Master

E4.14
SO16
ISE4.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2004

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Thursday, 6 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) : D.M. Brookes

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. A spectrogram $P(k,m)$ can be obtained by computing

$$P(k,m)=\left|\sum_{i=0}^{N-1} w(i)x(m-i)\exp\left(-\frac{2\pi j}{N}k(m-i)\right)\right|^2$$

where $x(n)$ is a speech signal and $w(i)$ is a window of length N.

(a) State the significance of the variables k and m. Hence give a brief description of the nature of the information captured in a spectrogram and purpose of a spectrogram in speech processing. [5]

(b) Which factors should be considered when choosing $w(i)$ and N? [2]

(c) Consider a special case when m is set to a particular value m_0. Show that [7]

$$P(k,m_0)=\left|\sum_{r=0}^{N-1} y_{m_0}(r)\exp\left(-\frac{2\pi j}{N}kr\right)\right|^2.$$

State the relationship between $y_{m_0}(r)$ and $x(n)$.

Using this result, describe and comment on the relationship of $P(k,m_0)$ to $w(i)$ and $x(n)$ in the frequency domain.

(d) Consider the modulated cosine wave given by: [6]

$$x(n)=\cos(0.4\pi n)\times(2+\cos(0.02\pi n))$$

Sketch $P(k,310)$ as a function of k for $0\le k<200$ when $N=200$ and the window function is given by:

(i) $w(i)=1$ for $i=0,1,\ldots,199$

(ii) $w(i)=1$ for $i=0,1,\ldots,19$

Comment on the differences between cases (i) and (ii).

2. (a) Draw the block diagram of a Differential Pulse Code Modulation (DPCM) speech coder in which the input and output signals are $s(n)$ and $d(n)$ respectively. Describe briefly how each of the blocks in the diagram operates and how a reduction in bit-rate for a given signal quality is achieved. [6]

(b) Consider a uniform quantizer having quantization intervals of width w. Show that the mean square quantization error is given by $w^2/12$. State clearly any assumptions you make. [2]

(c) Consider a uniform quantizer with output levels at values [6]

$$\pm 64(2k-1) \quad \textit{for } k = 1, \ldots, 128 .$$

Find the mean square value of a uniformly distributed signal occupying the following ranges of amplitude

(i) ± 11200

(ii) ± 400

and hence calculate the signal-to-quantization-noise ratio in dB for both cases.

(d) Consider a non-uniform quantizer with output levels at values of [6]

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

Show that for input values, x, in the range

$$\left(2^{r+5}-33\right) \le |x| < \left(2^{r+6}-33\right) \qquad r = 1, \ldots, 6$$

the exponent e will have the value r and the width of the quantization intervals will be 2^{e+1}.

For an input signal uniformly distributed in the range ±400, determine the possible values taken by e and the probability of each. Hence determine the mean square quantization error and the resultant signal-to-noise ratio in dB.

3. (a) Summarize the desirable properties of a feature set for use in a typical speech recognition application. [4]

(b) Multivariate Gaussian distributions are commonly used in speech recognition systems to model the variability of features within the states of a hidden Markov model. [6]

(i) Discuss the advantages and disadvantages of imposing on such models the restriction that the covariance matrices be diagonal and state the necessary conditions for diagonal covariance matrices.

(ii) State two approaches that are commonly employed in the design of speech recognition systems to reduce the disadvantages in part (i).

(c) Consider a particular feature vector

$$\mathbf{z} = \begin{pmatrix} z_1 & z_2 & \dots & z_4 \end{pmatrix}^T$$

where the z_i are independent identically distributed Gaussian random variables having zero mean and unit variance with a probability density function

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

(i) Derive an expression for the probability density function of the vector $\mathbf{z}$ and show that [5]

$$\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right) = \mathbf{I}$$

where $\mathrm{E}(\cdot)$ denotes the expectation operation and $\mathbf{I}$ is the identity matrix.

