Automatic Speech Recognition: Introduction

Peter Bell

Automatic Speech Recognition— ASR Lecture 1
13 January 2020
**Course details**

- **Lectures:** About 18 lectures
- **Labs:** Weekly lab sessions – using Python, Kaldi ([kaldi-asr.org](http://kaldi-asr.org) and OpenFst ([openfst.org](http://openfst.org))
  - Lab sessions in AT-3.09: Tuesdays 10:00, Wednesdays 10:00, Wednesdays 15:10, start week 2 (21/22 January)
  - Slots are allocated on Learn
- **Assessment:**
  - Exam in April or May (worth 70%)
  - Coursework (worth 30%, building on the lab sessions) (out on Thursday 13 February; in by Wednesday 18 March)
- **People:**
  - Lecturer: Peter Bell
  - TA: Andrea Carmantini

[http://www.inf.ed.ac.uk/teaching/courses/asr/](http://www.inf.ed.ac.uk/teaching/courses/asr/)
Your background

If you have taken:

- Speech Processing \textit{and} either of (MLPR or MLP)
  - Perfect!
- either of (MLPR or MLP) \textit{but not} Speech Processing
  (probably you are from Informatics)
  - You’ll require some speech background:
    - A couple of the lectures will cover material that was in Speech Processing
    - Some additional background study (including material from Speech Processing)
- Speech Processing \textit{but neither of} (MLPR or MLP)
  (probably you are from SLP)
  - You’ll require some machine learning background (especially neural networks)
    - A couple of introductory lectures on neural networks provided for SLP students
    - Some additional background study
Labs

- Series of weekly labs using Python, OpenFst and Kaldi
  - Labs are allocated on Learn
- Labs start week 2 (next week)
- Labs 1-4 will give you hands-on experience of building your own ASR system
  - **Note:** these labs are an important pre-requisite for the coursework
- Later labs will introduce you to Kaldi recipes for training acoustic models – useful if you will be doing an ASR-related research project
What is speech recognition?
What is speech recognition?
What is speech recognition?

Speech-to-text transcription
- Transform recorded audio into a sequence of words
- Just the words, no meaning.... But do need to deal with acoustic ambiguity: “Recognise speech?” or “Wreck a nice beach?”
- Speaker diarization: Who spoke when?
- Speech recognition: what did they say?
- Paralinguistic aspects: how did they say it? (timing, intonation, voice quality)
- Speech understanding: what does it mean?
Why is speech recognition difficult?
From a linguistic perspective

Many sources of variation

**Speaker** Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics
From a linguistic perspective

Many sources of variation

**Speaker** Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics

**Environment** Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)
From a linguistic perspective

Many sources of variation

**Speaker** Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics

**Environment** Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

**Style** Continuously spoken or isolated? Planned monologue or spontaneous conversation?
From a linguistic perspective

Many sources of variation

**Speaker**  Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics

**Environment**  Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

**Style**  Continuously spoken or isolated? Planned monologue or spontaneous conversation?

**Vocabulary**  Machine-directed commands, scientific language, colloquial expressions
From a linguistic perspective

Many sources of variation

**Speaker**  Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics

**Environment**  Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

**Style**  Continuously spoken or isolated? Planned monologue or spontaneous conversation?

**Vocabulary**  Machine-directed commands, scientific language, colloquial expressions

**Accent/dialect**  Recognise the speech of all speakers who speak a particular language
From a linguistic perspective

Many sources of variation

**Speaker** Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics

**Environment** Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

**Style** Continuously spoken or isolated? Planned monologue or spontaneous conversation?

**Vocabulary** Machine-directed commands, scientific language, colloquial expressions

**Accent/dialect** Recognise the speech of all speakers who speak a particular language

**Other paralinguistics** Emotional state, social class, ...
From a linguistic perspective

Many sources of variation

**Speaker**  Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics

**Environment**  Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

**Style**  Continuously spoken or isolated? Planned monologue or spontaneous conversation?

**Vocabulary**  Machine-directed commands, scientific language, colloquial expressions

**Accent/dialect**  Recognise the speech of all speakers who speak a particular language

**Other paralinguistics**  Emotional state, social class, ...

**Language spoken**  Estimated 7,000 languages, most with limited training resources; code-switching; language change
From a machine learning perspective

