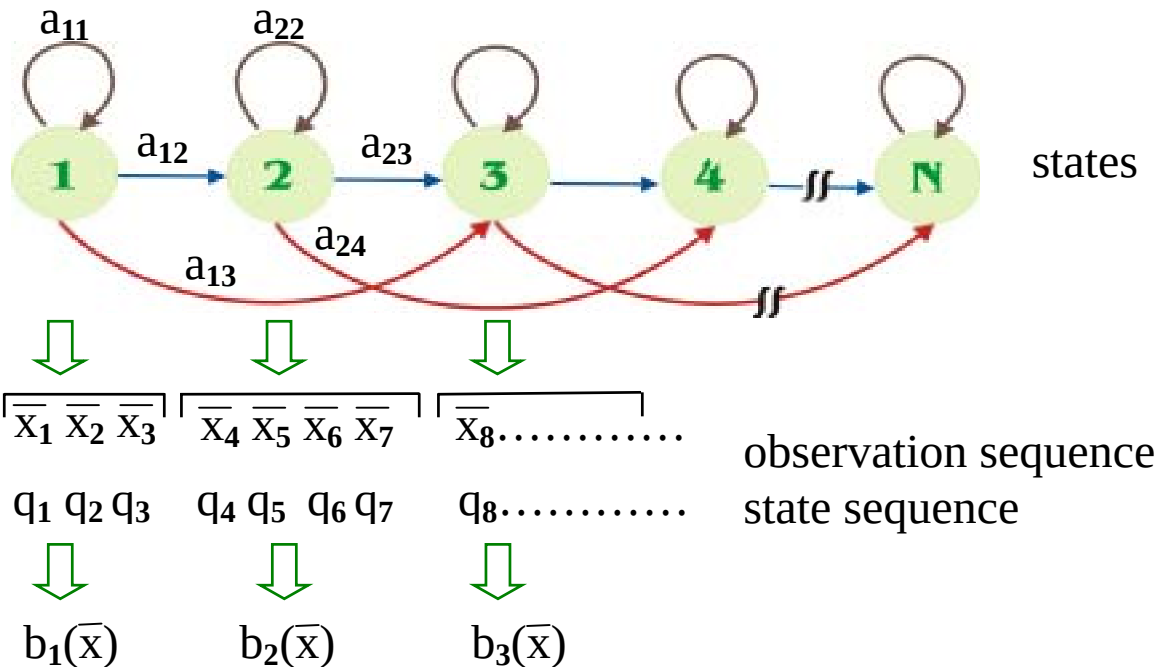
5.2 Fundamentals of Speech Recognition

Hidden Markov Models (HMM)



- **Formulation**

$\bar{x}_t = [x_1, x_2, \ldots x_D]^T$ feature vectors for frame at time t

$q_t = 1,2,3\ldots N$ state number for feature vector $\bar{x}_t$

$A = [a_{ij}]$, $a_{ij} = \text{Prob}[\, q_t = j \mid q_{t-1} = i \,]$

state transition probability

$B = [b_j(\bar{x}), j = 1,2,\ldots N]$ observation probability

$$b_j(\bar{x}) = \sum_{k=1}^{M} c_{jk} b_{jk}(\bar{x})$$

$b_{jk}(\bar{x})$: multi-variate Gaussian distortion for the k-th mixture of the j-th state

M : total number of mixtures

$$\sum_{k=1}^{M} c_{jk} = 1$$

$\pi = [\, \pi_1, \pi_2, \ldots \pi_N \,]$ initial probabilities

$\pi_i = \text{Prob}[q_1 = i]$

HMM : $(A, B, \pi) = \lambda$

Hidden Markov Models (HMM)

- **Double Layers of Stochastic Processes**
  - hidden states with random transitions for time warping
  - random output given state for random acoustic characteristics
- **Three Basic Problems**

(1) Given $\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots \overline{x}_t \ldots \overline{x}_T)$ and $\lambda = (A, B, \pi)$

find Prob $[\ \overline{X} \mid \lambda\ ]$

(2) Given $\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots \overline{x}_t \ldots \overline{x}_T)$ and $\lambda = (A, B, \pi)$

find a best state sequence $\overline{q} = (q_1, q_2, \ldots q_t, \ldots q_T)$

(3) Given $\overline{X}$, find best values for parameters in λ

such that Prob $[\ \overline{X} \mid \lambda\ ] = \max$

Feature Extraction (Front-end Signal Processing)

