E4.14
SO16
ISE4.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2007

MSc and EEE/ISE PART IV: MEng and ACGI

Corrected Copy

SPEECH PROCESSING

Thursday, 10 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

SPEECH PROCESSING

1. a) Figure 1.1 shows a block diagram of a Differential Pulse Code Modulation (DPCM) speech coder.

   i) Identify the input and output signals.

   ii) Describe the operation of each of the three blocks in the diagram and state how each of the three blocks enables a reduction in bit rate for a given signal quality.

   [6]

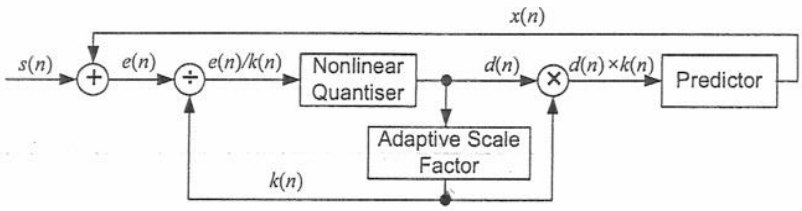

Figure 1.1

b) Show that the mean square quantization error for a uniform quantizer having quantization intervals w is $\frac{w^2}{12}$. State clearly any assumptions you make.

A uniform quantizer has output levels at values of $\pm 64(2k-1)$ for $k = 1, \ldots, 128$. Calculate the signal-to-noise ratio in dB for an input signal uniformly distributed in the range ± 8192. [6]

c) The quantization levels of a particular non-uniform quantizer are given by

$$\pm((2m+33) \times 2^e - 33) \quad \text{for } m = 0, 1, \ldots, 15 \text{ and } e = 0, 1, \ldots, 7.$$

For a signal x uniformly distributed in the range ± 400, determine the values taken by e and the corresponding range of x for each value of e. Give the probability of each value of e. Hence determine the mean square quantization error and the resultant signal-to-noise ratio in dB. [8]

2. Figure 2.1 shows a lossless tube model of the vocal tract with the glottis at the left and the lips at the right. U_G, U_1 and U_L are the z-transforms of the forward acoustic waves at the glottis, the glottis end of the first tube segment and the lips respectively. V_G, V_1 and V_L are the z-transforms of the corresponding reverse waves. A_G and A_1 are respectively the effective cross-sectional areas of the glottis and of the first tube segment.

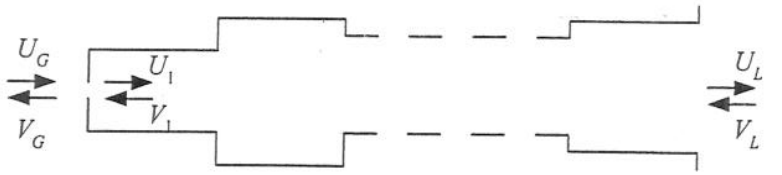

Figure 2.1

a) State the relevant relations for volume continuity and pressure continuity. [2]

b) Hence, show that

$$U_G = \frac{1}{1+r_0}\begin{pmatrix}1 & -r_0\end{pmatrix}\begin{pmatrix}U_1\\V_1\end{pmatrix}$$

and derive an expression for the reflection coefficient, r_0, in terms of A_G and A_1. [5]

c) Explain what value you would expect for r_0 during the closed-glottis interval and how you would expect it to vary during voiced speech as the glottis opens and closes. Explain why it is sometimes desirable to restrict LPC analysis to the closed-glottis intervals of the speech signal. [5]

d) Figure 2.2 shows the waveform of a vowel whose first three formants are at 420, 1333 and 1850 Hz respectively. The waveform is sampled at 8 kHz and the dashed lines indicate the closed-glottis interval (sample numbers 0 to 29) of the central larynx cycle. Figure 2.3 shows how the bandwidths of the three formants, labelled B1, B2 and B3 respectively, vary with the glottal reflection coefficient, r_0.

