IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2016

MSc and EEE/EIE PART IV: MEng and ACGI

Corrected copy

SPEECH PROCESSING

Thursday, 12 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

1. a) i)   For a discrete-time signal $x(n)$, write down the formula for its complex cepstrum $\hat{x}(n)$. [2]

       ii) Discuss the main consequences of performing signal processing in the complex cepstral domain and explain why this is potentially attractive in speech signal processing. [3]

    b) Consider a system represented by the rational z-transform

$$X(z) = \frac{A\prod_{k=1}^{M_a}\left(1 - a_k z^{-1}\right)\prod_{k=1}^{M_b}\left(1 - b_k^{-1}z^{-1}\right)}{\prod_{k=1}^{N}\left(1 - c_k z^{-1}\right)}$$

in which $|a_k|, |b_k|, |c_k| < 1$.

       i)   Identify the terms of this expression corresponding to

            •   zeros inside the unit circle in z,

            •   zeros outside the unit circle in z,

            •   poles inside the unit circle in z. [3]

       ii) Hence factorize $X(z)$ into 3 factors:

            •   a delay factor involving $z^{-\lambda}$,

            •   a maximum phase factor $X_{\max}(z)$

            •   a minimum phase factor $X_{\min}(z)$

            and write expressions for λ, $X_{\max}(z)$ and $X_{\min}(z)$. [3]

       iii) Given that $\hat{X}(z)$ is defined as the complex logarithm of $X(z)$, and adopting a suitable definition for the complex logarithm, write an expression for $\hat{X}(z)$ and hence write expressions for $\hat{x}(n)$ for the cases

            • $n = 0$

            • $n < 0$

            • $n > 0$.

            You may use the identity $\log(1-\alpha) = -\sum_{n=1}^{\infty}\dfrac{\alpha^n}{n}$ for $|\alpha| < 1$.

            [6]

       iv) Find the complex cepstrum $\hat{x}(n)$ of

$$x(n) = a^n u(n)$$

            in which $u(n)$ is the unit step function and $|a| < 1$. [3]

2.   a)   i)   Draw a block diagram that illustrates in overview the main components of an automatic speech recognition system.   [2]

         ii)   Explain what is meant by unigram, bigram and trigram language models. For a recognizer with a vocabulary of 30,000 words, state the number of probabilities that must be estimated for unigram, bigram and trigram language models.   [3]

         iii)   Consider a sequence of words denoted w and a speech signal denoted s. The conditional probability of the word sequence given the speech signal is $pr(w|s)$. Express this probability in terms of a language model and an acoustic model of speech production by employing Bayes' theorem.   [3]

   b)   In a particular application, the speech utterances are restricted to contain only the words 'red', 'blue' and 'green'. All utterances contain at least one of these words. Table 1 gives the frequencies $N(i, j)$ of each possible pair of successive words for which word i is followed by word j in a representative sample of utterances, i.e. $N(i, j)$ is the number of times word j follows word i. 'Start' and 'end' represent the start and end of an utterance.

The unigram probability for word j is denoted $p_u(j)$.

Bigram probabilities $p_b(i, j)$ for all relevant i, j are given by

$$p_b(i, j) = \begin{cases} \dfrac{N(i, j) - d(i)}{N(i)} & \text{for } N(i, j) > 2 \\ b(i) p_u(j) & \text{for } N(i, j) \leq 2 \end{cases}$$

where $N(i)$ is the number of times word i occurs.

The term $d(i)$ is defined as

$$d(i) = \begin{cases} 0 & \text{if } N(i, j) > 2 \text{ for all valid next words } j \\ 0.5 & \text{if } N(i, j) \leq 2 \text{ for all valid next words } j \end{cases}$$

and the term $b(i)$ is chosen so that

$$\sum_j p_b(i, j) = 1.$$

		Word j			
		'red	'blue'	'green'	'end'
Word i	'start'	10	10	30	0
	'red'	1	0	60	20
	'blue'	3	50	2	10
	'green'	67	5	10	20

Table 1 Word sequence frequencies

         i)   Calculate the unigram probabilities for each of the words 'red', 'blue', 'green' and 'end'.   [3]

ii)   Find the values $d(i)$ for all words i.   [2]

iii) Hence find the bigram probabilities $p_b(i, j)$.   [5]

iv)   Briefly explain why the term $d(i)$ is needed for bigram probability calculation and summarize its effect on the bigram probabilities.

