Version PCW/4

EGT3
ENGINEERING TRIPOS PART IIB

Friday 24 April 2015 2 to 3.30

Module 4F11

SPEECH AND LANGUAGE 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.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper.

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

1 (a) Draw a block diagram showing the generic architecture of a statistical speech recognition system based on sub-word acoustic units. State the function of each of the components. [15%]

(b) What are the basic modelling assumptions in using hidden Markov models (HMMs) to recognise speech? [10%]

(c) The forward-backward algorithm is used as part of the Baum-Welch estimation procedure for the parameters of an HMM.

(i) Define suitable forward and backward variables, and give methods for their recursive computation. [20%]

(ii) Show how the posterior probability of HMM state-occupation can be found. [10%]

(d) Describe how the Baum-Welch algorithm can be used to train the parameters of sub-word HMMs, such as context-independent phone HMMs, from a corpus of sentences with only word-level transcriptions. State how to deal with the case of multiple entries in the pronunciation dictionary for each word. How should the HMMs be initialised? [25%]

(e) It is proposed to improve the efficiency of the Baum-Welch training procedure by using a pruning mechanism. Two different types of pruning are considered, that are based on

(i) Forward probability values, or

(ii) Posterior probability of state occupation.

For each type of pruning, discuss how it may be implemented and the likely effect on computation for both short and long training utterances. [20%]

2 A speaker-independent continuous-speech recognition system based on hidden Markov models (HMMs) uses context-independent phone acoustic units, single Gaussian per state output probability density functions with diagonal covariance matrices, and mel-scale log energy filterbank features and a bigram language model. The system's vocabulary is 60,000 words. The word error rate is found to be too high for successful application deployment and a number of modifications have been proposed.

For each of the following changes, (1) give a brief description of the proposed change; (2) state how it would be expected to change the word error rate; and (3) describe any impact on the computational load and memory use of the resulting recognition system.

(a) Replace the log filterbank features by mel-frequency cepstral coefficients. [15%]

(b) Add differential features. [15%]

(c) Use of Gaussian mixture distributions. [20%]

(d) Use cross-word triphones with decision-tree based state tying. [25%]

(e) Use of a trigram language model. [25%]

3 In a text-to-speech synthesis system the linguistic analysis stage is followed by the speech synthesis stage.

(a) Describe briefly the motivation for this architecture. What is the input and output of each of these two stages? [10%]

(b) The linguistic analysis stage can itself be divided into multiple stages.

(i) Draw a block diagram of the linguistic analysis stages. [15%]

(ii) Describe the purpose of each linguistic analysis stage and give example input/output. [15%]

(c) Explain the term "synthesis unit" and briefly discuss why the choice of units is important. [10%]

(d) Give **three** ways in which the use of hidden Markov models (HMMs) in text-to-speech synthesis differs from the use of HMMs in speech recognition. [20%]

(e) Explain how the HMM independence assumptions can produce discontinuous feature sequences in HMM-based text-to-speech synthesis, and explain how generating consistent dynamic features overcomes this shortcoming. [20%]

(f) HMM state clustering is used in both text-to-speech synthesis and automatic speech recognition. Explain why the context clustering questions used in text-to-speech synthesis can make use of richer linguistic context. [10%]

4 (a) Define a semiring in the context of weighted finite state acceptors (WFSAs). [10%]

(b) Copy and complete Table 1 for the following three commonly used semirings. [20%]

Semiring	$\mathbb{K}$	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$
Probability					
Log					
Tropical					

Table 1

(c) Explain how a WFSA with arc weights in the probability semiring can be transformed for use under the log and the tropical semirings. Will the shortest path operation yield the same result in all three cases? [20%]

(d) The translation grammar in Table 2 is to be applied to the sentence '`a b c d e`', where `X1` and `X2` are non-terminals.

`a X1 e` $\rightarrow$ `A X1 E`	`b` $\rightarrow$ `B`
`a X1 c X2 e` $\rightarrow$ `A X2 C X1 E`	`c` $\rightarrow$ `C`
`b X1 d` $\rightarrow$ `B X1 D`	`d` $\rightarrow$ `D`
`a` $\rightarrow$ `A`	`e` $\rightarrow$ `E`

Table 2

(i) Draw the corresponding recursive transition network (RTN). [20%]

(ii) Draw the WFSA obtained by expansion of the corresponding RTN. [20%]

(e) Briefly give **two** advantages of using translation grammars. [10%]

END OF PAPER

THIS PAGE IS BLANK