EE4-14
EE9SO16

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2011

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Thursday, 19 May 2:30 pm

Time allowed: 3:00 hours

There are FOUR questions on this paper.

Answer ALL 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) : W. Dai

1. a) Briefly describe Linear Predictive Coding (LPC) and it's use in speech signal processing applications. Your description should summarize the many concepts in text and diagrams. It is not necessary to give or derive any mathematical formulae. [6]

b) A linear predictor for the speech signal $s(n)$ can be written

$$s(n) = Gu'(n) + \sum_{j=1}^{p} a_j s(n-j).$$

Explain what the terms $u'(n)$, G and a_j represent in this expression. [2]

c) Formulate a time domain expression for the prediction error and hence show that the prediction error is minimum when the following expression is satisfied:

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

where $\{F\}$ denotes the set of speech samples within a frame. [4]

d) i) Define the complex cepstral coefficients of a speech signal and briefly state their important properties for speech processing. [2]

ii) Derive a recursion to obtain the complex cepstral coefficients from the terms a_j as defined by parts b) and c) of this question. [5]

iii) Find the complex cepstral coefficients for the case when

$$[a_0, a_1, a_2, a_3, a_4] = [1.00, -0.20, -0.25, 0.18, 0.40].$$

[1]

2. a) It is usual to extract features from a speech signal when performing automatic speech recognition. Now consider Mel Frequency Cepstral Coefficients (MFCCs) as such speech features.

i) Explain, with the aid of appropriate diagrams, the procedure for calculating MFCCs from the speech signal. [3]

ii) Justify the use of MFCCs for automatic speech recognition in terms of their advantageous characteristics, including any important statistical properties. [3]

b) Consider 5 frames representing a segment of a speech signal, each frame comprising a vector of features. The frame rate is 100 Hz. We wish to train a 3-state Hidden Markov Model (HMM) using these 5 frames.

The state diagram of the HMM is shown in Fig. 2.1 and Table 1 shows the output probability of each frame from each state of the model.

For each of state $s = 1,2,3$ calculate the total probability that frame $\mathbf{x}_2$ corresponds to state s and that the model generates $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_5$. You should perform your calculations to 6 decimal places.

[8]

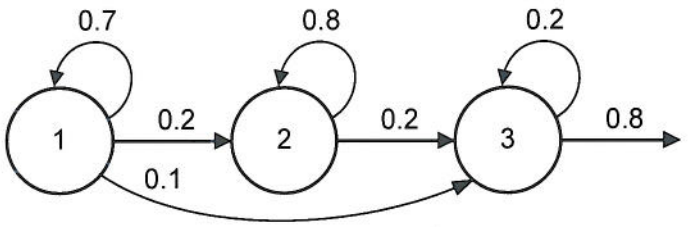

Figure 2.1

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

Table 1

c) In practical applications, speech signals are almost always degraded by some additive noise.

i) Briefly explain why additive noise is almost always present in speech signals of this type. [1]

ii) What impact would you expect additive noise to have on an automatic speech recognizer? Explain your reasoning. [1]

iii) Using qualitative explanations of the concepts, suggest two approaches to improve the robustness of an automatic speech recognizer to additive noise. [2]

d) Explain, using appropriate mathematical formulations, a process by which the statistical properties of the MFCCs can be improved by one or more linear transformations. [2]

3. a) It is known that poles in a system function $H(z)$ can cause resonant peaks in the magnitude of the system's frequency response. For the specific case of a single complex pole at radius $r < 1$, show that the 3 dB bandwidth b of the corresponding resonant peak in the frequency response can be approximated as $b = 2(1-r)$. Include an illustrative sketch and state the units of b. [3]

b) Consider a linear time-invariant system with system function

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}}.$$

i) Explain the concept of *bandwidth expansion* and show how to apply bandwidth expansion to $H(z)$. Describe the main application of bandwidth expansion in the context of speech processing. [3]

ii) Based on your explanation of bandwidth expansion and using the approximate expression for b as given in part a), write a new expression for the bandwidth of a resonant peak after bandwidth expansion. Clearly define all symbols in the new expression. [1]

c) Parts of a CELP speech coder are shown in Fig. 3.1 with speech excitation signal $r(n)$.

i) Write brief explanations of how this speech coder models the effects of:

- the vocal tract,
- the periodic components of $r(n)$,
- the non-periodic components of $r(n)$.

