### 5.2.4 Bass post-filter gain quantization

At 16.4 and 24.4 kbps, the bass post-filter gain is quantized on 2 bits. First the signal is reconstructed as defined in subclause 6.1.3. The bass post-filter is applied and the enhancement signal,  is computed as described in subclause 6.1.4.2.

The residual signal is then low pass filtered in time domain by a convolution with the impulse response , impulse response simulating the filtering done in CDLFB at the decoder side. Moreover the signal is adjusted by an estimated gain  corresponding to the attenuation factor of the anti-harmonic components. It is estimated as follows;

 (657)

The optimal gain adjustment is computed as:

 ()

The optimal gain adjustment is quantized on 2 bits as follows:

 ()

The quantization is adjusted in case the delta SNR provided by the estimated gain is detected as positive:

 ()

The index *k* is then modified in the following way:

 ()

The again adjustment is decoded as follows

 ()

### 5.2.5 Source Controlled VBR Coding

#### 5.2.5.1 Principles of VBR Coding

VBR coding [20] [21] describes a method that assigns different number of bits to a speech frame in the coded domain depending on the characteristics of the input speech signal. This method is often called source-controlled coding as well. Typically, a source-controlled coder encodes speech at different bit rates depending on how the current frame is classified, e.g., voiced, unvoiced, transient, or silence. Note that DTX operation can be combined with VBR coders in the same way as with Fixed Rate (FR) coders; the VBR operation is related to active speech segments.

The speech signal contains a varying amount of information across time, due to the way the human speech production system operates. Stationary voiced and unvoiced segments are good candidates to be encoded at lower bit-rates with minimal impact to voice quality. Transient speech contains information which is normally not well correlated to the past signal, and therefore hard to predict from the past. As a result, they are typically encoded at the higher bit-rates. Therefore it is reasonable to quantize each of the types of signals using only the necessary amount of bits, which has to be varied for maximal efficiency while at the same time minimizing the impact to voice quality.

The VBR solution provides narrowband and wideband coding using the bit rates 2.8, 7.2 and 8.0 kbps and produces an average bit rate at 5.9 kbps.

The Average Data Rate (ADR) control mechanism in subclause 5.2.5.5 relies on properties of the human speech production system, which do not apply well across different types of music signals. In such cases, the ADR for VBR mode starts approaching the most frequently used peak rate of 7.2 kbps. Due to the finer bit allocation, in comparison to Fixed Rate (FR) coding, VBR offers the advantage of a better speech quality at the same average active bit rate than FR coding at the given bit rate. The benefits of VBR can be exploited if the transmission network supports the transmission of speech frames (packets) of variable size, such as in LTE and UMTS networks.

#### 5.2.5.2 EVS VBR Encoder Coding Modes and Bit-Rates

The signal classification algorithm described in subclause 5.1.13 forms the basis for coding mode selection in the VBR encoder. In addition to the coding modes described in subclause 5.1.13, the VBR encoder introduces two low bit-rate (2.8 kbps) coding modes called PPP (Prototype Pitch Period) for voiced speech and NELP (Noise Excited Linear Prediction) for unvoiced speech. The Transition Coding (TC) mode uses 8 kbps and all other coding modes operate at 7.2 kbps. The VBR mode targets an average bit-rate of 5.9 kbps by adjusting the proportion of the low bit-rate (2.8 kbps) and high bit-rate (7.2, 8 kbps) frames for optimal voice quality.

Next, we describe the VBR specific algorithmic modules and the average rate control mechanism.

#### 5.2.5.3 Prototype-Pitch-Period (PPP) Encoding

##### 5.2.5.3.1 PPP Algorithm

It was the perceptual importance of the periodicity in voiced speech that motivated the development of the (Prototype Pitch Period) PPP coding technique. PPP exploits the fact that pitch-cycle waveforms in a voiced segment do not change quickly in a frame. This suggests that we do not have to transmit every pitch-cycle to the decoder; instead, we could transmit just a representative prototype pitch period. At the decoder, the non-transmitted pitch-cycle waveforms could then be derived by means of interpolation. In this way, very low bit-rate coding can be achieved while maintaining high quality reconstructed voiced speech. Figure 41 illustrates an example of how the pitch-cycle waveforms are extracted and interpolated. We will refer these pitch-cycle waveforms as Prototype Pitch Periods (abbreviated as PPPs).



Figure 41: Illustration of principles of PPP coding [20]

The quantization of PPP is carried out in frequency domain. Hence the time-domain signal is converted to a discrete-Fourier series (DFS), whose amplitudes and phases are independently quantized.

Figure 42 presents a high-level schematic diagram of PPP coding scheme. Front-end processing including LPC analysis is same for this scheme. The LSP quantization scheme for the PPP mode is the same as that of the Voiced Coding (VC) mode.



