary speech recognition, e.g., 1M names recognizer (Bell Labs), 500,000 words North American Business News (NAB) recognizer.

History

- mid 1990s: FSM library. Weighted transducers major component of almost all modern speech recognition and understanding systems. SVMs, kernel methods. Dictation systems, Dragon, IBM speaker-dependent system.
- 2000s: Broadcast News, conversational speech, e.g., Switchboard, Call Home, real-time large-vocabulary systems, unconstrained spoken-dialog systems, e.g., HMIHY.

History

(Juang and Rabiner, 1995)

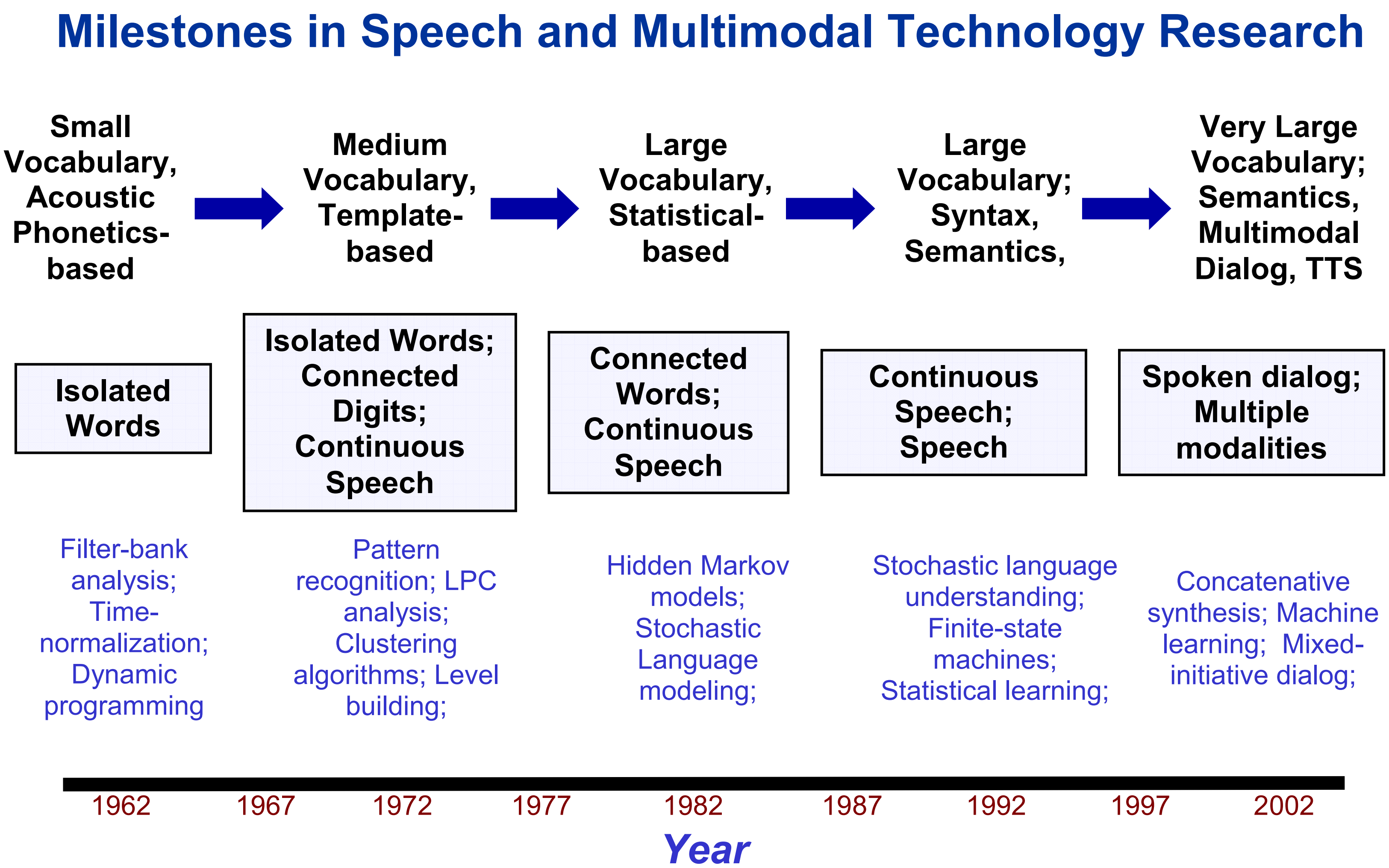


Figure 10 Milestones in Speech Recognition and Understanding Technology over the Past 40 Years.

Unscontrained Spoken-Dialog Systems

This Lecture

- Speech recognition problem
- Statistical formulation
- Acoustic features

This Lecture

- Speech recognition problem
- Statistical formulation
  - Maximum likelihood and maximum a posteriori
  - Statistical formulation of speech recognition
  - Components of a speech recognizer
- Acoustic features

Problem

- Data: sample drawn i.i.d. from set X according to some distribution D,

$$x_1, \ldots, x_m \in X.$$

- Problem: find distribution p out of a set $\mathcal{P}$ that best estimates D.

Maximum Likelihood

- Likelihood: probability of observing sample under distribution $p \in \mathcal{P}$, which, given the independence assumption is

$$\Pr[x_1, \ldots, x_m] = \prod_{i=1}^{m} p(x_i).$$

- Principle: select distribution maximizing sample probability

$$p_\star = \operatorname*{argmax}_{p \in \mathcal{P}} \prod_{i=1}^{m} p(x_i),$$

or $$p_\star = \operatorname*{argmax}_{p \in \mathcal{P}} \sum_{i=1}^{m} \log p(x_i).$$

Example: Bernoulli Trials

