AUTOMATIC SPEECH RECOGNITION

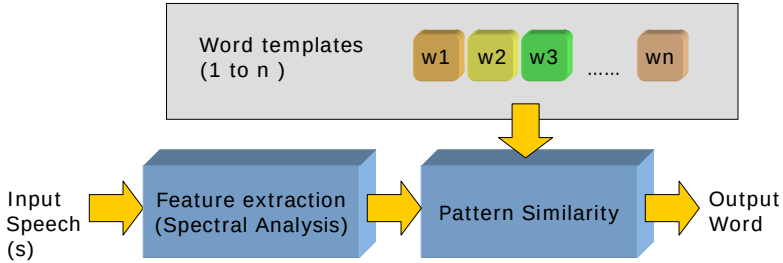
Lecture 8

Probability Review
Overview of Hidden Markov Models

This Lecture

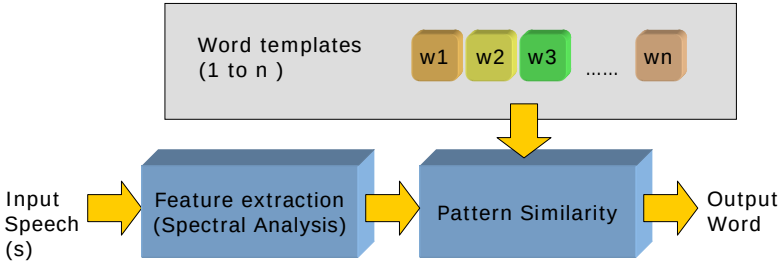
- Motivation
- Review of Probability
- Examples (somewhat) related to ASR
- Overview of HMMs

Word-based Template Matching Summary



1. Do spectral analysis to convert the input word to feature vectors
2. Compare the feature vectors for all the frames with each of the reference templates. Use DP matching to time align the patterns
3. Choose the output word which corresponds to the reference template with the minimum distance

How likely?



Suppose the distance between *S* and *W*i is *D*i.
*W*i yielding the minimal *D*i should be picked as the output word.

How likely that the input speech is *W*i?

To indicate how likely…

We need PROBABILITY !!!

A broad picture

Suppose we want to classify between two words, W1 and W2.

Feature vector: 1-dimensional … Let's say total Energy. 

(Obviously, a bad one)

Training Phase

Learn the value distribution of E from many examples of W1
→ building a model of E for W1

Learn the value distribution of E from many examples of W2
→ building a model of E for W2

Recognizing (Decoding) Phase

Compare Ex, the total energy of the word (X) to be recognized, with the two distributions.

Calculate the probabilities of X being W1 and X being W2. (Given that X must be either one of them)

Important Concepts

- Probability
- Conditional Probability
- Total probability theorem
- Bayes' Rule
- Random variable (R.V.)
  - discrete R.V.
  - continuous R.V.
- Expected Value and Variance
- Gaussian Random Variable
- Joint PDF

Probability

- A : any event
- P(A) : probability that the event A happens
- $0 \le P(A) \le 1$

Conditional Probability

- P(A|B) = probability of A, given that B has occurred.

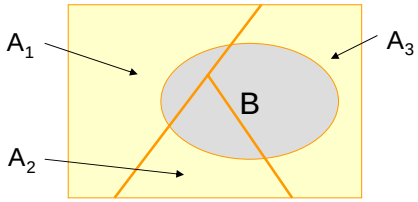
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Adjusting the universe to B

e.g. After finishing an ASR homework, a student has gone crazy and yell out one of the four cardinal vowels randomly.
P(the vowel is /i/) =?
Given that he/she yells out a front vowel,
P(/i/ | front vowel) =?

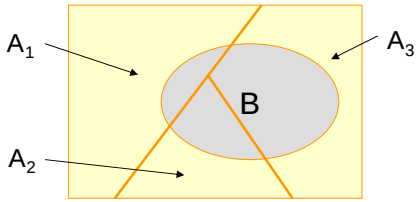
Total Probability Theorem

- Divide the universe into smaller partitions



$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

Bayes' Rule



$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(A_j)P(B \mid A_j)}$$

$P(A_i)$: "Prior" probability

Random Variable

- X : Random variable (R.V.)
- x : experimental value of the R.V. X

X	x	
Type of a vowel V	{/a/,/i/,...}	Discrete
# of syllable in a word W	{1,2,3,4,5,...}	Discrete
The first formant frequency	(0,Fs/2)	Continuous
$\log_{10}$ Energy of a signal	$(-\infty,\infty)$	Continuous
Probability of event A	[0,1]	Continuous

