Speech Recognition

Speech recognition is hard…

Inherent ambiguity

"It's not easy to wreck a nice beach"

vs

"It's not easy to recognize speech"

Pronunciation differences

"You say tomato, I say tomato"

In-user variability

How consistent are you, really?

Not to mention…

accents

Colloquialisms "jeet yet" "init"

noise

Homophonous Phrases

- "I scream, you scream, we all scream for ice cream!”
- “The boys are hoarse” and “The boy's a horse”
- “outstanding in the field” and “out standing in the field”
- “The good can decay many ways.” and “The good candy came anyways.”
- “The stuffy nose can lead to problems” and “The stuff he knows can lead to problems.”
- “Some others I've seen.” and “Some mothers I've seen.”
- “Oh, say! can you see by the dawn's early light.” “Jose can you see by the donzerly light?”

Speech Recognition Tasks

- Phones
- Words
- Sentences
- Meaning

Speech recognition as probabilistic inference

Goal: find the most likely word sequence, given a sound signal

$$P(words \mid signal) = \alpha P(signal \mid words) P(words)$$

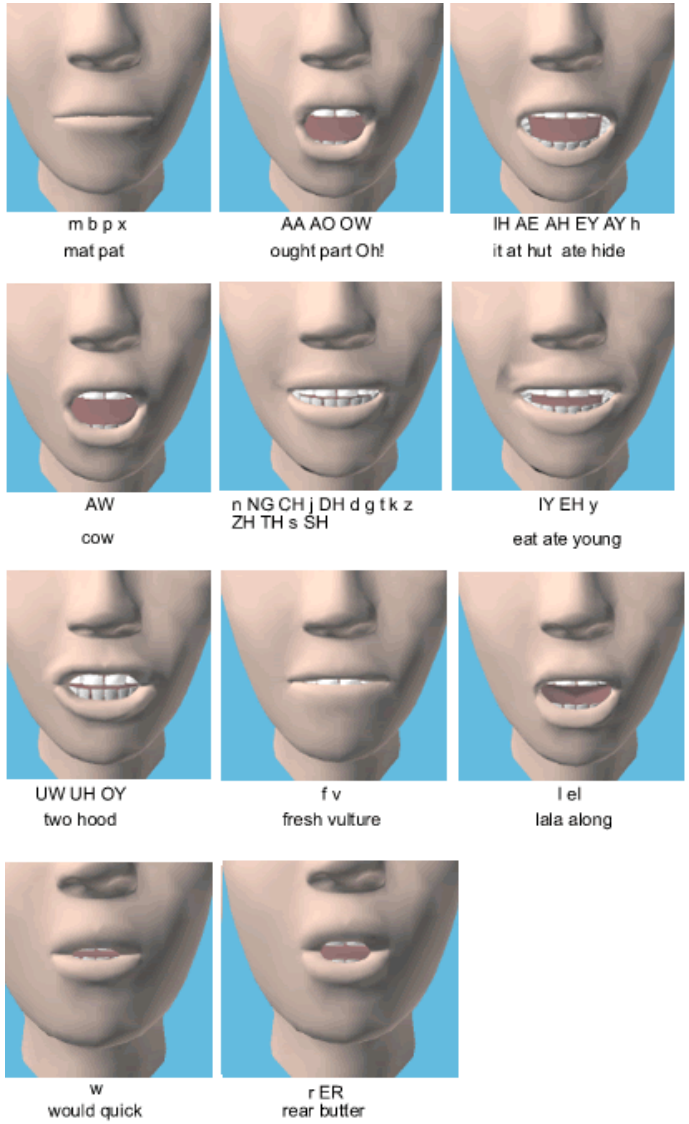
Acoustic mode Language model

Phones

Phones are the smallest unit of language

English is derived from 40-50 phones

Phone determined by configuration of articulators (lips, teeth, tongue, etc.)