- As a classification problem: very high dimensional output space
- As a sequence-to-sequence problem: very long input sequence (although limited re-ordering between acoustic and word sequences)
- Data is often noisy, with many “nuisance” factors of variation in the data
- Very limited quantities of training data available (in terms of words) compared to text-based NLP
- Manual speech transcription is very expensive (10x real time)
- Hierarchical and compositional nature of speech production and comprehension makes it difficult to handle with a single model
From a machine learning perspective

- As a classification problem: very high dimensional output space
- As a sequence-to-sequence problem: very long input sequence (although limited re-ordering between acoustic and word sequences)

Data is often noisy, with many “nuisance” factors of variation in the data

Very limited quantities of training data available (in terms of words) compared to text-based NLP

Manual speech transcription is very expensive (10x real time)

Hierachical and compositional nature of speech production and comprehension makes it difficult to handle with a single model
From a machine learning perspective

- As a classification problem: very high dimensional output space
- As a sequence-to-sequence problem: very long input sequence (although limited re-ordering between acoustic and word sequences)
- Data is often noisy, with many “nuisance” factors of variation in the data
From a machine learning perspective

- As a classification problem: very high dimensional output space
- As a sequence-to-sequence problem: very long input sequence (although limited re-ordering between acoustic and word sequences)
- Data is often noisy, with many "nuisance" factors of variation in the data
- Very limited quantities of training data available (in terms of words) compared to text-based NLP
  - Manual speech transcription is very expensive (10x real time)
From a machine learning perspective

- As a classification problem: very high dimensional output space
- As a sequence-to-sequence problem: very long input sequence (although limited re-ordering between acoustic and word sequences)
- Data is often noisy, with many “nuisance” factors of variation in the data
- Very limited quantities of training data available (in terms of words) compared to text-based NLP
  - Manual speech transcription is very expensive (10x real time)
- Hierarchical and compositional nature of speech production and comprehension makes it difficult to handle with a single model
Example: recognising TV broadcasts

MGB CHALLENGE

BBC Three showcase extravaganza.
We generally represent recorded speech as a sequence of acoustic feature vectors (observations), $X$ and the output word sequence as $W$. 
The speech recognition problem

- We generally represent recorded speech as a sequence of acoustic feature vectors (observations), $X$ and the output word sequence as $W$.
- At recognition time, our aim is to find the most likely $W$, given $X$. 

We generally represent recorded speech as a sequence of acoustic feature vectors (observations), \( X \) and the output word sequence as \( W \).

At recognition time, our aim is to find the most likely \( W \), given \( X \).

To achieve this, statistical models are trained using a corpus of labelled training utterances \((X^n, W^n)\).
Representing recorded speech (X)

Represent a recorded utterance as a sequence of feature vectors

Reading: Jurafsky & Martin section 9.3
Labelling speech (W)

Labels may be at different levels: words, phones, etc. Labels may be *time-aligned* – i.e. the start and end times of an acoustic segment corresponding to a label are known.

Reading: Jurafsky & Martin chapter 7 (especially sections 7.4, 7.5)
Two key challenges

In **training** the model:
- Aligning the sequences $X^n$ and $W^n$ for each training utterance
Two key challenges

In **training** the model:
Aligning the sequences $X^n$ and $W^n$ for each training utterance
In **training** the model:

Aligning the sequences $X^n$ and $W^n$ for each training utterance.
Two key challenges

In **training** the model:

- Aligning the sequences $X^n$ and $W^n$ for each training utterance.
Two key challenges

In training the model:
Aligning the sequences $X^n$ and $W^n$ for each training utterance
Two key challenges

In **training** the model:
Aligning the sequences $X^n$ and $W^n$ for each training utterance

In **performing recognition**:
Searching over all possible output sequences $W$ to find the most likely one
Two key challenges

In **training** the model:
Aligning the sequences $X^n$ and $W^n$ for each training utterance

In **performing recognition**:
Searching over all possible output sequences $W$ to find the most likely one

