Speech Understanding

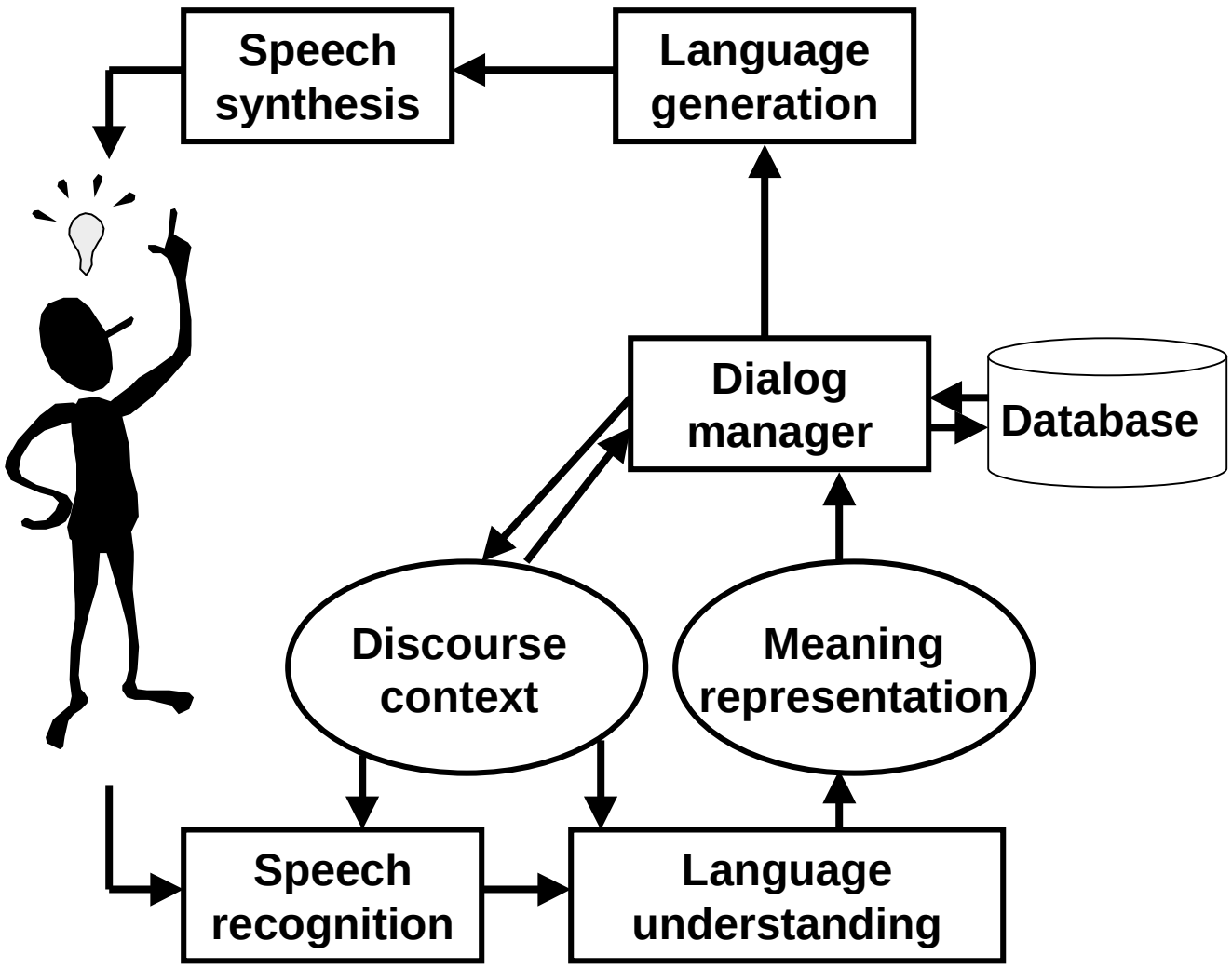
- 15-381 Artificial Intelligence
- Greg Aist

- Goals for this lecture:
  - Identify the main components involved in talking to your computer
  - Identify different knowledge sources and the role they play in speech recognition
  - Understand speech recognition in terms of probability
  - Understand speech recognition as search

Goal: Computers you can talk to

- Why ?
  - Speech is ~4 times faster than typing
  - Speech does not interfere with using your hands
  - Speech lets you move around
  - Speech is natural and requires less training
- How?

