E4.14
SO16
ISE4.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2008

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Corrected Copy
(+ Fig 1 - Spectrogram)

Thursday, 15 May 10:00 am

Time allowed: 3:00 hours

There are SIX questions on this paper.

Answer FOUR 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) :	P.L. Dragotti

This page is intentionally left blank.

Special Instructions for Invigilators: None

Information for Candidates:

Numbers in square brackets against the right margin of the following pages are a guide to the marking scheme.

1. Consider the utterance: 'Water shortage'

Write down a phonetic transcription of this utterance using the IPA phonetic alphabet given below. [4]

Explain what is meant by the term 'glottal stop' in the context of phonetics. State under which circumstances a glottal stop might occur. Indicate where a glottal stop might occur in the above utterance. [4]

The spectrogram in Figure 1 shows the above utterance. State whether this spectrogram is a narrow-band or wide-band spectrogram and justify your statement with an explanation. Suggest whether this is male or female speech signal and give your reasons. [7]

Label the spectrogram using your phonetic transcription of the utterance showing the boundaries between the phonemes clearly. [5]

Submit the spectrogram with your answer book.

Table of Vowels and Consonants
IPA Alphabet with example words in English

b	bad, lab
d	did, lady
f	find, if
g	give, flag
h	how, hello
j	yes, yellow
k	cat, back
l	leg, little
m	man, lemon
n	no, ten
ŋ	sing, finger
p	pet, map
r	red, try
s	sun, miss
ʃ	she, crash
t	tea, getting
tʃ	check, church
θ	think, both
ð	this, mother
v	voice, five
w	wet, window
z	zoo, lazy
ʒ	pleasure, vision
dʒ	just, large

i	bead
ɪ	bid
ɛ	bed
æ	bad
ɜ	bird
ə	about
ʌ	bud
u	food
ʊ	good
ɔ	born
ɒ	body
ɑ	bard

This page is intentionally left blank.

2. A speech signal is represented by the samples $s(n)$, $n = 0, 1, \ldots, N-1$. The prediction error of a p^{th} order linear predictor is defined by

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

where the a_k are the prediction coefficients.

(a) Show that the prediction coefficients that minimize E satisfy $\mathbf{Ra} = \mathbf{b}$ where the $(i,j)^{th}$ element of $\mathbf{R}$ is given by $r_{i,j} = \sum_{n=p}^{N-1} s(n-i)s(n-j)$, the i^{th} element of $\mathbf{b}$ is given by $b_i = r_{i,0}$ and $\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}^T$. [6]

(b) In the context of linear prediction, show how an unstable filter can be made stable and explain the effects on the magnitude and phase of such a filter. [4]

(c) Explain the differences between covariance LPC and autocorrelation LPC and state their relative advantages. [4]

[*This question continues on the next page.*]

(d) A signal sampled at 8 kHz consists of white noise plus three cosine waves at 800 Hz, 1 kHz and 1.2 kHz of relative amplitudes 1, 1 and 0.2. *Figure 2* shows the filter spectra, with a normalised frequency axis, resulting from 10th order LPC analysis under the following conditions (not necessarily in this order):

(i) Covariance LPC with frame lengths of 3.5 ms and 80 ms

(ii)Autocorrelation LPC using Hamming windows of length 3.5 ms and 80 ms.

Identify which plot corresponds with each of the four conditions. Give reasons for your choice and explain the factors that cause the differences between the plots. [6]

