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

Solution:
Steps in text-to-speech:

- Convert text to phonemes:
  - Check dictionary
  - try removing prefixes and suffixes
  - check dictionary, else apply pronunciation rules
  - add back prefix/suffix sounds including coarticulation
- Sort out prosodic variations of stress, pitch and duration
- Convert phonemes to synthesiser control parameters
  - split some phonemes up into sub-phonemes (e.g. diphthongs)
  - look up targets for formant frequencies and amplitude
  - look up rules for interpolating formant trajectories
- Generate acoustic signal
- Use a parallel or cascade formant synthesiser.
- Choose voicing mixture for each phoneme.

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]

Solution:
The cascade synthesiser contains a number of 2nd order resonant sections in series whereas the parallel synthesiser has them in parallel. The main difference between the two is that the parallel version has two separate input parameters for amplitude and bandwidth whereas the cascade version has only a single parameter. Because of this, the formant bandwidths in a cascade synthesiser are incorrect for consonants in which the point of acoustic excitation is partway along the vocal tract.

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]

Solution:
The noise block generates white noise for unvoiced sounds.
The $u'(n)$ block generates a periodic waveform with frequency fx approximating the time-derivative of volume flow through the glottis for voiced sounds. This is filtered by $(z - z0)$ to compensate for the output filter.
Each formant block mixes the two excitations according to the v_k parameter, multiplies by a gain, a_k , and passes the signal through a pole-pair to represent the formant resonance. The outputs from the formants are then added with alternating signs. The output filter is a low-pass filter with a cut-off of around 640 Hz to give the correct overall spectral shape for unvoiced sounds.

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]

Solution:
This arrangement involving alternating signs in the addition unit after the formant resonators avoids a sharp null in the spectrum at frequencies between adjacent formants with similar amplitudes due to phase cancellation.

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

Solution:
In the case of the vowel /ɑ/ all formants would be fully voiced such that $v_k = 1$. For the unvoiced fricative /s/ they would be unvoiced with $v_k = 0$. For the voiced fricative /z/ the low formants would be fully voiced but the voicing would be progressively reduced for higher formants as k increases.

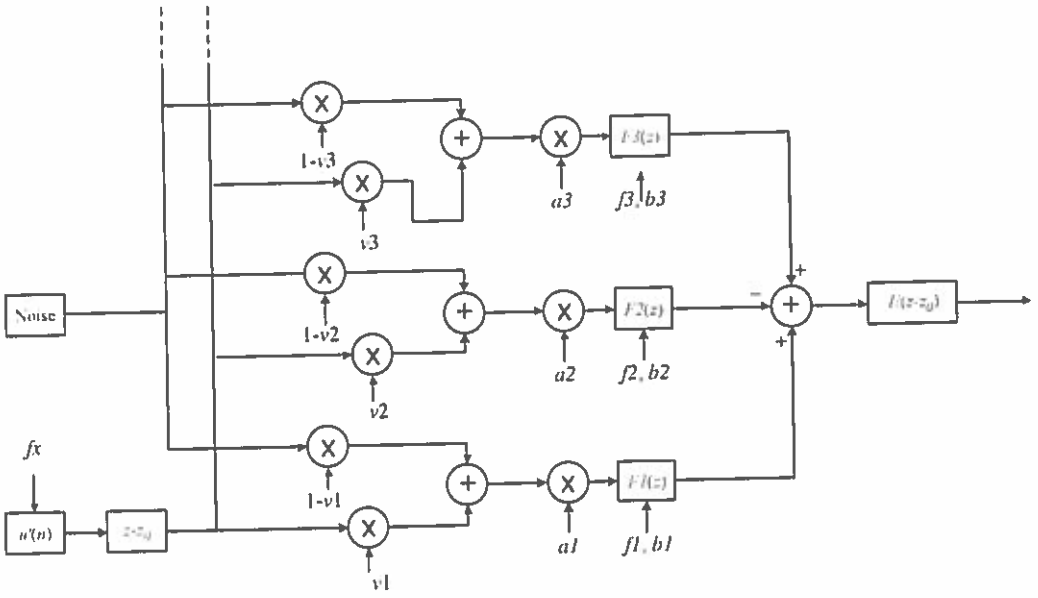


Figure 1.1 Parallel formant synthesiser

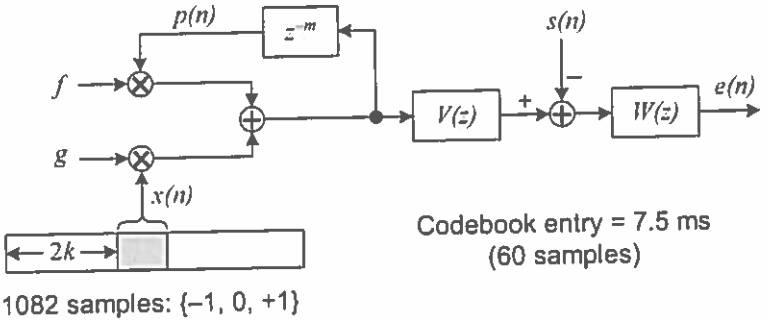
2. a) A signal $s(n)$ contains speech together with near-stationary background traffic noise giving an SNR of 25 dB. The speech component varies over time between voiced speech and unvoiced speech. Describe with the aid of appropriate diagrams the key characteristics that could be used to identify the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

