E4.14
SO16
ISE4.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2006

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Monday, 15 May 10:00 am

Time allowed: 3:00 hours

There are SIX questions on this paper.

Corrected Copy

Answer FOUR questions.

All questions carry equal marks

The figure for question 6 is on a separate sheet. If you answer this question please write your candidate number on the sheet and tie it into your answer book.

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

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 linear predictive coding of a speech signal $s(n)$ giving a vocal tract filter $V(z)$ and a prediction error filter $A(z)=\frac{1}{V(z)}$ with coefficients a_j.

(a) Explain why Line Spectrum Frequencies are more suitable for transmission than either the coefficients a_j or the roots of $A(z)$. [3]

(b) Write down symmetric and antisymmetric polynomials, $P(z)$ and $Q(z)$ respectively, formed from $A(z)$. Hence, state how the Line Spectrum Frequencies, f_l, are computed from the coefficients a_j and comment on the computational complexity of the computation. [5]

(c) Prove that the roots of $P(z)$ and $Q(z)$ lie on the unit circle in the z-plane. [6]

(d) Given $A(z)=1-0.8z^{-1}+0.6z^{-2}$, find the corresponding Line Spectrum Frequencies and make any relevant comments. [6]

2.

(a) In a speech recognizer, a speech utterance consists of T frames, $\mathbf{x}_1 \cdots \mathbf{x}_T$, and is compared with a Hidden Markov model having S states. The transition probability from state i to state j of the model is denoted by a_{ij} and the output probability density of frame t in state i is denoted by $d_i(\mathbf{x}_t)$.

Consider that frame 1 is in state 1 and frame t is in state s. Then, $B(t,s)$ is defined to be the highest probability density that the model generates frames $\mathbf{x}_1 \cdots \mathbf{x}_t$.

Explain fully how, for $t>1$, $B(t,s)$ can be expressed in terms of $B(t-1,i)$ for $i=1,2,\ldots,S$. Indicate the values that should be given to $B(1,i)$. [6]

(b) Consider the inequality $B(t,s) < k \times \max_r(B(t,r))$. 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. [5]

(c) A 6-frame utterance is compared with a 4-state Hidden Markov model. Table 1 shows the output probability density of each frame from each state of the model and Figure 1 shows the state diagram of the model including the transition probabilities. [9]

Using a pruning factor of $k=0.1$, determine the value of $B(6,4)$ and the state sequence to which it corresponds. You should perform all your calculations to six decimal places.

	Frame					
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
State 1	0.5	0.2	0.5	0.5	0.5	0.5
State 2	0.5	0.7	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

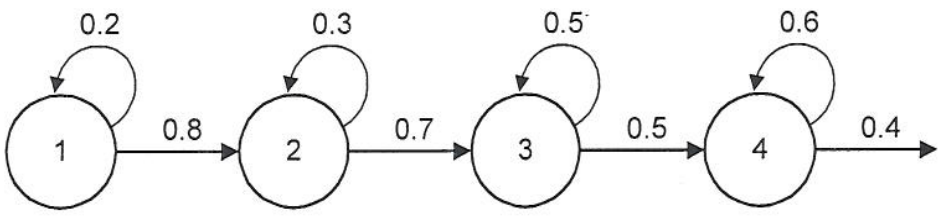


Figure 1

3.

A feature extraction subsystem to be used in a speech recognition system is shown in Figure 2. The input speech signal, $s(n)$, is sampled at 8 kHz and is divided into overlapping frames of length 24 ms with a frame rate of 100 Hz. The data labelled in the diagram as $\mathbf{q}$, $\mathbf{m}$ and $\mathbf{c}$ are defined to be $\mathbf{q} = [q_1, q_2, \ldots, q_N]^T$, $\mathbf{m} = [m_1, m_2, \ldots, m_N]^T$ and $\mathbf{c} = [c_1, c_2, \ldots, c_K]^T$ with $N = 24$ and $K = 12$. The block labelled DCT performs the discrete cosine transform defined by

$$c_k = \sum_{p=1}^{N} m_p \cos\left(\pi k \left(p - \tfrac{1}{2}\right) / N\right).$$

The mel filterbank calculates a vector of coefficients by applying a set of 24 filters whose peak frequencies in Hz are given by

$$f_i = \text{mel}^{-1}\left(0.04i \times \text{mel}(4000)\right) \qquad \text{for } i = 1, 2, \ldots, 24$$

where the mel frequency scale is defined by

$$\text{mel}(f) = 2295 \log_{10}(1 + f / 700\text{Hz}).$$

Each filter is triangular in shape, falling to zero at the peak frequencies of the two adjacent filters.