(ii) Let $\mathbf{x}$ be defined by $\mathbf{x} = \mathbf{A}\mathbf{z}$ where $\mathbf{A}$ is a non-singular matrix of dimension $N \times N$. Obtain an expression for the covariance matrix, $\mathbf{C}$, of $\mathbf{x}$. [5]

4. Consider a speech recognition system based on a hidden Markov model containing S states $\{s_1, s_2, \ldots, s_S\}$. The input speech samples are initially processed to give a sequence of T feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$.

(a) Describe the main features of the hidden Markov model. Include a description of the components of the model, a description of the operation of the model and definitions of the parameters of the model including, but not limited to, output probability density and transition probability. [4]

(b) $P(t,s)$ 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.

(i) Show that $P(t,s)$ can be expressed in terms of $P(t-1,k)$ for $k = 1, 2, \ldots, S$ and that $Q(t,s)$ can be expressed in terms of $Q(t+1,k)$ for $k = 1, 2, \ldots, S$. State the initialization of P and Q required for their recursive computation. [4]

(ii) Draw a labelled diagram of a left-to-right, no-skips, hidden Markov model with 4 states. The sequence of state transition probabilities is 0.1, 0.5 and 0.8 and the exit probability is 0.5. [4]

(iii) The output probability densities for each of six observed feature vectors are shown in Table 4.1. Determine the total probability of all alignments of the observation with the model for which frame 3 is in state 2 given that $\mathbf{x}_1$ is in state s_1 and $\mathbf{x}_6$ is in state s_4. [8]

Table 4.1

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
s_1	0.5	0.4	0.5	0.2	0.1	0.1
s_2	0.4	0.5	0.8	0.6	0.3	0.8
s_3	0.2	0.8	0.2	0.8	0.2	0.2
s_4	0.5	0.4	0.5	0.2	0.5	0.8

5. (a) Describe Line Spectrum Frequencies (LSF) in the context of speech coding. State their advantages and disadvantages for speech coding and show how LSFs are determined given a speech prediction filter [7]

$$V(z) = \frac{1}{1 - \sum_{j=1}^{p} a_j z^{-j}}.$$

(b) A discrete-time speech signal is denoted $s(n)$. Figure 5.1 shows part of the encoder in a Code-Excited Linear Prediction (CELP) speech transmission system. The re-synthesis error for sample n is defined as

$$\begin{aligned} e_k(n) &= s(n) - g_k y_k(n) \\ &= s(n) - g_k \sum_{i=0}^{n} h(i) x_k(n-i) \quad \text{for } n = 0,\ 1,\ 2,\ \ldots,\ N-1. \end{aligned}$$

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 re-synthesis error for the frame is given by

$$E_k = \sum_{n=0}^{N-1} e_k^2(n)$$

(i) 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 [6]

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

(ii) 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 $y_{k-1}(n-1)$. [3]

(iii) Determine an expression for the number of multiply operations required to compute $y_k(n)$ for all codebook entries under the conditions described in part (ii). Determine also an expression for the number of multiply operations required to compute $y_k(n)$ directly using convolution without the conditions described in part (ii). Comment on the difference. [4]

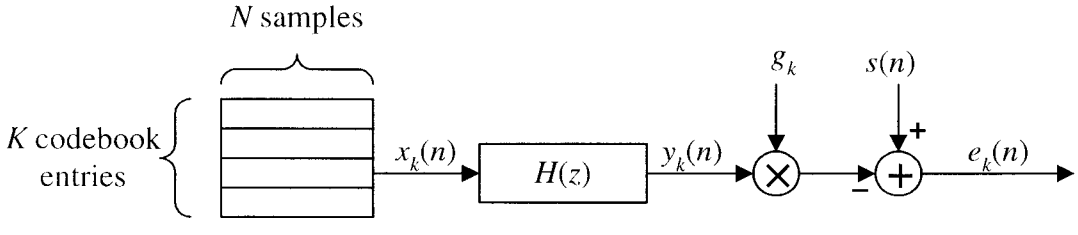


Figure 5.1

6. (a) Describe the source-filter model of speech production and draw a labelled block diagram. Describe how the model generates vowels and consonant sounds. Explain why a single excitation signal is usually inadequate when modelling consonant sounds. State briefly the physical process by which unvoiced excitation is generated in speech. [4]

(b) Describe briefly and draw block diagrams of the structure of (i) a cascade formant synthesiser and (ii) a parallel formant synthesiser and outline the advantages and disadvantages of the two types. [3]

(c) Consider the two transfer functions [7]

$$G(z) = \frac{1}{1 - 1.14z^{-1} + 0.9025z^{-2}} \quad \text{and} \quad H(z) = \frac{1}{1 + 0.81z^{-2}}.$$

Find the poles and zeros of $G(z)$ and $H(z)$. Plot and label them on the z-plane.

