Definition of the Opus Audio Codec
draft-ietf-codec-opus-00

Abstract

This document describes the Opus codec, designed for interactive speech and audio transmission over the Internet.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 18, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
Table of Contents

1.  Introduction .................................................. 3
2.  Opus Codec .................................................... 4
   2.1.  Source Code ............................................... 4
3.  Codec Modes ................................................... 5
   3.1.  Examples .................................................. 6
4.  Security Considerations ............................... 8
5.  IANA Considerations ............................................ 9
6.  Acknowledgments .............................................. 10
7.  Informative References ..................................... 11
Authors’ Addresses ............................................... 12
1. Introduction

We propose the Opus codec based on a linear prediction layer (LP) and an MDCT-based enhancement layer. The main idea behind the proposal is that the speech low frequencies are usually more efficiently coded using linear prediction codecs (such as CELP variants), while the higher frequencies are more efficiently coded in the transform domain (e.g. MDCT). For low sampling rates, the MDCT layer is not useful and only the LP-based layer is used. On the other hand, non-speech signals are not always adequately coded using linear prediction, so for music only the MDCT-based layer is used.

In this proposed prototype, the LP layer is based on the SILK [1] codec [SILK] and the MDCT layer is based on the CELT [2] codec [CELT].

This is a work in progress.
2. Opus Codec

In hybrid mode, each frame is coded first by the LP layer and then by the MDCT layer. In the current prototype, the cutoff frequency is 8 kHz. In the MDCT layer, all bands below 8 kHz are discarded, such that there is no coding redundancy between the two layers. Also both layers use the same instance of the range coder to encode the signal, which ensures that no "padding bits" are wasted. The hybrid approach makes it easy to support both constant bit-rate (CBR) and variable bit-rate (VBR) coding. Although the SILK layer used is VBR, it is easy to make the bit allocation of the CELT layer produce a final stream that is CBR by using all the bits left unused by the SILK layer.

The implementation of SILK-based LP layer is similar to the description in the SILK Internet-Draft [SILK] with the main exception that SILK was modified to use the same range coder as CELT. The implementation of the CELT-based MDCT layer is available from the CELT website and is a more recent version (0.8.1) of the CELT Internet-Draft [CELT]. The main changes include better support for 20 ms frames as well as the ability to encode only the higher bands using a range coder partially filled by the SILK layer.

In addition to their frame size, the SILK and CELT codecs require a look-ahead of 5.2 ms and 2.5 ms, respectively. SILK’s look-ahead is due to noise shaping estimation (5 ms) and the internal resampling (0.2 ms), while CELT’s look-ahead is due to the overlapping MDCT windows. To compensate for the difference, the CELT encoder input is delayed by 2.7 ms. This ensures that low frequencies and high frequencies arrive at the same time.

2.1. Source Code

The source code is currently available in a Git repository [3] which references two other repositories (for SILK and CELT). Some snapshots are provided for convenience at <http://people.xiph.org/~jm/ietfcodec/> along with sample files. Although the build system is very primitive, some instructions are provided in the toplevel README file. This is very early development so both the quality and feature set should greatly improve over time. In the current version, only 48 kHz audio is supported, but support for all configurations listed in Section 3 is planned.
3. Codec Modes

There are three possible operating modes for the proposed prototype:

1. A linear prediction (LP) mode for use in low bit-rate connections with up to 8 kHz audio bandwidth (16 kHz sampling rate)

2. A hybrid (LP+MDCT) mode for full-bandwidth speech at medium bitrates

3. An MDCT-only mode for very low delay speech transmission as well as music transmission.

Each of these modes supports a number of difference frame sizes and sampling rates. In order to distinguish between the various modes and configurations, we need to define a simple header that can be used in the transport layer (e.g. RTP) to signal this information. The following describes the proposed header.

The LP mode supports the following configurations (numbered from 00000...01011 in binary):

- 8 kHz: 10, 20, 40, 60 ms (00000...00011)
- 12 kHz: 10, 20, 40, 60 ms (00100...00111)
- 16 kHz: 10, 20, 40, 60 ms (01000...01011)

for a total of 12 configurations.

The hybrid mode supports the following configurations (numbered from 01100...01111):

- 32 kHz: 10, 20 ms (01100...01101)
- 48 kHz: 10, 20 ms (01110...01111)

for a total of 4 configurations.

The MDCT-only mode supports the following configurations (numbered from 10000...11101):

- 8 kHz: 2.5, 5, 10, 20 ms (10000...10011)
- 16 kHz: 2.5, 5, 10, 20 ms (10100...10111)
- 32 kHz: 2.5, 5, 10, 20 ms (11000...11011)
o 48 kHz: 2.5, 5, 10, 20 ms (11100...11111)

for a total of 16 configurations.

There is thus a total of 32 configurations, so 5 bits are necessary
to indicate the mode, frame size and sampling rate (MFS). This
leaves 3 bits for the number of frames per packets (codes 0 to 7):

o 0-2: 1-3 frames in the packet, each with equal compressed size

o 3: arbitrary number of frames in the packet, each with equal
compressed size (one size needs to be encoded)

o 4-5: 2-3 frames in the packet, with different compressed sizes,
which need to be encoded (except the last one)

o 6: arbitrary number of frames in the packet, with different
compressed sizes, each of which needs to be encoded

o 7: The first frame has this MFS, but others have different MFS.
Each compressed size needs to be encoded.

When code 7 is used and the last frames of a packet have the same
MFS, it is allowed to switch to another code for them.

The compressed size of the frames (if needed) is indicated -- usually
-- with one byte, with the following meaning:

o 0: No frame (DTX or lost packet)

o 1-251: Size of the frame in bytes

o 252-255: A second byte is needed. The total size is (size[1]*4)+
(size[0]%4)+252

The maximum size representable is 255*4+3+252=1275 bytes. For 20 ms
frames, that represents a bit-rate of 510 kb/s, which is really the
highest rate anyone would want to use in stereo mode (beyond that
point, lossless codecs would be more appropriate).

3.1. Examples

Simplest case: one packet
Four frames of the same compressed size:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   MFS   |0|0|0|               compressed data...              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Two frames of different compressed size:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   MFS   |0|1|1|               compressed data...              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Three frames of different _durations_:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 1st MFS |1|1|1|   frame size  | 2nd MFS |1|1|1|   frame size  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 3rd MFS |1|1|1|   frame size  |      compressed data...       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
4. Security Considerations

The codec needs to take appropriate security considerations into account, as outlined in [DOS] and [SECGUIDE]. It is extremely important for the decoder to be robust against malicious payloads. Malicious payloads must not cause the decoder to overrun its allocated memory or to take much more resources to decode. Although problems in encoders are typically rarer, the same applies to the encoder. Malicious audio stream must not cause the encoder to misbehave because this would allow an attacker to attack transcoding gateways.

In its current version, the Opus codec likely does NOT meet these security considerations, so it should be used with caution.
5. IANA Considerations

This document has no actions for IANA.
6. Acknowledgments

Thanks to all other developers, including Soeren Skak Jensen, Gregory Maxwell, Christopher Montgomery, Karsten Vandborg Soerensen, and Timothy Terriberry.
7. Informative References


Authors' Addresses

Jean-Marc Valin
Octasic Inc.
4101, Molson Street
Montreal, Quebec
Canada

Phone: +1 514 282-8858
Email: jean-marc.valin@octasic.com

Koen Vos
Skype Technologies S.A.
Stadsgaarden 6
Stockholm, 11645
SE

Phone: +46 855 921 989
Email: koen.vos@skype.net
Abstract

This document defines the Opus interactive speech and audio codec. Opus is designed to handle a wide range of interactive audio applications, including Voice over IP, videoconferencing, in-game chat, and even live, distributed music performances. It scales from low bitrate narrowband speech at 6 kb/s to very high quality stereo music at 510 kb/s. Opus uses both linear prediction (LP) and the Modified Discrete Cosine Transform (MDCT) to achieve good compression of both speech and music.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 30, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction ........................................ 5
   1.1. Notation and Conventions .......................... 6
2. Opus Codec Overview .................................. 8
   2.1. Control Parameters ............................... 10
      2.1.1. Bitrate ...................................... 10
      2.1.2. Number of Channels (Mono/Stereo) .......... 11
      2.1.3. Audio Bandwidth ............................... 11
      2.1.4. Frame Duration ............................... 11
      2.1.5. Complexity .................................. 11
      2.1.6. Packet Loss Resilience ....................... 12
      2.1.7. Forward Error Correction (FEC) .............. 12
      2.1.8. Constant/Variable Bitrate ................. 12
      2.1.9. Discontinuous Transmission (DTX) ........... 13
3. Internal Framing .................................... 14
   3.1. The TOC Byte .................................... 14
   3.2. Frame Packing ................................... 16
      3.2.1. Frame Length Coding ......................... 16
      3.2.2. Code 0: One Frame in the Packet ........... 17
      3.2.3. Code 1: Two Frames in the Packet, Each with Equal Compressed Size ................. 17
      3.2.4. Code 2: Two Frames in the Packet, with Different Compressed Sizes ................. 18
      3.2.5. Code 3: A Signaled Number of Frames in the Packet .. 18
   3.3. Examples ........................................ 21
   3.4. Receiving Malformed Packets ..................... 22
4. Opus Decoder ........................................ 24
   4.1. Range Decoder ................................... 24
      4.1.1. Range Decoder Initialization ............... 25
      4.1.2. Decoding Symbols .......................... 26
      4.1.3. Alternate Decoding Methods .................. 28
      4.1.4. Decoding Raw Bits ........................... 29
      4.1.5. Decoding Uniformly Distributed Integers .... 30
      4.1.6. Current Bit Usage .............................. 30
   4.2. SILK Decoder ..................................... 32
      4.2.1. SILK Decoder Modules ......................... 33
      4.2.2. LP Layer Organization ....................... 34
      4.2.3. Header Bits ................................. 36
      4.2.4. Per-Frame LBRR Flags ....................... 37
4.2.5. LBRR Frames ........................................... 37
4.2.6. Regular SILK Frames .................................... 38
4.2.7. SILK Frame Contents .................................... 38
  4.2.7.1. Stereo Prediction Weights ............................. 41
  4.2.7.2. Mid-only Flag ....................................... 43
  4.2.7.3. Frame Type .......................................... 44
  4.2.7.4. Subframe Gains ...................................... 45
  4.2.7.5. Normalized Line Spectral Frequency (LSF) and Linear Predictive Coding (LPC) Coefficients .... 47
  4.2.7.6. Long-Term Prediction (LTP) Parameters ............... 75
  4.2.7.7. Linear Congruential Generator (LCG) Seed ............ 87
  4.2.7.8. Excitation .......................................... 87
  4.2.7.9. SILK Frame Reconstruction ........................... 99
4.2.8. Stereo Unmixing ........................................ 102
4.2.9. Resampling ............................................ 103
4.3. CELT Decoder ............................................. 104
  4.3.1. Transient Decoding .................................... 109
  4.3.2. Energy Envelope Decoding ............................... 109
  4.3.3. Bit Allocation ......................................... 111
  4.3.4. Shape Decoding ........................................ 117
  4.3.5. Anti-Collapse Processing ............................... 121
  4.3.6. Denormalization ....................................... 122
  4.3.7. Inverse MDCT ......................................... 122
4.4. Packet Loss Concealment (PLC) ............................ 123
  4.4.1. Clock Drift Compensation ............................... 123
4.5. Configuration Switching ................................... 124
  4.5.1. Transition Side Information (Redundancy) ............... 125
  4.5.2. State Reset .......................................... 128
  4.5.3. Summary of Transitions ................................ 128
5. Opus Encoder .............................................. 131
  5.1. Range Encoder ........................................... 132
    5.1.1. Encoding Symbols .................................... 133
    5.1.2. Alternate Encoding Methods ........................... 134
    5.1.3. Encoding Raw Bits .................................... 135
    5.1.4. Encoding Uniformly Distributed Integers ............... 135
    5.1.5. Finalizing the Stream ................................ 135
    5.1.6. Current Bit Usage ................................... 136
  5.2. SILK Encoder ........................................... 136
    5.2.1. Sample Rate Conversion ............................... 136
    5.2.2. Stereo Mixing ....................................... 137
    5.2.3. SILK Core Encoder ................................... 137
  5.3. CELT Encoder ........................................... 150
    5.3.1. Pitch Prefilter ....................................... 150
    5.3.2. Bands and Normalization .............................. 151
    5.3.3. Energy Envelope Quantization ........................ 151
    5.3.4. Bit Allocation ....................................... 151
    5.3.5. Stereo Decisions ..................................... 152
    5.3.6. Time-Frequency Decision ............................. 153
1. Introduction

The Opus codec is a real-time interactive audio codec designed to meet the requirements described in [requirements]. It is composed of a linear prediction (LP)-based [LPC] layer and a Modified Discrete Cosine Transform (MDCT)-based [MDCT] layer. The main idea behind using two layers is that in speech, linear prediction techniques (such as Code-Excited Linear Prediction, or CELP) code low frequencies more efficiently than transform (e.g., MDCT) domain techniques, while the situation is reversed for music and higher speech frequencies. Thus a codec with both layers available can operate over a wider range than either one alone and, by combining them, achieve better quality than either one individually.

The primary normative part of this specification is provided by the source code in Appendix A. Only the decoder portion of this software is normative, though a significant amount of code is shared by both the encoder and decoder. Section 6 provides a decoder conformance test. The decoder contains a great deal of integer and fixed-point arithmetic which needs to be performed exactly, including all rounding considerations, so any useful specification requires domain-specific symbolic language to adequately define these operations. Additionally, any conflict between the symbolic representation and the included reference implementation must be resolved. For the practical reasons of compatibility and testability it would be advantageous to give the reference implementation priority in any disagreement. The C language is also one of the most widely understood human-readable symbolic representations for machine behavior. For these reasons this RFC uses the reference implementation as the sole symbolic representation of the codec.

While the symbolic representation is unambiguous and complete it is not always the easiest way to understand the codec's operation. For this reason this document also describes significant parts of the codec in English and takes the opportunity to explain the rationale behind many of the more surprising elements of the design. These descriptions are intended to be accurate and informative, but the limitations of common English sometimes result in ambiguity, so it is expected that the reader will always read them alongside the symbolic representation. Numerous references to the implementation are provided for this purpose. The descriptions sometimes differ from the reference in ordering or through mathematical simplification wherever such deviation makes an explanation easier to understand. For example, the right shift and left shift operations in the reference implementation are often described using division and multiplication in the text. In general, the text is focused on the "what" and "why" while the symbolic representation most clearly provides the "how".

1.1. Notation and Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [rfc2119].

Various operations in the codec require bit-exact fixed-point behavior, even when writing a floating point implementation. The notation "Q<n>", where n is an integer, denotes the number of binary digits to the right of the decimal point in a fixed-point number. For example, a signed Q14 value in a 16-bit word can represent values from -2.0 to 1.99993896484375, inclusive. This notation is for informational purposes only. Arithmetic, when described, always operates on the underlying integer. E.g., the text will explicitly indicate any shifts required after a multiplication.

Expressions, where included in the text, follow C operator rules and precedence, with the exception that the syntax "x**y" indicates x raised to the power y. The text also makes use of the following functions:

1.1.1. min(x,y)

The smallest of two values x and y.

1.1.2. max(x,y)

The largest of two values x and y.

1.1.3. clamp(lo,x,hi)

\[
\text{clamp(lo,x,hi) = max(lo,min(x,hi))}
\]

With this definition, if lo > hi, then lo is returned.

1.1.4. sign(x)

The sign of x, i.e.,

\[
\begin{align*}
\text{sign}(x) &= \begin{cases} 
-1, & x < 0, \\
0, & x = 0, \\
1, & x > 0. 
\end{cases}
\end{align*}
\]

1.1.5. abs(x)

The absolute value of x, i.e.,

\[
\text{abs}(x) = \text{sign}(x) \times x.
\]
1.1.6. floor(f)

The largest integer z such that z ≤ f.

1.1.7. ceil(f)

The smallest integer z such that z ≥ f.

1.1.8. round(f)

The integer z nearest to f, with ties rounded towards negative infinity, i.e.,

\[ \text{round}(f) = \text{ceil}(f - 0.5) \, . \]

1.1.9. log2(f)

The base-two logarithm of f.

1.1.10. ilog(n)

The minimum number of bits required to store a positive integer n in binary, or 0 for a non-positive integer n.

\[
\text{ilog}(n) = \begin{cases} 
0, & n \leq 0, \\
\text{floor}(\log_2(n)) + 1, & n > 0 
\end{cases}
\]

Examples:

- \( \text{ilog}(-1) = 0 \)
- \( \text{ilog}(0) = 0 \)
- \( \text{ilog}(1) = 1 \)
- \( \text{ilog}(2) = 2 \)
- \( \text{ilog}(3) = 2 \)
- \( \text{ilog}(4) = 3 \)
- \( \text{ilog}(7) = 3 \)
2. Opus Codec Overview

The Opus codec scales from 6 kb/s narrowband mono speech to 510 kb/s fullband stereo music, with algorithmic delays ranging from 5 ms to 65.2 ms. At any given time, either the LP layer, the MDCT layer, or both, may be active. It can seamlessly switch between all of its various operating modes, giving it a great deal of flexibility to adapt to varying content and network conditions without renegotiating the current session. The codec allows input and output of various audio bandwidths, defined as follows:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Audio Bandwidth</th>
<th>Sample Rate (Effective)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB (narrowband)</td>
<td>4 kHz</td>
<td>8 kHz</td>
</tr>
<tr>
<td>MB (medium-band)</td>
<td>6 kHz</td>
<td>12 kHz</td>
</tr>
<tr>
<td>WB (wideband)</td>
<td>8 kHz</td>
<td>16 kHz</td>
</tr>
<tr>
<td>SWB (super-wideband)</td>
<td>12 kHz</td>
<td>24 kHz</td>
</tr>
<tr>
<td>FB (fullband)</td>
<td>20 kHz (*)</td>
<td>48 kHz</td>
</tr>
</tbody>
</table>

Table 1

(*) Although the sampling theorem allows a bandwidth as large as half the sampling rate, Opus never codes audio above 20 kHz, as that is the generally accepted upper limit of human hearing.

Opus defines super-wideband (SWB) with an effective sample rate of 24 kHz, unlike some other audio coding standards that use 32 kHz. This was chosen for a number of reasons. The band layout in the MDCT layer naturally allows skipping coefficients for frequencies over 12 kHz, but does not allow cleanly dropping just those frequencies over 16 kHz. A sample rate of 24 kHz also makes resampling in the MDCT layer easier, as 24 evenly divides 48, and when 24 kHz is sufficient, it can save computation in other processing, such as Acoustic Echo Cancellation (AEC). Experimental changes to the band layout to allow a 16 kHz cutoff (32 kHz effective sample rate) showed potential quality degradations at other sample rates, and at typical bitrates the number of bits saved by using such a cutoff instead of coding in fullband (FB) mode is very small. Therefore, if an application wishes to process a signal sampled at 32 kHz, it should just use FB.

The LP layer is based on the SILK codec [SILK]. It supports NB, MB,
or WB audio and frame sizes from 10 ms to 60 ms, and requires an additional 5 ms look-ahead for noise shaping estimation. A small additional delay (up to 1.5 ms) may be required for sampling rate conversion. Like Vorbis [Vorbis-website] and many other modern codecs, SILK is inherently designed for variable-bitrate (VBR) coding, though the encoder can also produce constant-bitrate (CBR) streams. The version of SILK used in Opus is substantially modified from, and not compatible with, the stand-alone SILK codec previously deployed by Skype. This document does not serve to define that format, but those interested in the original SILK codec should see [SILK] instead.

The MDCT layer is based on the CELT codec [CELT]. It supports NB, WB, SWB, or FB audio and frame sizes from 2.5 ms to 20 ms, and requires an additional 2.5 ms look-ahead due to the overlapping MDCT windows. The CELT codec is inherently designed for CBR coding, but unlike many CBR codecs it is not limited to a set of predetermined rates. It internally allocates bits to exactly fill any given target budget, and an encoder can produce a VBR stream by varying the target on a per-frame basis. The MDCT layer is not used for speech when the audio bandwidth is WB or less, as it is not useful there. On the other hand, non-speech signals are not always adequately coded using linear prediction, so for music only the MDCT layer should be used.

A "Hybrid" mode allows the use of both layers simultaneously with a frame size of 10 or 20 ms and a SWB or FB audio bandwidth. The LP layer codes the low frequencies by resampling the signal down to WB. The MDCT layer follows, coding the high frequency portion of the signal. The cutoff between the two lies at 8 kHz, the maximum WB audio bandwidth. In the MDCT layer, all bands below 8 kHz are discarded, so there is no coding redundancy between the two layers.

The sample rate (in contrast to the actual audio bandwidth) can be chosen independently on the encoder and decoder side, e.g., a fullband signal can be decoded as wideband, or vice versa. This approach ensures a sender and receiver can always interoperate, regardless of the capabilities of their actual audio hardware. Internally, the LP layer always operates at a sample rate of twice the audio bandwidth, up to a maximum of 16 kHz, which it continues to use for SWB and FB. The decoder simply resamples its output to support different sample rates. The MDCT layer always operates internally at a sample rate of 48 kHz. Since all the supported sample rates evenly divide this rate, and since the the decoder may easily zero out the high frequency portion of the spectrum in the frequency domain, it can simply decimate the MDCT layer output to achieve the other supported sample rates very cheaply.

