EE4-14
EE9SO16

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2012

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Friday, 18 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. The diagrams in Figure 1.1 show signal flow graphs of alternative representations of the lossless tube model of the vocal tract, each using two tube sections. These are denoted structure A and structure B as labelled.

a) Using matrix notation, derive an expression for $\begin{bmatrix} V_1^+ \\ V_1^- \end{bmatrix}$ in terms of $\begin{bmatrix} V_2^+ \\ V_2^- \end{bmatrix}$ and another expression for $\begin{bmatrix} W_1^+ \\ W_1^- \end{bmatrix}$ in terms of $\begin{bmatrix} W_2^+ \\ W_2^- \end{bmatrix}$. [6]

b) Derive the transfer function $\frac{Y(z)}{X(z)}$. Also derive the transfer function $\frac{Q(z)}{P(z)}$. Hence derive the relationship between the two transfer functions. [7]

c) State the number of multiply operations and addition operations needed to compute each output sample for both structure A and also structure B. [2]

d) Consider now the structure B. Copy and complete the following function in MATLAB that computes the output sequence $\{q(n)\}$ given an input sequence $\{p(n)\}$. Use one FOR loop in your code to count through each sample.

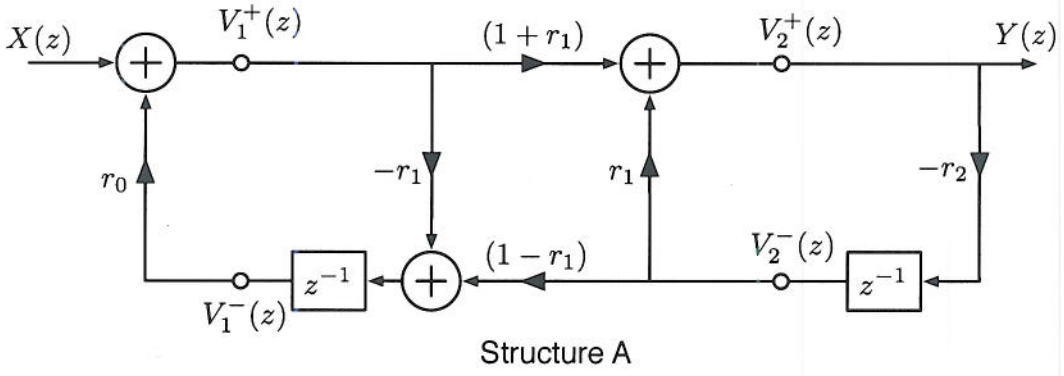
```
function q = losslessTubeB(p)

                    % COMMENT: p is a vector of input samples
                    % and q is a vector of output samples
    ...

for n=1:N           % COMMENT: this loop goes through each
                    % input sample and computes each
                    % corresponding output sample
    ...

end
```

Note that marks will not be deducted for syntax errors in the code providing that the operations are described unambiguously. It is not necessary to show any variable declarations or intialization code providing they are clear from the context. [5]



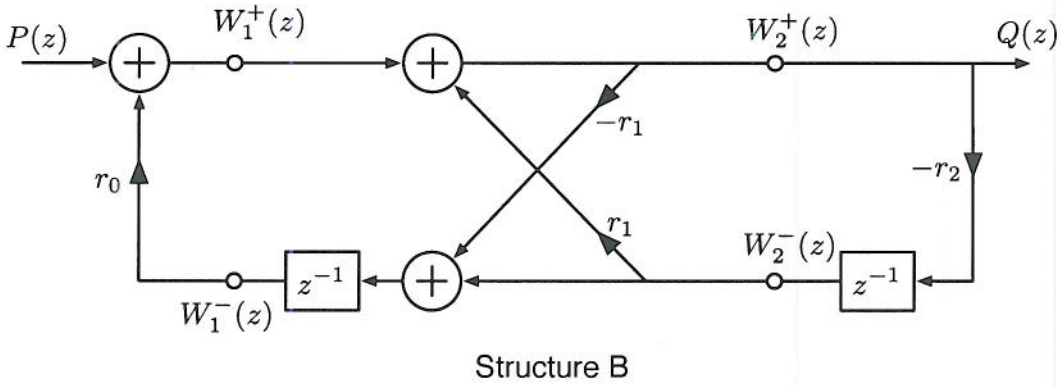


Figure 1.1 Lossless Tube Models

2. Consider a discrete-time speech signal $s(n)$, where n indicates the sample index. The speech signal $s(n)$ is initially not quantized and the values of $s(n)$ are distributed according to a continuous probability density function $p(s)$.