How to talk to a computer



– (Survey of the State of the Art in Human Language Technologies, ed. R. Cole; figures throughout except where noted)

Focus: Speech recognition

- Turning speech into text

Characterizing the problem

- From easy to hard...

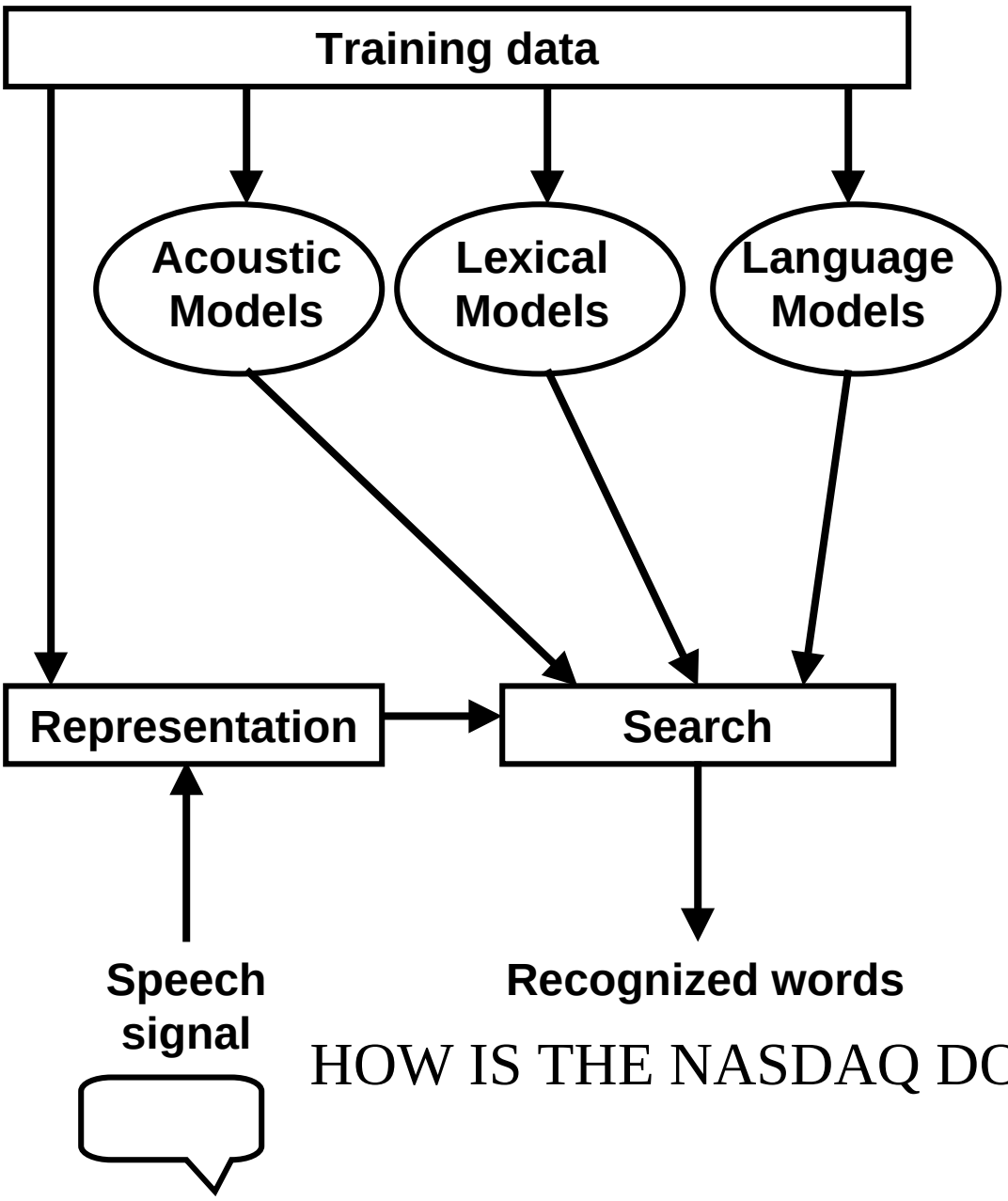
Parameters	Range
Speaking Mode	Isolated words to continuous speech
Speaking Style	Read speech to spontaneous speech
Enrollment	Speaker-dependent to Speaker-independent
Vocabulary	Small (< 20 words) to large (> 20,000 words)
Language Model	Finite-state to context-sensitive
Perplexity	Small (< 10) to large (> 100)
SNR	High (> 30 dB) to low (< 10 dB)
Transducer	Voice-cancelling microphone to telephone

