EEEM034/1112/7pg/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, 2012

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) The recognition task for responses to a yes-no question in a dialogue system is defined:

```
$answer = ( SENT-START [ SIL ] ( YES | NO ) [ SIL ] SENT-END )
```

where `|` denotes alternatives, `[ ]` denotes an optional part, and `SIL` represents silence.

i. Based on the above definition, draw the word network for this task.

ii. What is the purpose of the two `[ SIL ]` parts? **[15%]**

(b) The observations in this system are continuous, univariate (1D) values based on the signal's log energy in each 20 ms frame. The output pdfs for all states in the `YES` and `NO` models are represented by one gaussian each; the emitting state in the `SIL` model is represented by a mixture of two gaussians. The model parameters are given in Table 1.

i. What advantage does a mixture of gaussians have for modeling silence distributions, compared with a single gaussian?

ii. Sketch, on one graph for each model, the pdfs for the emitting states. **[20%]**

(c) The state transition matrices for the `YES`, `NO` and `SIL` models are:

$$A_{\text{YES}} = \left[\begin{array}{c|cccc} 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0.5 & 0.3 & 0.2 \\ 0 & 0 & 0 & 0.6 & 0.4 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad A_{\text{NO}} = \left[\begin{array}{c|cccc} 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0.8 & 0.1 & 0.1 \\ 0 & 0 & 0 & 0.6 & 0.4 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad A_{\text{SIL}} = \left[\begin{array}{c|cc} 0 & 1 & 0 \\ \hline 0 & 0.9 & 0.1 \\ \hline 0 & 0 & 0 \end{array}\right]$$

i. Give the term used to describe state topologies like that of the `YES` model.

ii. The trellis diagram for the 4-frame test utterance is outlined in Figure 1. Redraw the diagram showing all the paths with non-zero likelihood, including legal transitions between models based on the above grammar and state topologies.

iii. When using the Viterbi algorithm for decoding to determine the best path, write down the equation that would be used in the middle of the trellis (e.g., at t=3) to calculate the maximum cumulative likelihood, $\delta_t(j)$ for state j. Make sure the meaning of all symbols is clearly defined.

iv. Considering the output pdfs in your sketches and the legal transitions, what would be the most likely recognition output for the test utterance with observations $\mathcal{O} = \{21.8, 22.1, 24.3, 11.5\}$? Justify your answer. **[50%]**

(d) Propose **two** modifications to the design that would increase the dialogue system's success rate, assuming plenty of training data and processing power are available. **[15%]**

Question 1 continues on next page

Question 1 continued from previous page

B_{YES}			B_{NO}			B_{SIL}				
i	μ_i	σ_i^2	i	μ_i	σ_i^2	i	m	c_i	μ_i	σ_i^2
1	22	4	1	16	4	1	1	0.5	4	1
2	24	1	2	22	4	1	2	0.5	6	4
3	12	9	3	18	4					

Table 1: HMM parameters specifying the output pdfs, $b_i(o_t)$, for each state i (and mixture component m) of the model: mean μ, variance σ^2, and mixture weight c.

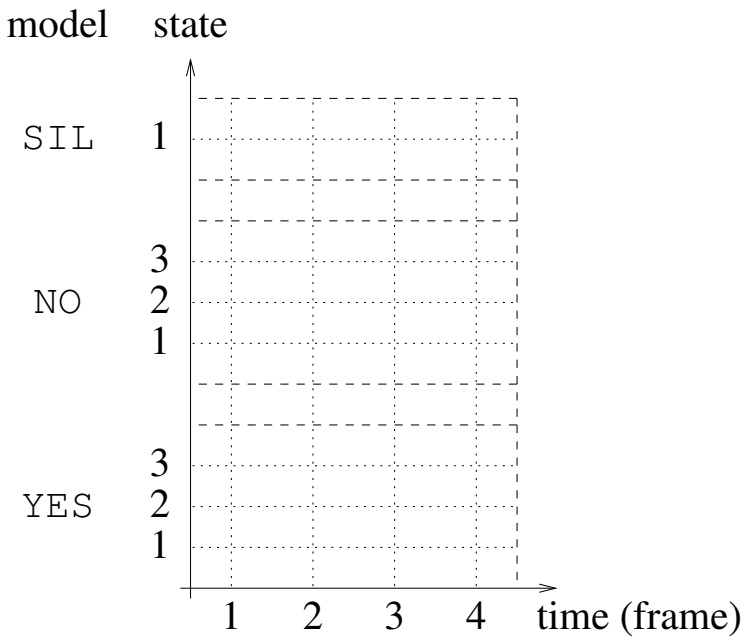

Figure 1: Trellis for recognizing responses to a yes-no question.