Next consider the quantization of the speech signal for the purpose of transmission such that $s(n)$ is quantized to N quantization levels $\{s_1, s_2, \ldots, s_N\}$.

a) i) Derive an expression for the mean square quantization noise due to this quantization operation. [4]

ii) Show that the mean square quantization noise can be minimized when

$$\int_{1/2(s_{i-1}+s_i)}^{1/2(s_i+s_{i+1})} (s - s_i) p(s) \mathrm{d}s = 0$$

where i is the quantization level index and takes values over a suitable range of the N quantization levels.

[6]

b) Next assume $N = 3$ and

$$p(s) = \begin{cases} 1 - |s| & \text{for} \quad |s| \leq 1 \\ 0 & \text{for} \quad |s| > 1 \end{cases}.$$

Show that the choice of quantization levels s_i that minimizes the mean square quantization noise is

$$s_i = \left\{ -\frac{1}{2}, \quad 0, \quad \frac{1}{2} \right\}.$$

[6]

c) Briefly explain the difference between scalar quantization and vector quantization. State whether scalar or vector quantization would be more efficient for speech signals, and give clear reasoning to support your statement. [4]

3. a) The principal elements of a Code-Excited Linear Predictor (CELP) speech coder include

- an input speech signal to be encoded, $s(n)$
- an excitation codebook with the k^{th} codeword denoted, $x_k(n)$
- a codebook gain factor, g_k
- a long-term predictor
- a vocal tract filter, $V(z)$
- a perceptual weighting filter, $W(z)$
- a coding error, $e(n)$.

i) Draw a labelled diagram of the CELP coder and write a step-by-step explanation of the encoder's operation. [6]

ii) The perceptually weighted output of the vocal tract filter in the CELP coder can be written as $g_k y_k(n) + q(n)$. Deduce and explain what $y_k(n)$ and $q(n)$ represent. [3]

iii) Derive an expression for g_k^{opt}, the value of g_k that minimizes E, where E is the energy of $e(n)$ in a frame $\{F\}$ of speech. [4]

iv) Show that, for $g_k = g_k^{\text{opt}}$, the energy E is given by

$$\sum_{n\in\{F\}} (t(n) - q(n))^2 - \frac{\left(\sum_{n\in\{F\}} (t(n) - q(n)) y_k(n)\right)^2}{\sum_{n\in\{F\}} y_k^2(n)}$$

where $t(n)$ represents the perceptually weighted speech signal.

[4]

b) The G.711 speech coder employs a-law or μ-law waveform coding. Compare and contrast G.711 and CELP speech coding techniques. [3]

4. A hidden Markov model (HMM) used in isolated word speech recognition is defined as having S states for which the transition probability from state i to state j is denoted a_{ij}. The speech signal to be recognized is represented by T feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, each such feature vector representing one frame of the speech signal.

a) Now consider the training of this HMM on some example speech for which the corresponding text is known.

i) State clearly the aim of the training procedure. [1]

ii) Describe an appropriate training procedure for this HMM. Include in your answer any relevant mathematical analysis and supporting diagrams. [5]

b) For an arbitrary state in the HMM, let the probability of a transition from the current state to the next state be denoted p. Show that the length of time, D, spent in the state has an average duration of $1/p$ frames . [4]

c) Five feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5$ are compared with a 3-state HMM. The output probability densities of the feature vectors for each state are shown in Table 1 . The state diagram of the HMM is shown in Figure 4.1 in which the labels on the arrows indicate the state transition probabilities. The maximum probability density that the model generates the sequence of feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$ from any sequence of states for which frame 1 is in state 1 and frame t is in state s is denoted $B(s,t)$.

