ENGINEERING TRIPOS PART IIB

Tuesday 26 April 2005 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.

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 parameters of a Linear Prediction (LP) model are to be estimated from a speech signal $s(n)$.

(a) Write the p^{th} order linear prediction in terms of the predictor coefficients $a_1, \ldots, a_p$, and write the expression for the prediction error sequence, $e(n)$. [10%]

(b) Show how the filter transfer function $\frac{S(z)}{E(z)}$ describes a system in which the speech signal is modelled as the output of a linear system excited by the prediction error sequence. Draw the corresponding linear filter. [20%]

(c) Referring to the Source-Filter model of speech production, explain the role of the prediction error sequence. [10%]

(d) The speech signal is bandpass filtered to a 20-4000 Hz range. Assuming that four formants are present in the signal, suggest an appropriate value of p and justify your answer. [10%]

(e) Explain how the LP spectrum is derived from the LP coefficients. [10%]

(f) Describe how to extract formant information from the LP spectrum. Describe how to extract formant information directly from the LP polynomial. [25%]

(g) Discuss how formant extraction based on LP analysis differs from formant extraction based on a smoothed discrete Fourier transform of speech. [15%]

2 (a) The analysis stage of a channel vocoder is shown in Fig. 1.

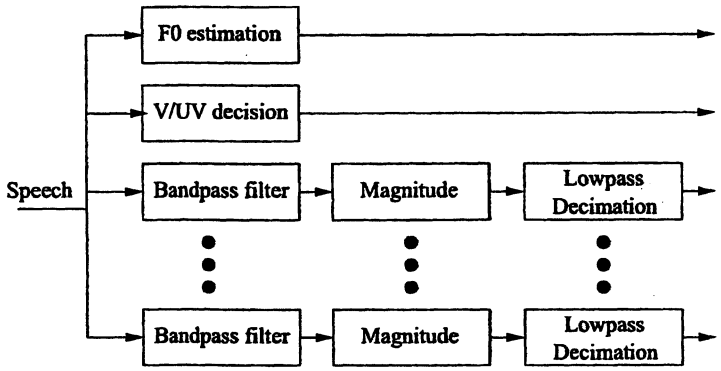

Fig. 1

(i) Draw the components of the synthesis stage. [15%]

(ii) Discuss the vocoder analysis and synthesis operations. [15%]

(iii) Discuss how overall synthesis quality could be improved by incorporating Mel frequency-scale filterbanks and exploiting auditory masking. [20%]

(b) Multipulse excitation is to be used for a linear prediction synthesis of speech. A single frame of N speech samples $s(n)$ is analysed to obtain a linear prediction filter with impulse response $h(n)$. The filter, and a single excitation impulse placed k samples from the start of the frame, are to be used to reconstruct the frame. By first finding the optimal amplitude of the impulse, show that the minimum sum squared reconstruction error is [35%]

$$\epsilon(k) = \left(\sum_{n=0}^{N-1} s^2(n)\right) - \frac{(\hat{r}_{sh}(k))^2}{\hat{r}_{hh}(0)}$$

$$\text{where} \quad \hat{r}_{sh}(k) = \sum_{n=0}^{N-1} s(n)h(n-k)$$

$$\text{and} \quad \hat{r}_{hh}(0) = \sum_{n=0}^{N-1} h^2(n-k)$$

(c) Compare multipulse excitation with the excitation in a CELP coder. [15%]

3 A training utterance, $\mathbf{y}_1 \ldots \mathbf{y}_T$, will be used in estimating the parameters of an N-state left-to-right Hidden Markov Model (HMM), $\mathcal{M} = \{N, \{a_{i,j}\}, \{b_j(\cdot)\}\}$, where each state has a single Gaussian observation distribution.

(a) Suggest a procedure to initialise the HMM parameters prior to training. [10%]

(b) Following initialisation, Viterbi training will be performed.

(i) Explain how the Viterbi algorithm can be used to find

$$\Phi_j(t) = \max_{s_1,\ldots,s_{t-1}} p(\mathbf{y}_1, \ldots, \mathbf{y}_t, s_t = j|\mathcal{M})$$

where s_t is the state occupied at time t. State how the algorithm is initialised. [15%]

(ii) Explain how to obtain the most likely state sequence $s_1, \ldots, s_T$. [10%]

(iii) Give the estimation equations for the parameters of the Gaussian distributions in terms of the segmented acoustic sequence. [20%]

(iv) Draw a flow chart describing the iterative application of the Viterbi training procedure. Suggest two different stopping criteria. [15%]

(c) After the Viterbi training of the single-Gaussian-per-state system has converged, the state output distributions are replaced by a mixture of Gaussians.

(i) Describe how Viterbi training can be used to estimate the parameters of the components of the mixture distributions. [20%]

(ii) Suggest a method to initialise this training procedure. [10%]

4 (a) Triphones are widely used in large-vocabulary speech recognition systems.

(i) Discuss why triphones are a good unit for use in a speech recognition system. [10%]

(ii) What issues must be addressed in using triphones in practical systems? [10%]

(b) Describe methods of state-level parameter tying for constructing triphones using

(i) Phonetic decision trees; [30%]

(ii) Bottom-up clustering. [20%]

In each case describe the basic operation of the technique, and list the advantages and disadvantages of the method for constructing triphone-based recognition systems.

(c) As part of the decision tree clustering procedure, a node splitting and merging cost must be defined. Single Gaussian distributions are used to model the data for the parent node p, and children nodes r and s. The change in likelihood for such a model can be computed from just the covariance matrices and the number of samples of data associated with p, r and s.

(i) Show how the covariance matrix at any node in the tree can be computed from knowledge of the state mean, covariance matrices and state occupation count of all triphone contexts, stating the assumptions that are made during this process. [20%]

(ii) Discuss the implications of using Gaussian mixtures to model each tree node during the decision tree clustering process. [10%]

5 (a) Define the term *perplexity* in statistical language modelling and show how perplexity may be computed given a language model and some test text. Explain why perplexity is a reasonable measure of language model quality. [20%]

(b) Describe a back-off N-gram language model, including parameter estimation using discounting and the computation of back-off weights. [30%]

(c) An N-gram language model has been trained on a corpus of text data. The training set perplexity is to be estimated solely from the language model parameters.

(i) Show how the perplexity of the training data can be computed for a unigram model. [15%]

(ii) Show how the training set perplexity calculation can be extended to a bigram model. [15%]

(d) Discuss how training set perplexity and test set perplexity differ. How would they vary with training set size, the number of language model parameters, and the use of discounting in language model estimation? [20%]

END OF PAPER