# Speech Communication, Spring 2006

## Lecture 5: Speech Recognition, Part II

Zheng-Hua Tan

Department of Communication Technology
Aalborg University, Denmark
zt@kom.aau.dk



# Part I: Hidden Markov model

- Hidden Markov model
  - How to evaluate an HMM – the forward algorithm
  - How to decode an HMM – the Viterbi algorithm
  - How to estimate HMM parameters – Baum-Welch Algorithm
- HMM based speech recognition

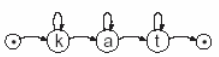
# Elements of an HMM

- **HMM is specified by:**

- states $q^i$


- transition probabilities $a_{ij}$

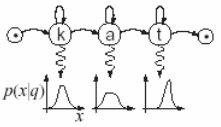
$$p(q_n^j \mid q_{n-1}^i) \equiv a_{ij}$$




| | k | a | t | • |
|---|---|---|---|---|
| • | 1.0 | 0.0 | 0.0 | 0.0 |
| k | 0.9 | 0.1 | 0.0 | 0.0 |
| a | 0.0 | 0.9 | 0.1 | 0.0 |
| t | 0.0 | 0.0 | 0.9 | 0.1 |

- emission distributions $b_i(x)$

$$p(x \mid q^i) \equiv b_i(x)$$




+ (initial state probabilities $p(q_1^i) \equiv \pi_i$ )

From Dan Ellis, 2004.



# Three basic HMM problems

1. **Scoring:** Given an observation sequence $\boldsymbol{O} = \{o_1, o_2, \ldots, o_T\}$ and a model $\lambda = \{\boldsymbol{A}, \boldsymbol{B}, \pi\}$, how to compute $P(\boldsymbol{O} \mid \lambda)$, the probability of the observation sequence? → The Forward-Backward Algorithm

2. **Matching:** Given an observation sequence $\boldsymbol{O} = \{o_1, o_2, \ldots, o_T\}$,how to choose a state sequence $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ which is optimum in some sense? → The Viterbi Algorithm

3. **Training:** How to adjust the model parameters $\lambda = \{A, B, \pi\}$ to maximize $P(\boldsymbol{O} \mid \lambda)$? → The Baum-Welch Re-estimation Procedures

# Problem 3: Training

- How to tune the model parameters $\lambda = \{A,B,\pi\}$ to maximize $P(\boldsymbol{O}|\lambda)$? - a learning problem
  - No efficient algorithm for global optimisation
  - Effective iterative algorithm for local optimisation: the Baum-Welch re-estimation
- Baum-Welch
  - = forward-backward algorithm (Baum, 1972)
  - is a special case of EM (expectation-maximization) algorithm
  - computes probabilities using current model $\lambda$;
  - refines $\lambda$ to $\bar{\lambda}$ such that $P(\boldsymbol{O}|\lambda)$ is locally maximised
  - uses $\alpha$ and $\beta$ from forward-backward algorithm



# Baum-Welch re-estimation

Define $\xi_t(i,j)$, the probability of being in state $i$ at time $t$, and state $j$ at time $t$+1, given $\lambda$ and $\boldsymbol{O}$, i.e.

$$\begin{aligned}\xi_t(i,j) &= P(q_t = i, q_{t+1} = j \mid \mathbf{O}, \lambda) \\ &= \frac{P(q_t = i, q_{t+1} = j, \mathbf{O} \mid \lambda)}{P(\mathbf{O} \mid \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}{P(\mathbf{O} \mid \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)}\end{aligned}$$

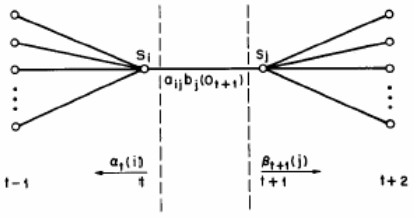



**Fig. 6.** Illustration of the sequence of operations required for the computation of the joint event that the system is in state $S_i$ at time $t$ and state $S_j$ at time $t + 1$.

# Baum-Welch Re-estimation (cont'd)

- Recall that $\gamma_t(i)$ is defined as the probability of being in state i at time t, given the entire observation sequence and the model, so

$$\gamma_t(i) = P(q_t = i \mid O, \lambda) = \sum_{j=1}^{N} P(q_t = i, q_{t+1} = j \mid O, \lambda) = \sum_{j=1}^{N} \xi_t(i,j)$$

- Sum $\gamma_t(i)$ and $\xi_t(i,j)$ over *t*, we have

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from state } i \text{ in } \mathbf{O}$$

$$(\sum_{t=1}^{T} \gamma_t(i) = \text{the expected number of times that state } i \text{ is visitied.})$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{expected number of transitions from state } i \text{ to state } j \text{ in } \mathbf{O}$$



