Automatic Speech Recognition (CS753)
Lecture 1: Introduction to Statistical Speech Recognition

Instructor: Preethi Jyothi
July 24, 2017
Course Specifics
Pre-requisites

Ideal Background:
Completed one of “Foundations of ML (CS 725)” or “Advanced ML (CS 726)” or “Foundations of Intelligent Agents (CS 747)” at IITB or have completed an ML course elsewhere.

Also acceptable as pre-req:
Completed courses in EE that deal with ML concepts.
Experience working on research projects that are ML-based.

Less ideal but still works:
Comfortable with probability, linear algebra and multivariable calculus. (Currently enrolled in CS 725.)
About the course (I)

Main Topics:

- Introduction to statistical ASR
- Acoustic models
  - Hidden Markov models
  - Deep neural network-based models
- Pronunciation models
- Language models (Ngram models, RNN-LMs)
- Decoding search problem (Viterbi algorithm, etc.)
About the course (II)

Course webpage:
www.cse.iitb.ac.in/~pjyothi/cs753

Reading:
All mandatory reading will be freely available online.
Reading material will be posted on the website.

Attendance:
Strongly advised to attend all lectures given there’s no fixed
textbook and a lot of the material covered in class will not be
on the slides

Audit requirements:
Complete all three assignments and score ≥40% on each of them
Evaluation — Assignments

**Grading:** 3 assignments + 1 mid-sem exam making up 50% of the grade.

**Format:**

1. One assignment will be almost entirely programming-based. The other two will contain a mix of problems to be solved by hand and programming questions.

2. Mid-sem and final exam will test concepts you’ve been taught in class.

**Late Policy:** 10% reduction in marks for every additional day past the due date. Submissions closed three days after the due date.
Evaluation — Final Project

**Grading:** Constitutes 25% of the total grade. (Exceptional projects could get extra credit. Details posted on website.)

**Team:** 2-3 members. Individual projects are highly discouraged.

**Project requirements:**

- Discuss proposed project with me on or before August 17th.
- Intermediate deadline: Project progress report. Due on September 28th.
- Finally, turn in: 4-5 page final report about methodology & detailed experiments
- Project presentation/demo
Evaluation — Final Project

About the Project:

- Could be implementation of ideas learnt in class, applied to real data (and/or to a new task)
- Could be a new idea/algorithm (with preliminary experiments)
- Excellent projects can turn into conference/workshop papers
Evaluation — Final Project

About the Project:

• Could be implementation of ideas learnt in class, applied to real data (and/or to a new task)
• Could be a new idea/algorithm (with preliminary experiments)
• Excellent projects can turn into conference/workshop papers

Sample project ideas:

• Detecting accents from speech
• Sentiment classification from voice-based reviews
• Language recognition from speech segments
• Audio search of speeches by politicians
Final Project Landscape (Spring ’17)

- Music Genre Classification
- Audio Synthesis Using LSTMs
- Emotion Recognition from speech
- Programming with speech-based commands
- Sanskrit Synthesis and Recognition
- Speaker Identification
- Swapping instruments in recordings
- Voice-based music player
- Speech synthesis & ASR for Indic languages
- InfoGAN for music
- Automatic authorised ASR
- Keyword spotting for continuous speech
- Tabla bol transcription
- Singer Identification
- Ad detection in live radio streams
- Speaker Verification
- Swapping in recordings
- End-to-end Audio-Visual Speech Recognition
- Speaker Adaptation
- Nationality detection from speech accents
- Programming with speech-based commands
Evaluation — Final Exam

Grading: Constitutes 25% of the total grade.

Syllabus: Will be tested on all the material covered in the course.

Academic Integrity Policy

• Write what you know.
• Use your own words.
• If you refer to *any* external material, *always* cite your sources. Follow proper citation guidelines.
• If you’re caught for plagiarism or copying, penalties are much higher than simply omitting that question.
• **In short:** Just not worth it. Don’t do it!

Image credit: https://www.flickr.com/photos/kurok/22196852451
Introduction to Speech Recognition
Exciting time to be an AI/ML researcher!

Image credit: http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html
Lots of new progress

How Google's AlphaGo Beat a Go World Champion

Here's what it's like to ride in Uber's self-driving car

Intel looks to a new chip to power the coming age of AI

Facebook, Amazon, Google, IBM, Microsoft form new AI alliance

What is speech recognition?
Why is it such a hard problem?
Automatic Speech Recognition (ASR)

- Automatic speech recognition (or speech-to-text) systems transform speech utterances into their corresponding text form, typically in the form of a word sequence.
Automatic Speech Recognition (ASR)

• Automatic speech recognition (or speech-to-text) systems transform speech utterances into their corresponding text form, typically in the form of a word sequence.