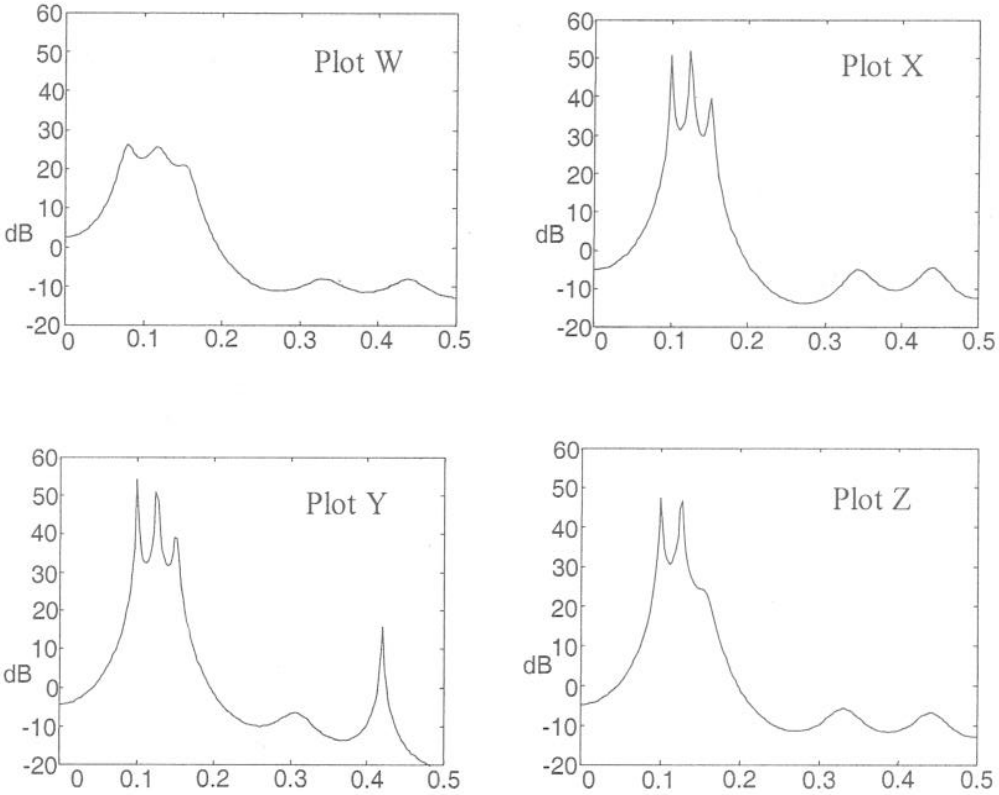

Figure 2 (All frequency axes are in normalized Hz)

3. *Figure 3* shows part of the encoder in a Code-Excited Linear Prediction (CELP) speech transmission system. The resynthesis error for sample n is defined as

$$e_k(n) = u(n) - g_k y_k(n) = u(n) - g_k \sum_{i=0}^{n} h(i) x_k(n-i) \text{ for } n = 0, 1, 2, \ldots, N-1.$$

In this expression, $x_k(n)$ denotes the n^{th} sample of the k_{th} codebook entry and g_k denotes the gain factor associated with that entry. The total resynthesis error for the frame is given by $E_k = \sum_{n=0}^{N-1} e_k^2(n)$.

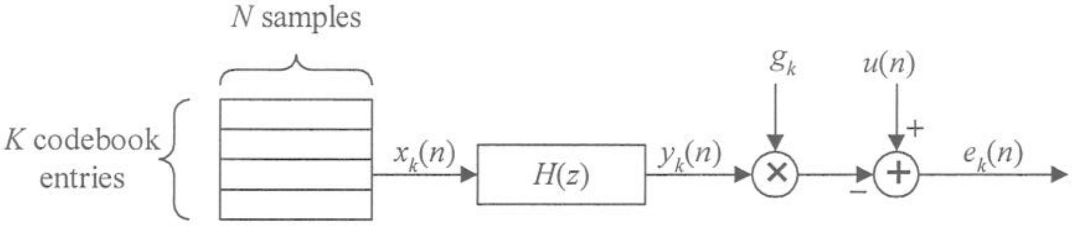

Figure 3

(a) Derive an expression for the value of g_k that minimizes E_k. Show that when this value of g_k is used, then E_k is given by [8]

$$E_k = \sum_{n=0}^{N-1} u^2(n) - \frac{\left(\sum_{n=0}^{N-1} u(n) y_k(n)\right)^2}{\sum_{n=0}^{N-1} y_k^2(n)}.$$

(b) Estimate the number of add/subtract and the number of multiply/divide operations needed to determine E_k for all codebook entries when $K = 1024$ and $N = 60$. [8]

(c) If consecutive codebook entries are related by $x_k(n) = x_{k-1}(n-1)$ for $n > 0$, obtain an expression relating $y_k(n)$ and $y_{k-1}(n-1)$. Estimate the number of add/subtract and the number of multiply/divide operations needed to determine E_k for all codebook entries when they are related in this way. [4]

4.