Table 1 gives the sample numbers and values of successive waveform maxima and minima taken from the open-glottis interval of this cycle. Estimate the value of r_0 during the open-glottis interval. You may assume without proof that the amplitude of a formant with bandwidth b Hz decays with a time constant of $(\pi b)^{-1}$ seconds. [8]

Sample	30	33	36	39	42	45	48	51	54	57	60	63
Value	-187	167	-167	94	-103	66	-23	52	-37	8	-21	14

Table 1

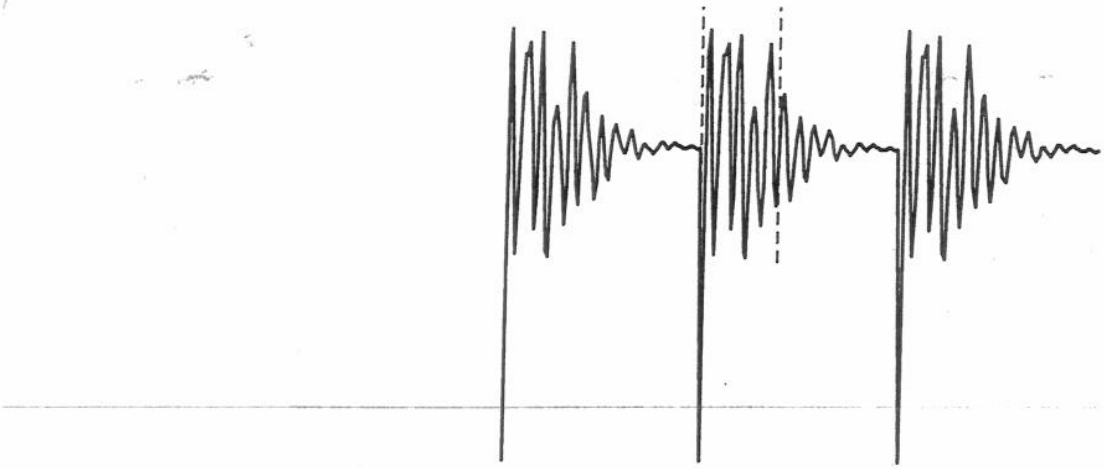

Figure 2.2

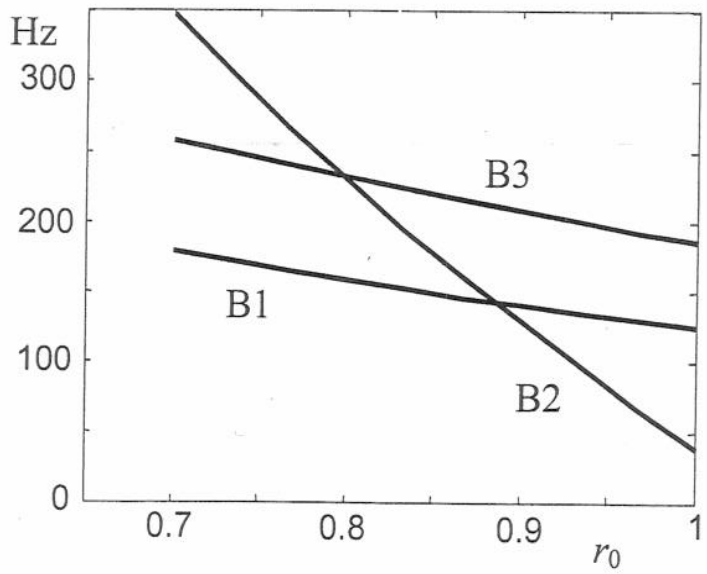

Figure 2.3

3. a) In a classification task, a feature vector $\mathbf{x}$ describing a speech signal is computed every 10 ms frame. The resulting feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ are 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 $d_i(\mathbf{x}_t)$.

Given that frame t corresponds to state s, we define:

i) The total probability that the model generates the frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ is $P(s,t)$;

ii) The total probability that the model generates the frames $\mathbf{x}_{t+1}, \mathbf{x}_2, \ldots, \mathbf{x}_T$ is $Q(s,t)$.