The **hidden Markov model** (HMM) provides a good solution to both problems
The Hidden Markov Model

- A simple but powerful model for mapping a sequence of continuous observations to a sequence of discrete outputs
- It is a **generative** model for the observation sequence
- Algorithms for training (forward-backward) and recognition-time decoding (Viterbi)
The Hidden Markov Model

- A simple but powerful model for mapping a sequence of continuous observations to a sequence of discrete outputs
- It is a **generative** model for the observation sequence
- Algorithms for training (forward-backward) and recognition-time decoding (Viterbi)
- Later in the course we will also look at newer all-neural, fully-differentiable “end-to-end” models
Hierarchical modelling of speech

**Generative Model**

"No right"

Utterance $W$

Word

Subword

HMM

Acoustics $X$

ASR Lecture 1  Automatic Speech Recognition: Introduction 16
If $X$ is the sequence of acoustic feature vectors (observations) and $W$ denotes a word sequence, the most likely word sequence $W^*$ is given by

$$W^* = \arg \max_W P(W | X)$$
If $X$ is the sequence of acoustic feature vectors (observations) and $W$ denotes a word sequence, the most likely word sequence $W^*$ is given by

$$W^* = \arg \max_W P(W | X)$$

Applying Bayes’ Theorem:

$$P(W | X) = \frac{p(X | W)P(W)}{p(X)}$$

$$\propto p(X | W)P(W)$$

$$W^* = \arg \max_W \underbrace{p(X | W)}_{\text{Acoustic model}} \underbrace{P(W)}_{\text{Language model}}$$
\[ W^* = \arg \max_{W} p(\mathbf{X} | W) P(W) \]

Use an acoustic model, language model, and lexicon to obtain the most probable word sequence \( W^* \) given the observed acoustics \( \mathbf{X} \).
Phones and Phonemes

- **Phonemes**
  - abstract unit defined by linguists based on contrastive role in word meanings (eg “cat” vs “bat”)
  - 40–50 phonemes in English

- **Phones**
  - speech sounds defined by the acoustics
  - many *allophones* of the same phoneme (eg /p/ in “pit” and “spit”)
  - limitless in number

Phones are usually used in speech recognition – but no conclusive evidence that they are the basic units in speech recognition

Possible alternatives: syllables, automatically derived units, ...
Example: TIMIT Corpus

- TIMIT corpus (1986)—first widely used corpus, still in use
  - Utterances from 630 North American speakers
  - Phonetically transcribed, time-aligned
  - Standard training and test sets, agreed evaluation metric (phone error rate)

- TIMIT phone recognition - label the audio of a recorded utterance using a sequence of phone symbols
  - Frame classification – attach a phone label to each frame data
  - Phone classification – given a segmentation of the audio, attach a phone label to each (multi-frame) segment
  - Phone recognition – supply the sequence of labels corresponding to the recorded utterance
Basic speech recognition on TIMIT

- Train a classifier of some sort to associate each feature vector with its corresponding label. Classifier could be
  - Neural network
  - Gaussian mixture model
  - ...

Then at run time, a label is assigned to each frame

- Questions
  - What’s good about this approach?
  - What the limitations? How might we address them?
How accurate is a speech recognizer?

String edit distance
- Use dynamic programming to align the ASR output with a reference transcription
- Three type of error: insertion, deletion, substitutions

Word error rate (WER) sums the three types of error. If there are $N$ words in the reference transcript, and the ASR output has $S$ substitutions, $D$ deletions and $I$ insertions, then:

$$\text{WER} = 100 \cdot \frac{S + D + I}{N} \%$$

Accuracy = 100 – WER%

For TIMIT, define phone error error rate analogously to word error rate

Speech recognition evaluations: common training and development data, release of new test sets on which different systems may be evaluated using word error rate
Recording Speech

Signal Analysis

Acoustic Model

Language Model

Search Space

Decoded Text (Transcription)

Training Data
Reading

- Jurafsky and Martin (2008). *Speech and Language Processing* (2nd ed.): Chapter 7 (esp 7.4, 7.5) and Section 9.3.
- General interest:
    http://www.economist.com/technology-quarterly/2017-05-01/language