After conversion to the common, desired output sample rate, the
decoder simply adds the output from the two layers together. To compensate for the different look-ahead required by each layer, the CELT encoder input is delayed by an additional 2.7 ms. This ensures that low frequencies and high frequencies arrive at the same time. This extra delay may be reduced by an encoder by using less look-ahead for noise shaping or using a simpler resampler in the LP layer, but this will reduce quality. However, the base 2.5 ms look-ahead in the CELT layer cannot be reduced in the encoder because it is needed for the MDCT overlap, whose size is fixed by the decoder.

Both layers use the same entropy coder, avoiding any waste from "padding bits" between them. The hybrid approach makes it easy to support both CBR and VBR coding. Although the LP layer is VBR, the bit allocation of the MDCT layer can produce a final stream that is CBR by using all the bits left unused by the LP layer.

2.1. Control Parameters

The Opus codec includes a number of control parameters which can be changed dynamically during regular operation of the codec, without interrupting the audio stream from the encoder to the decoder. These parameters only affect the encoder since any impact they have on the bit-stream is signaled in-band such that a decoder can decode any Opus stream without any out-of-band signaling. Any Opus implementation can add or modify these control parameters without affecting interoperability. The most important encoder control parameters in the reference encoder are listed below.

2.1.1. Bitrate

Opus supports all bitrates from 6 kb/s to 510 kb/s. All other parameters being equal, higher bitrate results in higher quality. For a frame size of 20 ms, these are the bitrate "sweet spots" for Opus in various configurations:

- 8-12 kb/s for NB speech,
- 16-20 kb/s for WB speech,
- 28-40 kb/s for FB speech,
- 48-64 kb/s for FB mono music, and
- 64-128 kb/s for FB stereo music.
2.1.2. Number of Channels (Mono/Stereo)

Opus can transmit either mono or stereo frames within a single stream. When decoding a mono frame in a stereo decoder, the left and right channels are identical, and when decoding a stereo frame in a mono decoder, the mono output is the average of the left and right channels. In some cases, it is desirable to encode a stereo input stream in mono (e.g., because the bitrate is too low to encode stereo with sufficient quality). The number of channels encoded can be selected in real-time, but by default the reference encoder attempts to make the best decision possible given the current bitrate.

2.1.3. Audio Bandwidth

The audio bandwidths supported by Opus are listed in Table 1. Just like for the number of channels, any decoder can decode audio encoded at any bandwidth. For example, any Opus decoder operating at 8 kHz can decode a FB Opus frame, and any Opus decoder operating at 48 kHz can decode a NB frame. Similarly, the reference encoder can take a 48 kHz input signal and encode it as NB. The higher the audio bandwidth, the higher the required bitrate to achieve acceptable quality. The audio bandwidth can be explicitly specified in real-time, but by default the reference encoder attempts to make the best bandwidth decision possible given the current bitrate.

2.1.4. Frame Duration

Opus can encode frames of 2.5, 5, 10, 20, 40 or 60 ms. It can also combine multiple frames into packets of up to 120 ms. For real-time applications, sending fewer packets per second reduces the bitrate, since it reduces the overhead from IP, UDP, and RTP headers. However, it increases latency and sensitivity to packet losses, as losing one packet constitutes a loss of a bigger chunk of audio. Increasing the frame duration also slightly improves coding efficiency, but the gain becomes small for frame sizes above 20 ms. For this reason, 20 ms frames are a good choice for most applications.

2.1.5. Complexity

There are various aspects of the Opus encoding process where trade-offs can be made between CPU complexity and quality/bitrate. In the reference encoder, the complexity is selected using an integer from 0 to 10, where 0 is the lowest complexity and 10 is the highest. Examples of computations for which such trade-offs may occur are:

- The order of the pitch analysis whitening filter [Whitening],
o The order of the short-term noise shaping filter,

o The number of states in delayed decision quantization of the residual signal, and

o The use of certain bit-stream features such as variable time-frequency resolution and the pitch post-filter.

2.1.6. Packet Loss Resilience

Audio codecs often exploit inter-frame correlations to reduce the bitrate at a cost in error propagation: after losing one packet several packets need to be received before the decoder is able to accurately reconstruct the speech signal. The extent to which Opus exploits inter-frame dependencies can be adjusted on the fly to choose a trade-off between bitrate and amount of error propagation.

2.1.7. Forward Error Correction (FEC)

Another mechanism providing robustness against packet loss is the in-band Forward Error Correction (FEC). Packets that are determined to contain perceptually important speech information, such as onsets or transients, are encoded again at a lower bitrate and this re-encoded information is added to a subsequent packet.

2.1.8. Constant/Variable Bitrate

Opus is more efficient when operating with variable bitrate (VBR), which is the default. When low-latency transmission is required over a relatively slow connection, then constrained VBR can also be used. This uses VBR in a way that simulates a "bit reservoir" and is equivalent to what MP3 (MPEG 1, Layer 3) and AAC (Advanced Audio Coding) call CBR (i.e., not true CBR due to the bit reservoir). In some (rare) applications, constant bitrate (CBR) is required. There are two main reasons to operate in CBR mode:

o When the transport only supports a fixed size for each compressed frame

o When encryption is used for an audio stream that is either highly constrained (e.g. yes/no, recorded prompts) or highly sensitive [SRTP-VBR]

Bitrate may still be allowed to vary, even with sensitive data, as long as the variation is not driven by the input signal (for example, to match changing network conditions). To achieve this, an application should still run Opus in CBR mode, but change the target rate before each packet.
2.1.9. Discontinuous Transmission (DTX)

Discontinuous Transmission (DTX) reduces the bitrate during silence or background noise. When DTX is enabled, only one frame is encoded every 400 milliseconds.
3. Internal Framing

The Opus encoder produces "packets", which are each a contiguous set of bytes meant to be transmitted as a single unit. The packets described here do not include such things as IP, UDP, or RTP headers which are normally found in a transport-layer packet. A single packet may contain multiple audio frames, so long as they share a common set of parameters, including the operating mode, audio bandwidth, frame size, and channel count (mono vs. stereo). This section describes the possible combinations of these parameters and the internal framing used to pack multiple frames into a single packet. This framing is not self-delimiting. Instead, it assumes that a lower layer (such as UDP or RTP [RFC3550] or Ogg [RFC3533] or Matroska [Matroska-website]) will communicate the length, in bytes, of the packet, and it uses this information to reduce the framing overhead in the packet itself. A decoder implementation MUST support the framing described in this section. An alternative, self-delimiting variant of the framing is described in Appendix B. Support for that variant is OPTIONAL.

All bit diagrams in this document number the bits so that bit 0 is the most significant bit of the first byte, and bit 7 is the least significant. Bit 8 is thus the most significant bit of the second byte, etc. Well-formed Opus packets obey certain requirements, marked [R1] through [R7] below. These are summarized in Section 3.4 along with appropriate means of handling malformed packets.

3.1. The TOC Byte

A well-formed Opus packet MUST contain at least one byte [R1]. This byte forms a table-of-contents (TOC) header that signals which of the various modes and configurations a given packet uses. It is composed of a configuration number, "config", a stereo flag, "s", and a frame count code, "c", arranged as illustrated in Figure 1. A description of each of these fields follows.

```
0 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+-+-
| config |s| c |
+-+-+-+-+-+-+-+-+-+-
```

Figure 1: The TOC Byte

The top five bits of the TOC byte, labeled "config", encode one of 32 possible configurations of operating mode, audio bandwidth, and frame size. As described, the LP (SILK) layer and MDCT (CELT) layer can be combined in three possible operating modes:
1. A SILK-only mode for use in low bitrate connections with an audio bandwidth of WB or less,

2. A Hybrid (SILK+CELT) mode for SWB or FB speech at medium bitrates, and

3. A CELT-only mode for very low delay speech transmission as well as music transmission (NB to FB).

The 32 possible configurations each identify which one of these operating modes the packet uses, as well as the audio bandwidth and the frame size. Table 2 lists the parameters for each configuration.

<table>
<thead>
<tr>
<th>Configuration Number(s)</th>
<th>Mode</th>
<th>Bandwidth</th>
<th>Frame Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0...3</td>
<td>SILK-only</td>
<td>NB</td>
<td>10, 20, 40, 60 ms</td>
</tr>
<tr>
<td>4...7</td>
<td>SILK-only</td>
<td>MB</td>
<td>10, 20, 40, 60 ms</td>
</tr>
<tr>
<td>8...11</td>
<td>SILK-only</td>
<td>WB</td>
<td>10, 20, 40, 60 ms</td>
</tr>
<tr>
<td>12...13</td>
<td>Hybrid</td>
<td>SWB</td>
<td>10, 20 ms</td>
</tr>
<tr>
<td>14...15</td>
<td>Hybrid</td>
<td>FB</td>
<td>10, 20 ms</td>
</tr>
<tr>
<td>16...19</td>
<td>CELT-only</td>
<td>NB</td>
<td>2.5, 5, 10, 20 ms</td>
</tr>
<tr>
<td>20...23</td>
<td>CELT-only</td>
<td>WB</td>
<td>2.5, 5, 10, 20 ms</td>
</tr>
<tr>
<td>24...27</td>
<td>CELT-only</td>
<td>SWB</td>
<td>2.5, 5, 10, 20 ms</td>
</tr>
<tr>
<td>28...31</td>
<td>CELT-only</td>
<td>FB</td>
<td>2.5, 5, 10, 20 ms</td>
</tr>
</tbody>
</table>

Table 2: TOC Byte Configuration Parameters

The configuration numbers in each range (e.g., 0...3 for NB SILK-only) correspond to the various choices of frame size, in the same order. For example, configuration 0 has a 10 ms frame size and configuration 3 has a 60 ms frame size.

One additional bit, labeled "s", signals mono vs. stereo, with 0 indicating mono and 1 indicating stereo.

The remaining two bits of the TOC byte, labeled "c", code the number of frames per packet (codes 0 to 3) as follows:
0: 1 frame in the packet
1: 2 frames in the packet, each with equal compressed size
2: 2 frames in the packet, with different compressed sizes
3: an arbitrary number of frames in the packet

This draft refers to a packet as a code 0 packet, code 1 packet, etc., based on the value of "c".

3.2. Frame Packing

This section describes how frames are packed according to each possible value of "c" in the TOC byte.

3.2.1. Frame Length Coding

When a packet contains multiple VBR frames (i.e., code 2 or 3), the compressed length of one or more of these frames is indicated with a one- or two-byte sequence, with the meaning of the first byte as follows:

0: No frame (discontinuous transmission (DTX) or lost packet)
1...251: Length of the frame in bytes
252...255: A second byte is needed. The total length is (second_byte*4)+first_byte

The special length 0 indicates that no frame is available, either because it was dropped during transmission by some intermediary or because the encoder chose not to transmit it. Any Opus frame in any mode MAY have a length of 0.

The maximum representable length is 255*4+255=1275 bytes. For 20 ms frames, this represents a bitrate of 510 kb/s, which is approximately the highest useful rate for lossily compressed fullband stereo music. Beyond this point, lossless codecs are more appropriate. It is also roughly the maximum useful rate of the MDCT layer, as quality no longer improves with additional bits due to limitations on the codebook sizes.

No length is transmitted for the last frame in a VBR packet, or for any of the frames in a CBR packet, as it can be inferred from the total size of the packet and the size of all other data in the packet. However, the length of any individual frame MUST NOT exceed 1275 bytes [R2], to allow for repacketization by gateways, conference
bridges, or other software.

3.2.2. Code 0: One Frame in the Packet

For code 0 packets, the TOC byte is immediately followed by N-1 bytes of compressed data for a single frame (where N is the size of the packet), as illustrated in Figure 2.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| config |s|0|0|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Compressed frame 1 (N-1 bytes)...          :          |
|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: A Code 0 Packet

3.2.3. Code 1: Two Frames in the Packet, Each with Equal Compressed Size

For code 1 packets, the TOC byte is immediately followed by the (N-1)/2 bytes of compressed data for the first frame, followed by (N-1)/2 bytes of compressed data for the second frame, as illustrated in Figure 3. The number of payload bytes available for compressed data, N-1, MUST be even for all code 1 packets [R3].

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| config |s|0|1|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Compressed frame 1 ((N-1)/2 bytes)... |
|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Compressed frame 2 ((N-1)/2 bytes)... |
|                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: A Code 1 Packet
3.2.4. Code 2: Two Frames in the Packet, with Different Compressed Sizes

For code 2 packets, the TOC byte is followed by a one- or two-byte sequence indicating the length of the first frame (marked N1 in Figure 4), followed by N1 bytes of compressed data for the first frame. The remaining N-N1-2 or N-N1-3 bytes are the compressed data for the second frame. This is illustrated in Figure 4. A code 2 packet MUST contain enough bytes to represent a valid length. For example, a 1-byte code 2 packet is always invalid, and a 2-byte code 2 packet whose second byte is in the range 252...255 is also invalid. The length of the first frame, N1, MUST also be no larger than the size of the payload remaining after decoding that length for all code 2 packets [R4]. This makes, for example, a 2-byte code 2 packet with a second byte in the range 1...251 invalid as well (the only valid 2-byte code 2 packet is one where the length of both frames is zero).