Derive expressions for $P(s,t)$ and $Q(s,t)$ in terms of $P(i,t-1)$ and $Q(i,t+1)$ respectively, where i ranges from 1 to S. Indicate the values that should be used for $P(s,1)$ and $Q(s,T)$. [6]

b) A 3-state Hidden Markov model is trained using 5 frames from a speech utterance. Table 1 shows the output probability of each frame from each state of the model and Figure 3.1 shows the state diagram of the model including the transition probabilities. For each of $s = 1, 2, 3$, calculate the total probability that frame $\mathbf{x}_2$ corresponds to state s and that the model generates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5$. You should perform your calculations to 6 decimal places.

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

Table 1

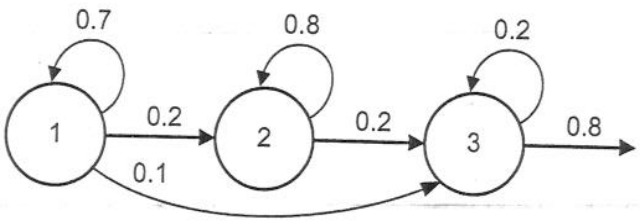

Figure 3.1

[8]

c) i) Explain how and why a language model can be usefully employed in connected speech recognition. State why it is advantageous to use shared states. [3]

ii) Part of a language model is shown in Figure 3.2. Redraw this using an equivalent branching tree structure and label all probabilities on the tree. Assume that the language contains only the words shown on the figure.

[3]

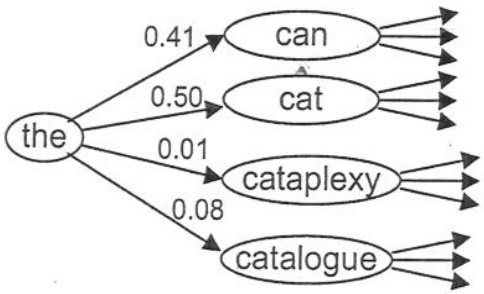

Figure 3.2

4. a) Consider a speech utterance containing both voiced and unvoiced speech recorded in background noise such that the SNR is 40 dB. Design an algorithm to locate and identify the regions of voiced speech and the regions of unvoiced speech. State clearly the steps of the algorithm and describe the processing performed at each step. You may present your design employing flow graphs and/or MATLAB code and/or text descriptions.

State the reasoning on which your design is based. Credit will be given for sound reasoning and clear descriptions. [8]

b) Draw a block diagram of a CELP encoder. With reference to your block diagram, describe briefly how voiced speech and unvoiced speech is modelled in a CELP encoder. [4]

c) Part of a CELP encoder is shown in Figure 4.1 in which the codebook has K entries of length N samples. The kth codevector, $x_k(n)$, is filtered by $H(z)$, multiplied by a gain, g_k, and subtracted from a target signal $t(n)$ to generate the error signal $e_k(n)$. Let $\breve{g}_k$ be defined as the value of g_k that minimizes $E_k = \sum_{n=0}^{N-1} e_k^2(n)$

i) Find an expression for $\breve{g}_k$ in terms of $y_k(n)$ and $t(n)$

ii) Find an expression for E_k when $g_k = \breve{g}_k$.

[8]

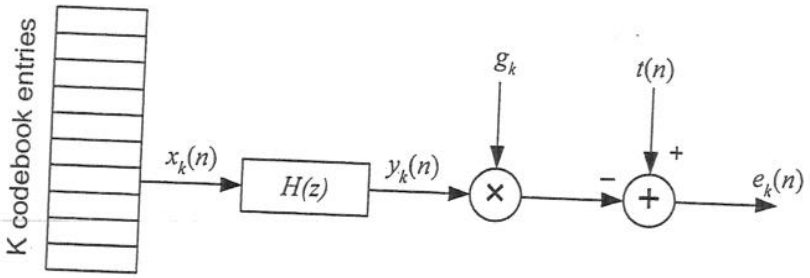

Figure 4.1

5. a) Explain the important differences between covariance LPC and autocorrelation LPC for performing the analysis of a speech signal of several seconds duration. [4]