Draw the lattice showing the states vertically and the frames horizontally and on the lattice draw the feasible paths from $(\text{state}, \text{frame}) = (1,1)$ to $(3,5)$. Determine the maximum probability alignment of the frames to the states of the HMM and the value of $B(3,5)$ corresponding to this alignment. [10]

	frame $\mathbf{x}_1$	frame $\mathbf{x}_2$	frame $\mathbf{x}_3$	frame $\mathbf{x}_4$	frame $\mathbf{x}_5$
state 1	0.5	0.2	0.6	0.4	0.5
state 2	0.5	0.7	0.3	0.1	0.5
state 3	0.5	0.5	0.1	0.6	0.6

Table 1 Output probabilities.

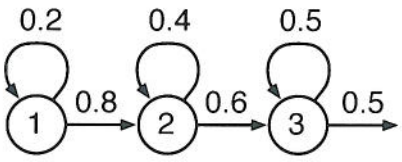


Figure 4.1 Hidden Markov model.

master copy

IEE4-14

SOLUTIONS 2012

1. The diagrams in Figure 1.1 show signal flow graphs of alternative representations of the lossless tube model of the vocal tract, each using two tube sections. These are denoted structure A and structure B as labelled.

a) Using matrix notation, derive an expression for $\begin{bmatrix} V_1^+ \\ V_1^- \end{bmatrix}$ in terms of $\begin{bmatrix} V_2^+ \\ V_2^- \end{bmatrix}$ and another expression for $\begin{bmatrix} W_1^+ \\ W_1^- \end{bmatrix}$ in terms of $\begin{bmatrix} W_2^+ \\ W_2^- \end{bmatrix}$. [6]

Solution:

In structure A we have

$$
\begin{aligned}
V_2^+ &= (1+r_1)V_1^+ + r_1 V_2^- \\
\Rightarrow V_1^+ &= \frac{1}{1+r_1}\left(V_2^+ - r_1 V_2^-\right) \\
V_1^- &= z^{-1}\left(-r_1 v_1^+ + (1-r_1)V_2^-\right) = \frac{z^{-1}}{1+r_1}\left(-r_1 V_2^+ + V_2^-\right) \\
\begin{bmatrix} V_1^+ \\ V_1^- \end{bmatrix} &= \frac{1}{1+r_1}\begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix}\begin{bmatrix} V_2^+ \\ V_2^- \end{bmatrix}.
\end{aligned}
$$

In structure B we have

$$
\begin{aligned}
W_2^+ &= W_1^+ + r_1 W_2^- \Rightarrow W_1^+ = W_2^+ - r_1 W_2^- \\
W_1^- &= z^{-1}\left(-r_1 W_2^+ + W_2^-\right) \\
\begin{bmatrix} W_1^+ \\ W_1^- \end{bmatrix} &= \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix}\begin{bmatrix} W_2^+ \\ W_2^- \end{bmatrix}
\end{aligned}
$$

b) Derive the transfer function $\frac{Y(z)}{X(z)}$. Also derive the transfer function $\frac{Q(z)}{P(z)}$. Hence derive the relationship between the two transfer functions. [7]

Solution:

For structure B we have

$$
\begin{aligned}
P &= \begin{bmatrix} 1 & -r_0 \end{bmatrix}\begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix}\begin{bmatrix} 1 \\ -r_2 z^{-1} \end{bmatrix} Q \\
&= \left(1 + r_1(r_0+r_2)z^{-1} + r_0 r_2 z^{-2}\right) Q.
\end{aligned}
$$

