Chapter 7

Spectral Amplitude Quantisation

7.1 Introduction

The majority of the spectral characteristics of the speech in the PS SB-LPC are modeled by the 10^{th} order LPC filter. The remaining characteristics are provided by the spectral amplitudes of the LP residual signal. These characteristics are required as the LPC filter does not model successfully all the spectral components of the speech signal. The spectral modeling of the LPC filter is limited by three factors; filter order, all pole modeling and speech stationary assumptions. Because of these assumptions aimed at reducing bit rate, accurate analysis and quantisation of spectral amplitude information is required.

7.2 Background

The main difficulty when quantising spectral amplitude information is that the number of amplitudes to be quantised is dependent upon the pitch length P of the pitch cycle. The number of amplitudes N present is given by

$$N = \frac{P}{f_s} \times f_c = 0.4625\ P \tag{7.1}$$

where f_s is the sampling frequency of 8000 Hz and f_c is the cut off frequency typically 3700 Hz. In the PS SB-LPC the pitch cycle length is assumed to vary between 15

to 150 samples applying (7.1) results in a vector length of between 7 to 70. Suitable quantisation schemes must be designed to take into the account the long length of such a vector and the variation in its length as a function of the pitch. Typically VQ techniques is the preferred method for low bit rate speech coders due to its enhanced performance. Normal VQ routines are applied to fixed vector lengths so alternative methods have been produced in order to solve this problem.

The 2.4 kbps MELP [48] coder uses the fact that low frequency components are more important than high frequencies and therefore the corresponding perceptually important components should be quantised more accurately than the rest. This coder utilises VQ to quantise the first 10 spectral amplitudes with the remaining amplitudes set to unity. This method results in lower distortion when the pitch length is shorter such as for females and children but for longer pitch lengths such as in male speech the number of harmonics can be high and as a result only a small number of the spectral amplitudes are transmitted with accuracy. For example a pitch length of 100 samples produces a vector length of over 46 entries, as only accurate information on the first ten values of this vector are kept only a small amount of information up to 800 Hz is utilised which can lead to rather severe quality degradation.

In [3] a Mel-Scale Transformation was used to translate the variable length amplitude vectors to a fixed length. This method divides the spectrum into frequency bands, the amplitudes contained in the bands are averaged and quantised as a single value. The frequency bands are selected by a Mel-Scale measure which takes into account the variation in sensitivity of the human ear with frequency. In theory, the bands are of equal perceptual performance. As none of the spectral amplitudes are discarded this results in less synthetic speech quality degradation especially in male speech where the number of spectral amplitudes to be quantised is high. The variable dimension spectral vector x of length L is converted into a fixed dimension vector z of length N as

$$z(m) = \frac{1}{u_m - l_m + 1} \sum_{k=l_m}^{u_m} x^2(k) \tag{7.2}$$

where $x(k)$ and $z(m)$ denote the k^{th} and the m^{th} elements of the vector x and z respectively and l_m and u_m denote the lower and upper harmonic bounds of the m^{th}

spectral band $[K\frac{m}{M} - 0.5]$ and $[K\frac{m+1}{M} - 1.5]$ (with $u_{M-1} = K - 1$) respectively. The fixed length vector z can be quantised through VQ and if it contains enough bands the warping technique should be transparent. However to obtain good speech quality over N, typically 20, or more bands for male speech are needed, requiring many bits to quantise this number of values. Although capable of providing good speech quality at higher bit rates, for low bit rates this method is not efficient.

Optimal vector quantisation of variable-dimension vectors in principle is feasible by using a set of fixed dimension VQ codebooks. However such a multi-codebook approach would be excessive in storage and computational complexity. Variable Dimension Vector Quantisation (VDVQ) [4]-[6] attempts to solve this problem by transforming the codevectors of the quantisers codebook instead of the input vectors. VDVQ uses a single fixed dimension universal codebook covering the entire range of input vector dimensions. This technique aims to reduce the quantisation distortion as the input vectors are not subjected to any transformation which in general create losses.

However this method requires extensive and elaborate training processes to produce the universal codebook with a very large number of training vectors especially with a codebook of high dimension. This method was tried by [67] but gave poor results.