- **Problem**: find most likely Bernoulli distribution, given sequence of coin flips

$$H, T, T, H, T, H, T, H, H, H, T, T, \ldots, H.$$

- **Bernoulli distribution**: $p(H) = \theta, p(T) = 1 - \theta.$

- **Likelihood**: $\begin{aligned} l(p) &= \log \theta^{N(H)}(1-\theta)^{N(T)} \\ &= N(H)\log\theta + N(T)\log(1-\theta). \end{aligned}$

- **Solution**: l is differentiable and concave;

$$\frac{dl(p)}{d\theta} = \frac{N(H)}{\theta} - \frac{N(T)}{1-\theta} = 0 \Leftrightarrow \theta = \frac{N(H)}{N(H)+N(T)}.$$

Example: Gaussian Distribution

- Problem: find most likely Gaussian distribution, given sequence of real-valued observations

$$3.18, 2.35, .95, 1.175, \ldots$$

- Normal distribution: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$
- Likelihood: $l(p) = -\frac{1}{2} m \log(2\pi\sigma^2) - \sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2}.$
- Solution: l is differentiable and concave;

$$\frac{\partial p(x)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \frac{\partial p(x)}{\partial \sigma^2} = 0 \Leftrightarrow \sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2 - \mu^2.$$

Properties

- Problems:
  - the underlying distribution may not be among those searched.
  - overfitting: number of examples too small wrt number of parameters.

Maximum A Posteriori (MAP)

- Principle: select the most likely hypothesis $h \in H$ given the sample, with some *prior distribution* over the hypotheses, $\Pr[h]$,

$$\begin{aligned} h_\star &= \operatorname*{argmax}_{h \in H} \Pr[h \mid S] \\ &= \operatorname*{argmax}_{h \in H} \frac{\Pr[S|h] \Pr[h]}{\Pr[S]} \\ &= \operatorname*{argmax}_{h \in H} \Pr[S \mid h] \Pr[h]. \end{aligned}$$