Phones

Phones are the smallest unit of language

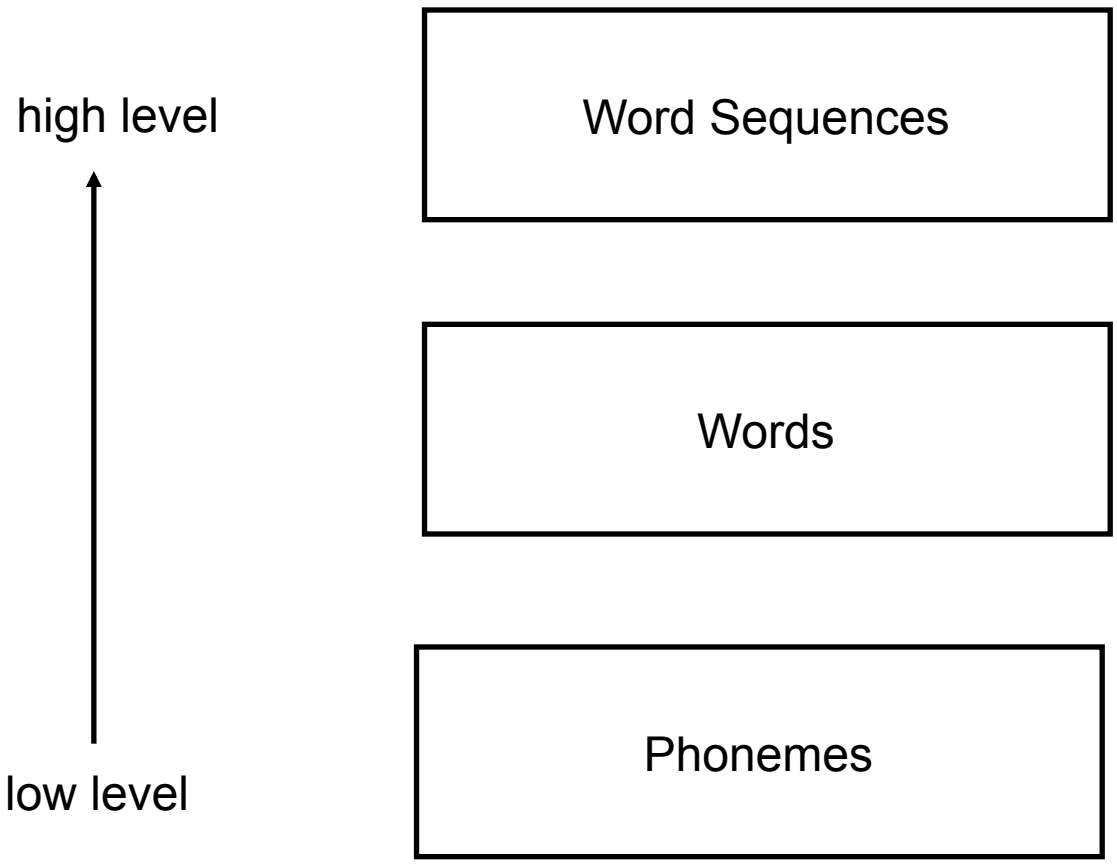
English is derived from 40-50 phones

Phone determined by configuration of articulators (lips, teeth, tongue, etc.)

[iy]	beat	[b]	bet	[p]	pet
[ih]	bit	[ch]	Chet	[r]	rat
[ey]	bet	[d]	debt	[s]	set
[ao]	bought	[hh]	hat	[th]	thick
[ow]	boat	[hv]	high	[dh]	that
[er]	Bert	[l]	let	[w]	wet
[ix]	roses	[ng]	sing	[en]	button
⋮	⋮	⋮	⋮	⋮	⋮

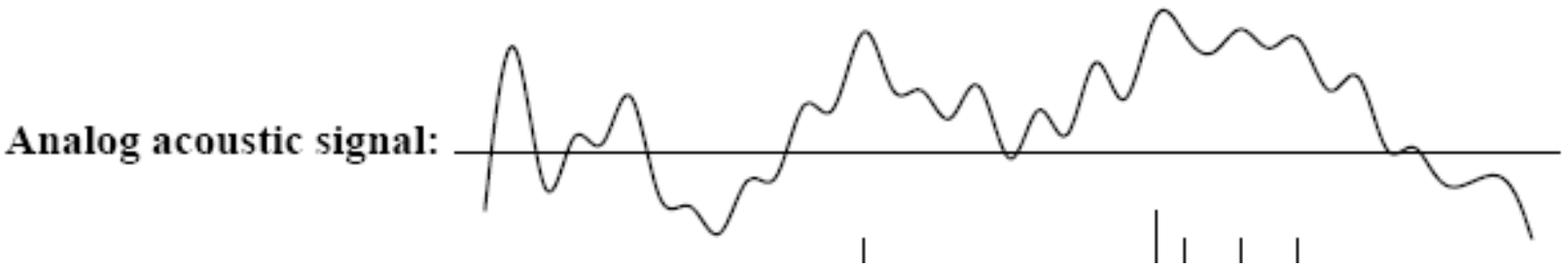
E.g., "ceiling" is [s iy l ih ng] / [s iy l ix ng] / [s iy l en]