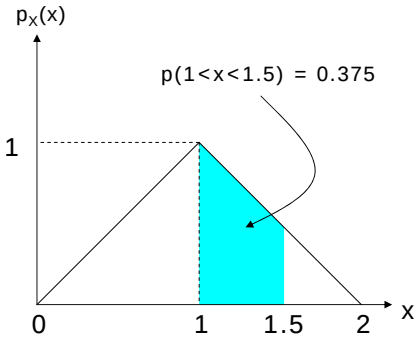
Probability Density Function (PDF)

- $p_X(x)$ is a density of probability measure on the event space for random variable X

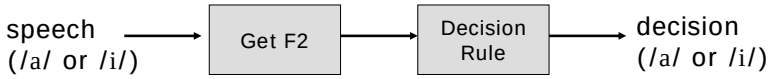
$$P(a < x \le b) = \int_a^b p(x)dx$$

$$P(-\infty < x \le \infty) = \int_{-\infty}^{\infty} p(x)dx = 1$$

$$P(x = a) = 0$$

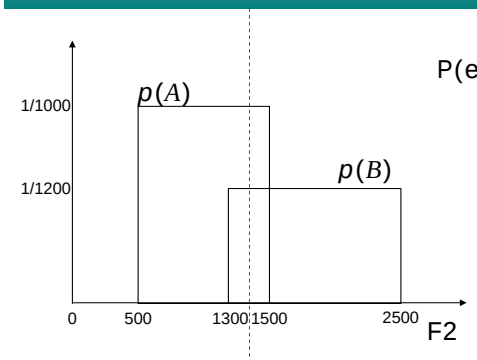


Example



- Let
  - A= F2 of /a/
  - B= F2 of /i/
- Suppose:
  - A uniformly distributes between 500 and 1500 Hz
  - B uniformly distributes between 1300 and 2500Hz
- We want to classify an incoming vowel by the following rule:
  - if (F2 >1400Hz) choose /i/; else choose /a/;
- If both vowels are equally likely to be produced, what is P(error)?

Example



P(err) = P(choose /a/ | /i/)P(/i/)
+P(choose /i/ | /a/)P(/a/)

= P(F2<1400 | /i/)x0.5
+P(F2>1400 | /a/)x0.5

= P(B<1400 | /i/)x0.5
+P(A>1400 | /a/)x0.5

= 100x(1/1200)x0.5
+100x(1/1000)x0.5

= 0.092

Expected Value and Variance

- Expected value

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx \qquad E[g(x)] = \int_{-\infty}^{\infty} g(x)\,p(x)dx$$

- Variance (σ^2) (Standard Deviation = σ)

$$Var[x] = E[(x - E[x])^2] = \sigma^2 = \int_{-\infty}^{\infty} (x - E[x])^2\, p(x)dx$$

$$E[(x - E[x])^2] = E[x^2] - (E[x])^2$$

Expected Value

- E[a] = a; a is a constant.
- E[aX+b] = aE[X]+b
- E[X+Y] = E[X]+E[Y]
- Var[a] = 0
- Var[aX+b] = a^{2}Var[X]

Conditional Expected Value

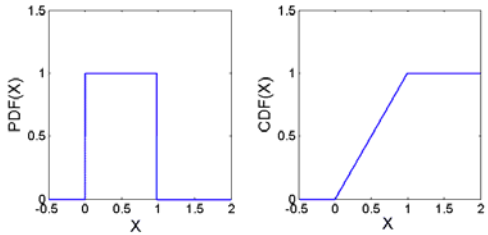
$$E[x \mid A] = \int_{-\infty}^{\infty} x p(x \mid A) dx$$

$$E[g(x) \mid A] = \int_{-\infty}^{\infty} g(x) p(x \mid A) dx$$

Cumulative Distribution Function

- The Cumulative Distribution Function (CDF) $p_{x\leq}$ provides the probability $P(X \leq x_o)$of the event that the value of an R.V. X does not exceed x_o

$$p_{x\leq}(x_o) = \int_{-\infty}^{x_o} p(x) dx$$



Gaussian (Normal) Random Variable

- X is normal (Gaussian): $X \sim N(\mu,\sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