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| config |s|1|0| N1 (1-2 bytes):
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Compressed frame 1 (N1 bytes)...          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Compressed frame 2...                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Figure 4: A Code 2 Packet
```

3.2.5. Code 3: A Signaled Number of Frames in the Packet

Code 3 packets signal the number of frames, as well as additional padding, called "Opus padding" to indicate that this padding is added at the Opus layer, rather than at the transport layer. Code 3 packets MUST have at least 2 bytes [R6,R7]. The TOC byte is followed by a byte encoding the number of frames in the packet in bits 2 to 7 (marked "M" in Figure 5), with bit 1 indicating whether or not Opus padding is inserted (marked "p" in Figure 5), and bit 0 indicating VBR (marked "v" in Figure 5). M MUST NOT be zero, and the audio duration contained within a packet MUST NOT exceed 120 ms [R5]. This limits the maximum frame count for any frame size to 48 (for 2.5 ms frames), with lower limits for longer frame sizes. Figure 5 illustrates the layout of the frame count byte.
When Opus padding is used, the number of bytes of padding is encoded in the bytes following the frame count byte. Values from 0...254 indicate that 0...254 bytes of padding are included, in addition to the byte(s) used to indicate the size of the padding. If the value is 255, then the size of the additional padding is 254 bytes, plus the padding value encoded in the next byte. There MUST be at least one more byte in the packet in this case [R6,R7]. The additional padding bytes appear at the end of the packet, and MUST be set to zero by the encoder to avoid creating a covert channel. The decoder MUST accept any value for the padding bytes, however.

Although this encoding provides multiple ways to indicate a given number of padding bytes, each uses a different number of bytes to indicate the padding size, and thus will increase the total packet size by a different amount. For example, to add 255 bytes to a packet, set the padding bit, p, to 1, insert a single byte after the frame count byte with a value of 254, and append 254 padding bytes with the value zero to the end of the packet. To add 256 bytes to a packet, set the padding bit to 1, insert two bytes after the frame count byte with the values 255 and 0, respectively, and append 254 padding bytes with the value zero to the end of the packet. By using the value 255 multiple times, it is possible to create a packet of any specific, desired size. Let P be the number of header bytes used to indicate the padding size plus the number of padding bytes themselves (i.e., P is the total number of bytes added to the packet). Then P MUST be no more than N-2 [R6,R7].

In the CBR case, let R=N-2-P be the number of bytes remaining in the packet after subtracting the (optional) padding. Then the compressed length of each frame in bytes is equal to R/M. The value R MUST be a non-negative integer multiple of M [R6]. The compressed data for all M frames follows, each of size R/M bytes, as illustrated in Figure 6.
In the VBR case, the (optional) padding length is followed by M-1 frame lengths (indicated by \"N1\" to \"N[M-1]\" in Figure 7), each encoded in a one- or two-byte sequence as described above. The packet MUST contain enough data for the M-1 lengths after removing the (optional) padding, and the sum of these lengths MUST be no larger than the number of bytes remaining in the packet after decoding them [R7]. The compressed data for all M frames follows, each frame consisting of the indicated number of bytes, with the final frame consuming any remaining bytes before the final padding, as illustrated in Figure 6. The number of header bytes (TOC byte, frame count byte, padding length bytes, and frame length bytes), plus the signaled length of the first M-1 frames themselves, plus the signaled length of the padding MUST be no larger than N, the total size of the packet.
# Interactive Audio Codec

## 3.3. Examples

**Simplest case, one NB mono 20 ms SILK frame:**

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| config |s|1|1|1|p|     M     | Padding length (Optional)     :                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:               Compressed frame 1 (N1 bytes)...                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:               Compressed frame 2 (N2 bytes)...                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:               ...                                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:               Compressed frame M...                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:                 Opus Padding (Optional)...                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: A VBR Code 3 Packet

**Two FB mono 5 ms CELT frames of the same compressed size:**

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  compressed data...                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8
3.4. Receiving Malformed Packets

A receiver MUST NOT process packets which violate any of the rules above as normal Opus packets. They are reserved for future applications, such as in-band headers (containing metadata, etc.). Packets which violate these constraints may cause implementations of _this_ specification to treat them as malformed, and discard them.

These constraints are summarized here for reference:

[R1] Packets are at least one byte.

[R2] No implicit frame length is larger than 1275 bytes.

[R3] Code 1 packets have an odd total length, \( N \), so that \((N-1)/2\) is an integer.

[R4] Code 2 packets have enough bytes after the TOC for a valid frame length, and that length is no larger than the number of bytes remaining in the packet.
[R5] Code 3 packets contain at least one frame, but no more than 120 ms of audio total.

[R6] The length of a CBR code 3 packet, $N$, is at least two bytes, the number of bytes added to indicate the padding size plus the trailing padding bytes themselves, $P$, is no more than $N-2$, and the frame count, $M$, satisfies the constraint that $(N-2-P)$ is a non-negative integer multiple of $M$.

[R7] VBR code 3 packets are large enough to contain all the header bytes (TOC byte, frame count byte, any padding length bytes, and any frame length bytes), plus the length of the first $M-1$ frames, plus any trailing padding bytes.
4. Opus Decoder

The Opus decoder consists of two main blocks: the SILK decoder and the CELT decoder. At any given time, one or both of the SILK and CELT decoders may be active. The output of the Opus decode is the sum of the outputs from the SILK and CELT decoders with proper sample rate conversion and delay compensation on the SILK side, and optional decimation (when decoding to sample rates less than 48 kHz) on the CELT side, as illustrated in the block diagram below.

4.1. Range Decoder

Opus uses an entropy coder based on range coding [range-coding] [Martin79], which is itself a rediscovery of the FIFO arithmetic code introduced by [coding-thesis]. It is very similar to arithmetic encoding, except that encoding is done with digits in any base instead of with bits, so it is faster when using larger bases (i.e., a byte). All of the calculations in the range coder must use bit-exact integer arithmetic.

Symbols may also be coded as "raw bits" packed directly into the bitstream, bypassing the range coder. These are packed backwards starting at the end of the frame, as illustrated in Figure 12. This reduces complexity and makes the stream more resilient to bit errors, as corruption in the raw bits will not desynchronize the decoding process, unlike corruption in the input to the range decoder. Raw bits are only used in the CELT layer.
Each symbol coded by the range coder is drawn from a finite alphabet and coded in a separate "context", which describes the size of the alphabet and the relative frequency of each symbol in that alphabet.

Suppose there is a context with n symbols, identified with an index that ranges from 0 to n-1. The parameters needed to encode or decode symbol k in this context are represented by a three-tuple (fl[k], fh[k], ft), all 16-bit unsigned integers, with

\[ 0 \leq f[k] < f[h[k]] \leq ft \leq 65535. \]

The values of this tuple are derived from the probability model for the symbol, represented by traditional "frequency counts". Because Opus uses static contexts these are not updated as symbols are decoded. Let f[i] be the frequency of symbol i. Then the three-tuple corresponding to symbol k is given by

\[
\begin{align*}
    fl[k] &= \sum_{i=0}^{k-1} f[i], \\
    fh[k] &= fl[k] + f[k], \\
    ft &= \sum_{i=0}^{n-1} f[i]
\end{align*}
\]

The range decoder extracts the symbols and integers encoded using the range encoder in Section 5.1. The range decoder maintains an internal state vector composed of the two-tuple (val, rng), representing the difference between the high end of the current range and the actual coded value, minus one, and the size of the current range, respectively. Both val and rng are 32-bit unsigned integer values.

4.1.1. Range Decoder Initialization

Let b0 be an 8-bit unsigned integer containing first input byte (or containing zero if there are no bytes in this Opus frame). The
decoder initializes rng to 128 and initializes val to \((127 - (b_0>>1))\), where \((b_0>>1)\) is the top 7 bits of the first input byte. It saves the remaining bit, \((b_0&1)\), for use in the renormalization procedure described in Section 4.1.2.1, which the decoder invokes immediately after initialization to read additional bits and establish the invariant that \(\text{rng} > 2^{23}\).

4.1.2. Decoding Symbols

Decoding a symbol is a two-step process. The first step determines a 16-bit unsigned value \(fs\), which lies within the range of some symbol in the current context. The second step updates the range decoder state with the three-tuple \((fl[k], fh[k], ft)\) corresponding to that symbol.

The first step is implemented by \text{ec_decode()}\) (entdec.c), which computes

\[
\text{val} = \frac{fs}{\text{rng}} \cdot \frac{ft}{ft} - \min\left(\frac{ft}{\text{rng}} + 1, ft\right).
\]

The divisions here are integer division.

The decoder then identifies the symbol in the current context corresponding to \(fs\); i.e., the value of \(k\) whose three-tuple \((fl[k], fh[k], ft)\) satisfies \(fl[k] \leq fs < fh[k]\). It uses this tuple to update \(\text{val}\) according to

\[
\text{val} = \text{val} - \frac{\text{rng}}{ft} \cdot (ft - fh[k]).
\]

If \(fl[k]\) is greater than zero, then the decoder updates \(\text{rng}\) using

\[
\text{rng} = \frac{\text{rng}}{ft} \cdot (fh[k] - fl[k]).
\]

Otherwise, it updates \(\text{rng}\) using

\[
\text{rng} = \text{rng} - \frac{\text{rng}}{ft} \cdot (ft - fh[k]).
\]

Using a special case for the first symbol (rather than the last symbol, as is commonly done in other arithmetic coders) ensures that all the truncation error from the finite precision arithmetic
accumulates in symbol 0. This makes the cost of coding a 0 slightly smaller, on average, than its estimated probability indicates and makes the cost of coding any other symbol slightly larger. When contexts are designed so that 0 is the most probable symbol, which is often the case, this strategy minimizes the inefficiency introduced by the finite precision. It also makes some of the special-case decoding routines in Section 4.1.3 particularly simple.

After the updates, implemented by ec_dec_update() (entdec.c), the decoder normalizes the range using the procedure in the next section, and returns the index k.

4.1.2.1. Renormalization

To normalize the range, the decoder repeats the following process, implemented by ec_dec_normalize() (entdec.c), until rng > 2**23. If rng is already greater than 2**23, the entire process is skipped. First, it sets rng to (rng<<8). Then it reads the next byte of the Opus frame and forms an 8-bit value sym, using the left-over bit buffered from the previous byte as the high bit and the top 7 bits of the byte just read as the other 7 bits of sym. The remaining bit in the byte just read is buffered for use in the next iteration. If no more input bytes remain, it uses zero bits instead. See Section 4.1.1 for the initialization used to process the first byte. Then, it sets

\[ \text{val} = ((\text{val}<<8) + (255-\text{sym})) \& 0x7FFFFFFF \]

It is normal and expected that the range decoder will read several bytes into the raw bits data (if any) at the end of the packet by the time the frame is completely decoded, as illustrated in Figure 13. This same data MUST also be returned as raw bits when requested. The encoder is expected to terminate the stream in such a way that the decoder will decode the intended values regardless of the data contained in the raw bits. Section 5.1.5 describes a procedure for doing this. If the range decoder consumes all of the bytes belonging to the current frame, it MUST continue to use zero when any further input bytes are required, even if there is additional data in the current packet from padding or other frames.
4.1.3. Alternate Decoding Methods

The reference implementation uses three additional decoding methods that are exactly equivalent to the above, but make assumptions and simplifications that allow for a more efficient implementation.

4.1.3.1. ec_decode_bin()

The first is ec_decode_bin() (entdec.c), defined using the parameter ftb instead of ft. It is mathematically equivalent to calling ec_decode() with ft = (1<<ftb), but avoids one of the divisions.

4.1.3.2. ec_dec_bit_logp()

The next is ec_dec_bit_logp() (entdec.c), which decodes a single binary symbol, replacing both the ec_decode() and ec_dec_update() steps. The context is described by a single parameter, logp, which is the absolute value of the base-2 logarithm of the probability of a "1". It is mathematically equivalent to calling ec_decode() with ft = (1<<logp), followed by ec_dec_update() with the 3-tuple (fl[k] = 0, fh[k] = (1<<logp) - 1, ft = (1<<logp)) if the returned value of fs is less than (1<<logp) - 1 (a "0" was decoded), and with (fl[k] = (1<<logp) - 1, fh[k] = ft = (1<<logp)) otherwise (a "1" was decoded). The implementation requires no multiplications or divisions.

4.1.3.3. ec_dec_icdf()

The last is ec_dec_icdf() (entdec.c), which decodes a single symbol with a table-based context of up to 8 bits, also replacing both the ec_decode() and ec_dec_update() steps, as well as the search for the decoded symbol in between. The context is described by two parameters, an icdf ("inverse" cumulative distribution function).
The function is mathematically equivalent to calling ec_decode() with ft = (1<<ftb), using the returned value fs to search the table for the first entry where fs < (1<<ftb)-icdf[k], and calling ec_dec_update() with fl[k] = (1<<ftb) - icdf[k-1] (or 0 if k == 0), fh[k] = (1<<ftb) - idcf[k], and ft = (1<<ftb). Combining the search with the update allows the division to be replaced by a series of multiplications (which are usually much cheaper), and using an inverse CDF allows the use of an ftb as large as 8 in an 8-bit table without any special cases. This is the primary interface with the range decoder in the SILK layer, though it is used in a few places in the CELT layer as well.

Although icdf[k] is more convenient for the code, the frequency counts, f[k], are a more natural representation of the probability distribution function (PDF) for a given symbol. Therefore this document lists the latter, not the former, when describing the context in which a symbol is coded as a list, e.g., {4, 4, 4, 4}/16 for a uniform context with four possible values and ft = 16. The value of ft after the slash is always the sum of the entries in the PDF, but is included for convenience. Contexts with identical probabilities, f[k]/ft, but different values of ft (or equivalently, ftb) are not the same, and cannot, in general, be used in place of one another. An icdf table is also not capable of representing a PDF where the first symbol has 0 probability. In such contexts, ec_dec_icdf() can decode the symbol by using a table that drops the entries for any initial zero-probability values and adding the constant offset of the first value with a non-zero probability to its return value.

4.1.4. Decoding Raw Bits

The raw bits used by the CELT layer are packed at the end of the packet, with the least significant bit of the first value packed in the least significant bit of the last byte, filling up to the most significant bit in the last byte, continuing on to the least significant bit of the penultimate byte, and so on. The reference implementation reads them using ec_dec_bits() (entdec.c). Because the range decoder must read several bytes ahead in the stream, as described in Section 4.1.2.1, the input consumed by the raw bits may overlap with the input consumed by the range coder, and a decoder MUST allow this. The format should render it impossible to attempt to read more raw bits than there are actual bits in the frame, though a decoder may wish to check for this and report an error.
4.1.5. Decoding Uniformly Distributed Integers

The function ec_dec_uint() (entdec.c) decodes one of \( ft \) equiprobable values in the range 0 to \((ft - 1)\), inclusive, each with a frequency of 1, where \( ft \) may be as large as \((2^{**32} - 1)\). Because ec_decode() is limited to a total frequency of \((2^{**16} - 1)\), it splits up the value into a range coded symbol representing up to 8 of the high bits, and, if necessary, raw bits representing the remainder of the value. The limit of 8 bits in the range coded symbol is a trade-off between implementation complexity, modeling error (since the symbols no longer truly have equal coding cost), and rounding error introduced by the range coder itself (which gets larger as more bits are included). Using raw bits reduces the maximum number of divisions required in the worst case, but means that it may be possible to decode a value outside the range 0 to \((ft - 1)\), inclusive.

ec_dec_uint() takes a single, positive parameter, \( ft \), which is not necessarily a power of two, and returns an integer, \( t \), whose value lies between 0 and \((ft - 1)\), inclusive. Let \( ftb = ilog(ft - 1) \), i.e., the number of bits required to store \((ft - 1)\) in two’s complement notation. If \( ftb \) is 8 or less, then \( t \) is decoded with \( t = ec_decode(ft) \), and the range coder state is updated using the three-tuple \((t, t + 1, ft)\).

If \( ftb \) is greater than 8, then the top 8 bits of \( t \) are decoded using

\[
t = ec_decode(((ft - 1) >> (ftb - 8)) + 1),
\]

the decoder state is updated using the three-tuple \((t, t + 1, ((ft - 1) >> (ftb - 8)) + 1)\), and the remaining bits are decoded as raw bits, setting

\[
t = (t << (ftb - 8)) | ec_dec_bits(ftb - 8) .
\]

If, at this point, \( t \geq ft \), then the current frame is corrupt. In that case, the decoder should assume there has been an error in the coding, decoding, or transmission and SHOULD take measures to conceal the error (e.g. saturate to \( ft-1 \) or use the PLC) and/or report to the application that the error has occurred.

4.1.6. Current Bit Usage

The bit allocation routines in the CELT decoder need a conservative upper bound on the number of bits that have been used from the current frame thus far, including both range coder bits and raw bits. This drives allocation decisions that must match those made in the encoder. The upper bound is computed in the reference implementation.
to whole-bit precision by the function ec_tell() (entcode.h) and to fractional 1/8th bit precision by the function ec_tell_frac() (entcode.c). Like all operations in the range coder, it must be implemented in a bit-exact manner, and must produce exactly the same value returned by the same functions in the encoder after encoding the same symbols.