Consider also the following three combinations of these transfer functions

(i) $Y_1(z) = G(z)H(z)$ (ii) $Y_2(z) = G(z) + H(z)$ (iii) $Y_3(z) = G(z) - H(z)$.

Find the poles and zeros of $Y_1(z)$, $Y_2(z)$ and $Y_3(z)$ and comment on the relationship between the three transfer functions. Add these poles and zeros to your z-plane plot and label them.

For each of the three expressions sketch a graph showing its magnitude response over the range $\omega = 0$ to π.

Explain and comment on the relationships of the combinations (i), (ii) and (iii) to cascade and parallel formant synthesis.

(d) A particular speech signal contains a formant with bandwidth 32 Hz at a frequency of 912 Hz. Explain how this formant would be modelled in a cascade formant synthesiser and determine any relevant parameters. The sampling frequency is 8 kHz. [3]

(e) The excitation signal used to drive each formant filter in a parallel formant synthesiser consists of a mixture of a periodic and a random signal. Figure 6.1 shows how the ratio of these two signals is varied for each formant using a Voicing Control parameter. Explain how this voicing control would be integrated into a parallel formant synthesiser and illustrate your explanation using the specific example for which the Voicing Control = 0.5. [3]

State an example phoneme for which the Voicing Control would be close to 0.5.

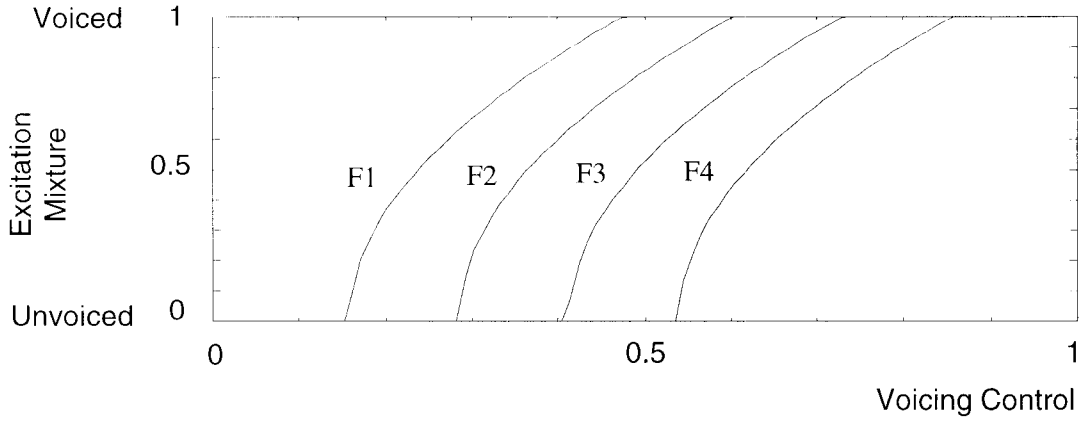

Voiced
1
Excitation
Mixture
0.5
F1
F2
F3
F4
Unvoiced
0
0
0.5
1
Voicing Control

Figure 6.1

solutions -
Speech Processing 2004

1.

(a) The variables k and m represent the frequency bin index and the time frame index respectively. A spectrogram captures a time-frequency representation of the speech signal and facilitates the visualization of the time variations and frequency variations in speech.

(b) There are three main competing considerations:

(i) Good time resolution requires a short window whose DFT has a wide central lobe.

(ii) Good frequency resolution requires a long window whose DFT has a narrow central lobe.

(iii) Good dynamic range requires a window with well-supressed sidelobes. For a given window length, this will result in a wider central lobe and hence worse frequency resolution.

