Final

E4.14
SO16

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2010

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Thursday, 29 April 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) :	P.L. Dragotti

SPEECH PROCESSING

1. Consider a signal $s(n)$ for $n = 0,\ 1,\ \ldots,\ N-1$. In order to predict the samples of this signal, a p^{th} order prediction filter is employed for which the prediction error is given by

$$\varepsilon = \sum_{n=0}^{N-1} e^2(n)$$

where $e(n)$ is the difference between the speech signal at sample instant n and the corresponding prediction. Assume that the prediction filter is stable and its coefficients are real.

a) Write an expression for ε in terms of $s(n)$ and the coefficients of the prediction filter. [4]

b) Show that the prediction error is minimized when the coefficients

$$\mathbf{a} = (a_1,\ a_2,\ \ldots,\ a_p)^T$$

of the prediction filter satisfy the equation

$$\Phi\mathbf{a} = \mathbf{c}.$$

State definitions of Φ and $\mathbf{c}$. [8]

c) Now consider the signal $s(n)$ given by

$$s(n) = \cos(2\pi f n) + v(n)$$

where $v(n)$ represents white noise with zero mean and

$$E\left(v(n)v(n+k)\right) = \begin{cases} \sigma^2 & \text{when } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

and E is the expectation operator.

Determine expressions for the elements of Φ and $\mathbf{c}$ for the case of $p = 2$ under the assumption that N is very large and that the summations involved can be replaced by expected values. [8]

2. Consider a discrete-time speech signal $s(n)$ corrupted by additive noise $v(n)$. The noisy speech signal can be written in the time domain as $y(n) = s(n) + v(n)$ and in the frequency domain as $Y(l,k) = S(l,k) + V(l,k)$ for time-frame index l and frequency index k.

a) i) Explain what is meant by spectral variance of a signal in this context.

ii) Write down the mathematical definition for the spectral variance of $y(n)$, denoted $\phi_y(l,k)$.

[5]

b) Assume that it is only possible to observe $y(n)$.

i) One method to compute an estimate of the spectral variance of the noise, $\hat{\phi}_v(l,k)$, employs a Voice Activity Detector (VAD). Explain this method in detail. In addition, state and explain in detail one other method.

Include in your answer a description of each technique and give mathematical expressions for the estimators.

ii) Compare and contrast the advantages and disadvantages of these alternative methods.

[8]

c) Consider a device which performs noise reduction on a noisy speech signal. The device uses the method of amplitude spectral subtraction.

i) Show that the output signal of the device at time-frame l and frequency index k can be written

$$Z(l,k) = H(l,k)Y(l,k)$$

and write down an expression for $H(l,k)$ in terms of $\hat{\phi}_v(l,k)$. [3]

ii) It is not known what method the device uses to form the estimate $\hat{\phi}_v(l,k)$. Design a test signal to probe the device such that, by observing the output of the device in response to your test signal input, the method of estimation of $\hat{\phi}_v(l,k)$ can be deduced. [4]

3. a) Draw a fully labelled sketch of a p^{th} order lossless tube model of the vocal tract in the human speech production system. Write in one paragraph an explanation highlighting the key characteristics of the lossless tube model. [5]

b) Consider the junction between two sections of the lossless tube model with cross-sectional areas A and B respectively. Sketch an illustrative diagram and derive an expression in matrix form relating the forward and reverse waves in the two tube sections.

State the definition of the reflection coefficients in terms of the cross-sectional area of the tubes in this model. Write down and justify an appropriate value for the reflection coefficient at the input of the lossless tube model. [7]

c) Derive the transfer function of the p^{th} order model in terms of the reflection coefficients and show that the transfer function is all-pole. [5]

d) Sketch a signal flow graph for a complete lossless tube model employing 2 tube sections. The signal flow graph should contain delay elements, multipliers and addition nodes. [3]

4. Consider a hidden Markov model speech recognition system in which the number of states employed in each model is S. During a recognition test, a sequence of T speech frames is observed $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. The probability density of the hidden Markov model generating an observation frame $\mathbf{x}$ from state s is $d_s(\mathbf{x})$ and the transition probability from state i to state j is a_{ij}.

a) Let $B(t,s)$ be the highest probability density of generating the sequence

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$$

from any sequence of states for which frame 1 is in state 1 and frame t is in state s.

