# Introduction to Speech Recognition

Steve Renals & Hiroshi Shimodaira

Automatic Speech Recognition— ASR Lecture 1
13 January 2014

# Automatic Speech Recognition — ASR

## Course details

- About 15 lectures
- Some lab coursework: build a large vocabulary ASR system using HTK (worth 15%)
- Literature review coursework (worth 15%)
- An exam in April or May (worth 70%)
- Books and papers:
  - Jurafsky & Martin (2008), *Speech and Language Processing*, Pearson Education (2nd edition). (J&M)
  - Some general review and tutorial articles
  - Readings for specific topics
- If you haven't taken Speech Processing...
  — read J&M, chapter 7 (Phonetics)

http://www.inf.ed.ac.uk/teaching/courses/asr/

# Automatic Speech Recognition — ASR

## Course content

- Introduction to statistical speech recognition
- The basics
  - Speech signal processing
  - Acoustic modelling with HMMs
  - Pronunciations and language models
  - Search
- Advanced topics:
  - Adaptation
  - (Deep) neural networks
  - Discriminative training

http://www.inf.ed.ac.uk/teaching/courses/asr/

# Automatic Speech Recognition — ASR

## Course content

- Introduction to statistical speech recognition
- The basics
  - Speech signal processing
  - **Acoustic modelling with HMMs**
  - Pronunciations and language models
  - Search
- Advanced topics:
  - Adaptation
  - **(Deep) neural networks**
  - Discriminative training

http://www.inf.ed.ac.uk/teaching/courses/asr/

# Overview

## Introduction to Speech Recognition

Today

- Overview
- Statistical Speech Recognition
- Hidden Markov Models (HMMs)

http://www.inf.ed.ac.uk/teaching/courses/asr/

# What is ASR?

**Speech-to-text transcription**

- Transform recorded audio into a sequence of words
- Just the words, no meaning....
- But: "Will the new display recognise speech?" or "Will the nudist play 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)

How would ASR be useful?
Potential applications?

# 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?

# Spontaneous vs. Planned

Oh [laughter] he he used to be pretty crazy but I think now that he's kind of gotten his act together now that he's mentally uh sharp he he doesn't go in for that anymore.

Dictated Imitated Spontaneous

# Linguistic Knowledge or Machine Learning?

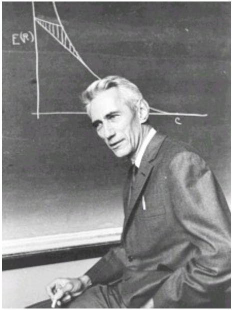
- Intense effort needed to derive and encode linguistic rules that cover all the language
- Very difficult to take account of the variability of spoken language with such approaches
- Data-driven machine learning: Construct simple models of speech which can be learned from large amounts of data (thousands of hours of speech recordings)

# Statistical Speech Recognition

Thomas Bayes (1701-1761)

А. А. Марков (1886).

AA Markov (1856-1922)




Claude Shannon (1916-2001)

# 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:

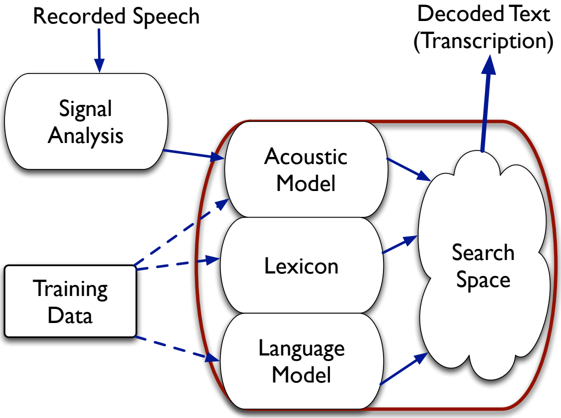
$$\begin{aligned} P(\mathbf{W} \mid \mathbf{X}) &= \frac{p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W})}{p(\mathbf{X})} \\ &\propto p(\mathbf{X} \mid \mathbf{W}) P(\mathbf{W}) \\ \mathbf{W}^* &= \arg\max_{\mathbf{W}} \underbrace{p(\mathbf{X} \mid \mathbf{W})}_{\text{Acoustic model}} \quad \underbrace{P(\mathbf{W})}_{\text{Language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

# Statistical speech recognition

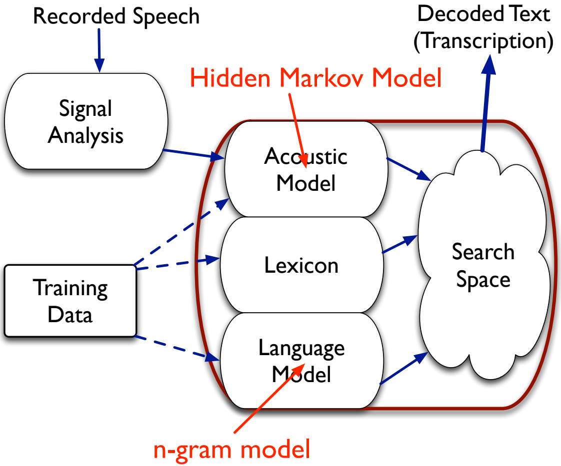
Statistical models offer a statistical "guarantee" — see the licence conditions of the best known automatic dictation system, for example:

> *Licensee understands that* **speech recognition is a statistical process** *and that* **recognition errors are inherent in the process**. *Licensee acknowledges that it is licensee's responsibility to* **correct recognition errors before using the results of the recognition**.