(c) Let r = N–1–i, giving

$$P(k,m_0)=\left|\exp\left(-\frac{2\pi j}{N}k(m_0-N+1)\right)\sum_{r=0}^{N-1}y_{m_0}(r)\exp\left(-\frac{2\pi j}{N}kr\right)\right|^2$$

with

$$y_{m_0}(r)=w(N-1-r)x(m_0-N+1+r)$$

and the modulus of the first exp() term is unity.

The term $y_{m_0}(r)$ is the product of two discrete-time signals and its DFT is therefore the convolution of their corresponding periodic DFTs. The two signals are a time-shifted version of x and a time-reversed window w. Thus a vertical slice through the spectrogram gives the convolution of the "true" spectrum of x with the fourier transform of the window function. This is a smeared version of the "true" spectrum

(d) The carrier period is 5 samples while the modulation period is 100 samples. The long window (i) has sufficient frequency resolution to detect the modulation at 0.4*pi +/- 0.02*pi whereas the short window (ii) does not.

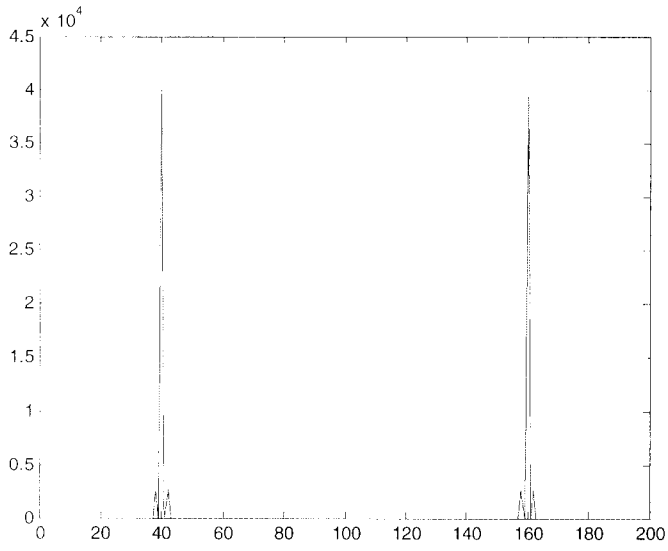


Figure 1: N = 200, the two components are both visible

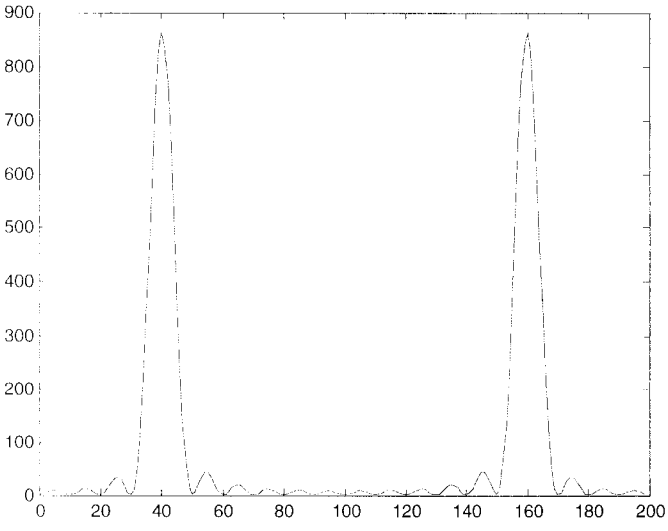
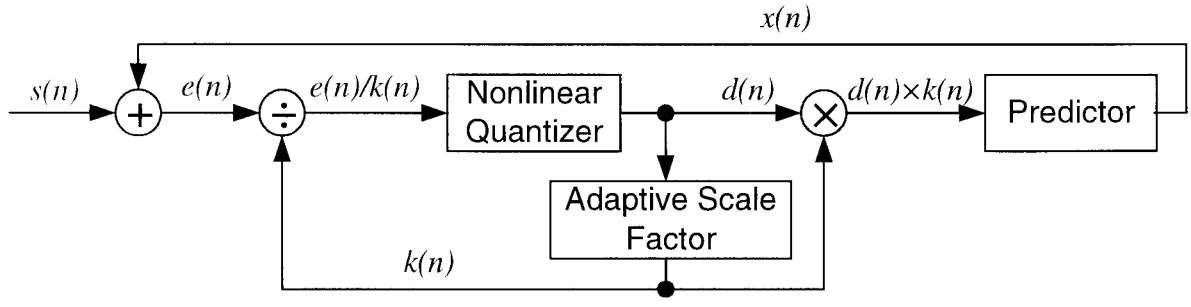


Figure 2: N = 20, only the carrier is visible because of the poor frequency resolution

(a)



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 quantisation error will be correspondingly reduced. The ratio of the variances is the prediction gain.

