Digital Speech Processing, Midterm
Dec. 15, 2006, 10:10-12:10

- OPEN EVERYTHING
- 除專有名詞可用英文以外，所有文字說明一律以中文為限，未用中文者不計分
- Total points: 165
- Note that you don't need to be able to answer all the questions.

---

1. (10) Explain the concept of "Corpus-based Text-to-Speech Synthesis", how it works and why it is good.

2. (25) Given a HMM $\lambda = (A, B, \pi)$, an observation sequence $\overline{O} = o_1 o_2 ... o_t ... o_T$ and a state sequence $\overline{q} = q_1 q_2 ... q_t ... q_T$

   (a) (10) Formulate and describe the forward algorithm to evaluate $P(\overline{O} \mid \lambda)$. Explain how it works.

   (b) (10) Formulate and describe the Viterbi algorithm to find the best state sequence $\overline{q}^* = q_1^* q_2^* ... q_t^* ... q_T^*$ giving the highest probability $Prob(\overline{q}^*, \overline{O} \mid \lambda)$. Explain how it works.

   (c) (5) Now in order to recognize L words $w_1, w_2, ... w_L$ each with an HMM respectively, $\lambda_1, \lambda_2, ... \lambda_L$ it is well known that one can use either the forward algorithm or the Viterbi algorithm,

$$\arg\max_k P(\overline{O} \mid \lambda_k) \approx \arg\max_k P(\overline{q}^*, \overline{O} \mid \lambda_k)$$

   Explain why and discuss the difference between them.

3. (10) Write down the procedures for LBG algorithm and discuss why and how it is better than the K-means algorithm.

4. (10) Explain: in designing the decision tree to train tri-phone models, how the information theory is used to split a node n into two nodes a and b.

5. (10) In Classification and Regression Trees (CART), one can use composite questions instead of simple questions only. Write down what you know about this.

6. (10) The perplexity of a language source S is

$$PP(S) = 2^{H(S)}, H(S) = -\sum_i p(x_i)\log\left[p(x_i)\right],$$

where x_i is a word in the language, Explain why $PP(S)$ is the estimate of the branching factor for the language assuming a "virtual vocabulary"'?

7. (10) Explain the detailed principles and process for Katz smoothing.

8. (10) Given a set of events $\{x_i, i = 1, 2, \ldots, M\}$, $\{p(x_i), i = 1, 2, \ldots, M\}$ and $\{q(x_i), i = 1, 2, \ldots, M\}$ are two probability distributions. What is the Kullback-Leibler(KL) distance between $p(x_i)$ and $q(x_i)$ and what does it mean?

9. (10)
   (a) (5) What are the voiced/unvoiced speech signals and their time-domain waveform characteristics?
   (b) (5) What is pitch in speech signals and how is it related to the tones in Mandarin Chinese?

10. (10) The Hamming window has much lower sidelobes but wider mainlobe as compared to the rectangular window. Why is it good for front-end feature extraction for speech recognition?

11. (10) For large vocabulary continuous speech recognition, explain how the Viterbi algorithm can be performed such that the knowledge from the acoustic models, lexicon and language model can be efficiently integrated?

12. (15) Under what kind of condition a heuristic search is admissible? Show or explain why?

13. (15)
   (a) (8) Explain why Maximum Likelihood Linear Regression (MLLR) approaches can adjust a set of speaker-independent acoustic models to a new speaker with very limited quantity of adaptation data, but the performance is saturated at relatively lower accuracy?

   (b) (7) Explain why tree-structured classes can be helpful here.

14. (10) In Latent Semantic Analysis the elements w_{ij} of the word-document matrix $\bar{W}$ is

$$w_{ij} = (1 - \varepsilon_i) \frac{c_{ij}}{n_j}$$

Where c_{ij} is the number of times the word w_i occurs in the document d_j, n_j is the total number of words in d_j, and

$$\varepsilon_i = -\frac{1}{\log N} \sum_{j=1}^{N} \left(\frac{c_{ij}}{t_i}\right) \log\left(\frac{c_{ij}}{t_i}\right), \qquad t_i = \sum_{j=1}^{N} c_{ij},$$

where N is the total number of documents. Explain the meaning of all these parameters.