The SB-LPC [67] used a peak picking algorithm to transform the vector to a fixed length. This method selects spectral amplitudes according to their perceptual importance. It was found to produce good quality speech. The PS SB-LPC [63] attempted to utilise this method to quantise the spectral amplitudes, although good quality synthetic speech could not be produced when this method was used in conjunction with the routines outlined in Section 4.5.3.4. The following section describes the actions taken to accurately quantise the spectral amplitudes based on the peak picking method.

7.3 Quantisation Of Spectral Amplitudes

The current method of quantising the spectral amplitudes in the PS SB-LPC was summarised in Section 4.5.3.4 is to pick and quantise fourteen amplitudes from the first and last cycle in the frame. The fourteen amplitudes per cycle of the remaining

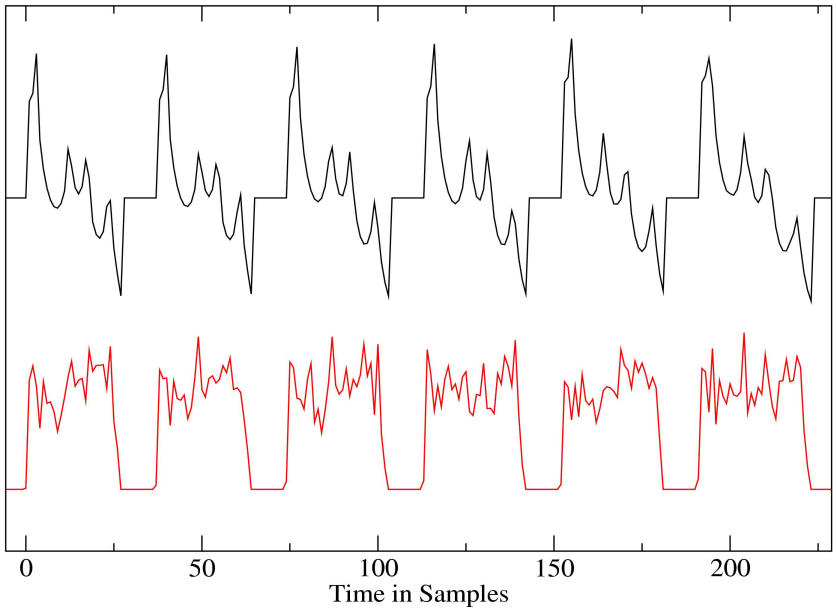

Figure 7.1: Log amplitude of LP spectrum (top) and corresponding spectral amplitudes (bottom)

cycles are effectively selected by a Joint Quantisation Interpolation (JQI) across the frame. This method was originally designed for effectively quantising the LSFs. This is possible because the LSF spectrum shape generally varies in a deterministic pattern for each cycle in the speech frame. The spectral amplitudes however do not represent a smooth shape, they describe the speech signal after the vocal tract information has been de-convolved and are almost noise like in shape. This can be clearly seen in Figure 7.1 which shows a log amplitude plot of the LP spectrum and corresponding spectral amplitudes for several cycles of male voiced speech.

Currently the edge vector containing twenty eight amplitudes in total from the first and last cycle in the frame is quantised with 24 bits and the shape vector with 6 bits. If there are two cycles in a frame then the edge vector is not used and the values are quantised directly. Male speech with its longer pitch values frequently contains only two cycles in a frame, this typically means that for a considerable segments of male speech twenty percent of the bits allocated for spectral amplitude quantisation are not used.

An experiment was initiated to see the effect of raising the bit rate on the JQI method. The edge vectors were re quantised at 27 and 30 bits, at four stages of 8,8,8,3 and 8,8,8,6 bits respectively. The effect of this increase in bit rate is shown as Figure 7.2.