(b) Assuming that the quantisation error is uniformly distributed in the range ±½w with a pdf $p(e)=w^{-1}$, we can calculate the mean square quantisation error as:

$$\int_{-½w}^{+½w} e^2 w^{-1} de = \left[\frac{e^3}{3w} \right]_{-½w}^{+½w} = \frac{w^2}{12}$$

(c) Quantisation interval = 128 and hence mean square error = 1365.3

The mean square signal level is:

(i) $1/12 \times 22400^2 = 4.1813 \times 10^7$ hence SNR = 44.86 dB

(ii) $1/12 \times 800^2 = 5.33 \times 10^4$ hence SNR = 15.92 dB or 15.7 dB if you take account of the incompletely filled bins where |x| > 384.

(d) The highest value for e=r–1 is $|x| = 63 \times 2^{r-1} - 33 = 31.5 \times 2^r - 33$ when m=15 and the lowest value for e=r is $|x| = 33 \times 2^r - 33$ when m=0. Thus the threshold between e=r–1 and e=r is at $|x| = 32 \times 2^r - 33 = 2^{r+5} - 33$. It follows that e will equal r for

$$2^{r+5} - 33 \le |x| < 2^{r+6} - 33$$

As a special case, e=0 for |x| < 31.

Thus for x in the range ±400, we have

e	|x|	prob	Quant Step	MSE	probxMSE
0	0 to 31	31/400	2	4/12	0.025833
1	31 to 95	64/400	4	16/12	0.2133
2	95 to 223	128/400	8	64/12	1.7067
3	223 to 400	177/400	16	256/12	9.44
Total:		400/400			11.385

Hence SNR = 10*log10(5.33x10^4 / 11.385) = 36.71 dB

since power in the uniform distribution on the interval [a,b] is given by $(b-a)^2/12 = 800^2/12 = 5.33 \times 10^4$

3.

a)

- Compact representation of the speech.
- Elements are independent and similar range.
- High between-class variability, low within-class variability
- Low computational complexity to compute
- Robust to noise

b)

(i) The advantages of diagonal covariances are:

- reduced number of training examples required since there are fewer parameters to train.
- computation of log probabilities is much quicker. If the number of features is F, log probability calculation requires $O(F^2)$ multiplications for a full covariance matrix but only O(F) for a diagonal covariance matrix.

The disadvantage of diagonal covariances is that you are making a, generally false, assumption that the elements of the parameter vector are independent and hence the performance of the recognizer is degraded.

(ii) One approach is to select parameters that are as close as possible to being independent, such as the mel cepstral coefficients which benefit from the partial orthogonalization of the DCT. Another approach is to explicitly decorrelate the features by employing a linear transformation on feature vector $\mathbf{x}$ of the form $\mathbf{y} = \mathbf{F}^T\mathbf{x}$ where the transformation matrix $\mathbf{F}$ is found from the eigenvalues of the average within-state covariance matrix.

c)

(i) Since the components of **z** are independent, their joint p.d.f. is just the product of the individual density functions:

$$p(\mathbf{z}) = \prod_{i=1}^{N} (2\pi)^{-½} \exp\left(-½ z_i^2\right) = (2\pi)^{-½N} \exp\left(-½ \sum_{i=1}^{N} z_i^2\right) = (2\pi)^{-½N} \exp\left(-½ \mathbf{z}^T\mathbf{z}\right).$$

Then $\mathrm{E}\left(z_i z_j\right) = 1$ for the diagonal elements (i=j) since we are told that z has unit variance and $\mathrm{E}\left(z_i z_j\right) = 0$ for the off-diagonal elements $\left(i \neq j\right)$ since z_i and z_j are independent so that $\mathrm{E}\left(z_i z_j\right) = \mathrm{E}\left(z_i\right)\mathrm{E}\left(z_j\right)$ and the vector z is zero mean. Hence we obtain the identity matrix.