SEE NEXT PAGE

2. (a) A set of continuous HMMs trained on recorded examples of English phonemes is to be used to generate synthesized speech from text. The text is first converted from words into a series of phonemes. For example, the text "sunny" would become /sʌnɪ/ in phonetic symbols, or [s,uh,n,iy] using the names of the corresponding models, shown in Figure 2. Given those models, draw the combined HMM for this example. **[20%]**

(b) The output probabilities associated with each state are in the form of a multivariate gaussian pdf with diagonal covariance over the 13-dimensional acoustic feature space. The acoustic features are composed of 12 linear prediction (LP) coefficients plus the log energy from the windowed frames of speech taken every 10 ms. The log magnitude from the spectra at the mean LP features of the dominant states in the phone models [s], [n] and [iy] are shown in Figure 3, next to a measured spectrum from one speech frame of the same phone.

i. Which plot describes the phone [iy]?

ii. Which plot describes the phone [s]?

Provide justification and reasons to explain your choice in each case. **[20%]**

(c) What is the main difference between the LP representation of the spectral envelope and that derived from cepstral coefficients? **[10%]**

(d) In addition to the spectral envelope (filter) information from the state pdfs, the synthesizer needs an excitation (source) signal in order to produce a speech waveform. Suggest a suitable type of excitation:

i. for the phone [iy];

ii. for the phone [s]. **[15%]**

(e) i. To the nearest 100 ms, how long would you expect the synthesized utterance to be?

ii. Hence, based on your answer to part (a), what would be the average time spent in each state? Express your answer in terms of both milliseconds and frames.

[10%]

(f) If the mean of the gaussian pdf were used to define the filter characteristic for each state during synthesis, how would this affect the dynamics of the output speech pattern? **[10%]**

(g) **Either** briefly suggest ways to improve naturalness of the output speech,
or describe in one short paragraph the relationship between the LP coefficients and the acoustical properties of the recorded speaker's vocal tract. **[15%]**

Question 2 continues on next page

FIGURES FOR QUESTION 2 CONTINUED FROM PREVIOUS PAGE

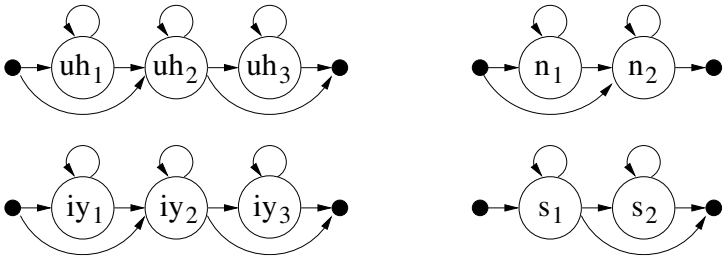

Figure 2: State diagrams of the phone models: (top, left) [uh], (right) [n], (bottom, left) [iy], (right) [s].