b) For a speech signal $s(n)$, state how the covariance matrix $\boldsymbol{\Phi}$ and each of its elements ϕ_{ij} is formed. Discuss any symmetry properties of $\boldsymbol{\Phi}$ for the cases of autocorrelation LPC and covariance LPC. [4]

c) Derive the Yule-Walker equations for covariance LPC as the solution to a minimization problem. [6]

d) Consider a signal $s(n) = sin(2\pi fn)$ for $f = 0.1$.

i) Find the values of the elements of the covariance matrix of dimension 2×2. [2]

ii) Hence use 2nd order covariance LPC to obtain the predictor coefficients a_1 and a_2 and write down the corresponding time domain expression for the prediction of $s(n)$. [2]

iii) Draw a labelled sketch of the roots of the predictor polynomial and comment on your sketch. [2]

6. a) Describe the process of diphone-based concatenative text-to-speech synthesis. Define what is meant by a diphone in this context. Explain why such a synthesizer needs to control the pitch and duration of the synthesized speech without affecting the formant frequencies. [4]

b) Consider a segment of speech $s(n)$ of duration 1 second sampled at 8 kHz containing a time-invariant vowel with pitch 100 Hz. Pitch marks are located at sample numbers $n = 40,\ 120,\ 200,\ \ldots,\ 7960$.

i) Explain, with detailed reference to the samples of $s(n)$, how to increase the duration of $s(n)$ to 1.2 seconds without affecting the pitch or formant frequencies. [3]

ii) Explain the effect of removing the central 10 samples of each pitch cycle, i.e. outputting samples, $5 : 74,\ 85 : 154,\ 165 : 234,\ \ldots$, on

- the pitch,
- the formant frequencies,
- the overall duration.

[3]

iii) Explain the effect on the pitch, formant frequencies and duration of outputting the samples at a sample frequency of 10 kHz instead of 8 kHz. [3]

iv) It is desired to change the pitch to 176 Hz, increase all formant frequencies by 10% and change the signal duration to 0.852 seconds. Describe the sequence of operations required and state the output sample frequency required. Given that the first output pitch mark is at sample number 40, determine which two input samples contribute to the output sample number 150. [7]

Solutions - 2007

Master

SPEECH PROCESSING

1. a) i) The input is $s(n)$ and the output is $d(n)$.

ii) The non-linear quantizer has quantization levels that are optimal for a fixed-variance gaussian pdf. The quantization levels are closer together at low signal levels: this reduces the mean square quantization error since these levels occur with higher probability.

The adaptive scale factor normalizes the input to the quantizer so that its variance remains approximately constant at the value required by the quantizer design. The scale factor is reduced each time the quantizer output gives zero and increased each time the quantizer output gives a large value. The scale factor thus converges to a value proportional to the rms value of $e(n)$ and the scaled quantization error will be proportional to the variance of $e(n)$.

The predictor uses previously transmitted information to predict the value of $s(n)$. If the prediction is a good one, the signal e(n) will have a smaller variance than s(n) and the quantization error will be correspondingly reduced. The ratio of the variances is the prediction gain.

b) Assuming that the quantization error is uniformly distributed in the range $\pm\frac{1}{2}w$ with a uniform pdf $p(e) = \frac{1}{w}$, we can calculate the mean square quantization error as:

$$\int_{-1/2w}^{+1/2w} e^2 w^{-1} de = \left[\frac{e^3}{3w}\right]_{-1/2w}^{+1/2w} = \frac{w^2}{12}$$

Quantization interval = 128 and hence mean square error = 1365.3.
The mean square signal level is: $\frac{1}{12}(8192 * 2)^2 = 2.24 \times 10^7$ hence SNR = 42.15 dB

c) The ranges of e are given by

e	$\|x\|$	Prob	MSE	Prob × MSE
0	0 to 31	31/400	4/12	0.0258
1	31 to 95	64/400	16/12	0.2133
2	95 to 223	128/400	64/12	1.7067
3	223 to 400	177/400	256/12	9.44
Total		400/400		11.385

Mean square of uniformly distributed signal is $\frac{1}{2}400^2$. Thus SNR is 38.5 dB.