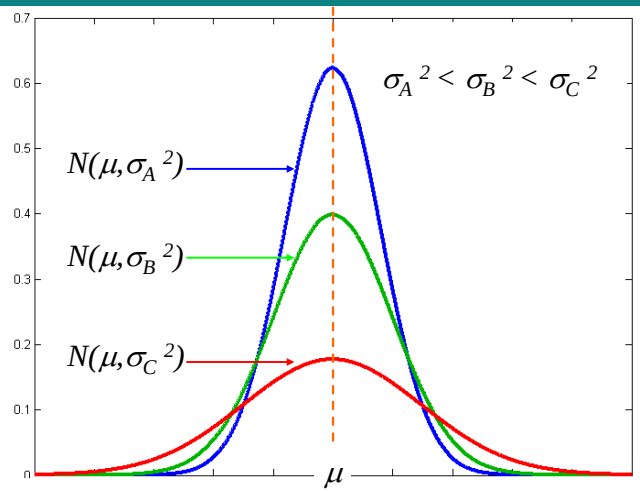
$$E[x] = \mu$$
$$Var[x] = \sigma^2$$

- X is Standard normal (Standard Gaussian): $X \sim N(0,1)$ when $\mu=0$, $\sigma^2=1$

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}$$

$$E[x] = 0$$
$$Var[x] = 1$$

Normal PDF



Standard Normal Variable

- Normality is preserved by linear transformation. Calculation involving the normal variable is usually done in terms of standard normal.
- Let $Y=aX+b$,
  if $X\sim N(\mu,\sigma^2)$ → $Y\sim N(a\mu+b,a^2\sigma^2)$
- Let $Z=(X-\mu)/\sigma$,
  if $X\sim N(\mu,\sigma^2)$ → $Z\sim N(0,1)$: Standard Normal

Standard Normal CDF

- X is a normal R.V.
- If we want to find $P(X \le x_o)$, convert to a standard normal Z and then use Standard Normal CDF. (tabulated in the form of Φ)

$$P(X \le x_o) = P(\sigma Z + \mu \le x_o) = P(Z \le \frac{x_o - \mu}{\sigma}) = \Phi(\frac{x_o - \mu}{\sigma})$$

- $\Phi(z_o)$ is the total area under the standard normal PDF curve from $-\infty$ to z_o

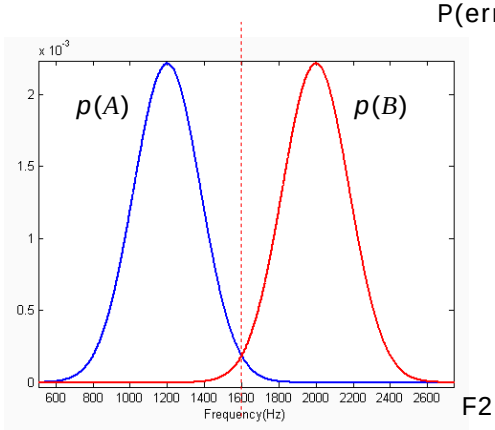
$$\Phi(z_o) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{zo} e^{\frac{-t^2}{2}} dt$$

Example

speech (/ɔ/ or /ɤ/) → Get F2 → Decision Rule → decision (/ɔ/ or /ɤ/)

- Let
  - A= F2 of /ɔ/
  - B= F2 of /ɤ/
- Suppose:
  - A is normal with mean 1200Hz and SD 180
  - B is normal with mean 2000 Hz and SD 180
- We want to classify an incoming vowel by the following rule:
  - if (F2 >1600Hz) choose /ɤ/; else choose /ɔ/;
- If both vowels are equally likely to be produced, what is P(error)?

Example



P(err) = P(choose /a/ | /i/)P(/i/)
+P(choose /i/ | /a/)P(/a/)

= P(F2<1600 | /i/)x0.5
+P(F2>1600 | /a/)x0.5

= (1-Φ((1600-1200)/180))x0.5
+ (1-Φ((2000-1600)/180))x0.5

= (1-Φ(2.2)) x0.5
+ (1-Φ(2.2)) x0.5

= 1-Φ(2.2) = 0.0139

Joint PDF

- Since we hardly use 1-dimensional feature vector, we mostly use joint PDFs to model the value distributions of multiple random variables.