- An easy problem:
  Training a computer to
  learn when you say “yes” or
  “no”
- A hard problem:
  Transcribing conference calls

What sources of knowledge are available?

- Acoustic information: pitch, ...
- Phonetic and phonemic: what are the basic sounds of the language?
- Lexical: what are the words?
- Morphological: -ing, -ed
- Syntactic: what is the grammar?
- Semantic: what do sentences mean?
- Pragmatic: what do sentences mean in context?

Components of a speech recognizer



Language model

- How?
- Suprisingly, a “simple” method works well...
- Model three-word sequences
  - called trigrams
- Use two words in a row to predict next word
  - President Bill _____
  - in the ____
- Conditional probability: P(Clinton| President Bill)

Lexical model

- Represent words
- Spelling and sound
  - *cat* /k ae td/
  - *read* /r iy d/
  - *read*(2) /r eh dd/
- Notice that the sound of a word is represented as a sequence of smaller sounds (phonemes)

Acoustic model

- Represent phonemes in terms of (features derived from) the speech signal
- Essentially, fit a statistical model to (a transformed version of) the speech data, yielding a vector of coefficients
- To the recognizer, speech looks like a sequence of vectors

Speech recognition in terms of probability

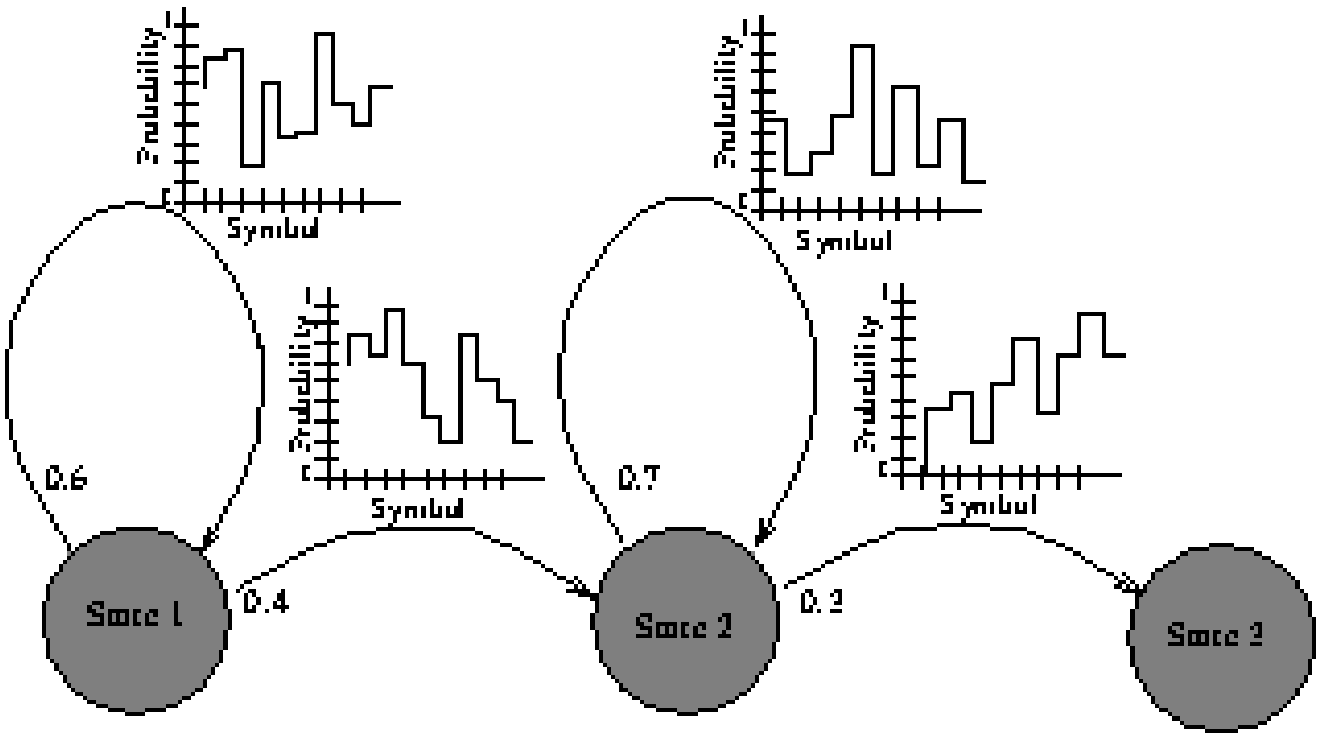
- Let W = the sequence of words
- Let A = the acoustic signal -- what was said
- We want to find W such that P(W|A) is maximized.
- How?