[3]

ii) Assume that the values of k and f in this coder are determined so as to minimize the value of

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

Show that the optimum value of f for a given value of k is given by

$$f_{opt}(k) = \frac{\sum_0^{N-1} r(n)r(n-k)}{\sum_0^{N-1} r^2(n-k)}.$$

[4]

iii) For $f(k) = f_{opt}(k)$, find and simplify the corresponding expression for

$$E_{opt}(k) = E(k, f_{opt}).$$

[3]

iv) The non-zero values of $r(n)$ are defined in Table 1. Except for the values of n shown in the table, $r(n) = 0$. For the case $N = 40$, calculate $f_{opt}(k)$ and $E_{opt}(k)$. [3]

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

Table 1

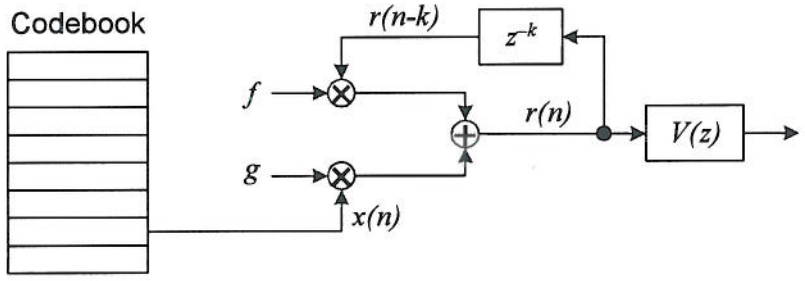

Figure 3.1

4. a) The utterance "Lunchtime doubly so" spoken by a male talker is shown in the spectrograms of Figs. 4.1 and 4.2.

i) Which one of Figs. 4.1 or 4.2 corresponds to a wideband spectrogram? Give your reasoning. [2]

ii) Choose one of the two spectrograms and estimate the fundamental frequency of the speech at 0.9 s. Explain your method of estimation. [2]

iii) Draw a sketch of a spectrogram showing only the formant tracks during the utterance of "time". Comment on the key characteristics observed. [3]

b) Consider a lossless tube in which the sound pressure and volume velocity are governed by

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t} \tag{4.1}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \tag{4.2}$$

where $p(x,t)$ is the sound pressure in the tube at position x and time t, $u(x,t)$ is the volume velocity (flow) at position x and time t, ρ is the density of air in the tube, c is the speed of sound and $A(x,t)$ is the cross-sectional area of the tube. In the context of modelling the vocal tract, $x = 0$ corresponds to the position of the glottis and $x = l$ corresponds to the position of the lips.

i) Re-write these expressions for the case of a tube with constant cross-sectional area $A(x,t) = A$. [1]

ii) For the case of constant cross-sectional area A, show that

$$u(x,t) = u^+(t - x/c) - u^-(t + x/c)$$
$$p(x,t) = \frac{\rho c}{A}\left(u^+(t - x/c) + u^-(t + x/c)\right)$$

are solutions to the above equations (4.1) and (4.2) in which $u^+(t - x/c)$ and $u^-(t + x/c)$ correspond to travelling waves in the positive and negative directions respectively. [3]

iii) Next consider the boundary conditions

$$u(0,t) = U_G(\Omega)e^{j\Omega t}$$
$$p(l,t) = 0,$$

with Ω representing frequency.

