EE4-14
EE9-SO16

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2015

MSc and EEE/EIE PART IV: MEng and ACGI

SPEECH PROCESSING

Corrected Copy

Thursday, 14 May 10:00 am

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) Describe the processing steps involved in a text-to-speech synthesis system. [7]

b) Now consider synthesis of speech signals.

i) Explain the relative merits and differences between a cascade formant synthesiser and a parallel formant synthesiser. [4]

ii) The structure of a parallel formant speech synthesiser is shown in Figure 1.1. Describe the function of each block in the diagram. [5]

iii) Explain why the outputs from $F1(z)$, $F2(z)$, and $F3(z)$ are added with alternating positive and negative signs and illustrate your answer with a typical vowel output spectrum. [2]

iv) Explain and justify how the values of $v1$, $v2$, and $v3$ would differ for the phonemes /a/, /s/ and /z/. [2]

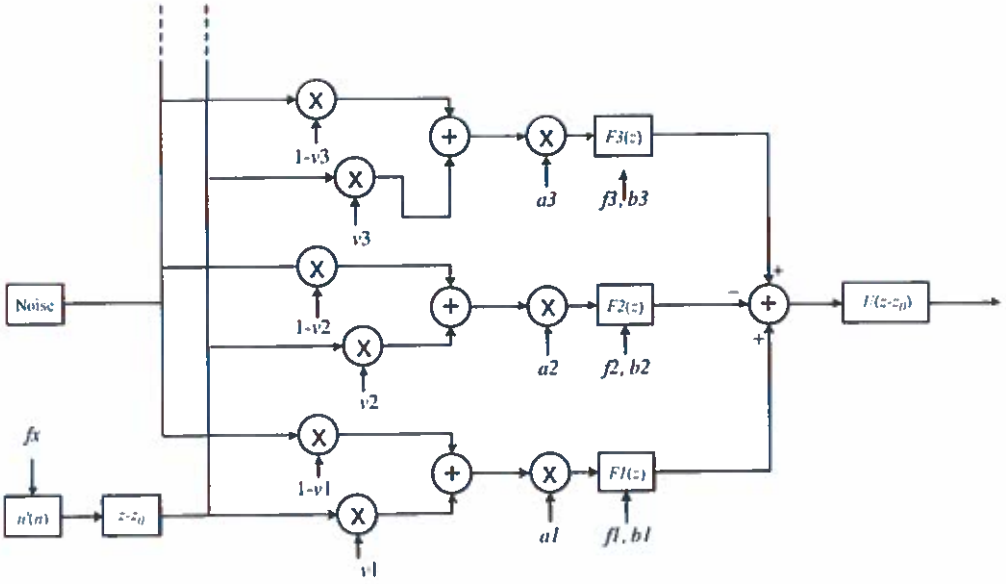

Figure 1.1 Parallel formant synthesiser

2. a) A signal $s(n)$ contains speech together with near-stationary background traffic noise giving an SNR of 25 dB. The speech component varies over time between voiced speech and unvoiced speech. Describe with the aid of appropriate diagrams the key characteristics that could be used to identify the time segments of $s(n)$ containing:

- only voiced speech with background noise;
- only unvoiced speech with background noise;
- only background noise. [6]

b) Code Excited Linear Prediction (CELP) is an established method for speech coding. Draw and label a detailed block diagram of a CELP encoder. With reference to your block diagram, describe briefly how voiced speech and unvoiced speech are coded by a CELP encoder. [6]

c) State the relevant properties of an algebraic codebook and explain the advantages and disadvantages of using an algebraic codebook as an excitation signal in CELP. [3]

d) For use in CELP speech coding, an adaptive postfilter is proposed of the form

$$H_p(z) = (1 - \mu z^{-1}) \frac{1 - \sum_{k=1}^{p} \gamma_1^k a_k z^{-k}}{1 - \sum_{k=1}^{p} \gamma_2^k a_k z^{-k}}$$

in which a_k are predictor coefficients. Deduce the meaning of the parameters of $H_p(z)$ and hence explain the function of this filter. Include relevant illustrative diagrams and/or plots in your explanation. [5]