When $t > 1$, explain how $B(t,s)$ can be expressed in terms of $B(t-1,i)$ for $i = 1, 2, \ldots, S$ and, for each of these values of i, state what values should be assigned to $B(1,i)$. [5]

b) A 3-state hidden Markov model is shown in Fig. 4.1. For each state, the only non-zero transition probabilities are labelled on the arrows. The feature vector employed in this example is a special case containing only one element. The probability density in frame s is given by

$$d_s(x) = \frac{\exp(-|x - m_s|/k_s)}{2k_s}$$

where m_s and k_s are the state-dependent parameters shown below each state in Fig. 4.1.

For the observation sequence

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5\} = \{1, 1.5, 2.5, 1, 1\}$$

determine the value of $B(5,3)$ and the state sequence to which it corresponds. You should perform all your calculations to 5 decimal places.

[15]

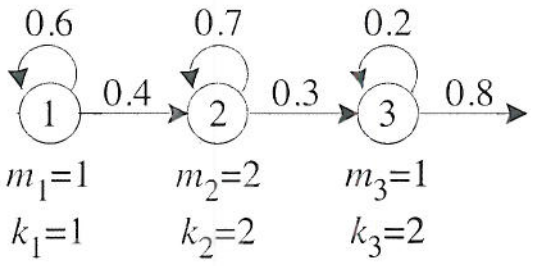

Figure 4.1

E4.14
S016

SPEECH PROCESSING – Solutions 2010

1. a)

$$\varepsilon = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} \left(s(n) - \sum_{j=1}^{p} a_j s(n-j) \right)^2$$

b) The minimum sum squared error is obtained when $\frac{\partial E}{\partial a_i} = 0$ for $i = 1, 2, \ldots, p$.

We obtain

$$\frac{\partial E}{\partial a_i} = \sum_{n=0}^{N-1} \frac{\partial (e^2(n))}{\partial a_i} = \sum_{n=0}^{N-1} 2e(n) \frac{\partial e(n)}{\partial a_i} = -2 \sum_{n=0}^{N-1} e(n)s(n-i).$$

Then for each i we write

$$\begin{aligned}
&\sum_{n=0}^{N-1} e(n)s(n-i) = 0 \\
&\sum_{n=0}^{N-1} \left(s(n)s(n-i) - \sum_{j=1}^{p} a_j s(n-j)s(n-i) \right) = 0 \\
&\sum_{n=0}^{N-1} \sum_{j=0}^{p} a_j \sum_{n=0}^{N-1} s(n-j)s(n-i) = \sum_{n=0}^{N-1} s(n)s(n-i) \\
&\sum_{j=1}^{p} \phi_{ij} a_j = \phi_{i0}
\end{aligned}$$

where $\phi_{ij} = \sum_{n=0}^{N-1} s(n-j)s(n-i)$. The vector $\mathbf{c}$ is defined as $c_i = \phi_{i0}$.

c)

$$\begin{aligned}
\phi_{ij} &= E\left(s(n-i)s(n-j)\right) \\
&= E\left(\cos(2\pi f(n-i))\cos(2\pi f(n-j))\right)\sigma^2\delta_{ij} \\
&= \frac{1}{2}E\left(\cos(2\pi f(2n-i-j))\cos(2\pi f(i-j))\right)\sigma^2\delta_{ij} \\
&= \frac{1}{2}\cos(2\pi f(j-i))\sigma^2\delta_{ij} \\
c_i = \phi_{i0} &= \frac{1}{2}cos(2\pi fi).
\end{aligned}$$

2. a) i) For any signal $x(n)$ with time-frequency domain representation $X(l,k)$, the spectral variance is a measure of the signal power at time-frame l and frequency k.

ii)
$$\phi_y(l,k) = E\left[|Y(l,k)|^2\right]$$

b) Voice activity detection (VAD): determine the segments of the signal during which only noise is present. Then update the noise model according to:

$$\hat{\phi}_v(l,k) = \begin{cases} \hat{\phi}_v(l-1,k), & \text{if speech is active} \\ \alpha\hat{\phi}_v(l-1,k) + (1-\alpha)|Y(l,k)|^2, & \text{if speech is not active.} \end{cases}$$

The term α has to be chosen according to the desired level of smoothing and can be time and frequency varying.