• Many downstream applications of ASR:
  • Speech understanding: comprehending the semantics of text
  • Audio information retrieval: searching speech databases
  • Spoken translation: translating spoken language into foreign text
  • Keyword search: searching for specific content words in speech

• Other related tasks include speaker recognition, speaker diarization, speech detection, etc.
History of ASR

RADIO REX (1922)
History of ASR

SHOEBOX (IBM, 1962)

1 word

Freq.
detector
History of ASR

1922

Freq. detector

1 word

1932

16 words

Isolated word recognition

1942

1952

1962

1972

1982

1992

2002

2012

HARPY (CMU, 1976)
History of ASR

1922
- Freq. detector
1932
- 1 word
1942
- Isolated word recognition
1952
- 16 words
1962
- Connected speech
1972
- 1000 words
1982
- HIDDEN MARKOV MODELS (1980s)
1992
2002
2012
History of ASR

Deep Neural Network Based Systems (>2010)

- 1 word: Freq. detector
- 16 words: Isolated word recognition
- 1000 words: Connected speech
- 10K+ words: LVCSR systems

Timeline:
- 1922
- 1932
- 1942
- 1952
- 1962
- 1972
- 1982
- 1992
- 2002
- 2012

Innovations:
- Amazon Echo
- Cortana
- Siri
- "Ok Google"
History of ASR

What’s next?

- 1 word
- 16 words
- 1000 words
- 10K+ words
- 1M+ words

Freq. detector
Isolated word recognition
Connected speech
LVCSR systems
DNN-based systems

VOICE IS THE NEXT BIG PLATFORM, UNLESS YOU HAVE AN ACCENT
This can’t be blamed on ASR

“Call me a taxi"

From now on, I’ll call you ‘Taxi’. OK?

Cancel   Yes
ASR is the front-engine
Why is ASR a challenging problem?

Variabilities in different dimensions:

**Style**: Read speech or spontaneous (conversational) speech? Continuous natural speech or command & control?

**Speaker characteristics**: Rate of speech, accent, prosody (stress, intonation), speaker age, pronunciation variability even when the same speaker speaks the same word.

**Channel characteristics**: Background noise, room acoustics, microphone properties, interfering speakers.

**Task specifics**: Vocabulary size (very large number of words to be recognized), language-specific complexity, resource limitations.
Noisy channel model

Claude Shannon
1916-2001
Noisy channel model applied to ASR

Claude Shannon  
1916-2001

Fred Jelinek  
1932-2010
Statistical Speech Recognition

Let $O$ represent a sequence of acoustic observations (i.e. $O = \{O_1, O_2, \ldots, O_t\}$ where $O_i$ is a feature vector observed at time $t$) and $W$ denote a word sequence. Then, the decoder chooses $W^*$ as follows:

$$W^* = \arg \max_W \Pr(W|O)$$

$$= \arg \max_W \frac{\Pr(O|W) \Pr(W)}{\Pr(O)}$$

This maximisation does not depend on $\Pr(O)$. So, we have

$$W^* = \arg \max_W \Pr(O|W) \Pr(W)$$
Statistical Speech Recognition

\[ W^* = \arg \max_W \Pr(O|W) \Pr(W) \]

\( \Pr(O|W) \) is referred to as the “acoustic model”

\( \Pr(W) \) is referred to as the “language model”

speech signal \rightarrow Acoustic Feature Generator \rightarrow SEARCH \rightarrow Acoustic Model \rightarrow Language Model

word sequence \( W^* \)
Example: Isolated word ASR task

**Vocabulary:**
10 digits (zero, one, two, ...), 2 operations (plus, minus)

**Data:**
Speech utterances corresponding to each word sample from multiple speakers

Recall the acoustic model is \( \Pr(O|W) \): direct estimation is impractical (why?)

Let’s parameterize \( \Pr_\alpha(O|W) \) using a Markov model with parameters \( \alpha \). Now, the problem reduces to estimating \( \alpha \).
Isolated word-based acoustic models