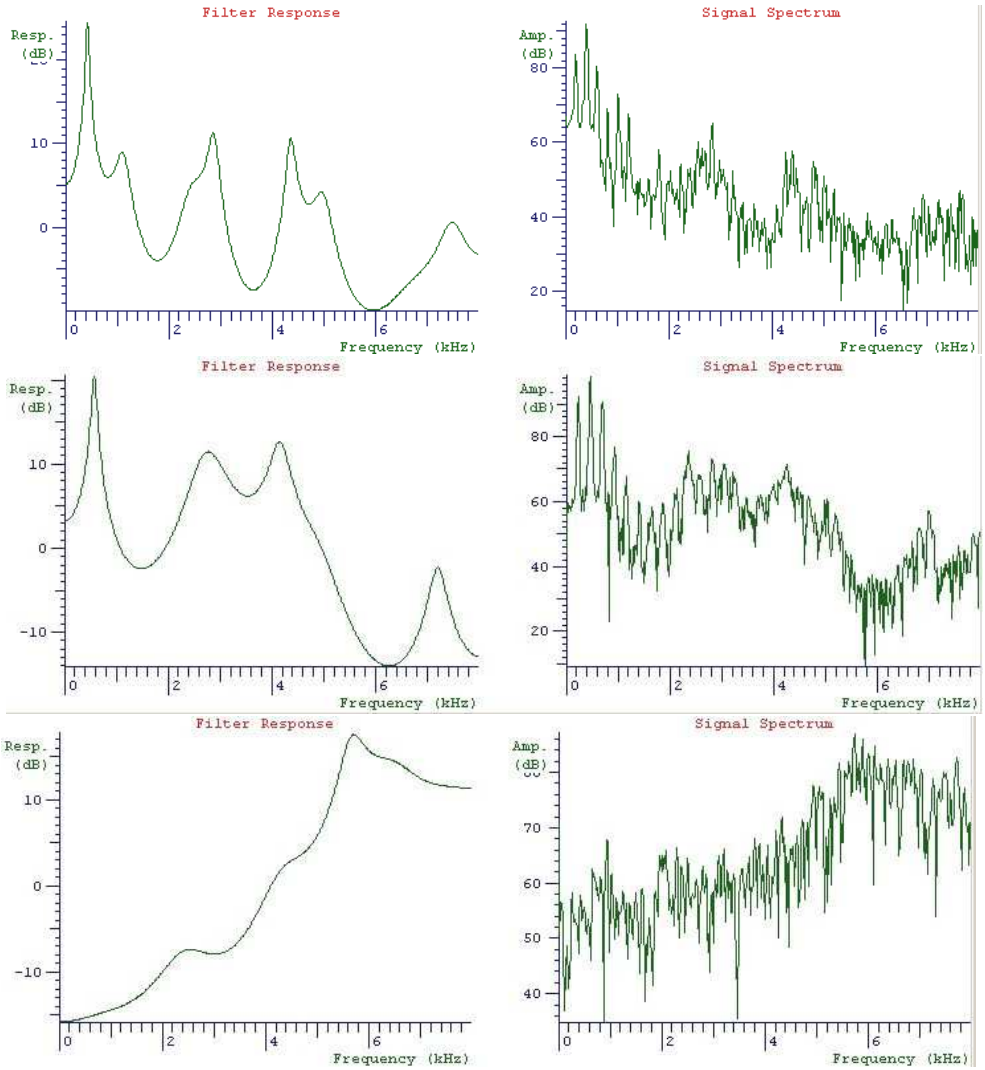

Figure 3: LP filter response (left) and magnitude spectrum (right) of three short speech segments.

SEE NEXT PAGE

3. (a) During development of an HMM-based recognition system, what is the purpose of training? **[10%]**

(b) Given the state-transition probability matrix A below for a prototype HMM, which incorporates the entry and exit probabilities, sketch the topology of the model in the usual way:

- show emitting states as circles containing the state number
- denote start and end null nodes as smaller filled circles
- draw arrows to indicate the permissible transitions

$$A = \left[\begin{array}{c|cccc} 0 & 0.75 & 0.25 & 0 & 0 \\ \hline 0 & 0.6 & 0.3 & 0.1 & 0 \\ 0 & 0 & 0.5 & 0.4 & 0.1 \\ 0 & 0 & 0 & 0.7 & 0.3 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right], \qquad B = \begin{array}{c} \begin{array}{ccccc} \oplus & \ominus & \otimes & \oslash & \odot \end{array} \\ \left[\begin{array}{ccccc} 0.4 & 0.3 & 0.2 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.0 & 0.2 \\ 0.2 & 0.0 & 0.1 & 0.3 & 0.4 \end{array}\right] \end{array}$$

[15%]

(c) Using this prototype, the best state sequence $X = \{1, 1, 1, 1, 2, 3, 3, 3, 3, 3, 3\}$ was found by the Viterbi algorithm for an observation sequence $\mathcal{O} = \{\ominus, \oslash, \otimes, \oplus, \otimes, \odot, \oslash, \oplus, \oslash, \odot, \oslash\}$, including a final transition to the (non-emitting) exit node. According to Viterbi training:

i. calculate the new state-transition matrix A^{Vit}, including entry and exit probabilities
ii. calculate the new output probabilities B^{Vit} **[45%]**

(d) i. From your values of A^{Vit}, sketch the topology of the newly re-estimated model.
ii. Comment on any changes from the original prototype HMM. **[15%]**

(e) i. What is the main difference between Baum-Welch training and Viterbi training?
ii. With reference to the A^{Vit} matrix that you re-estimated by Viterbi training in part (c), describe qualitatively how you would expect the values re-estimated by Baum-Welch training to differ. **[15%]**

SEE NEXT PAGE

4. (a) What is the *likelihood ratio* in decision making and how can it be applied in speaker verification?

[**10%**]

(b) What is the difference between *impostor* distribution and *world population* distribution?

[**10%**]

(c) What is the role of score normalisation in biometric systems?

[**10%**]

(d) Describe three score normalisation methods.

[**10%**]

(e) Write down the expression for the score normalisation method by *mapping to designated means*.

[**10%**]

(f) Consider a biometric personal identity verification system that takes a voice biometric measurement on a client and evaluates its consistency with the claimed client identity. The outcome of the consistency test is expressed in terms of a one-dimensional score. It is known that the distribution of the scores for the class of impostors is uniform on the interval $[3, 5]$ whereas the client score distribution is assumed to be triangular, given as

$$p(s|\text{client}) = \begin{cases} (s+4)\frac{1}{16} & \text{for } s\,\epsilon\,[-4, 0] \\ -(s-4)\frac{1}{16} & \text{for } s\,\epsilon\,[0, 4] \\ 0 & \text{elsewhere} \end{cases}$$

i. How will the above two distributions change as a result of the score normalisation method by *mapping to the designated means*, set to 0 and 1 for the client and imposter distributions respectively?

ii. Assuming that the decision threshold is set at threshold $t = 0.5$, what will be the false acceptance and false rejection rates?

iii. Find the minimum error threshold separating the normalised scores of clients and impostors.

iv. Estimate the error for the decision threshold derived in part f(iii).

[**50%**]