Figure 42: Block diagram of PPP coding scheme [20]

After computing the residual signal, a PPP is extracted from the residual signal on a frame-wise basis. The length of the PPP is determined by the lag estimate supplied by the pitch estimator. Special attention needs to be paid to the boundaries of the PPP during the extraction process. Since each PPP will undergo circular rotation in the alignment process (as will be seen later), the energy around the PPP boundaries needs to be minimized to prevent discontinuities after circular rotation (where the left side of the PPP meets the right side). Such minimization can be accomplished by slightly jittering the location of the extraction. After the extraction, the PPP of the current frame needs to be time-aligned with the PPP extracted from the previous frame. Specifically, the current PPP is circularly shifted until it has the maximum cross-correlation with the previous PPP. The alignment process serves two purposes: Firstly, it facilitates PPP quantization especially in low-bit-rate predictive quantization schemes, and secondly it facilitates the construction of the whole frame of excitation in the synthesis procedure which will be discussed next. The aligned PPPs are then quantized and transmitted to the decoder. To create a full-frame of excitation from the current and previous PPPs, we need to compute an instantaneous phase track which is designed carefully so as to achieve maximum time-synchrony with the original residual signal. For this purpose, a cubic phase track can be employed along with four boundary conditions:

(1) the initial lag value,

(2) the initial phase offset,

(3) the final lag value and

(4) the final phase offset.

Having created the entire frame of quantized residual, the decoder concludes its operation with LPC synthesis filter and memory updates.

##### 5.2.5.3.2 Amplitude Quantization

The DFS amplitude is first normalized by two scaling factors at the encoder – one for the low band and one for the high band. The two resulting scaling factors are vector-quantized in the logarithmic domain and transmitted to the decoder. The normalized spectrum is non-uniformly-downsampled/averaged to transform a variable dimension vector (pitch lag dependent) to a fixed dimension vector. The Equivalent Rectangular Bandwidth (ERB) auditory scale is used for the downsampling/averaging process which helps to model the frequency-dependent frequency resolution of the human auditory system and removes perceptually irrelevant information in the spectra. The downsampled spectrum is split into a low band and high band spectra, each of which is separately quantized. At the decoder, the low and the high band spectra are first reconstructed from the bit-stream transmitted from the encoder. The two spectra are then combined and sent to a non-uniform spectral upsampler. Afterwards, the scaling factors recovered from the bit-stream are used to denormalize the upsampled spectrum to reconstruct the quantized spectrum. Both the scaling factors and the spectra are quantized differentially to ensure minimal bit consumption.

##### 5.2.5.3.3 Phase Quantization

The phase spectrum of the current PPP can be readily derived by combining the phase spectrum of the previous PPP and the contributions from a simplified pitch contour derived from the previous and current frame pitch lags. Finally, the number of samples needed to align the pitch pulse of the quantized PPP with that of the original residual signal is computed and sent to the decoder. This helps with closed loop pitch search in the subsequent ACELP frame.

#### 5.2.5.4 Noise-Excited-Linear-Prediction (NELP) Encoding

The objective of NELP coding is to accurately capture unvoiced segments of speech using a minimal number of bits per frame. Front-end processing including LPC analysis is same for this scheme. The LSP quantization scheme for the NELP mode is the same as that of the Unvoiced Coding (UC) mode. The resulting residual signal is then divided into smaller sub-frames whose gains are computed and quantized. The quantized gains are applied to a randomly generated sparse excitation which is then shaped by a set of bandpass filters. The spectral characteristics of the shaped excitation are analyzed and compared to the spectral characteristics of the original residual signal. Based on this analysis, a filter is chosen to further shape the spectral characteristics of the excitation to achieve optimal performance.

#### 5.2.5.5 Average Data Rate (ADR) Control for the EVS VBR mode

This section describes an ADR control mechanism which ensures the compliance of ADR requirements under a wide variety of language and noise type mix. The average rate control is done by a combination of threshold changes and the change of the high rate and low rate frame selection pattern. A set of thresholds referred to as bump-up thresholds play a key role in the rate control algorithm. When coding PPP frames, the encoder runs a set of checks to verify whether the given frame is suited for PPP mode of coding. If the set of checks fail the PPP coding module decides the given frame is not suitable for PPP coding and the frame is coded as a high rate frame (H-frame). This rejection is referred to as a “bump-up”. There are two types of bump-ups used in the PPP coding module.

1. Open loop bump-ups: Bump-up is performed before the prototype pitch period wave form is coded. For example if the pitch lag difference between the current and previous frame is more than a certain threshold, the PPP coding module decides that the current frame is not suitable for PPP coding.

2. Closed-loop bump-ups: This is done after the prototype pitch period waveform is coded. For example if the energy ratio of the prototype pitch waveform before and after the quantization is not within a certain set of thresholds, the PPP coding module abandons the PPP mode of coding and subsequently that frame is coded using high rate coding. In general the close loop bump-ups make sure the PPP coding is in good quality.