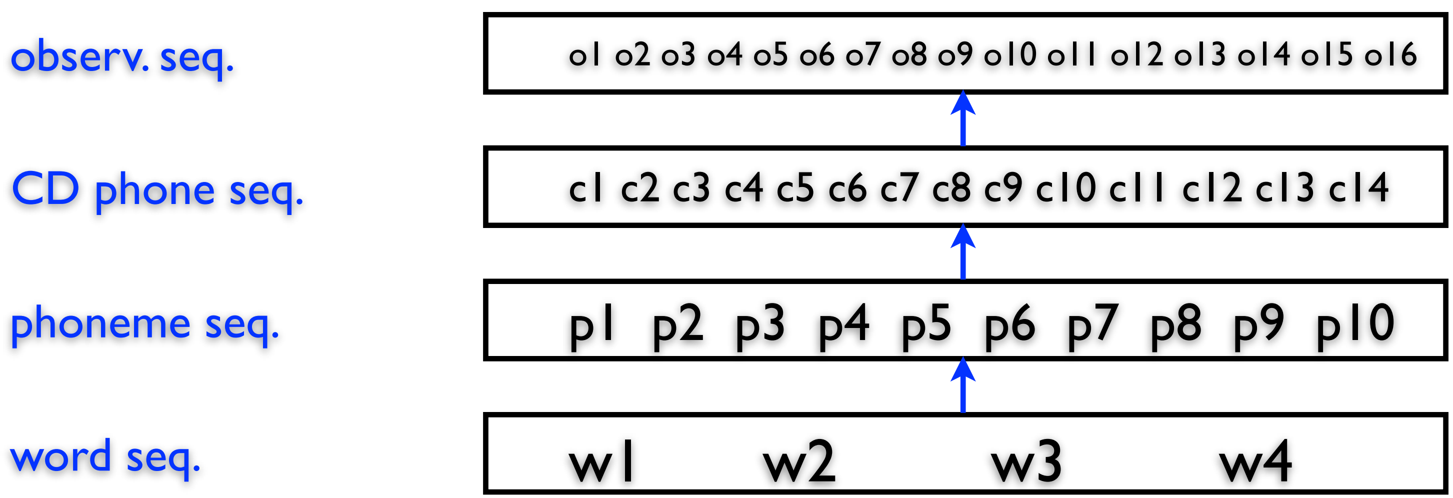
- Note: for a uniform prior, MAP coincides with maximum likelihood.

This Lecture

- Speech recognition problem
- Statistical formulation
  - Maximum likelihood and maximum a posteriori
  - Statistical formulation of speech recognition
  - Components of a speech recognizer
- Acoustic features

General Ideas

- Probabilistic formulation: given a spoken utterance, find the most likely transcription.
- Decomposition: mapping from spoken utterances to word sequences decomposed into intermediate units.



Statistical Formulation

(Bahl, Jelinek, and Mercer, 1983)

- Observation sequence produced by signal processing system: $o = o_1 \ldots o_m$.
- Sequence of words over alphabet Σ : $w = w_1 \ldots w_k$.
- Formulation (maximum a posteriori decoding):

$$
\begin{aligned}
\hat{w} &= \operatorname*{argmax}_{w \in \Sigma^*} \Pr[w \mid o] \\
&= \operatorname*{argmax}_{w \in \Sigma^*} \frac{\Pr[o \mid w] \Pr[w]}{\Pr[o]} \\
&= \operatorname*{argmax}_{w \in \Sigma^*} \underbrace{\Pr[o \mid w]}_{\text{acoustic \& pronunciation model}} \underbrace{\Pr[w]}_{\text{language model}}.
\end{aligned}
$$

Fred Jelinek

18 November 1932 - 14 September 2010

Components

- Acoustic and pronunciation model:

$$\Pr(o \mid w) = \sum_{d,c,p} \Pr(o \mid d) \Pr(d \mid c) \Pr(c \mid p) \Pr(p \mid w).$$

acoustic model

- $\Pr(o \mid d)$: observation seq. $\leftarrow$ distribution seq.
- $\Pr(d \mid c)$: distribution seq. $\leftarrow$ CD phone seq.
- $\Pr(c \mid p)$: CD phone seq. $\leftarrow$ phoneme seq.
- $\Pr(p \mid w)$: phoneme seq. $\leftarrow$ word seq.