(a) Describe LSFs and their use in speech coding. Include in your description the advantages and disadvantages of using LSFs. [3]

Show how LSFs are computed from the vocal tract filter employing polynomial expressions of the form [3]

$$P(z) = A(z) + z^{-(p+1)}A^*(z^{*-1})$$
$$Q(z) = A(z) - z^{-(p+1)}A^*(z^{*-1})$$

where the terms have their usual meaning.

Show that the roots of $P(z)$ and $Q(z)$ lie on the unit circle in the z-plane. [4]

(b) State the desirable characteristics of feature vectors to be used for speech recognition and explain the extent to which these characteristics are found in Mel-frequency cepstral coefficients. [4]

Define Mel-frequency cepstral coefficients and give a detailed description of how these coefficients are computed using appropriate illustrations. [3]

You may use the definition of Mel in terms of frequency f Hz as

$$\mathrm{Mel(f)} = 2595 \log_{10}(1 + \mathrm{f} / 700) \ .$$

Draw a labelled sketch of the magnitude response as a function of frequency in Hz of each channel of a Mel-frequency filter bank having 4 bands uniformly spaced in Mel and covering the range 0 to 4000 Hz. [3]

5. *Figure 4* shows a block diagram of a Differential Pulse Code Modulation (DPCM) speech coder in which the input and output signals are labelled $s(n)$ and $d(n)$ respectively.

(a) Describe how each of the three blocks in the diagram enables a reduction in bit rate for a given signal quality. [6]

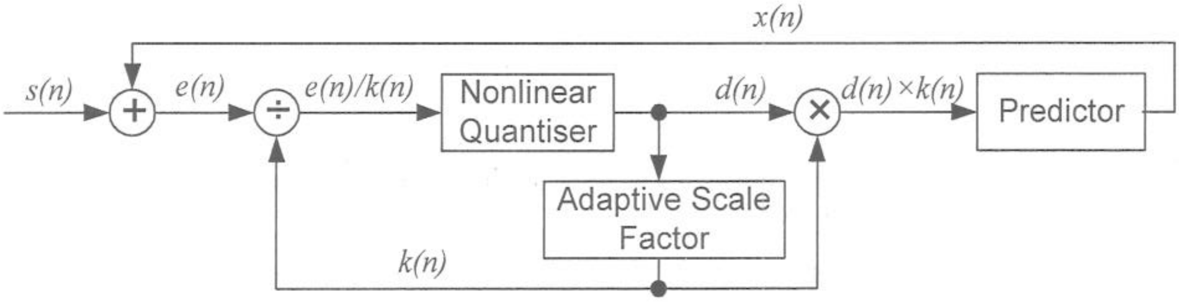

Figure 4

[*This question continues on the next page.*]

(b) The block diagram of *Figure 5* represents an adaptive quantizer. An incoming sample $w(n)$ is divided by a scale factor, $k(n)$, before being quantized to one of 3 code values by the block marked Q. The quantization thresholds are at $x(n) = \pm½$. The inverse quantiser, Q^{-1}, generates a signal $y(n)$ which only takes values of −1, 0 and +1. This is multiplied by $k(n)$ to give the output signal $z(n)$.

Before processing the following sample, the scale factor, $k(n)$, is increased or decreased by a factor that depends on the code value at the output of the Q block.