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

Solution:
Features such as pitch, and format frequencies and bandwidths (or related spectral properties) are a potential basis for such a separation to be performed. Other well reasoned proposals will also get full credit. Separation of the unvoiced speech from the noise is the more ambiguous case and well reasoned consideration of this case is necessary to achieve full marks.

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]

Solution
Block diagram



Both voiced and unvoiced speech employ a vocal tract filter $V(z)$. For voiced speech, the excitation is provided (mainly) by the adaptive codebook (long-term predictor) while for unvoiced speech the fixed (algebraic) codebook provides the excitation. The gain factors f and g control the mix.

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]

Solution
An algebraic codebook is understood to contain sample elements with values $-1, 0, +1$ only with typically $p(-1) = p(+1) = 0.1$ and $p(0) = 0.9$. This gives rise to a sparse set of pulses for the excitation and is used, together with an adaptive codebook element to generate the speech excitation input to the vocal tract filter. The main advantage is low computational complexity during the exhaustive codebook search. The main disadvantages is a reduction in sound quality - this may be mitigated by several post-processing techniques described in the literature.

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]

Solution

Consider the second element first. Some similarity can be seen with the study of bandwidth expansion filters and also of perceptual weighting filters. As stated the a_k are predictor coefficients such that the numerator and denominator both contain polynomials in z that, for $\gamma_1 = \gamma_2$ give unit gain. For $\gamma_1 < \gamma_2$ this fraction becomes a filter with peaks and values in the magnitude spectrum aligned to the peaks and valleys in the speech spectrum (as represented by the predictor polynomial). The effect is to attenuate the spectral components in the valleys of the coded speech magnitude spectrum without significantly effecting the speech. Such a postfilter can reduce the background noise in a CELP coder and improved the naturalness of the decoded speech.

In the first element, we see a simple highpass filtering operation with corner freq determined by μ. This provides a spectral tilt and also improves the operation of the postfilter.

The postfilter is called adaptive because it's freq response varies according to the speech spectrum.

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 $d_i(\mathbf{x}_t)$.

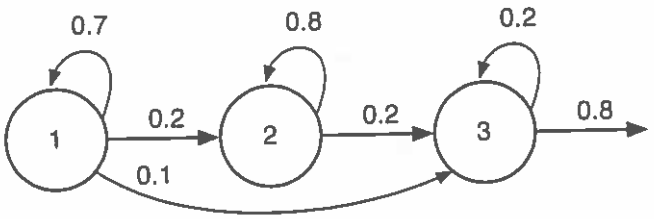


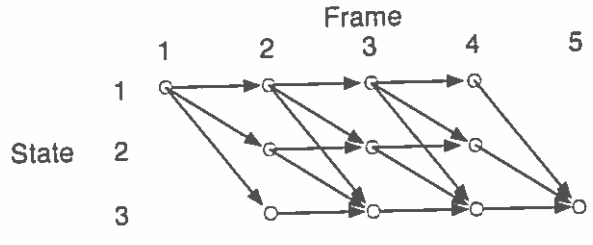
Figure 3.1 State diagram

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

Solution:
The choice of frame rate is a trade-off the needs to target sufficiently long frames that features can be accurately estimated but not so long as to violate the assumption of quasi-stationarity of the speech in each frame.

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

Solution:
The lattice is drawn as:



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]

Solution:

For $P(s,t)$, any alignment of frames $1,\ldots,t$ must allocate frame $t-1$ to one of the states i in the range $1,\ldots,S$. Thus

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

A corresponding formula for Q follows as

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

The initial and final conditions arise from consideration that frame 1 must be in state 1 and frame T must be in state S so that

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

$$Q(s,T) = \begin{cases} a_{s=} & s = S \\ 0 & \textit{otherwise.} \end{cases}$$

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]

Solution:
For the total probability, we require $P(3,2)$ and $Q(3,2)$.