Bayes' rule applied

- $P(W|A) = \frac{P(W) * P(A|W)}{P(A)}$
- We just want to maximize this probability, not compute it...
- Maximize
  $P(W) * P(A|W)$
- What is P(W)?
  Language model
- What is P(A|W)
  Acoustic model

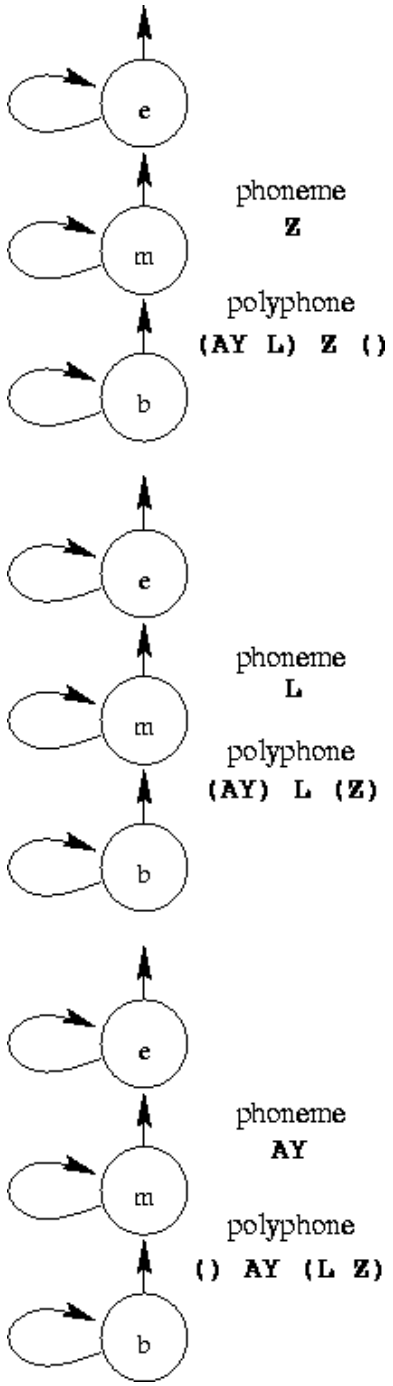
A bit more about the acoustic model

- Markov model with unknown state



- Hidden Markov Model (HMM) for one phoneme
- Initial, transition, and output probabilities