(a) State the processing function of each of the blocks in Figure 2 and discuss the strengths and weakness of this feature extraction subsystem for speech recognition. [5]

(b) Determine the number of DFT outputs that contribute to (i) m_2 and (ii) m_{20}. [4]

(c) Now consider that the speech signal, $s(n)$, has been corrupted by being passed through a channel whose effect is that of a filter having a frequency response $H(j\omega)$. Describe the way in which the mel-cepstral coefficients c_k are affected by the channel. Derive a formula for the amount by which c_k is altered under the assumption that each filterbank output is scaled by the channel response at its peak frequency. [8]

(d) Explain why the performance of speech recognition systems operating over the telephone can often be improved by subtracting from **c** its time-average. [3]

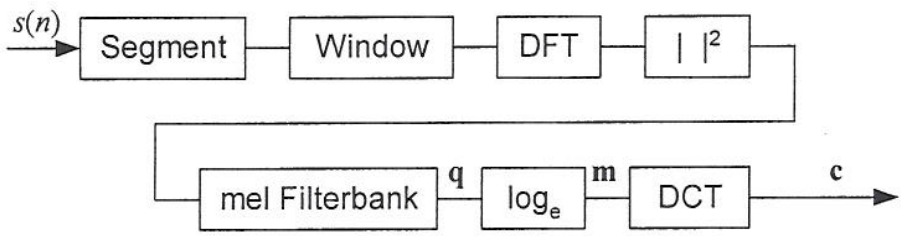


Figure 2

4.

(a) Briefly describe what is meant by language modelling in speech recognition. [2]

Explain what is meant by unigram, bigram and trigram language models. In each case give the number of transition probabilities that must be estimated for a recogniser with a vocabulary of 20,000 words. [3]

The conditional probability of a particular word sequence, w, given an input speech utterance, s, is denoted $pr(w|s)$. Describe how Bayes' Theorem may be used to express this probability in terms of a language model and an acoustic model of speech production. [3]

(b) Consider security passwords consisting only of the digits 1, 2 and 3 and containing at least one digit. Thus "23", "1", "333" and "1212321" are all valid passwords. Table 2 gives the frequencies of each possible pair of successive words in a representative sample of passwords. "Start" and "end" represent the start and end of a password.

(i) Calculate the unigram probabilities for each of 1, 2, 3 and "end". [3]

(ii) Bigram probabilities are calculated according to the following formula:

$$p(i,j) = \begin{cases} \dfrac{N(i,j) - d(i)}{N(i)} & \text{for } N(i,j) > 2 \\ b(i)p(j) & \text{for } N(i,j) \le 2. \end{cases}$$

The number of occurrences in the training data of word i followed by word j is $N(i,j)$ and the bigram probability of word j given word i is $p(i,j)$. The number of occurrences in the training data of word j is $N(j)$ and $p(j)$ is the unigram probability for word j. The discounting factor, $d(i)$ equals zero if $N(i,j) > 2$ for all valid next states j, and equals 0.5 if $N(i,j) \le 2$ for any valid j. The factor $b(i)$ is chosen so that

$$\sum_j p(i,j) = 1$$

Calculate the values of $d(i)$ and $b(i)$ for each i and all the bigram probabilities $p(i,j)$. [5]

(iii) Explain why a different formula is used in (ii) above according to whether $N(i,j)$ is greater than or less than 2. What is the effect of the discounting scheme on bigram probabilities? [4]

$N(i,j)$		Following Word, j			
		1	2	3	end
Initial Word, i	start	10	10	30	0
	1	1	0	60	20
	2	3	50	2	10
	3	67	5	10	20

Table 2

5.

(a) Briefly describe the p^{th} order lossless tube model of the vocal tract and give a labelled diagram of the model. [5]

(b) State how the cross-sectional area of the tubes is related to the reflection coefficients for the model. [2]

(c) Derive the transfer function of the model in terms of the reflection coefficients by considering segment delays and segment junctions. [4]

(d) Consider the signal flow graph representation of a section of the model at the boundary between tube k and tube $k+1$ shown in Figure 3 in which the elements labelled τ_k and τ_{k+1} represent time delays.