ec_tell() is guaranteed to return ceil(ec_tell_frac()/8.0). In various places the codec will check to ensure there is enough room to contain a symbol before attempting to decode it. In practice, although the number of bits used so far is an upper bound, decoding a symbol whose probability model suggests it has a worst-case cost of p 1/8th bits may actually advance the return value of ec_tell_frac() by p-1, p, or p+1 1/8th bits, due to approximation error in that upper bound, truncation error in the range coder, and for large values of ft, modeling error in ec_dec_uint().

However, this error is bounded, and periodic calls to ec_tell() or ec_tell_frac() at precisely defined points in the decoding process prevent it from accumulating. For a range coder symbol that requires a whole number of bits (i.e., for which ft/(fh[k] - fl[k]) is a power of two), where there are at least p 1/8th bits available, decoding the symbol will never cause ec_tell() or ec_tell_frac() to exceed the size of the frame ("bust the budget"). In this case the return value of ec_tell_frac() will only advance by more than p 1/8th bits if there was an additional, fractional number of bits remaining, and it will never advance beyond the next whole-bit boundary, which is safe, since frames always contain a whole number of bits. However, when p is not a whole number of bits, an extra 1/8th bit is required to ensure that decoding the symbol will not bust the budget.

The reference implementation keeps track of the total number of whole bits that have been processed by the decoder so far in the variable nbits_total, including the (possibly fractional) number of bits that are currently buffered, but not consumed, inside the range coder. nbits_total is initialized to 9 just before the initial range renormalization process completes (or equivalently, it can be initialized to 33 after the first renormalization). The extra two bits over the actual amount buffered by the range coder guarantees that it is an upper bound and that there is enough room for the encoder to terminate the stream. Each iteration through the range coder's renormalization loop increases nbits_total by 8. Reading raw bits increases nbits_total by the number of raw bits read.

4.1.6.1. ec_tell()

The whole number of bits buffered in rng may be estimated via lg = ilog(rng). ec_tell() then becomes a simple matter of removing these
bits from the total. It returns (nbits_total - lg).

In a newly initialized decoder, before any symbols have been read, this reports that 1 bit has been used. This is the bit reserved for termination of the encoder.

4.1.6.2. ec_tell_frac()

ec_tell_frac() estimates the number of bits buffered in rng to fractional precision. Since rng must be greater than 2**23 after renormalization, lg must be at least 24. Let

\[
    r_{Q15} = \text{rng} \gg (\text{lg}-16)
\]

so that 32768 ≤ r_{Q15} < 65536, an unsigned Q15 value representing the fractional part of rng. Then the following procedure can be used to add one bit of precision to lg. First, update

\[
    r_{Q15} = (r_{Q15} \times r_{Q15}) \gg 15
\]

Then add the 16th bit of r_{Q15} to lg via

\[
    \text{lg} = 2 \times \text{lg} + (r_{Q15} \gg 16)
\]

Finally, if this bit was a 1, reduce r_{Q15} by a factor of two via

\[
    r_{Q15} = r_{Q15} \gg 1
\]

so that it once again lies in the range 32768 ≤ r_{Q15} < 65536.

This procedure is repeated three times to extend lg to 1/8th bit precision. ec_tell_frac() then returns (nbits_total*8 - lg).

4.2. SILK Decoder

The decoder’s LP layer uses a modified version of the SILK codec (herein simply called "SILK"), which runs a decoded excitation signal through adaptive long-term and short-term prediction synthesis filters. It runs at NB, MB, and WB sample rates internally. When used in a SWB or FB Hybrid frame, the LP layer itself still only runs in WB.
4.2.1. SILK Decoder Modules

An overview of the decoder is given in Figure 14.

The decoder feeds the bitstream (1) to the range decoder from Section 4.1, and then decodes the parameters in it (2) using the procedures detailed in Sections 4.2.3 through 4.2.7.8.5. These parameters (3, 4, 5) are used to generate an excitation signal (see Section 4.2.7.8.6), which is fed to an optional long-term prediction (LTP) filter (voiced frames only, see Section 4.2.7.9.1) and then a short-term prediction filter (see Section 4.2.7.9.2), producing the decoded signal (6). For stereo streams, the mid-side representation is converted to separate left and right channels (7). The result is finally resampled to the desired output sample rate (e.g., 48 kHz) so that the resampled signal (8) can be mixed with the CELT layer.
4.2.2. LP Layer Organization

Internally, the LP layer of a single Opus frame is composed of either a single 10 ms regular SILK frame or between one and three 20 ms regular SILK frames. A stereo Opus frame may double the number of regular SILK frames (up to a total of six), since it includes separate frames for a mid channel and, optionally, a side channel. Optional Low Bit-Rate Redundancy (LBRR) frames, which are reduced-bitrate encodings of previous SILK frames, may be included to aid in recovery from packet loss. If present, these appear before the regular SILK frames. They are in most respects identical to regular, active SILK frames, except that they are usually encoded with a lower bitrate. This draft uses "SILK frame" to refer to either one and "regular SILK frame" if it needs to draw a distinction between the two.

Logically, each SILK frame is in turn composed of either two or four 5 ms subframes. Various parameters, such as the quantization gain of the excitation and the pitch lag and filter coefficients can vary on a subframe-by-subframe basis. Physically, the parameters for each subframe are interleaved in the bitstream, as described in the relevant sections for each parameter.

All of these frames and subframes are decoded from the same range coder, with no padding between them. Thus packing multiple SILK frames in a single Opus frame saves, on average, half a byte per SILK frame. It also allows some parameters to be predicted from prior SILK frames in the same Opus frame, since this does not degrade packet loss robustness (beyond any penalty for merely using fewer, larger packets to store multiple frames).

Stereo support in SILK uses a variant of mid-side coding, allowing a mono decoder to simply decode the mid channel. However, the data for the two channels is interleaved, so a mono decoder must still unpack the data for the side channel. It would be required to do so anyway for Hybrid Opus frames, or to support decoding individual 20 ms frames.

Table 3 summarizes the overall grouping of the contents of the LP layer. Figures 15 and 16 illustrate the ordering of the various SILK frames for a 60 ms Opus frame, for both mono and stereo, respectively.
Table 3: Organization of the SILK layer of an Opus frame

<table>
<thead>
<tr>
<th>Symbol(s)</th>
<th>PDF(s)</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice Activity Detection (VAD)</td>
<td>{1, 1}/2</td>
<td></td>
</tr>
<tr>
<td>flags</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBRR flag</td>
<td>{1, 1}/2</td>
<td></td>
</tr>
<tr>
<td>Per-frame LBRR flags</td>
<td>Table 4</td>
<td>Section 4.2.4</td>
</tr>
<tr>
<td>LBRR Frame(s)</td>
<td>Section 4.2.7</td>
<td>Section 4.2.4</td>
</tr>
<tr>
<td>Regular SILK Frame(s)</td>
<td>Section 4.2.7</td>
<td></td>
</tr>
</tbody>
</table>

Figure 15: A 60 ms Mono Frame
4.2.3. Header Bits

The LP layer begins with two to eight header bits, decoded in `silk_Decode()` (dec_API.c). These consist of one Voice Activity Detection (VAD) bit per frame (up to 3), followed by a single flag indicating the presence of LBRR frames. For a stereo packet, these first flags correspond to the mid channel, and a second set of flags is included for the side channel.

Figure 16: A 60 ms Stereo Frame
Because these are the first symbols decoded by the range coder and because they are coded as binary values with uniform probability, they can be extracted directly from the most significant bits of the first byte of compressed data. Thus, a receiver can determine if an Opus frame contains any active SILK frames without the overhead of using the range decoder.

4.2.4. Per-Frame LBRR Flags

For Opus frames longer than 20 ms, a set of LBRR flags is decoded for each channel that has its LBRR flag set. Each set contains one flag per 20 ms SILK frame. 40 ms Opus frames use the 2-frame LBRR flag PDF from Table 4, and 60 ms Opus frames use the 3-frame LBRR flag PDF. For each channel, the resulting 2- or 3-bit integer contains the corresponding LBRR flag for each frame, packed in order from the LSB to the MSB.

<table>
<thead>
<tr>
<th>Frame Size</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 ms</td>
<td>{0, 53, 53, 150}/256</td>
</tr>
<tr>
<td>60 ms</td>
<td>{0, 41, 20, 29, 41, 15, 28, 82}/256</td>
</tr>
</tbody>
</table>

Table 4: LBRR Flag PDFs

A 10 or 20 ms Opus frame does not contain any per-frame LBRR flags, as there may be at most one LBRR frame per channel. The global LBRR flag in the header bits (see Section 4.2.3) is already sufficient to indicate the presence of that single LBRR frame.

4.2.5. LBRR Frames

The LBRR frames, if present, contain an encoded representation of the signal immediately prior to the current Opus frame as if it were encoded with the current mode, frame size, audio bandwidth, and channel count, even if those differ from the prior Opus frame. When one of these parameters changes from one Opus frame to the next, this implies that the LBRR frames of the current Opus frame may not be simple drop-in replacements for the contents of the previous Opus frame.

For example, when switching from 20 ms to 60 ms, the 60 ms Opus frame may contain LBRR frames covering up to three prior 20 ms Opus frames, even if those frames already contained LBRR frames covering some of the same time periods. When switching from 20 ms to 10 ms, the 10 ms Opus frame can contain an LBRR frame covering at most half the prior
20 ms Opus frame, potentially leaving a hole that needs to be concealed from even a single packet loss (see Section 4.4). When switching from mono to stereo, the LBRR frames in the first stereo Opus frame MAY contain a non-trivial side channel.

In order to properly produce LBRR frames under all conditions, an encoder might need to buffer up to 60 ms of audio and re-encode it during these transitions. However, the reference implementation opts to disable LBRR frames at the transition point for simplicity. Since transitions are relatively infrequent in normal usage, this does not have a significant impact on packet loss robustness.

The LBRR frames immediately follow the LBRR flags, prior to any regular SILK frames. Section 4.2.7 describes their exact contents. LBRR frames do not include their own separate VAD flags. LBRR frames are only meant to be transmitted for active speech, thus all LBRR frames are treated as active.

In a stereo Opus frame longer than 20 ms, although the per-frame LBRR flags for the mid channel are coded as a unit before the per-frame LBRR flags for the side channel, the LBRR frames themselves are interleaved. The decoder parses an LBRR frame for the mid channel of a given 20 ms interval (if present) and then immediately parses the corresponding LBRR frame for the side channel (if present), before proceeding to the next 20 ms interval.

4.2.6. Regular SILK Frames

The regular SILK frame(s) follow the LBRR frames (if any). Section 4.2.7 describes their contents, as well. Unlike the LBRR frames, a regular SILK frame is coded for each time interval in an Opus frame, even if the corresponding VAD flags are unset. For stereo Opus frames longer than 20 ms, the regular mid and side SILK frames for each 20 ms interval are interleaved, just as with the LBRR frames. The side frame may be skipped by coding an appropriate flag, as detailed in Section 4.2.7.2.

4.2.7. SILK Frame Contents

Each SILK frame includes a set of side information that encodes

- The frame type and quantization type (Section 4.2.7.3),
- Quantization gains (Section 4.2.7.4),
- Short-term prediction filter coefficients (Section 4.2.7.5),
- A Line Spectral Frequencies (LSF) interpolation weight (Section 4.2.7.5.5),

- Long-term prediction filter lags and gains (Section 4.2.7.6), and

- A linear congruential generator (LCG) seed (Section 4.2.7.7).

The quantized excitation signal (see Section 4.2.7.8) follows these at the end of the frame. Table 5 details the overall organization of a SILK frame.
<table>
<thead>
<tr>
<th>Symbol(s)</th>
<th>PDF(s)</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo Prediction Weights</td>
<td>Table 6</td>
<td>Section 4.2.7.1</td>
</tr>
<tr>
<td>Mid-only Flag</td>
<td>Table 8</td>
<td>Section 4.2.7.2</td>
</tr>
<tr>
<td>Frame Type</td>
<td>Section 4.2.7.3</td>
<td></td>
</tr>
<tr>
<td>Subframe Gains</td>
<td>Section 4.2.7.4</td>
<td></td>
</tr>
<tr>
<td>Normalized LSF Stage-1</td>
<td>Table 14</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized LSF Stage-2</td>
<td>Section 4.2.7.5.2</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized LSF Interpolation Weight</td>
<td>Table 26</td>
<td>20 ms frame</td>
</tr>
<tr>
<td>Primary Pitch Lag</td>
<td>Section 4.2.7.6.1</td>
<td>Voiced frame</td>
</tr>
<tr>
<td>Subframe Pitch Contour</td>
<td>Table 32</td>
<td>Voiced frame</td>
</tr>
<tr>
<td>Periodicity Index</td>
<td>Table 37</td>
<td>Voiced frame</td>
</tr>
<tr>
<td>LTP Filter</td>
<td>Table 38</td>
<td>Voiced frame</td>
</tr>
<tr>
<td>LTP Scaling</td>
<td>Table 42</td>
<td>Section 4.2.7.6.3</td>
</tr>
<tr>
<td>LCG Seed</td>
<td>Table 43</td>
<td></td>
</tr>
<tr>
<td>Excitation Rate Level</td>
<td>Table 45</td>
<td></td>
</tr>
<tr>
<td>Excitation Pulse Counts</td>
<td>Table 46</td>
<td></td>
</tr>
<tr>
<td>Excitation Pulse Locations</td>
<td>Section 4.2.7.8.3</td>
<td>Non-zero pulse count</td>
</tr>
<tr>
<td>Excitation LSBs</td>
<td>Table 51</td>
<td>Section 4.2.7.8.2</td>
</tr>
<tr>
<td>Excitation Signs</td>
<td>Table 52</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Order of the symbols in an individual SILK frame
4.2.7.1. Stereo Prediction Weights

A SILK frame corresponding to the mid channel of a stereo Opus frame begins with a pair of side channel prediction weights, designed such that zeros indicate normal mid-side coupling. Since these weights can change on every frame, the first portion of each frame linearly interpolates between the previous weights and the current ones, using zeros for the previous weights if none are available. These prediction weights are never included in a mono Opus frame, and the previous weights are reset to zeros on any transition from mono to stereo. They are also not included in an LBRR frame for the side channel, even if the LBRR flags indicate the corresponding mid channel was not coded. In that case, the previous weights are used, again substituting in zeros if no previous weights are available since the last decoder reset (see Section 4.5.2).

To summarize, these weights are coded if and only if

- This is a stereo Opus frame (Section 3.1), and
- The current SILK frame corresponds to the mid channel.

The prediction weights are coded in three separate pieces, which are decoded by silk_stereo_decode_pred() (stereo_decode_pred.c). The first piece jointly codes the high-order part of a table index for both weights. The second piece codes the low-order part of each table index. The third piece codes an offset used to linearly interpolate between table indices. The details are as follows.

Let \( n \) be an index decoded with the 25-element stage-1 PDF in Table 6. Then let \( i_0 \) and \( i_1 \) be indices decoded with the stage-2 and stage-3 PDFs in Table 6, respectively, and let \( i_2 \) and \( i_3 \) be two more indices decoded with the stage-2 and stage-3 PDFs, all in that order.

<table>
<thead>
<tr>
<th>Stage</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>{7, 2, 1, 1, 1, 10, 24, 8, 1, 1, 3, 23, 92, 23, 3, 1, 1, 8, 24, 10, 1, 1, 1, 2, 7}/256</td>
</tr>
<tr>
<td>Stage 2</td>
<td>{85, 86, 85}/256</td>
</tr>
<tr>
<td>Stage 3</td>
<td>{51, 51, 52, 51, 51}/256</td>
</tr>
</tbody>
</table>

Table 6: Stereo Weight PDFs
Then use n, i0, and i2 to form two table indices, wi0 and wi1, according to

\[
\begin{align*}
wi0 &= i0 + 3*(n/5) \\
wi1 &= i2 + 3*(n\%5)
\end{align*}
\]

where the division is integer division. The range of these indices is 0 to 14, inclusive. Let \( w_{Q13}[i] \) be the \( i \)'th weight from Table 7. Then the two prediction weights, \( w_{0\_Q13} \) and \( w_{1\_Q13} \), are

\[
\begin{align*}
w_{1\_Q13} &= w_{Q13}[wi1] \\
&\quad + (((w_{Q13}[wi1+1] - w_{Q13}[wi1]) \times 6554) >> 16) \times (2*i3 + 1) \\
w_{0\_Q13} &= w_{Q13}[wi0] \\
&\quad + (((w_{Q13}[wi0+1] - w_{Q13}[wi0]) \times 6554) >> 16) \times (2*i1 + 1) \\
&\quad - w_{1\_Q13}
\end{align*}
\]

N.b., \( w_{1\_Q13} \) is computed first here, because \( w_{0\_Q13} \) depends on it. The constant 6554 is approximately 0.1 in Q16. Although \( wi0 \) and \( wi1 \) only have 15 possible values, Table 7 contains 16 entries to allow interpolation between entry \( wi0 \) and \( (wi0 + 1) \) (and likewise for \( wi1 \)).
<table>
<thead>
<tr>
<th>Index</th>
<th>Weight (Q13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-13732</td>
</tr>
<tr>
<td>1</td>
<td>-10050</td>
</tr>
<tr>
<td>2</td>
<td>-8266</td>
</tr>
<tr>
<td>3</td>
<td>-7526</td>
</tr>
<tr>
<td>4</td>
<td>-6500</td>
</tr>
<tr>
<td>5</td>
<td>-5000</td>
</tr>
<tr>
<td>6</td>
<td>-2950</td>
</tr>
<tr>
<td>7</td>
<td>-820</td>
</tr>
<tr>
<td>8</td>
<td>820</td>
</tr>
<tr>
<td>9</td>
<td>2950</td>
</tr>
<tr>
<td>10</td>
<td>5000</td>
</tr>
<tr>
<td>11</td>
<td>6500</td>
</tr>
<tr>
<td>12</td>
<td>7526</td>
</tr>
<tr>
<td>13</td>
<td>8266</td>
</tr>
<tr>
<td>14</td>
<td>10050</td>
</tr>
<tr>
<td>15</td>
<td>13732</td>
</tr>
</tbody>
</table>

Table 7: Stereo Weight Table

4.2.7.2. Mid-only Flag

A flag appears after the stereo prediction weights that indicates if only the mid channel is coded for this time interval. It appears only when

- This is a stereo Opus frame (see Section 3.1),
- The current SILK frame corresponds to the mid channel, and
Either

* This is a regular SILK frame where the VAD flags (see Section 4.2.3) indicate that the corresponding side channel is not active.

* This is an LBRR frame where the LBRR flags (see Section 4.2.3 and Section 4.2.4) indicate that the corresponding side channel is not coded.

It is omitted when there are no stereo weights, for all of the same reasons. It is also omitted for a regular SILK frame when the VAD flag of the corresponding side channel frame is set (indicating it is active). The side channel must be coded in this case, making the mid-only flag redundant. It is also omitted for an LBRR frame when the corresponding LBRR flags indicate the side channel is coded.

When the flag is present, the decoder reads a single value using the PDF in Table 8, as implemented in silk_stereo_decode_mid_only() (stereo_decode_pred.c). If the flag is set, then there is no corresponding SILK frame for the side channel, the entire decoding process for the side channel is skipped, and zeros are fed to the stereo unmixing process (see Section 4.2.8) instead. As stated above, LBRR frames still include this flag when the LBRR flag indicates that the side channel is not coded. In that case, if this flag is zero (indicating that there should be a side channel), then Packet Loss Concealment (PLC, see Section 4.4) SHOULD be invoked to recover a side channel signal. Otherwise, the stereo image will collapse.

