Automatic Speech Recognition: Introduction

Steve Renals & Hiroshi Shimodaira

Automatic Speech Recognition— ASR Lecture 1
15 January 2018

Automatic Speech Recognition — ASR

Course details

- **Lectures:** About 18 lectures, plus a couple of extra lectures on basic introduction to neural networks
- **Labs:** Weekly lab sessions – using Kaldi (`kaldi-asr.org`) to build speech recognition systems.
  - Lab sessions in AT-4.12: Tuesdays 10:00, Wednesdays 10:00, Wednesdays 15:10, start week 2 (23/24 January)
  - Select one lab session at `https://doodle.com/poll/gxmh9kwp3a8espxx`
- **Assessment:**
  - Exam in April or May (worth 70%)
  - Coursework (worth 30%, building on the lab sessions): out on Monday 12 February; in by Wednesday 14 March
- **People:**
  - Lecturers: Steve Renals and Hiroshi Shimodaira
  - TAs: Joachim Fainberg and Ondrej Klejch

Your background

If you have taken:

- Speech Processing *and* either of (MLPR or MLP)
  - Perfect!
- either of (MLPR or MLP) *but not* Speech Processing
  - 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 *but neither of* (MLPR or MLP)
  - You'll require some machine learning background (especially neural networks)
    - A couple of introductory lectures on neural networks
    - Some additional background study

Labs

- Series of weekly labs using Kaldi.
- Labs start week 2 (next week)
- **Note:** Training speech recognisers can take time
  - ASR training in some labs will not finish in an hour...
  - Give yourself plenty of time to complete the coursework, don't leave it until the last couple of days

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?

Variability in speech recognition

Several sources of variation

Size Number of word types in vocabulary, perplexity

Variability in speech recognition

Several sources of variation

Size Number of word types in vocabulary, perplexity

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

Variability in speech recognition

Several sources of variation

Size Number of word types in vocabulary, perplexity

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

Acoustic environment Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

Variability in speech recognition

Several sources of variation

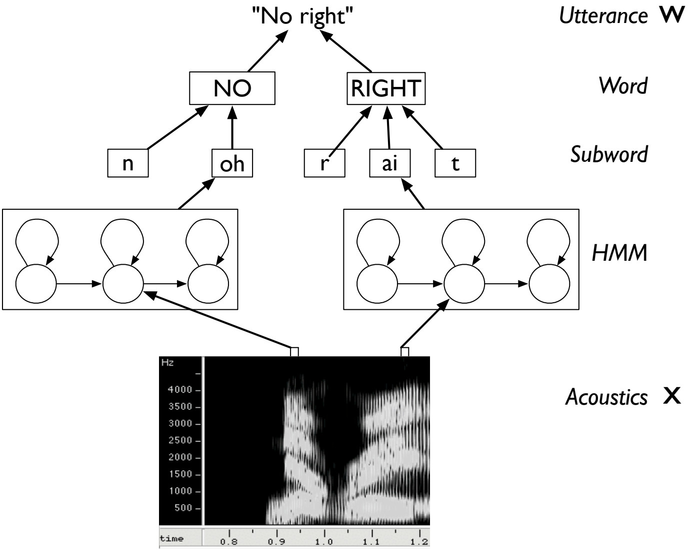
Size Number of word types in vocabulary, perplexity

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

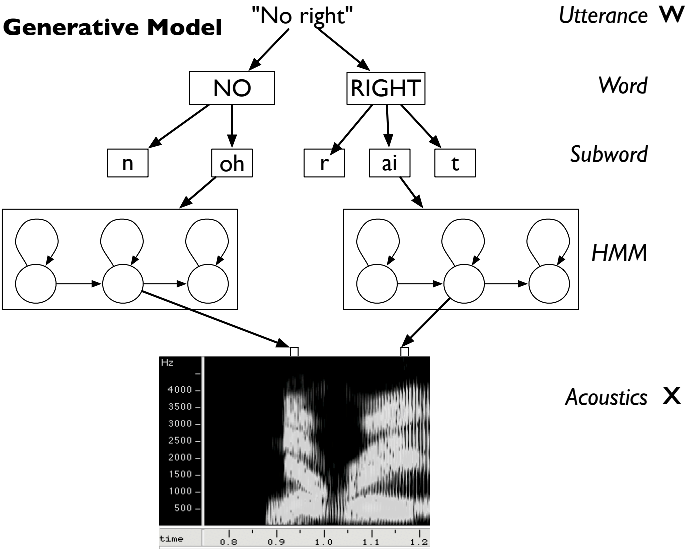
Acoustic environment Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

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

Hierarchical modelling of speech



Hierarchical modelling of speech



“Fundamental Equation of Statistical Speech Recognition”

If $\mathbf{X}$ is the sequence of acoustic feature vectors (observations) and $\mathbf{W}$ denotes a word sequence, the most likely word sequence $\mathbf{W}^*$ is given by