- Language model: $\Pr(w)$, distribution over word seq.

Notes

- Formulation does not match the way speech recognition errors are typically measured: edit-distance between hypothesis and reference transcription.

This Lecture

- Speech recognition problem
- Statistical formulation
    - Maximum likelihood and maximum a posteriori
    - Statistical formulation of speech recognition
    - Components of a speech recognizer
- Acoustic features

Acoustic Observations

- Discretization
  - time: local spectral analysis of the speech waveform at regular intervals,

    $$t = t_1, \ldots, t_m, \quad t_{i+1} - t_i = 10\text{ms (typically)}.$$

    Parameter vectors

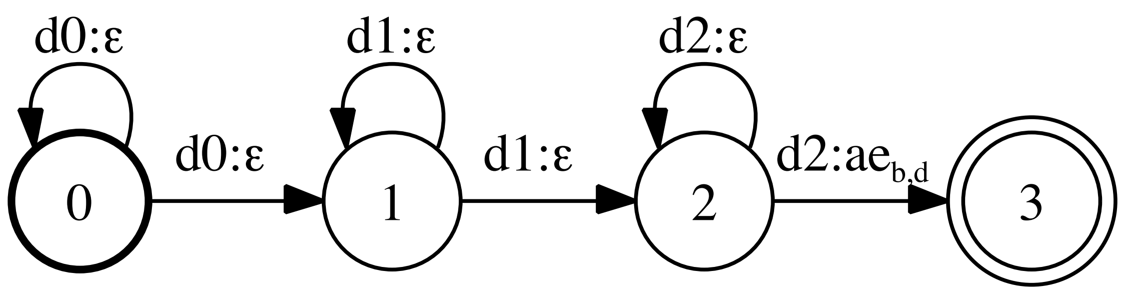
    $$o = o_1 \ldots o_m, \quad o_i \in \mathbb{R}^N, N = 39 \text{ (typically)}.$$

  - magnitude.
- Note: other perceptual information, e.g., visual information is ignored.

Acoustic Model

(Rabiner and Juang, 1993)

- Three-state hidden Markov models (HMMs)



- Distributions:
  - Full covariance multivariate Gaussians:

$$\Pr[\omega] = \frac{1}{(2\pi)^{N/2}|\sigma|^{1/2}} e^{-\frac{1}{2}(\omega-\mu)^T \sigma^{-1}(\omega-\mu)}.$$

  - Diagonal covariance Gaussian mixture.
  - Semi-continuous, tied mixtures.

Context-Dependent Model

(Lee, 1990; Young et al., 1994)

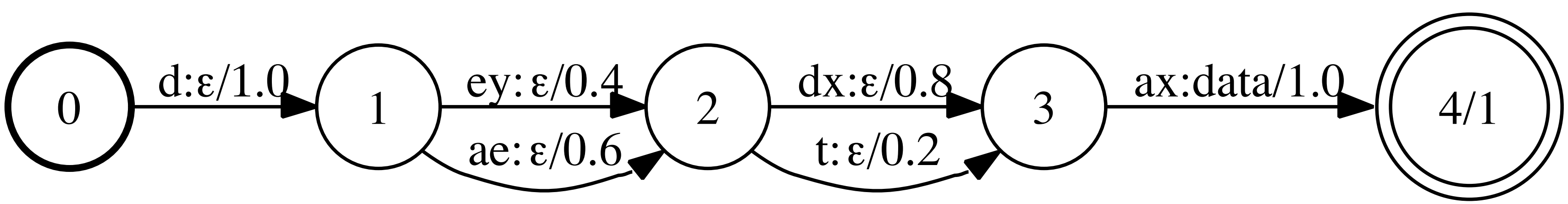
- Idea:
  - phoneme pronunciation depends on environment (allophones, co-articulation).
  - model phone in context $\rightarrow$ better accuracy.
- Context-dependent rules:
  - Context-dependent units: $ae/b\_\_\_d \rightarrow ae_{b,d}$.
  - Allophonic rules: $t/V'\_\_\_V \rightarrow dx$.
  - Complex contexts: regular expressions.

