EEE116 – Multimedia Systems 2007/08
Solutions to tutorial problem sheet 4 (Week 5)

(Q.7) Derive the Huffman code for the following message:

A E B A B D A A C D C D A A B C E A E A A B B D A A C A E D B A A C A E D C B

Calculate the average bits per symbol and compare with Shannon's Entropy value.

First count the frequency of occurrence (f) for each of the different symbols and compute the probabilities:

	f	P_i
A	15	0.3846
B	7	0.1795
C	6	0.1538
D	6	0.1538
E	5	0.1282
Total	39	1

Huffman Codes

	P	Merge 1	Merge 2	Merge 3		Huffman code
A	0.39				0	0
B	0.18		1 0.33	1 0.61	1	111
C	0.15		0			110
D	0.15	1 0.28		0		101
E	0.13	0				100

Average bits per symbol = $\sum_{i=1}^{5} n_i P_i$ = 1(0.39)+3(0.18)+3(0.15)+3(0.15)+3(0.13)

= 2.23 bits/sample

Now compute the Shannon entropy by computing the $-\log_2(P_i)$ and computing the products of $-\log_2(P_i)$ and P_i for each symbol. The sum of all products give the Shannon's entropy.

	f	P_i	$-\log_2(P_i)$	$-P_i\log_2(P_i)$
A	15	0.3846	1.3785	0.5302
B	7	0.1795	2.4780	0.4448
C	6	0.1538	2.7004	0.4155
D	6	0.1538	2.7004	0.4155
E	5	0.1282	2.9635	0.3799
Total	39	1		2.1858

Shannon's entropy = 2.19 bits per sample (this is the theoretical minimum)

Efficiency = (Shannon's Entropy)/ (average code length)
= 2.19/2.23 = 98.2%

Compression ratio = Original code length / new code length
= 3 / 2.19 = 1.37:1

(Q.12) A 25-value of data set is shown below.

17 17 17 18 18 17 16 16 16 16 16 17 18 19 19 19 19 20 20 20 20 20 19 18 18

(i) Compute the Shannon's entropy value for the data set.

First count all the symbols to find the frequency of occurrence and

symbol	f	P
17	5	0.2
18	5	0.2
16	5	0.2
19	5	0.2
20	5	0.2

In this case all the probabilities are equal. In such a case you can compute the Entropy without using the formula as follows:
Entropy = $-\log_2(p) = -\log_2(0.2) = \log_2(5) = 2.32$ bits/sample.

Try to prove this using the formula:

(ii) The data is first represented in digital format using fixed length codes. How many bits are required for a codeword using the fixed length code?

Using N bits, we can represent 2^N different symbols.
Therefore, $2^N=5$.
$N = \log_2(5) = 2.32$
That means we need 3 bits.

(iii) A simple lossless compression model is used for reduction of the average code lengths. A simple lossless compression model consists of DPCM followed by the Huffman coding. Derive the DPCM of this data set. What is the Shannon's entropy value for the new data set? How much compression is achieved by the DPCM process?

The original data set:
17 17 17 18 18 17 16 16 16 16 16 17 18 19 19 19 19 20 20 20 20 20 19 18 18

Now consider the DPCM data set:

Current DPCM= current original value – previous original value.
For the first data point we assume the previous value was 0.

17 0 0 1 0 -1 -1 0 0 0 0 1 1 1 0 0 0 1 0 0 0 0 -1 -1 0

Now count the symbols in the new data set to create the table:

symbol	f_i	p_i	$-\log_2(p_i)$	$-p_i \log_2(p_i)$
17	1	0.04	4.64	0.186
0	15	0.6	0.74	0.442
1	5	0.2	2.32	0.464
-1	4	0.16	2.64	0.423
Totals	25	1		1.51

The Shannon's entropy value is 1.51 bits/sample.

We measure compression by computing the compression ratio. The generic formula for computing the compression ratio is:

Compression ratio = (original size) / (new size)

The size can be measured in total bits, average code length or the entropy.

The original data set requires 3 bits/sample.
The new data set requires only 2 bits/sample.

Therefore, the compression is 3/ 2 = 1.5:1.

(iv) Devise the Huffman code for the new data set.

symbol	p_i	Merge 1		Merge 2		Merge 3		code
0	0.6					1	1.0	1
1	0.2			1	0.4	0		01
-1	0.16	1	0.2	0				001
17	0.04	0						000

The derived code is 0=1 1=01 -1=001 17=000

(v) What is the efficiency of the devised Huffman code?

We need to compute the average code length first.

Average code length = $\sum_{i=1}^{4} n_i P_i$ = 1(0.6)+2(0.2)+3(0.16)+3(0.04)

= 1.6 bits/sample

The efficiency = (Shannon's Entropy)/ (average code length)
= 1.51/1.6 = 94.4%

(vi) Estimate the inter-sample redundancy and the coding redundancy removed by the DPCM and Huffman coding, respectively?

To answer this question we have to compute the redundancy of the representations at the three points of the compression model. Remember the compression model we used?

Original data → DPCM modelled data → Huffman code data

The final entropy of the DPCM modelled data set is 1.51 bits/sample.
The entropy of the original data set is 2.32 bits/sample

We redefine the efficiency in order to address the data modelling process. In fact, in this case the efficiency measures the efficiency of the data distribution.

Redundancy = 1-(Shannon's Entropy of the model data)/(Shannon's Entropy of the original data)
= 1-(1.51/2.32) = 34.9%

This redundancy is due to inter-sample redundancy.

If we don't use the Huffman code, the DPCM modelled data can be represented by using 2 bits/sample fixed length code.
Therefore the redundancy in the DPCM representation is
= 1- efficiency
= 1-(Shannon's Entropy)/ (average code length)
= 1-(1.51/2) = 24.5%
That means the fixed length code is a 24.5% redundant representation compared to the theoretical minimum.

The final Huffman coded representation uses 1.6 bits per sample.
The redundancy in the final representation 1- efficiency
= 1-(1.51/1.6) = 5.6%

That means the Huffman coding is 5.6% redundant representation compared to the theoretical minimum.

That means we can reduce the coding redundancy from 24.5% to 5.6%

(vii) What is the overall compression ratio?

Initial bit rate = 3 bits per sample
Final bit rate = 1.6 bite sample
Compression ratio is 3/1.6 = 1.87:1.