Paper Number(s): **E4.14**
SO16
ISE4.17

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE
UNIVERSITY OF LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2002

MSc and EEE/ISE PART IV: M.Eng. and ACGI

SPEECH PROCESSING

Thursday, 25 April 10:00 am

There are SIX questions on this paper.

Answer FOUR questions.

Corrected Copy

Time allowed: 3:00 hours

Examiners responsible:

First Marker(s): Brookes,D.M.

Second Marker(s): Naylor,P.A.

Special Instructions for Candidates: None

Special Instructions for Invigilators: None

1. A filter representing a single vocal tract formant in a system using a sample frequency of 8 kHz is defined by

$$H(z) = \frac{1}{1 - 1.4z^{-1} + 0.98z^{-2}}.$$

(a) Find the poles of $H(z)$ and its gain in dB at 0 Hz, 1 kHz and 4 kHz.

A modified filter is defined by $G(z) = H(z/k)$ where $0 < k < 1$. Determine its gain in dB at 0 Hz, 1 kHz and 4 kHz for the case when $k = 0.95$.

Draw a dimensioned sketch showing the frequency response in dB of both $H(z)$ and $G(z)$. [6]

(b) Explain why, in a speech coder, the filter coefficients derived from LPC analysis are often modified in the manner of part (a) before they are transmitted. [4]

(c) A spectral weighting filter is defined by

$$P_k(z) = \frac{H(z/k)}{H(z/0.95)}.$$

If k has the specific value 0.6, determine the filter's gain in dB at 0 Hz, 1 kHz and 4 kHz and draw a dimensioned sketch of its frequency response in dB. [5]

(d) In a speech coder, the transmitted parameters are selected to minimize the energy of the weighted synthesis error where $P_k(z)$ is used as the weighting filter. Explain the reason for using such a weighting filter and explain its effect on the spectrum of the synthesis error in the receiver. Say how varying the value of k will affect the action of the filter. [5]

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 $e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$ and the total prediction error is defined as $E = \sum_{n=0}^{N-1} e^2(n)$.

We define the vector $\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}^T$ and the polynomial $A(z) = 1 - \sum_{j=1}^{p} a_j z^{-j}$.

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

(b) The polynomials $P(z)$ and $Q(z)$ are defined by

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

Show that $Q(1) = \overset{0}{\cancel{1}}$ and that either $P(-1) = \overset{0}{\cancel{1}}$ or $Q(-1) = \overset{0}{4}$. [3]

corrected at 10.25 by M.B

(c) Show that for the case $A(z) = 1 + r^2 z^{-2}$ with $0 < r < 1$, the complex roots of $P(z)$ and $Q(z)$ are at $z = \exp(\pm j\theta)$ where $2\cos(\theta) = \pm(1 - r^2)$. Sketch a dimensioned graph of θ versus r for the range $0 < r < 1$ showing the values of θ that lie in the range $0 \le \theta \le \pi$. [7]

(d) In speech coders, the coefficients a_j are commonly transformed into Line Spectrum Frequencies before being transmitted. Explain which properties of the Line Spectrum Frequencies make them more suitable for transmission than either the coefficients a_j or the roots of $A(z)$. [4]

3. (a) *Figure 3.1* shows a simplified block diagram of a Code-Excited Linear Prediction (CELP) speech coder. Explain how the coder models the periodic and non-periodic components of the excitation signal, $r(n)$. [2]

(b) In a certain speech coder, the values of k and f are chosen to minimize the energy:

$$E(k,f)=\sum_{n=0}^{N-1}\left(r(n)-r(n-k)f\right)^2 .$$

Show that for a given value of k, the optimum value of f and the resultant energy are given by

$$\hat{f}(k)=\frac{\sum r(n)r(n-k)}{\sum r^2(n-k)} \text{ and } \hat{E}(k)=\sum r^2(n)-\frac{\left(\sum r(n)r(n-k)\right)^2}{\sum r^2(n-k)}$$

where all summations are over the range 0 to $N-1$. [5]