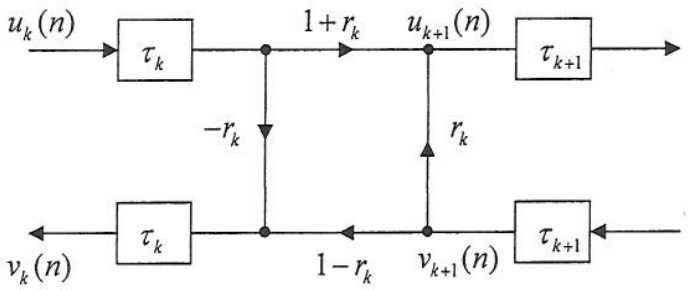


Figure 3

(i) State the value of the time delays τ_k and τ_{k+1} in samples. [1]

(ii) Draw the signal flow graph for a complete lossless tube model with order $N=2$ using the boundary conditions for the glottis and lips respectively of [3]

$$u_1(n) = \frac{1+r_0}{2} u_G(n) + r_0 v_1(t) \text{ and}$$

$$v_N(n+\tau_N) = -r_L u_N(n-\tau_N) \qquad u_L(n) = (1+r_L) u_N(n-\tau_N).$$

The volume velocity at the glottis and lips are $u_G(n)$ and $u_L(n)$ respectively.

(iii) Derive the transfer function of the model directly from the signal flow graph in terms of the reflection coefficients and compare your answer to (c) above. [5]

6.

Consider the utterance: 'Winter Olympics'

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

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

The spectrogram in Figure 4 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 and give your reasons. [7]

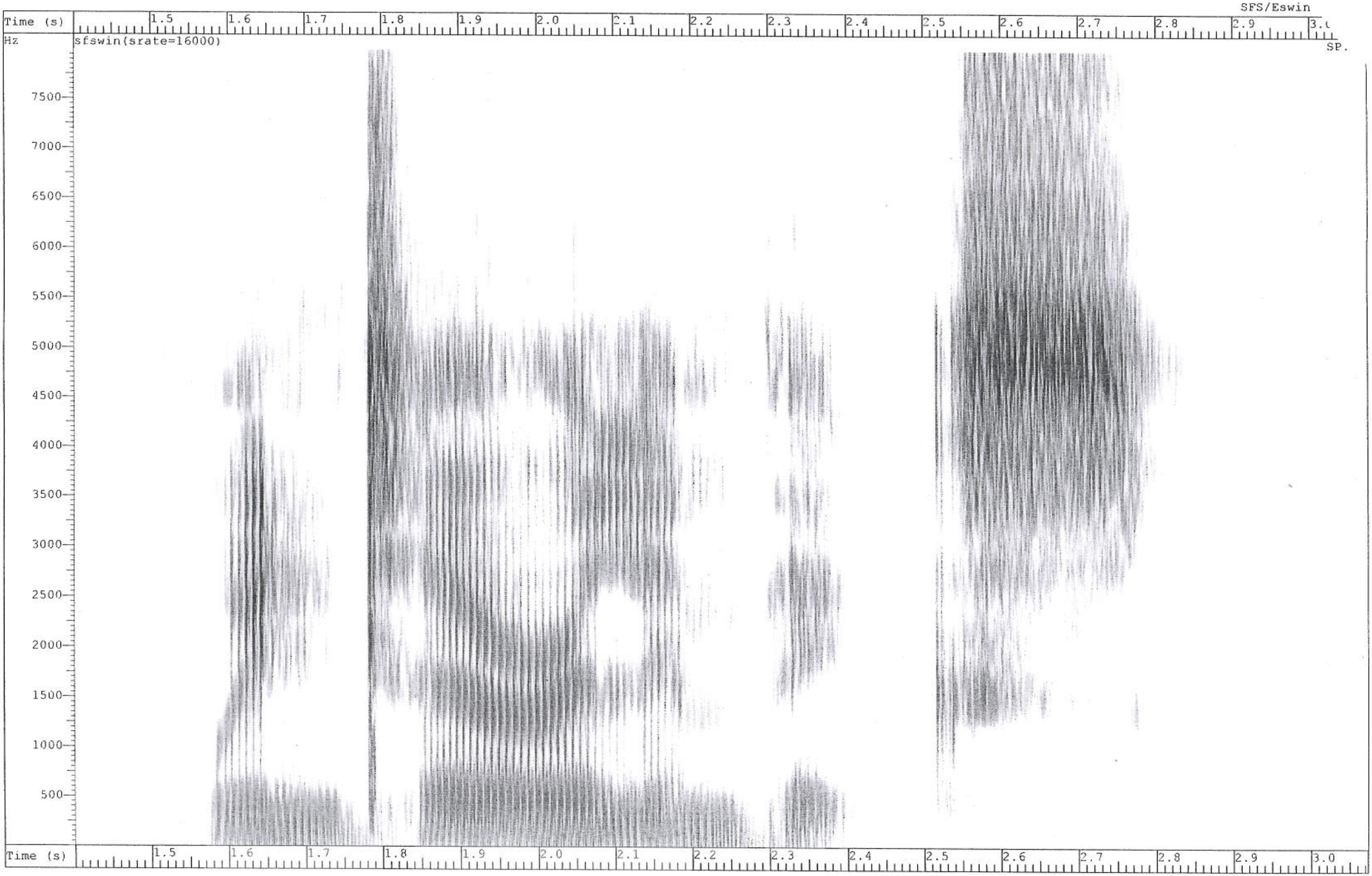
Label the spectrogram using your phonetic transcription of the utterance showing the boundaries between the phonemes clearly. [5]