Speech Recognition: Task Layers



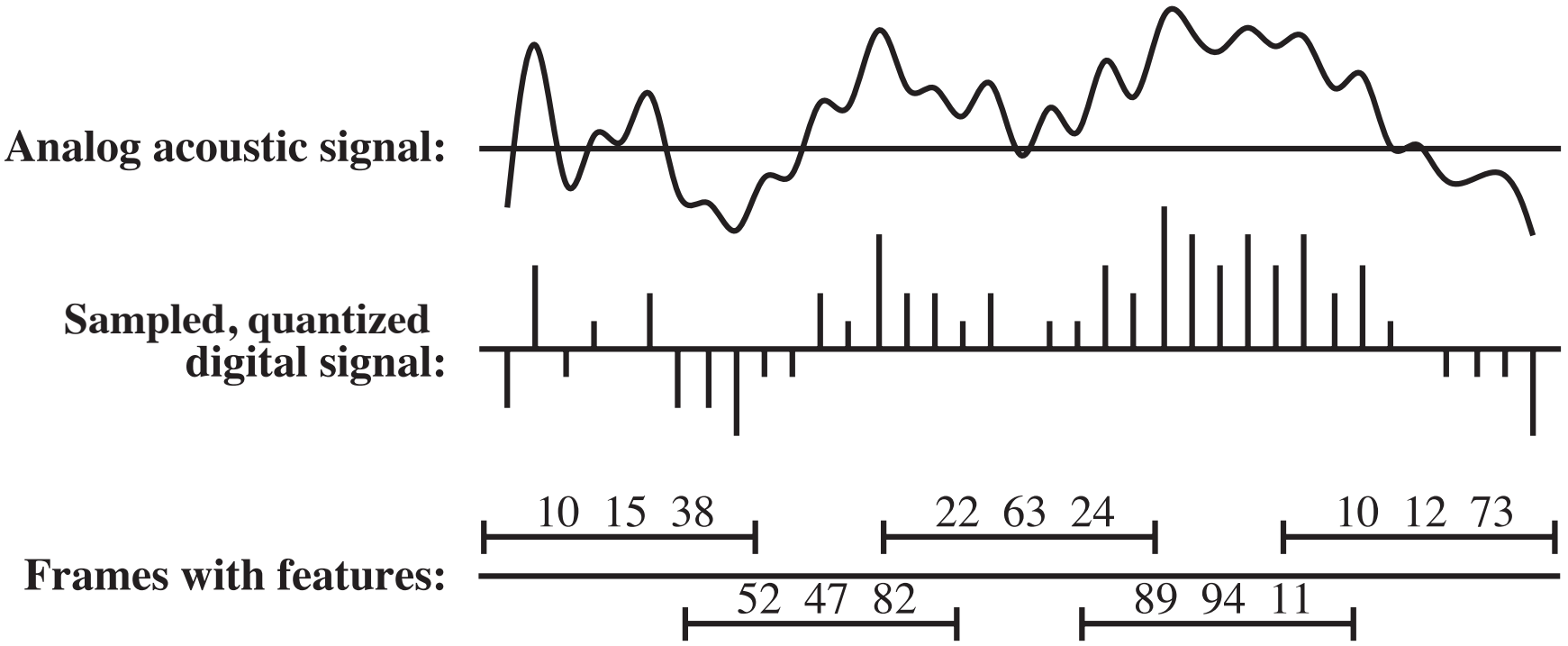
Speech Signal

- How do we recognize phones?
  - What is a speech signal anyway?



Speech Signal

Raw signal is the microphone displacement as a function of time; processed into overlapping 10ms frames, each described by features



Phone model

- Problem #1: Feature space
  - Say there are *n* features, each of which has 256 values… what is the problem here?

Phone model

- Problem #1: Feature space
- Solution:
  - Vector quantization: Divide feature space into some number of regions, named C1…C255, for example

Phone model

- Problem #2: Phone state

  Say the phone [t] a few times…
  what do you notice about the sound you produce?

- Solution:

  Three State Model

Phone Model

- Problem #3: Co-articulation

  Say the words "soft" and "sweet" a few times
  paying specific attention to the shape of your mouth…
  What do you notice about the "s" sound?