3. a) Consider a 3-state Hidden Markov Model with S states that is to be used for speech recognition. Feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ are extracted from the speech signal every 20 ms. The model is trained using 5 frames from a speech utterance. Table 1 shows the output probability of each frame from each state of the model. The probabilities of transitions between states and to the same state are shown in the state diagram of the model in Figure 3.1. 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 $d_i(\mathbf{x}_t)$.

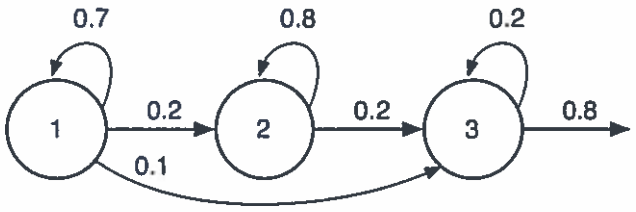

Figure 3.1 State diagram

i) Discuss the factors that should be taken into account when choosing the frame rate. [2]

ii) Sketch the alignment lattice showing all allowable transitions between states. [4]

iii) For an alignment in which frame t is aligned to state s, the total probability that the model generates the frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ is $P(s,t)$ and the total probability that the model generates the frames $\mathbf{x}_{t+1}, \mathbf{x}_2, \ldots, \mathbf{x}_T$ is $Q(s,t)$. Derive recursive expressions for $P(s,t)$ and $Q(s,t)$ in terms of $P(i,t-1)$ and $Q(i,t+1)$ respectively, where $i = 1, 2, \ldots, S$ and include values for $P(s,1)$ and $Q(s,T)$. [4]

iv) Calculate the total probability that frame $\mathbf{x}_2$ corresponds to state s for $s = 1$, 2, 3 and that the model generates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5$. Show the relevant probability calculations for $P(s,t)$ and $Q(s,t)$. Use 6 decimal places for calculations. [7]

v) Explain the meaning of *language modelling* in automatic speech recognition. Describe the main principles of operation of such a language model and include any relevant diagrams. [3]

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
s_1	0.5	0.4	0.3	0.1	0.5
s_2	0.3	0.1	0.7	0.2	0.2
s_3	0.2	0.4	0.5	0.4	0.5

Table 1 Output probabilities

4. Consider linear prediction of a speech signal $s(n)$ using LPC of order p and prediction coefficients a_k for $k = 1, \ldots, p$ and an LPC model system given by

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}.$$

a) Consider the properties of speech signals.

i) What properties of speech signals result in LPC being a commonly chosen basis for methods of encoding speech. [2]

ii) Briefly describe the characteristics of the prediction residual in the case when the prediction is ideal for the cases of both voiced and unvoiced speech. [2]

iii) Summarise the method of prediction of pitch periodicity in LPC-based speech coding methods. [2]

b) Now consider a particular frame containing a segment of the speech signal $s(n)$ for which we define

$$\varphi(i,k) = \sum_{m} s(m-i)s(m-k).$$

i) Write an expression for the squared prediction error ε in this frame in terms of a_k and s. [2]

ii) Show how the prediction coefficients can be chosen to minimize ε and write an expression for ε in terms of φ and a_k. [4]

c) Consider the case for a frame of L samples in which $s(m)$ is zero except within the interval $0 \leq m \leq L-1$. For the range $0 \leq m \leq L-1+p$, state which samples indices could be expected to have large prediction error and explain why. [4]

d) Now consider a particular frame of speech for which the first 3 values of the autocorrelation function are 51, 45, 29.

Determine the optimal linear prediction coefficients in $H(z)$ for 2nd order LPC and find the corresponding minimum squared error. [4]