Submit the spectrogram with your answer book.

Table of 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

Figure 4


SFS/Eswin
Time (s)
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
Hz
sfswin(srate=16000)
SP.
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
Time (s)
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0

E4.14 / IE4.17 / 5016

1. (a)

The major advantages of the LSF coefficients are:

- Less sensitive to quantisation errors than the LPC coefficients
- Unlike the pole positions, they have a natural order. This makes coding much more efficient since each coefficient normally lies in a small range.
- Stability is easy to detect and is preserved when interpolating between two sets of coefficients.
- They are closely related to formant frequencies and amplitudes and therefore naturally encode the most important aspect of the spectrum.

(b)

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A^*(z^{*-1}) \\ &= 1 - (a_1 + a_p)z^{-1} - (a_2 + a_{p-1})z^{-2} - \ldots - (a_p + a_1)z^{-p} + z^{-(p+1)} \\ Q(z) &= A(z) - z^{-(p+1)} A^*(z^{*-1}) \\ &= 1 - (a_1 - a_p)z^{-1} - (a_2 - a_{p-1})z^{-2} - \ldots - (a_p - a_1)z^{-p} - z^{-(p+1)} \end{aligned}$$

If the roots of $P(z)$ are at $e^{2\pi j f_i}$ for $i = 1,3,\ldots$ and those of $Q(z)$ are at $e^{2\pi j f_i}$ for $i = 0,2,\ldots$ with $f_{i+1} > f_i \geq 0$ then the LSF frequencies are defined as $f_1, f_2, \ldots, f_p$.
Note that it is always true that $f_0 = +1$ and $f_{p+1} = -1$.

Computational load is relatively high because it is necessary to find the angle of the roots around the unit circle.

(c)

Proof

$$\begin{aligned} P(z) = 0 \quad &\Leftrightarrow \quad A(z) = -z^{-(p+1)} A^*(z^{*-1}) \quad \Leftrightarrow \quad H(z) = -1 \\ Q(z) = 0 \quad &\Leftrightarrow \quad A(z) = +z^{-(p+1)} A^*(z^{*-1}) \quad \Leftrightarrow \quad H(z) = +1 \end{aligned}$$

$$\text{where} \quad H(z) = \frac{A(z)}{z^{-(p+1)} A^*(z^{*-1})} = z \prod_{i=1}^{p} \frac{(1 - x_i z^{-1})}{z^{-1}(1 - x_i^* z)} = z \prod_{i=1}^{p} \frac{(z - x_i)}{(1 - x_i^* z)}$$

here the x_i are the roots of $A(z)=V^{-1}(z)$. It turns out that providing all the x_i lie inside the unit circle, the absolute values of the terms making up $H(z)$ are either all > 1 or else all < 1. Taking | | of a typical term:

$$\left|\frac{(z-x_i)}{(1-x_i^* z)}\right| > 1 \quad \Leftrightarrow \quad \left|1-x_i^* z\right| < \left|z-x_i\right|$$