```
+---------------+
| PDF           |
+---------------+
| {192, 64}/256 |
+---------------+
```

Table 8: Mid-only Flag PDF

### 4.2.7.3. Frame Type

Each SILK frame contains a single "frame type" symbol that jointly codes the signal type and quantization offset type of the corresponding frame. If the current frame is a regular SILK frame whose VAD bit was not set (an "inactive" frame), then the frame type symbol takes on a value of either 0 or 1 and is decoded using the first PDF in Table 9. If the frame is an LBRR frame or a regular SILK frame whose VAD flag was set (an "active" frame), then the value of the symbol may range from 2 to 5, inclusive, and is decoded using
the second PDF in Table 9. Table 10 translates between the value of the frame type symbol and the corresponding signal type and quantization offset type.

<table>
<thead>
<tr>
<th>VAD Flag</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>{26, 230, 0, 0, 0, 0}/256</td>
</tr>
<tr>
<td>Active</td>
<td>{0, 0, 24, 74, 148, 10}/256</td>
</tr>
</tbody>
</table>

Table 9: Frame Type PDFs

<table>
<thead>
<tr>
<th>Frame Type</th>
<th>Signal Type</th>
<th>Quantization Offset Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Inactive</td>
<td>Low</td>
</tr>
<tr>
<td>1</td>
<td>Inactive</td>
<td>High</td>
</tr>
<tr>
<td>2</td>
<td>Unvoiced</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>Unvoiced</td>
<td>High</td>
</tr>
<tr>
<td>4</td>
<td>Voiced</td>
<td>Low</td>
</tr>
<tr>
<td>5</td>
<td>Voiced</td>
<td>High</td>
</tr>
</tbody>
</table>

Table 10: Signal Type and Quantization Offset Type from Frame Type

4.2.7.4. Subframe Gains

A separate quantization gain is coded for each 5 ms subframe. These gains control the step size between quantization levels of the excitation signal and, therefore, the quality of the reconstruction. They are independent of and unrelated to the pitch contours coded for voiced frames. The quantization gains are themselves uniformly quantized to 6 bits on a log scale, giving them a resolution of approximately 1.369 dB and a range of approximately 1.94 dB to 88.21 dB.

The subframe gains are either coded independently, or relative to the gain from the most recent coded subframe in the same channel. Independent coding is used if and only if
This is the first subframe in the current SILK frame, and

Either

* This is the first SILK frame of its type (LBRR or regular) for this channel in the current Opus frame, or

* The previous SILK frame of the same type (LBRR or regular) for this channel in the same Opus frame was not coded.

In an independently coded subframe gain, the 3 most significant bits of the quantization gain are decoded using a PDF selected from Table 11 based on the decoded signal type (see Section 4.2.7.3).

<table>
<thead>
<tr>
<th>Signal Type</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>(32, 112, 68, 29, 12, 1, 1, 1)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>(2, 17, 45, 60, 62, 47, 19, 4)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>(1, 3, 26, 71, 94, 50, 9, 2)/256</td>
</tr>
</tbody>
</table>

Table 11: PDFs for Independent Quantization Gain MSB Coding

The 3 least significant bits are decoded using a uniform PDF:

<table>
<thead>
<tr>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(32, 32, 32, 32, 32, 32, 32, 32)/256</td>
</tr>
</tbody>
</table>

Table 12: PDF for Independent Quantization Gain LSB Coding

These 6 bits are combined to form a value, gain_index, between 0 and 63. When the gain for the previous subframe is available, then the current gain is limited as follows:

\[
\text{log}_2\text{gain} = \max(\text{gain_index}, \text{previous_log_gain} - 16) .
\]

This may help some implementations limit the change in precision of their internal LTP history. The indices which this clamp applies to cannot simply be removed from the codebook, because previous_log_gain will not be available after packet loss. The clamping is skipped after a decoder reset, and in the side channel if the previous frame in the side channel was not coded, since there is no value for

previous_log_gain available. It MAY also be skipped after packet loss.

For subframes which do not have an independent gain (including the first subframe of frames not listed as using independent coding above), the quantization gain is coded relative to the gain from the previous subframe (in the same channel). The PDF in Table 13 yields a delta_gain_index value between 0 and 40, inclusive.

<table>
<thead>
<tr>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>{6, 5, 11, 31, 132, 21, 8, 4, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 13: PDF for Delta Quantization Gain Coding

The following formula translates this index into a quantization gain for the current subframe using the gain from the previous subframe:

\[
\log_{\text{gain}} = \text{clamp}(0, \max(2*\text{delta_gain_index} - 16, \text{previous_log_gain} + \text{delta_gain_index} - 4), 63) .
\]

silk_gains_dequant() (gain_quant.c) dequantizes log_gain for the k’th subframe and converts it into a linear Q16 scale factor via

\[
\text{gain}_{Q16}[k] = \text{silk_log2lin}((0x1D1C71*\text{log_gain}>>16) + 2090)
\]

The function silk_log2lin() (log2lin.c) computes an approximation of \(2^{\text{inLog}_{Q7}/128.0}\), where inLog_Q7 is its Q7 input. Let \(i = \text{inLog}_{Q7}>>7\) be the integer part of inLogQ7 and \(f = \text{inLog}_{Q7}\&127\) be the fractional part. Then

\[
(1<<i) + ((-174*f*(128-f)>>16)+f)*((1<<i)>>7)
\]

yields the approximate exponential. The final Q16 gain values lies between 81920 and 168610208, inclusive (representing scale factors of 1.25 to 25728, respectively).

4.2.7.5. Normalized Line Spectral Frequency (LSF) and Linear Predictive Coding (LPC) Coefficients

A set of normalized Line Spectral Frequency (LSF) coefficients follow the quantization gains in the bitstream, and represent the Linear Predictive Coding (LPC) coefficients for the current SILK frame. Once decoded, the normalized LSFs form an increasing list of Q15
values between 0 and 1. These represent the interleaved zeros on the upper half of the unit circle (between 0 and ρi, hence "normalized") in the standard decomposition [line-spectral-pairs] of the LPC filter into a symmetric part and an anti-symmetric part (P and Q in Section 4.2.7.5.6). Because of non-linear effects in the decoding process, an implementation SHOULD match the fixed-point arithmetic described in this section exactly. An encoder SHOULD also use the same process.

The normalized LSFs are coded using a two-stage vector quantizer (VQ) (Section 4.2.7.5.1 and Section 4.2.7.5.2). NB and MB frames use an order-10 predictor, while WB frames use an order-16 predictor, and thus have different sets of tables. After reconstructing the normalized LSFs (Section 4.2.7.5.3), the decoder runs them through a stabilization process (Section 4.2.7.5.4), interpolates them between frames (Section 4.2.7.5.5), converts them back into LPC coefficients (Section 4.2.7.5.6), and then runs them through further processes to limit the range of the coefficients (Section 4.2.7.5.7) and the gain of the filter (Section 4.2.7.5.8). All of this is necessary to ensure the reconstruction process is stable.

4.2.7.5.1. Normalized LSF Stage 1 Decoding

The first VQ stage uses a 32-element codebook, coded with one of the PDFs in Table 14, depending on the audio bandwidth and the signal type of the current SILK frame. This yields a single index, I1, for the entire frame, which

1. Indexes an element in a coarse codebook,

2. Selects the PDFs for the second stage of the VQ, and

3. Selects the prediction weights used to remove intra-frame redundancy from the second stage.

The actual codebook elements are listed in Table 23 and Table 24, but they are not needed until the last stages of reconstructing the LSF coefficients.
### Table 14: PDFs for Normalized LSF Stage-1 Index Decoding

<table>
<thead>
<tr>
<th>Audio Bandwidth</th>
<th>Signal Type</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB or MB</td>
<td>Inactive or unvoiced</td>
<td>(44, 34, 30, 19, 21, 12, 11, 3, 3, 2, 16, 2, 2, 1, 5, 2, 1, 3, 3, 1, 1, 2, 2, 2, 3, 1, 9, 9, 2, 7, 2, 1)/256</td>
</tr>
<tr>
<td>NB or MB</td>
<td>Voiced</td>
<td>(1, 10, 1, 8, 3, 8, 8, 14, 13, 14, 1, 14, 12, 13, 11, 11, 12, 11, 10, 10, 11, 10, 8, 9, 8, 7, 8, 1, 1, 6, 1, 6, 5)/256</td>
</tr>
<tr>
<td>WB</td>
<td>Inactive or unvoiced</td>
<td>(31, 21, 3, 17, 1, 8, 17, 4, 1, 18, 16, 4, 2, 3, 1, 10, 1, 3, 16, 11, 16, 2, 2, 3, 2, 11, 1, 4, 9, 8, 7, 3)/256</td>
</tr>
<tr>
<td>WB</td>
<td>Voiced</td>
<td>(1, 4, 16, 5, 18, 11, 5, 14, 15, 1, 3, 12, 13, 14, 14, 6, 14, 12, 2, 6, 1, 12, 12, 11, 10, 3, 10, 5, 1, 1, 1, 3)/256</td>
</tr>
</tbody>
</table>

4.2.7.5.2. Normalized LSF Stage 2 Decoding

A total of 16 PDFs are available for the LSF residual in the second stage: the 8 (a...h) for NB and MB frames given in Table 15, and the 8 (i...p) for WB frames given in Table 16. Which PDF is used for which coefficient is driven by the index, \( I_1 \), decoded in the first stage. Table 17 lists the letter of the corresponding PDF for each normalized LSF coefficient for NB and MB, and Table 18 lists the same information for WB.
<table>
<thead>
<tr>
<th>Codebook</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>{1, 1, 1, 15, 224, 11, 1, 1, 1}/256</td>
</tr>
<tr>
<td>b</td>
<td>{1, 1, 2, 34, 183, 32, 1, 1, 1}/256</td>
</tr>
<tr>
<td>c</td>
<td>{1, 1, 4, 42, 149, 55, 2, 1, 1}/256</td>
</tr>
<tr>
<td>d</td>
<td>{1, 1, 8, 52, 123, 61, 8, 1, 1}/256</td>
</tr>
<tr>
<td>e</td>
<td>{1, 3, 16, 53, 101, 74, 6, 1, 1}/256</td>
</tr>
<tr>
<td>f</td>
<td>{1, 3, 17, 55, 90, 73, 15, 1, 1}/256</td>
</tr>
<tr>
<td>g</td>
<td>{1, 7, 24, 53, 74, 67, 26, 3, 1}/256</td>
</tr>
<tr>
<td>h</td>
<td>{1, 1, 18, 63, 78, 58, 30, 6, 1}/256</td>
</tr>
</tbody>
</table>

Table 15: PDFs for NB/MB Normalized LSF Stage-2 Index Decoding

<table>
<thead>
<tr>
<th>Codebook</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>{1, 1, 1, 9, 232, 9, 1, 1, 1}/256</td>
</tr>
<tr>
<td>j</td>
<td>{1, 1, 2, 28, 186, 35, 1, 1, 1}/256</td>
</tr>
<tr>
<td>k</td>
<td>{1, 1, 3, 42, 152, 53, 2, 1, 1}/256</td>
</tr>
<tr>
<td>l</td>
<td>{1, 1, 10, 49, 126, 65, 2, 1, 1}/256</td>
</tr>
<tr>
<td>m</td>
<td>{1, 4, 19, 48, 100, 77, 5, 1, 1}/256</td>
</tr>
<tr>
<td>n</td>
<td>{1, 1, 14, 54, 100, 72, 12, 1, 1}/256</td>
</tr>
<tr>
<td>o</td>
<td>{1, 1, 15, 61, 87, 61, 25, 4, 1}/256</td>
</tr>
<tr>
<td>p</td>
<td>{1, 7, 21, 50, 77, 81, 17, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 16: PDFs for WB Normalized LSF Stage-2 Index Decoding

<table>
<thead>
<tr>
<th>I1</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td></td>
<td>a a a a a a a a a a</td>
</tr>
<tr>
<td>---</td>
<td>---------------------</td>
</tr>
<tr>
<td>1</td>
<td>b d b c c b c b b b</td>
</tr>
<tr>
<td>2</td>
<td>c b b b b b b b b b</td>
</tr>
<tr>
<td>3</td>
<td>b c c c c b c b b b</td>
</tr>
<tr>
<td>4</td>
<td>c d d d c c c c c</td>
</tr>
<tr>
<td>5</td>
<td>a f d d c c c c b b</td>
</tr>
<tr>
<td>6</td>
<td>a c c c c c c c c b</td>
</tr>
<tr>
<td>7</td>
<td>c d g e e e f e f f</td>
</tr>
<tr>
<td>8</td>
<td>c e f f e f e g e e</td>
</tr>
<tr>
<td>9</td>
<td>c e e h e f e f f e</td>
</tr>
<tr>
<td>10</td>
<td>e d d d c d c c c c</td>
</tr>
<tr>
<td>11</td>
<td>b f f g e f e f f f</td>
</tr>
<tr>
<td>12</td>
<td>c h e g f f f f f f</td>
</tr>
<tr>
<td>13</td>
<td>c h f f f f f g f e</td>
</tr>
<tr>
<td>14</td>
<td>d d f e e f e f e e</td>
</tr>
<tr>
<td>15</td>
<td>c d d f e e e e e e</td>
</tr>
<tr>
<td>16</td>
<td>c e e g e f e f f f</td>
</tr>
<tr>
<td>17</td>
<td>c f e g f f f e f e</td>
</tr>
<tr>
<td>18</td>
<td>c h e f e f e f f f</td>
</tr>
<tr>
<td>19</td>
<td>c f e g h g f g f e</td>
</tr>
<tr>
<td>20</td>
<td>d g h e g f f g e f</td>
</tr>
<tr>
<td>21</td>
<td>c h g e e e f e f f</td>
</tr>
<tr>
<td>22</td>
<td>e f f e g g f g f e</td>
</tr>
<tr>
<td>23</td>
<td>c f f g f g e g e e</td>
</tr>
</tbody>
</table>
Table 17: Codebook Selection for NB/MB Normalized LSF Stage-2 Index Decoding

<table>
<thead>
<tr>
<th>I</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>i i i i i i i i i i i i i i i i</td>
</tr>
<tr>
<td>1</td>
<td>k l l l l l k k k k j j j j i</td>
</tr>
<tr>
<td>2</td>
<td>k n n l p m m n k n m n n m l l</td>
</tr>
<tr>
<td>3</td>
<td>i k j k k j j j j i i i i i j</td>
</tr>
<tr>
<td>4</td>
<td>i o n m o m p n m m n m m m m l</td>
</tr>
<tr>
<td>5</td>
<td>i l n n m l l n l l l l l l k m</td>
</tr>
<tr>
<td>6</td>
<td>i i i i i i i i i i i i i i i i i</td>
</tr>
<tr>
<td>7</td>
<td>i k o l p k n l m n n m l l k l</td>
</tr>
<tr>
<td>8</td>
<td>i o k o o m n m o m m n l l l</td>
</tr>
<tr>
<td>9</td>
<td>k j i i i i i i i i i i i i i i i i</td>
</tr>
<tr>
<td>10</td>
<td>i j i i i i i i i i i i i i i i j</td>
</tr>
<tr>
<td>11</td>
<td>k k l m n l l l l l l k k j l</td>
</tr>
</tbody>
</table>
Decoding the second stage residual proceeds as follows. For each coefficient, the decoder reads a symbol using the PDF corresponding to \( I_1 \) from either Table 17 or Table 18, and subtracts 4 from the result to give an index in the range -4 to 4, inclusive. If the
index is either -4 or 4, it reads a second symbol using the PDF in Table 19, and adds the value of this second symbol to the index, using the same sign. This gives the index, I2[k], a total range of -10 to 10, inclusive.

+-------------------------------+
| PDF                          |
+-------------------------------+
| (156, 60, 24, 9, 4, 2, 1)/256 |
+-------------------------------+

Table 19: PDF for Normalized LSF Index Extension Decoding

The decoded indices from both stages are translated back into normalized LSF coefficients in silk_NLSF_decode() (NLSF_decode.c). The stage-2 indices represent residuals after both the first stage of the VQ and a separate backwards-prediction step. The backwards prediction process in the encoder subtracts a prediction from each residual formed by a multiple of the coefficient that follows it. The decoder must undo this process. Table 20 contains lists of prediction weights for each coefficient. There are two lists for NB and MB, and another two lists for WB, giving two possible prediction weights for each coefficient.
The prediction is undone using the procedure implemented in
silk_NLSF_residual_dequant() (NLSF_decode.c), which is as follows.
Each coefficient selects its prediction weight from one of the two
lists based on the stage-1 index, I1. Table 21 gives the selections
for each coefficient for NB and MB, and Table 22 gives the selections
for WB. Let d_LPC be the order of the codebook, i.e., 10 for NB and
MB, and 16 for WB, and let pred_Q8[k] be the weight for the k'th
coefficient selected by this process for 0 <= k < d_LPC-1. Then, the
stage-2 residual for each coefficient is computed via

\[
\text{res}_Q10[k] = (k+1 < d_{\text{LPC}} \ ? \ (\text{res}_Q10[k+1] \cdot \text{pred}_Q8[k]) >> 8 : 0) \\
+ (((I2[k] << 10) - \text{sign}(I2[k]) \cdot 102) \cdot \text{qstep}) >> 16)
\]

Table 20: Prediction Weights for Normalized LSF Decoding

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>179</td>
<td>116</td>
<td>175</td>
<td>68</td>
</tr>
<tr>
<td>1</td>
<td>138</td>
<td>67</td>
<td>148</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>140</td>
<td>82</td>
<td>160</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>148</td>
<td>59</td>
<td>176</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>151</td>
<td>92</td>
<td>178</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>149</td>
<td>72</td>
<td>173</td>
<td>117</td>
</tr>
<tr>
<td>6</td>
<td>153</td>
<td>100</td>
<td>174</td>
<td>85</td>
</tr>
<tr>
<td>7</td>
<td>151</td>
<td>89</td>
<td>164</td>
<td>90</td>
</tr>
<tr>
<td>8</td>
<td>163</td>
<td>92</td>
<td>177</td>
<td>118</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>174</td>
<td>136</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>196</td>
<td>151</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td>182</td>
<td>142</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td>198</td>
<td>160</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td>192</td>
<td>142</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td>182</td>
<td>155</td>
</tr>
</tbody>
</table>

where qstep is the Q16 quantization step size, which is 11796 for NB and MB and 9830 for WB (representing step sizes of approximately 0.18 and 0.15, respectively).

<table>
<thead>
<tr>
<th>11</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5 6 7 8</td>
<td>-------------</td>
</tr>
<tr>
<td>A B A A A A A A A</td>
<td></td>
</tr>
<tr>
<td>B A A A A A A A A</td>
<td></td>
</tr>
<tr>
<td>A A A A A A A A A</td>
<td></td>
</tr>
<tr>
<td>B B B A A A A B A</td>
<td></td>
</tr>
<tr>
<td>A B A A A A A A A</td>
<td></td>
</tr>
<tr>
<td>B A B B A A A B A</td>
<td></td>
</tr>
<tr>
<td>A B B A A B B A A</td>
<td></td>
</tr>
<tr>
<td>A A B B A B A B B</td>
<td></td>
</tr>
<tr>
<td>A A B B A A B B B</td>
<td></td>
</tr>
<tr>
<td>A A A A A A A A A</td>
<td></td>
</tr>
<tr>
<td>A B A B B B B B A</td>
<td></td>
</tr>
<tr>
<td>A B A B B B B B A</td>
<td></td>
</tr>
<tr>
<td>A B B B A B B B A</td>
<td></td>
</tr>
<tr>
<td>A B B B A A B B B</td>
<td></td>
</tr>
<tr>
<td>A B B A A B A B A</td>
<td></td>
</tr>
<tr>
<td>A A B B A A B A B</td>
<td></td>
</tr>
<tr>
<td>A A A A A B B B A</td>
<td></td>
</tr>
<tr>
<td>A A B B A A A A A</td>
<td></td>
</tr>
</tbody>
</table>

Table 21: Prediction Weight Selection for NB/MB Normalized LSF Decoding

<table>
<thead>
<tr>
<th>I1</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>C C C C C C C C C C C C C D</td>
</tr>
<tr>
<td>1</td>
<td>C C C C C C C C C C C C C C</td>
</tr>
<tr>
<td>2</td>
<td>C C D C C D D D C D D D C C</td>
</tr>
<tr>
<td>3</td>
<td>C C C C C C C C C C C C C C</td>
</tr>
<tr>
<td>4</td>
<td>C D D C D C D D C D D D D C</td>
</tr>
<tr>
<td>5</td>
<td>C C D C C C C C C C C C C C</td>
</tr>
<tr>
<td>6</td>
<td>D C C C C C C C C C C C D C D</td>
</tr>
<tr>
<td>7</td>
<td>C D D C C C D D D C D C D C D</td>
</tr>
</tbody>
</table>
Reconstructing the Normalized LSF Coefficients

Once the stage-1 index \( I1 \) and the stage-2 residual \( \text{res}_Q10[] \) have been decoded, the final normalized LSF coefficients can be reconstructed.

The spectral distortion introduced by the quantization of each LSF coefficient varies, so the stage-2 residual is weighted accordingly, using the low-complexity Inverse Harmonic Mean Weighting (IHMW) function proposed in [laroia-icassp]. The weights are derived directly from the stage-1 codebook vector. Let \( \text{cb1}_Q8[k] \) be the \( k \)'th entry of the stage-1 codebook vector from Table 23 or Table 24. Then for \( 0 \leq k < d_LPC \) the following expression computes the square of the weight as a Q18 value:

\[
\text{w}_2_{Q18}[k] = \left( \frac{1024}{\text{cb1}_Q8[k] - \text{cb1}_Q8[k-1]} + \frac{1024}{\text{cb1}_Q8[k+1] - \text{cb1}_Q8[k]} \right) \ll 16 ,
\]

where \( \text{cb1}_Q8[-1] = 0 \) and \( \text{cb1}_Q8[d_LPC] = 256 \), and the division is integer division. This is reduced to an unsquared, Q9 value using the following square-root approximation:

\[
i = \text{ilog(}w_2_{Q18}[k]) \\
f = (w_2_{Q18}[k]>>(i-8)) \ & \ 127 \\
y = ((i\&1) \ ? \ 32768 : 46214) \ >> ((32-i)>>1) \\
w_9[k] = y + ((213*f*y)>>16)
\]

The constant 46214 here is approximately the square root of 2 in Q15. The \( \text{cb1}_Q8[] \) vector completely determines these weights, and they may be tabulated and stored as 13-bit unsigned values (with a range of 1819 to 5227, inclusive) to avoid computing them when decoding. The reference implementation already requires code to compute these weights on unquantized coefficients in the encoder, in \text{silknLSF_VQ_weights_laroia}() (NLSF_VQ_weights_laroia.c) and its callers, so it reuses that code in the decoder instead of using a pre-computed table to reduce the amount of ROM required.
<table>
<thead>
<tr>
<th>1</th>
<th>15</th>
<th>32</th>
<th>55</th>
<th>77</th>
<th>101</th>
<th>125</th>
<th>151</th>
<th>175</th>
<th>201</th>
<th>225</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>19</td>
<td>42</td>
<td>66</td>
<td>89</td>
<td>114</td>
<td>137</td>
<td>162</td>
<td>184</td>
<td>209</td>
<td>230</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>25</td>
<td>50</td>
<td>72</td>
<td>97</td>
<td>120</td>
<td>147</td>
<td>172</td>
<td>200</td>
<td>223</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>44</td>
<td>69</td>
<td>90</td>
<td>114</td>
<td>135</td>
<td>159</td>
<td>180</td>
<td>205</td>
<td>225</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>22</td>
<td>53</td>
<td>80</td>
<td>106</td>
<td>130</td>
<td>156</td>
<td>180</td>
<td>205</td>
<td>228</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>25</td>
<td>44</td>
<td>64</td>
<td>90</td>
<td>115</td>
<td>142</td>
<td>168</td>
<td>196</td>
<td>222</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>24</td>
<td>62</td>
<td>82</td>
<td>100</td>
<td>120</td>
<td>145</td>
<td>168</td>
<td>190</td>
<td>214</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>31</td>
<td>50</td>
<td>79</td>
<td>103</td>
<td>120</td>
<td>151</td>
<td>170</td>
<td>203</td>
<td>227</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>29</td>
<td>45</td>
<td>65</td>
<td>106</td>
<td>124</td>
<td>150</td>
<td>171</td>
<td>196</td>
<td>224</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>49</td>
<td>75</td>
<td>97</td>
<td>121</td>
<td>142</td>
<td>165</td>
<td>186</td>
<td>209</td>
<td>229</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>25</td>
<td>52</td>
<td>70</td>
<td>93</td>
<td>116</td>
<td>143</td>
<td>166</td>
<td>192</td>
<td>219</td>
</tr>
<tr>
<td>12</td>
<td>26</td>
<td>34</td>
<td>62</td>
<td>75</td>
<td>97</td>
<td>118</td>
<td>145</td>
<td>167</td>
<td>194</td>
<td>217</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
<td>33</td>
<td>56</td>
<td>70</td>
<td>91</td>
<td>113</td>
<td>143</td>
<td>165</td>
<td>196</td>
<td>223</td>
</tr>
<tr>
<td>14</td>
<td>21</td>
<td>34</td>
<td>51</td>
<td>72</td>
<td>97</td>
<td>117</td>
<td>145</td>
<td>171</td>
<td>196</td>
<td>222</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>29</td>
<td>50</td>
<td>67</td>
<td>90</td>
<td>117</td>
<td>144</td>
<td>168</td>
<td>197</td>
<td>221</td>
</tr>
<tr>
<td>16</td>
<td>22</td>
<td>31</td>
<td>48</td>
<td>66</td>
<td>95</td>
<td>117</td>
<td>146</td>
<td>168</td>
<td>196</td>
<td>222</td>
</tr>
<tr>
<td>17</td>
<td>24</td>
<td>33</td>
<td>51</td>
<td>77</td>
<td>116</td>
<td>134</td>
<td>158</td>
<td>180</td>
<td>200</td>
<td>224</td>
</tr>
<tr>
<td>18</td>
<td>21</td>
<td>28</td>
<td>70</td>
<td>87</td>
<td>106</td>
<td>124</td>
<td>149</td>
<td>170</td>
<td>194</td>
<td>217</td>
</tr>
<tr>
<td>19</td>
<td>26</td>
<td>33</td>
<td>53</td>
<td>64</td>
<td>83</td>
<td>117</td>
<td>152</td>
<td>173</td>
<td>204</td>
<td>225</td>
</tr>
<tr>
<td>20</td>
<td>27</td>
<td>34</td>
<td>65</td>
<td>95</td>
<td>108</td>
<td>129</td>
<td>155</td>
<td>174</td>
<td>210</td>
<td>225</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>26</td>
<td>72</td>
<td>99</td>
<td>113</td>
<td>131</td>
<td>154</td>
<td>176</td>
<td>200</td>
<td>219</td>
</tr>
<tr>
<td>22</td>
<td>34</td>
<td>43</td>
<td>61</td>
<td>78</td>
<td>93</td>
<td>114</td>
<td>155</td>
<td>177</td>
<td>205</td>
<td>229</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>29</td>
<td>54</td>
<td>97</td>
<td>124</td>
<td>138</td>
<td>163</td>
<td>179</td>
<td>209</td>
<td>229</td>
</tr>
<tr>
<td>24</td>
<td>30</td>
<td>38</td>
<td>56</td>
<td>89</td>
<td>118</td>
<td>129</td>
<td>158</td>
<td>178</td>
<td>200</td>
<td>231</td>
</tr>
</tbody>
</table>
Table 23: NB/MB Normalized LSF Stage-1 Codebook Vectors

<table>
<thead>
<tr>
<th>I1</th>
<th>Codebook (Q8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15</td>
</tr>
<tr>
<td>0</td>
<td>7 23 38 54 69 85 100 116 131 147 162 178 193 208 223 239</td>
</tr>
<tr>
<td>1</td>
<td>13 25 41 55 69 83 98 112 127 142 157 171 187 203 220 236</td>
</tr>
<tr>
<td>2</td>
<td>15 21 34 51 61 78 92 106 126 136 152 167 185 205 225 240</td>
</tr>
<tr>
<td>3</td>
<td>10 21 36 50 63 79 95 110 126 141 157 173 189 205 221 237</td>
</tr>
<tr>
<td>4</td>
<td>17 20 37 51 59 78 89 107 123 134 150 164 184 205 224 240</td>
</tr>
<tr>
<td>5</td>
<td>10 15 32 51 67 81 96 112 129 142 158 173 189 204 220 236</td>
</tr>
<tr>
<td>6</td>
<td>8 21 37 51 65 79 98 113 126 138 155 168 179 192 209 218</td>
</tr>
<tr>
<td>7</td>
<td>12 15 34 55 63 78 87 108 118 131 148 167 185 203 219 236</td>
</tr>
<tr>
<td>8</td>
<td>16 19 32 36 56 79 91 108 118 136 154 171 186 204 220 237</td>
</tr>
<tr>
<td>9</td>
<td>11 28 43 58 74 89 105 120 135 150 165 180 196 211 226 241</td>
</tr>
<tr>
<td>10</td>
<td>6 16 33 46 60 75 92 107 123 137 156 169 185 199 214 225</td>
</tr>
<tr>
<td>11</td>
<td>11 19 30 44 57 74 89 105 121 135 152 169 186 202 218 234</td>
</tr>
<tr>
<td>12</td>
<td>12 19 29 46 57 71 88 100 120 132 148 165 182 199 216 233</td>
</tr>
</tbody>
</table>
Given the stage-1 codebook entry cb1_Q8[], the stage-2 residual res_Q10[], and their corresponding weights, w_Q9[], the reconstructed normalized LSF coefficients are

\[
NLSF_Q15[k] = \text{clamp}(0, (cb1_Q8[k]<<7) + (res_Q10[k]<<14)/w_Q9[k], 32767),
\]

Table 24: WB Normalized LSF Stage-1 Codebook Vectors

| 13 | 17 23 35 46 56 77 92 106 123 134 152 167 185 204 222 237 |
| 14 | 14 17 45 53 63 75 89 107 115 132 151 171 188 206 221 240 |
| 15 | 9 16 29 40 56 71 88 103 119 137 154 171 189 205 222 237 |
| 16 | 16 19 36 48 57 76 87 105 118 132 150 167 185 202 218 236 |
| 17 | 12 17 29 54 71 81 94 104 126 136 149 164 182 201 221 237 |
| 18 | 15 28 47 62 79 97 115 129 142 155 168 180 194 208 223 238 |
| 19 | 8 14 30 45 62 78 94 111 127 143 159 175 192 207 223 239 |
| 20 | 17 30 49 62 79 92 107 119 132 145 160 174 190 204 220 235 |
| 21 | 14 19 36 45 61 76 91 108 121 138 154 172 189 205 222 238 |
| 22 | 12 18 31 45 60 76 91 107 123 138 154 171 187 204 221 236 |
| 23 | 13 17 31 43 53 70 83 103 114 131 149 167 185 203 220 237 |
| 24 | 17 22 35 42 58 78 93 110 125 139 155 170 188 206 224 240 |
| 25 | 8 15 34 50 67 83 99 115 131 146 162 178 193 209 224 239 |
| 26 | 13 16 41 66 73 86 95 111 128 137 150 163 183 206 225 241 |
| 27 | 17 25 37 52 63 75 92 102 119 132 144 160 175 191 212 231 |
| 28 | 19 31 49 65 83 100 117 133 147 161 174 187 200 213 227 242 |
| 29 | 18 31 52 68 88 103 117 126 138 149 163 177 192 207 223 239 |
| 30 | 16 29 47 61 76 90 106 119 133 147 161 176 193 209 224 240 |
| 31 | 15 21 35 50 61 73 86 97 110 119 129 141 175 198 218 237 |
where the division is integer division. However, nothing in either
the reconstruction process or the quantization process in the encoder
thus far guarantees that the coefficients are monotonically
increasing and separated well enough to ensure a stable filter
[Kabal86]. When using the reference encoder, roughly 2% of frames
violate this constraint. The next section describes a stabilization
procedure used to make these guarantees.

4.2.7.5.4. Normalized LSF Stabilization

The normalized LSF stabilization procedure is implemented in
\texttt{silk\_NLSF\_stabilize()} (\texttt{NLSF\_stabilize.c}). This process ensures that
consecutive values of the normalized LSF coefficients, \texttt{NLSF\_Q15[]},
are spaced some minimum distance apart (predetermined to be the 0.01
percentile of a large training set). Table 25 gives the minimum
spacings for NB and MB and those for WB, where row \texttt{k} is the minimum
allowed value of \texttt{NLSF\_Q15[k]}-\texttt{NLSF\_Q15[k-1]}. For the purposes of
computing this spacing for the first and last coefficient,
\texttt{NLSF\_Q15[-1]} is taken to be 0, and \texttt{NLSF\_Q15[d\_LPC]} is taken to be
32768.
The procedure starts off by trying to make small adjustments which attempt to minimize the amount of distortion introduced. After 20 such adjustments, it falls back to a more direct method which guarantees the constraints are enforced but may require large adjustments.

Let $\text{NDeltaMin}_Q15[k]$ be the minimum required spacing for the current audio bandwidth from Table 25. First, the procedure finds the index
i where \( NLSF_{Q15}[i] - NLSF_{Q15}[i-1] - NDeltaMin_{Q15}[i] \) is the smallest, breaking ties by using the lower value of \( i \). If this value is non-negative, then the stabilization stops; the coefficients satisfy all the constraints. Otherwise, if \( i == 0 \), it sets \( NLSF_{Q15}[0] \) to \( NDeltaMin_{Q15}[0] \), and if \( i == d_{LPC} \), it sets \( NLSF_{Q15}[d_{LPC}-1] \) to \( (32768 - NDeltaMin_{Q15}[d_{LPC}]) \). For all other values of \( i \), both \( NLSF_{Q15}[i-1] \) and \( NLSF_{Q15}[i] \) are updated as follows:

\[
\begin{align*}
\text{min\_center}_{Q15} &= (NDeltaMin_{Q15}[i]>>1) + \sum_{k=0}^{d_{LPC}-1} NDeltaMin_{Q15}[k] \\
\text{max\_center}_{Q15} &= 32768 - (NDeltaMin_{Q15}[i]>>1) - \sum_{k=i+1}^{d_{LPC}} NDeltaMin_{Q15}[k] \\
\text{center\_freq}_{Q15} &= \text{clamp}(\text{min\_center}_{Q15}[i], \\
&(NLSF_{Q15}[i-1] + NLSF_{Q15}[i] + 1)>>1, \\
&\text{max\_center}_{Q15}[i]) \\
NLSF_{Q15}[i-1] &= \text{center\_freq}_{Q15} - (NDeltaMin_{Q15}[i]>>1) \\
NLSF_{Q15}[i] &= NLSF_{Q15}[i-1] + NDeltaMin_{Q15}[i].
\end{align*}
\]

Then the procedure repeats again, until it has either executed 20 times or has stopped because the coefficients satisfy all the constraints.

After the 20th repetition of the above procedure, the following fallback procedure executes once. First, the values of \( NLSF_{Q15}[k] \) for \( 0 <= k < d_{LPC} \) are sorted in ascending order. Then for each value of \( k \) from 0 to \( d_{LPC}-1 \), \( NLSF_{Q15}[k] \) is set to

\[
\text{max}(NLSF_{Q15}[k], NLSF_{Q15}[k-1] + NDeltaMin_{Q15}[k]) .
\]

Next, for each value of \( k \) from \( d_{LPC}-1 \) down to 0, \( NLSF_{Q15}[k] \) is set to

\[
\text{min}(NLSF_{Q15}[k], NLSF_{Q15}[k+1] - NDeltaMin_{Q15}[k+1]) .
\]

There is no need to check if the coefficients satisfy all the constraints before applying this fallback procedure. If they do, then it will not change their values.
4.2.7.5.5. Normalized LSF Interpolation

For 20 ms SILK frames, the first half of the frame (i.e., the first two subframes) may use normalized LSF coefficients that are interpolated between the decoded LSFs for the most recent coded frame (in the same channel) and the current frame. A Q2 interpolation factor follows the LSF coefficient indices in the bitstream, which is decoded using the PDF in Table 26. This happens in silk_decode_indices() (decode_indices.c). After either

- An uncoded regular SILK frame in the side channel, or
- A decoder reset (see Section 4.5.2),

the decoder still decodes this factor, but ignores its value and always uses 4 instead. For 10 ms SILK frames, this factor is not stored at all.