# Baum-Welch re-estimation formulas

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } i \text{ at time } (t=1) = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions from state } i} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} = \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \overset{s.t.\, o_t = v_k}{=} \frac{\sum_{t=1}^{T} P(O, q_t = i \mid \lambda) \cdot \delta(o_t, v_k)}{\sum_{t=1}^{T} P(O, q_t = i \mid \lambda)}$$

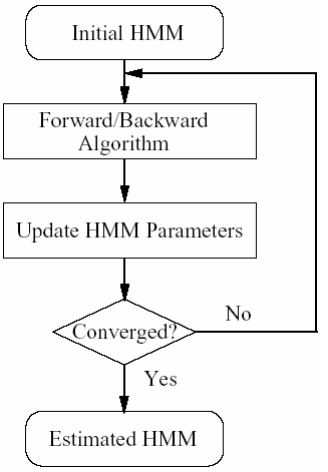
$$\delta(o_t, v_k) = \begin{cases} 1 & o_t = v_k \\ 0 & \text{otherwise} \end{cases}$$

# Parameter re-estimation process

1. Initialize $\lambda = \{A, B\}$
2. Compute $\alpha, \beta,$ and $\xi$
3. Estimate $\overline{\lambda} = \{\overline{A}, \overline{B}\}$ from $\xi$
4. Replace $\lambda$ with $\overline{\lambda}$
5. If not converged go to 2

It can be shown that

$$P(O \mid \overline{\lambda}) > P(O \mid \lambda) \text{ unless } \overline{\lambda} = \lambda$$






# Continuous density HMMs

- Replaces the discrete observation probabilities, $b_j(k)$, by a continuous PDF (probability density function) $b_j(x)$
- The PDF $b_j(x)$ is often represented as a mixture of Gaussians:

$c_{jk}$ is the mixture weight, $c_{jk} \geq 0$, and $\sum_{k=1}^{M} c_{jk} = 1$

$$b_j(\boldsymbol{x}) = \sum_{k=1}^{M} c_{jk} N[\boldsymbol{x}, \boldsymbol{\mu}_{jk}, \Sigma_{jk}] \qquad 1 \leq j \leq N$$

$N$ is the normal density

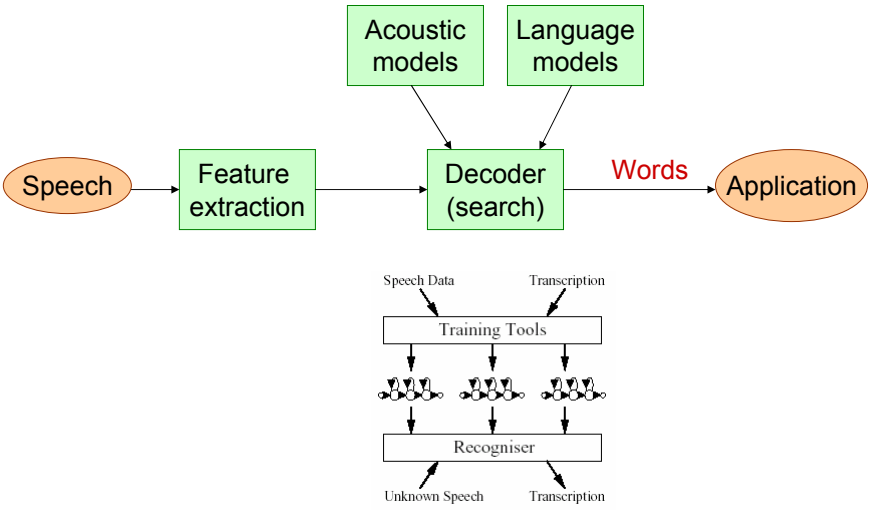
The mean and covariance matrix associated with state $j$ and mixture $k$

# Part II: HMM based ASR

- Hidden Markov model
  - How to evaluate an HMM – the forward algorithm
  - How to decode an HMM – the Viterbi algorithm
  - How to estimate HMM parameters – Baum-Welch Algorithm
- HMM based speech recognition
  - HTK (Steve Young, 1996)