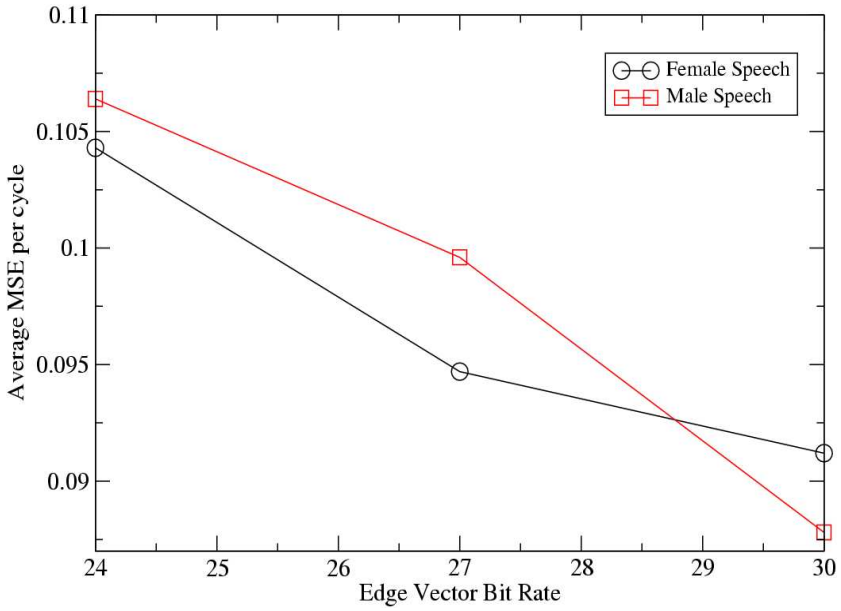

Figure 7.2: Average MSE per cycle of JQI spectral amplitudes for male and female speech

The current peak picking algorithm see Section 4.5.3.4 can only pick a maximum of fourteen amplitudes. The first two amplitudes and no more than three peak amplitudes and one amplitude either side of the peaks; fourteen in total. For testing purposes the maximum number of peaks is set at five peaks therefore seventeen amplitudes can be (but rarely are) picked for each cycle. For experimental purposes this vector can be considered as the target as a 10^{th} order LP filter is considered to contain two formants per peak. The error is measured against this maximum of seventeen amplitudes only as the other amplitudes such as those corresponding to formant valleys do not contain perceptually important information. The average MSE per cycle is measured as

$$MSE = \frac{1}{N}\sum_{i=0}^{N-1}(x(i) - \hat{x}(i))^2 \tag{7.3}$$

where x and $\hat{x}$ are the target unquantised amplitude vector and quantised picked amplitude vector per cycle respectively.

In Figure 7.2 for both male and female speech the average MSE falls with increasing bit rate of the edge vectors, however the fall is greater for male than female speech. It is believed that this due to the fact that female speech uses the shape vector more frequently than male due its greater number of cycles per frame. This increase in bit rate of the edge vectors has a greater effect on male speech as its amplitudes are quantised directly without the influence of the shape vector. As male speech with its greater pitch length, frequently has only two cycles per frame which do not require the use of a shape vector during quantisation.

The number of spectral amplitudes for any given frame is demonstrated in Figure 7.3 which shows Frame A with two cycles and a Frame B with six cycles. It is clear that despite the variation in the number of cycles between the two frames that the number of spectral amplitudes for a given frame size is similar despite the number of cycles per frame. This is a direct consequence of (7.1).

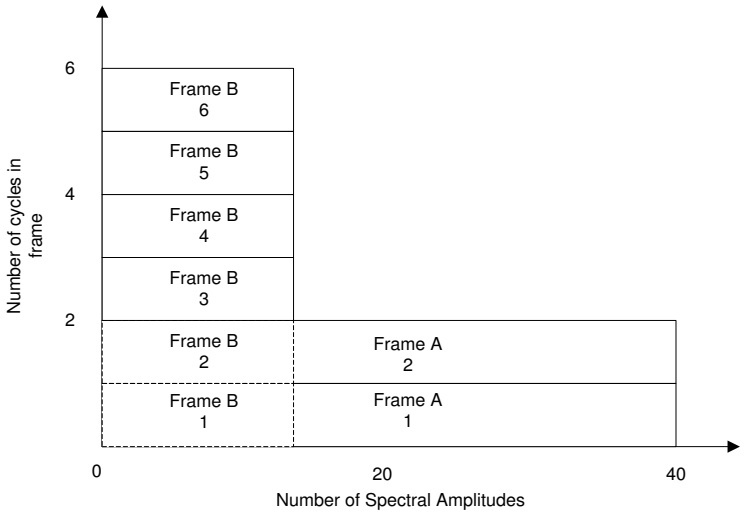

Figure 7.3: Spectral Amplitude allocation for two frames containing differing numbers of cycles