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X})$$

"Fundamental Equation of Statistical Speech Recognition"

If $\mathbf{X}$ is the sequence of acoustic feature vectors (observations) and $\mathbf{W}$ denotes a word sequence, the most likely word sequence $\mathbf{W}^*$ is given by

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X})$$

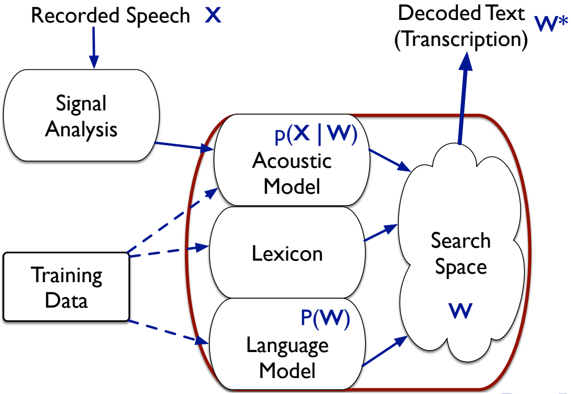
Applying Bayes' Theorem:

$$\begin{aligned} P(\mathbf{W} \mid \mathbf{X}) &= \frac{p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W})}{p(\mathbf{X})} \\ &\propto p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W}) \\ \mathbf{W}^* &= \arg\max_{\mathbf{W}} \underbrace{p(\mathbf{X} \mid \mathbf{W})}_{\text{Acoustic model}} \quad \underbrace{P(\mathbf{W})}_{\text{Language model}} \end{aligned}$$

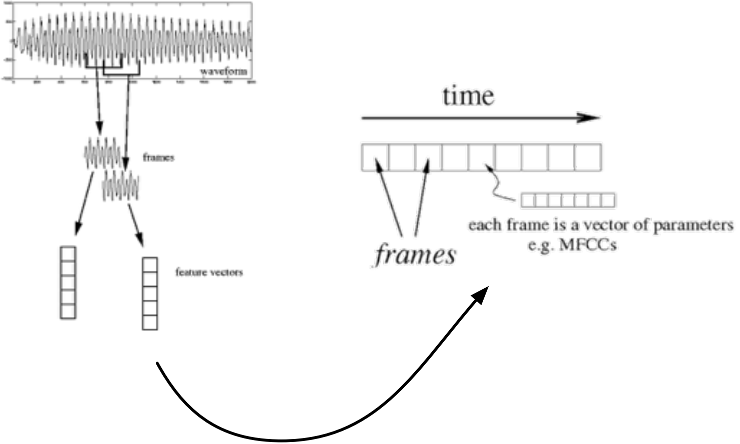
Speech Recognition Components

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W})$$

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

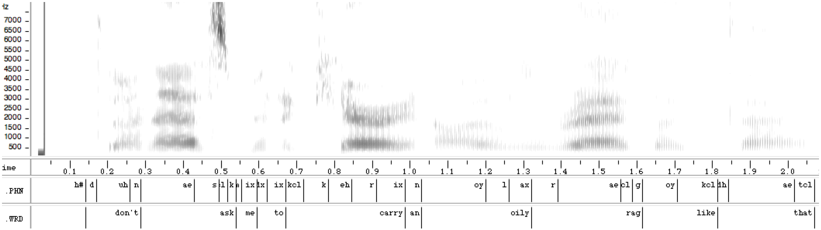


Representing recorded speech (X)



Represent a recorded utterance as a sequence of *feature vectors*

Reading: Jurafsky & Martin section 9.3



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)

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, ...

(Slide taken from Martin Cooke from long ago)

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
  - ...

  The at test time, a label is assigned to each frame
- Questions
  - What's good about this approach?
  - What the limitations? How might we address them?

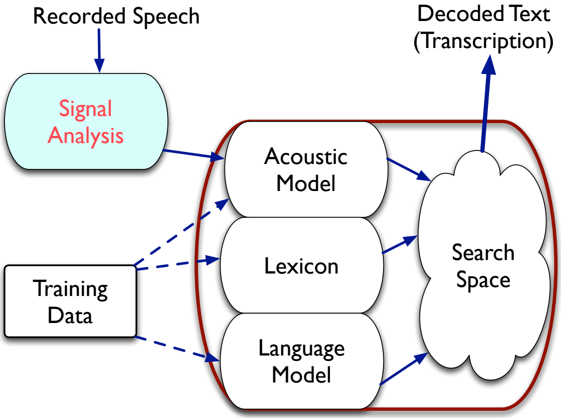
Evaluation

- 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:

$$\mathrm{WER} = 100 \cdot \frac{S + D + I}{N} \% \qquad \mathrm{Accuracy} = 100 - \mathrm{WER}\%$$

- For TIMIT, define phone error error rate analagously 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

Next Lecture



Reading

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