E4.14
SO16
ISE4.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2003

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

**SPEECH PROCESSING**

Monday, 19 May 10:00 am

Time allowed: 3:00 hours

**There are SIX questions on this paper.**

**Answer FOUR questions.**

**Corrected Copy**

**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

1. (a) Describe the source-filter model of speech production. Give an illustrative block diagram. [4]

(b) The filter, $V(z)$, in the source-filter model can be characterised by several alternative equivalent sets of parameters. State the properties of such parameter sets that would be desirable for use in: [4]
(i) speech recognition,
(ii) speech coding.

(c) For a second order vocal tract filter defined by $V(z) = \dfrac{1}{1 - a_1 z^{-1} - a_2 z^{-2}}$ with coefficients [6]
$a_1 = 0.98$ and $a_2 = -(0.98)^2$, calculate the frequency of the formant $f_1$. Also calculate the gain of the vocal tract filter at this frequency $f_1$.

Recalculate the gain of the filter at frequency $f_1$ after a 1% increase in $a_2$, keeping $a_1$ fixed.

(d) Consider the hypothesis that the set of reflection coefficients would be more robust to small errors in the coefficient values than the set of prediction coefficients. Test this hypothesis for a 1% change in one of the reflection coefficients while keeping the other fixed for the initial vocal tract filter, $V(z)$, given in part (c). [5]
(You may assume that reflection coefficient $r_0 = 1$.)

Comment on these results in relation to the choice of parameters used for coding speech. [1]

2. Consider a Hidden Markov model having $S$ states. Consider also $T$ feature vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_T$, extracted from non-overlapping frames of a speech signal. The frame duration is 10 ms. Let $a_{ij}$ represent the transition probability from state $i$ to state $j$ of the model and let $d_i(\mathbf{x}_t)$ represent the output probability of frame $t$ in state $i$.

Let $P(s,t)$ be defined as the total probability density that the model generates $\mathbf{x}_1, \ldots, \mathbf{x}_t$ summed over all alignments having frame $t$ in state $s$. Let $Q(s,t)$ be defined as the total probability density that the model generates $\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T$ summed over all alignments having frame $t$ in state $s$.

(a) With the aid of a suitable diagram, explain why $P(s,t)$ depends on no other value of $P$ except $P(\{s\}, t-1)$, where $s$ is a particular state and $\{s\}$ is the set of states. [3]

State recursive expressions for $P(s,t)$ and $Q(s,t)$. [4]

Hence write out the recursive algorithm for calculating $P(s,t)$ including any initialization. You may use any reasonable notation, MATLAB code or pseudo code provided that the meaning of each statement is clear and unambiguous. [3]

(b) Feature vectors from 5 frames of speech are used to train a 3-state Hidden Markov model with the states referred to as States 1, 2 and 3.

Table 2.1 shows the output probability of each frame from each state of the model and Figure 2.1 shows the state diagram of the model including the transition probabilities.

For each of $i = 1$, 2 and 3, calculate the total probability that frame $\mathbf{x}_4$ corresponds to state $i$ and that the model generates $\mathbf{x}_1, \ldots, \mathbf{x}_5$. [10]

| | Frame | | | | |
|---|---|---|---|---|---|
| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\mathbf{x}_5$ |
| State 1 | 0.5 | 0.1 | 0.8 | 0.6 | 0.7 |
| State 2 | 0.4 | 0.4 | 0.2 | 0.3 | 0.6 |
| State 3 | 0.3 | 0.2 | 0.5 | 0.2 | 0.5 |

Table 2.1

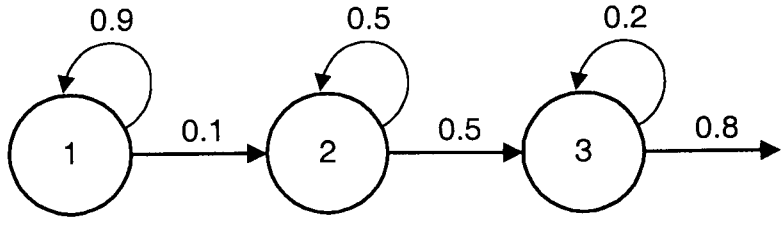


Figure 2.1

3. (a) Consider a continuous density, left-to-right, no-skips, hidden Markov model (HMM) of an isolated word for speech recognition.

[4]

(i) State the main characteristics of a Markov model and explain why it is called hidden.
(ii) Explain the meaning of left-to-right, no-skips.
(iii) Explain the meaning of continuous density.

(b) For a particular state, $i$, in an HMM of the type of part (a), the probability of a transition from state $i$ to the next state is $p$. Derive an expression for the expected value of $D$, the number of frames in state $i$. Include all relevant working. [4]

(c) Define Mel-frequency cepstral coefficients and give a detailed description of how these coefficients are computed using appropriate illustrations. [4]

Using the definition of Mel in terms of frequency $f$ Hz as [4]

$$\mathrm{Mel(f)} = 2595 \log_{10}(1 + \mathrm{f} \,/\, 700)$$

draw a labelled sketch of the magnitude response as a function of frequency in Hz of each channel of a Mel-frequency filterbank having 4 bands uniformly spaced in Mel and covering the range 0 to 4000 Hz.