(c) The signal $r(n)$ is zero except for the following values of n:

n	–100	–95	–50	–45	0	5
$r(n)$	9	4	10	4	6	3

If N=40, determine $\hat{f}(k)$ and $\hat{E}(k)$ for $k=$ 45, 50, 60 and 100. [6]

(d) In a particular speech coder using a sample frequency of 8 kHz, the optimum values of k are found for each of three ranges such that $20\le k_1\le 39$, $40\le k_2\le 79$, $80\le k_3\le 143$. The final value of k is then chosen as follows

$$\hat{k}=\begin{cases} k_3 & \text{if } F(k_3)>\max\left(1.6F(k_1),1.6F(k_2)\right)\\ k_2 & \text{if } F(k_2)\ge\max\left(1.6F(k_1),F(k_3)/1.6\right)\\ k_1 & \text{otherwise}\end{cases}$$

where $F(k)=\sum r(n)^2-\hat{E}(k)$.

Explain the principle underlying this decision procedure and illustrate your answer by showing how it would affect the decision in the example of part (c) above. Explain why the quality of the decoded speech will be improved by using this procedure rather than by just selecting the k that minimises $\hat{E}(k)$. [7]

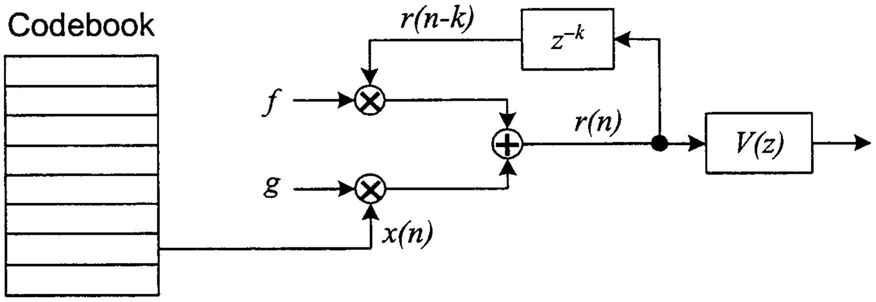


Figure 3.1

4. (a) Outline the major steps that a speech synthesis system must follow when converting a sentence from text into a sequence of phonemes. [4]

(b) *Figure 4.1* shows how a synthesis system varies the frequency of a particular formant at phoneme boundaries. Between phoneme 1 and phoneme 2 the frequency changes between the steady state values v_1 and v_2 as shown by the line labelled $f_2(t)$. The transition times at the end of phoneme 1 and the start of phoneme 2 are q_1 and p_2 respectively and the frequency at the boundary between the phonemes is $f_2(t_2) = b_2$. The formulae in *Table 4.1* for b_2, q_1 and p_2 make use of synthesis parameters that are taken from a lookup table whose entries for the phonemes /s/ and /u/ are listed in *Table 4.2*. In the formulae, the subscripts indicate whether a parameter is taken from the entry for phoneme 1 or phoneme 2 and, according to which of r_1 and r_2 is the greater, the expressions from either the first or the second row apply.

Determine the values of b_2, q_1 and p_2 for (i) the phoneme sequence /su/ and (ii) the sequence /us/. [3]

(c) Within phoneme 2, the final value for the frequency $f(t)$ is obtained by linearly interpolating between the functions $f_2(t)$ and $f_3(t)$. If the boundaries of phoneme 2 are at t_2 and t_3 as shown in *Figure 4.1*, then $f(t)$ is given by

$$f(t) = \frac{t_3 - t}{t_3 - t_2} f_2(t) + \frac{t - t_2}{t_3 - t_2} f_3(t) \qquad \text{for } t_2 \le t \le t_3.$$

Calculate the minimum value of $f(t)$ and draw a dimensioned sketch of $f_2(t)$, $f_3(t)$ and $f(t)$ for $t_2 \le t \le t_3$ for each of the following instances of the phoneme sequence /sus/: (i) $t_3 - t_2 = 120$ ms, (ii) $t_3 - t_2 = 50$ ms and (iii) $t_3 - t_2 = 30$ ms. [7]