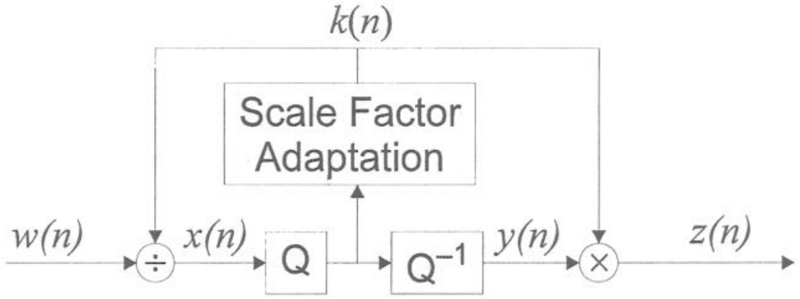

Figure 5

The probability density function for the input signal $w(n)$ is given by:

$$p(w) = ½e^{-|w|}$$

The error at sample number n is defined by $e(n) = z(n) - w(n)$. Calculate the expected value of $e^2(n)$ as a function of $k(n)$ and show that its minimum value occurs when $k(n)$ = 2. [7]

(c) The scale factor is adapted at each sample according to the following rule:

$$k(n+1) = \begin{cases} k(n) \times e^{-a} & \text{for } y(n) = 0 \\ k(n) \times e^{+b} & \text{for } y(n) = \pm 1 \end{cases}$$

where a and b are positive constants. Determine the ratio b/a such that the expected value of

$$\log_e(k(n+1)) - \log_e(k(n))$$

is zero when k has the value 2. Assume that the signal $w(n)$ is uncorrelated. [7]

Note: The following integrals may be helpful:

$$\int_t^{\infty} we^{-w}dw = (1+t)e^{-t}, \qquad \int_t^{\infty} w^2e^{-w}dw = (2+2t+t^2)e^{-t}$$

6. In a speech recognition system, an observation sequence comprising T frames $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ is to be compared with a hidden Markov model containing S states. For the hidden Markov model, the probability density of generating an observation frame $\mathbf{x}$ from state s is $d_s(\mathbf{x})$ and the transition probability from state i to state j is a_{ij}.

$P(t,s)$ is defined to be the total probability density of all possible alignments of frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ with $\mathbf{x}_1$ in state 1 and $\mathbf{x}_t$ in state s. Similarly, $Q(t,s)$ is defined to be the total probability density of all possible alignments of frames $\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T$ given that $\mathbf{x}_t$ is in state s and $\mathbf{x}_T$ is in state S. The probability density $d_s(\mathbf{x}_t)$ is included in $P(t,s)$ but not in $Q(t,s)$.

(a) Show that $P(t,s)$ can be expressed in terms of $P(t-1,k)$ for $k=1,\ldots,S$ and that $Q(t,s)$ can be expressed in terms of $Q(t+1,k)$ for $k=1,\ldots,S$. [9]

(b) A six-frame observation is to be compared with a four-state model whose transition probabilities are shown in *Figure 6*. The values of $d_s(\mathbf{x}_t)$ are given in the following table:

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
State 1	0.5	0.4	0.5	0.2	0.1	0.1
State 2	0.4	0.5	0.8	0.6	0.3	0.8
State 3	0.2	0.8	0.2	0.8	0.2	0.2
State 4	0.5	0.4	0.5	0.2	0.5	0.8

Determine the total probability of all alignments of the observation with the model for which frame 3 is in state 2. Show all your working. [11]

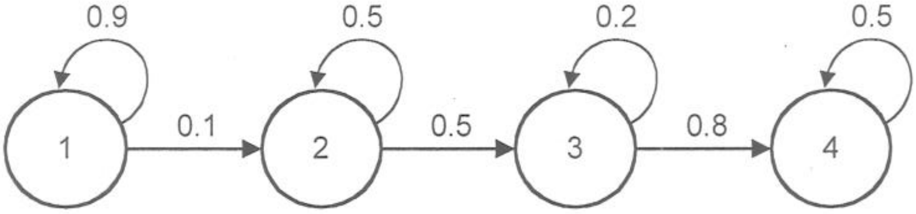

Figure 6