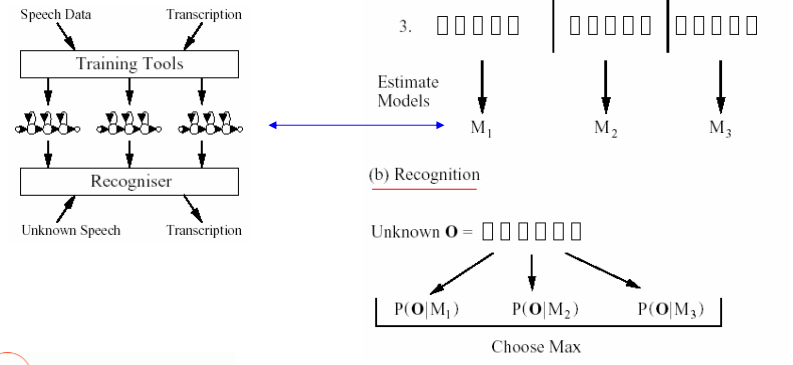


# Speech recognition system



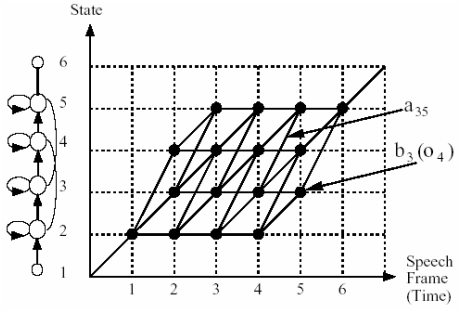
# Training and test procedures for IVR

From (Young et al. 1996)






# The Viterbi algorithm for IVR



# Continuous speech recognition

- The allowed sequence of phoneme-based HMMs is defined by a finite state network and all of the words are placed in a loop

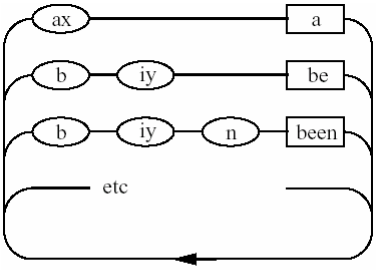



Fig. 1.7 Recognition Network for Continuously Spoken Word Recognition



# Token passing

- An alternative formulation of the Viterbi algorithm is used called the *Token Passing Model*. In brief, the token passing model makes the concept of a state alignment path explicit.
- The key steps in this algorithm are as follows
  - Pass a copy of every token in state *i* to all connecting states *j*, incrementing the log probability of the copy by $log[a_{ij}] + log[b_j(\boldsymbol{o}(t)]$ .
  - Examine the tokens in every state and discard all but the token with the highest probability.

# Grammar

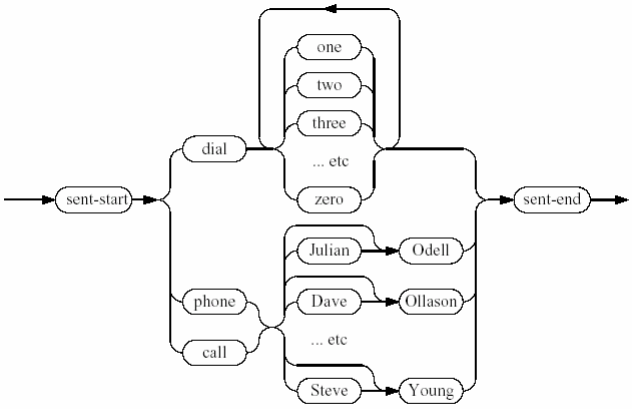
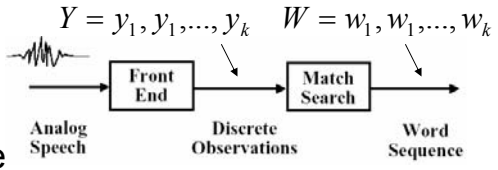



Fig. 3.1 Grammar for Voice Dialling



# Continuous speech recognition

- Goal:
  - Given acoustic data
  - Find word sequence
  - Such that $P(W|Y)$ is maximized
- Bayes Rule:



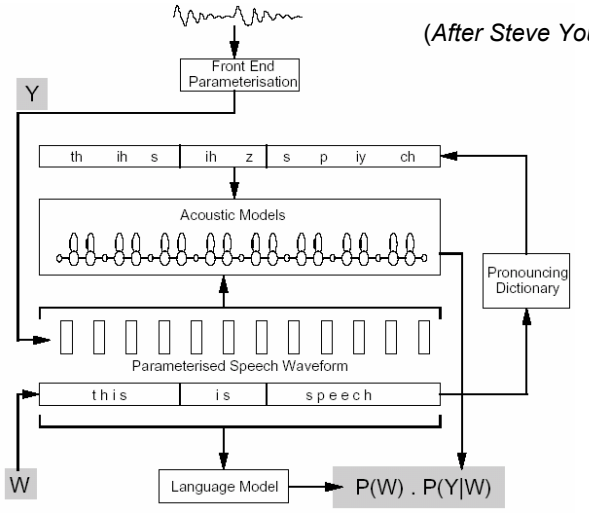

*Acoustic model* *Language model*

$$P(W \mid Y) = \frac{P(Y \mid W) \cdot P(W)}{P(Y)}$$

*P(Y) is a constant for a complete sentence*

# Overview of HMM based ASR

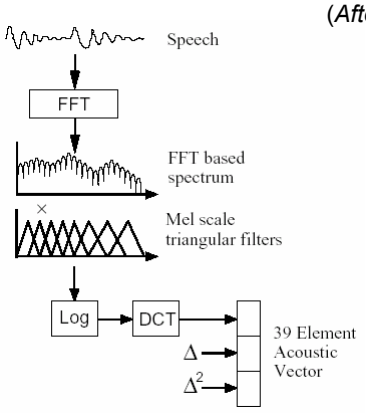
(*After Steve Young, 1996*)