Joint PDF

- $p_{XY}(x,y)$ is a joint density of probability measure on the event space for random variables X and Y

$$P(A) = \iint_A p_{XY}(x,y)dxdy$$

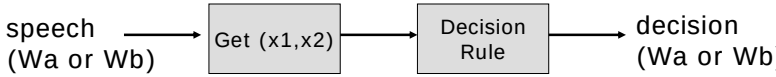
$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)dy$$

$$p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x,y)dx$$

Independence

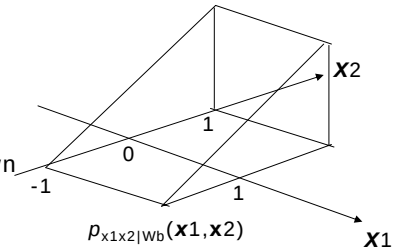
- When $p(X|Y) = p(X)$, it is said that the R.V. X is independent of Y. This leads to $p(XY)=p(X)p(Y)$
- E.g.:
  - A speaker says a phrase with two words
  - X is the duration in ms. of the first word.
  - Y is the duration in ms. of the second word.
  - if we do not know the total duration, knowing Y does not affect our knowledge of X
    - $p(X|Y) = p(X)$
  - Suppose we want to find P($X>500$,$Y<500$)→ multiply P($X>500$) and P($Y<600$) directly.

Example



- Vocab: Wa, Wb
- Feature vector: ($\boldsymbol{X}1=x1,\boldsymbol{X}2=x2$)
- Decision rule: Pick Wi giving maximal $p_{\boldsymbol{X}1\boldsymbol{X}2|Wi}(x1,x2|Wi)$
- Models: (from training data)
  - $p_{\boldsymbol{X}1|Wa}$ ~ Uniform over (-1,1)
  - $p_{\boldsymbol{X}2|Wa}$ ~ Uniform over (-1,1)

  (Given Wa, $\boldsymbol{X}1$ and $\boldsymbol{X}2$ are independent.)
  - $p_{\boldsymbol{X}1\boldsymbol{X}2|Wb}(\boldsymbol{X}1,\boldsymbol{X}2|Wb)$ ~ as shown



Example

- 4 test speech tokens:
  - (0.5,-0.3), (0.2,0,2), (-0.8,0.7), (-0.5,-0.5)
- What should the decisions be?

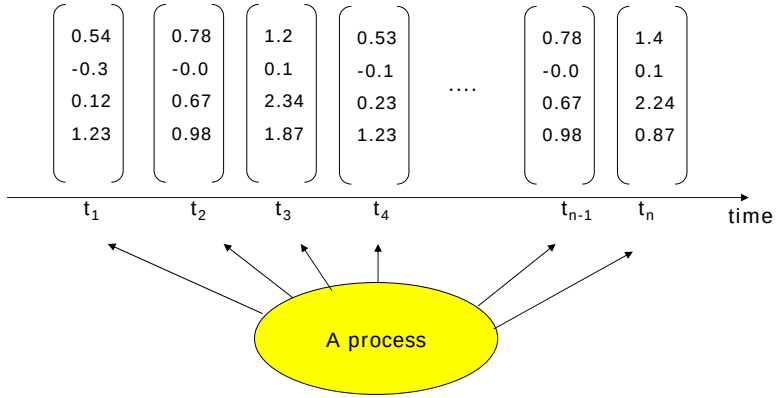
- What if the decision rule is changed to:

  Pick Wi giving maximal $p_{Wi|x1x2}$ (Wi|x1,x2)

  and we know that Wb occurs twice as often as Wa?

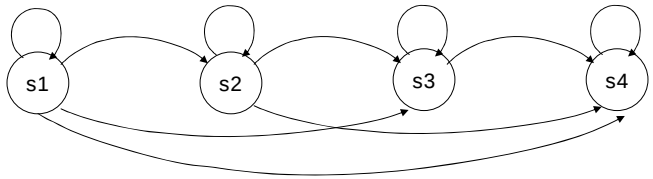
Hidden Markov Models

- HMMs are statistical models for processes that generate sequence of numbers.



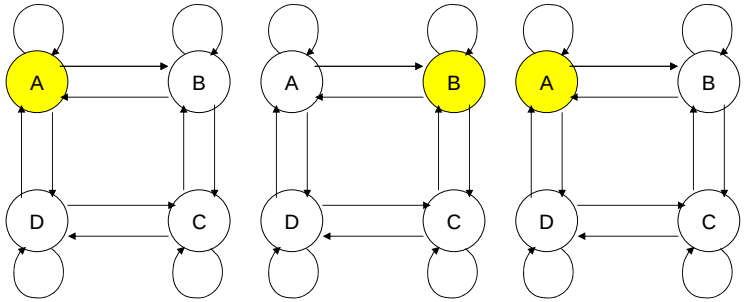
Hidden Markov Models

- Each process is described using a “Finite State Machine”.