(d) State the desirable characteristics of feature vectors to be used for speech recognition and explain the extent to which these characteristics are found in Mel-frequency cepstral coefficients. [4]

4. (a) Using not more than two sides of paper, give a descriptive overview of text-to-speech synthesis. Include a clear statement of the steps involved, an explanation of each step and indicate the important issues that must be considered at each step. [8]

(b) In the text-to-speech task, converting words into phoneme sequences requires the use of a combination of the following three elements: [6]

- a word dictionary
- a morph dictionary
- a set of letter-to-sound rules.

Explain the role and importance of each of element.

(c) Figure 4.1 shows a simplified state diagram for the sequences of morphs that can be concatenated to form a word. Each state is labelled with a morph category and the transitions are labelled with an applicable "cost". Determine all possible decompositions of the word "uninformed" and the total cost of each decomposition. The dictionary of available morphs is given in Table 4.1. [6]

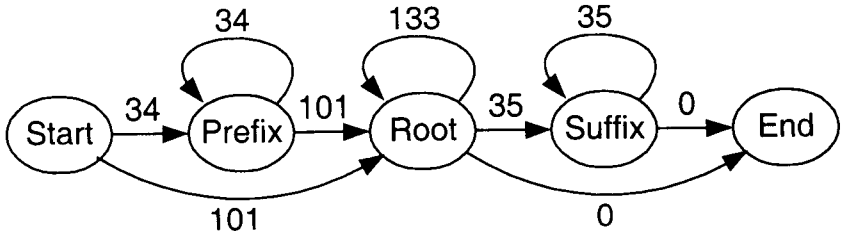


Figure 4.1

| Category | Morphs |
|---|---|
| Prefix | in, un |
| Root | in, inform, form, formed |
| Suffix | ed |

Table 4.1

5. (a) Derive the Yule-Walker equations for covariance LPC as the solution to a minimisation problem and give an expression for the elements of the covariance matrix $\Phi$ that they contain. State any symmetry properties of $\Phi$. [5]

Describe how covariance LPC can be used to analyse the formants of a speech signal of several seconds duration. [4]

(b) For frequency $f = 0.1$ and with $v(n)$ representing white noise with zero mean and variance $\sigma^2$, let the discrete-time signal $x(n)$ be defined as

$$x(n) = \sin(2\pi fn) + v(n).$$

The frequency of $x(n)$ can be detected by performing 2nd order covariance LPC analysis and then determining the frequency at which the resulting poles are located. However, the effect of additive noise $v(n)$ may degrade the accuracy of this frequency estimation.

To determine the magnitude of any such degradation, carry out the following steps using 2nd order covariance LPC analysis and then compare the estimated frequency with the true value of $f = 0.1$ :

(i) Determine the elements of $\Phi$ under the assumption that the summations involved may be replaced by expected values. [5]

(ii) Determine the LPC predictor coefficients and hence the estimated value of $f$ for the case $\sigma^2 = 0.1$. State the percentage error in the frequency estimate. [6]

6. Describe and illustrate the lossless tube model of the vocal tract. State any assumptions implicit in the model and typical values for the length of each segment and the number of segments for a sampling frequency of 11 kHz. [6]

State the features of a real vocal tract that are omitted from the lossless tube model. [2]

Consider two adjacent segments of the lossless tube model, Section 1 and Section 2 with cross-sectional areas $A_1$ and $A_2$ respectively as shown in Figure 6.1. The terms $U, V, W$ and $X$ represent the volume flow rates of the acoustic waves travelling in the directions indicated by arrows. The acoustic pressure associated with $U$ is given by

$$\rho c \frac{U}{A_1}$$

where $\rho$ is the density of air and $c$ is the speed of sound.

Derive a matrix equation expressing $\begin{bmatrix} V \\ W \end{bmatrix}$ in terms of $\begin{bmatrix} U \\ X \end{bmatrix}$. [4]

Explain what is meant by the reflection coefficient and give an expression for the reflection coefficient in terms of $A_1$ and $A_2$. [2]

In a simplified case, a particular frame of a speech signal is found to be sufficiently well modelled by a lossless tube model with 2 sections and with the transfer function of the corresponding vocal tract filter, $V(z)$, given by

$$V(z) = \frac{6z^{-1}}{1 - 0.8z^{-1} + 0.8z^{-2}}.$$

Draw a labelled sketch of the lossless tube model showing how the cross-sectional area varies along the length of the vocal tract, as described by $V(z)$. You may assume that the glottis is closed. [6]

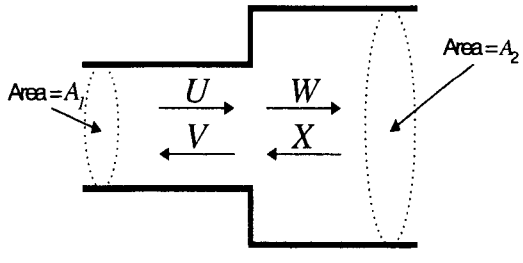


Figure 6.1