Pronunciation Dictionary

- Phonemic transcription
  - Example: word *data* in American English.

data	D ey dx ax	0.32
data	D ey t ax	0.08
data	D ae dx ax	0.48
data	D ae t ax	0.12

- Representation



Language Model

- Definition: probabilistic model for sequences of words $w = w_1 \ldots w_k$.
  - By the chain rule,

$$\Pr[w] = \prod_{i=1}^{k} \Pr[w_i \mid w_1 \ldots w_{i-1}].$$

- Modeling simplifications:
  - Clustering of histories:

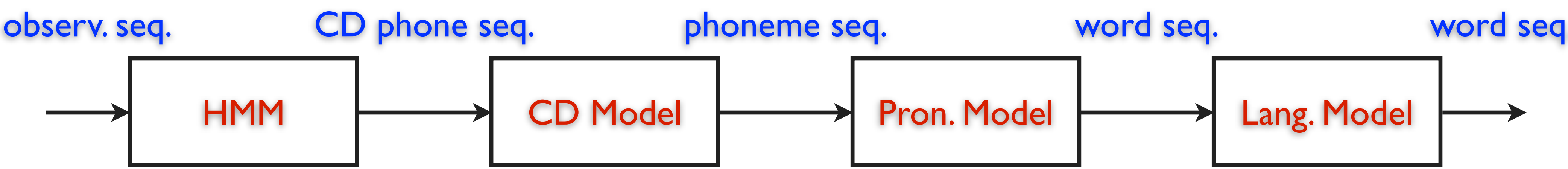
$$(w_1, \ldots, w_{i-1}) \mapsto c(w_1, \ldots, w_{i-1}).$$

  - Example: *n*th order Markov assumption,

$$\forall i, \Pr[w_i \mid w_1 \ldots w_{i-1}] = \Pr[w_i \mid h_i], \ |h_i| \leq n - 1.$$

Recognition Cascade

- Combination of components



- Viterbi approximation

$$\hat{w} = \operatorname*{argmax}_{w} \sum_{d,c,p} \Pr[o \mid d] \Pr[d \mid c] \Pr[c \mid p] \Pr[p \mid w] \Pr[w]$$

$$\approx \operatorname*{argmax}_{w} \max_{d,c,p} \Pr[o \mid d] \Pr[d \mid c] \Pr[c \mid p] \Pr[p \mid w] \Pr[w].$$

Speech Recognition Problems

- Learning: how to create accurate models for each component?
- Search: how to efficiently combine models and determine best transcription?
- Representation: compact data structure for the computational representation of the models.

→ common representation and algorithmic framework based on weighted transducers (next lectures).