The PPP coding module decides the current speech frame as clean speech or noisy speech by comparing the current frame’s SNR against a threshold. This threshold is referred to as the clean and noisy speech threshold () which is localized to the PPP coding module. If the SNR of the current frame is more than  then the current frame is classified as a clean speech frame or as a noisy speech frame otherwise. For each of the two cases (clean and noisy) there are two sets of bump up thresholds, resulting four sets of bump up thresholds collectively. For clean speech the two bump up threshold sets consist of a strict set  (a strict set of bump up thresholds resulting more bump ups) and a relaxed set (a relaxed set of bump up thresholds resulting less bump ups). Similarly there are two similar bump up threshold sets available for noisy speech  and . Clean speech bump up threshold sets are more stricter compared to corresponding noisy speech bump up threshold sets ( creates more bump ups compared to  and similarly  creates more bump ups compared to ).



ADR Control mechanism is introduced based on multiple steps depending on long term ADR, short term ADR (ADR during last 600 frames) and the target rate. Following rate control mechanisms are picked based on the long term and short term average rates to achieve the desired average rate.

a. Change the threshold () which classifies the speech as clean or noisy. Increasing  classifies more frames as noisy speech and reduces the number frames classified as clean. At this point  and  threshold sets are used for bump ups. However by classifying more frames as noisy speech, more frames will use the threshold set  thus lesser number of bump ups will occur.

b. Use a low rate (L) and high rate (H) frame pattern which generates more low rate frames. For example we can set the default pattern to LHH and change the pattern to LLH to obtain more L frames, which reduces the ADR.

c. Use relaxed PPP bump up threshold sets (  and ) for both clean and noisy speech. This reduces the number of bump ups thus more L frames are generated. In item (a) we increased the threshold () which classifies the speech frames as clean or noisy without relaxing the corresponding PPP bump up thresholds (used  and ).

d. Impact the open loop voicing and un-voicing decisions to reduce the rate by increasing PPP and NELP frames.

e. Apart from rate reduction mechanisms, the algorithm utilizes a speech quality improvement strategy if the global rate is less than the target rate by a specific margin. To achieve that a percentage of the L frames are sent to H frames to improve the speech quality.

The objective of the ADR control algorithm is to keep the average data rate of the VBR mode at 5.9 kbps and not to exceed it by 5%. The target rate is set by the algorithm (e.g. 5.9 kbps) and short term and long term average rates are computed to control different actions given in items (a) through (e) above.

The average rate during last N frames or the short term average rate ( average of last 600 active frames) is used to compute the long term average rate as follows.  is updated after every N active frames. The value  is selected as 0.98.

 ()

The rate control is done in multiple steps. If the long term average rate  is larger than the target rate, the clean and noise decision threshold  (in above (a)) is increased in steps. If the  cannot be reduced with the maximum  value, the encoder will use the LLH pattern to generate more quarter rate PPP frames. If the  is not reduced below the target rate, bump up thresholds are relaxed (item (c) above). Finally the open loop voiced and unvoiced decisions are relaxed such that the number of PPP and NELP frames will be increased to control the average rate.

Once the  is reduced below the target rate, the rate control mechanisms are relaxed gradually. First the open loop decision thresholds are restored to the values before the aggressive rate control. The next step is to revert to the lesser aggressive bump up thresholds (and ). If the long term average rate is still well under the target rate, frame selection pattern LLH is reverted back to LHH and then the noise decision threshold  is gradually reduced to its default value.

If the long term average rate  is well under control and below a minimum target rate speech quality improvement algorithm in (e) is exercised by converting some of the low rate PPP frames (L frames) into high rate H frames.

Figure 43 shows the flow chart of the average rate control based on this method. Table 51 contains the summary of the terms used in the flowchart in figure 43.

Table 51: Summary of the terms used in the flowchart

|  |  |
| --- | --- |
| **Symbol** | **Description** |
|  | Set to 1 if the LLH pattern is used. Set to 0 if LHH pattern is used |
|  | Maximum value for the clean and noisy speech decision threshold |
|  | Set to 1 if the relaxed bump up thresholds is used. Set to 0 otherwise |
|  | Long term average data rate |
|  | Target average date rate |
|  | Rate tolerance 1. (e.g. set to 0.1 kbps for a target rate of 6.1 kbps target rate) |
|  | Rate tolerance 1. (e.g. set to 0.05 kbps for a target rate of 6.1 kbps target rate) |
|  | Set size to increase the  ( value for the clean and noisy speech decision threshold) |
|  | Set size to decrease the  ( value for the clean and noisy speech decision threshold) |
|  | Set to 1 if open loop voicing/unvoicing decisions are changed to have more PPP and NELP frames. Set to 0 otherwise. |



Figure 43: Average rate control in Variable Bit-Rate Coding