$$P(1,1) = 0.5$$
$$P(1,2) = 0.5 \times 0.7 \times 0.4 = 0.14$$
$$P(2,2) = 0.5 \times 0.2 \times 0.1 = 0.01$$
$$P(3,2) = 0.5 \times 0.1 \times 0.3 = 0.02$$

$$Q(3,5) = 0.8$$
$$Q(1,4) = 0.1 \times 0.5 \times 0.8 = 0.04$$
$$Q(2,4) = 0.2 \times 0.5 \times 0.8 = 0.08$$
$$Q(3,4) = 0.2 \times 0.5 \times 0.8 = 0.08$$
$$Q(1,3) = 0.7 \times 0.1 \times 0.04 + 0.2 \times 0.2 \times 0.08 + 0.1 \times 0.4 \times 0.08 = 0.0092$$
$$Q(2,3) = 0.8 \times 0.2 \times 0.08 + 0.2 \times 0.4 \times 0.08 = 0.0192$$
$$Q(3,3) = 0.2 \times 0.4 \times 0.08 = 0.0064$$
$$Q(1,2) = 0.7 \times 0.3 \times 0.0092 + 0.2 \times 0.7 \times 0.0192 + 0.1 \times 0.5 \times 0.0064 = 0.00494$$
$$Q(2,2) = 0.8 \times 0.7 \times 0.0192 + 0.2 \times 0.5 \times 0.0064 = 0.011392$$
$$Q(3,2) = 0.2 \times 0.5 \times 0.0064 = 0.00064$$

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]

Solution:
Language model is a process by which knowledge of the structure of language can be incorporated into a speech recognizer. Such models may vary in complexity upwards from a simple unigram model. The language model will output the prior probability of each hypothesized word. This is multiplied by the acoustic model product probability to give an overall score.

	x_1	x_2	x_3	x_4	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 ideal case when the prediction is exact 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]

Solution:
Speech formats arise due to acoustic resonances in the vocal tract and these can be modelled using a complex conjugate pole pair per resonance. Resonances, and their associated harmonic structure are predictable with high accuracy using an auto-regressive structure - the key structure of LPC. LPC can be computationally efficient.
If the prediction of $H(z)$is exact, then the prediction residual will be an impulse train for voiced speech and white noise for unvoiced speech.
Notably, LPC does not predict the quasi period nature of pitch periodicity. This is handled separately in speech coders by a long-term predictor (sometimes called the adaptive codebook).

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 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. [4]

Solution

$$\varepsilon = \sum_m e^2(m) = \sum_m \left(s(n) - \sum_{k=1}^{p} a_k s(m-k) \right)^2$$

From the results of setting the partial derivate w.r.t. to $a_k = 0$, the minimum ε is obtained when a_k are chosen as solutions to

$$\sum_m s(m-i)s(m) = \sum_{k=1}^{p} a_k \sum_m s(m-i)s(m-k)$$
$$1 \le i \le p.$$

Next we can write for optimal $a_k = \bar{a}_k$

$$\varepsilon = \sum_m \left(s(n) - \sum_{k=1}^{p} a_k s(m-k) \right)^2$$
$$= \sum_m s^2(m) - \sum_{k=1}^{p} \bar{a}_k \sum_m s(m)s(m-k)$$

and using

$$\varphi(i,k) = \sum_m s(m-i)s(m-k)$$

leads to

$$\varepsilon = \varphi(0,0) - \sum_{k=1}^{p} \bar{a}_k \varphi(0,k).$$

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

Solution
Since we assume $s(m)$ is zero except within the interval $0 \le m \le L-1$, the sample indices $0 \le m \le p-1$ involve predictions of non-zero sample values from a weighted combination of previous sample values, at least some of which are zero. By a similar argument, sample indices $L \le m \le p-1$ involve predictions of zero sample values from a weighted combination of previous sample values, at least some of which are non-zero. In both these ranges we can expect poor prediction accuracy.

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]

Solution

In this case we can use the autocorrelation method for LPC formulated as

$$\begin{bmatrix} R(0) & R(1) \\ R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \end{bmatrix}.$$

The autocorrelation matrix is $\begin{bmatrix} 51 & 45 \\ 45 & 51 \end{bmatrix}$ which has determinant 576 and inverse $\begin{bmatrix} 0.089 & -0.078 \\ -0.078 & 0.089 \end{bmatrix}$.

Hence $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.089 & -0.078 \\ -0.078 & 0.089 \end{bmatrix} \begin{bmatrix} 45 \\ 29 \end{bmatrix} = \begin{bmatrix} 1.743 \\ -0.978 \end{bmatrix}$.

Thus we can write

$$H(z) = \frac{1}{1 - 1.7425z^{-1} + 0.978z^{-2}}.$$

The minimum mean squared prediction error is given by

$$\varepsilon = R(0) - \sum_{k=1}^{p} \tilde{a}_k R(k) = 51 - 1.743 * 45 + 0.978 * 29 = 0.93.$$