EE4-14
EE9-SO16

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2018

MSc and EEE/EIE PART IV: MEng and ACGI

SPEECH PROCESSING

Corrected copy

Wednesday, 9 May 10:00 am

Time allowed: 3:00 hours

There are FOUR questions on this paper.

Answer ALL questions.

All questions carry equal marks

Any special instructions for invigilators and information for candidates are on page 1.

Examiners responsible First Marker(s) : P.A. Naylor
Second Marker(s) : W. Dai

SPEECH PROCESSING

1. Consider a p^{th} order lossless tube model of the vocal tract in the human speech production system.

   a) Draw a fully labelled sketch of this model and briefly explain its key characteristics. [4]

   b) Consider the junction between two sections of the lossless tube model for which the cross-sectional area either side of the junction is different. Derive expressions relating the forward and reverse acoustic waves in the two tube sections. Write your expressions also in matrix form. [5]

   c) State the definition of the reflection coefficients in terms of the cross-sectional area of the tubes in this model. State an appropriate value for the reflection coefficient at the lips. [3]

   d) Sketch a signal flow graph for a complete lossless tube model employing 2 tube sections. The signal flow graph should contain delay elements, multipliers and addition nodes. [4]

   e) Now consider the glottal volume velocity as a function of time t, denoted $u_g(t)$. Draw an illustrative sketch of $u_g(t)$ over a representative time duration for a vowel.

   The definition of the LF Model includes

$$u'_g(t) = \begin{cases} e^{at}\sin(bt) & 0 \le t < t_e \\ c + de^{-ft} & t_e \le t < 1. \end{cases}$$

   Add to your above sketch of $u_g(t)$ a time-aligned sketch of the corresponding function $u'_g(t)$ and label $t = 0,\ t_e,\ 1$. [4]

2. In a speech recognition task, a particular hidden Markov model (HMM) with S states includes the transition probability from state i to state j which are denoted a_{ij}. Consider features computed from T frames representative of the speech, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ which are compared with the HMM. The output probability density is denoted by $d_i(\mathbf{x}_t)$ for frame t in state i. It can be assumed that frame 1 is in state 1.

a) Write down an expression for the probability density associated with the segment of the alignment path between frame $(t-1)$ in state $(s-1)$ to frame t in state s, given that frame t of the speech signal is in state s.
[3]

b) Let $B(t,s)$ be defined as the highest probability density that the model generates frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$. For the case when $t > 1$, explain fully how $B(t,s)$ can be expressed in terms of $B(t-1,i)$ for $i = 1, 2, \ldots, S$. State the value of $B(1,i)$.
[4]

c) Consider the inequality $B(t,s) < k \times \max(B(t,r))$. Explain how this inequality can be used to reduce the computational complexity of the speech recognizer and outline the criteria that should be used in choosing the factor k.
[6]

d) A 6-frame utterance is compared with a 4-state Hidden Markov model. Table 1 shows the output probability density of each frame in each state of the model and Figure 2.1 shows the state diagram of the model including the transition probabilities. Determine the value of $B(6,4)$ and the state sequence to which it corresponds given that the computation complexity reduction technique of part c) is employed with $k = 0.15$. At each appropriate step, indicate precisely the effect of pruning. Perform all your calculations to at least six decimal places. Draw and label the resulting alignment lattice.
[7]

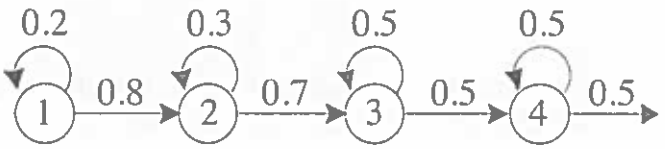

Figure 2.1

	Frame $\mathbf{x}_1$	Frame $\mathbf{x}_2$	Frame $\mathbf{x}_3$	Frame $\mathbf{x}_4$	Frame $\mathbf{x}_5$	Frame $\mathbf{x}_6$
State 1	0.5	0.2	0.5	0.5	0.5	0.5
State 2	0.5	0.6	0.3	0.1	0.5	0.5
State 3	0.5	0.5	0.1	0.6	0.6	0.5
State 4	0.5	0.5	0.5	0.1	0.4	0.4

Table 1

3. a) Consider a linear predictor with order p and prediction coefficients a_k. Consider this linear predictor being applied to a segment containing N samples of a speech signal $s(n)$, indexed beginning with $n = 0$. Let the prediction error be denoted E.

i) Show that

$$E = \sum_{n=p}^{N-1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k) \right)^2 .$$

[3]

ii) In the above expression for E, explain the reason why the lower limit of the first summation begins at $n = p$. [2]

iii) Show that the prediction coefficients that minimize E satisfy the equation

$$\mathbf{Ra} = \mathbf{b}$$

and, for this solution, give expressions for the elements of matrix $\mathbf{R}$ and vectors $\mathbf{a}$ and $\mathbf{b}$. [4]

b) i) State the main application of Line Spectral Frequencies (LSFs) in the context of speech processing and explain the advantages and disadvantages of LSFs in that application. [2]

ii) Consider a particular segment of speech for which the vocal tract transfer function is $V(z)$. Describe how LSFs would be obtained from $V(z)$. Include the mathematical details in your description. [4]

iii) Consider the vocal tract transfer function given by

$$V(z) = \frac{1}{1 - 0.95z^{-1} + 0.45z^{-2}}.$$

Find the LSFs corresponding to $V(z)$ and draw a representative sketch of the LSFs. [5]

4. a) i) Consider a single complex pole forming part of an all-pole system function $H(z)$. This single complex pole has radius $r < 1$. Show that the 3 dB bandwidth b of the corresponding resonant peak in the magnitude frequency response can be approximated by $b = 2(1 - r)$. Include an illustrative sketch and state the units of b. [3]

ii) Consider a linear time-invariant system with system function

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + a_4 z^{-4}}.$$

Show how the process known as *bandwidth expansion* can be applied to $H(z)$. Describe the main application of such bandwidth expansion in the context of speech processing. [3]

iii) With reference to part (a)(i), apply bandwidth expansion using the approximate expression for b as given in part a) to write a new expression for the bandwidth of a resonant peak after bandwidth expansion. Clearly define all symbols in the new expression. [3]

b) Consider a speech recognition system using mel-frequency cepstrum coefficients, $\mathbf{c}_t$ as features, computed in appropriately short time-frames t, with

$$\mathbf{c}_t = [c_{t,0},\ c_{t,1},\ \ldots,\ c_{t,P}].$$

i) Show mathematically how the mel-frequency cepstrum coefficients are calculated. [4]

ii) It is decided also to compute the first-order time derivatives of the mel-frequency cepstrum coefficients. The simple difference approximation of the first-order time derivative can be written

$$\Delta c_{t,p} = c_{t,p} - c_{t-1,p}$$

for $p = 0,\ 1,\ \ldots,\ P$. However, this was found to be too inaccurate.

Propose and derive a formula for an alternative improved scheme to find $\Delta c_{t,p}$ [3]

iii) State the 2 important advantages of using time-derivates of the features in a speech recognition system. [4]