More about HMMs used for speech recognition

Steven Wegmann
ICSI

4 April 2012

HMM-based speech recognition has been the dominant methodology for nearly 25 years

The main HMM assumptions are strong:

- Conditional independence of observations
- Parametric form for output distributions

Strong assumptions: good practical consequences

- Model estimation tractable: Baum-Welch algorithm
- Inference tractable: Forward + Viterbi algorithms

Strong assumptions: bad practical consequence

- Speech data do not satisfy these assumptions

The statistical approach that we use

Some notation:

- W = sequence of words (transcription)
- X = sequence of frames (acoustic observations = cepstral features)

Given a particular acoustic utterance, x, our goal is to compute for every possible transcription w the probability

$$P(W = w \mid X = x)$$

We use these probabilities to *decode* or *recognize* the utterance x by selecting the most likely transcription w^{recog} via:

$$w^{recog} = \arg\max_{w} P(W = w \mid X = x)$$

The statistical approach that we use (cont'd)

We use Bayes' Theorem

$$P(w \mid x) = \frac{P(x \mid w)P(w)}{P(x)}$$

This decomposes the problem into two probability models

- The *acoustic model* gives $P(x \mid w)$
- The *language model* gives $P(w)$
- The term $P(x)$ is constant so it is (usually) ignored

The Viterbi algorithm is the key to decoding

We use hidden Markov models (HMM) for the acoustic model

Each word is expanded into a sequence of phonemes using a dictionary

Each phoneme is modeled using a HMM

- We actually model phonemes in context (eg. triphones)
- Roughly speaking, we group together collections of triphones and model the groups: eg. 64k possible triphones $\rightarrow$ 5k models

HMM: formal definition

A HMM consists of two synchronized stochastic processes

- An unobservable Markov chain of hidden states Q_t
- An observed process X_t

Each Q_t is a discrete random variable, while the X_t can be either a discrete or continuous random variable

The hidden chain 'explains' the observed process, because each Q_t emits X_t

HMM: formal definition (cont'd)

The hidden finite Markov chain Q_t:

- Takes its values in the states $\{1, \dots, L\}$.
- Has $L \times L$ transition matrix $A_\theta = (a_\theta(i,j) = p_\theta(q_j \mid q_i))$

Conditional independence:

- $\{X_t\}$ are conditionally independent given $\{Q_i\}$
- Given Q_t, X_t is independent of Q_j with $j \neq t$

Stationarity:

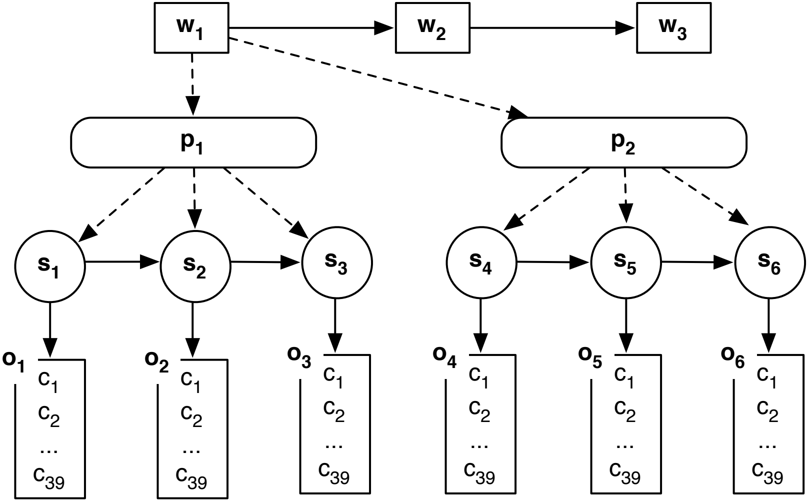
- The distribution of $X_t \mid Q_t = q_l$ does not depend on t
- We call these the *output distributions* for the states

HMM: formal definition (cont'd)

We showed that these assumptions imply

$$P(x) = \sum_{q} P(q, x) = \sum_{q} \prod_{t} P(x_t \mid q_t) P(q)$$

A depiction of the model



Notation: $s = q$ states, $o = x$ observations

Typical HMM transition structure