(d) Explain the main differences between, and relative advantages of, a diphone synthesiser and a formant synthesiser. Explain how the variation of $f(t)$ with phoneme length that was illustrated in part (c) allows a formant synthesiser to model an aspect of human speech that is not modelled by a diphone synthesiser. [6]

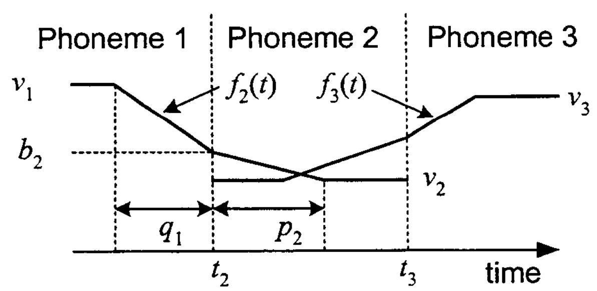


Figure 4.1

	b_2	q_1	p_2
$r_1 < r_2$	$w_2 + k_2 v_1$	l_2	h_2
$r_1 \ge r_2$	$w_1 + k_1 v_2$	h_1	l_1

Table 4.1

	r	v	w	k	l	h
/s/	6	1720 Hz	860 Hz	0.4	50 ms	20 ms
/u/	2	1200 Hz	600 Hz	0.55	30 ms	35 ms

Table 4.2

5. In a speaker identification system, the probability density of the input feature vector, $\mathbf{x}$, is modelled for each speaker as a multivariate gaussian:

$$p(\mathbf{x}) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right)$$

where N is the dimension of the feature vector, $\mathbf{m} = E[\mathbf{x}]$ is its mean and $\mathbf{C} = E[\mathbf{x}\mathbf{x}^T] - \mathbf{m}\mathbf{m}^T$ is its covariance matrix.

(a) Estimate the number of arithmetic operations needed to calculate $\ln(p(\mathbf{x}))$ when $\mathbf{C}$ is (i) an arbitrary matrix, (ii) a diagonal matrix and (iii) the identity matrix. You may assume that any required quantities that do not depend on $\mathbf{x}$ have been precalculated. Each addition, subtraction, multiplication or division counts as one operation. [4]

(b) A linear transformation is applied to the input feature vectors to form

$$\mathbf{y} = F\mathbf{x}.$$

Obtain expressions for the mean vector and the covariance matrix of $\mathbf{y}$ in terms of $\mathbf{m}$ and $\mathbf{C}$. [3]

(c) In a particular application, $N = 2$, the mean feature vectors for speakers A and B are $(2 \quad 6)^T$ and $(5 \quad 10)^T$ respectively and the covariance matrices equal the identity. Show that if $\mathbf{x} = (x_1 \quad x_2)^T$, the probability for speaker A that x_1 is less than a threshold T is given by: [2]

$$pr(x_1 < T) = \Phi(T-2) \text{ where } \Phi(y) = (2\pi)^{-1/2}\int_{t=-\infty}^{y}\exp(-\tfrac{1}{2}t^2).$$

(d) An unknown speaker is identified as speaker A if $x_1 < T$ and as speaker B if $x_1 \geq T$. If the prior probabilities of the speakers both equal 0.5, determine the probability of correct identification when (i) $T = 3$, (ii) $T = 3.5$ and (iii) $T = 4$. *Table 5.1* lists selected values of $\Phi(T)$. [4]

(e) The system is modified by transforming the input data as $\mathbf{y} = F\mathbf{x}$ using the matrix $F = \begin{pmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{pmatrix}$. If the speaker is identified as speaker A if $y_1 < T$ and as speaker B if $y_1 \geq T$, state what value of T should be used and determine the probability of correct identification. [4]

(f) Explain briefly why it can be advantageous in a speech or speaker recognition system to apply a linear transformation to the input feature vectors before they are used as the input of a pattern matching process. [3]

T	0	0.5	1.0	1.5	2.0	2.5
$\Phi(T)$	0.5	0.6915	0.8413	0.9332	0.9772	0.9938

Table 5.1

6. (a) An utterance consists of T frames, $\mathbf{x}_1, \ldots, \mathbf{x}_T$, 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 $d_i(\mathbf{x}_t)$.