Let $u^+(t - x/c)$ and $u^-(t + x/c)$ be of the form

$$u^+(t - x/c) = K^+ e^{j\Omega(t - x/c)}$$
$$u^-(t + x/c) = K^- e^{j\Omega(t + x/c)}.$$

Find expressions for K^+ and K^- and hence write down solutions for $u(x,t)$ and $p(x,t)$ in terms of $U_G(\Omega)$. [6]

iv) Hence, sketch a plot of the magnitude of the volume velocity transfer function defined as the ratio in the frequency domain between the volume velocity at the lips and the volume velocity at the glottis. [3]

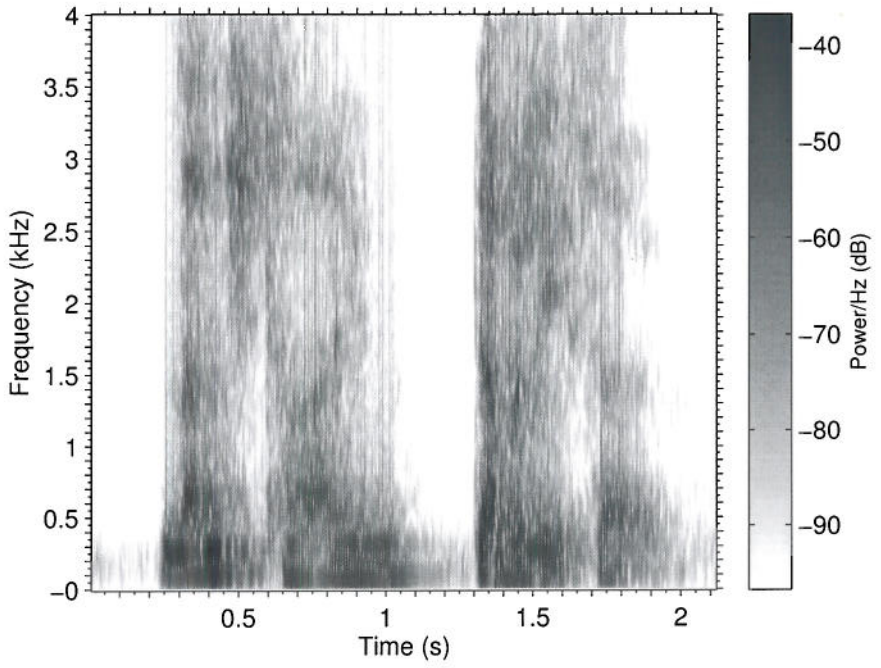

Figure 4.1

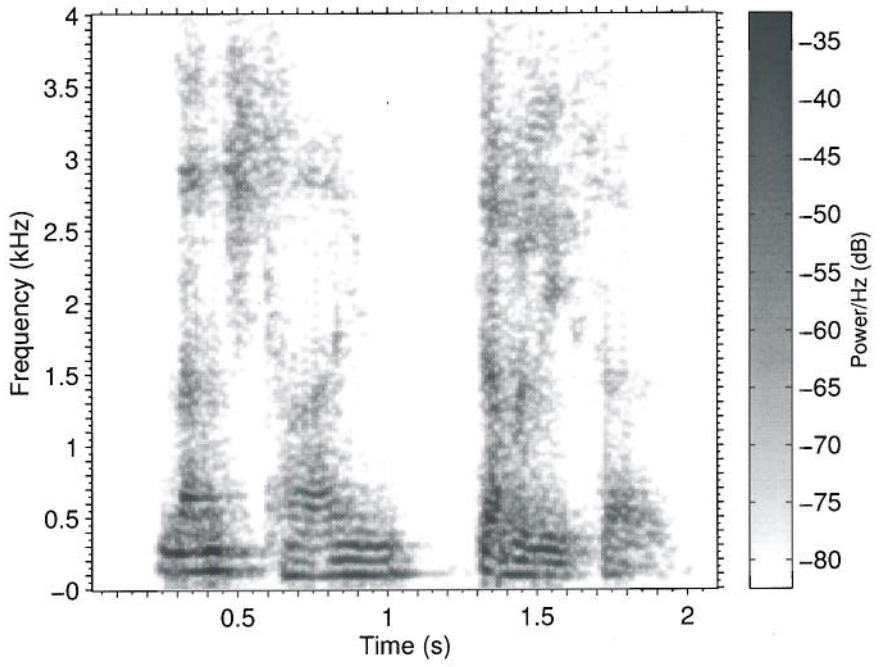

Figure 4.2

SPEECH PROCESSING Solutions 2011

1. a) LPC is used as a method of spectral estimation from which characteristic features of speech can be obtained. The LPC coefficients are the denominator coefficients of an all pole (AR) transfer function. A close link can be derived between the AR system and the vocal tract transfer function. The LPC coefficients may be used in speech coding (often after conversion to Line Spectral Frequencies) and speech recognition (often after conversion to cepstral coefficients.)

b)
$$s(n) = Gu'(n) + \sum_{j=1}^{p} a_j s(n-j)$$

$u'(n)$ represents the derivative of glottal flow.
G represents a gain factor.
a_j represent the predictor coefficients, or the coefficients of the denominator polynomial of the vocal tract transfer function.

c) Under the assumption that the second term dominates, the prediction error is given by
$$e(n) = s(n) - \sum_{j=1}^{p} a_j s(n-j).$$

The sum squared prediction error is $Q_E = \sum_{n\in\{F\}} e^2(n)$ which is differentiated w.r.t. each a_j giving

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

The optimum values of a_j are then found from

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

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

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

d) i) Cepstrum is defined as inverse Fourier transform of log spectrum.

$$c_n = \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \log\left(V(e^{j\omega})\right) e^{j\omega n} d\omega.$$

It can be computed either from roots of the prediction filter polynomial or from the coefficients of the prediction filter polynomial:

- good at discriminating between different phonemes
- fairly independent of each other

- have approximately Gaussian distributions for a particular phoneme.

ii)

$$C(z) = \log(G) - \log(A(z))$$
$$A(z)zC'(z) = -zA'(z)$$
$$\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)\left(z \sum_{m=0}^{\infty} -mc_m z^{-(m+1)}\right) = -z \sum_{n=1}^{p} na_n z^{-(n+1)}$$
$$\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)\left(\sum_{m=1}^{\infty} mc_m z^{-m}\right) = \sum_{n=1}^{p} na_n z^{-n}$$
$$\sum_{n=1}^{\infty} nc_n z^{-n} - \sum_{k=1}^{p}\sum_{m=1}^{\infty} mc_m a_k z^{-(m+k)} = \sum_{n=1}^{p} na_n z^{-n}$$
$$\sum_{n=1}^{\infty} nc_n z^{-n} = \sum_{n=1}^{p} na_n z^{-n} + \sum_{k=1}^{p}\sum_{n=k+1}^{\infty} (n-k)c_{(n-k)}a_k z^{-n}.$$

Now take the coefficient of z^{-n} in the above equation noting that

$$n \geq k+1 \Rightarrow k \leq n-1$$

to give

$$nc_n = na_n + \sum_{k=1}^{\min(p,n-1)} c_{(n-k)}a_k$$
$$\Rightarrow c_n = a_n + \frac{1}{n}\sum_{k=1}^{\min(p,n-1)} c_{(n-k)}a_k$$

iii)

$$c = \begin{pmatrix} 1.0000 & 0.3000 & -0.1167 & -0.0000 & 0.4086 \end{pmatrix}$$

2. a)

- Divide signal into overlapping 25 ms segments at 10 ms intervals. Frame rate = 100 frames/s (Hz)
- Apply Hamming window and take FFT
- Smooth the spectrum with a mel filter bank. Mel filter bank concentrates data values in the more significant part of the spectrum
- Take the log of the mel spectrum. Variations in signal level just cause a DC shift in the log spectrum. Gaussian approximation is more nearly true for log spectrum than for the power spectrum directly
- Discrete Cosine Transform (DCT). Reduces correlation between coefficients. Compresses information into fewer low-order coefficients. Output is the mel-cepstrum
- DC component is ignored to make it independent of signal level
- First and Second time derivatives. Provide additional information about how the spectrum is changing with time
- Result is a 39 element feature vector. Sometimes 38 if the log energy is not included

b) 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

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

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:

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

$$Q(s,T) = \begin{cases} a_{S=} & \text{for} \quad s=S \\ 0 & \text{otherwise} \end{cases}$$

The lattice is as shown in Fig. 2.1

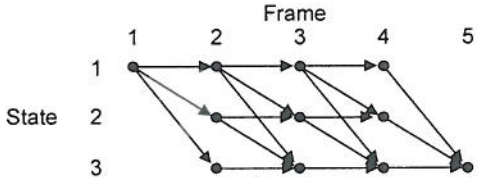

Figure 2.1

The we compute P and Q as follows.

$$
\begin{aligned}
&\text{P(1,1)} = 0.5 \\
&\text{P(1,2)} = 0.5 \times 0.7 \times 0.5 = 0.175 \\
&\text{P(2,2)} = 0.5 \times 0.2 \times 0.1 = 0.01 \\
&\text{P(3,2)} = 0.5 \times 0.1 \times 0.4 = 0.02
\end{aligned}
$$

$$
\begin{aligned}
&\text{Q(3,5)} = 0.8 \\
&\text{Q(1,4)} = 0.1 \times 0.5 \times 0.8 = 0.04 \\
&\text{Q(2,4)} = 0.2 \times 0.5 \times 0.8 = 0.08 \\
&\text{Q(3,4)} = 0.2 \times 0.5 \times 0.8 = 0.08 \\
&\text{Q(1,3)} = 0.7 \times 0.1 \times 0.04 + 0.2 \times 0.2 \times 0.08 + 0.1 \times 0.3 \times 0.08 = 0.0084 \\
&\text{Q(2,3)} = 0.8 \times 0.2 \times 0.08 + 0.2 \times 0.3 \times 0.08 = 0.0176 \\
&\text{Q(3,3)} = 0.2 \times 0.3 \times 0.08 = 0.0048 \\
&\text{Q(1,2)} = 0.7 \times 0.3 \times 0.0084 + 0.2 \times 0.8 \times 0.0176 + 0.1 \times 0.5 \times 0.0048 = 0.0048 \\
&\text{Q(2,2)} = 0.8 \times 0.8 \times 0.0192 + 0.2 \times 0.5 \times 0.0064 = 0.0117 \\
&\text{Q(3,2)} = 0.2 \times 0.5 \times 0.0048 = 0.0005
\end{aligned}
$$

$$
\begin{aligned}
&\text{P(1,2)Q(1,2)} = 0.175 \times 0.0048 = 0.000844 \\
&\text{P(2,2)Q(2,2)} = 0.01 \times 0.0117 = 0.000117 \\
&\text{P(3,2)Q(3,2)} = 0.02 \times 0.0005 = 0.0000096
\end{aligned}
$$

c) i) Noise from the ambient acoustics as well as electrical noise and interference are ubiquitous.

ii) Noise modifies the feature vectors of speech so that they match less well to those feature vectors used during training.

iii) Speech enhancement might be used to reduce the level of noise in noisy speech prior to ASR, the ASR having been trained on clean speech. However, such speech enhancement processing is known to add artefacts (musical noise) that will likely damage ASR performance. Alternatively, ASR can be trained on noisy speech but in this case the number of different noisy types that can be considered must be kept small in order to maintain computational complexity within tractable limits.

d) The feature vectors $\mathbf{x}$ are ideally independent but in reality not so. Their properties can be improved by matrix transformation of the type

$$\mathbf{y} = \mathbf{F}^T \mathbf{x}$$

with $\mathbf{F} = \mathbf{Y}\mathbf{D}^{1/2}$ where $\mathbf{Y}$ and $\mathbf{D}$ are matrices of eigenvectors and eigenvalues respectively of the average of the within state covariance matrices.