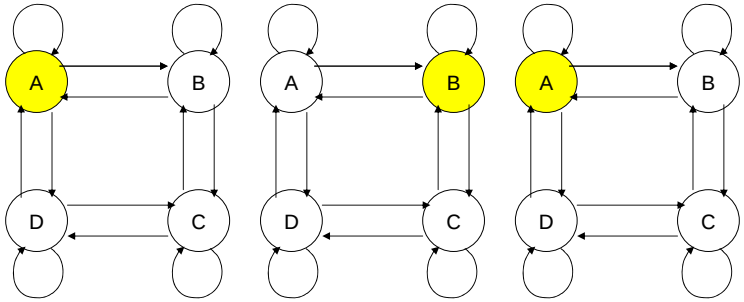
What is HMM?

- The model assumes that underlying process can be in one of a number of states at any time instant.



What is HMM?

- Let q_i = the state that the process is in at time i

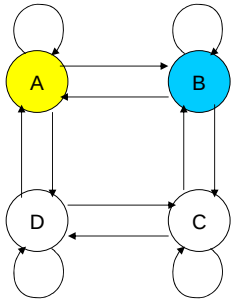


q_1 = 'A' q_2 = 'B' q_3 = 'C'

State sequence $Q = \{q_1,q_2,q_3,...,q_n\}$

What is HMM?

- The state that the process is in at any time instant is dependent only on the state it was in at the previous instant. (Markovian)



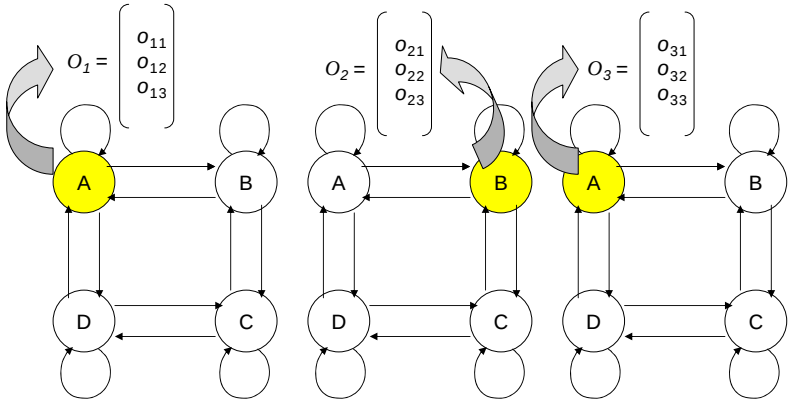
$$P(q_t \mid q_{t-1}) = P(q_t \mid q_{t-1}q_{t-2}q_{t-3}...q_0)$$

What is HMM?

- At each time instant the process generates an observation using probability distribution that is specific to the state that it is in.

What is HMM?

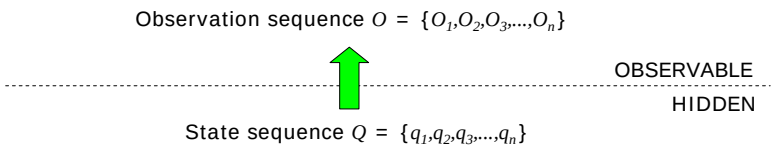
Observation sequence $O = \{O_1,O_2,O_3,...,O_n\}$



q_1 = 'A' q_2 = 'B' q_3 = 'C'

What is HMM?

- The generated observations are all that we can see.
- The actual state sequences that happened are not directly observable.

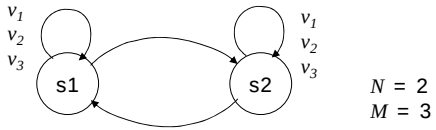


Discrete and Continuous Models

- HMMs can be classified into discrete models or continuous models
  - observable events assigned to each state are discrete, e.g. codewords after vector quantization
    → discrete model
  - observable events assigned to each state are continuous, e.g. many acoustic measurements without vector quantization
    → continuous model

Elements of an HMM

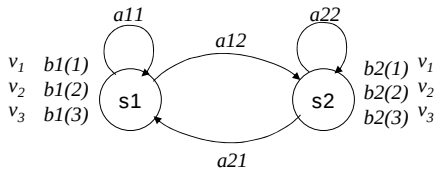
- N: number of states in the model
  - states, $\boldsymbol{S} = \{s_1,s_2,..s_N\}$
  - state at time t, $q_t \in \boldsymbol{S}$
- M: number of observation symbols (assume discrete observation)
  - observation symbols, $\boldsymbol{V} = \{v_1,v_2,...,v_M\}$
  - observation at time t, $o_t \in \boldsymbol{V}$