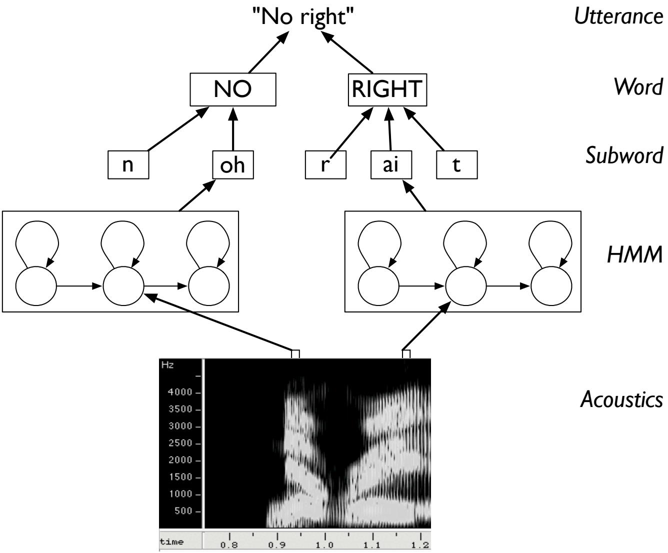
# Statistical Speech Recognition



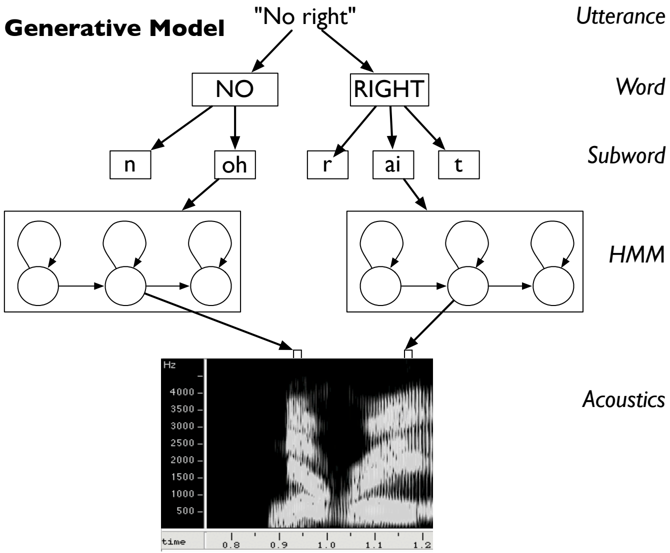
# Statistical Speech Recognition



# Hierarchical modelling of speech



# Hierarchical modelling of speech



# Data

- The statistical framework is based on learning from data
- Standard corpora with agreed evaluation protocols very important for the development of the ASR field

# Data

- The statistical framework is based on learning from data
- Standard corpora with agreed evaluation protocols very important for the development of the ASR field
- 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)

# Data

- The statistical framework is based on learning from data
- Standard corpora with agreed evaluation protocols very important for the development of the ASR field
- 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)
- Many standard corpora released since TIMIT: DARPA Resource Management, read newspaper text (eg Wall St Journal), human-computer dialogues (eg ATIS), broadcast news (eg Hub4), conversational telephone speech (eg Switchboard), multiparty meetings (eg AMI)
- Corpora have real value when closely linked to evaluation benchmark tests (with new test data from the same domain)

# Evaluation

- How accurate is a speech recognizer?

# Evaluation

- How accurate is a speech recognizer?
- Use dynamic programming to align the ASR output with a reference transcription
- Three type of error: insertion, deletion, substitution

# Evaluation

- How accurate is a speech recognizer?
- Use dynamic programming to align the ASR output with a reference transcription
- Three type of error: insertion, deletion, substitution
- 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:

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

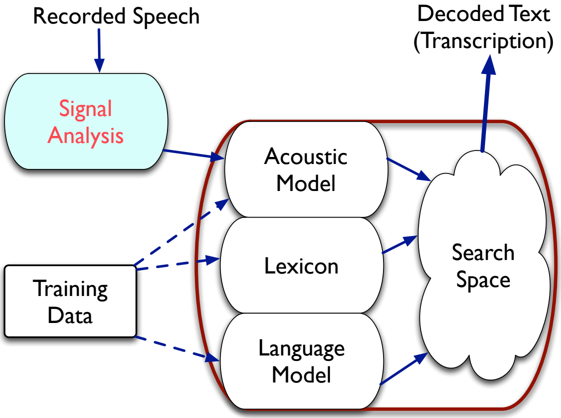
# Evaluation

- How accurate is a speech recognizer?
- Use dynamic programming to align the ASR output with a reference transcription
- Three type of error: insertion, deletion, substitution
- 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:

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

- Speech recognition evaluations: common training and development data, release of new test sets on which different systems may be evaluated using word error rate
  - NIST evaluations enabled an objective assessment of ASR research, leading to consistent improvements in accuracy
  - May have encouraged incremental approaches at the cost of subduing innovation ("Towards increasing speech recognition error rates")

# Next Lecture



# Reading

- Jurafsky and Martin (2008). *Speech and Language Processing* (2nd ed.): Chapter 9 to end of sec 9.3.
- Renals and Hain (2010). “Speech Recognition”, *Computational Linguistics and Natural Language Processing Handbook*, Clark, Fox and Lappin (eds.), Blackwells. (on website)