$P(s,t)$ is defined to be the total probability that the model generates frames $\mathbf{x}_1, \ldots, \mathbf{x}_t$ and that frame t corresponds to state s. $Q(s,t)$ is defined to be the total probability that the model generates frames $\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T$ given that frame t corresponds to state s. 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 $Q(s,T)$. [5]

(b) A 5-frame utterance is used to train a 3-state Hidden Markov model. *Table 6.1* shows the output probability of each frame from each state of the model and *Figure 6.1* shows the state diagram of the model including the transition probabilities.

For each of $s = 1$, 2 and 3, calculate the total probability that frame $\mathbf{x}_3$ corresponds to state s and that the model generates $\mathbf{x}_1, \ldots, \mathbf{x}_5$. You should perform your calculations to 5 decimal places. [10]

(c) In the Baum-Welch training procedure, the feature vector mean for state s is re-estimated as a weighted average of the input frames:

$$\mathbf{m}_s = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_n(s,t) \times \mathbf{x}_{n,t}}{\sum_{n=1}^{N}\sum_{t=1}^{T_n} A_n(s,t)} \quad \text{where } A(t,s) = \frac{P(s,t) \times Q(s,t)}{\sum_{j=1}^{S}\left(P(j,t) \times Q(j,t)\right)}$$

and the subscript n refers to the n^{th} training word in all cases. Explain the significance of the terms $A(s,t)$ and justify their use as weights in the expression for $\mathbf{m}_s$. [5]

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

Table 6.1

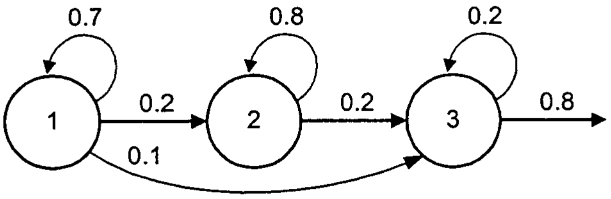


Figure 6.1

E4.14
Sol6
ISE 4.17

Solutions to Speech Processing 2002

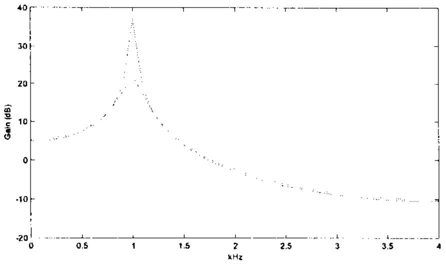
1. (a) The poles of $H(z)$ are at $0.7 \pm 0.7j$.

Freq	z	z^2	Gain^{-1}	Gain	dB
0 Hz	1	1	0.58	1.7	4.2
1 kHz	$(1+j)/\sqrt{2}$	j	0.014	70.7	37
4 kHz	-1	1	3.38	0.3	−10.6

The poles of $G(z) = \left(1 - 1.33z^{-1} + 0.884z^{-2}\right)^{-1}$

are at $0.95(0.7 \pm 0.7j) = 0.665 \pm 0.665j$.

Freq	z	z^2	Gain^{-1}	Gain	dB
0 Hz	1	1	0.554	1.8	5.1
1 kHz	$(1+j)/\sqrt{2}$	j	0.081	12.2	21.8
4 kHz	-1	1	3.21	0.311	−10.1



[6]

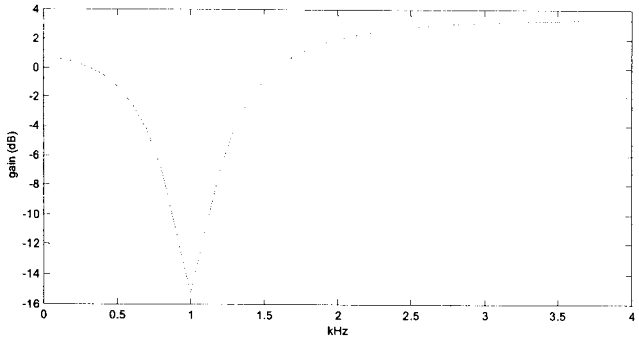
(b) The LPC anlysis may generate filters with poles very close to the unit circle. The bandwidth expansion procedure ensures that these are moved towards the origin. Advantages are:

- Eliminates incorrectly sharp resonances that may arise in LPC analysis
- Less sensitivity to quantisation errors for most alternative parametrisations including LSF coefficients.
- Reduced range of LSF coefficients gives more efficient coding.

[4]

(c) The poles of $H(z/0.6) = \left(1 - 0.84z^{-1} + 0.3528z^{-2}\right)^{-1}$ are at $0.6(0.7 \pm 0.7j) = 0.42 \pm 0.42j$.

Freq	z	z^2	**Gain^{-1}**	**Gain**	**dB**	**P dB**
0 Hz	1	1	0.5128	1.9501	5.8010	0.6783
1 kHz	$(1+j)/\sqrt{2}$	j	0.4723	2.1175	6.5165	–15.234
4 kHz	-1	1	2.1928	0.456	-6.82	3.3222



[5]

(d) The effect of the weighting filter is to down-weight errors at the formant frequency where the speech amplitude is high while leaving the gain at other frequencies at around 0 dB. The consequence of this is that when searching for the best excitation signal, the coder will tend to ignore errors at the formant frequencies and concentrate on reducing the error spectrum at other frequencies where noise will be more audible because it is not masked by high speech energy. The larger the value of *k*, the less effect the weighting filter will have.

[5]

2. (a) Differentiating the expression for E gives:

$$\tfrac{1}{2}\frac{\partial E}{\partial a_k} = \sum_{n=0}^{N-1} e(n)\frac{\partial e(n)}{\partial a_k} = -\sum_{n=0}^{N-1} e(n)s(n-k)$$

$$= \sum_{m=1}^{p} a_m \sum_{n=0}^{N-1} s(n-m)s(n-k) - \sum_{n=0}^{N-1} s(n)s(n-k)$$

Setting this to zero gives $\sum_{m=1}^{p} r_{km}a_m = b_k$ for $k = 1, 2, \ldots, p$

In matrix form, this is $\mathbf{Ra} = \mathbf{b}$ [6]

(b) Firstly: $Q(1) = A(1) - 1^{-(p+1)}A(1) = 0$

Also:

$$P(-1) = A(-1) + (-1)^{-(p+1)}A(-1) = \begin{cases} 0 & \text{if } p \text{ even} \\ 2A(-1) & \text{if } p \text{ odd} \end{cases}$$

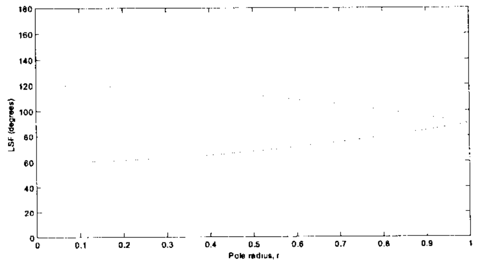
$$Q(-1) = A(-1) - (-1)^{-(p+1)}A(-1) = \begin{cases} 2A(-1) & \text{if } p \text{ even} \\ 0 & \text{if } p \text{ odd} \end{cases}$$ [3]

Thus one or other of P and Q has a root at -1.

(c) Since we know the real roots of $P(z)$ and $Q(z)$, we can factor them out to give a quadratic in each case:

$$P(z) = 1 + r^2z^{-1} + r^2z^{-2} + z^{-3} = \left(1 + z^{-1}\right)\left(1 - \left(1 - r^2\right)z^{-1} + z^{-2}\right)$$
$$Q(z) = 1 - r^2z^{-1} + r^2z^{-2} - z^{-3} = \left(1 - z^{-1}\right)\left(1 + \left(1 - r^2\right)z^{-1} + z^{-2}\right)$$

The roots of the quadratic $\left(1 - 2\cos(\theta)z^{-1} + z^{-2}\right)$ are at $z = \exp(\pm j\theta)$ so if $2\cos(\theta) = \pm\left(1 - r^2\right)$, these are the roots of P and Q. The two roots with positive imaginary parts are plotted below (in degrees) as a function of r.