2. a)

$$U_G - V_G = U_1 - V_1$$

$$\frac{U_G + V_G}{A_G} = \frac{U_1 + V_1}{A_1}$$

b) Substitute $V_G = U_G - U_1 + V_1$ into $A_1(U_G + V_G) = A_G)U_1 + V_1)$ giving

$$\begin{aligned} A_1(2U_G - U_1 + V_1) &= A_G(U_1 + V_1) \\ 2A_1U_G &= U_1(A_G + A_1) + V_1(A_G - A_1) \\ U_G &= \frac{1}{2A_1}\begin{pmatrix} A_G + A_1 & A_G - A_1 \end{pmatrix}\begin{pmatrix} U_1 \\ V_1 \end{pmatrix} \end{aligned}$$

Hence

$$\frac{1}{1+r_0} = A_G + A_1 2A_1 \Rightarrow r_0 = \frac{2A_1}{A_G + A_1} - 1 = \frac{A_1 - A_G}{A_1 + A_G}$$

From which

$$U_G = \frac{1}{1+r_0}\begin{pmatrix} 1 & -r_0 \end{pmatrix}\begin{pmatrix} U_1 \\ V_1 \end{pmatrix}.$$

c) When $A_G = 0$ we have $r_0 = 1$ so this is the value we would expect during the closed phase. During the open-glottis interval, r_0 will decrease but will remain positive since A_G will never exceed A_1. Restricting LPC to the closed phase allows a more accurate estimate of the vocal tract since the acoustic input is equal to zero.

d) From the values given in the table, the period of the dominant oscillation is 6 samples; this corresponds to 1333 Hz and so arises from the second formant. During the closed phase, we expect the envelope to decay exponentially. Therefore if we plot the absolute value of the peaks against time, we should get a straight line

$$y = k\exp(-\pi bt) = k\exp(-\pi bn/f_s) \Rightarrow \ln(y) = \ln(k) + n \times -\pi b/f_s.$$

By estimating the gradient we get $-\pi b/f_s = -0.0888$ and therefore $b = 226$ Hz. From the graph given in the question, we get $r_0 \approx 0.81$.

3. a) To calculate $P(s,t)$, we observe that any alignment of frames 1, ..., t must allocate frame $t-1$ to one of the states i in the range 1, ..., S. Thus

$$P(s,t) = \sum_{i=1}^{S} P(i,t-1) \times p(\text{frame } t \text{ is in state } s | \text{frame } t-1 \text{ is in state } i$$
$$= \sum_{i=1}^{S} P(i,t-1) \times a_{i,s} \times d_s(\mathbf{x}_t)$$

The development of

$$Q(s,t) = \sum_{i=1}^{S} a_{s,i} \times d_i(\mathbf{x}_{t+1}) \times Q(i,t+1)$$

follows a similarly.

Since we always assume that frames 1 and T must be in states 1 and S respectively, the initial conditions for the recursions are:

$$P(s,1) = \begin{cases} d_1(\mathbf{x}_1) & \text{for} \quad s=1 \\ 0 & \text{otherwise} \end{cases}$$
$$Q(s,T) = \begin{cases} a_{S=} & \text{for} \quad s=S \\ 0 & \text{otherwise} \end{cases}$$

b) It is useful to draw the alignment lattice as follows:

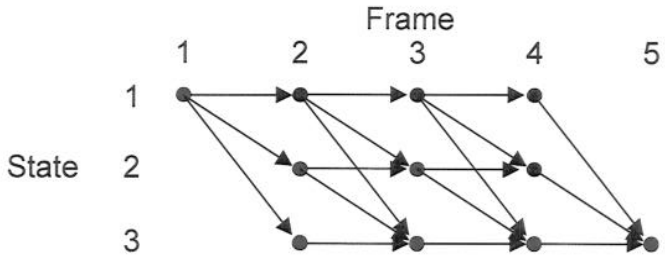

Figure 3.1

P(1,1) $= 0.5$
P(1,2) $= 0.5 \times 0.7 \times 0.5 = 0.175$
P(2,2) $= 0.5 \times 0.2 \times 0.1 = 0.01$
P(3,2) $= 0.5 \times 0.1 \times 0.4 = 0.02$

Q(3,5) $= 0.8$
Q(1,4) $= 0.1 \times 0.5 \times 0.8 = 0.04$
Q(2,4) $= 0.2 \times 0.5 \times 0.8 = 0.08$
Q(3,4) $= 0.2 \times 0.5 \times 0.8 = 0.08$
Q(1,3) $= 0.7 \times 0.1 \times 0.04 + 0.2 \times 0.2 \times 0.08 + 0.1 \times 0.4 \times 0.08 = 0.0092$
Q(2,3) $= 0.8 \times 0.2 \times 0.08 + 0.2 \times 0.4 \times 0.08 = 0.0192$
Q(3,3) $= 0.2 \times 0.4 \times 0.08 = 0.0064$
Q(1,2) $= 0.7 \times 0.3 \times 0.0092 + 0.2 \times 0.8 \times 0.00192 + 0.1 \times 0.5 \times 0.0064 = 0.005324$
Q(2,2) $= 0.8 \times 0.8 \times 0.0192 + 0.2 \times 0.5 \times 0.0064 = 0.012928$
Q(3,2) $= 0.2 \times 0.4 \times 0.08 = 0.0064$

$P(1,2)Q(1,2) = 0.175 \times 0.005324 = 0.000932$
$P(2,2)Q(2,2) = 0.01 \times 0.012928 = 0.000129$
$P(3,2)Q(3,2) = 0.02 \times 0.0064 = 0.000128$

c) i) If several words begin with the same sequence of phonemes it is wasteful to create distinct states for each one. Instead we create a branching tree for all words starting with the same phonemes.

ii) The tree is shown in Figure 3.2

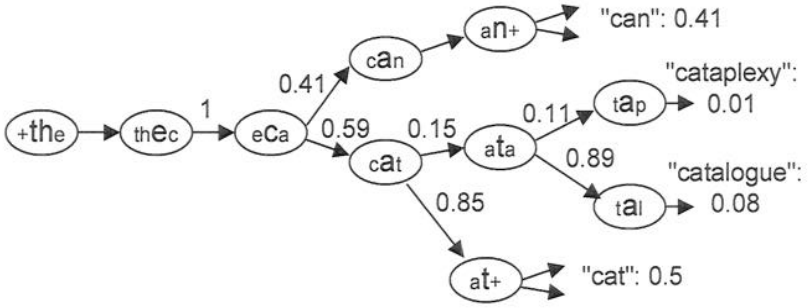

Figure 3.2

4. a) Reasoning should be based on an explanation of the differences and similarities between voiced, unvoiced speech and silence. Features such as format frequencies and bandwidths (or related spectral properties) are an obvious basis for such an algorithm. Other well reasoned proposals will also get full credit.

b) The periodicity (due primarily to voiced excitation) is modelled using a long-term predictor. The residue after removing the long-term predictor output is modelled using a stochastic codebook.

c) i) We have $e_k(n) = t(n) - g_k y_k(n)$

$$1/2\frac{\partial E_k}{\partial g_k} = \sum_n e_k(n)\frac{\partial e_k}{\partial g_k} = -\sum_n e_k(n)y_k(n) = g_k\sum_n y_k^2(n) - \sum_n y_k(n)t(n)$$