$$\Leftrightarrow \quad \left(1-x_i^* z\right)\left(1-x_i^* z\right)^* < \left(z-x_i\right)\left(z-x_i\right)^*$$

$$\Leftrightarrow \quad \left(1-x_i^* z\right)\left(1-x_i z^*\right) < \left(z-x_i\right)\left(z^*-x_i^*\right)$$

$$\Leftrightarrow \quad 1-x_i^* z-x_i z^* + x_i x_i^* zz^* < zz^* - x_i^* z - x_i z^* + x_i x_i^*$$

$$\Leftrightarrow \quad 1-x_i x_i^* - zz^* + x_i x_i^* zz^* < 0$$

$$\Leftrightarrow \quad \left(1-\left|x_i\right|^2\right)\left(1-\left|z\right|^2\right) < 0 \quad \Leftrightarrow \quad \left|z\right| > 1 \quad \text{since each } \left|x_i\right| < 1$$

Thus each term is greater or less than 1 according to whether |z|>1 or |z|<1. Hence |*H*(z)|=1 if and only if |z|=1 and so the roots of *P(z)* and *Q(z)* must lie on the unit circle.

(d)

$$A(z) = 1 - 0.8z^{-1} + 0.6z^{-2}$$

$$P(z) = 1 - 0.2z^{-1} - 0.2z^{-2} + z^{-3}$$

$$Q(z) = 1 - 1.4z^{-1} + 1.4z^{-2} - z^{-3}$$

The LSFs are found from factorizing the cubic equations.
$Q(z)$ can be factorized by noting that $z = 1$ is always a root of $Q(z)$. We can therefore write

$$Q(z) = 1 - 1.4z^{-1} + 1.4z^{-2} - z^{-3} = 0$$

$$(z-1)(z^2 - 0.4z + 1) = 0$$

And so the roots are at $z = 1$ and at $z = 0.2 \pm j\sqrt{3.84}/2$.

Thus the LSFs from $Q(z)$ are: $\begin{cases} 0 \\ \frac{1}{2\pi}\tan^{-1}\left(\frac{0.98}{0.2}\right) = 0.218 \\ \frac{1}{2\pi}\tan^{-1}\left(\frac{-0.98}{0.2}\right) = 0.782 \end{cases}$

$P(z)$ can be factorized by noting that $z = -1$ is a root of $P(z)$. We can therefore write

$$P(z) = 1 - 0.2z^{-1} - 0.2z^{-2} + z^{-3} = 0$$

$$(z+1)(z^2 - 1.2z + 1) = 0$$

And so the roots are at $z = -1$ and at $z = 0.6 \pm j\sqrt{2.56}/2 = 0.6 \pm j0.8$.

Thus the LSFs from $P(z)$ are: $\begin{cases} 0.5 \\ \frac{1}{2\pi}\tan^{-1}\left(\frac{0.8}{0.6}\right) = 0.1476 \\ \frac{1}{2\pi}\tan^{-1}\left(\frac{-0.8}{0.6}\right) = 0.8524 \end{cases}$

It can be verified that the LSF are interlaced.

2. (a)

The best path for $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ with frame t in state s must have frame t–1 in one of the states, say state i. Since the sub-path $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}$ must also be optimum, we must have:

$$B(t,s) = B(t-1,i) \times a_{is} \times d_s(\mathbf{x}_t)$$

Since $B(t,s)$ represents the probability density of the *best* path, we must have:

$$B(t,s) = \max_{1 \le i \le S} \left(B(t-1,i) \times a_{is} \times d_s(\mathbf{x}_t) \right)$$

Since we require frame 1 to be in state 1, $B(1,s) = d_1(\mathbf{x}_1)$ if s=1 and 0 otherwise.

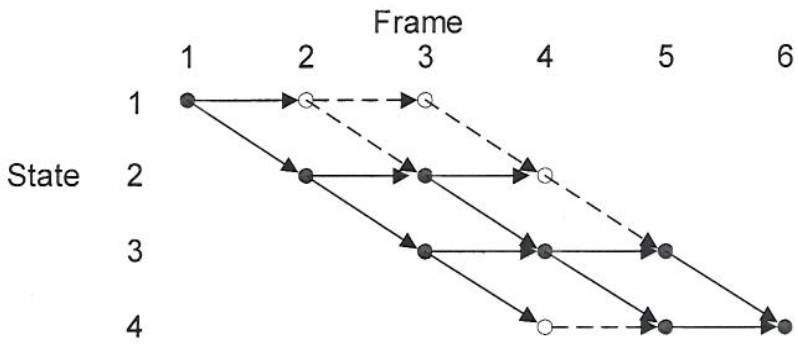
(b)

The maximisation of part (a) above involves a great deal of calculation if S is large. We can reduce this by eliminating from consideration, values of i for which $B(t-1,i)$ is small. To achieve this, after calculating $B(t,s)$ for all s, we delete (or "prune") all those less than $k \times \max_s B(t,s)$ for some factor k<1.