Instead of allocating six bits to a shape vector which interpolates across a noise like signal, it may be more efficient to allocate the six bits to a scheme which allocates these

bits on a block based scheme. The method to attempt this is known here as Block Amps Quantisation (BAQ), the next section will describe the steps taken to implement this idea.

7.3.1 Algorithm Specifics

An amplitude peak picking scheme was first implemented in [67] for the quantisation of spectral amplitudes in the SB-LPC. This method selected a number of spectral amplitudes according to perceptual importance, the other amplitudes are deemed to be of little importance perceptually and are set to one. The selected amplitudes are chosen as:

1. The first two spectral amplitudes since LP modeling can be poor in this area and lower frequencies are more important perceptually
2. Spectral amplitudes under a peak in the LPC filter frequency response ensuring formants are well represented.
3. Spectral amplitudes either side of the peaks in the spectrum due to errors in LSF quantisation which can make the formant positions shift by a few samples
4. The remaining spectral amplitudes corresponding to valleys in the spectrum are considered to be unimportant perceptually and are set to one.

This method of selecting the amplitudes forms the basis of the peak picking in the BAQ method. It is considered for a 10^{th} LPC filter there are only five peaks available for a cycle of speech as each pair of LPC coefficients describes one formant. The maximum number of peaks $Peak_{max}$ which can be picked for each cycle is set at five. The first two amplitudes are always selected and five peaks plus the two amplitudes either side of these five peaks which gives a maximum total of seventeen amplitudes picked.

The number of selected amplitudes allocated per frame is set as $Amps_{fr}$. For example, if $Amps_{fr}$ is set to thirty and there are three cycles in the frame, then the number of amplitudes per cycle $Amps_{cyc}$ is 10. If the number of peaks found in each cycle is 3 then this results in eleven amplitudes being picked - the first two plus the three peaks with

one either side - per cycle and the total number of amplitudes for the frame is thirty three. As in this example $Amps_{fr}$ is set to thirty therefore three amplitudes must be de-selected. The amplitudes to be removed are selected from the following list which shows the degree of importance. For example we would start at the bottom of the list and move up until the correct number has been removed. If the number of amplitudes picked is less than $Amps_{fr}$ then more amplitudes are selected according to the list.

1. First two spectral amplitudes per cycle.
2. The peaks in the cycle.
3. Amplitudes either side of the peaks.
4. Further amplitudes either side of peaks.

If the variation in cycle sizes per frame was found to be greater then six according to (6.7) then the number of amplitudes per cycle are found from a simple ratio of the number of harmonics per cycle compared to the total number of harmonics for the frame, the cycle with the largest number of harmonics is given the largest value of $Amps_{cyc}$.

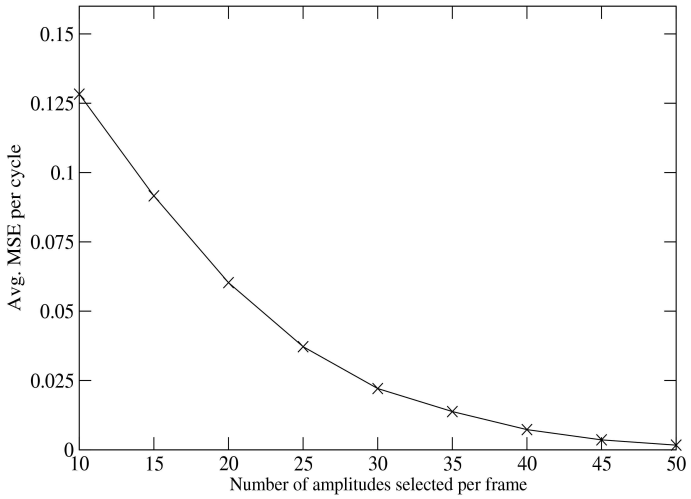

Figure 7.4: Average MSE per cycle with variation in the number of unquantised amplitudes selected per frame

