ENGINEERING TRIPOS PART IIB

ELECTRICAL AND INFORMATION SCIENCES TRIPOS PART II

Wednesday 30 April 2003 2.30 to 4

Module 4F11

SPEECH PROCESSING

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin*

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

(TURN OVER

1 The source-filter model of speech production separates the speech signal into two parts referred to as the excitation source and the filter representing the vocal tract.

(a) The excitation signal is commonly generated using two types of signal source.

(i) Name the class of speech sound associated with each of the two signal sources. [10%]

(ii) For each of the two classes draw an example of a typical waveform and the associated spectrum. For each case describe the characteristics of the signal in the time and frequency domain. [25%]

(b) The filter representing the human vocal-tract determines the envelope of the speech spectrum. Linear predictor filters are often used to model this envelope.

(i) Describe the basic operation of a linear predictor. [15%]

(ii) The filter parameters are optimised on a 25ms frame of the speech signal. Outline an optimisation criterion used for finding linear prediction coefficients. You should clearly state the cost function to be optimised. What is the consequence of using this criterion in terms of matching the spectrum? [25%]

(c) How could the source-filter model described in sections (a) and (b) be used to synthesize a speech signal? Describe the limitations of such an approach. [25%]

2 Speech coding algorithms allow a considerable reduction in the number of bits per second required to represent the speech signal.

(a) Name *four* commonly used design criteria for speech coders. [20%]

(b) Assessing the quality of the speech output from a coder is a difficult and expensive process. What are the key factors in speech coder quality assessment? Describe *two* techniques commonly used to assess these factors. [30%]

(c) The LPC10 coder is a standard speech coder for encoding speech sampled at 8 kHz. Linear prediction coefficients (LPC) are used to model the spectral envelope. A complete set of parameters are transmitted every 180 samples.

(i) Describe the parameters transmitted at each frame. Give an indication of the typical number of bits associated with each parameter. [15%]

(ii) Using your estimates from part (c)(i) compute the overall bit-rate. [10%]

(iii) In order to reduce the bit-rate the LPC are quantised. Briefly describe a quantisation scheme that would allow a significant reduction in the number of bits used for encoding the LPC while maintaining good speech quality. [25%]

3 The front-end feature extraction process is to be designed for a hidden Markov model based large vocabulary speech recognition system.

(a) What are the desirable attributes for the front-end of a speech recognition system? [20%]

(b) An initial proposal is to use Mel-frequency cepstral coefficients (MFCCs) as the front-end. Describe the steps for obtaining a set of MFCC parameters from a 16KHz digitised waveform. At each stage in the process describe what attribute of the feature vector is being altered and give typical values for the size of the feature vector. [35%]

(c) It is decided to add delta and delta-delta coefficients to the feature vector. Describe how these features are obtained and why they may be useful in a large vocabulary speech recognition system. You should also state any disadvantages in adding these features. [25%]

(d) Practical speech recognition systems are required to operate in situations where there may be high levels of background noise. For the feature extraction process described in part (b) comment on how background noise will affect the feature vector obtained. The background noise may be assumed to be additive to, and independent of, the speech signal. [20%]

4 An automatic speech recognition system is to be designed using hidden Markov models (HMMs). Three proposals are being considered. The systems are:

(i) monophone models with a single full-covariance Gaussian distribution as the output distribution for each state;

(ii) monophone models with an M-component diagonal covariance matrix Gaussian mixture model as the output distribution for each state;

(iii) cross-word triphone models with a single diagonal covariance matrix Gaussian distribution as the output distribution for each state.

For all systems a d-dimensional feature vector is used. The size of the phone set to be used is 50.

(a) Give expressions for the log-likelihood of the output distribution of a particular state, j, generating an observation vector, $\boldsymbol{y}$, for systems (i) and (ii). Compare these two forms of output distribution in terms of the number of model parameters per state, ability to model the training data for that state, and computational cost in calculating the log-likelihood. Clearly state any assumptions made. [35%]

(b) The third system uses cross-word triphones as the acoustic unit.

(i) What are cross-word triphone models? [15%]

(ii) One approach to generating a cross-word triphone system is to use decision-tree tying. Why is parameter tying required for generating cross-word triphone models? Describe the main attributes of decision-tree tying. [30%]

(c) The total number of model parameters in systems (ii) and (iii) are set up to be approximately the same. Contrast how these two systems handle the co-articulation problem in speech recognition. [20%]

5 Part of the dictionary for a phone-based speech recognition system is given below.

```
FREE   f r iy
TREE   t r iy
TREAT  t r iy t
```

A recogniser based on the Viterbi algorithm is to be used.

(a) Briefly describe how the Viterbi algorithm can be implemented for isolated word recognition using the *token passing* algorithm. Your description should include how the tokens are initialised and how the tokens are propagated at each time instance. A description of traceback is not required. [25%]

(b) Draw the phone network that allows arbitrary length sequences of the three words given above to be recognised. The network should require that silence, modelled using the `sil` model, is placed at the start and the end of each sequence and optionally between words. A linear lexicon should be used. [20%]

(c) Due to computational constraints *beam search* is to be used.

(i) Briefly describe beam search and why it is useful for decreasing the computational load of speech recognition systems. [20%]

(ii) A unigram language model is to be added to improve the recognition performance. Mark on the phone network of part (b) where the language model probability should be incorporated. Clearly state the reason for your choice. [15%]

(d) A tree-structured lexicon is to be used in addition to beam search. Draw the tree structured lexicon associated with the three word dictionary above (the language model probability may be ignored). How does tree-structuring the lexicon reduce the computational load of the recogniser? [20%]

END OF PAPER