(ii) $\mathbf{C} = \mathrm{E}\left(\mathbf{x}\mathbf{x}^T\right) = \mathrm{E}\left(\mathbf{A}\mathbf{z}\mathbf{z}^T\mathbf{A}^T\right) = \mathbf{A}\,\mathrm{E}\left(\mathbf{z}\mathbf{z}^T\right)\mathbf{A}^T = \mathbf{A}\mathbf{A}^T$

4.

(a) A Hidden Markov Model for a word must specify the following parameters for state *s*: The mean and variance for each of the *F* elements of the parameter vector: μ_s and σ_s^2. These allow us to calculate $d_s(\mathbf{x})$: the output probability density of input frame $\mathbf{x}$ in state *s*. The transition probabilities $a_{s,j}$ to every possible successor state. $a_{s,j}$ is often zero for all *j* except *j=s* and *j=s+1* it is then called a *left-to-right, no skips* model. For a Hidden Markov Model with *S* states we therefore have around $(2F+1)S$ parameters. A typical word might have *S*=15 and *F*=39 giving 1200 parameters in all.

(b)

(i)

$$P(t,s) = d_s(\mathbf{x}_t)\sum_{k=1}^{S} a_{ks} P(t-1,k)$$

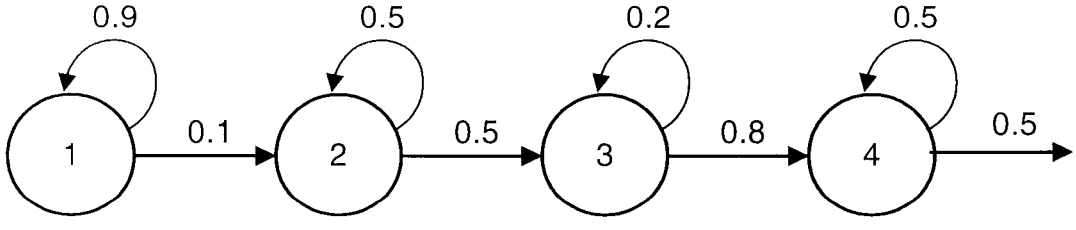
Every alignment going through state s at time t must go though some state, say k, at time t-1. Thus the total probability of all alignments going through state s at time t can be obtained by adding up P(t-1,k) for all k and multiplying by the probability of a transition from state k to state s. We must then multiply by $d_s(x_t)$ to include the output probability at time t.

We initialise this recursion by setting $P(1,1)=d_1(x_1)$

$$Q(t,s) = \sum_{k=1}^{S} a_{sk} d_k(\mathbf{x}_{t+1}) Q(t+1,k)$$

This is the same argument as above but working in reverse time. We need to initialise $Q(T,S)=a_{S=}$ but this does not affect the optimal alignment.

(ii)



(iii) We need to calculate $P(3,2) \times Q(3,2)$

$P(1,1) = 0.5$
$P(2,1) = 0.4\times(0.9\times0.5) = 0.18$
$P(2,2) = 0.5\times(0.1\times0.5) = 0.025$
$P(3,2) = 0.8\times(0.1\times0.18 + 0.5\times0.025) = 0.0244$

$Q(6,4)=0.5$
$Q(5,4)=0.5\times0.8\times0.5=0.2$
$Q(5,3)=0.8\times0.8\times0.5=0.32$
$Q(4,4)=0.5\times0.5\times0.2 = 0.05$
$Q(4,3)=0.2\times0.2\times0.32+0.8\times0.5\times0.2 = 0.0928$
$Q(4,2)=0.5\times0.2\times0.32 = 0.032$
$Q(3,2)=0.5\times0.6\times0.032+0.5\times0.8\times0.0928 = 0.0467$

Hence $P(3,2) \times Q(3,2) = 0.00114$

5

(a) Line Spectrum Frequencies are features of speech that are well suited to speech coding. They can be obtained from the coefficients of the predictor filter as shown below. They have the advantages/disadvantages of :

- Easy to check if filter is stable
- Good for interpolation between frames
- Smooth trajectory in time
- Strong relationship to formant peaks in the speech spectrum
- Additional computation requirements.