Before quantisation can take place it is important to determine what is the optimum number of $Amps_{fr}$ that can be quantised without causing significant distortion. The value of $Amps_{fr}$ was varied in the range of 10 to 50 in the speech coder and the average MSE per cycle measured using unquantised values of $\hat{x}$ in (7.3), the results are shown as Figure 7.4

As expected as the number of amplitudes per frame increases the MSE falls in value. The steepness of the curve begins to decrease in the region of 25 amplitudes onwards. The synthetic speech produced was evaluated perceptually for the various values of $Amps_{fr}$, it was found that only a small amount of distortion was present when a value of 30 amplitudes per frame was selected, when a value of 50 amplitudes was selected there was little or no discernible distortion present in the unquantised BAQ method.

7.3.2 MSVQ Experiments

Various bit rate configurations were trained using MSVQ routines, they are shown in shown in Table 7.1. They were implemented in the coder using the M-best tree search of Section 3.7.2.2 where M was set to a value of 8 and the average MSE per cycle was found using (7.3) for various values of selected amplitudes per frame. The MSE results are shown as Figure 7.5. Figure 7.5 shows that initially 26 amplitudes per frame gives

Number of Bits	Bit Allocation
22	6,4,4,4,4
24	6,6,4,4,4
26	6,6,6,4,4
28	6,6,6,6,4
30	6,6,6,6,6
32	8,6,6,6,6
34	8,8,6,6,6
36	8,8,8,6,6

Table 7.1: MSVQ bit allocation for Figure 7.5

the lowest average MSE per cycle. But as the bit rates for 30 and 34 amplitudes per

frame respectively reach 1 bit per amplitude they give a superior performance.

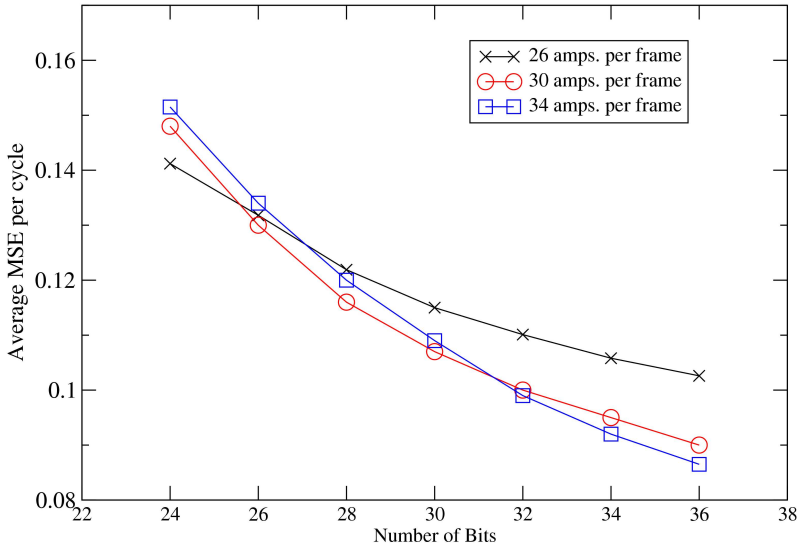

Figure 7.5: Variation in average MSE per cycle at various bit rates during testing

