ENGINEERING TRIPOS PART IIB

Tuesday 8 May 2012 9 to 10.30

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.*

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

pcw03

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

(a) What are the desirable attributes for the front-end analysis of a speech recognition system? [15%]

(b) It is proposed to use Mel-frequency cepstral coefficients (MFCCs) as the front-end representation. Describe the steps for obtaining MFCC parameters from a speech waveform sampled at 8 kHz. At each step describe how the feature vector is being altered and state the dimensionality of the representation. [30%]

(c) Briefly compare MFCCs to cepstral coefficients obtained from a linear prediction model of speech. [15%]

(d) It is now proposed to add both 'delta' and 'delta-delta' coefficients to the feature vector. Describe how these are obtained and why they may be useful in a HMM-based speech recognition system. What are the disadvantages of using such additional features? [20%]

(e) Speech recognition systems that operate with telephone channels need to adequately deal with variations in the channel characteristics between different callers. If the effect of the channel can be modelled as a fixed linear filter for each call, describe how the MFCC representation will be affected by the channel and suggest a method of channel compensation. [20%]

2 A large-vocabulary continuous-speech recognition system is to be constructed. Initially the system is to use a set of context-independent monophone hidden Markov models (HMMs), a unigram language model and a linear lexicon organisation. A recogniser based on the Viterbi algorithm is to be used.

(a) Draw a diagram of the phone-level network structure, including the language model probabilities, used in the Viterbi search. [15%]

(b) Describe how the network needs to be altered to include bigram language model probabilities. [10%]

(c) Describe how the Viterbi algorithm is used with this network structure. Include how the word-level result is generated. [20%]

(d) Describe how the beam search algorithm can be used to reduce the computational load of the Viterbi algorithm. [15%]

(e) The monophone HMMs are to be replaced by cross-word triphone HMMs.

(i) Explain what is meant by cross-word triphones and what are the potential advantages of such models. [15%]

(ii) How would the network structure be altered if cross-word triphone acoustic models are used in place of monophone HMMs? [15%]

(iii) Describe one possible way of further reducing the computational load while using cross-word triphone HMMs. [10%]

3 (a) Discuss the role of automatic performance metrics in the development of statistical machine translation systems. Explain why the performance metrics used in automatic speech recognition, such as the word error rate, are not suitable for use in developing machine translation systems. [20%]

(b) Give the formula for the BLEU score. For simplicity, you may omit the Brevity Penalty. Suggest how the BLEU score could be modified to reflect the presence of synonyms and discuss briefly how this could make the BLEU score less sensitive to minor variations in translation. [20%]

(c) A pair of sentences $f_1^J = f_1 \dots f_J$ and $e_1^I = e_1 \dots e_I$ are known to be translations. The word alignment hidden Markov model (HMM) takes the form

$$P(f_1^J, a_1^J, J|e_1^I) = P_L(J|I) \prod_{j=1}^{J} P_T(f_j|e_{a_j}) P(a_j|a_{j-1}, I)$$

where $a_1^J = a_1 \dots a_J$ specifies the word-to-word alignment.

(i) Define the forward and backward probabilities for this model. [10%]

(ii) Show how the forward and backward probabilities can be used to find the probabilities $P(f_1^J|e_1^I)$, $P(a_j = i, f_1^J|e_1^I)$ and $P(a_j = i|f_1^J, e_1^I)$. [20%]

(iii) Give the parameter update equation for the word translation distribution $P_T(f|e)$. [10%]

(iv) Discuss how the component distributions of the word alignment HMM could be made **context-dependent**. *Hint*: Consider the differences between the monophone and triphone acoustic HMMs used in speech recognition. [20%]

4 (a) Using semiring operations, explain how path weights are calculated by a Weighted Finite State Acceptor (WFSA) and explain how path weights contribute to the weight assigned to a string by a WFSA. [25%]

(b) The weight assigned to a string x by a WFSA C is denoted $[[C]](x)$. For a string x which is accepted by the WFSAs A and B, give the equations for $[[C]](x)$ where

(i) C is the union of A and B, i.e. $C = A \cup B$ [10%]

(ii) C is the intersection of A and B, i.e. $C = A \cap B$ [10%]

(iii) C is the concatenation of A and B, i.e. $C = A \otimes B$ [10%]

(c) A system is required that can transform uncased word sequences into correctly cased word sequences. As an example, the system should transform the uncased word sequence 'i went to london' into the cased word sequence 'I went to London'.

The uncased word sequence is denoted $u_1^n = u_1 \ldots u_n$ and the cased sequence is denoted $c_1^n = c_1 \ldots c_n$. A model for the probability distribution $P(u_1^n, c_1^n)$ is to be designed.

(i) Show how a model for $P(u_1^n, c_1^n)$ can be derived as

$$\prod_{i=1}^{n} P_T(u_i|c_i) P(c_i|c_{i-1})$$

stating any conditional independence assumptions used in the derivation. Note that c_0 is a sentence-start symbol. [15%]

(ii) Discuss how a large corpus of correctly cased text could be used to estimate the component distributions $P_T(u|c)$ and $P(c_i|c_{i-1})$. [15%]

(iii) Given an uncased word sequence u_1^n, the system produces a cased word sequence as

$$\underset{c_1^n}{\text{argmax}} \log \prod_{i=1}^{n} P_T(u_i|c_i) P(c_i|c_{i-1})$$

Discuss how this operation can be done using WFSAs and WFSA operations. [15%]

END OF PAPER

pcw03