Setting the partial derivatives to zero gives

$$\breve{g}_k = \frac{\sum_n y_k(n)t(n)}{\sum_n y_k^2(n)}$$

ii) The optimal E_k is then given by

$$E_{k(opt)} = \sum_n (t(n) - g_k y_k(n))^2 = \sum_n t^2(n) - 2g_k\sum_n t(n)y_k(n) + g_k^2\sum_n y_k^2(n)$$

$$= \sum_n t^2(n) - 2\frac{\left(\sum_n t(n)y_k(n)\right)^2}{\sum_n y_k^2(n)} + \frac{\left(\sum_n t(n)y_k(n)\right)^2}{\sum_n y_k^2(n)} = \sum_n t^2(n) - \frac{\left(\sum_n t(n)y_k(n)\right)^2}{\sum_n y_k^2(n)}$$

5. a) Autocorrelation LPC divides the speech into frames of about 20 ms duration using a tapered window. The frame length in covariance LPC can be much shorter (2 ms) and tapered windows are not used. Autocorrelation LPC uses only data from within the limits of the analysis frame whereas covariance LPC requires $p-1$ preceding samples. The autocorrelation method always gives stable poles whereas covariance may give unstable poles. These can be reflected inside the unit circle without changing the magnitude of the frequency response of the filter.

b) $\phi_{ij} = \sum_{n\in\{F\}} s(n-i)s(n-j)$

In autocorrelation LPC the limits on the summation are $\pm\infty$ giving Toeplitz symmetric. In covariance LPC the limits are 0 to $N-1$ giving symmetry but not Toeplitz symmetry.

c)

$$e(n) = s(n) - \sum_{j=1}^{p} a_j s(n-j) = s(n) - a_1 s(n-1) - a_2 s(n-2) - \ldots - a_p s(n-p)$$

$$Q_E = \sum_{n\in\{F\}} e^2(n)$$

$$\frac{\partial Q_E}{\partial a_i} = \sum_{n\in\{F\}} \frac{\partial\left(e^2(n)\right)}{\partial a_i} = \sum_{n\in\{F\}} 2e(n)\frac{\partial e(n)}{\partial a_i} = -\sum_{n\in\{F\}} 2e(n)s(n-i)$$

$$\sum_{n\in\{F\}} e(n)s(n-i) = 0 \quad \text{for} \quad i = 1,\ldots,p$$

$$\Rightarrow \sum_{n\in\{F\}} \left(s(n)s(n-i) - \sum_{j=1}^{p} a_j s(n-j)s(n-i) \right) = 0 \quad \text{for} \quad i = 1,\ldots,p$$

$$\Rightarrow \sum_{j=1}^{p} a_j \sum_{n\in\{F\}} s(n-j)s(n-i) = \sum_{n\in\{F\}} s(n)s(n-i)$$

$$\Rightarrow \sum_{j=1}^{p} \phi_{ij} a_j = \phi_{i0} \quad \textit{where} \quad \phi_{ij} = \sum_{n\in\{F\}} s(n-i)s(n-j)$$

In matrix form:

$$\mathbf{\Phi a} = \mathbf{c} \quad \Rightarrow \quad \mathbf{a} = \mathbf{\Phi}^{-1}\mathbf{c} \quad \text{providing } \mathbf{\Phi}^{-1} \text{ exists}$$

d) Consider a signal $s(n) = sin(2\pi fn)$ for $f = 0.1$.

i)

$$\begin{aligned}
\phi_{ij} &= E\left(s(n-i)\,s(n-j)\right) \\
&= E\left(\sin\left(2\pi f(n-i)\right)\sin\left(2\pi f(n-j)\right)\right) \\
&= {}^1\!/_2 E\left(\cos\left(2\pi f(-i+j)\right) - \cos\left(2\pi f(2n-j-i)\right)\right) \\
&= {}^1\!/_2 \cos\left(2\pi f(j-i)\right) \\
c_i &= \phi_{i0} = {}^1\!/_2 \cos\left(2\pi f i\right) \\
\phi_{ij} &= \begin{bmatrix} 0.5 & 0.405 \\ 0.405 & 0.5 \end{bmatrix} \\
\phi_{i0} &= \begin{bmatrix} 0.405 \\ 0.155 \end{bmatrix}
\end{aligned}$$

ii)

$$\begin{aligned}
\mathbf{a} &= \boldsymbol{\Phi}^{-1}\mathbf{c} \\
&= \frac{1}{0.087}\begin{bmatrix} 0.5 & -0.405 \\ -0.405 & 0.5 \end{bmatrix}\begin{bmatrix} 0.405 \\ 0.155 \end{bmatrix} = \begin{bmatrix} 1.618 \\ -1 \end{bmatrix}
\end{aligned}$$