- Solution:

  Tri-Phone Model

Phone Model

- Problem #3: Co-articulation

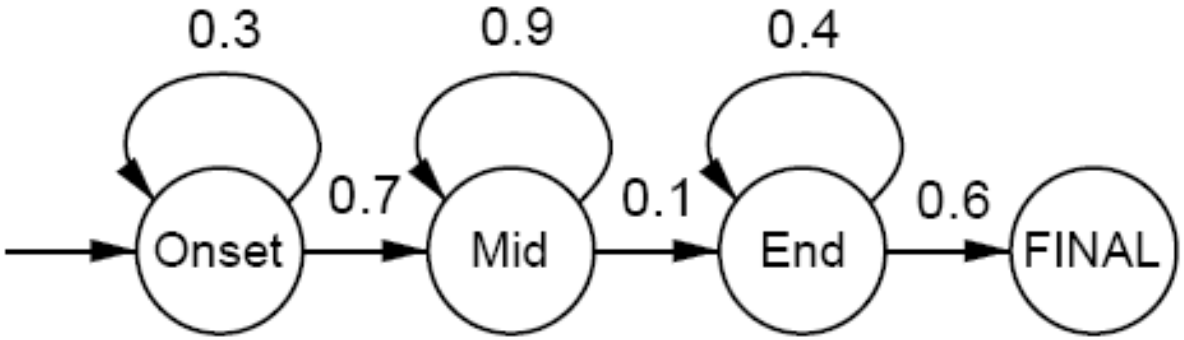
  Say the words "soft" and "sweet" a few times
  paying specific attention to the shape of your mouth…
  What do you notice about the "s" sound?

- Solution:

Implications of triphone + 3-state model ➔ increases state space to 3n * n^2 = 3n^3

Phone Model Example

Phone HMM for [m]:



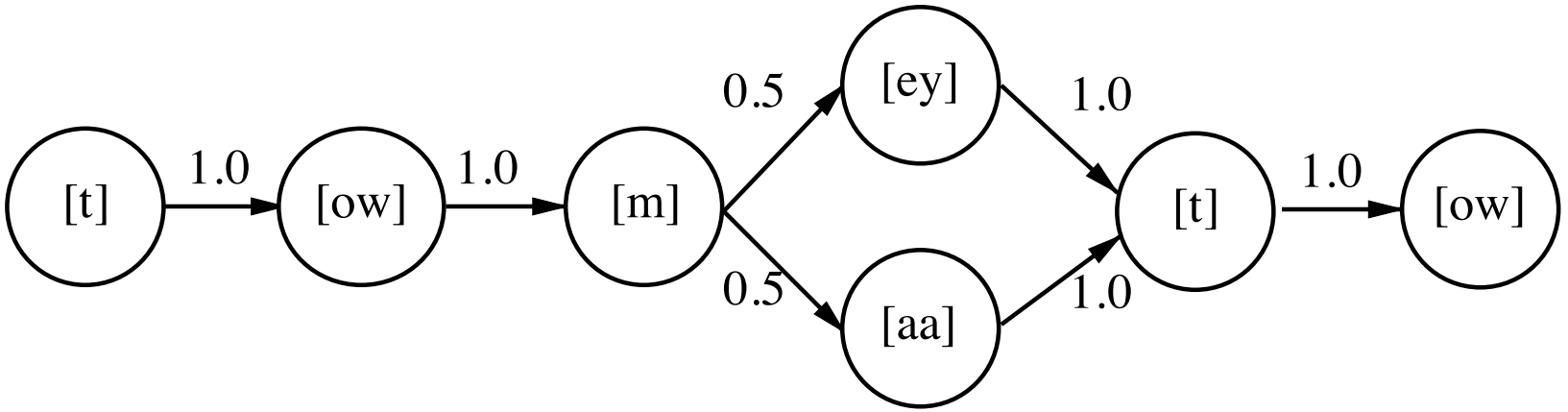
Output probabilities for the phone HMM:

Onset:	Mid:	End:
C1: 0.5	C3: 0.2	C4: 0.1
C2: 0.2	C4: 0.7	C6: 0.5
C3: 0.3	C5: 0.1	C7: 0.4