Similarly, for structure A we have

$$
X(1+r_1) = \left(1 + r_1(r_0+r_2)z^{-1} + r_0 r_2 z^{-2}\right) Y.
$$

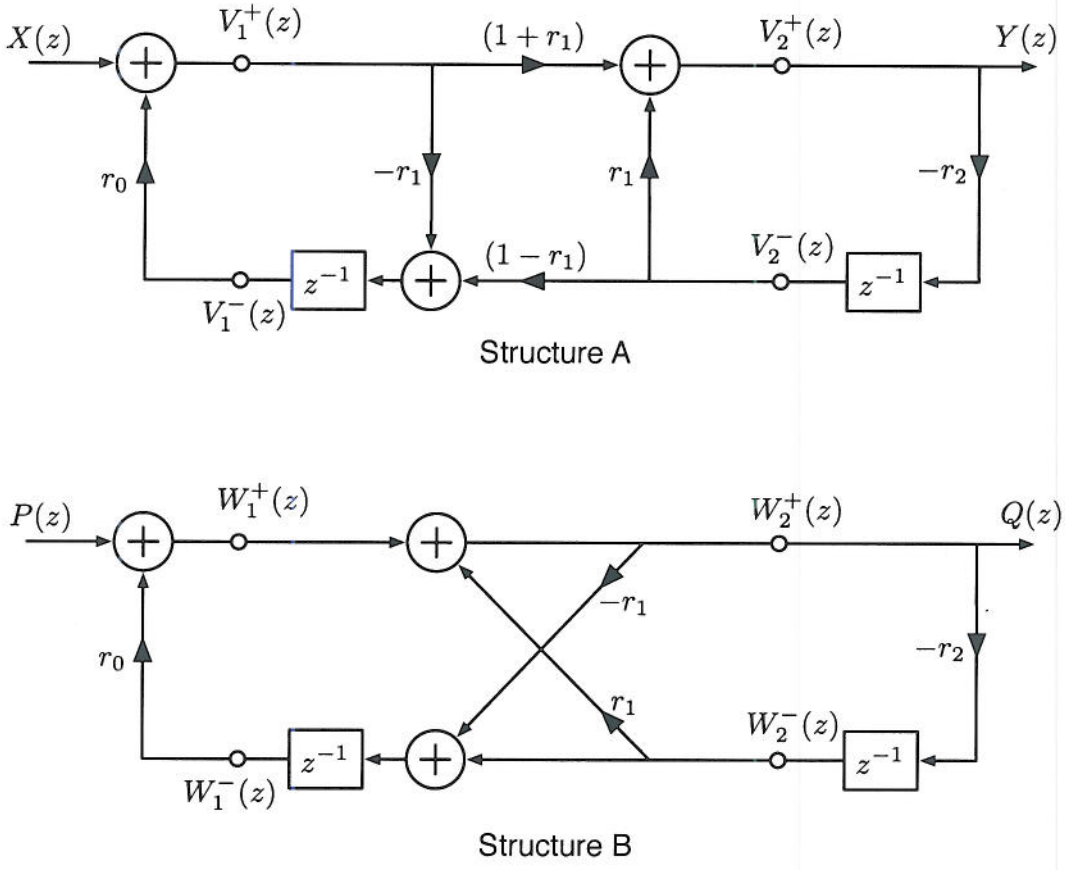


Figure 1.1 Lossless Tube Models

Hence

$$\frac{Y}{X} = (1 + r_1)\frac{Q}{P}$$

c) State the number of multiply operations and addition operations needed to compute each output sample for both structure A and also structure B. [2]

Solution:

Structure A requires 6x and 3+. Structure B requires 4x and 3+.

d) Consider now the structure B. Copy and complete the following function in MATLAB that computes the output sequence $\{q(n)\}$ given an input sequence $\{p(n)\}$. Use one FOR loop in your code to count through each sample.

```
function q = losslessTubeB(p)

                    % COMMENT: p is a vector of input samples
                    % and q is a vector of output samples
    ...

for n=1:N           % COMMENT: this loop goes through each
                    % input sample and computes each
                    % corresponding output sample
    ...

end
```

Note that marks will not be deducted for syntax errors in the code providing that the operations are described unambiguously. It is not necessary to show any variable declarations or intialization code providing they are clear from the context. [5]

Solution:

```
for n=1:N
    pp = p(n);
        wp = pp + r(0)*wm(1);
        wp = wp + r(1)*wm(2);
    wm(1) = wm(2) - r(1)*wp;
    q(n) = wp;
    wm(2) = -r(2)*wp;
end
```

2. Consider a discrete-time speech signal $s(n)$, where n indicates the sample index. The speech signal $s(n)$ is initially not quantized and the values of $s(n)$ are distributed according to a continuous probability density function $p(s)$.