So that the predictor can be written

$$s(n) = 1 + 1.618s(n-1) - s(n-2)$$

and

$$S(z) = \frac{1}{1 - 1.618z^{-1} + z^{-2}}$$

iii) The poles occur on the unit circle so describe a marginally stable oscillation at a frequency of $0.1 \times 2\pi$.

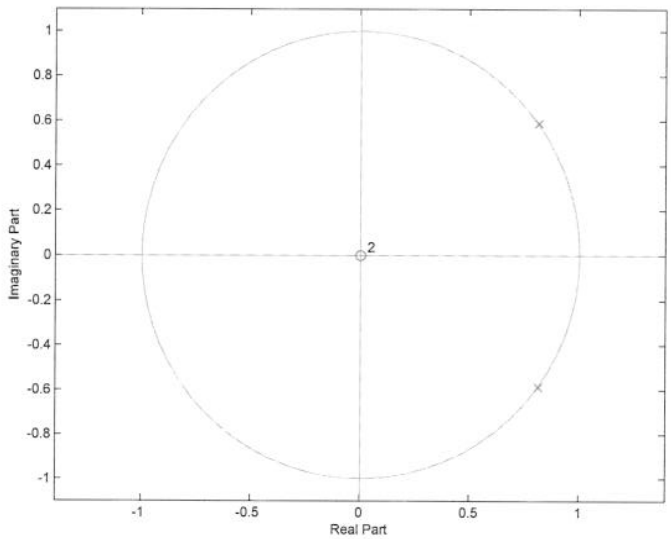

Figure 5.1

6. a) We create the output speech waveform by joining together segments of recorded speech. The base segments are diphones: i.e. they extend from the centre of one phoneme to the centre of the next.

In synthesized speech, the duration, amplitude and pitch of each phoneme must be controlled by the synthesizer. The pitch follows a smooth contour, generally rising to the first stressed syllable of a sentence and falling thereafter. The amplitude broadly follows the same contour as the pitch, rising slightly for each stressed syllable. The durations are adjusted so that the intervals between stressed syllables are roughly uniform. If the required values of duration, amplitude and pitch differ from those present in the recoded diphone, they must be adjusted accordingly.

b) The samples are output in the order $0:399,\ 320:719,\ 640:1039,\ \ldots$ The effects may be summarized as:

	Pitch	Formants	Duration
(i)	$\times 1$	$\times 1$	$\times 1.2$
(ii)	$\times 1.14$	$\times 1$	$\times 0.875$
(iii)	$\times 1.25$	$\times 1.25$	$\times 0.8$

c) We can determine the required transformations by choosing the sample rate to adjust the formant frequencies, reducing the cycle length to adjust the pitch and finally replicating cycles to adjust the duration:

	Pitch	Formants	Duration
Fs $\times 1.1$	$\times 1.1$	$\times 1.1$	$\times 0.909$
Remove 30/80 samples	$\times 1.6$	$\times 1$	$\times 0.625$
Repeat alternate cycles	$\times 1$	$\times 1$	$\times 1.5$
Total	$\times 1.76$	$\times 1.1$	$\times 0.852$

Thus the new output frequency is 8800 Hz.

Since alternate cycles are replicated, the output pitch cycles centred at samples 40, 90, 140 and 190 are formed from the input cycles centred at samples 40, 120, 120 and 200. Output sample number 150 will be formed from the +10th sample of the input cycle centred at 120 and the -40th sample of the input cycle centred at 200. It thus contains contributions from input samples 130 and 160.