```
+---------------------------+
| PDF                      |
+---------------------------+
| {13, 22, 29, 11, 181}/256 |
+---------------------------+
```

Table 26: PDF for Normalized LSF Interpolation Index

Let \( n2_{Q15}[k] \) be the normalized LSF coefficients decoded by the procedure in Section 4.2.7.5, \( n0_{Q15}[k] \) be the LSF coefficients decoded for the prior frame, and \( w_{Q2} \) be the interpolation factor. Then the normalized LSF coefficients used for the first half of a 20 ms frame, \( n1_{Q15}[k] \), are

\[
n1_{Q15}[k] = n0_{Q15}[k] + (w_{Q2}*(n2_{Q15}[k] - n0_{Q15}[k]) >> 2) \, .
\]

This interpolation is performed in silk_decode_parameters() (decode_parameters.c).

4.2.7.5.6. Converting Normalized LSFs to LPC Coefficients

Any LPC filter \( A(z) \) can be split into a symmetric part \( P(z) \) and an anti-symmetric part \( Q(z) \) such that

\[
d_{LPC}
A(z) = 1 - \sum_{k=1}^{\infty} a[k] * z^{-k} = 1 - (P(z) + Q(z)) / 2
\]

with

\[ P(z) = A(z) + z^{d_LPC-1} A(z) \]
\[ Q(z) = A(z) - z^{d_LPC-1} A(z) \]

The even normalized LSF coefficients correspond to a pair of conjugate roots of \( P(z) \), while the odd coefficients correspond to a pair of conjugate roots of \( Q(z) \), all of which lie on the unit circle. In addition, \( P(z) \) has a root at \( \pi \) and \( Q(z) \) has a root at \( 0 \). Thus, they may be reconstructed mathematically from a set of normalized LSF coefficients, \( n[k] \), as

\[
P(z) = (1 + z^{-1}) \sum_{k=0}^{d_LPC/2-1} (1 - 2\cos(\pi n[k]) z^{-2} + z^{-1})
\]
\[
Q(z) = (1 - z^{-1}) \sum_{k=0}^{d_LPC/2-1} (1 - 2\cos(\pi n[2k+1]) z^{-2} + z^{-1})
\]

However, SILK performs this reconstruction using a fixed-point approximation so that all decoders can reproduce it in a bit-exact manner to avoid prediction drift. The function silk_NLSF2A() (NLSF2A.c) implements this procedure.

To start, it approximates \( \cos(\pi n[k]) \) using a table lookup with linear interpolation. The encoder SHOULD use the inverse of this piecewise linear approximation, rather than the true inverse of the cosine function, when deriving the normalized LSF coefficients. These values are also re-ordered to improve numerical accuracy when constructing the LPC polynomials.
<table>
<thead>
<tr>
<th>Coefficient</th>
<th>NB and MB</th>
<th>WB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Table 27: LSF Ordering for Polynomial Evaluation

The top 7 bits of each normalized LSF coefficient index a value in the table, and the next 8 bits interpolate between it and the next value. Let \( i = (n[k] >> 8) \) be the integer index and \( f = (n[k] \& 255) \) be the fractional part of a given coefficient. Then the re-ordered, approximated cosine, \( c_{Q17[\text{ordering}[k]]} \), is

\[
c_{Q17[\text{ordering}[k]]} = (\cos_{Q12}[i] \times 256 + (\cos_{Q12}[i+1]-\cos_{Q12}[i]) \times f + 4) >> 3,
\]

where \( \text{ordering}[k] \) is the \( k \)’th entry of the column of Table 27.
corresponding to the current audio bandwidth and \( \cos_{Q12}[i] \) is the \( i \)'th entry of Table 28.

<table>
<thead>
<tr>
<th></th>
<th>+0</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4096</td>
<td>4095</td>
<td>4091</td>
<td>4085</td>
</tr>
<tr>
<td>4</td>
<td>4076</td>
<td>4065</td>
<td>4052</td>
<td>4036</td>
</tr>
<tr>
<td>8</td>
<td>4017</td>
<td>3997</td>
<td>3973</td>
<td>3948</td>
</tr>
<tr>
<td>12</td>
<td>3920</td>
<td>3889</td>
<td>3857</td>
<td>3822</td>
</tr>
<tr>
<td>16</td>
<td>3784</td>
<td>3745</td>
<td>3703</td>
<td>3659</td>
</tr>
<tr>
<td>20</td>
<td>3613</td>
<td>3564</td>
<td>3513</td>
<td>3461</td>
</tr>
<tr>
<td>24</td>
<td>3406</td>
<td>3349</td>
<td>3290</td>
<td>3229</td>
</tr>
<tr>
<td>28</td>
<td>3166</td>
<td>3102</td>
<td>3035</td>
<td>2967</td>
</tr>
<tr>
<td>32</td>
<td>2896</td>
<td>2824</td>
<td>2751</td>
<td>2676</td>
</tr>
<tr>
<td>36</td>
<td>2599</td>
<td>2520</td>
<td>2440</td>
<td>2359</td>
</tr>
<tr>
<td>40</td>
<td>2276</td>
<td>2191</td>
<td>2106</td>
<td>2019</td>
</tr>
<tr>
<td>44</td>
<td>1931</td>
<td>1842</td>
<td>1751</td>
<td>1660</td>
</tr>
<tr>
<td>48</td>
<td>1568</td>
<td>1474</td>
<td>1380</td>
<td>1285</td>
</tr>
<tr>
<td>52</td>
<td>1189</td>
<td>1093</td>
<td>995</td>
<td>897</td>
</tr>
<tr>
<td>56</td>
<td>799</td>
<td>700</td>
<td>601</td>
<td>501</td>
</tr>
<tr>
<td>60</td>
<td>401</td>
<td>301</td>
<td>201</td>
<td>101</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>-101</td>
<td>-201</td>
<td>-301</td>
</tr>
<tr>
<td>68</td>
<td>-401</td>
<td>-501</td>
<td>-601</td>
<td>-700</td>
</tr>
<tr>
<td>72</td>
<td>-799</td>
<td>-897</td>
<td>-995</td>
<td>-1093</td>
</tr>
<tr>
<td>76</td>
<td>-1189</td>
<td>-1285</td>
<td>-1380</td>
<td>-1474</td>
</tr>
<tr>
<td>80</td>
<td>-1568</td>
<td>-1660</td>
<td>-1751</td>
<td>-1842</td>
</tr>
</tbody>
</table>
Table 28: Q12 Cosine Table for LSF Conversion

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>-1931</td>
<td>-2019</td>
<td>-2106</td>
<td>-2191</td>
</tr>
<tr>
<td>88</td>
<td>-2276</td>
<td>-2359</td>
<td>-2440</td>
<td>-2520</td>
</tr>
<tr>
<td>92</td>
<td>-2599</td>
<td>-2676</td>
<td>-2751</td>
<td>-2824</td>
</tr>
<tr>
<td>96</td>
<td>-2896</td>
<td>-2967</td>
<td>-3035</td>
<td>-3102</td>
</tr>
<tr>
<td>100</td>
<td>-3166</td>
<td>-3229</td>
<td>-3290</td>
<td>-3349</td>
</tr>
<tr>
<td>104</td>
<td>-3406</td>
<td>-3461</td>
<td>-3513</td>
<td>-3564</td>
</tr>
<tr>
<td>108</td>
<td>-3613</td>
<td>-3659</td>
<td>-3703</td>
<td>-3745</td>
</tr>
<tr>
<td>112</td>
<td>-3784</td>
<td>-3822</td>
<td>-3857</td>
<td>-3899</td>
</tr>
<tr>
<td>116</td>
<td>-3920</td>
<td>-3948</td>
<td>-3973</td>
<td>-3997</td>
</tr>
<tr>
<td>120</td>
<td>-4017</td>
<td>-4036</td>
<td>-4052</td>
<td>-4065</td>
</tr>
<tr>
<td>124</td>
<td>-4076</td>
<td>-4085</td>
<td>-4091</td>
<td>-4095</td>
</tr>
<tr>
<td>128</td>
<td>-4096</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Given the list of cosine values, silk_NLSF2A_find_poly() (NLSF2A.c) computes the coefficients of P and Q, described here via a simple recurrence. Let \( p_{Q16[k][j]} \) and \( q_{Q16[k][j]} \) be the coefficients of the products of the first \((k+1)\) root pairs for P and Q, with \( j \) indexing the coefficient number. Only the first \((k+2)\) coefficients are needed, as the products are symmetric. Let 

- \( p_{Q16[0][0]} = q_{Q16[0][0]} = 1<<16, \)
- \( p_{Q16[0][1]} = -c_{Q17[0]}, \)
- \( q_{Q16[0][1]} = -c_{Q17[1]}, \)

and \( d_2 = d_{LPC}/2. \) As boundary conditions, assume \( p_{Q16[k][j]} = q_{Q16[k][j]} = 0 \) for all \( j < 0. \) Also, assume 

- \( p_{Q16[k][k+2]} = p_{Q16[k][k]} \)
- \( q_{Q16[k][k+2]} = q_{Q16[k][k]} \)

(because of the symmetry). Then, for \( 0 < k < d_2 \) and \( 0 \leq j \leq k+1, \)

\[
\begin{align*}
    p_{Q16[k][j]} &= p_{Q16[k-1][j]} + p_{Q16[k-1][j-2]} \\
                   &\quad - ((c_{Q17[2*k]}*p_{Q16[k-1][j-1]} + 32768)>>16), \\
    q_{Q16[k][j]} &= q_{Q16[k-1][j]} + q_{Q16[k-1][j-2]} \\
                   &\quad - ((c_{Q17[2*k+1]}*q_{Q16[k-1][j-1]} + 32768)>>16).
\end{align*}
\]

The use of Q17 values for the cosine terms in an otherwise Q16 expression implicitly scales them by a factor of 2. The multiplications in this recurrence may require up to 48 bits of
precision in the result to avoid overflow. In practice, each row of
the recurrence only depends on the previous row, so an implementation
does not need to store all of them.

silk_NLSF2A() uses the values from the last row of this recurrence to
reconstruct a 32-bit version of the LPC filter (without the leading
1.0 coefficient), \( a_{32\_Q17}[k] \), \( 0 \leq k < d2 \):

\[
a_{32\_Q17}[k] = -(q_{Q16[d2-1][k+1]} - q_{Q16[d2-1][k]})
- (p_{Q16[d2-1][k+1]} + p_{Q16[d2-1][k]}) ,
\]

\[
a_{32\_Q17}[d\_LPC-k-1] = (q_{Q16[d2-1][k+1]} - q_{Q16[d2-1][k]})
- (p_{Q16[d2-1][k+1]} + p_{Q16[d2-1][k]}) .
\]

The sum and difference of two terms from each of the \( p_{Q16} \) and \( q_{Q16} \)
coefficient lists reflect the \((1 + z^{-1})\) and \((1 - z^{-1})\) factors of
\( P \) and \( Q \), respectively. The promotion of the expression from \( Q16 \) to
\( Q17 \) implicitly scales the result by 1/2.

4.2.7.5.7. Limiting the Range of the LPC Coefficients

The \( a_{32\_Q17}[] \) coefficients are too large to fit in a 16-bit value,
which significantly increases the cost of applying this filter in
fixed-point decoders. Reducing them to Q12 precision doesn’t incur
any significant quality loss, but still does not guarantee they will
fit. silk_NLSF2A() applies up to 10 rounds of bandwidth expansion to
limit the dynamic range of these coefficients. Even floating-point
decoders SHOULD perform these steps, to avoid mismatch.

For each round, the process first finds the index \( k \) such that
\( \text{abs}(a_{32\_Q17}[k]) \) is largest, breaking ties by choosing the lowest
value of \( k \). Then, it computes the corresponding Q12 precision value,
\( \text{maxabs}_{Q12} \), subject to an upper bound to avoid overflow in subsequent
computations:

\[
\text{maxabs}_{Q12} = \min((\text{maxabs}_{Q17} + 16) \gg 5, 163838) .
\]

If this is larger than 32767, the procedure derives the chirp factor,
\( \text{sc}_{Q16}[0] \), to use in the bandwidth expansion as

\[
\text{sc}_{Q16}[0] = 65470 - \frac{(\text{maxabs}_{Q12} - 32767) \ll 14
\text{maxabs}_{Q12} \times (k+1)) \gg 2}{\text{maxabs}_{Q12} \times (k+1)) \gg 2}
\]

where the division here is integer division. This is an
approximation of the chirp factor needed to reduce the target
coefficient to 32767, though it is both less than 0.999 and, for
\( k > 0 \) when \( \text{maxabs}_{Q12} \) is much greater than 32767, still slightly too
large. The upper bound on maxabs_Q12, 163838, was chosen because it is equal to \((2^{31} - 1) \gg 14 + 32767\), i.e., the largest value of maxabs_Q12 that would not overflow the numerator in the equation above when stored in a signed 32-bit integer.

silk_bwexpander_32() (bwexpander_32.c) performs the bandwidth expansion (again, only when maxabs_Q12 is greater than 32767) using the following recurrence:

\[
\begin{align*}
a32_{Q17}[k] &= (a32_{Q17}[k] \cdot sc_{Q16}[k]) \gg 16 \\
sc_{Q16}[k+1] &= (sc_{Q16}[0] \cdot sc_{Q16}[k] + 32768) \gg 16
\end{align*}
\]

The first multiply may require up to 48 bits of precision in the result to avoid overflow. The second multiply must be unsigned to avoid overflow with only 32 bits of precision. The reference implementation uses a slightly more complex formulation that avoids the 32-bit overflow using signed multiplication, but is otherwise equivalent.

After 10 rounds of bandwidth expansion are performed, they are simply saturated to 16 bits:

\[
a32_{Q17}[k] = \text{clamp}(-32768, (a32_{Q17}[k] + 16) \gg 5, 32767) \ll 5 .
\]

Because this performs the actual saturation in the Q12 domain, but converts the coefficients back to the Q17 domain for the purposes of prediction gain limiting, this step must be performed after the 10th round of bandwidth expansion, regardless of whether or not the Q12 version of any coefficient still overflows a 16-bit integer. This saturation is not performed if maxabs_Q12 drops to 32767 or less prior to the 10th round.

4.2.7.5.8. Limiting the Prediction Gain of the LPC Filter

The prediction gain of an LPC synthesis filter is the square-root of the output energy when the filter is excited by a unit-energy impulse. Even if the Q12 coefficients would fit, the resulting filter may still have a significant gain (especially for voiced sounds), making the filter unstable. silk_NLSF2A() applies up to 18 additional rounds of bandwidth expansion to limit the prediction gain. Instead of controlling the amount of bandwidth expansion using the prediction gain itself (which may diverge to infinity for an unstable filter), silk_NLSF2A() uses silk_LPC_inverse_pred_gain_QA() (LPC_inv_pred_gain.c) to compute the reflection coefficients associated with the filter. The filter is stable if and only if the magnitude of these coefficients is sufficiently less than one. The reflection coefficients, rc[k], can be computed using a simple
Levinson recurrence, initialized with the LPC coefficients $a[d_{LPC}-1][n] = a[n]$, and then updated via
\[
rc[k] = -a[k][k],
\]
\[
a[k][n] - a[k][k-n-1]*rc[k]
\]
\[
a[k-1][n] = \frac{2}{1 - rc[k]}. 
\]

However, silk_LPC_inverse_pred_gain_QA() approximates this using fixed-point arithmetic to guarantee reproducible results across platforms and implementations. Since small changes in the coefficients can make a stable filter unstable, it takes the real Q12 coefficients that will be used during reconstruction as input. Thus, let
\[
a_{32\_Q12}[n] = (a_{32\_Q17}[n] + 16) >> 5
\]
be the Q12 version of the LPC coefficients that will eventually be used. As a simple initial check, the decoder computes the DC response as
\[
d_{PLC-1}
\]
\[
DC_{resp} = \sum_{n=0}^{d_{PLC-1}} a_{32\_Q12}[n]
\]
and if $DC_{resp} > 4096$, the filter is unstable.

Increasing the precision of these Q12 coefficients to Q24 for intermediate computations allows more accurate computation of the reflection coefficients, so the decoder initializes the recurrence via
\[
a_{32\_Q24}[d_{LPC}-1][n] = a_{32\_Q12}[n] << 12.
\]

Then for each $k$ from $d_{LPC}-1$ down to 0, if $\text{abs}(a_{32\_Q24}[k][k]) > 16773022$, the filter is unstable and the recurrence stops. The constant 16773022 here is approximately 0.99975 in Q24. Otherwise, row $k-1$ of $a_{32\_Q24}$ is computed from row $k$ as
\[
\begin{align*}
rc\_Q31[k] &= -a32\_Q24[k][k] << 7 , \\
div\_Q30[k] &= (1<<30) - (rc\_Q31[k]*rc\_Q31[k] >> 32) , \\
b1[k] &= ilog(div\_Q30[k]) , \\
b2[k] &= b1[k] - 16 , \\
(1<<29) - 1 \quad \text{inv\_Qb2[k]} &= \frac{1}{div\_Q30[k] >> (b2[k]+1)} , \\
err\_Q29[k] &= (1<<29) \\
&- ((div\_Q30[k]<<((15-b2[k]))*inv\_Qb2[k] >> 16) , \\
gain\_Qb1[k] &= ((inv\_Qb2[k] << 16) \\
&+ (err\_Q29[k]*inv\_Qb2[k] >> 13)) , \\
num\_Q24[k-1][n] &= a32\_Q24[k][n] \\
&- ((a32\_Q24[k][k-n-1]*rc\_Q31[k] + (1<<30)) >> 31) , \\
a32\_Q24[k-1][n] &= (num\_Q24[k-1][n]*gain\_Qb1[k] \\
&+ (1<<((b1[k]-1))) >> b1[k] ,
\end{align*}
\]

where 0 <= n < k. Here, rc\_Q30[k] are the reflection coefficients. div\_Q30[k] is the denominator for each iteration, and gain\_Qb1[k] is its multiplicative inverse (with b1[k] fractional bits, where b1[k] ranges from 20 to 31). inv\_Qb2[k], which ranges from 16384 to 32767, is a low-precision version of that inverse (with b2[k] fractional bits). err\_Q29[k] is the residual error, ranging from -32763 to 32392, which is used to improve the accuracy. The values t\_Q24[k-1][n] for each n are the numerators for the next row of coefficients in the recursion, and a32\_Q24[k-1][n] is the final version of that row. Every multiply in this procedure except the one used to compute gain\_Qb1[k] requires more than 32 bits of precision, but otherwise all intermediate results fit in 32 bits or less. In practice, because each row only depends on the next one, an implementation does not need to store them all.

If abs(a32\_Q24[k][k]) <= 16773022 for 0 <= k < d\_LPC, then the filter is considered stable. However, the problem of determining stability is ill-conditioned when the filter contains several reflection coefficients whose magnitude is very close to one. This fixed-point algorithm is not mathematically guaranteed to correctly classify filters as stable or unstable in this case, though it does very well in practice.

On round i, 1 <= i <= 18, if the filter passes these stability checks...
checks, then this procedure stops, and the final LPC coefficients to use for reconstruction in Section 4.2.7.9.2 are

\[ a_{Q12}[k] = (a_{32Q17}[k] + 16) \gg 5 . \]

Otherwise, a round of bandwidth expansion is applied using the same procedure as in Section 4.2.7.5.7, with

\[ sc_{Q16}[0] = 65536 - (2^{<<i}) . \]

During the 15th round, \( sc_{Q16}[0] \) becomes 0 in the above equation, so \( a_{Q12}[k] \) is set to 0 for all \( k \), guaranteeing a stable filter.

4.2.7.6. Long-Term Prediction (LTP) Parameters

After the normalized LSF indices and, for 20 ms frames, the LSF interpolation index, voiced frames (see Section 4.2.7.3) include additional LTP parameters. There is one primary lag index for each SILK frame, but this is refined to produce a separate lag index per subframe using a vector quantizer. Each subframe also gets its own prediction gain coefficient.

4.2.7.6.1. Pitch Lags

The primary lag index is coded either relative to the primary lag of the prior frame in the same channel, or as an absolute index. Absolute coding is used if and only if

- This is the first SILK frame of its type (LBRR or regular) for this channel in the current Opus frame,
- The previous SILK frame of the same type (LBRR or regular) for this channel in the same Opus frame was not coded, or
- That previous SILK frame was coded, but was not voiced (see Section 4.2.7.3).

With absolute coding, the primary pitch lag may range from 2 ms (inclusive) up to 18 ms (exclusive), corresponding to pitches from 500 Hz down to 55.6 Hz, respectively. It is comprised of a high part and a low part, where the decoder reads the high part using the 32-entry codebook in Table 29 and the low part using the codebook corresponding to the current audio bandwidth from Table 30. The final primary pitch lag is then

\[ \text{lag} = \text{lag}_{\text{high}} \times \text{lag}_{\text{scale}} + \text{lag}_{\text{low}} + \text{lag}_{\text{min}} \]

where \( \text{lag}_{\text{high}} \) is the high part, \( \text{lag}_{\text{low}} \) is the low part, and
lag_scale and lag_min are the values from the "Scale" and "Minimum Lag" columns of Table 30, respectively.

<table>
<thead>
<tr>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>{3, 3, 6, 11, 21, 30, 32, 19, 11, 10, 12, 13, 13, 12, 11, 9, 8, 7, 6, 4, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 29: PDF for High Part of Primary Pitch Lag

<table>
<thead>
<tr>
<th>Audio Bandwidth</th>
<th>PDF</th>
<th>Scale</th>
<th>Minimum Lag</th>
<th>Maximum Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>{64, 64, 64, 64}/256</td>
<td>4</td>
<td>16</td>
<td>144</td>
</tr>
<tr>
<td>MB</td>
<td>{43, 42, 43, 43, 42, 43}/256</td>
<td>6</td>
<td>24</td>
<td>216</td>
</tr>
<tr>
<td>WB</td>
<td>{32, 32, 32, 32, 32, 32, 32}/256</td>
<td>8</td>
<td>32</td>
<td>288</td>
</tr>
</tbody>
</table>

Table 30: PDF for Low Part of Primary Pitch Lag

All frames that do not use absolute coding for the primary lag index use relative coding instead. The decoder reads a single delta value using the 21-entry PDF in Table 31. If the resulting value is zero, it falls back to the absolute coding procedure from the prior paragraph. Otherwise, the final primary pitch lag is then

\[
\text{lag} = \text{previous\_lag} + (\text{delta\_lag\_index} - 9)
\]

where previous_lag is the primary pitch lag from the most recent frame in the same channel and delta_lag_index is the value just decoded. This allows a per-frame change in the pitch lag of -8 to +11 samples. The decoder does no clamping at this point, so this value can fall outside the range of 2 ms to 18 ms, and the decoder must use this unclamped value when using relative coding in the next SILK frame (if any). However, because an Opus frame can use relative coding for at most two consecutive SILK frames, integer overflow should not be an issue.
Table 31: PDF for Primary Pitch Lag Change

After the primary pitch lag, a "pitch contour", stored as a single entry from one of four small VQ codebooks, gives lag offsets for each subframe in the current SILK frame. The codebook index is decoded using one of the PDFs in Table 32 depending on the current frame size and audio bandwidth. Tables 33 through 36 give the corresponding offsets to apply to the primary pitch lag for each subframe given the decoded codebook index.