[7]

(d) 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.

[4]

3. (a) The periodic component of $r(n)$ is modelled by the upper feedback loop that involves a delay of z^{-k}. The aperiodic component is modelled by selecting the entry from a fixed codebook that gives the lowest resynthesis error. [2]

(b) To find the optimum *f*, we differentiate *E* and set it to zero:

$$\frac{\partial E}{\partial f} = -2\sum (r(n) - fr(n-k))r(n-k)$$

$$\Rightarrow \quad \sum r(n)r(n-k) - \hat{f}\sum r^2(n-k) = 0$$

$$\Rightarrow \quad \hat{f} = \frac{\sum r(n)r(n-k)}{\sum r^2(n-k)}$$

Substituting this value into the expression for *E* gives

$$\hat{E} = \sum (r(n) - \hat{f}r(n-k))^2 = \sum r^2(n) - 2\hat{f}\sum r(n)r(n-k) + \hat{f}^2 \sum r^2(n-k)$$

$$= \sum r^2(n) - 2\frac{\left(\sum r(n)r(n-k)\right)^2}{\sum r^2(n-k)} + \frac{\left(\sum r(n)r(n-k)\right)^2}{\sum r^2(n-k)} = \sum r^2(n) - \frac{\left(\sum r(n)r(n-k)\right)^2}{\sum r^2(n-k)}$$

[5]

(c) At most two terms in the summations are non-zero. This gives

k	45	50	60	100
$\hat{f}$	24/16=1.5	72/116=0.62	0	66/97=0.68
$\hat{E}(k)$	9	0.31	45	0.09
$F(k)$	36	44.69	0	44.91

Thus $k = 100$ gives the lowest value of $\hat{E}(k)$. [6]

(d) We have k_1 arbitrary with $F(k_1) = 0$, $k_2 = 50$ with $F(k_2) = 44.69$ and $k_3 = 100$ with $F(k_3) = 44.91$.

We have $\begin{array}{l} F(k_3) < 1.6F(k_2) \\ F(k_2) > 1.6F(k_1) \quad F(k_2) > 0.625F(k_3) \end{array}$

Hence the first condition is not satisfied but the second is and so *k*$_2$ is selected.

The decision procedure guards against selecting a pitch period that is twice the correct one. Each of the three ranges spans a single octave so that an autocorrelation peak that corresponds to a harmonic of another peak will be in a different range. The algorithm will choose the pitch period of the fundamental even if its $F(k)$ value is as much as 2 dB less that that of its harmonic. [7]

The reason for wanting to avoid pitch period doubling is that it will introduce many harmonics that should not be there and so will not generate a good signal.

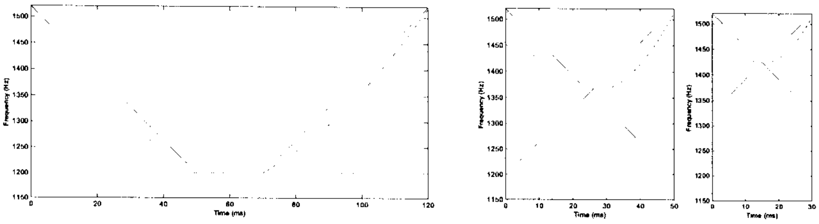
4. (a) Major steps in text-to-speech:

- Convert non-words (e.g. acronyms, numbers, dates) to words
- Convert words to phonemes using both a dictionary and rules
- Insert pauses between phrases
- Sort out prosodic variations of stress, pitch and duration
- Generate acoustic signal
  - Use a concatenative or a rule-based synthesiser. [4]

(b) For /su/, the first phoneme has the higher rank, so we use the second line of formulae. These give: $b_2 = 1520\text{ Hz}$, $q_1 = 20\text{ ms}$, $p_2 = 50\text{ ms}$.