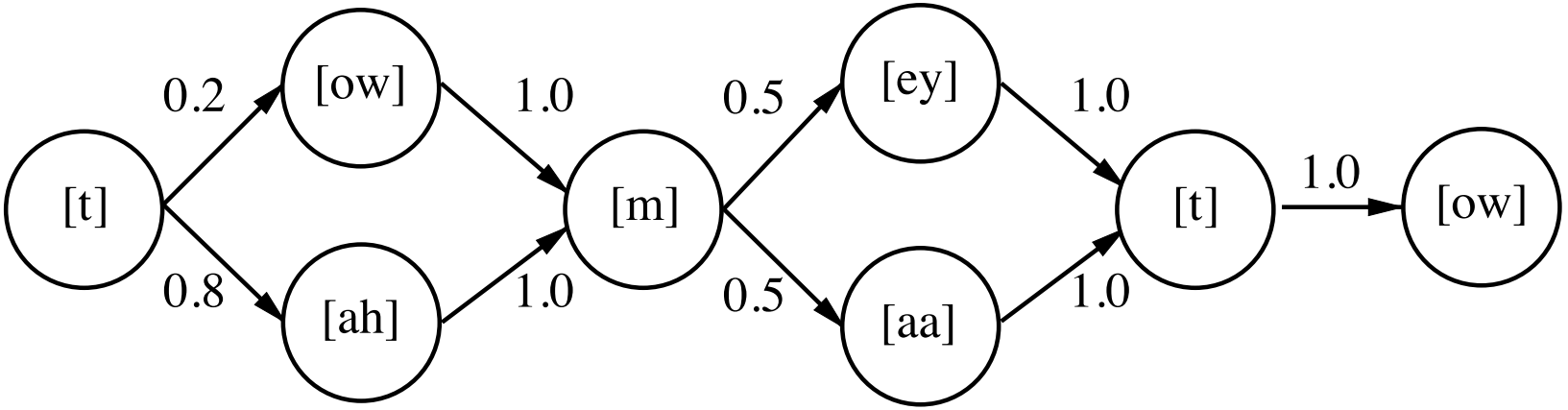
Modeling words

- Given phones, how could you model words?

(a) Word model with dialect variation:

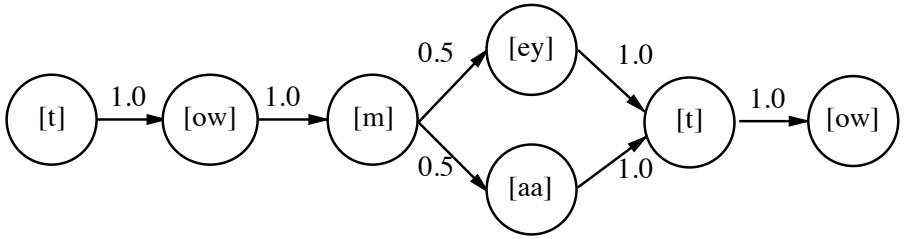


(b) Word model with coarticulation and dialect variations

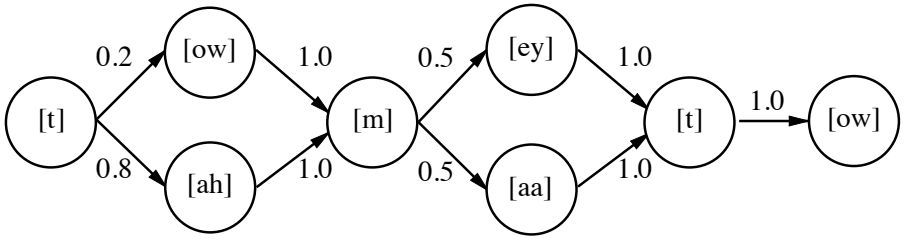


Where do these models come from?

(a) Word model with dialect variation:



(b) Word model with coarticulation and dialect variations



Isolated Words

- Phone models + word models give $P(e_{1:t}|word)$ for isolated word

$$P(word \mid e_{1:t}) = \alpha P(e_{1:t} \mid word) P(word)$$

How do you find P(word)?

Continuous Speech

- A sequence of isolated word recognitions?