Next consider the quantization of the speech signal for the purpose of transmission such that $s(n)$ is quantized to N quantization levels $\{s_1, s_2, \ldots, s_N\}$.

a) i) Derive an expression for the mean square quantization noise due to this quantization operation. [4]

Solution:

The range of signal from $\frac{1}{2}(s_{i-1}+s_i)$ *as far as* $\frac{1}{2}(s_i+s_{i+1})$ *will be quantized to the value* s_i. *Let us denote these limits as* a_i *and* b_i *respectively. The mean square error is then given by*

$$E = \int_{-\infty}^{\infty} (s - q(s))^2 \, p(s) \mathrm{d}s$$

where $q(.)$ *denotes the quantization function. It is next necessary to split this integral into the sum of ranges for which* $q(s)$ *is constant so as to obtain*

$$E = \sum_{i=1}^{N} \left(\int_{a_i}^{b_i} (s - s_i)^2 \, p(s) \mathrm{d}s \right).$$

ii) Show that the mean square quantization noise can be minimized when

$$\int_{1/2(s_{i-1}+s_i)}^{1/2(s_i+s_{i+1})} (s - s_i) p(s) \mathrm{d}s = 0$$

where i is the quantization level index and takes values over a suitable range of the N quantization levels.

[6]

Solution:

The minimum is found by setting the partial derivatives w.r.t. each s_i *to zero. We obtain*

$$\frac{\partial E}{\partial s_i} = \frac{1}{2} \left((b_{i-1} - s_{i-1})^2 \, p(b_{i-1}) - (a_i - s_i)^2 \, p(a_i) + (b_i - s_i)^2 \, p(b_i) - (a_{i+1} - s_{i+1})^2 \, p(a_{i+1}) \right)$$

$$- 2 \int_{a_1}^{b_i} (s - s_i) \, p(s) \mathrm{d}s = 0$$

using

$$\frac{\partial b_{i-1}}{\partial s_i} = \frac{\partial a_i}{\partial s_i} = \frac{\partial b_i}{\partial s_i} = \frac{\partial a_{i+1}}{\partial s_i} = \frac{1}{2}.$$

The proof follows from that fact that the terms in the first bracket of the above equation sums to zero since

$$a_i = b_{i-1} = \frac{s_i + s_{i-1}}{2}$$

which leaves only the second (integral) term, which shows what is required.

b) Next assume $N = 3$ and

$$p(s) = \begin{cases} 1 - |s| & \text{for} \quad |s| \leq 1 \\ 0 & \text{for} \quad |s| > 1 \end{cases}.$$

Show that the choice of quantization levels s_i that minimizes the mean square quantization noise is

$$s_i = \left\{ -\frac{1}{2}, \quad 0, \quad \frac{1}{2} \right\}.$$

[6]

Solution:

A good starting point is to see that $p(s)$ is symmetric around the origin and therefore the choice of quantization levels should also be symmetric, written as $\{-a, 0, a\}$. The task is then to find a.

The problem formulation, taking into account $p(s)$ and the expression for the quantization levels from (ii), leads to finding the value of a that satisfies

$$\int_{a/2}^{1} (s - a)(1 - s)\, \mathrm{d}s = 0.$$

This gives

$$2a^3 - 9a^2 + 12a - 4 = 2(a - 1/2)(a - 2)(a - 2) = 0$$

leading to the only possible solution $a = 1/2$.

c) Briefly explain the difference between scalar quantization and vector quantization. State whether scalar or vector quantization would be more efficient for speech signals, and give clear reasoning to support your statement. [4]

Solution:

Vector quantization is more efficient than scalar quantization as a fundamental result of rate-distortion theory. Furthermore, since speech samples are statistically dependent, this dependency can be exploited by jointly quantizing blocks of samples. However, this is not normally done in practice.

3. a) The principal elements of a Code-Excited Linear Predictor (CELP) speech coder include

- an input speech signal to be encoded, $s(n)$
- an excitation codebook with the k^{th} codeword denoted, $x_k(n)$
- a codebook gain factor, g_k
- a long-term predictor
- a vocal tract filter, $V(z)$
- a perceptual weighting filter, $W(z)$
- a coding error, $e(n)$.

i) Draw a labelled diagram of the CELP coder and write a step-by-step explanation of the encoder's operation. [6]

Solution:

The steps in CELP coding can be summarized as:

A. *LPC analysis of speech frame*

B. *Using each codebook entry in turn as the input, synthesise speech with the LPC derived filter, subtract the original speech and apply a perceptual weighting filter to downweight noise at formant frequencies. Choose the codebook entry giving the lowest error energy.*