The choice of k is a compromise. If k is too large, the computational saving will be slight since few states will be pruned. If k is too small then it is possible that the pruning will delete one of the states that in fact lies along the optimum path. [3]

(c)

The possible paths through the State-Frame lattice are shown below. The white circles show nodes that are pruned.



B(1,1) = 0.5
B(2,1) = 0.5 × 0.2 × 0.2 = 0.002 : **pruned** because < 0.1 B(2,2)
B(2,2) = 0.5 × 0.8 × 0.7 = 0.28
B(3,2) = 0.28 × 0.3 × 0.3 = 0.0252
B(3,3) = 0.28 × 0.7 × 0.1 = 0.0196
B(4,2) = 0.0252 × 0.3 × 0.1 = 0.000756 : **pruned** because < 0.1 B(4,3)
B(4,3) = max{0.0252 × 0.7 = 0.01764, 0.0196 × 0.5 = 0.0098} × 0.6 = 0.010584
B(4,4) = 0.0196 × 0.5 × 0.1 = 0.00098 : **pruned** because < 0.1 B(4,3)
B(5,3) = 0.010584 × 0.5 × 0.6 = 0.003175
B(5,4) = 0.010584 × 0.5 × 0.4 = 0.002117
B(6,4) = max{0.003175 × 0.5 = 0.001588, 0.002117 × 0.6 = 0.0012702} × 0.4 = 0.0006352

The best path is 1,2,2,3,3.4 This may be obtained by tracing the underlined values backwards from B(6,4).

3.

(a)

Segment	Divides the incoming speech into overlapping frames of 10 to 30 ms duration. The duration needs to be long enough to give adequate spectral resolution but short enough to detect speech sounds of short duration. The DFT calculation is more efficient if the segment length is an exact power of 2.
Window	The segment is multiplied by a window function in the time domain (omitting this step is equivalent to choosing a rectangular window). The speech spectrum is convolved with that of the window function. The choice of window function is a compromise between the width of the central lobe (and hence the spectral resolution) and the height of the sidelobes (which cause artefacts in the spectrum).
DFT and | |2	These steps calculate the energy spectrum of the windowed segment.
Filterbank	The filterbank smooths the spectrum by taking a weighted average of several adjacent frequency bins. The widths of the filters vary according to the mel scale with narrow filters at low frequencies and wider ones at high frequencies.
log	We take the log of the filterbank outputs because the coefficients corresponding to a particular speech sound then follow an approximately gaussian probability density function.
DCT	We take the discrete cosine transform of the log energies in order to reduce the correlations between coefficients. This allows us to model the coefficient distributions as independent gaussians or gaussian mixtures.
Advantages	features can be extracted at a suitable rate to follow variations in speech; spectrally based, therefore independent of the phase errors of the channel or speech input system; mel filter bank gives substantially independent coefficients with high discriminative capability.
Disadvantages	Time-frequency resolution tradeoff Quasi-stationary assumption for speech within a frame not always valid
Complexity	OK for real time.

(b) The DFT length is 24 ms giving a DFT frequency increment of 41.7 Hz. mel(4000)=1898 giving a mel scale increment of 75.92 between successive filter peaks.

(i) f1=75.9 mel=55.4 Hz and f3=227.8 mel=179.7 Hz. These correspond to DFT bin numbers 1.3 and 4.3. Hence three bins: 2,3 and 4 contribute to the filter output.

(ii) f19=1442.4 mel =2275.9 Hz and f21=1594.3 mel=2765.6 Hz. These correspond to DFT bin numbers 54.6 and 66.4. Hence 12 bins: 55 to 66 contribute to the filter output.

(c) The speech spectrum is multiplied by the channel frequency response and this product is then convolved with the window spectrum. Providing the channel response is smooth the elements of the q vector will be multiplied by a factor that is approximately equal to $|H|^2$ evaluated at the peak frequencies of the corresponding filters. It follows that the log of these factors will be added to the elements of m and the DCT of the log added to **c**.