[2]

3.

a) Speaker recognition aims to determine the most likely identity of an unknown talker from an example of their speech by comparing *features* of the unknown talker's speech to *features* previously extracted from the speech of a set of talkers with known identity. The identity of the unknown talker is determined as the closest matching known talker.

Describe the processing steps employed to extract appropriate *features* from speech. Discuss the desired properties of such *features*. Give examples of appropriate choices for such *features* for the task of speaker recognition. Include any relevant diagrams to clarify your description.

[6]

b) At a particular time, the *features* from part (a) are stored in a column vector $z = (z_1, z_2, \ldots, z_N)^T$ where the elements of z are assumed independent identically distributed Gaussian random variables having zero mean and unit variance with a probability density function

$$p(z_i) = (2\pi)^{-1/2} \exp\left(-\frac{z_i^2}{2}\right)$$

and ^{T} indicates matrix transpose.

i) Derive an expression for the probability density function of the vector z and show that the expected value $E\left(zz^T\right)$ is equal to the identity matrix. [4]

ii) Consider a non-singular $N \times N$ matrix A. If $x = Az$, obtain an expression for the covariance matrix $C = E\left(xx^T\right)$. [4]

iii) Derive an expression in terms of C for the natural log of the probability density function of the vector x, where x is defined in part (ii) above. [6]

Note: You may assume that if the probability density function of z is $f(z)$, then the probability density function of $x = Az$ is given by $f\left(A^{-1}x\right) \times |A|^{-1}$.

4. a) Speech enhancement can be performed in the STFT domain using the formula

$$Z(l,k) = H(l,k)Y(l,k)$$

in which l is the time-frame index, k is the frequency index, $Y(l,k)$ is the noisy speech signal and $H(l,k) = \sqrt{1 - \dfrac{\hat{\phi}_v(l,k)}{|Y(l,k)|^2}}$.

   i) State what $\hat{\phi}_v$ represents and explain two alternative methods for finding $\hat{\phi}_v$ in a speech enhancement system. Discuss the relative merits of each approach. [5]

   ii) What additional parameter(s) does model-based spectral enhancement depend on and how might it/they be estimated? [2]

   iii) The values of $Y(l,k)$ over a particular interval are given in Table 1. Assuming $\hat{\phi}_v = 0.16$ and is constant, on which time-frequency indices does the method fail and why? How can this be overcome?

[6]

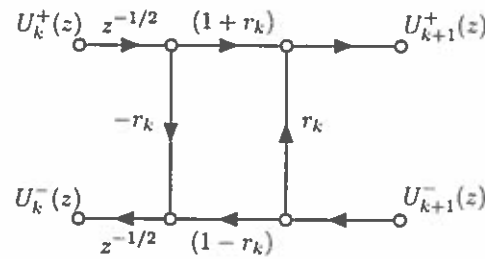|	$l=1$	$l=2$
$k=1$	0.332-0.009i	-0.277+1.695i
$k=2$	0.668+0.482i	-0.226+0.580i
$k=3$	-0.103-0.342i	0.254+0.359i
$k=4$	2.069+1.284i	-2.912+0.738i
$k=5$	-0.009+1.260i	-0.583-0.120i

Table 1 STFT analysis of noisy speech $Y(l,k)$

Figure 4.1

b) Consider the lossless tube model of the vocal tract for which a partial sketch diagram is shown in Fig. 4.1.

   i) Draw a complete labelled diagram of a lossless tube model of order 4 and describe the model and its relationship to the human speech production system. [4]

   ii) For the partial model shown in Fig. 4.1, find expressions for $U_k^+(z)$ and $U_k^-(z)$ in terms of $U_{k+1}^+(z)$ and $U_{k+1}^-(z)$. [3]