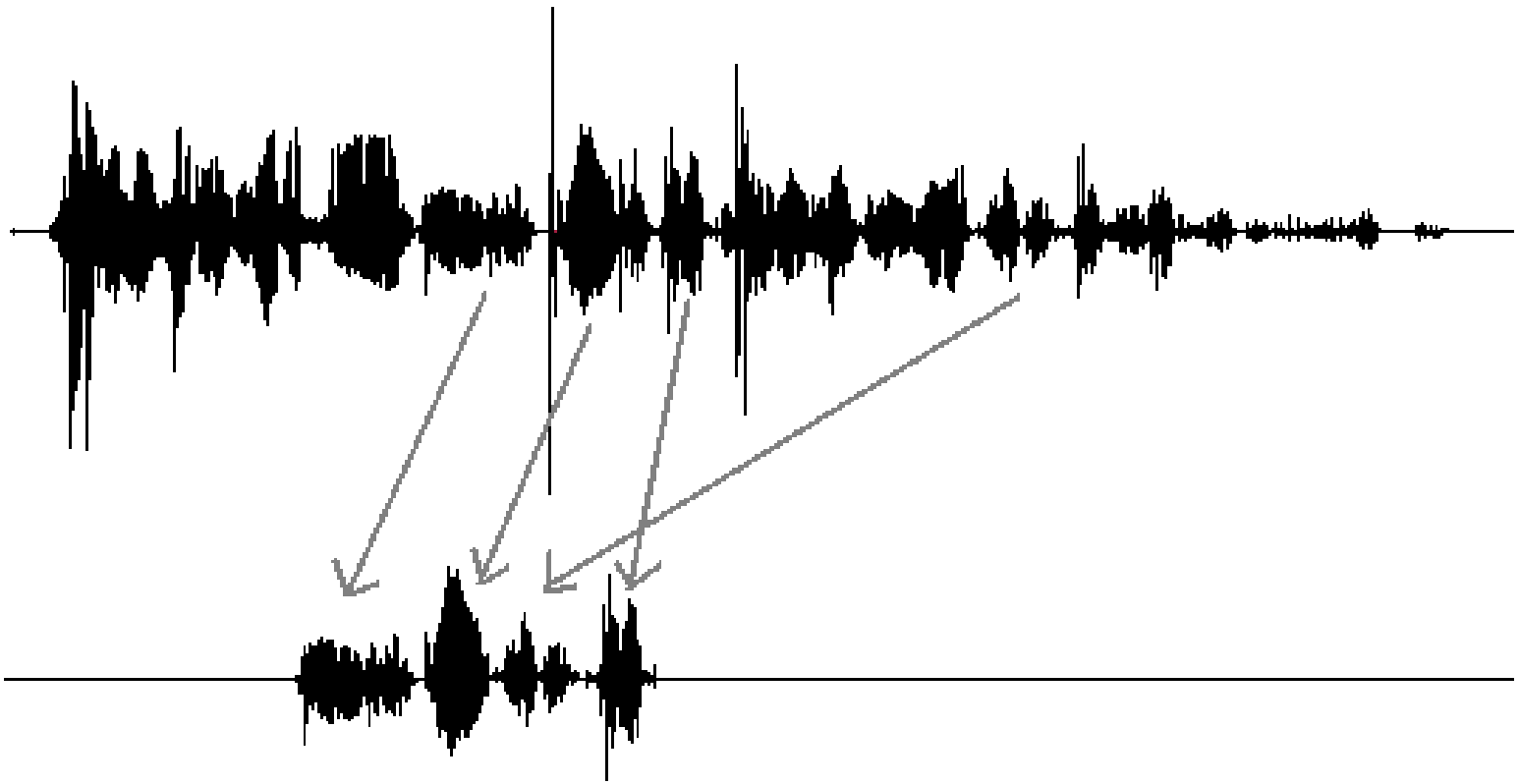
HMM-Based Speech Synthesis

Erica Cooper

CS4706

Spring 2011

Concatenative Synthesis

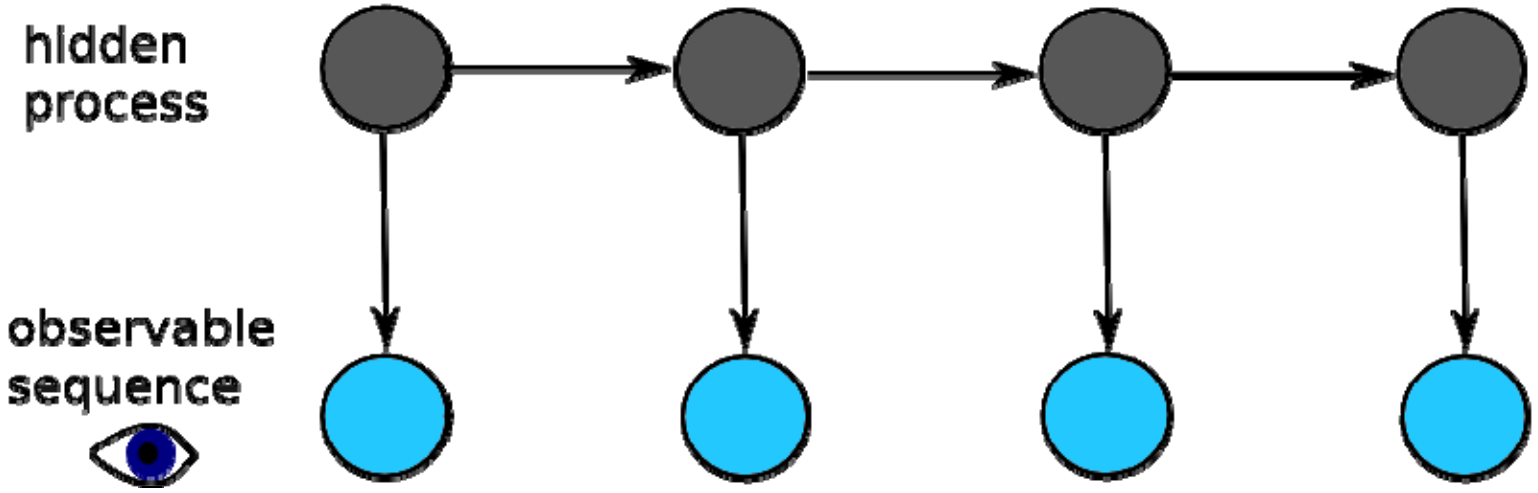


HMM Synthesis

- A parametric model