- Challenges:
  - The sequence of most likely words is not the most likely sequence of words
  - segmentation

Language Model

- What is the prior probability of a sequence of words? $P(w_1 \ldots w_n) =$

Word	Unigram count	Previous words							
		of	in	is	on	to	from	model	agent
the	33508	3833	2479	832	944	1365	597	28	24
on	2573	1	0	33	2	1	0	0	6
of	15474	0	0	29	1	0	0	88	7
to	11527	0	4	450	21	4	16	9	82
is	10566	3	6	1	4	2	1	47	127
model	752	8	1	0	1	14	0	6	4
agent	2100	10	3	3	2	3	0	0	36
idea	241	0	0	0	0	0	0	0	0

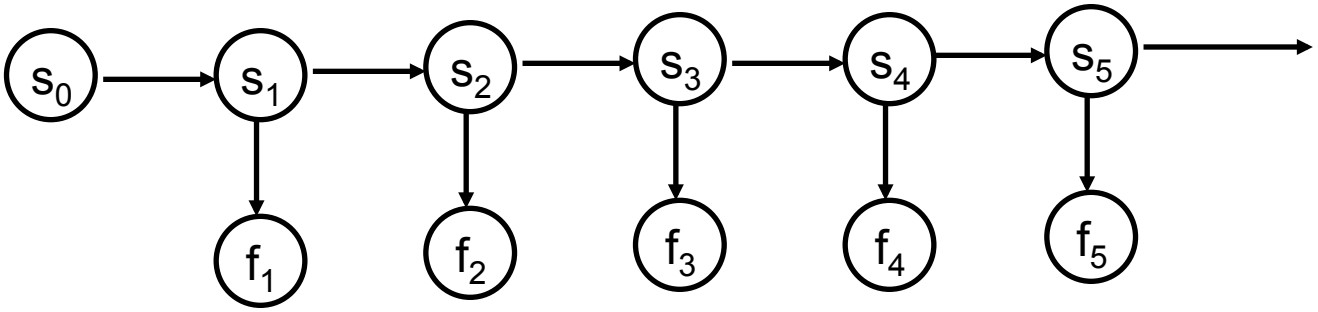
Figure 15.21 A partial table of unigram and bigram counts for the words in this book. "The" is the most common single word with a count of 33,508 (out of 513,893 total words). The bigram "of the" is the most common, at 3,833. Some counts are higher than expected (e.g. 4 for "on is") because the bigram counts ignore punctuation: one sentence might end with "on" and the next begin with "is."

Putting it all together…

- What are the states in the combined model?

- How many states are there?

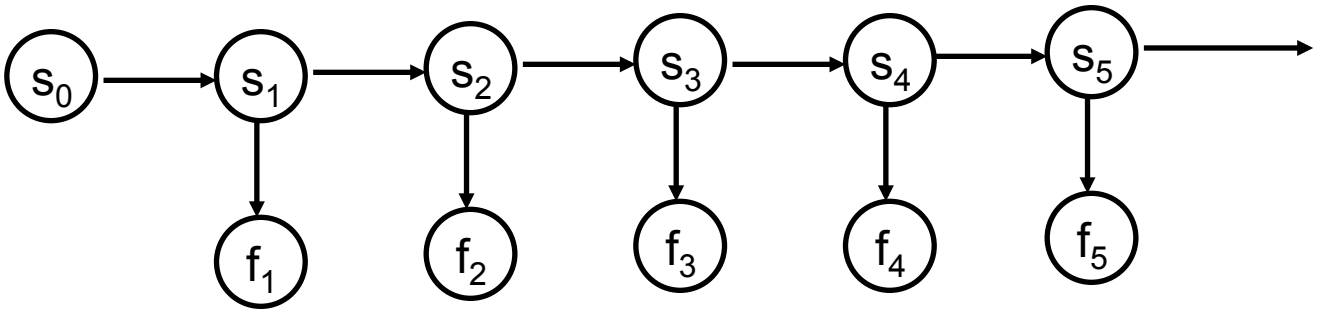
- How do we find the most likely sequence of words?



Inference Tasks

- Filtering: $P(X_t|e_{0:t})$
  - Decision making in the here and now
- Prediction: $P(X_{t+k}|e_{0:t})$
  - Trying to plan the future
- Smoothing: $P(X_k|e_{0:t})$ for $0<=k<t$
  - "Revisionist history" (essential for learning)
- Most Likely Explanation (MLE): $\text{argmax}_{x1:t}P(x_{1:t}|e_{1:t})$
  - e.g., speech recognition

Finding the most likely sequence of words



Find most likely sequence of states, then map to words

State of the Art?

- IBM's Via Voice
- DragonFly Naturally Speaking
- Integrated into Widows Vista… well sort of
- http://www.youtube.com/watch?v=kX8oYoYy2Gc

Applications of Speech Recognition

- IChef
- News at Seven on Stage
- Jackie