Given

$$A(z) = G \times V^{-1}(z) = 1 - \sum_{j=1}^{p} a_j z^{-j} = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_p z^{-p}$$

Form symmetric and asymmetric polynomials

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

$V(z)$ is stable if and only if the roots of P(z) and Q(z) all lie on the unit circle and they are interleaved.

If the roots of $P(z)$ are at $e^{2\pi j f_i}$ for i=1,3,... and those of $Q(z)$ are at $e^{2\pi j f_i}$ for i=0,2,... with $f_{i+1} > f_i \geq 0$ then the LSF frequencies are given by $f_1, f_2, \ldots, f_p$.

(b)

(i)

$$½ \frac{\partial E}{\partial g_k} = \sum_{n=0}^{N-1} e(n) \frac{\partial e(n)}{\partial g_k} = -\sum_{n=0}^{N-1} \left(s(n) - g_k y_k(n) \right) y_k(n)$$

Equating this with zero to find the turning point gives

$$g_k = \frac{\sum_{n=0}^{N-1} s(n) y_k(n)}{\sum_{n=0}^{N-1} y_k^2(n)}$$

Substituting this back in the expression for *E* gives

$$E_k = \sum_{n=0}^{N-1}\left(s(n) - g_k y_k(n)\right)^2 = \sum_{n=0}^{N-1} s^2(n) - g_k\left(2\sum_{n=0}^{N-1} s(n)y_k(n) - g_k\sum_{n=0}^{N-1} y_k^2(n)\right)$$

$$= \sum_{n=0}^{N-1} s^2(n) - g_k\left(2\sum_{n=0}^{N-1} s(n)y_k(n) - \sum_{n=0}^{N-1} s(n)y_k(n)\right)$$

$$= \sum_{n=0}^{N-1} s^2(n) - g_k\sum_{n=0}^{N-1} s(n)y_k(n) = \sum_{n=0}^{N-1} s^2(n) - \frac{\left(\sum_{n=0}^{N-1} s(n)y_k(n)\right)^2}{\sum_{n=0}^{N-1} y_k^2(n)}$$

(ii)

$$y_k(n) = \sum_{j=0}^{n} h(j)x_k(n-j) = h(n)x_k(0) + \sum_{j=0}^{n-1} h(j)x_{k-1}(n-j-1) = h(n)x_k(0) + y_{k-1}(n-1)$$

(iii) The direct calculation of $y_k(n) = \sum_{j=0}^{n} h(j)x_k(n-j)$ requires $n+1$ multiplies. Therefore for n = 0, ..., N-1 we require $\sum_{n=0}^{N-1} n+1 = \sum_{n=1}^{N} n = N(N+1)/2$. This is the cost for each of K codebook entries so total is K times this quantity.

The calculation of $y_k(n) = h(n)x_k(0) + y_{k-1}(n-1)$ requires:

- k=0: full computation of $y_0(n) = \sum_{j=0}^{n} h(j)x_0(n-j)$: $n+1$ multiplies for each of n=0, ..., N-1 giving $N(N+1)/2$ as above.
- k=1, ..., K: recursive computation: 1 multiply for each of n=0,...,N-1 giving N multiplies in total.

Therefore the conditions of part (ii) reduce the number of multiplies by

$$(K-1)\left(\frac{N(N+1)}{2} - N\right)$$

6

(a) Source-filter model: bookwork.

For a vowel sound, the only source of acoustic excitation is the larynx and there are no constrictions in the vocal tract that close off the track or that induce turbulence. For a consonant, there are one or more constrictions in the vocal tract that do either close it off or induce turbulence. A consonant may have one or more of the following sources of excitation: vocal fold vibration, turbulence, explosive release of pressure.

The single excitation + all-pole vocal tract model is less good for consonants because (i) they may have multiple excitation sources at different points in the vocal tract, (ii) the portion of the vocal tract that is behind the excitation source will introduce zeros into the vocal tract transfer function.

