Version PCW/3

EGT3
ENGINEERING TRIPOS PART IIB

Monday 28 April 2014 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

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

1 A large-vocabulary continuous-speech recognition system is to be designed. Initially the system is to use a set of context-independent monophone hidden Markov models (HMMs), a bigram 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) Briefly describe how the Viterbi algorithm is used with this network structure. Include how the word-level result is generated. [20%]

(c) If a trigram language model is to be used instead of a bigram, explain why the complexity of the network structure significantly increases. Suggest **two** ways of efficient decoding using a trigram language model. [15%]

(d) It is now suggested that the context independent models are replaced by cross-word triphone models.

(i) What is meant by cross-word triphones? Compare the modelling of co-articulation in a cross-word triphone system using Gaussian output distributions to a system using monophones with Gaussian mixture output distributions. [15%]

(ii) Why is parameter-tying usually used in constructing cross-word triphones? Explain how decision-tree state-tying operates, and what advantages it offers for estimating cross-word triphone systems. [25%]

(iii) What are the disadvantages of using cross-word triphone models? [10%]

2 Hidden Markov models (HMMs) are to be trained using Baum-Welch re-estimation for an isolated word speech recognition task with each word modelled initially by a single HMM with a Gaussian state output distribution. A particular N-state HMM, with parameter set λ, is to be trained on an observation sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2 \ldots \mathbf{o}_T\}$.

(a) Explain why a 39-dimensional feature vector based on mel frequency cepstral coefficients (MFCCs), log energy and the first and second differentials of these values is often used in HMM-based speech recognition systems. [15%]

(b) The forward probability is defined as

$$\alpha_j(t) = p(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, x(t) = j \,|\, \lambda)$$

where $x(t)$ denotes the state occupied at time t. Define a corresponding backwards probability, $\beta_j(t)$, and show how it can be computed recursively. [20%]

(c) How can $\alpha_j(t)$ and $\beta_j(t)$ be combined to find $L_j(t)$, which is defined as the posterior probability of state occupation given the observation sequence? [10%]

(d) Write down maximum-likelihood re-estimation formulae for the HMM mean parameters. How are the formulae extended if multiple training sequences are used? [15%]

(e) The HMM state output distributions are now modified to use a mixture of Gaussians with diagonal covariances.

(i) By viewing a mixture distribution as a set of parallel states, describe how the posterior probability of mixture component occupation can be computed. Write down the re-estimation formulae for both the mean vectors and the component weights. [25%]

(ii) After a number of iterations of training this model, it was discovered that all the variance parameters of one of the Gaussian components were very small. Explain why this might happen, what impact it might have on the trained HMM, and suggest one approach to avoid this problem. [15%]

3 (a) Give **two** reasons why n-gram language models are useful for speech and language processing systems. [10%]

(b) Give the equation for the probability assigned to a word sequence $w_1 \ldots w_M$ by an n-gram language model of order N. Derive this equation from the general form of a predictive language model, explaining all approximations used, and including sentence boundary markers. [20%]

(c) A trigram language model is to be estimated from a large text collection. Counts $f(w_i, w_j, w_k)$ are collected for all trigrams w_i, w_j, w_k in the collection.

(i) Give the equation for an *interpolated* language model based on a linear combination of trigram, bigram, and unigram counts. Discuss how the weights might be shared between different trigram contexts to allow for robust weight estimation. [20%]

(ii) Explain how *deleted interpolation* can be used to set values for the interpolation weights. [20%]

(iii) Give the equation for a trigram language model based on the *stupid back-off* approach. Compare the stupid back-off approach to the interpolated language model. [20%]

(d) Define *perplexity* and explain how it is computed over a corpus of test sentences. Discuss whether perplexity is suitable for assessing the quality of the interpolated language model and the stupid back-off language model. [10%]

4 A sentence-aligned parallel text corpus is to be used to estimate word alignment models for use in a statistical machine translation system.

(a) Define the terms *fertility* and *distortion* as they are used in word alignment. [15%]

(b) Compare the modelling power and computational tractability of IBM Model-1 and IBM Model-4. [15%]

(c) A sentence $e_1^I = e_1 \dots e_I$ and its translation $f_1^J = f_1 \dots f_J$ are taken from the parallel text corpus.

(i) Explain how the alignment process $a_1^J = a_1 \dots a_J$ indicates translation equivalence between words in the two sentences. [10%]

(ii) Derive the alignment likelihood $P(f_1^J, a_1^J, J|e_1^I)$ under IBM Model-2. [15%]

(iii) Derive an expression for calculation of the probability $P(a_j = i|f_1^J, e_1^I)$ under IBM Model-2 and show how it can be used to update the component distributions of the model. [25%]

(d) Suppose the parallel text consists of French and English travel documents. Suggest **two** automatic methods that could be used to create a English-French dictionary for word to word translation. Suggest a procedure to include phrase translations in the dictionary. [20%]

END OF PAPER

THIS PAGE IS BLANK