- Can train on mixed data from many speakers
- Model takes up a very small amount of space
- Speaker adaptation

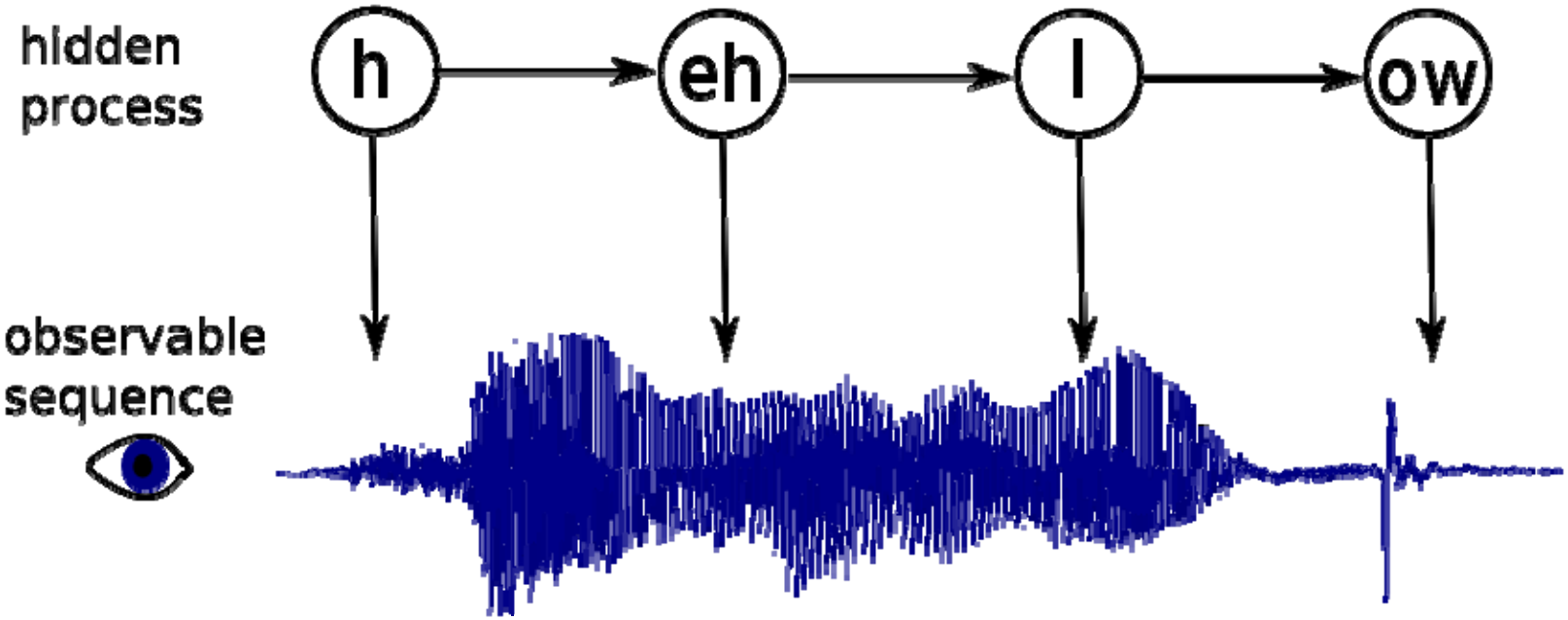
HMMs

- Some hidden process has generated some visible observation.