(b) Both types of formant synthesiser use a 2-pole filter to generate the transfer function peak associated with each formant. In the cascade synthesiser, a single excitation signal is passed through each formant filter in turn to generate the speech. The resultant transfer function is therefore the product of those of the individual formant filters and will be all-pole.

In a parallel formant synthesiser, individual excitation signals are sent to each formant filter and their outputs are added together and passed through an output filter to give the synthesised speech. the resultant transfer function will include both poles and zeros. Because each formant filter has its own excitation signal, the amplitude and degree of voicing can be adjusted individually for each.

The cascade synthesiser is an excellent model for vowels: the relationship between the formant amplitudes and bandwidths will automatically be correct. For consonants however, the transfer function will not be correct and, for voiced consonants, the same transfer function will, incorrectly, be applied to the two excitation sources.

(c) Poles of *G(z)* are at $z = 0.57 \pm 0.76j = 0.95 \prec \pm 53.1°$ and those of H(z) are at $z = \pm 0.9\text{j}$.

$$G(z) + H(z) = \left(H^{-1}(z) + G^{-1}(z)\right)G(z)H(z) = (2 - 1.14z^{-1} + 1.7125z^{-2})G(z)H(z)$$

$$G(z) - H(z) = \left(H^{-1}(z) - G^{-1}(z)\right)G(z)H(z) = (-1.14z^{-1} + 0.0925z^{-2})G(z)H(z)$$

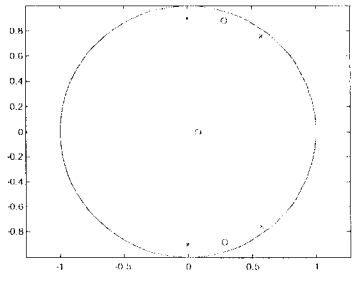
All three expressions have the same poles: namely the poles of G(z) and those of H(z). The zeros are as follows:

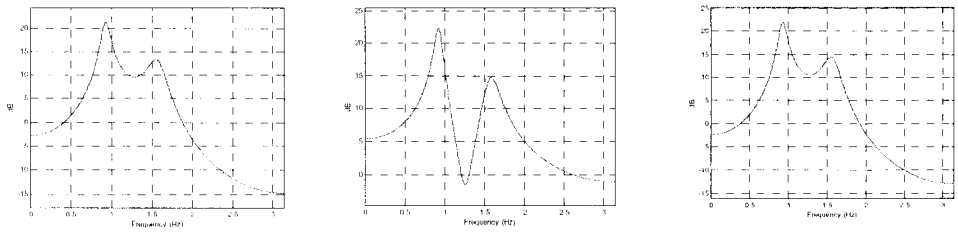
(i) No zeros

(ii) $0.285 \pm 0.88j = 0.925 \prec \pm 72°$

(iii) 0.0811

The diagram shows the poles and zeros for all three expressions. It can be seen that *G(z)+H(z)* contains a zero close to the unit circle that will cause a deep dip in the transfer function.





It can be seen that the third graph is a much better approximation to the all-pole transfer function than the second one. The formant filter outputs are added with alternating signs in order to eliminate the deep nulls that would otherwise occur between adjacent formants.

(d) A complex conjugate pole pair is required given by

$$G(z) = \frac{1}{(1-e^{-\pi b+2j\pi f}z^{-1})(1-e^{-\pi b-2j\pi f}z^{-1})} = \frac{1}{1-2e^{-\pi b}\cos(2\pi f)z^{-1}+e^{-2\pi b}z^{-2}}$$

with f = 912/8000 = 0.114 and b = 32/8000 = 0.004 giving

$$G(z) = \frac{1}{1-1.4897z^{-1}+0.9752z^{-2}}$$

(e) Since the turbulent excitation originates partway along the vocal tract whereas the periodic excitation originates at the larynx a different transfer function should be applied to the two signals. In practice, the proportion of larynx excitation is greater for low frequency formants than for high frequency formants. For a voiced consonant such as /z/, the voicing control would be around 0.5: from the graph we can see that F1 would be fully voiced whereas F4 would be fully unvoiced.