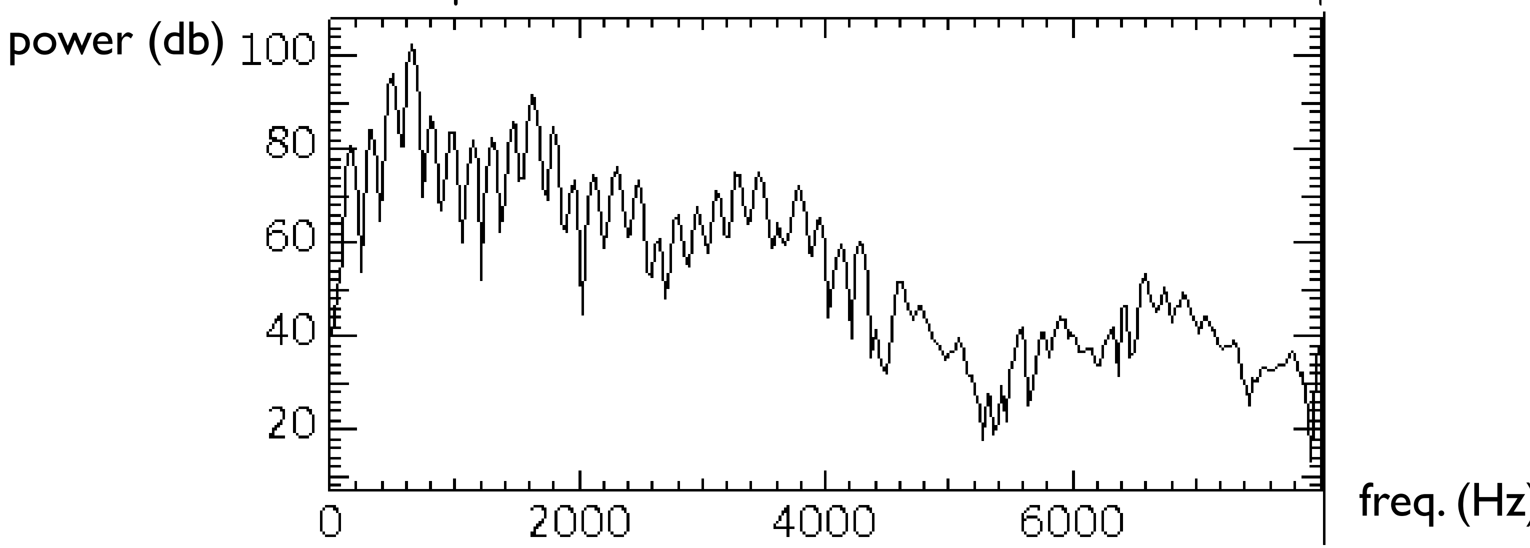
This Lecture

- Speech recognition problem
- Statistical formulation
- Acoustic features

Feature Selection

- Short-time Fourier analysis:

$$\log \left| \int x(t) w(t-\tau) e^{-i\omega t} \, dt \right|$$



Short-time (25 msec. Hamming window) spectrum of /ae/.

- Idea: find smooth approximation eliminating large variations over short frequency intervals.

Cepstral Coefficients

- Let $x(\omega)$ denote the Fourier transform of the signal.
- Definition: the 13 cepstral coefficients are the energy and the 12 first coefficients of the expansion

$$\log |x(\omega)| = \sum_{n=-\infty}^{\infty} c_n e^{-in\omega}.$$

- Other coefficients: 13 first-order (delta-cepstra) and 13 second-order (delta-delta cepstra) differentials.

Mel Frequency Cepstral Coefficients

(Stevens and Volkman, 1940)

- Refinement: non-linear scale, approximation of human perception of distance between frequencies, e.g., mel frequency scale:

$$f_{\mathrm{mel}} = 2595 \log_{10}(1 + f/700).$$

- MFCCs:
  - signal first transformed using the Mel frequency band.
  - extraction of cepstral coefficients.

Other Refinements

- Speaker/Channel adaptation:
  - mean cepstral subtraction.
  - vocal tract normalization.
  - linear transformations.

References

- Bahl, L. R., Jelinek, F., and Mercer, R. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 5(2), 179-190.

- Biing-Hwang Juang and Lawrence R. Rabiner. *Automatic Speech Recognition - A Brief History of the Technology*. Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

- Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

- Kai-Fu Lee. Context-Dependent Phonetic Hidden Markov Models for Continuous Speech Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 38(4): 599-609, 1990.

- Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

References

- S.S. Stevens and J. Volkman. The relation of pitch to frequency. American Journal of Psychology, 53:329, 1940.

- Steve Young, J. Odell, and Phil Woodland. Tree-Based State-Tying for High Accuracy Acoustic Modelling. In *Proceedings of ARPA Human Language Technology Workshop*, Morgan Kaufmann, San Francisco, 1994.