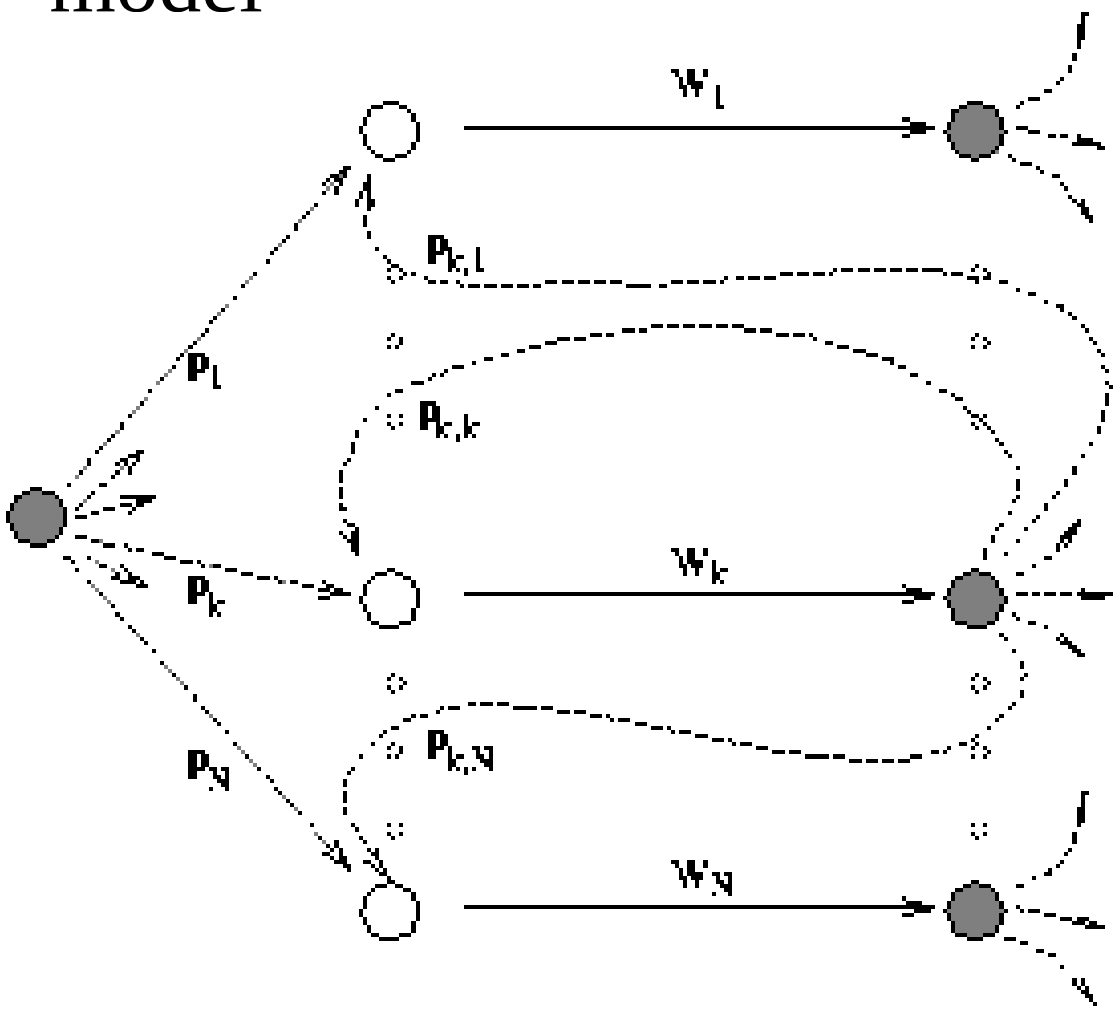
Hidden Markov Model for the word *aisles*



- http://www.cs.cmu.edu/~monika/thesis/html/node17.html

Speech recognition as search

- (In principle) Build a graph from the language model, the lexical model, and the acoustic model



How well might various search techniques work?