C. *Transmit: LPC coefficients (or equivalent), Codebook index, Gain*

D. *At the receiver, regenerate speech using the corresponding codebook entry and the LPC filter.*

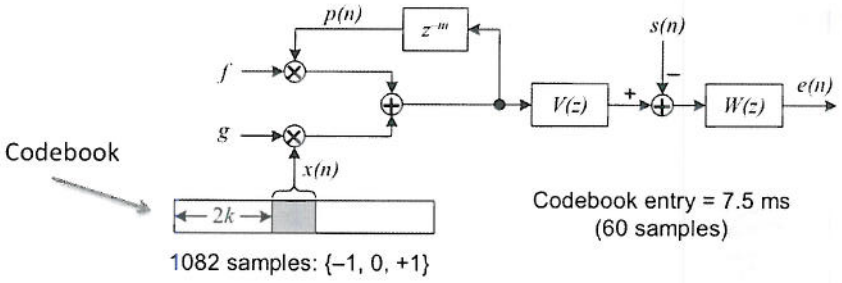


Figure 3.1 CELP Coder

ii) The perceptually weighted output of the vocal tract filter in the CELP coder can be written as $g_k y_k(n) + q(n)$. Deduce and explain what $y_k(n)$ and $q(n)$ represent. [3]

Solution:

The CELP coder includes two additive elements of excitation to the vocal tract filter: the long-term predictor output and the excitation codebook. Given that g is defined as the codebook gain, it is implied that y_k must be the output of the vocal tract filter in response to the codebook whereas q must therefore be the response of the vocal tract filter to the long-term predictor.

iii) Derive an expression for g_k^{opt}, the value of g_k that minimizes E, where E is the energy of $e(n)$ in a frame $\{F\}$ of speech. [4]

Solution:

The first step is to define the sum squared energy to be minimized:

$$E = \sum_n e^2(n) = \sum_n (gy(n) + q(n) - t(n))$$

where $t(n)$ is the perceptually weighted speech signal as indicated in the question.

The optimum value of g is obtained by differentiating w.r.t. g:

$$\frac{1}{2}\frac{\partial E}{\partial g} = \sum_n y(n)e(n) = \sum_n gy^2(n) - \sum_n y(n)\,(t(n) - q(n)) = 0$$

$$g^{opt} = \frac{\sum_n y(n)\,(t(n) - q(n))}{\sum_n y^2(n)}$$

iv) Show that, for $g_k = g_k^{\text{opt}}$, the energy E is given by

$$\sum_{n\in\{F\}} (t(n) - q(n))^2 - \frac{\left(\sum_{n\in\{F\}} (t(n) - q(n))\, y_k(n)\right)^2}{\sum_{n\in\{F\}} y_k^2(n)}$$

where $t(n)$ represents the perceptually weighted speech signal.

[4]

Solution:

Using the value of g^{opt} from the previous question part, we can substitute it into the expression for E to obtain:

$$\begin{aligned} E_{opt} &= g^{opt}\left(g^{opt}\sum_n y^2(n) - 2\sum_n y(n)\,(t(n) - q(n))\right) + \sum_n (t(n) - q(n))^2 \\ &= g^{opt}\left(-\sum_n y(n)\,(t(n) - q(n))\right) + \sum_n (t(n) - q(n))^2 \\ &= \sum_n (t(n) - q(n))^2 - \frac{\left(\sum_n y(n)\,(t(n) - q(n))\right)^2}{\sum_n y^2(n)}. \end{aligned}$$

b) The G.711 speech coder employs a-law or μ-law waveform coding. Compare and contrast G.711 and CELP speech coding techniques. [3]

Solution:

G.711 uses non-uniform quantization to achieve an efficient waveform encoding at 64 kBits/s with a sound quality close to 14 bits/sample linear PCM. CELP is a model-based encoding exploiting perceptual motivations and can achieve between 4.8 an 12.2 kBits/s with varying levels of quality, but always at least toll quality.

4. A hidden Markov model (HMM) used in isolated word speech recognition is defined as having S states for which the transition probability from state i to state j is denoted a_{ij}. The speech signal to be recognized is represented by T feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, each such feature vector representing one frame of the speech signal.

a) Now consider the training of this HMM on some example speech for which the corresponding text is known.

i) State clearly the aim of the training procedure. [1]

ii) Describe an appropriate training procedure for this HMM. Include in your answer any relevant mathematical analysis and supporting diagrams. [5]

Solution:

The training procedure should determine the state transition probabilities and the probability distributions of the features in each state. The probability distributions are typically assumed to be Gaussian and are represented in such cases by the mean and variance for each feature.