(d) The frequency response of telephone microphones is very variable, but within a given telephone call, the response is constant. The time-average of **c** will be the cepstrum of the microphone/channel combination added to the average cepstrum of the speech. Subtracting

this average from the **c** parameter vectors will remove the effects of the microphone and channel.

The speech recognition system must be trained on speech that has undergone the same processing, i.e. the cepstral mean must also be subtracted during training.

At some frequencies, the channel response may be very low. Any additive noise at these frequencies that is not affected by the channel will worsen the SNR and hence increase the variability of the c coefficients. It may be necessary to compensate for this by increasing the variances within the speech model.

4.

(a) The language model expresses the probability of a word sequence $w = w_1, w_2, \cdots$ as

$$pr(w) = \prod_i pr(w_i \mid w_1, w_2, \cdots, w_{i-1}) .$$

In a unigram model, $pr(w_i \mid w_1, w_2, \cdots, w_{i-1}) = f(w_i)$ implying that the probability of any word is independent of the words that preceded it in the sentence.

In a bigram model, $pr(w_i \mid w_1, w_2, \cdots, w_{i-1}) = f(w_i, w_{i-1})$ and in a trigram model $pr(w_i \mid w_1, w_2, \cdots, w_{i-1}) = f(w_i, w_{i-1}, w_{i-2})$. Thus in these models the probability of a word depends on either the previous or the two previous words.

For a 20000 word vocabulary, there are 2×10^4 unigram probabilities, 4×10^8 bigram probabilities and 8×10^{12} trigram probabilities.

From Bayes' theorem, $pr(w \mid s) = pd(s \mid w)ds \times pr(w) \div pd(s)ds$.

The *language model* gives $pr(w)$, the prior probability of the word sequence occurring. The *acoustic model of speech production* gives $pd(s \mid w)$, the probability that the word sequence *w* would generate the observed input speech signal *s*.

(b)

(i) Totalling each column to obtain *N(j)* gives 81, 65, 102 and 50 respectively for a total of 298 tokens. The unigram probabilities are therefore:

1: 81/298=0.272, 2: 65/298=0.218, 3: 102/298=0.342, "end": 50/298=0.168

(ii) We first calculate *d(i)* according to whether or not all valid successors occur 3 or more times. Note that *d*("start")=0 since "end" is not a valid successor of "start".

b(i) represents the total probability allocated to the infrequent successors divided by the sum of the corresponding unigram probabilities:

$$b(1) = \frac{1+0+0.5+0.5}{81} \times \frac{1}{0.272+0.218} = 0.0504 \qquad b(2) = \frac{0.5+0.5+2+0.5}{65} \times \frac{1}{0.342} = 0.157$$

We can now complete the table:

p(i,j)		Second Word, *j*						
		1	2	3	end	N(i)	d(i)	b(i)
Initial Word, *i*	start	0.2	0.2	0.6	0	50	0	
	1	0.014	0.011	0.735	0.241	81	0.5	0.050
	2	0.038	0.762	0.054	0.146	65	0.5	0.15
	3	0.657	0.049	0.098	0.196	102	0	

(iii) Bigram probabilities cannot be estimated accurately for sequences that do not occur in the training data or that occur only very rarely. For these cases, the probabilities are therefore assumed to be proportional to the corresponding unigram probabilities with a scaling constant, $b(i)$ chosen to ensure that the probabilities of all successors sum to unity.

These rare bigrams are likely to be under-represented in the training data (certainly true for valid bigrams that do not occur at all in the training data).We therefore "steal" some of the probability from the more common bigrams using the discounting factor $d(i)$. The effect of this is to reduce the probability of common bigrams and increase that of rare ones.

5.

