UNIVERSITY OF CAMBRIDGE

ENGINEERING TRIPOS PART IIB

Tuesday 20 April 2004 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 In a text-to-speech synthesis system, the linguistic analysis stage is followed by the speech signal synthesis stage.

(a) Briefly discuss the motivation for this architecture. What is the output of each stage? [10%]

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

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

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

(iii) For each of the linguistic analysis stages identify those that require expert knowledge and describe the form of this knowledge. [20%]

(c) The selection of an appropriate speech signal synthesis stage depends on several factors.

(i) Name **three** factors that should be considered when designing a system for a particular speech synthesis task. [10%]

(ii) Explain the term "synthesis unit" and briefly discuss why the choice of units is important. [15%]

(iii) Discuss the differences between formant synthesis and pitch synchronous overlap and add (PSOLA) in terms of the criteria named in (c)(i) and the selection of synthesis units. [15%]

2 In a speech analysis system, the source filter model of speech production forms the basis of the signal analysis.

(a) Draw a block diagram of the source filter model of speech production. Briefly discuss the model components and their physiological interpretation. [20%]

(b) What do the terms "*formant*" and "*pitch*" mean? Relate these terms to the block diagram drawn in (a). [15%]

(c) The most commonly used techniques for formant detection are either based on cepstral analysis or on linear prediction analysis.

(i) Discuss the fundamental differences between the two approaches. Briefly describe the implications for formant detection. [30%]

(ii) Describe how the real cepstrum of a speech signal can be computed. [20%]

(d) The analysis system is to be used with telephone bandwidth speech to detect four formant frequencies using cepstral analysis. Due to computational constraints speech frames are taken at 50ms intervals. Discuss potential problems of the configuration with the type of speech data used. [15%]

3 (a) What are the basic assumptions in using hidden Markov models (HMMs) for recognising speech data? [20%]

(b) A set of HMMs is to be trained using Baum-Welch re-estimation for an isolated word digit recognition task. A particular HMM, $\mathcal{M}$, with a single Gaussian per state output distribution, is to be estimated on a training sequence of observation vectors $\boldsymbol{y}_1 \ldots \boldsymbol{y}_T$. The forward probability is defined as

$$\alpha_j(t) = p(\boldsymbol{y}_1 \ldots \boldsymbol{y}_t, s(t) = j|\mathcal{M})$$

where $s(t) = j$ denotes that state j is occupied at time t.

(i) Define a corresponding backwards probability $\beta_j(t)$. [10%]

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

(iii) Hence, write down a re-estimation formula for the mean parameters of the HMM. [10%]

(c) If the HMMs are now to be trained on continuously spoken strings of digits, discuss how Baum-Welch re-estimation could still be used for this task. [20%]

(d) It is proposed to improve the efficiency of the Baum-Welch training procedure by using a pruning mechanism. Two stages of pruning are suggested based on

(i) forward probability values, and

(ii) the posterior probability of state occupation.

For each stage of pruning, discuss how a pruning mechanism might be implemented and comment on the effect on the mechanism on the training computation needed. [20%]

4 A 60,000 word vocabulary speech recognition system uses monophone HMMs with Gaussian mixture output distributions along with a unigram language model. The system uses a front-end analysis based on MFCCs with first and second order derivatives.

(a) For this speech recognition system

(i) What are the features of a basic search algorithm that finds the most likely recognition hypothesis using a standard linear network to represent the lexicon. [20%]

(ii) Describe how beam pruning could be included in the search. [15%]

(iii) Discuss the possible advantages of using a tree-based lexicon. [10%]

(b) It is proposed to improve the system by changing either the form of the acoustic models or the language model. For each of the four proposed modifications listed below: briefly describe the proposed change; discuss the effect on the computation required for a search with beam pruning; and state with reasons if a multi-pass strategy might be preferred.

(i) Word-internal triphones. [10%]

(ii) Cross-word triphones. [15%]

(iii) Syllable-based acoustic models. [15%]

(iv) Trigram language model. [15%]

5 N-gram language models are widely used in large vocabulary speech recognition systems.

(a) What is meant by an N-gram language model? Why are N-gram language models effective for speech recognition? [25%]

(b) What is the issue with directly using the maximum likelihood estimates of N-gram language model parameters for speech recognition systems? [10%]

(c) What is meant by discounting and back-off for N-gram language models? How can these methods improve the robustness of N-gram estimates? [30%]

(d) The number of parameters in a trigram language model is to be limited in order to save memory. Three alternative methods listed below have been proposed. For each proposed change suggest how effective it might be in controlling the language model size and the impact on recognition performance.

(i) Use of a bigram model instead of a trigram. [10%]

(ii) Only retain trigrams which occur more than once in the training corpus. [10%]

(iii) Only retain trigrams which give a higher probability than a use of a bigram back-off. [15%]

END OF PAPER