The training procedure comprises a first step of initial alignment such as uniform partition of frames to states or a partitioning based on the peaks in the Euclidean distance between neighbouring frames in the feature space. This is followed by a second step of re-estimation that could be performed by Viterbi re-estimation and/or Baum-Welch re-estimation. For full marks, students are expected to give complete descriptions of (at least) one of these methods including supporting analysis (bookwork). Additional credit is given for consideration of triphone models in preference to phones.

b) For an arbitrary state in the HMM, let the probability of a transition from the current state to the next state be denoted p. Show that the length of time, D, spent in the state has an average duration of $1/p$ frames . [4]

Solution:

$$pr(D=n) = p(1-p)^{n-1} \Rightarrow E(D) = \sum_{n=1}^{\infty} np(1-p)^{n-1} = \frac{1}{p}.$$

This result is obtained from differentiating $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \Rightarrow \sum_{n=0}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$.

c) Five feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5$ are compared with a 3-state HMM. The output probability densities of the feature vectors for each state are shown in Table 1 . The state diagram of the HMM is shown in Figure 4.1 in which the labels on the arrows indicate the state transition probabilities. The maximum probability density that the model generates the sequence of feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$ from any sequence of states for which frame 1 is in state 1 and frame t is in state s is denoted $B(s,t)$.

Draw the lattice showing the states vertically and the frames horizontally and on the lattice draw the feasible paths from (state, frame) $= (1,1)$ to $(3,5)$. Determine the maximum probability alignment of the frames to the states of the HMM and the value of $B(3,5)$ corresponding to this

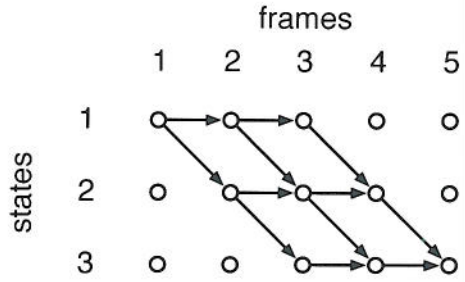


Figure 4.2 Lattice

alignment. [10]

	frame $\mathbf{x}_1$	frame $\mathbf{x}_2$	frame $\mathbf{x}_3$	frame $\mathbf{x}_4$	frame $\mathbf{x}_5$
state 1	0.5	0.2	0.6	0.4	0.5
state 2	0.5	0.7	0.3	0.1	0.5
state 3	0.5	0.5	0.1	0.6	0.6

Table 1 Output probabilities.

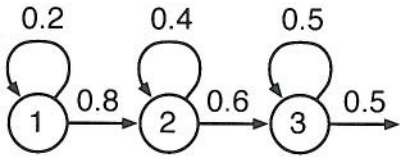


Figure 4.1 Hidden Markov model.

Solution:

The lattice and feasible alignment paths are shown in Fig. 4.2.

Computing the best probabilities for the lattice leads to:

$$
\begin{aligned}
&B(1,1) = 0.5 \\
&B(1,2) = 0.5 \times 0.2 \times 0.2 = 0.02 \\
&B(2,2) = 0.5 \times 0.8 \times 0.7 = \mathbf{0.28} \\
&B(1,3) = 0.02 \times 0.2 \times 0.6 = 0.0024 \\
&B(2,3) = \max(0.28 \times 0.4 \times 0.3, 0.02 \times 0.8 \times 0.3) = \mathbf{0.0336} \\
&B(3,3) = 0.28 \times 0.6 \times 0.1 = 0.0168 \\
&B(2,4) = 0.0336 \times 0.4 \times 0.1 = 0.0013 \\
&B(3,4) = \max(0.0336 \times 0.6 \times 0.6, 0.0168 \times 0.5 \times 0.6) = \mathbf{0.0121} \\
&B(3,5) = \max(0.0013 \times 0.6 \times 0.6, 0.0121 \times 0.5 \times 0.6) = 0.0036
\end{aligned}
$$

and the alignments path for the five frames is states $\{1,2,2,3,3\}$.