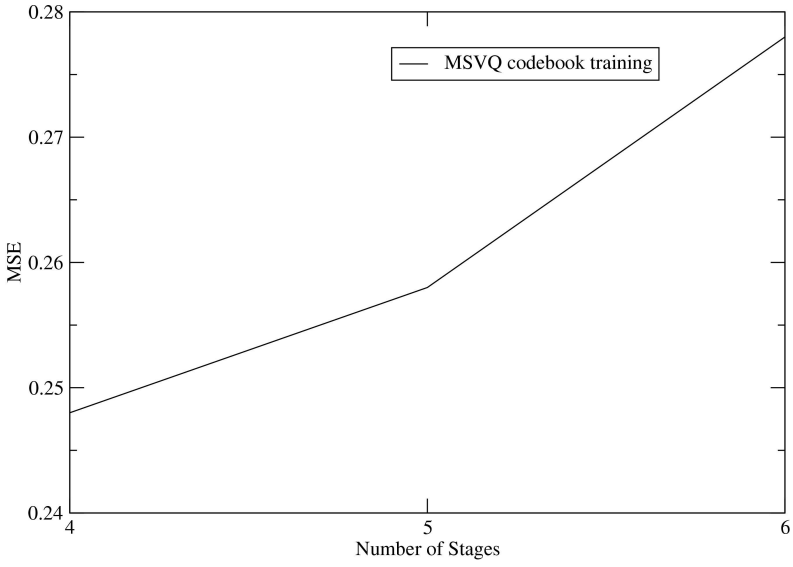

Figure 7.6: MSE variation in quantiser training for BAQ methods in Table 7.2

It was found previously that at a bit rate of 1 bit per amplitude was optimum when quantising spectral amplitudes in the SB-LPC [67]. When these values were evaluated perceptually little difference could be found between selecting 30 and 34 amplitudes

per frame. Therefore 30 spectral amplitudes per frame at a resultant bit rate of 1 bit per amplitude was selected for the BAQ method.

To compare the relative performance of the quantisation configurations a spectral amplitude database of 50,000 sets was used. For the purpose of comparison this size of speech database is effective and should provide reliable results. For a given bit rate the MSVQ can differ in stages. The structure of the quantiser affects complexity and memory storage and affects performance. A lower performance usually results when more structure is imposed on the codebooks but at the benefit of reduced complexity and storage. MSVQ quantisers have been trained all using 30 bits for various stages from 4 to 6. The training results are plotted in Figure 7.6 for the BAQ configurations shown in Table 7.2. It can be seen from Figure 7.6 that as more structure is imposed on the codebooks the error rises during the training process.

Method	Number of stages	Bit allocation	Complexity	Memory	MSE testing	MSE testing interleaved
BAQ	4	8, 8, 8, 6	145920	24960	0.0983	0.0914
BAQ	5	6,6,6,6,6	63360	9600	0.1031	0.0952
BAQ	6	5,5,5,5,5,5	39360	5760	0.1068	0.9986
JQI	3+1	8, 8, 8 + 6	130560	23040	0.1054	N/A

Table 7.2: Comparison of MSVQ structures when quantising spectral amplitude information at a 30-bit bit rate.

The memory and complexity requirements in Table 7.2 are found from (3.46) and (3.49) respectively. The number of stages in (3.49) during training was set at M equal to 8. A comparison can be made between the methods shown here and the JQI method of quantising the amplitudes, these are shown in Table 7.2. The MSE results are found using (7.3). For comparison the MSE results are also shown for the JQI method at 24 bits plus 6 bit shape vector. The BAQ method at four and five stages gives lower error values than the JQI method. The BAQ method at six stages gives a similar error value to JQI but at a much lower memory and complexity requirement.

7.3.3 Interleaving Of Spectral Amplitudes

During peak picking several selected amplitudes per cycle are selected for quantisation. These selected amplitudes from each cycle are placed into an array and vector quantised. Such arrays are demonstrated in parts (a) and (b) of Figure 7.7 which shows the selected amplitudes for frames of two and three cycles respectively.