Transition probabilities denoted by $a_{ij}$ from state $i$ to state $j$

Observation vectors $O_t$ are generated from the probability density $b_j(O_t)$

Isolated word-based acoustic models

For an $O=\{O_1, O_2, \ldots, O_6\}$ and a state sequence $Q=\{0, 1, 1, 2, 3, 4\}$:

$$Pr(O, Q|W = \text{‘one’}) = a_{01} b_1(O_1) a_{11} b_1(O_2) \ldots$$

$$Pr(O|W = \text{‘one’}) = \sum_Q Pr(O, Q|W = \text{‘one’})$$
Isolated word recognition

one: \[\text{Figure 2.1: Standard topology used to represent a phone HMM.}\]

\[
\begin{align*}
\text{O}_1 & \rightarrow a_{11} \rightarrow b_1(\cdot) \rightarrow \text{O}_2 \\
\text{O}_2 & \rightarrow a_{12} \rightarrow b_2(\cdot) \rightarrow \text{O}_3 \\
\text{O}_3 & \rightarrow a_{21} \rightarrow b_3(\cdot) \rightarrow \text{O}_4 \\
\text{O}_4 & \rightarrow a_{31} \rightarrow b_4(\cdot) \rightarrow \text{O}_5 \\
\end{align*}
\]

\[
\begin{align*}
\Pr(O|W = \text{'one'}) & \\
\Pr(O|W = \text{'two'}) & \\
\text{Pick } \arg\max \Pr(O|W = w) & \\
\text{What are we assuming about } \Pr(W)? & \\
\Pr(O|W = \text{'plus'}) & \\
\Pr(O|W = \text{'minus'}) &
\]

acoustic features

O

Pick arg max \( \Pr(O|W = w) \)

What are we assuming about \( \Pr(W) \)?
Isolated word recognition

one: $0 \xrightarrow{a_{01}} 1 \xrightarrow{a_{12}} 2 \xrightarrow{a_{23}} 3 \xrightarrow{a_{34}} 4$

two: $0 \xrightarrow{a_{01}} 1 \xrightarrow{a_{12}} 2 \xrightarrow{a_{23}} 3 \xrightarrow{a_{34}} 4$

... acoustic features...

plus: $0 \xrightarrow{a_{01}} 1 \xrightarrow{a_{12}} 2 \xrightarrow{a_{23}} 3 \xrightarrow{a_{34}} 4$

minus: $0 \xrightarrow{a_{01}} 1 \xrightarrow{a_{12}} 2 \xrightarrow{a_{23}} 3 \xrightarrow{a_{34}} 4$

$Pr(O|W = 'one')$

$Pr(O|W = 'two')$

Is this approach scalable?

$Pr(O|W = 'plus')$

$Pr(O|W = 'minus')$
Why are word-based models not scalable?

Example

Words

Phonemes

Pronunciation model maps words to phoneme sequences
Recall: Statistical Speech Recognition

\[ W^* = \arg \max_W \Pr(O|W) \Pr(W) \]

Diagram:
- Speech signal
- Acoustic Feature Generator
- SEARCH
- Acoustic Model
- Language Model
- Word sequence
- \( W^* \)
Statistical Speech Recognition

\[ W^* = \arg \max_W \Pr(O|W) \Pr(W) \]

- **speech signal**
- **Acoustic Feature Generator**
- **Search**
- **Acoustic Model (phonemes)**
- **Pronunciation Model**
- **Language Model**

**word sequence**

**W***
Evaluate an ASR system

Quantitative metric: Error rates computed on an unseen test set by comparing $W^*$ (decoded output) against $W_{\text{ref}}$ (reference sentence) for each test utterance

- Sentence/Utterance error rate (trivial to compute!)
- Word/Phone error rate
Evaluate an ASR system

Word/Phone error rate (ER) uses the Levenshtein distance measure: What are the minimum number of edits (insertions/deletions/substitutions) required to convert $W^*$ to $W_{ref}$?

On a test set with $N$ instances:

\[
ER = \frac{\sum_{j=1}^{N} \text{Ins}_j + \text{Del}_j + \text{Sub}_j}{\sum_{j=1}^{N} \ell_j}
\]

\text{Ins}_j, \text{Del}_j, \text{Sub}_j \text{ are number of insertions/deletions/substitutions in the } j^{th} \text{ ASR output}

\ell_j \text{ is the total number of words/phones in the } j^{th} \text{ reference}
Course Overview

- Speaker Adaptation
- Hybrid HMM-DNN Systems
- Deep Neural Networks
- Hidden Markov Models
- Acoustic Model (phones)
- Pronunciation Model
- G2P/feature-based models
- Ngram/RNN LMs

Input: speech signal

1. Acoustic Feature Generator
2. Properties of speech sounds
3. Acoustic Signal Processing
4. SEARCH
5. word sequence W*
6. Language Model
7. Ngram/RNN LMs
Course Overview

- Speaker Adaptation
- Hybrid HMM-DNN Systems
- Deep Neural Networks
- Hidden Markov Models
- Acoustic Model (phones)
- Acoustic Feature Generator
- Properties of speech sounds
- Acoustic Signal Processing
- Search algorithms
- Pronunciation Model
- Language Model
- G2P/feature-based models
- Ngram/RNN LMs
- word sequence \( W^* \)