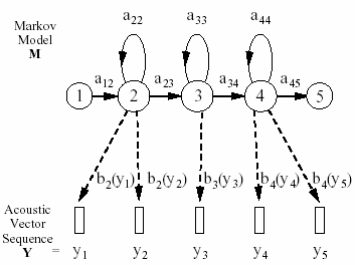

# MFCC-based front-end processing

(*After Steve Young, 1996*)



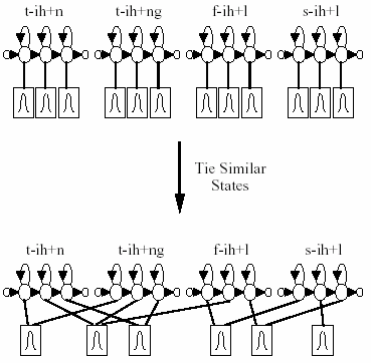
# HMM-based phone model

(*After Steve Young, 1996*)

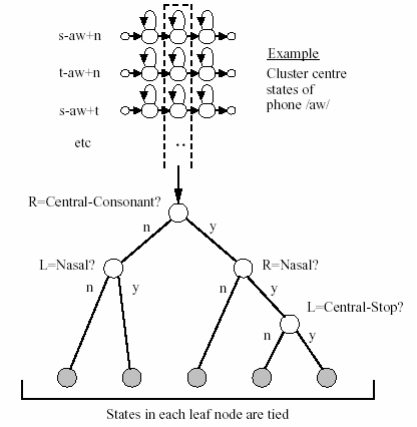





# Triphone models and state tying



# Decision tree clustering






# Variability in the speech signal

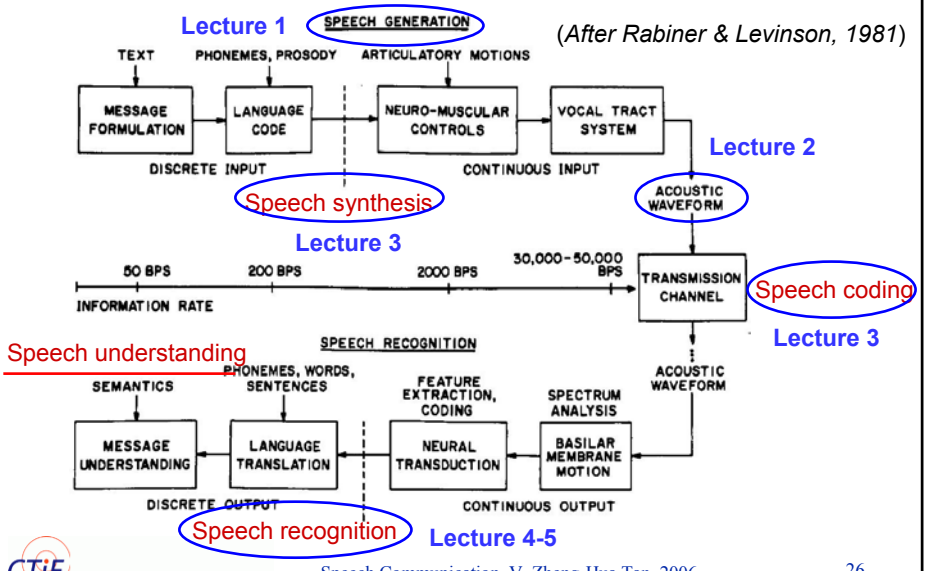
- Most noticeable factors that determine accuracy are variations in context, in speaker, and in environment.
- Speech recogniser can be very accurate for a particular speaker, in a particular language and speaking style, in a particular environment, and limited to a particular task.
- But it remains a research challenge to build a recogniser that can understand anyone's speech, in any language, on any topic, in any free-flowing style, and in any speaking environment
- Accuracy and robustness are the ultimate measures for the success of ASR

# Variability

- Context variability
  - It is easy to recognise speech.
  - It is easy to wreck a nice beach.
- Style variability
  - Isolated, continuous, spontaneous
- Speaker variability – human vocal tract
  - Speaker-dependent vs. speaker-independent
  - Speaker-adaptation
- Environmental variability
  - Multistyle training



# Human speech communication process



# Summary

- Hidden Markov model
- HMM-based ASR

- Next lectures: : Language Processing and Speech Understanding by Tom Brøndsted.