- **Pre-emphasis**

  $H(z) = 1 - az^{-1}, \quad 0 << a < 1$

  $x[n] = x'[n] - ax'[n-1]$

  - pre-emphasis of spectrum at higher frequencies

- **Windowing**

  $$Q_n = \sum_{m=-\infty}^{\infty} T\{x[m]\}w[n-m]$$

  $T\{\bullet\}$: some operator

  $w[m]$: window shape

  - Rectangular window

  $$w[m] = \begin{cases} 1, & 0 < m \le L-1 \\ 0, & \text{else} \end{cases}$$

  Hamming window

  $$w[m] = \begin{cases} 0.54 - 0.46\cos\left[\dfrac{2\pi m}{L-1}\right], & 0 \le m \le L-1 \\ 0, & \text{else} \end{cases}$$

  window length/shift/shape

- **Endpoint Detection (Speech/Silence Discrimination)**

  - short-time energy

  $$E_n = \sum_{m=-\infty}^{\infty} (x[m])^2 w[n-m]$$

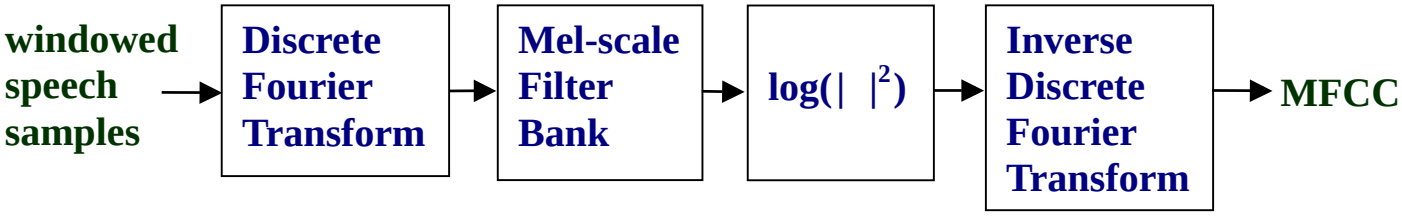
  short-time zero-crossing rate

  $$Z_n = \sum_{m=-\infty}^{\infty} (\mathrm{sgn}[m] - \mathrm{sgn}[m-1])w[n-m]$$

  - double thresholds/mixed references/adaptive parameters

Feature Extraction (Front-end Signal Processing)

- **Mel Frequency Cepstral Coefficients (MFCC)**

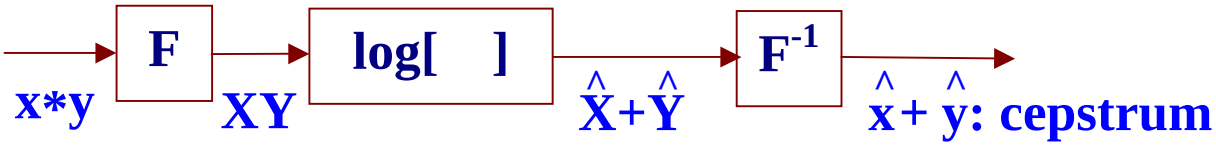


  - Mel-scale Filter Bank

    triangular shape in frequency/overlapped uniformly spaced below 1 kHz

    logarithmic scale above 1 kHz

  - Cepstral Analysis



- **Energy/Pitch and Delta Coefficients**
  - short-time energy
  - pitch detection
  - 1st/2nd order differences

Language Modeling

$W = (w_1, w_2, w_3, \ldots, w_i, \ldots w_R)$ a word sequence

- Evaluation of P(W)

$$P(W) = P(w_1) \prod_{i=2}^{R} P(w_i | w_1, w_2, \ldots w_{i-1})$$

- Assumption:

$$P(w_i | w_1, w_2, \ldots w_{i-1}) = P(w_i | w_{i-N+1}, w_{i-N+2}, \ldots w_{i-1})$$