3. a) The approximation comes from the geometric relationships for the 3 dB bandwidth arising from the right angle triangle shown in Fig.

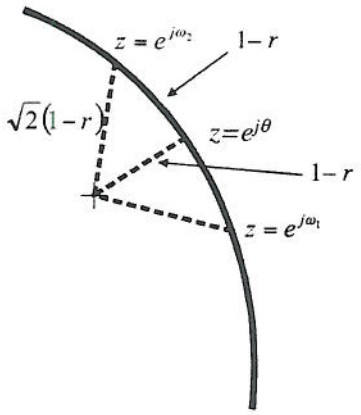

Figure 3.1

b) i) From the original $H(z)$, we can form a new filter by multiplying coefficients a_i and b_i by k^i for some $k < 1$ to give

$$G(z) = H(z/k) = \frac{b_0 + b_1 k z^{-1} + b_2 k z^{-2}}{a_0 + a_1 k z^{-1} + a_2 k z^{-2} + a_3 k z^{-3}}$$

such that if $H(z)$ has a pole/zero at z_0 then $G(z)$ will have one at kz_0 so that all poles and zeros will be moved towards the origin by a factor k.

ii) Given a pole in $H(z)$ with bandwidth $b = 2(1-r)$, the corresponding pole after bandwidth expansion has bandwidth

$$2(1-kr) = b + 2r(1-k).$$

c) i) The vocal tract is modelled by the all-pole filter $V(z)$. 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.

ii) By setting the derivative of E to zero we obtain

$$\begin{aligned}
\frac{\partial E}{\partial f} &= -2\sum (r(n) - f r(n-k))\, r(n-k) \\
&\Rightarrow \sum r(n) r(n-k) - f_{opt} \sum r^2(n-k) = 0 \\
f_{opt} &= \frac{\sum r(n) r(n-k)}{r^2(n-k)}
\end{aligned}$$

iii)

$$\begin{aligned}
E_{opt} &= \sum (r(n) - f_{opt} r(n-k))^2 \\
&= \sum r^2(n) - 2 f_{opt} \sum r(n) r(n-k) + f_{opt}^2 \sum r^2(n-k) \\
&= \sum r^2(n) - \frac{(\sum r(n) r(n-k))^2}{\sum r^2(n-k)}
\end{aligned}$$

iv) Only two terms are non-zero in the summation. The computations are shown in Table 1.

k	45	50	60	100
f_{opt}	24/16=1.5	72/116=0.62	0	66/97=0.68
E_{opt}	9	0.31	45	0.09

Table 1

4. a) i) The wideband spectrogram is Fig. 4.1 and the narrowband is Fig. 4.2. Narrowband spectrograms have better frequency resolution and in this case the harmonics of the fundamental are clearly resolved.

ii) Using Fig. 4.2, the harmonics of the fundamental at t=0.9 s are spaced by approximately 110 Hz. This corresponds to the fundamental frequency.

iii) The sketch is shown in Fig. 4.1. The key characteristics are that of a diphthong showing the transition between two phonemes.

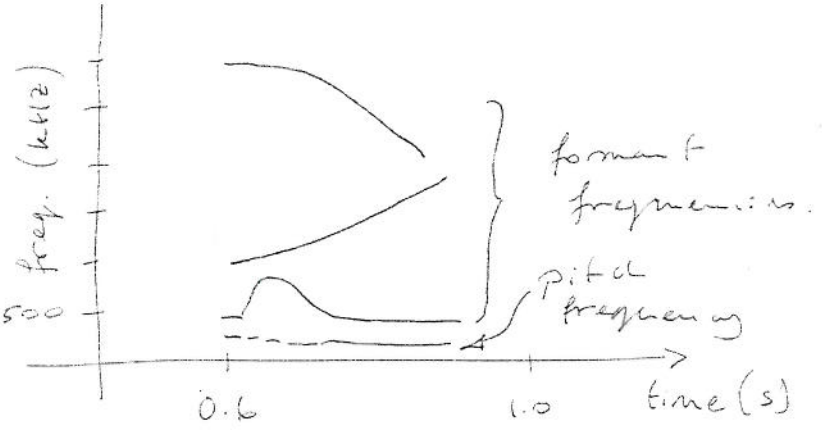

Figure 4.1

b) i) For constant cross-sectional area we obtain

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial (u)}{\partial t}$$
$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial (p)}{\partial t}.$$

