ENGINEERING TRIPOS PART IIB

Tuesday 25 April 2006 9 to 10.30

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.*

There are no attachments.

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS
Engineering Data Book
CUED approved calculator allowed

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

1 The automatic analysis of speech signals can be based on the source-filter model of speech production.

(a) Draw a block diagram of the source-filter model. Briefly discuss the role each component plays in modelling speech. [20%]

(b) Give **two** of the simplifying assumptions underlying the application of the source-filter model in speech signal analysis. [15%]

(c) Describe how Block Processing is used in applying stationary spectral analysis techniques, such as the DFT, to a speech signal. [15%]

(d) Describe how to use the DFT to implement a triangular filterbank. [20%]

(e) Explain the term 'Mel-scale'. [10%]

(f) Explain how the triangular filterbank analysis procedure of part (d) can be modified to implement a Mel-scale filterbank. [20%]

2 This question concerns Linear Prediction Vocoders.

(a) Draw a block diagram of a simple Linear Prediction Vocoder. [15%]

(b) Describe the parameters which need to be encoded and transmitted for each frame of speech. Explain your choice of representation for the Linear Prediction filter used in your description of the LP Vocoder parameter set. [25%]

(c) The parameters of the Linear Prediction filter are extracted from a sequence of windowed frames of speech. Describe a parameter updating procedure which can be used in synthesis to reduce the effect of discontinuities introduced by the analysis procedure. [25%]

(d) In the analysis stage of the Linear Prediction Vocoder, an estimate of the pitch will be extracted from the prediction residual signal.

(i) Explain why the prediction residual signal is useful for pitch extraction. [15%]

(ii) Describe how to extract the pitch from the residual. [20%]

3 (a) Briefly explain the term *formant*. [10%]

(b) Briefly describe the articulatory features associated with the location of the first and second formants of vowels. [10%]

(c) Describe the articulatory process and spectral characteristics of voiced stops. [15%]

(d) Describe how to extract formants from the smoothed linear prediction spectrum. [20%]

(e) Describe how formant frequencies can be found by factoring the linear predictor polynomial. Justify the use of the roots of the linear predictor polynomial in estimating formant frequencies. [25%]

(f) Explain what is meant by a *formant tracker* and describe a heuristic procedure for constructing a formant tracker from formant measurements made from individual speech frames. [20%]

4 It is required to estimate the parameters of a *Hidden Markov Model* (HMM), $\mathcal{M}$, from some training data, $Y = y_1, \ldots, y_T$, using maximum likelihood estimation. The HMM has N emitting states, with state 0 being the initial state and state $N+1$ the final state. Associated with each emitting state is a Gaussian mixture distribution with M full covariance components.

(a) List the assumptions when using this HMM for speech recognition. [20%]

(b) Given that the forward probability is defined as

$$\alpha_j(t) = p(y_1, \ldots, y_t, s(t) = j | \mathcal{M})$$

where $s(t)$ denotes the state occupied at time t, how can a suitable backward probability, $\beta_j(t)$, be defined and recursively calculated? [20%]

(c) Show how $\alpha_j(t)$ and $\beta_j(t)$ can be used to find the posterior probability of state occupation, $L_j(t)$. [20%]

(d) Briefly describe how this method can be extended to find the posterior probability of mixture component occupation, $L_{jm}(t)$. [10%]

(e) Give Baum-Welch re-estimation formulae for the Gaussian covariance parameters and mixture weights. [20%]

How would these formulae be altered if multiple sequences of training vectors were to be used to estimate $\mathcal{M}$? [10%]

5 A large-vocabulary speaker-independent continuous speech recognition system is based on the use of context-independent phone hidden Markov models and a unigram language model. It is suggested that a suitable decoder would use the Viterbi algorithm and a beam search pruning mechanism with a linear lexicon organisation.

(a) Give a description, including equations, of the suggested method for decoding. Describe how the method could be improved by the use of a *tree-structured* lexicon. [50%]

(b) A number of modifications have been suggested to reduce the word error rate of the system by changing the acoustic and language models. For each modification listed below, briefly discuss the effect on the decoding problem and how the original decoder would need to be modified to incorporate these more complex models. In each case state with reasons whether a multi-pass decoder may be more appropriate.

(i) Word-internal triphone models. [10%]

(ii) Cross-word triphone models. [15%]

(iii) Bigram language model. [10%]

(iv) Trigram language model. [15%]

END OF PAPER