Occurrence of a word depends on previous N–1 words only

N-gram language models

$N = 2$	:	bigram	$P(w_i \mid w_{i-1})$
$N = 3$	:	tri-gram	$P(w_i \mid w_{i-2}, w_{i-1})$
$N = 4$	:	four-gram	$P(w_i \mid w_{i-3}, w_{i-2}, w_{i-1})$
⋮			
$N = 0$	:	unigram	$P(w_i)$

probabilities estimated from a training text database

example : tri-gram model

$$P(W) = P(w_1)\, P(w_2 | w_1) \prod_{i=3}^{N} P(w_i | w_{i-2}, w_{i-1})$$

Language Modeling

- Evaluation of N-gram model parameters

unigram

$$P(w^i) = \frac{N(w^i)}{\sum_{j=1}^{V} N(w^j)}$$

w^i : a word in the vocabulary

V : total number of different words in the vocabulary

$N(\bullet)$ number of counts in the training text database

bigram

$$P(w^j|w^k) = \frac{N(<w^k,w^j>)}{N(w^k)}$$

$< w^k, w^j >$: a word pair

trigram

$$P(w^j|w^k,w^m) = \frac{N(<w^k,w^m,w^j>)}{N(<w^k,w^m>)}$$

smoothing – estimation of probabilities of rare events by statistical approaches

- Word-class-based Language Models

$P(w^j|w^k) \rightarrow P[C(w^j)|C(w^k)] \bullet Prob[w^j|C(w^j)]$

$C(w^j)$: the word-class including the word w^j

smoothing effect

word clustering algorithm

Language Modeling

- Segment-pattern-based Language Models

  segment patterns automatically extracted from training text database (e.g. using statistical approaches) taken as words in language modeling

- Language Models integrating grammatical knowledge from natural language processing

  N-grams are primarily based on local statistics grammatical knowledge provides more information about global structure

Large Vocabulary Continuous Speech Recognition

$W = (w_1, w_2, \ldots w_R)$ a word sequence

$\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots \overline{x}_T)$ feature vectors for a speech utterance

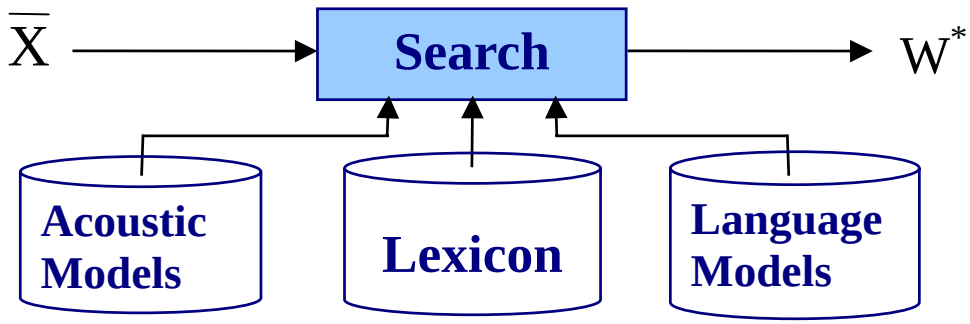
$$W^* = \underset{W}{\text{Arg Max}}\ \text{Prob}(W|\overline{X}) \quad \text{MAP principle}$$

$$\text{Prob}(W|\overline{X}) = \frac{\text{Prob}(\overline{X}|W) \bullet P(W)}{P(\overline{X})} = \max$$

$$\text{Prob}(\overline{X}|W) \bullet P(W) = \max$$

by HMM by language model

- **A Search Process Based on Three Knowledge Sources**



- Acoustic Models : HMMs for basic voice units (e.g. phonemes)
- Lexicon : a database of all possible words in the vocabulary, each word including its pronunciation in terms of component basic voice units
- Language Models : based on words in the lexicon

Example: See Figs. 7.3, 7.5, pp. 307, 309 of Becchetti

Word Recognition and Keyword Spotting

- **Word Recognition**
  - word models (HMMs) constructed by basic voice units (HMMs)

    Example: See Fig. 1.3, p. 15 of Becchetti

- **Keyword Spotting**
  - filler models
  - confidence measures

$$C = \frac{\mathrm{Prob}(\overline{X} \mid \lambda_{\mathbf{w}})}{\mathrm{Prob}(\overline{X} \mid \lambda'_{\mathbf{w}})}$$

$\lambda_{\mathbf{w}}$: model for the keyword

$\lambda'_{\mathbf{w}}$: background model for the keyword

Ref: 1.3, 3.3, 3.4, 4.2, 4.3, 6.4, 7.2, 7.3 of Becchetti