$N = 2$
$M = 3$

Elements of an HMM

- $\boldsymbol{A} = \{a_{ij}\}$: state transition probability distribution
  - $a_{ij} = \mathrm{P}(q_{t+1}=s_j|q_t=s_i);\ 1 \le i,j \le N$
- $\boldsymbol{B} = \{b_j(k)\}$: observation symbol probability distribution in state j
  - $b_j(k) = \mathrm{P}(v_k \text{ at } t|q_t=s_i);\ 1 \le j \le N,\ 1 \le k \le M$
- $\boldsymbol{\pi} = \{\pi_i\}$: initial state distribution
  - $\pi_i = \mathrm{P}(q_1=s_i);\ 1 \le i \le N$

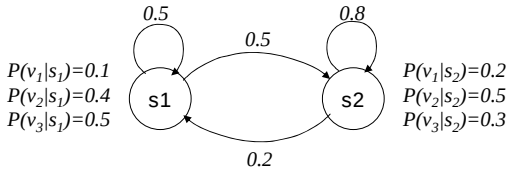


Example

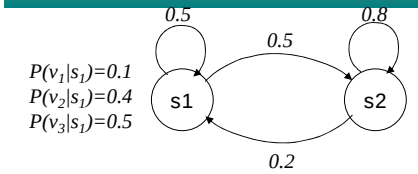
$$A = \begin{bmatrix} a11 & a12 \\ a21 & a22 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix}$$

$$B = \begin{bmatrix} b1(v1) & b1(v2) & b1(v3) \\ b2(v1) & b2(v2) & b2(v3) \end{bmatrix} = \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

$$\pi = \{a01, a02\} = \{0.5, 0.5\}$$



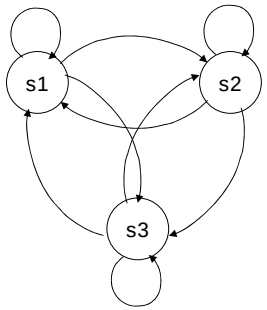
Example



Observe $O = \{v2, v1, v3\}, P(O \mid A, B, \pi) = ?$

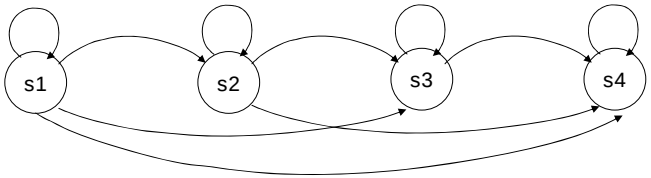
Possible state sequences, Q	$P(O \mid A, B, \pi, q)$
$\{s1, s1, s1\}$	$0.5\times0.4\times0.5\times0.1\times0.5\times0.5$
$\{s1, s1, s2\}$	$0.5\times0.4\times0.5\times0.1\times0.5\times0.3$
$\{s1, s2, s1\}$	$0.5\times0.4\times0.5\times0.2\times0.2\times0.5$
$\{s1, s2, s2\}$	$0.5\times0.4\times0.5\times0.2\times0.8\times0.3$
$\{s2, s1, s1\}$	$0.5\times0.5\times0.2\times0.1\times0.5\times0.5$
$\{s2, s1, s2\}$	$0.5\times0.5\times0.2\times0.1\times0.5\times0.3$
$\{s2, s2, s1\}$	$0.5\times0.5\times0.8\times0.2\times0.2\times0.5$
$\{s2, s2, s2\}$	$0.5\times0.5\times0.8\times0.2\times0.8\times0.3$

Ergodic Model



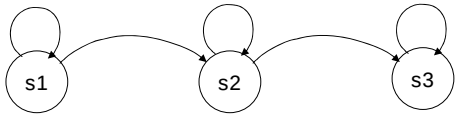
- An HMM is called ergodic or fully-connected model when every state of the model can be reached (in a single step) from every other state of the model.

Left-to-Right Model

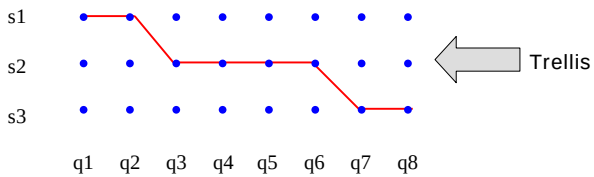


- The underlying state sequence associated with the model has the property that, as time increases, the state index increases.
- The left-to-right model exhibits the desirable property of being readily able to model speech whose properties change over time in a successive manner.