Minimum statistics approach: This technique is based on the assumption that during a speech pause, or within brief periods between words and even syllables, the speech energy is close to zero. As a result, a short-term power spectrum estimate of the noisy signal, even during speech activity, decays frequently due to the noise power. Thus, by tracking the temporal spectral minimum without distinguishing between speech presence and speech absence, the noise power in a specific frequency band can be estimated.

$$\hat{\phi}_y(l,k) = \alpha\hat{\phi}_y(l-1,k) + (1-\alpha)|Y(l,k)|^2$$
$$\hat{\phi}_v(l,k) = \min\left\{\hat{\phi}_y(l,k),\ \hat{\phi}_y(l-1,k),\ \ldots,\ \hat{\phi}_y(l-D+1,k)\right\}$$

c) i) For amplitude spectral subtraction, we have

$$\begin{aligned} Z(l,k) &= Y(l,k) - \sqrt{\hat{\phi}_v(l,k)} \\ &= Y(l,k)\left(1 - \frac{\sqrt{\hat{\phi}_v(l,k)}}{Y(l,k)}\right) \\ &= Y(l,k)H(l,k) \end{aligned}$$

ii) Here we are aiming to determine whether or not the noise model is updated continuously or only during non-speech periods. We can consider a test signal containing two speech utterances separated in time by a short pause. To this, we can add the 'noise' of a swept sine wave. If the noise model is updated only during speech pauses, the sine wave will not be significantly suppressed during the speech utterances but will be strongly suppressed during and shortly after the pause. If the noise model is updated continuously then the sine wave will be suppressed throughout. The former case indicates VAD-based noise modelling whereas the later indicates a minimum statistics approach.

[Full marks will be awarded for well-reasoned responses].

3. a) The lossless tube model represents the vocal tract using a concatenation of lossless tube sections each of constant cross-sectional area. The length of each section corresponds to the distance travelled by sound in one half sampling period. The number of sections corresponds to the order of the model. Within each section, we consider a forward wave representing the flow of air from left to right as well as a reverse wave representing the flow of air from right to left. The flows are modelled in terms of the volume velocity. At each section junction, reflections occur according to the nature of the change of cross-sectional area.

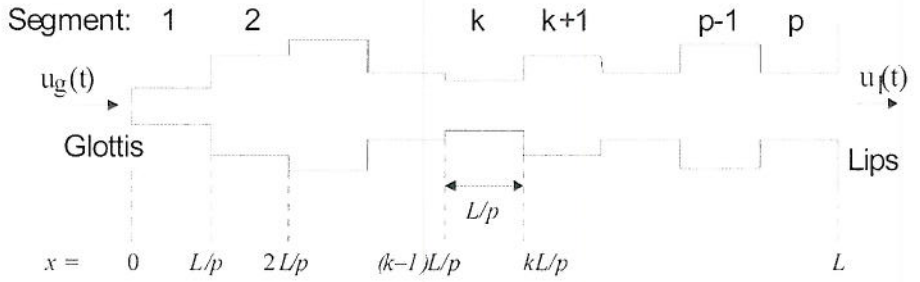

Figure 3.1

b) The interface between two sections of the lossless tube model can be illustrated as shown in Fig. 3.2. Let the forward waves be U and W and let the reverse