For /us/, the /s/ is again the higher ranked phoneme so the values stay the same but p_2 and q_1 swap: $b_2 = 1520\text{ Hz}$, $q_1 = 50\text{ ms}$, $p_2 = 20\text{ ms}$. [3]

(c) The graphs show $f_2(t)$, $f_3(t)$ and $f(t)$ for durations of 120 ms, 50 ms and 30 ms:



The minimum reached is $1200\text{ Hz}, 1520-320\times 50/100 = 1360\text{ Hz}$ and $1520-320\times 30/100 = 1424\text{ Hz}$. [7]

(d) A diphone synthesiser generates speech by concatenating segments of real speech that stretch from the centre of one phoneme to the centre of the next. It needs a mechanism to adjust the pitch and duration of each segment to ensure that these characteristics change smoothly. A formant synthesiser generates speech by using rules that relate phoneme sequences to spectral characteristics such as formant frequencies.

Diphone synthesisers are computationally much simpler and, since they are based on actual speech recordings, they reproduce accurately the detailed transitions between phonemes. However because they generally have only one version of each phoneme transition, they cannot alter this in response to changing environments or different prosody. [6]

The behaviour of the formant synthesiser in part (c) models the characteristic of human speech that toungue position targets, and hence formant frequency targets are not attained fully in rapid speech.

5. (a) We have $\ln(p(\mathbf{x})) = -\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) + k$ where the constant k does not depend on $\mathbf{x}$. For the three types of $\mathbf{C}$ we get:

Calculation	Arbitrary	Diagonal	Identity	
$(\mathbf{x}-\mathbf{m})$	N	N	N	subtract
$(-\frac{1}{2}\mathbf{C}^{-1})(\mathbf{x}-\mathbf{m})$	N^2	N	0	multiply
	N^2-N	0	0	add
$(\mathbf{x}-\mathbf{m})^T \times \ldots$	N	N	N	multiply
	N–1	N–1	N–1	add
$\ldots + k$	1	1	1	add
Total	$2N^2+2N$	$4N$	$3N$	

[4]

(b) $\mathbf{m}_y = E[\mathbf{Fx}] = \mathbf{Fm}$

$$\mathbf{C}_y = E[\mathbf{y}\mathbf{y}^T] - \mathbf{m}_y\mathbf{m}_y^T = E[\mathbf{F}\mathbf{x}\mathbf{x}^T\mathbf{F}^T] - \mathbf{m}_y\mathbf{m}_y^T = \mathbf{FCF}^T + \mathbf{F}\mathbf{m}\mathbf{m}^T\mathbf{F}^T - \mathbf{m}_y\mathbf{m}_y^T = \mathbf{FCF}^T$$ [3]

(c) If the covariance matrix is the identity then the two components x_1 and x_2 are independent unit variance gaussians. Therefore

$$pr(x_1 < T) = (2\pi)^{-\frac{1}{2}} \int_{x=-\infty}^{T} \exp\left(-\tfrac{1}{2}(x-m_1)^2\right)dx = (2\pi)^{-\frac{1}{2}} \int_{t=-\infty}^{T-m_1} \exp\left(-\tfrac{1}{2}t^2\right)dx = \Phi(T-m_1)$$ [2]

(d) If the true identity is A, then the probability of correct recognition is $\Phi(T-m_1)$ while if the true identity is B, the probability of correct recognition is $1-\Phi(T-m_2) = \Phi(m_2-T)$. Since the prior probabilities of A and B both equal ½, the overall probability of correct recognition is

$(\Phi(T-m_1) + \Phi(m_2-T))/2$

(i) $T = 3$: $(\Phi(1)+\Phi(2))/2 = 0.9093$

(ii) $T = 3.5$: $(\Phi(1.5)+\Phi(1.5))/2 = 0.9332$ [4]

(iii) $T = 4$: $(\Phi(2)+\Phi(1))/2 = 0.9093$

Unsurprisingly, the optimum choice for T is midway between m_1 and m_2.