Representing State Diagram by Trellis

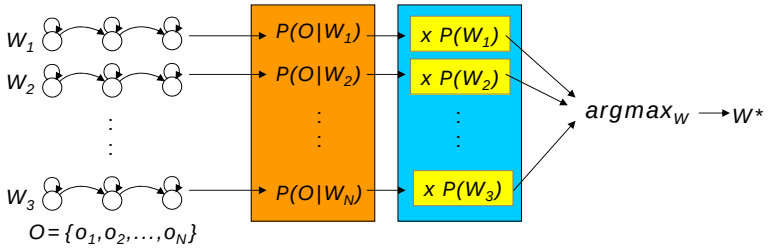


$Q = \{ s_1, s_1, s_2, s_2, s_2, s_2, s_3, s_3 \}$

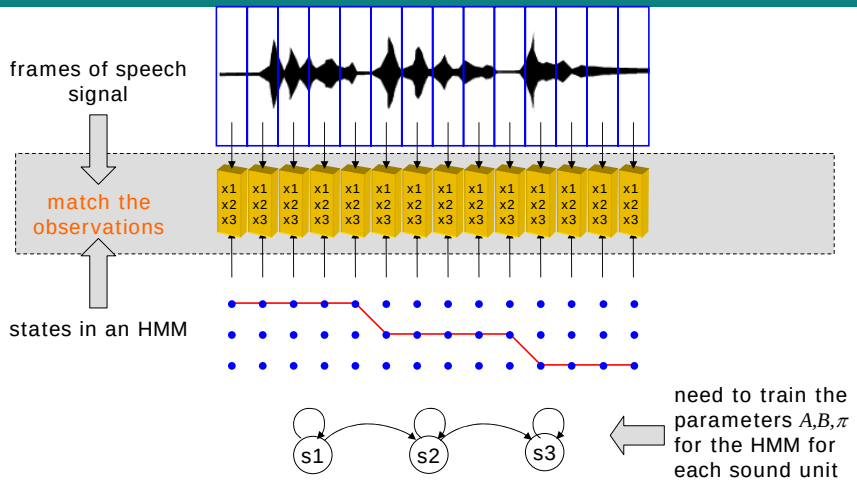


Modeling words with HMMs

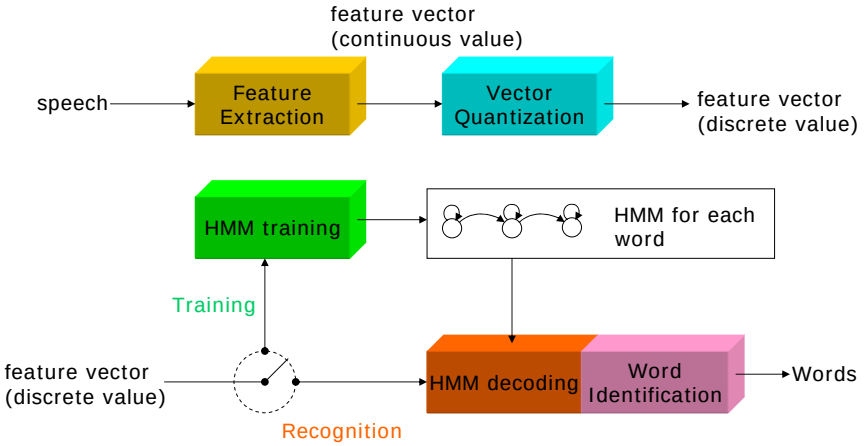
- We seek to find $W^* = argmax_W(P(W|O))$
- $W^* = argmax_W(P(O|W)P(W)/P(O))$
- W is an element of set $\{W_1, W_2, W_3, .., W_N\}$
- So, we compare $P(O|W_i)$, $i=1,2,...,N$, and pick Wi that gives maximal $P(O|W_i)$.