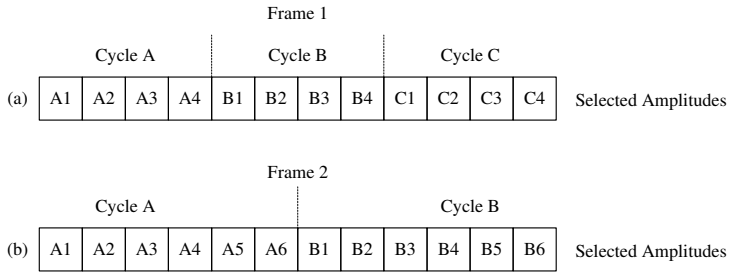

Figure 7.7: Example array of selected amplitudes for quantisation for (a) frame of two and (b) three cycles

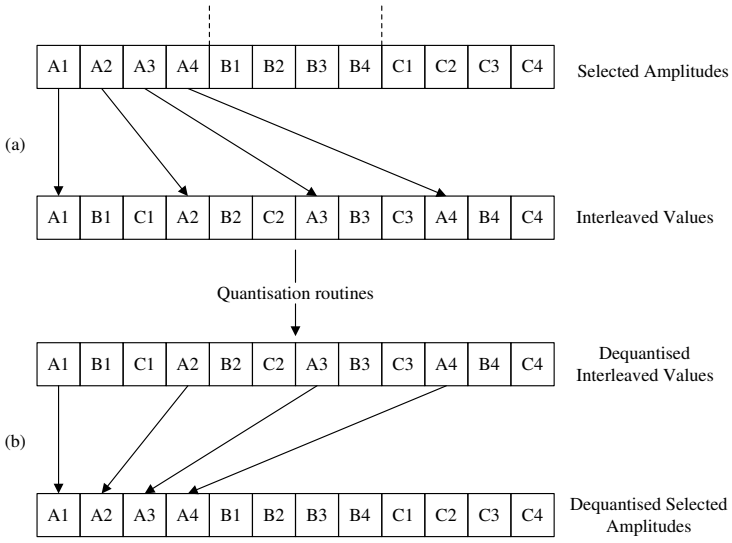

Figure 7.8: Interleaving of selected amplitudes. (a) The array is interleaved before quantisation and (c) the dequantised values

These arrays are fairly uncorrelated to each other as amplitudes of expected similar values are placed at different points in the arrays to be quantised. For example in

the peak picking algorithm the first two amplitudes are always selected. The values therefore of A1, A2, B1, B2, C1 and C2 in frame 1 are at different points to A1, A2, B1 and B2 of frame 2. As the number of cycles per frame can vary from 1 to 12 and 50,000 amplitude vectors are used for quantisation training this factor may have a considerable effect on quantisation performance.

It would be better to group similar amplitudes together before quantisation is carried out. This procedure is illustrated in parts (a) and (b) of Figure 7.8 where values which are likely to be similar are grouped together before quantisation by interleaving. This would produce arrays of spectral amplitudes which are more highly correlated for frames of varying cycle numbers. The interleaved MSE results from testing are shown in Table 7.2 there is a clear improvement in all cases when the spectral amplitudes are interleaved before quantisation.

Interleaving substantially improves the MSE error results. After interleaving all number of stages of the BAQ method give lower error values than the JQI method. When examined perceptually the interleaved BAQ method with the number of stages at four and five gave a similar level of performance. Given that a five stage interleaved BAQ method gives a considerable complexity and memory saving against a four stage implementation, the five stage BAQ method with interleaving was chosen as the method to be used in the PS SB-LPC for quantising the spectral amplitudes.

7.4 Bit Allocation Of Coder

This chapter has detailed the successful design and implementation of spectral amplitude quantisation routine. This routine can be integrated along with the other quantised parameters detailed in the Section 4.5.3. A suggested bit rate allocation for this coder is presented in Table 7.3.

Parameters	Bits per 20ms frame	kbps
LPC	36	1.8
Spectral Amplitudes	30	1.5
PCW pitch length and voicing	16	0.8
Energy	14	0.7
Total	96	4.8

Table 7.3: Example bit allocation for 4.8 kbps PS SB-LPC

7.5 Concluding Remarks

This chapter has detailed the steps taken to carry out quantisation of pitch synchronous spectral amplitude information in a sinusoidal coder. Previous research in the area was presented and compared to current research. A new quantisation method was presented, known as Block Amplitude Quantisation (BAQ). When this method was used to quantise the spectral amplitude information in the PS SB-LPC significant benefits over previous methods were found. By carrying out interleaving on sets of the spectral amplitude information before quantisation, the sets became more highly correlated which resulted in more efficient quantisation.