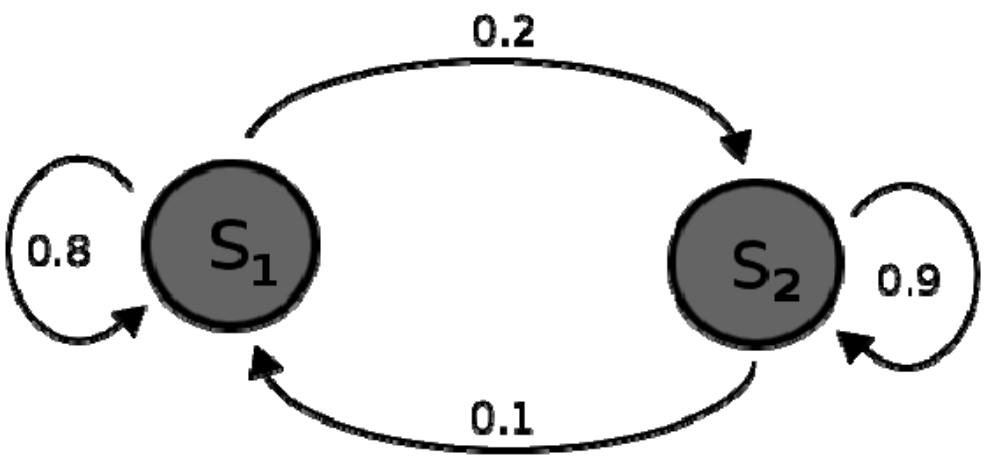
HMMs

- Some hidden process has generated some visible observation.



HMMs

- Hidden states have transition probabilities and emission probabilities.



	S_1	S_2
e_1	0.8	0.1
e_2	0.1	0.2
e_3	0.1	0.7

HMM Synthesis

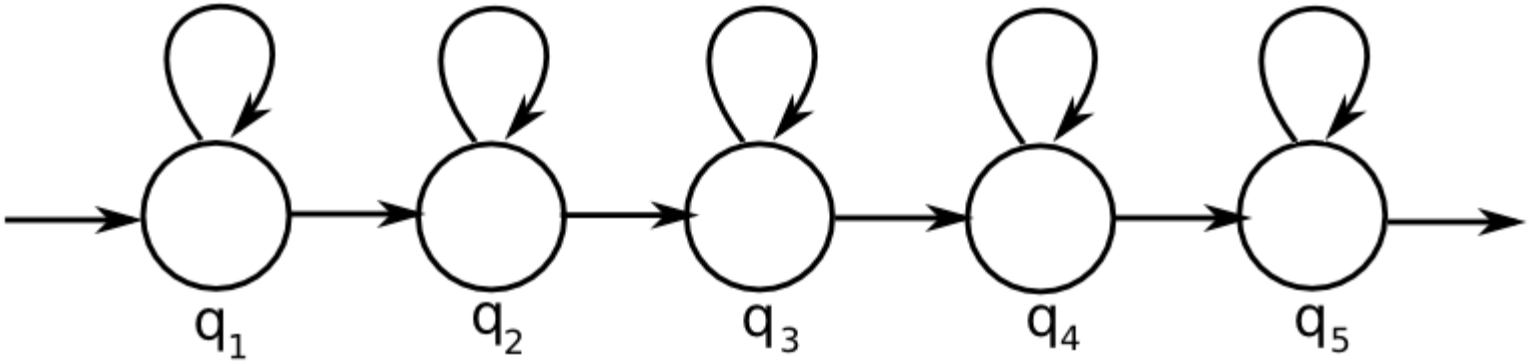
- Every phoneme+context is represented by an HMM.

  The cat is on the mat.
  The cat is near the door.

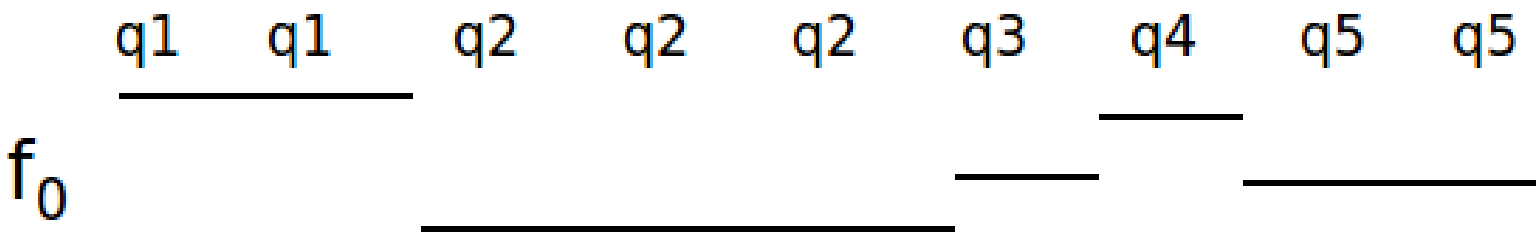
  < phone=/th/, next_phone=/ax/, word='the',
  next_word='cat', num_syllables=6, ... >

- Acoustic features extracted: f0, spectrum, duration
- Train HMM with these examples.