Acoustic Modeling using HMM

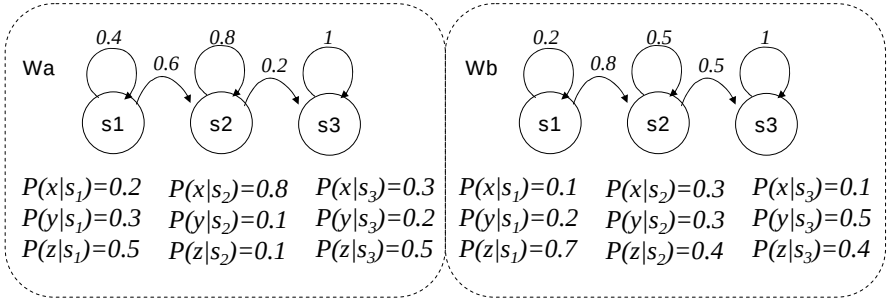


Speech Recognizer Based on discrete HMMs



Example

- 1D feature vector consisting of symbols {x,y,z}
- Vocab: {Wa,Wb}
- Observations: O = {x, x, z, y}
- Models: (already obtained from training)



$P(x\|s_1)=0.2$	$P(x\|s_2)=0.8$	$P(x\|s_3)=0.3$
$P(y\|s_1)=0.3$	$P(y\|s_2)=0.1$	$P(y\|s_3)=0.2$
$P(z\|s_1)=0.5$	$P(z\|s_2)=0.1$	$P(z\|s_3)=0.5$

$P(x\|s_1)=0.1$	$P(x\|s_2)=0.3$	$P(x\|s_3)=0.1$
$P(y\|s_1)=0.2$	$P(y\|s_2)=0.3$	$P(y\|s_3)=0.5$
$P(z\|s_1)=0.7$	$P(z\|s_2)=0.4$	$P(z\|s_3)=0.4$

- Decision rule: Pick Wi that maximize P(O|Wi)
- Constraints: 1st observation generated by s1. The last by s3.

Modeling subword units with HMMS

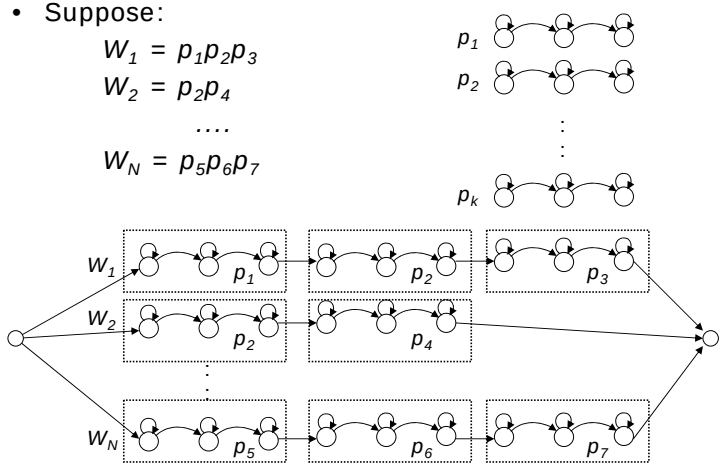
- Suppose:

$W_1 = p_1 p_2 p_3$

$W_2 = p_2 p_4$

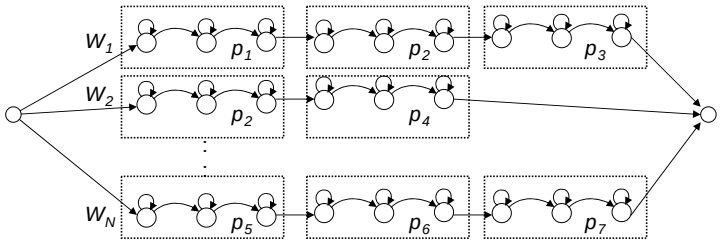
....

$W_N = p_5 p_6 p_7$



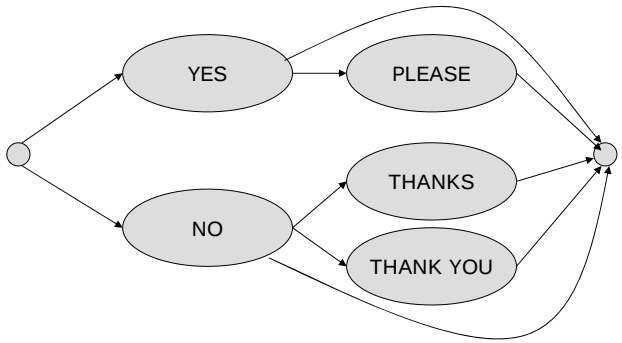
Modeling subword units with HMMS

- Find the best state sequence $Q = \{q_1,q_2,q_3,...,q_n\}$



Continuous Speech Recognizer

- Suppose a user can say:
  "Yes" / "Yes, please" / "No" / "No. Thanks" / "No. Thank you"
- Find the best state sequence $Q = \{q_1,q_2,q_3,...,q_n\}$



Continuous Speech Recognizer

- Suppose the system need to recognize a bit string without knowing its length.