(a) bookwork

(b) $r_k = \dfrac{A_{k+1} - A_k}{A_{k+1} + A_k}$

(c)

$$\begin{pmatrix} U_g \\ V_g \end{pmatrix} = \frac{z^{½p}}{\prod_{k=0}^{p}(1+r_k)} \prod_{k=0}^{p-1} \begin{pmatrix} 1 & -r_k z^{-1} \\ -r_k & z^{-1} \end{pmatrix} \times \begin{pmatrix} 1 \\ -r_p \end{pmatrix} U_l$$

Ignoring Vg we can write

$$U_g = \frac{z^{½p}}{\prod_{k=0}^{p}(1+r_k)} \prod_{k=0}^{p-1} \begin{pmatrix} 1 & -r_k z^{-1} \end{pmatrix} \times \begin{pmatrix} 1 \\ -r_p \end{pmatrix} U_l$$

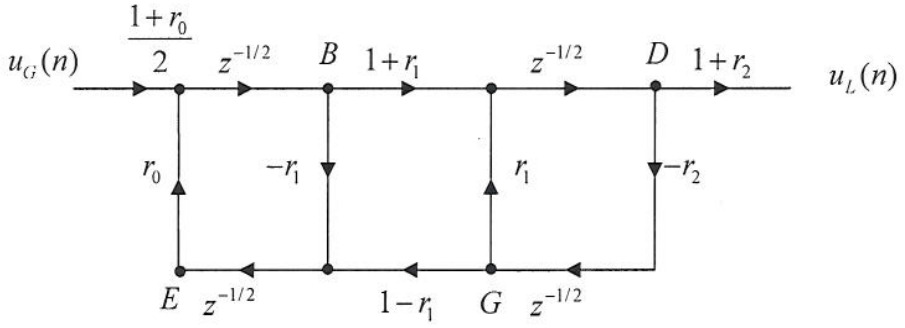
giving

$$\frac{U_l}{U_g} = \frac{\prod_{k=0}^{p}(1+r_k) z^{-½p}}{\prod_{k=0}^{p-1} \begin{pmatrix} 1 & -r_k z^{-1} \end{pmatrix} \times \begin{pmatrix} 1 \\ -r_p \end{pmatrix}}$$

which can be written in the form

$$V(z) = \frac{U_l}{U_g} = \frac{G z^{-½p}}{1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_p z^{-p}}.$$

(d)
(i) delays are half one sample period.
(ii)



$$B = \frac{1+r_0}{2} u_g z^{-1/2} + r_0 E z^{-1/2}$$

$$D = (1+r_1) B z^{-1/2} + r_1 G z^{-1/2}$$

$$G = -r_2 D z^{-1/2}$$

$$E = (1-r_1) G z^{-1/2} - r_1 B z^{-1/2}$$

$$D = (1+r_1) B z^{-1/2} - r_1 r_2 D z^{-1/2} \;\therefore\; B = \frac{D(1+r_1 r_2 z^{-1})}{(1+r_1) z^{-1/2}}$$

$$E = -(1-r_1) r_2 D z^{-1} - r_1 B z^{-1/2}$$

$$B = \frac{1+r_0}{2} u_g z^{-1/2} - r_0 r_2 (1-r_1) D z^{-3/2} - r_0 r_1 B z^{-1}$$

$$B = \frac{\frac{1+r_0}{2} u_g z^{-1/2} - r_0 r_2 (1-r_1) D z^{-3/2}}{1 + r_0 r_1 z^{-1}}$$

$$D = \frac{-B(1+r_0 r_1 z^{-1}) + \frac{1+r_0}{2} u_g z^{-1/2}}{r_0 r_2 (1-r_1) z^{-3/2}}$$

$$= \frac{\frac{1}{2}(1+r_0)(1+r_1) u_g z^{-1}}{1 + (r_1 r_2 + r_0 r_1) z^{-1} + r_0 r_2 z^{-2}}$$

and

$$u_L = (1+r_2) D$$

$$\therefore \frac{u_L}{u_g} = \frac{\frac{1}{2}(1+r_0)(1+r_1)(1+r_2) z^{-1}}{1 + (r_1 r_2 + r_0 r_1) z^{-1} + r_0 r_2 z^{-2}}$$

6.

Winter Olympics wɪnt3ː əʊlɪmpɪks

Note here that the aim is to assess overview understanding of phonetic transcription and not detailed knowledge of IPA phonetic alphabet. Hence full marks will be given when good understanding is shown even if errors are present in the transcription.

The neutral C3 vowel ə is usually known as "schwar": it usually occurs in unstressed syllables. The 'o' of Olympics əʊ is likely to be spoken as a schwar.

A wideband spectrogram because the time resolution is high. High enough to see the pitch period during voicing.

A count of 11 pitch periods in a time of 0.1 seconds indicates a fundamental frequency of 110 Hz. Thus we deduce it is male speech.