- Bad: Depth-first
- Good
  - A*
  - Viterbi search
    Most probable path through
    the model, given the observations

Viterbi search

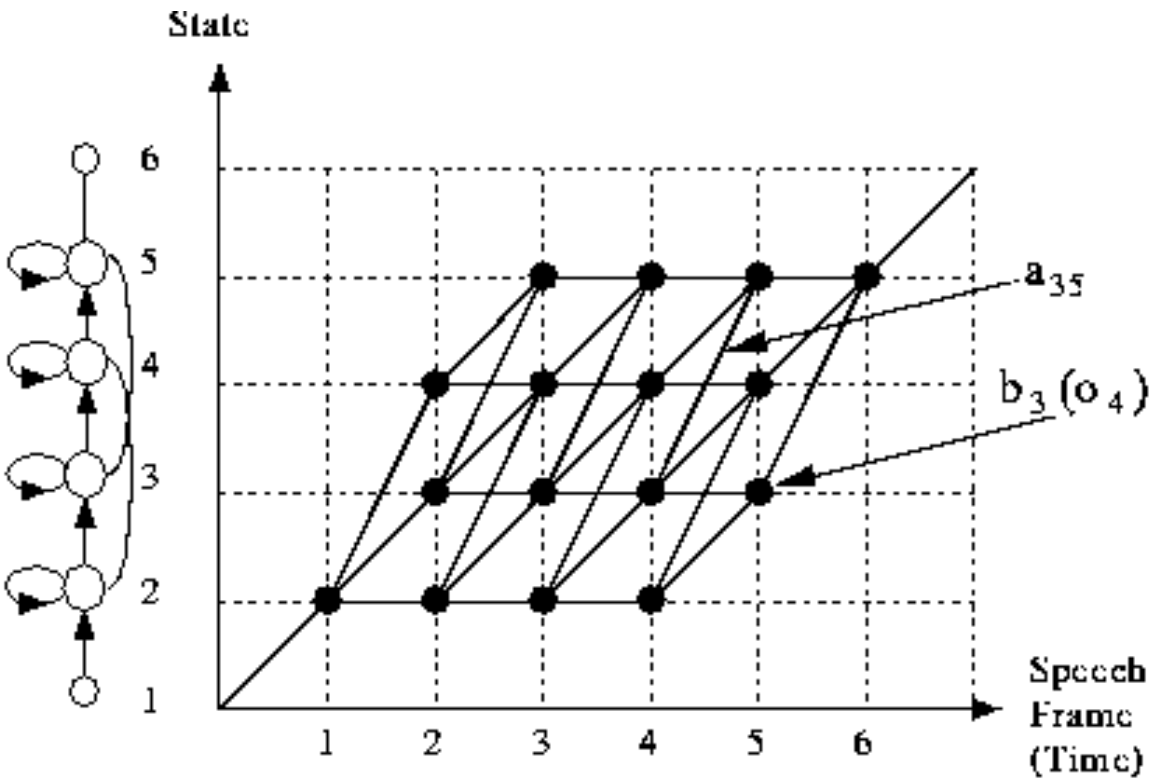


Fig. 1.6 The Viterbi Algorithm for Isolated Word Recognition

- Find the best path through the model
- http://crg7.uwaterloo.ca/software/HTKBook/node8.html

Beam search

- Great! Unfortunately...
  - With a large vocabulary, the state space is very large
  - Only expand “good” states: states with probabilities that are already high. (beam search.)

What can you do with speech understanding?

- Save money
  Credit card numbers (AT&T)
  How may I help you? (AT&T)
  Train schedules (Phillips)
- Help make money
  Stock quotes over the phone (Schwab)
- Save lives
  Voice-activated car phone dialing
- Help make a better life
  Learning to read
  Learning a foreign language
  Home automation
  ($30 voice-activated
  light switch)

http://www.dynamic-living.com/voice_activated_switch.htm

Still a lot of work!

- Given money and time,
  a company can build you a good dialog system
  AT&T, Dragon, IBM, Lernout & Hauspie, Lucent, Nuance, Phillips, SpeechWorks, ...
- Given less money and less time, you can build a simple dialog system
  Oregon Graduate Institute spoken dialog toolkit
  CMU Sphinx II speech recognizer
  -- open source from SourceForge
  Microsoft Agent
- Less gets re-used than we would like
- Still difficult for non-experts (Not as simple as making your own web page)
- Better accuracy, robust to noisy rooms, spontaneous speech, ...