ii) This is shown by substitution of

$$u(x,t) = u^+(t-x/c) - u^-(t+x/c)$$
$$\frac{\partial u}{\partial t} = u'(t-x/c) - u'(t+x/c)$$
$$p(x,t) = \frac{\rho c}{A}\left(u^+(t-x/c) + u^-(t+x/c)\right)$$
$$\frac{\partial p}{\partial x} = \frac{\rho c}{A}\left(-\frac{1}{c}u'(t-x/c) + \frac{1}{c}u'(t+x/c)\right)$$

into the solution to part (i).

iii)

$$u(0,t) = U_G(\Omega)e^{j\Omega t} = K^+e^{j\Omega t} - K^-e^{j\Omega t}$$
$$p(l,t) = 0 = \frac{\rho c}{A}\left(K^+e^{j\Omega(t-l/c)} + K^-e^{j\Omega(t+l/c)}\right)$$

so that

$$K^+ = U_G(\Omega)\frac{e^{2j\Omega l/c}}{1+e^{2j\Omega l/c}}$$
$$K^- = -\frac{U_G(\Omega)}{1+e^{2j\Omega l/c}}$$

and finally

$$u(x,t) = \left(\frac{e^{j\Omega(2l-x)/c}+e^{j\Omega x/c}}{1+e^{j\Omega 2l/c}}\right)U_G(\Omega)e^{j\Omega t} = \frac{\cos(\Omega(l-x)/c)}{\cos(\Omega l/c)}U_G(\Omega)e^{j\Omega t}$$
$$p(x,t) = \frac{\rho c}{A}\left(\frac{e^{j\Omega(2l-x)/c}-e^{j\Omega x/c}}{1+e^{2j\Omega l/c}}\right)U_G(\Omega)e^{j\Omega t} = j\frac{\rho c}{A}\frac{\sin(\Omega(l-x)/c)}{\cos(\Omega l/c)}U_G(\Omega)e^{j\Omega t}.$$

iv) From the boundary condition, we have $U(0,\Omega) = U_G(\Omega)$. For the value at the lips, we have can obtain

$$u(l,t) = \frac{1}{\cos(\Omega l/c)}U_G(\Omega)e^{j\Omega t} = U(l,\Omega)e^{j\Omega t}$$

so that the transfer function is given by

$$V(j\Omega) = \frac{1}{\cos(\Omega l/c)}.$$

The sketch is shown in Fig. 4.2.

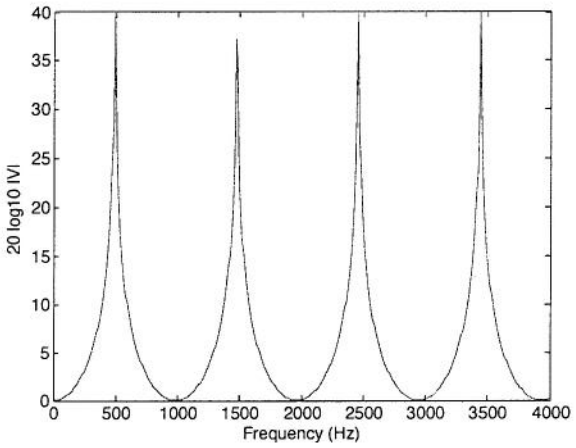
40
35
30
25
20
15
10
5
0
20 log10 |V|
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency (Hz)

Figure 4.2