<table>
<thead>
<tr>
<th>Audio Bandwidth</th>
<th>SILK Frame Size</th>
<th>Codebook Size</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>10 ms</td>
<td>3</td>
<td>{143, 50, 63}/256</td>
</tr>
<tr>
<td>NB</td>
<td>20 ms</td>
<td>11</td>
<td>{68, 12, 21, 17, 19, 22, 30, 24, 17, 16, 10}/256</td>
</tr>
<tr>
<td>MB or WB</td>
<td>10 ms</td>
<td>12</td>
<td>{91, 46, 39, 19, 14, 12, 8, 7, 6, 5, 5, 4}/256</td>
</tr>
<tr>
<td>MB or WB</td>
<td>20 ms</td>
<td>34</td>
<td>{33, 22, 18, 16, 15, 14, 14, 13, 13, 10, 9, 9, 8, 6, 6, 5, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 32: PDFs for Subframe Pitch Contour
<table>
<thead>
<tr>
<th>Index</th>
<th>Subframe Offsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td>1</td>
<td>1 0</td>
</tr>
<tr>
<td>2</td>
<td>0 1</td>
</tr>
</tbody>
</table>

Table 33: Codebook Vectors for Subframe Pitch Contour: NB, 10 ms Frames

<table>
<thead>
<tr>
<th>Index</th>
<th>Subframe Offsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>1</td>
<td>2 1 0 -1</td>
</tr>
<tr>
<td>2</td>
<td>-1 0 1 2</td>
</tr>
<tr>
<td>3</td>
<td>-1 0 0 1</td>
</tr>
<tr>
<td>4</td>
<td>-1 0 0 0</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 1</td>
</tr>
<tr>
<td>6</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>7</td>
<td>1 1 0 0</td>
</tr>
<tr>
<td>8</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>9</td>
<td>0 0 0 -1</td>
</tr>
<tr>
<td>10</td>
<td>1 0 0 -1</td>
</tr>
</tbody>
</table>

Table 34: Codebook Vectors for Subframe Pitch Contour: NB, 20 ms Frames
<table>
<thead>
<tr>
<th>Index</th>
<th>Subframe Offsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td>1</td>
<td>0 1</td>
</tr>
<tr>
<td>2</td>
<td>1 0</td>
</tr>
<tr>
<td>3</td>
<td>-1 1</td>
</tr>
<tr>
<td>4</td>
<td>1 -1</td>
</tr>
<tr>
<td>5</td>
<td>-1 2</td>
</tr>
<tr>
<td>6</td>
<td>2 -1</td>
</tr>
<tr>
<td>7</td>
<td>-2 2</td>
</tr>
<tr>
<td>8</td>
<td>2 -2</td>
</tr>
<tr>
<td>9</td>
<td>-2 3</td>
</tr>
<tr>
<td>10</td>
<td>3 -2</td>
</tr>
<tr>
<td>11</td>
<td>-3 3</td>
</tr>
</tbody>
</table>

Table 35: Codebook Vectors for Subframe Pitch Contour: MB or WB, 10 ms Frames

<table>
<thead>
<tr>
<th>Index</th>
<th>Subframe Offsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>1</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 0 0</td>
</tr>
<tr>
<td>3</td>
<td>-1 0 0 0</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 1</td>
</tr>
<tr>
<td>5</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>6</td>
<td>-1 0 0 1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>-2</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>-2</td>
</tr>
<tr>
<td>13</td>
<td>-2</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>-3</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>-3</td>
</tr>
<tr>
<td>19</td>
<td>-4</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>-4</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>-5</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>26</td>
<td>-6</td>
</tr>
<tr>
<td>27</td>
<td>-5</td>
</tr>
<tr>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>-7</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 36: Codebook Vectors for Subframe Pitch Contour: MB or WB, 20 ms Frames

The final pitch lag for each subframe is assembled in 
silk_decode_pitch() (decode_pitch.c). Let lag be the primary pitch lag for the current SILK frame, contour_index be index of the VQ codebook, and lag_cb[contour_index][k] be the corresponding entry of the codebook from the appropriate table given above for the k’th subframe. Then the final pitch lag for that subframe is

\[ \text{pitch_lags}[k] = \text{clamp}(\text{lag_min}, \text{lag} + \text{lag_cb}[\text{contour_index}][k], \text{lag_max}) \]

where lag_min and lag_max are the values from the "Minimum Lag" and "Maximum Lag" columns of Table 30, respectively.

4.2.7.6.2. LTP Filter Coefficients

SILK uses a separate 5-tap pitch filter for each subframe, selected from one of three codebooks. The three codebooks each represent different rate-distortion trade-offs, with average rates of 1.61 bits/subframe, 3.68 bits/subframe, and 4.85 bits/subframe, respectively.

The importance of the filter coefficients generally depends on two factors: the periodicity of the signal and relative energy between the current subframe and the signal from one period earlier. Greater periodicity and decaying energy both lead to more important filter coefficients, and thus should be coded with lower distortion and higher rate. These properties are relatively stable over the duration of a single SILK frame, hence all of the subframes in a SILK frame choose their filter from the same codebook. This is signaled with an explicitly-coded "periodicity index". This immediately follows the subframe pitch lags, and is coded using the 3-entry PDF from Table 37.
The indices of the filters for each subframe follow. They are all coded using the PDF from Table 38 corresponding to the periodicity index. Tables 39 through 41 contain the corresponding filter taps as signed Q7 integers.

### Table 37: Periodicity Index PDF

<table>
<thead>
<tr>
<th>Periodicity Index</th>
<th>Codebook Size</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>{185, 15, 13, 13, 9, 9, 6, 6}/256</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>{57, 34, 21, 20, 15, 13, 12, 13, 10, 10, 9, 8, 7, 8}/256</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>{15, 16, 14, 12, 12, 12, 11, 11, 11, 10, 9, 9, 9, 9, 8, 8, 8, 8, 7, 7, 6, 6, 5, 4, 5, 4, 4, 4, 3, 4, 3, 2}/256</td>
</tr>
</tbody>
</table>

### Table 38: LTP Filter PDFs

Table 39: Codebook Vectors for LTP Filter, Periodicity Index 0

<table>
<thead>
<tr>
<th>Index</th>
<th>Filter Taps (Q7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4 6 24 7 5</td>
</tr>
<tr>
<td>1</td>
<td>0 0 2 0 0</td>
</tr>
<tr>
<td>2</td>
<td>12 28 41 13 -4</td>
</tr>
<tr>
<td>3</td>
<td>-9 15 42 25 14</td>
</tr>
<tr>
<td>4</td>
<td>1 -2 62 41 -9</td>
</tr>
<tr>
<td>5</td>
<td>-10 37 65 -4 3</td>
</tr>
<tr>
<td>6</td>
<td>-6 4 66 7 -8</td>
</tr>
<tr>
<td>7</td>
<td>16 14 38 -3 33</td>
</tr>
</tbody>
</table>
Table 40: Codebook Vectors for LTP Filter, Periodicity Index 1

<table>
<thead>
<tr>
<th>Index</th>
<th>Filter Taps (Q7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13 22 39 23 12</td>
</tr>
<tr>
<td>1</td>
<td>-1 36 64 27 -6</td>
</tr>
<tr>
<td>2</td>
<td>-7 10 55 43 17</td>
</tr>
<tr>
<td>3</td>
<td>1 1 8 1 1</td>
</tr>
<tr>
<td>4</td>
<td>6 -11 74 53 -9</td>
</tr>
<tr>
<td>5</td>
<td>-12 55 76 -12 8</td>
</tr>
<tr>
<td>6</td>
<td>-3 3 93 27 -4</td>
</tr>
<tr>
<td>7</td>
<td>26 39 59 3 -8</td>
</tr>
<tr>
<td>8</td>
<td>2 0 77 11 9</td>
</tr>
<tr>
<td>9</td>
<td>-8 22 44 -6 7</td>
</tr>
<tr>
<td>10</td>
<td>40 9 26 3 9</td>
</tr>
<tr>
<td>11</td>
<td>-7 20 101 -7 4</td>
</tr>
<tr>
<td>12</td>
<td>3 -8 42 26 0</td>
</tr>
<tr>
<td>13</td>
<td>-15 33 68 2 23</td>
</tr>
<tr>
<td>14</td>
<td>-2 55 46 -2 15</td>
</tr>
<tr>
<td>15</td>
<td>3 -1 21 16 41</td>
</tr>
</tbody>
</table>

Table 40: Codebook Vectors for LTP Filter, Periodicity Index 1
<table>
<thead>
<tr>
<th>4</th>
<th>-9 19 94 29 -9</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0 12 99 6 4</td>
</tr>
<tr>
<td>6</td>
<td>8 -19 102 46 -13</td>
</tr>
<tr>
<td>7</td>
<td>3 2 13 3 2</td>
</tr>
<tr>
<td>8</td>
<td>9 -21 84 72 -18</td>
</tr>
<tr>
<td>9</td>
<td>-11 46 104 -22 8</td>
</tr>
<tr>
<td>10</td>
<td>18 38 48 23 0</td>
</tr>
<tr>
<td>11</td>
<td>-16 70 83 -21 11</td>
</tr>
<tr>
<td>12</td>
<td>5 -11 117 22 -8</td>
</tr>
<tr>
<td>13</td>
<td>-6 23 117 -12 3</td>
</tr>
<tr>
<td>14</td>
<td>3 -8 95 28 4</td>
</tr>
<tr>
<td>15</td>
<td>-10 15 77 60 -15</td>
</tr>
<tr>
<td>16</td>
<td>-1 4 124 2 -4</td>
</tr>
<tr>
<td>17</td>
<td>3 38 84 24 -25</td>
</tr>
<tr>
<td>18</td>
<td>2 13 42 13 31</td>
</tr>
<tr>
<td>19</td>
<td>21 -4 56 46 -1</td>
</tr>
<tr>
<td>20</td>
<td>-1 35 79 -13 19</td>
</tr>
<tr>
<td>21</td>
<td>-7 65 88 -9 -14</td>
</tr>
<tr>
<td>22</td>
<td>20 4 81 49 -29</td>
</tr>
<tr>
<td>23</td>
<td>20 0 75 3 -17</td>
</tr>
<tr>
<td>24</td>
<td>5 -9 44 92 -8</td>
</tr>
<tr>
<td>25</td>
<td>1 -3 22 69 31</td>
</tr>
<tr>
<td>26</td>
<td>-6 95 41 -12 5</td>
</tr>
<tr>
<td>27</td>
<td>39 67 16 -4 1</td>
</tr>
</tbody>
</table>

Table 41: Codebook Vectors for LTP Filter, Periodicity Index 2

4.2.7.6.3. LTP Scaling Parameter

An LTP scaling parameter appears after the LTP filter coefficients if and only if

- This is a voiced frame (see Section 4.2.7.3), and
- Either
  * This SILK frame corresponds to the first time interval of the current Opus frame for its type (LBRR or regular), or
  * This is an LBRR frame where the LBRR flags (see Section 4.2.4) indicate the previous LBRR frame in the same channel is not coded.

This allows the encoder to trade off the prediction gain between packets against the recovery time after packet loss. Unlike absolute-coding for pitch lags, regular SILK frames that are not at the start of an Opus frame (i.e., that do not correspond to the first 20 ms time interval in Opus frames of 40 or 60 ms) do not include this field, even if the prior frame was not voiced, or (in the case of the side channel) not even coded. After an uncoded frame in the side channel, the LTP buffer (see Section 4.2.7.9.1) is cleared to zero, and is thus in a known state. In contrast, LBRR frames do include this field when the prior frame was not coded, since the LTP buffer contains the output of the PLC, which is non-normative.

If present, the decoder reads a value using the 3-entry PDF in Table 42. The three possible values represent Q14 scale factors of 15565, 12288, and 8192, respectively (corresponding to approximately 0.95, 0.75, and 0.5). Frames that do not code the scaling parameter use the default factor of 15565 (approximately 0.95).
4.2.7.7. Linear Congruential Generator (LCG) Seed

As described in Section 4.2.7.8.6, SILK uses a linear congruential generator (LCG) to inject pseudorandom noise into the quantized excitation. To ensure synchronization of this process between the encoder and decoder, each SILK frame stores a 2-bit seed after the LTP parameters (if any). The encoder may consider the choice of seed during quantization, and the flexibility of this choice lets it reduce distortion, helping to pay for the bit cost required to signal it. The decoder reads the seed using the uniform 4-entry PDF in Table 43, yielding a value between 0 and 3, inclusive.

4.2.7.8. Excitation

SILK codes the excitation using a modified version of the Pyramid Vector Quantization (PVQ) codebook [PVQ]. The PVQ codebook is designed for Laplace-distributed values and consists of all sums of K signed, unit pulses in a vector of dimension N, where two pulses at the same position are required to have the same sign. Thus the codebook includes all integer codevectors \( y \) of dimension \( N \) that satisfy

\[
\sum_{j=0}^{N-1} |y[j]| = K.
\]

Unlike regular PVQ, SILK uses a variable-length, rather than fixed-length, encoding. This encoding is better suited to the more Gaussian-like distribution of the coefficient magnitudes and the non-uniform distribution of their signs (caused by the quantization...
SILK also handles large codebooks by coding the least significant bits (LSBs) of each coefficient directly. This adds a small coding efficiency loss, but greatly reduces the computation time and ROM size required for decoding, as implemented in `silk_decode_pulses()` (decode_pulses.c).

SILK fixes the dimension of the codebook to \( N = 16 \). The excitation is made up of a number of "shell blocks", each 16 samples in size. Table 44 lists the number of shell blocks required for a SILK frame for each possible audio bandwidth and frame size. 10 ms MB frames nominally contain 120 samples (10 ms at 12 kHz), which is not a multiple of 16. This is handled by coding 8 shell blocks (128 samples) and discarding the final 8 samples of the last block. The decoder contains no special case that prevents an encoder from placing pulses in these samples, and they must be correctly parsed from the bitstream if present, but they are otherwise ignored.

<table>
<thead>
<tr>
<th>Audio Bandwidth</th>
<th>Frame Size</th>
<th>Number of Shell Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>10 ms</td>
<td>5</td>
</tr>
<tr>
<td>MB</td>
<td>10 ms</td>
<td>8</td>
</tr>
<tr>
<td>WB</td>
<td>10 ms</td>
<td>10</td>
</tr>
<tr>
<td>NB</td>
<td>20 ms</td>
<td>10</td>
</tr>
<tr>
<td>MB</td>
<td>20 ms</td>
<td>15</td>
</tr>
<tr>
<td>WB</td>
<td>20 ms</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 44: Number of Shell Blocks Per SILK Frame

4.2.7.8.1. Rate Level

The first symbol in the excitation is a "rate level", which is an index from 0 to 8, inclusive, coded using the PDF in Table 45 corresponding to the signal type of the current frame (from Section 4.2.7.3). The rate level selects the PDF used to decode the number of pulses in the individual shell blocks. It does not directly convey any information about the bitrate or the number of pulses itself, but merely changes the probability of the symbols in Section 4.2.7.8.2. Level 0 provides a more efficient encoding at low rates generally, and level 8 provides a more efficient encoding at high rates generally, though the most efficient level for a particular SILK frame may depend on the exact distribution of the...
coded symbols. An encoder should, but is not required to, use the most efficient rate level.

<table>
<thead>
<tr>
<th>Signal Type</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive or Unvoiced</td>
<td>{15, 51, 12, 46, 45, 13, 33, 27, 14}/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>{33, 30, 36, 17, 34, 49, 18, 21, 18}/256</td>
</tr>
</tbody>
</table>

Table 45: PDFs for the Rate Level

4.2.7.8.2. Pulses Per Shell Block

The total number of pulses in each of the shell blocks follows the rate level. The pulse counts for all of the shell blocks are coded consecutively, before the content of any of the blocks. Each block may have anywhere from 0 to 16 pulses, inclusive, coded using the 18-entry PDF in Table 46 corresponding to the rate level from Section 4.2.7.8.1. The special value 17 indicates that this block has one or more additional LSBs to decode for each coefficient. If the decoder encounters this value, it decodes another value for the actual pulse count of the block, but uses the PDF corresponding to the special rate level 9 instead of the normal rate level. This process repeats until the decoder reads a value less than 17, and it then sets the number of extra LSBs used to the number of 17’s decoded for that block. If it reads the value 17 ten times, then the next iteration uses the special rate level 10 instead of 9. The probability of decoding a 17 when using the PDF for rate level 10 is zero, ensuring that the number of LSBs for a block will not exceed 10. The cumulative distribution for rate level 10 is just a shifted version of that for 9 and thus does not require any additional storage.
Table 46: PDFs for the Pulse Count

4.2.7.8.3. Pulse Location Decoding

The locations of the pulses in each shell block follow the pulse counts, as decoded by silk_shell_decoder() (shell_coder.c). As with the pulse counts, these locations are coded for all the shell blocks before any of the remaining information for each block. Unlike many other codecs, SILK places no restriction on the distribution of pulses within a shell block. All of the pulses may be placed in a
single location, or each one in a unique location, or anything in between.

The location of pulses is coded by recursively partitioning each block into halves, and coding how many pulses fall on the left side of the split. All remaining pulses must fall on the right side of the split. The process then recurses into the left half, and after that returns, the right half (preorder traversal). The PDF to use is chosen by the size of the current partition (16, 8, 4, or 2) and the number of pulses in the partition (1 to 16, inclusive). Tables 47 through 50 list the PDFs used for each partition size and pulse count. This process skips partitions without any pulses, i.e., where the initial pulse count from Section 4.2.7.8.2 was zero, or where the split in the prior level indicated that all of the pulses fell on the other side. These partitions have nothing to code, so they require no PDF.
<table>
<thead>
<tr>
<th>Pulse Count</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{126, 130}/256</td>
</tr>
<tr>
<td>2</td>
<td>{56, 142, 58}/256</td>
</tr>
<tr>
<td>3</td>
<td>{25, 101, 104, 26}/256</td>
</tr>
<tr>
<td>4</td>
<td>{12, 60, 108, 64, 12}/256</td>
</tr>
<tr>
<td>5</td>
<td>{7, 35, 84, 87, 37, 6}/256</td>
</tr>
<tr>
<td>6</td>
<td>{4, 20, 59, 86, 63, 21, 3}/256</td>
</tr>
<tr>
<td>7</td>
<td>{3, 12, 38, 72, 75, 42, 12, 2}/256</td>
</tr>
<tr>
<td>8</td>
<td>{2, 8, 25, 54, 73, 59, 27, 7, 1}/256</td>
</tr>
<tr>
<td>9</td>
<td>{2, 5, 17, 39, 63, 65, 42, 18, 4, 1}/256</td>
</tr>
<tr>
<td>10</td>
<td>{1, 4, 12, 28, 49, 63, 54, 30, 11, 3, 1}/256</td>
</tr>
<tr>
<td>11</td>
<td>{1, 4, 8, 20, 37, 55, 57, 41, 22, 8, 2, 1}/256</td>
</tr>
<tr>
<td>12</td>
<td>{1, 3, 7, 15, 28, 44, 53, 48, 33, 16, 6, 1, 1}/256</td>
</tr>
<tr>
<td>13</td>
<td>{1, 2, 6, 12, 21, 35, 47, 48, 40, 25, 12, 5, 1, 1}/256</td>
</tr>
<tr>
<td>14</td>
<td>{1, 1, 4, 10, 17, 27, 37, 47, 43, 33, 21, 9, 4, 1, 1}/256</td>
</tr>
<tr>
<td>15</td>
<td>{1, 1, 1, 8, 14, 22, 33, 40, 43, 38, 28, 16, 8, 1, 1, 1}/256</td>
</tr>
<tr>
<td>16</td>
<td>{1, 1, 1, 1, 13, 18, 27, 36, 41, 41, 34, 24, 14, 1, 1, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 47: PDFs for Pulse Count Split, 16 Sample Partitions
<table>
<thead>
<tr>
<th>Pulse Count</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{127, 129}/256</td>
</tr>
<tr>
<td>2</td>
<td>{53, 149, 54}/256</td>
</tr>
<tr>
<td>3</td>
<td>{22, 105, 106, 23}/256</td>
</tr>
<tr>
<td>4</td>
<td>{11, 61, 111, 63, 10}/256</td>
</tr>
<tr>
<td>5</td>
<td>{6, 35, 86, 88, 36, 5}/256</td>
</tr>
<tr>
<td>6</td>
<td>{4, 20, 59, 87, 62, 21, 3}/256</td>
</tr>
<tr>
<td>7</td>
<td>{3, 13, 40, 71, 73, 41, 13, 2}/256</td>
</tr>
<tr>
<td>8</td>
<td>{3, 9, 27, 53, 70, 56, 28, 9, 1}/256</td>
</tr>
<tr>
<td>9</td>
<td>{3, 8, 19, 37, 57, 61, 44, 20, 6, 1}/256</td>
</tr>
<tr>
<td>10</td>
<td>{3, 7, 15, 28, 44, 54, 49, 33, 17, 5, 1}/256</td>
</tr>
<tr>
<td>11</td>
<td>{1, 7, 13, 22, 34, 46, 48, 38, 28, 14, 4, 1}/256</td>
</tr>
<tr>
<td>12</td>
<td>{1, 1, 11, 22, 27, 35, 42, 47, 33, 25, 10, 1, 1}/256</td>
</tr>
<tr>
<td>13</td>
<td>{1, 1, 6, 14, 26, 37, 43, 43, 37, 26, 14, 6, 1, 1}/256</td>
</tr>
<tr>
<td>14</td>
<td>{1, 1, 4, 10, 20, 31, 40, 42, 40, 31, 20, 10, 4, 1, 1}/256</td>
</tr>
<tr>
<td>15</td>
<td>{1, 1, 3, 8, 16, 26, 35, 38, 38, 35, 26, 16, 8, 3, 1, 1}/256</td>
</tr>
<tr>
<td>16</td>
<td>{1, 1, 2, 6, 12, 21, 30, 36, 38, 36, 30, 21, 12, 6, 2, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 48: PDFs for Pulse Count Split, 8 Sample Partitions
<table>
<thead>
<tr>
<th>Pulse Count</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{127, 129}/256</td>
</tr>
<tr>
<td>2</td>
<td>{49, 157, 50}/256</td>
</tr>
<tr>
<td>3</td>
<td>{20, 107, 109, 20}/256</td>
</tr>
<tr>
<td>4</td>
<td>{11, 60, 113, 62, 10}/256</td>
</tr>
<tr>
<td>5</td>
<td>{7, 36, 84, 87, 36, 6}/256</td>
</tr>
<tr>
<td>6</td>
<td>{6, 24, 57, 82, 60, 23, 4}/256</td>
</tr>
<tr>
<td>7</td>
<td>{6, 14, 29, 47, 61, 52, 30, 14, 3}/256</td>
</tr>
<tr>
<td>8</td>
<td>{5, 18, 39, 64, 68, 42, 16, 4}/256</td>
</tr>
<tr>
<td>9</td>
<td>{1, 15, 23, 35, 51, 50, 40, 30, 10, 1}/256</td>
</tr>
<tr>
