EEEM034/1011/5pg/Req:FB

UNIVERSITY OF SURREY ©

Faculty of Engineering and Physical Sciences
Department of Electronic Engineering

Undergraduate and Postgraduate Programmes in Electronic Engineering

Module EEEM034; 15 credits

Speaker & Speech Recognition

Level HEM Examination

Time allowed: Two hours

Semester 2, 2011

Read these instructions!

Answer **three questions** out of four.
If you do more, your best three will count.
All questions carry equal credit.

Where appropriate the mark carried by an individual part of a question is indicated in square brackets [ ].

Additional materials: Formula booklet

Answer three questions

1. (a) Briefly explain the source-filter theory of speech production, making reference to the following aspects:

    - types of source for representing voiced and unvoiced speech sounds
    - the main acoustical elements of the filter
    - the assumptions that are embodied in the theory
    - an illustration of the theory for a typical vowel sound, using both time-domain and frequency-domain plots

    [20%]

   (b) There are three speech sounds, or phonemes, in the word "mass": /m/, /æ/ and /s/. Answer all parts **for each sound**.

    i. State the manner and place of articulation, as in the IPA chart.
    ii. What is the origin of the sound source?
    iii. Describe the articulatory configuration, highlighting which parts of the anatomy determine the filter response.
    iv. Suggest one characteristic of the speech signal or short-time speech spectrum that you would expect to observe.

    [60%]

   (c) Qualitatively describe one method for extracting relevant phonetic information from a speech signal into features suitable for ASR.

    [20%]

2. (a) i. For what purpose is the Viterbi algorithm used with HMMs in ASR?
   ii. In what sense can the Viterbi algorithm be described as *recursive*? **[10%]**

(b) For an observation sequence $\mathcal{O}=\boldsymbol{o}_1^T=\{o_1..o_T\}$ and any state sequence $X=\boldsymbol{x}_1^T=\{x_1..x_T\}$, a continuous HMM, λ, can be used to compute $P(X|\lambda)$ and $p(\mathcal{O}|X,\lambda)$. Hence, how can we calculate $p(\mathcal{O},X|\lambda)$ for given state and observation sequences? **[10%]**

(c) For a given model λ, maximum cumulative likelihood is defined at state i and time t: $\delta_t(i) = \max_{\boldsymbol{x}_1^t} p\left(\boldsymbol{o}_1^t, \boldsymbol{x}_1^{t-1}, x_t = i\right)$ for all $i \in \{1..N\}$ and $t \in \{1..T\}$.

   i. Give a similar expression for the best state sequence X^* up to the final point.
   ii. Using $\delta_t(i)$'s definition above, write the expression for $\delta_{t+1}(j)$ in state j and time $t+1$, in terms of $p\left(\boldsymbol{o}_1^t, \boldsymbol{x}_1^{t-1}, x_t = i\right)$.
   iii. Hence, considering the model's assumption that $p\left(o_{t+1}, x_{t+1} = j|\boldsymbol{o}_1^t, \boldsymbol{x}_1^{t-1}, x_t = i\right) = P\left(x_{t+1} = j|x_t = i\right) p\left(o_{t+1}|x_{t+1} = j\right)$, derive an expression for $\delta_{t+1}(j)$ in terms of $\delta_t(i)$ and elements of $\lambda = \{A, B\}$. **[20%]**

(d) A two-emitting-state continuous HMM has parameters $\lambda = \{A, B\}$ defined:

$$A = \left(\begin{array}{c|cc|c} 0 & 0.8 & 0.2 & 0 \\ \hline 0 & 0.5 & 0.4 & 0.1 \\ 0 & 0.1 & 0.7 & 0.2 \\ \hline 0 & 0 & 0 & 0 \end{array}\right), \qquad \mu = \begin{pmatrix} -1.2 \\ 1.8 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

where the output pdfs, $B=\{b_i\}$, are univariate Gaussian distributions $\mathcal{N}(o_t;\mu_i,\Sigma_i)$ with mean μ_i and variance Σ_i.

   i. Sketch this HMM's state topology, including the entry and exit null states.
   ii. Give the term used to describe this type of topology.

[15%]

(e) An observation sequence $\mathcal{O}=\boldsymbol{o}_1^2=\{-1.0, 2.0\}$ has output probabilities as in Table 1.

   i. Calculate $\delta_t(i)$ for both emitting states $i \in \{1,2\}$ and time frames, $t \in \{1,2\}$.
   ii. Calculate $p(\mathcal{O}, X^*|\lambda)$ for the best complete state sequence X^* (i.e., incorporating the transition into the exit node).
   iii. Determine the path of the best state sequence X^*.
   iv. Comment on the decoded occupation of states along X^* in relation to your expectations, based on the observations and state topology. **[45%]**

state	o_1	o_2
1	0.229	0.042
2	0.040	0.279

Table 1: Output likelihoods $b_i(o_t)$ for each state i and observations at time frames $t = \{1, 2\}$.

