Speech Recognition
Lecture 1: Introduction

Mehryar Mohri
Courant Institute of Mathematical Sciences
mohri@cims.nyu.edu
Logistics

- **Prerequisites**: basics in analysis of algorithms and probability. No specific knowledge about signal processing.

- **Workload**: 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 Spectrogram

IT'S HARD TO RECOGNIZE SPEECH

IT'S HARD TO RECOGNIZE SPEECH

IT'S HARD TO RECOGNIZE SPEECH
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 **unconstraining**, 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

In addition information about how to t.v. biological weapons is increasingly available.
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


- **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)

Milestones in Speech and Multimodal Technology Research

Small Vocabulary, Acoustic Phonetics-based

Medium Vocabulary, Template-based

Isolated Words; Connected Digits; Continuous Speech

Connected Words; Continuous Speech

Large Vocabulary, Statistical-based

Hidden Markov models; Stochastic Language modeling;

Continuous Speech; Speech

Stochastic language understanding; Finite-state machines; Statistical learning;

Very Large Vocabulary; Semantics, Multimodal Dialog, TTS

Spoken dialog; Multiple modalities

Filter-bank analysis; Time-normalization; Dynamic programming

Pattern recognition; LPC analysis; Clustering algorithms; Level building;

Concatenative synthesis; Machine learning; Mixed-initiative dialog;


Year

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

$$p^* = \arg\max_{p \in \mathcal{P}} \prod_{i=1}^{m} p(x_i),$$

or

$$p^* = \arg\max_{p \in \mathcal{P}} \sum_{i=1}^{m} \log p(x_i).$$
Relative Entropy Formulation

- **Empirical distribution**: distribution \( \hat{p} \) that assigns to each point the frequency of its occurrence in the sample.

- **Lemma**: \( p^* \) has maximum likelihood \( l(p^*) \) iff

\[
p^* = \arg\min_{p \in \mathcal{P}} D(\hat{p} \| p).
\]

- **Proof**:

\[
D(\hat{p} \| p) = \sum_{z_i \text{ observed}} \hat{p}(z_i) \log \hat{p}(z_i) - \sum_{z_i \text{ observed}} \hat{p}(z_i) \log p(z_i)
= -H(\hat{p}) - \sum_{z_i} \frac{\text{count}(z_i)}{N_0} \log p(z_i)
= -H(\hat{p}) - \frac{1}{N_0} \log \prod_{z_i} p(z_i)^{\text{count}(z_i)}
= -H(\hat{p}) - \frac{1}{N_0} l(p).
\]
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**: \( l(p) = \log \theta^N(H) (1 - \theta)^N(T) \)
  \[ = N(H) \log \theta + N(T) \log(1 - \theta). \]

- **Solution**: \( l \) is differentiable and concave;

  \[ \frac{dl(p)}{d\theta} = \frac{N(H)}{\theta} - \frac{N(T)}{\theta} = 0 \iff \theta = \frac{N(H)}{N(H) + N(T)}. \]
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: 
\[ l(p) = -\frac{1}{2} m \log(2\pi \sigma^2) - \frac{1}{2} \sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2}. \]

Solution: \( l \) is differentiable and concave;

\[ \frac{\partial p(x)}{\partial \mu} = 0 \iff \mu = \frac{1}{m} \sum_{i=1}^{m} x_i \quad \frac{\partial p(x)}{\partial \sigma^2} = 0 \iff \sigma^2 = \frac{1}{m} \sum_{i=1}^{m} x_i^2 - \mu^2. \]
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, \( \Pr[h] \),

\[
h^* = \arg\max_{h \in H} \Pr[h | S]
\]

\[
= \arg\max_{h \in H} \frac{\Pr[S|h] \Pr[h]}{\Pr[S]}
\]

\[
= \arg\max_{h \in H} \Pr[S | h] \Pr[h].
\]

**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.

```
observ. seq.       o1 o2 o3 o4 o5 o6 o7 o8 o9 o10 o11 o12 o13 o14 o15 o16
CD phone seq.     c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14
phoneme seq.      p1  p2  p3  p4  p5  p6  p7  p8  p9  p10
word seq.         w1  w2  w3  w4
```
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 \( \Sigma : w = w_1 \ldots w_k \).
- **Formulation** (maximum a posteriori decoding):

\[
\hat{w} = \arg\max_{w \in \Sigma^*} \Pr[w \mid o]
\]

\[
= \arg\max_{w \in \Sigma^*} \frac{\Pr[o \mid w] \Pr[w]}{\Pr[o]}
\]

\[
= \arg\max_{w \in \Sigma^*} \Pr[o \mid w] \Pr[w].
\]

acoustic & pronunciation model  language model
Components

- **Acoustic and pronunciation model:**

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

- \(Pr(o \mid d)\): 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

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.

(Rabiner and Juang, 1993)
Context-Dependent Model

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

- **Idea:**
  - phoneme pronunciation depends on environment (allophones, co-articulation).
  - model phone in context → 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.

<table>
<thead>
<tr>
<th>pronunciation</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>D ey dx ax</td>
</tr>
<tr>
<td>data</td>
<td>D ey t ax</td>
</tr>
<tr>
<td>data</td>
<td>D ae dx ax</td>
</tr>
<tr>
<td>data</td>
<td>D ae t ax</td>
</tr>
</tbody>
</table>

- 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

\[
\hat{w} = \arg \max_w \sum_{d,c,p} \Pr[o \mid d] \Pr[d \mid c] \Pr[c \mid p] \Pr[p \mid w] \Pr[w]
\]

\approx \arg \max_w \max_{d,c,p} \Pr[o \mid d] \Pr[d \mid c] \Pr[c \mid p] \Pr[p \mid w] \Pr[w].

- Viterbi approximation
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.
This Lecture

- Speech recognition problem
- Statistical formulation
- Acoustic features
Feature Selection

- **Short-time spectral analysis:**

\[
\log \left| \int g(\tau) x(t + \tau) e^{-i2\pi f \tau} d\tau \right|
\]

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

**Idea:** find a 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:

  \[ f_{\text{mel}} = 2595 \log_{10}(1 + f/700). \]

- **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


References