(e) We obtain $\mathbf{m}_{y,A} = \mathbf{Fm}_A = (6 \quad 2)^T$ and $\mathbf{m}_{y,B} = \mathbf{Fm}_B = (11 \quad 2)^T$

We choose T to be midway between 6 and 11, that is, 8.5. [4]

Thus probability of correct recognition is now $(\Phi(2.5)+\Phi(2.5))/2 = 0.9938$

(f) A linear transformation can be used

(i) To make the within class covariance matrices approximately diagonal thus saving computation as in part (a). [3]

(ii) To compress the useful discriminative information into a smaller number of parameters. This saves computation and can also, in practice, improve performance.

6. (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

$$P(s,t) = \sum_{i=1}^{S} P(i,t-1) \times \text{prob(frame } t \text{ is in state } s \mid \text{frame } t-1 \text{ is in state } i)$$

$$= \sum_{i=1}^{S} P(i,t-1) \times a_{is} \times d_s(\mathbf{x}_t)$$

By a similar argument

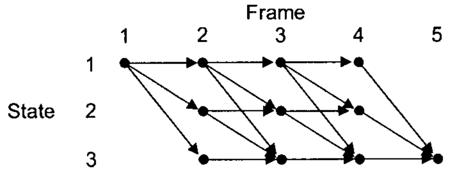
$$Q(s,t) = \sum_{i=1}^{S} a_{si} \times d_i(\mathbf{x}_{t+1}) \times Q(i,t+1)$$

Under the assumption that frames 1 and *T* must be in states 1 and *S* respectively, the initial conditions for the recursions are: [5]

$$P(s,1) = d_1(\mathbf{x}_1) \quad \text{for} \quad s = 1 \quad \text{else} \quad = 0$$

$$Q(s,T) = a_{S=} \quad \text{for} \quad s = S \quad \text{else} \quad = 0$$

(b) The possible paths through the State-Frame lattice are shown below.



P(1,1) = 0.5
P(1,2) = 0.5 × 0.7 × 0.5 = 0.175
P(2,2) = 0.5 × 0.2 × 0.1 = 0.01
P(3,2) = 0.5 × 0.1 × 0.4 = 0.02
P(1,3) = 0.175 × 0.7 × 0.3 = 0.03675
P(2,3) =(0.175 × 0.2 + 0.01 × 0.8) × 0.8 = 0.0344
P(3,3) = (0.175 × 0.1 + 0.01 × 0.2 + 0.02 × 0.2) × 0.5 = 0.01175

[10]

Q(3,5) = 0.8
Q(1,4) = 0.1 × 0.5 × 0.8 = 0.04
Q(2,4) = 0.2 × 0.5 × 0.8 = 0.08
Q(3,4) = 0.2 × 0.5 × 0.8 = 0.08
Q(1,3) = 0.7 × 0.1 × 0.04 + 0.2 × 0.1 × 0.08 + 0.1 × 0.4 × 0.08 = 0.0076
Q(2,3) = 0.8 × 0.1 × 0.08 + 0.2 × 0.4 × 0.08 = 0.0128
Q(3,3) = 0.2 × 0.4 × 0.08 = 0.0064

P×Q(1,3) = 0.03675 × 0.0076 = 0.00028
P×Q (2,3) = 0.0344 ×0.0128 = 0.00044
P×Q (3,3) = 0.01175 × 0.0064 = 0.00008

(c) The numerator of $A(s,t)$ is the quantity calculated in part (b), that is, the probability that the model generates the utterance and that frame *t* is in stte *s*. The numerator is the sum of these over all states which is just the total probability that the model generates the utterance. It follows that $A(s,t)$ is therefore the conditional probability that frame *t* is in state *s* given that the model generates the utterance.

The expression for $\mathbf{m}_s$ is a weighted average taken over all the training data frames. Each frame is weighted by the conditional probability that it corresponded to state s. This training procedure in which a training frame contributes to all the states to a greater or lesser extent gives more accurate results than Viterbi training which assigns each training frame to only a single state. [5]