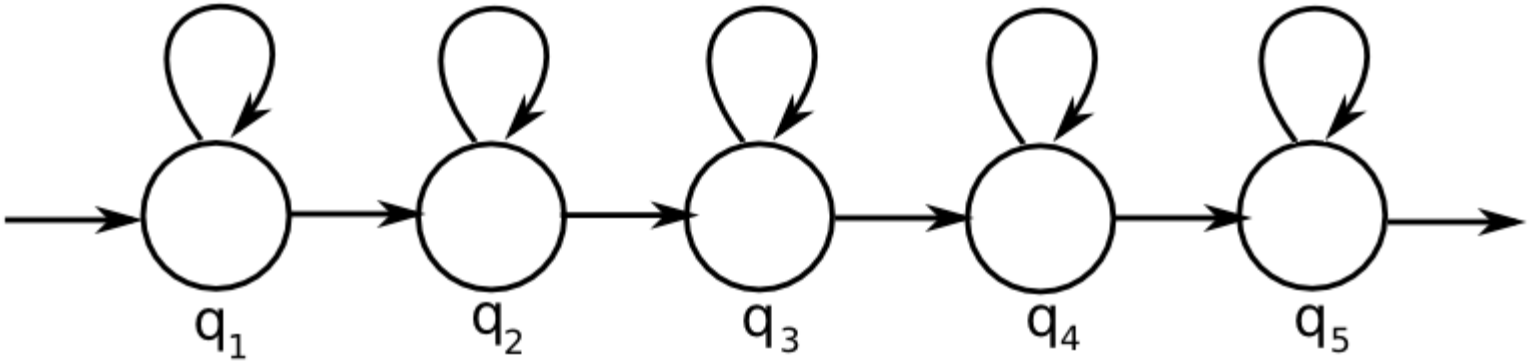
HMM Synthesis



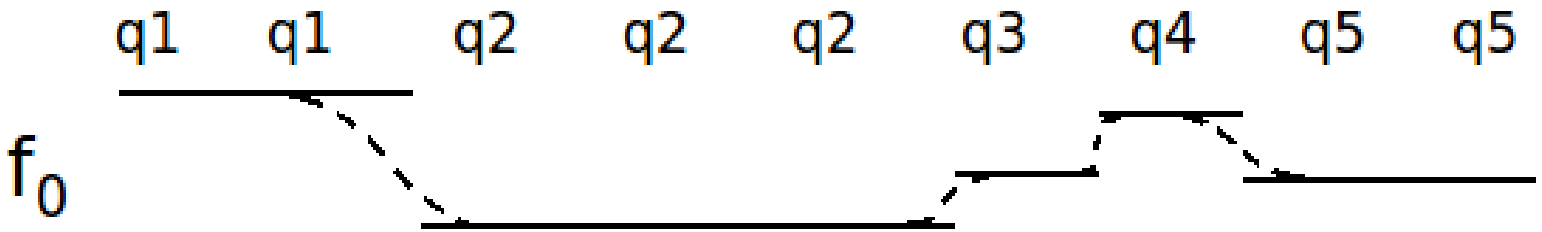
- Each state outputs acoustic features (a spectrum, an f_0, and duration)



HMM Synthesis



- Each state outputs acoustic features (a spectrum, an f_0, and duration)



HMM Synthesis

- Many contextual features = data sparsity
- Cluster similar-sounding phones
- e.g: 'bog' and 'dog'
  the /aa/ in both have similar acoustic features, even though their context is a bit different
- Make one HMM that produces both, and was trained on examples of both.

Experiments: Google, Summer 2010

- Can we train on lots of mixed data? (~1 utterance per speaker)
- More data vs. better data

- 15k utterances from Google Voice Search as training data

`ace hardware rural supply`

More Data vs. Better Data

- Voice Search utterances filtered by speech recognition confidence scores

50%, 6849 utterances

75%, 4887 utterances

90%, 3100 utterances

95%, 2010 utterances

99%, 200 utterances





Future Work

- Speaker adaptation
- Phonetically-balanced training data
- Listening experiments
- Parallelization
- Other sources of data
- Voices for more languages

Reference

- http://hts.sp.nitech.ac.jp