3. (a) What is the purpose of performing Baum-Welch training on an HMM, λ? **[10%]**

(b) For an observation sequence $\mathcal{O}=\boldsymbol{o}_1^T=\{o_1, .., o_T\}$, forward and backward likelihoods are defined as $\alpha_t(i) = P\left(\boldsymbol{o}_1^t, x_t = i|\lambda\right)$ and $\beta_t(i) = P\left(\boldsymbol{o}_{t+1}^T|x_t = i, \lambda\right)$ respectively, in state i at time t. How can they be used to express:

i. the occupation likelihood $\gamma_t(i) = P\left(x_t = i|\boldsymbol{o}_1^T, \lambda\right)$?

ii. the transition likelihood $\xi_t(i,j) = P\left(x_{t-1} = i, x_t = j|\boldsymbol{o}_1^T, \lambda\right)$? **[15%]**

(c) A three-emitting-state discrete HMM with initial parameters $\lambda = \{A, B\}$ as in Table 2 is updated by Baum-Welch re-estimation using a single training example: $\mathcal{O} = \{\clubsuit, \diamondsuit\}$. Calculate $\alpha_t(i)$ for $i \in \{1, 2, 3\}$ and $t \in \{1, 2\}$. **[25%]**

(d) Use your values of $\alpha_t(i)$ with those of $\beta_t(i)$ in Table 2 and $P(\mathcal{O}|\lambda)$=0.0028 to compute $\gamma_t(i)$ for $i \in \{1, 2, 3\}$ and $t \in \{1, 2\}$. **[25%]**

(e) i. Given that $\hat{b}_i(k) = \left(\sum_t \gamma_t(i)\omega_t(k)\right) / \sum_t \gamma_t(i)$ where $\omega_t(k)$ is a binary indicator for observation o_t and type of observation k, derive output probabilities $\hat{b}_i(k = \clubsuit)$ to update the third column of the B matrix for $i \in \{1, 2, 3\}$, showing all your working.

ii. Comment on how these new $\hat{b}_i$ values compare to the initial parameters b_i.

[25%]

$$A = \left(\begin{array}{c|ccc|c} 0 & 0.8 & 0.2 & 0 & 0 \\ \hline 0 & 0.7 & 0.2 & 0.1 & 0 \\ 0 & 0 & 0.6 & 0.3 & 0.1 \\ 0 & 0 & 0 & 0.9 & 0.1 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad B = \left(\begin{array}{cccc} \spadesuit & \heartsuit & \clubsuit & \diamondsuit \\ 0.4 & 0.5 & 0.1 & 0.0 \\ 0.0 & 0.1 & 0.8 & 0.1 \\ 0.1 & 0.6 & 0.0 & 0.3 \end{array}\right)$$

Table 2: Initial HMM parameters $\lambda = \{A, B\}$ for state transition matrix A and output matrix B with observation types $k \in \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$.

	$\beta_1(i)$	$\beta_2(i)$
state 1	0.005	0.0
state 2	0.015	0.1
state 3	0.027	0.1

Table 3: Backward likelihood $\beta_t(i)$ for each state i and observations o_1 and o_2 at $t = \{1, 2\}$.

4. (a) Explain what an authenticator is. List the major authenticator types and give an example for each category.

[15%]

(b) Identify advantages and disadvantages of voice biometrics.

[15%]

(c) Draw a block diagram for a speaker recognition system, describing each component and its function.

[15%]

(d) Sketch a receiver operating characteristic (ROC) curve and explain its purpose.

[10%]

(e) A speaker model is characterised by a covariance matrix $\Phi = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$.
A speech utterance to be tested against the model has covariance matrix
$\Sigma = \Phi + \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix}$. Determine the range of permissible values of a for Σ to remain a covariance matrix.

[25%]

(f) Using the Bhattacharrya distance

$$J = \ln \frac{|\frac{1}{2}(\Phi + \Sigma)|}{\sqrt{|\Phi||\Sigma|}}$$

as a matching criterion, determine whether an access claim posed by the speaker verification problem defined by the covariance matrices Φ and Σ (given in part (e)) will be accepted against a threshold $t = 0.3$ with $a = 2$.

[20%]

Examiners: Dr. P.J.B. Jackson, Prof. J. Kittler
External examiners: Prof. L. Cuthbert, Prof. P. Rees