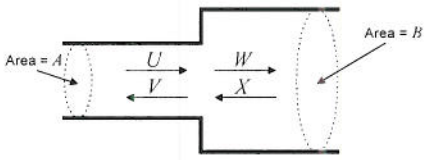

Figure 3.2

waves be V and X. Then from flow continuity we have

$$(U-V)=(W-X)$$

and from pressure continuity we have

$$\frac{\rho c}{A}(U+V)=\frac{\rho c}{B}(W+X).$$

Combining these in matrix form leads to

$$\begin{pmatrix} 1 & -1 \\ B & B \end{pmatrix}\begin{pmatrix} U \\ V \end{pmatrix}=\begin{pmatrix} 1 & -1 \\ A & A \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$

The reflection coefficients are written

$$r=\frac{B-A}{B+A}$$

Reflection coefficient at the input (glottis): $r_0 = 1$ when the glottis is closed, or for an assumption of zero input.

c) We obtain

$$\begin{pmatrix} U_g \\ V_g \end{pmatrix} = \frac{z^{1/2\,p}}{\prod_{k=0}^{p-1}(1+r_k)} \prod_{k=0}^{p-1} \begin{pmatrix} 1 & -r_k z^{-1} \\ -r_k & z^{-1} \end{pmatrix} \times \begin{pmatrix} 1 \\ -r_p \end{pmatrix} U_l$$

which results in a transfer function of the form

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

This can be seen to contain only zeros, ignoring the delay.

d) The signal flow graph for a single interface is shown in Fig. 3.3. For a model comprising two sections, two of these elements will be connected in cascade.

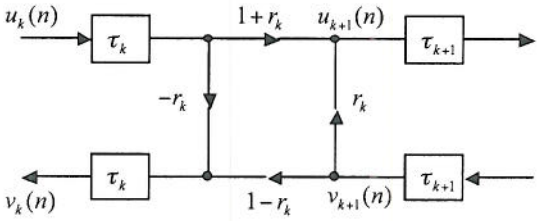

Figure 3.3

4. a) The best path for $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ with t in state s must have frame $t-1$ in one of the other states i. Since the alignment $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}$ must also be an optimal alignment

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

Since $B(t,s)$ represents the probability density of the best path, we can therefore write

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

Frame 1 is always aligned to state 1 and therefore

$$B(1,s) = \begin{cases} d_1(\mathbf{x}_1) & \text{if } s = 1 \\ 0 & \text{otherwise.} \end{cases}$$

b) The possible paths through the lattice alignment of frames to states is shown in Fig. 4.1.

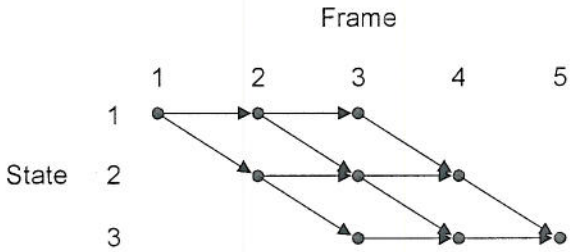

Figure 4.1

$B(1,1) = \underline{0.5}$

$B(2,1) = 0.5 \times 0.6 \times 0.30327 = \underline{0.09098}$

$B(2,2) = 0.5 \times 0.4 \times 0.19470 = 0.03894$

$B(3,1) = 0.09098 \times 0.6 \times 0.11157 = 0.00609$

$B(3,2) = \max(\underline{0.09098 \times 0.4 = 0.03639}, 0.03894 \times 0.7 = 0.02726) * 0.19470 = \underline{0.00709}$

$B(3,3) = 0.03894 \times 0.3 \times 0.11809 = 0.00138$

$B(4,2) = \max(0.00609 \times 0.4 = 0.00243, \underline{0.00709 \times 0.7 = 0.00496}) \times 0.15163 = \underline{0.00075}$

$B(4,3) = \max(0.00709 \times 0.3 = 0.002126, 0.00138 \times 0.2 = 0.000276) \times 0.25 = 0.00053$

$B(5,3) = \max(\underline{0.00075 \times 0.3 = 0.00023}, 0.00053 \times 0.2 = 0.00011) \times 0.25 = \underline{0.00006}$

The aligned path is therefore 1, 1, 2, 2, 3 by tracing back from $B(5,3)$.