<td>10</td>
<td>{1, 1, 21, 32, 42, 52, 46, 41, 18, 1, 1}/256</td>
</tr>
<tr>
<td>11</td>
<td>{1, 6, 16, 27, 36, 42, 42, 36, 27, 16, 6, 1}/256</td>
</tr>
<tr>
<td>12</td>
<td>{1, 5, 12, 21, 31, 38, 40, 38, 31, 21, 12, 5, 1}/256</td>
</tr>
<tr>
<td>13</td>
<td>{1, 3, 9, 17, 26, 34, 38, 38, 34, 26, 17, 9, 3, 1}/256</td>
</tr>
<tr>
<td>14</td>
<td>{1, 3, 7, 14, 22, 29, 34, 36, 34, 29, 22, 14, 7, 3, 1}/256</td>
</tr>
<tr>
<td>15</td>
<td>{1, 2, 5, 11, 18, 25, 31, 35, 35, 31, 25, 18, 11, 5, 2, 1}/256</td>
</tr>
<tr>
<td>16</td>
<td>{1, 1, 4, 9, 15, 21, 28, 32, 34, 32, 28, 21, 15, 9, 4, 1, 1}/256</td>
</tr>
</tbody>
</table>

Table 49: PDFs for Pulse Count Split, 4 Sample Partitions
### Table 50: PDFs for Pulse Count Split, 2 Sample Partitions

#### 4.2.7.8.4. LSB Decoding

After the decoder reads the pulse locations for all blocks, it reads the LSBs (if any) for each block in turn. Inside each block, it reads all the LSBs for each coefficient in turn, even those where no
pulses were allocated, before proceeding to the next one. For 10 ms MB frames, it reads LSBs even for the extra 8 samples in the last block. The LSBs are coded from most significant to least significant, and they all use the PDF in Table 51.

<table>
<thead>
<tr>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>{136, 120}/256</td>
</tr>
</tbody>
</table>

Table 51: PDF for Excitation LSBs

The number of LSBs read for each coefficient in a block is determined in Section 4.2.7.8.2. The magnitude of the coefficient is initially equal to the number of pulses placed at that location in Section 4.2.7.8.3. As each LSB is decoded, the magnitude is doubled, and then the value of the LSB added to it, to obtain an updated magnitude.

4.2.7.8.5. Sign Decoding

After decoding the pulse locations and the LSBs, the decoder knows the magnitude of each coefficient in the excitation. It then decodes a sign for all coefficients with a non-zero magnitude, using one of the PDFs from Table 52. If the value decoded is 0, then the coefficient magnitude is negated. Otherwise, it remains positive.

The decoder chooses the PDF for the sign based on the signal type and quantization offset type (from Section 4.2.7.3) and the number of pulses in the block (from Section 4.2.7.8.2). The number of pulses in the block does not take into account any LSBs. Most PDFs are skewed towards negative signs because of the quantization offset, but the PDFs for zero pulses are highly skewed towards positive signs. If a block contains many positive coefficients, it is sometimes beneficial to code it solely using LSBs (i.e., with zero pulses), since the encoder may be able to save enough bits on the signs to justify the less efficient coefficient magnitude encoding.

<table>
<thead>
<tr>
<th>Signal Type</th>
<th>Quantization Offset Type</th>
<th>Pulse Count</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>0</td>
<td>(2, 254)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>1</td>
<td>(207, 49)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>2</td>
<td>(189, 67)/256</td>
</tr>
<tr>
<td>State</td>
<td>Rate</td>
<td>Quality</td>
<td>Rate</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>3</td>
<td>(179, 77)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>4</td>
<td>(174, 82)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>5</td>
<td>(163, 93)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>6 or more</td>
<td>(157, 99)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>0</td>
<td>(58, 198)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>1</td>
<td>(245, 11)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>2</td>
<td>(238, 18)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>3</td>
<td>(232, 24)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>4</td>
<td>(225, 31)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>5</td>
<td>(220, 36)/256</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>6 or more</td>
<td>(211, 45)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>0</td>
<td>(1, 255)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>1</td>
<td>(210, 46)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>2</td>
<td>(190, 66)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>3</td>
<td>(178, 78)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>4</td>
<td>(169, 87)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>5</td>
<td>(162, 94)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>6 or more</td>
<td>(152, 104)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>0</td>
<td>(48, 208)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>1</td>
<td>(242, 14)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>2</td>
<td>(235, 21)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>3</td>
<td>(224, 32)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>4</td>
<td>(214, 42)/256</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>5</td>
<td>(205, 51)/256</td>
</tr>
<tr>
<td>Excitation Type</td>
<td>Range</td>
<td>Probability</td>
<td>Value</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>6 or more</td>
<td>(190, 66)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>0</td>
<td>(1, 255)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>1</td>
<td>(162, 94)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>2</td>
<td>(152, 104)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>3</td>
<td>(147, 109)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>4</td>
<td>(144, 112)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>5</td>
<td>(141, 115)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>6 or more</td>
<td>(138, 118)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>0</td>
<td>(8, 248)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>1</td>
<td>(203, 53)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>2</td>
<td>(187, 69)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>3</td>
<td>(176, 80)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>4</td>
<td>(168, 88)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>5</td>
<td>(161, 95)/256</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>6 or more</td>
<td>(154, 102)/256</td>
</tr>
</tbody>
</table>

Table 52: PDFs for Excitation Signs

4.2.7.8.6. Reconstructing the Excitation

After the signs have been read, there is enough information to reconstruct the complete excitation signal. This requires adding a constant quantization offset to each non-zero sample, and then pseudorandomly inverting and offsetting every sample. The constant quantization offset varies depending on the signal type and quantization offset type (see Section 4.2.7.3).
Let $e_{\text{raw}}[i]$ be the raw excitation value at position $i$, with a magnitude composed of the pulses at that location (see Section 4.2.7.8.3) combined with any additional LSBs (see Section 4.2.7.8.4), and with the corresponding sign decoded in Section 4.2.7.8.5. Additionally, let seed be the current pseudorandom seed, which is initialized to the value decoded from Section 4.2.7.7 for the first sample in the current SILK frame, and updated for each subsequent sample according to the procedure below. Finally, let offset\_Q23 be the quantization offset from Table 53.

Then the following procedure produces the final reconstructed excitation value, $e_{Q23}[i]$:

$$
e_{Q23}[i] = (e_{\text{raw}}[i] << 8) - \text{sign}(e_{\text{raw}}[i]) \times 20 + \text{offset}_{Q23};$$

$$
\text{seed} = (196314165 \times \text{seed} + 907633515) \& 0xFFFFFFFF;
\text{e}_{Q23}[i] = (\text{seed} \& 0x80000000) \neq \text{e}_{Q23}[i] : \text{e}_{Q23}[i];$$

$$
\text{seed} = (\text{seed} + e_{\text{raw}}[i]) \& 0xFFFFFFFF;
$$

When $e_{\text{raw}}[i]$ is zero, $\text{sign}()$ returns 0 by the definition in Section 1.1.4, so the factor of 20 does not get added. The final $e_{Q23}[i]$ value may require more than 16 bits per sample, but will not require more than 23, including the sign.

### 4.2.7.9. SILK Frame Reconstruction

The remainder of the reconstruction process for the frame does not need to be bit-exact, as small errors should only introduce proportionally small distortions. Although the reference implementation only includes a fixed-point version of the remaining steps, this section describes them in terms of a floating-point

<table>
<thead>
<tr>
<th>Signal Type</th>
<th>Quantization Offset Type</th>
<th>Quantization Offset (Q23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>Low</td>
<td>25</td>
</tr>
<tr>
<td>Inactive</td>
<td>High</td>
<td>60</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>Low</td>
<td>25</td>
</tr>
<tr>
<td>Unvoiced</td>
<td>High</td>
<td>60</td>
</tr>
<tr>
<td>Voiced</td>
<td>Low</td>
<td>8</td>
</tr>
<tr>
<td>Voiced</td>
<td>High</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 53: Excitation Quantization Offsets
version for simplicity. This produces a signal with a nominal range of \(-1.0\) to \(1.0\).

`silk_decode_core()` (decode_core.c) contains the code for the main reconstruction process. It proceeds subframe-by-subframe, since quantization gains, LTP parameters, and (in 20 ms SILK frames) LPC coefficients can vary from one to the next.

Let \(a_{Q12}[k]\) be the LPC coefficients for the current subframe. If this is the first or second subframe of a 20 ms SILK frame and the LSF interpolation factor, \(w\_Q2\) (see Section 4.2.7.5.5), is less than 4, then these correspond to the final LPC coefficients produced by Section 4.2.7.5.8 from the interpolated LSF coefficients, \(n1\_Q15[k]\) (computed in Section 4.2.7.5.5). Otherwise, they correspond to the final LPC coefficients produced from the uninterpolated LSF coefficients for the current frame, \(n2\_Q15[k]\).

Also, let \(n\) be the number of samples in a subframe (40 for NB, 60 for MB, and 80 for WB), \(s\) be the index of the current subframe in this SILK frame (0 or 1 for 10 ms frames, or 0 to 3 for 20 ms frames), and \(j\) be the index of the first sample in the residual corresponding to the current subframe.

### 4.2.7.9.1. LTP Synthesis

Voiced SILK frames (see Section 4.2.7.3) pass the excitation through an LTP filter using the parameters decoded in Section 4.2.7.6 to produce an LPC residual. The LTP filter requires LPC residual values from before the current subframe as input. However, since the LPC coefficients may have changed, it obtains this residual by "rewhitening" the corresponding output signal using the LPC coefficients from the current subframe. Let \(\text{out}[i]\) for \((j - \text{pitch\_lags}[s] - d\_LPC - 2) \leq i < j\) be the fully reconstructed output signal from the last \((\text{pitch\_lags}[s] + d\_LPC + 2)\) samples of previous subframes (see Section 4.2.7.9.2), where \(\text{pitch\_lags}[s]\) is the pitch lag for the current subframe from Section 4.2.7.6.1. During reconstruction of the first subframe for this channel after either

- An uncoded regular SILK frame (if this is the side channel), or
- A decoder reset (see Section 4.5.2),

\(\text{out}[\cdot]\) is rewhitened into an LPC residual, \(\text{res}[i]\), via
4.0*LTP_scale_Q14
res[i] = ----------------- * clamp(-1.0, gain_Q16[s])
        d_LPC-1
        a_Q12[k]
out[i] - \ out[i-k-1] * --------, 1.0) .
        /_
        k=0
        4096.0

This requires storage to buffer up to 306 values of out[i] from previous subframes. This corresponds to WB with a maximum pitch lag of 18 ms * 16 kHz samples, plus 16 samples for d_LPC, plus 2 samples for the width of the LTP filter.

Let e_Q23[i] for j <= i < (j + n) be the excitation for the current subframe, and b_Q7[k] for 0 <= k < 5 be the coefficients of the LTP filter taken from the codebook entry in one of Tables 39 through 41 corresponding to the index decoded for the current subframe in Section 4.2.7.6.2. Then for i such that j <= i < (j + n), the LPC residual is

\[
res[i] = \frac{4}{2.0^{**23}} \sum_{k=0}^{4} e_Q23[i] \cdot b_Q7[k] \cdot \text{res[i - pitch_lags[s] + 2 - k]} .
\]

For unvoiced frames, the LPC residual for j <= i < (j + n) is simply a normalized copy of the excitation signal, i.e.,

\[
res[i] = \frac{e_Q23[i]}{2.0^{**23}}
\]

4.2.7.9.2. LPC Synthesis

LPC synthesis uses the short-term LPC filter to predict the next output coefficient. For i such that (j - d_LPC) <= i < j, let lpc[i] be the result of LPC synthesis from the last d_LPC samples of the previous subframe, or zeros in the first subframe for this channel after either

- An uncoded regular SILK frame (if this is the side channel), or
- A decoder reset (see Section 4.5.2).

Then for i such that j <= i < (j + n), the result of LPC synthesis

The current subframe is

\[
\text{lpc}[i] = \frac{\text{gain}_\text{Q16}[i]}{65536.0} \times \text{res}[i] + \sum_{k=0}^{d_\text{LPC}-1} \frac{\text{lpc}[i-k-1]}{4096.0} \times a_\text{Q12}[k].
\]

The decoder saves the final d_LPC values, i.e., lpc[i] such that \((j + n - d_\text{LPC}) \leq i < (j + n)\), to feed into the LPC synthesis of the next subframe. This requires storage for up to 16 values of lpc[i] (for WB frames).

Then, the signal is clamped into the final nominal range:

\[
\text{out}[i] = \text{clamp}(-1.0, \text{lpc}[i], 1.0).
\]

This clamping occurs entirely after the LPC synthesis filter has run. The decoder saves the unclamped values, lpc[i], to feed into the LPC filter for the next subframe, but saves the clamped values, out[i], for rewhitening in voiced frames.

4.2.8. Stereo Unmixing

For stereo streams, after decoding a frame from each channel, the decoder must convert the mid-side (MS) representation into a left-right (LR) representation. The function silk_stereo_MS_to_LR (stereo_MS_to_LR.c) implements this process. In it, the decoder predicts the side channel using a) a simple low-passed version of the mid channel, and b) the unfiltered mid channel, using the prediction weights decoded in Section 4.2.7.1. This simple low-pass filter imposes a one-sample delay, and the unfiltered mid channel is also delayed by one sample. In order to allow seamless switching between stereo and mono, mono streams must also impose the same one-sample delay. The encoder requires an additional one-sample delay for both mono and stereo streams, though an encoder may omit the delay for mono if it knows it will never switch to stereo.

The unmixing process operates in two phases. The first phase lasts for 8 ms, during which it interpolates the prediction weights from the previous frame, prev_w0_Q13 and prev_w1_Q13, to the values for the current frame, w0_Q13 and w1_Q13. The second phase simply uses these weights for the remainder of the frame.

Let mid[i] and side[i] be the contents of out[i] (from Section 4.2.7.9.2) for the current mid and side channels, respectively, and let left[i] and right[i] be the corresponding stereo output channels. If the side channel is not coded (see
Section 4.2.7.2), then side[i] is set to zero. Also let j be defined as in Section 4.2.7.9, n1 be the number of samples in phase 1 (64 for NB, 96 for MB, and 128 for WB), and n2 be the total number of samples in the frame. Then for i such that j <= i < (j + n2), the left and right channel output is

\[
\begin{align*}
\text{prev}_w0_{Q13} & = \frac{\text{w0}_{Q13} - \text{prev}_w0_{Q13}}{8192.0} + \frac{\min(i - j, n1) \cdot \text{p0}}{8192.0 \cdot n1}, \\
\text{prev}_w1_{Q13} & = \frac{\text{w1}_{Q13} - \text{prev}_w1_{Q13}}{8192.0} + \frac{\min(i - j, n1) \cdot \text{p0}}{8192.0 \cdot n1}, \\
\text{p0} & = \frac{\text{mid}[i-2] + 2 \cdot \text{mid}[i-1] + \text{mid}[i]}{4.0}, \\
\text{left}[i] & = \text{clamp}(-1.0, (1 + \text{w1}) \cdot \text{mid}[i-1] + \text{side}[i-1] + \text{w0} \cdot \text{p0}, 1.0), \\
\text{right}[i] & = \text{clamp}(-1.0, (1 - \text{w1}) \cdot \text{mid}[i-1] - \text{side}[i-1] - \text{w0} \cdot \text{p0}, 1.0).
\end{align*}
\]

These formulas require two samples prior to index j, the start of the frame, for the mid channel, and one prior sample for the side channel. For the first frame after a decoder reset, zeros are used instead.

4.2.9. Resampling

After stereo unmixing (if any), the decoder applies resampling to convert the decoded SILK output to the sample rate desired by the application. This is necessary when decoding a Hybrid frame at SWB or FB sample rates, or whenever the decoder wants the output at a different sample rate than the internal SILK sampling rate (e.g., to allow a constant sample rate when the audio bandwidth changes, or to allow mixing with audio from other applications). The resampler itself is non-normative, and a decoder can use any method it wants to perform the resampling.

However, a minimum amount of delay is imposed to allow the resampler to operate, and this delay is normative, so that the corresponding delay can be applied to the MDCT layer in the encoder. A decoder is always free to use a resampler which requires more delay than allowed for here (e.g., to improve quality), but it must then delay the output of the MDCT layer by this extra amount. Keeping as much delay as possible on the encoder side allows an encoder which knows it will never use any of the SILK or Hybrid modes to skip this delay. By contrast, if it were all applied by the decoder, then a decoder which
processes audio in fixed-size blocks would be forced to delay the output of CELT frames just in case of a later switch to a SILK or Hybrid mode.

Table 54 gives the maximum resampler delay in samples at 48 kHz for each SILK audio bandwidth. Because the actual output rate may not be 48 kHz, it may not be possible to achieve exactly these delays while using a whole number of input or output samples. The reference implementation is able to resample to any of the supported output sampling rates (8, 12, 16, 24, or 48 kHz) within or near this delay constraint. Some resampling filters (including those used by the reference implementation) may add a delay that is not an exact integer, or is not linear-phase, and so cannot be represented by a single delay at all frequencies. However, such deviations are unlikely to be perceptible, and the comparison tool described in Section 6 is designed to be relatively insensitive to them. The delays listed here are the ones that should be targeted by the encoder.

+-----------------+----------------------+
| Audio Bandwidth | Delay in millisecond |
|-----------------+----------------------|
| NB              | 0.538                |
| MB              | 0.692                |
| WB              | 0.706                |
+-----------------+----------------------+

Table 54: SILK Resampler Delay Allocations

NB is given a smaller decoder delay allocation than MB and WB to allow a higher-order filter when resampling to 8 kHz in both the encoder and decoder. This implies that the audio content of two SILK frames operating at different bandwidths are not perfectly aligned in time. This is not an issue for any transitions described in Section 4.5, because they all involve a SILK decoder reset. When the decoder is reset, any samples remaining in the resampling buffer are discarded, and the resampler is re-initialized with silence.

4.3. CELT Decoder

The CELT layer of Opus is based on the Modified Discrete Cosine Transform [MDCT] with partially overlapping windows of 5 to 22.5 ms. The main principle behind CELT is that the MDCT spectrum is divided into bands that (roughly) follow the Bark scale, i.e., the scale of the ear's critical bands [Zwicker61]. The normal CELT layer uses 21 of those bands, though Opus Custom (see Section 6.2) may use a
different number of bands. In Hybrid mode, the first 17 bands (up to 8 kHz) are not coded. A band can contain as little as one MDCT bin per channel, and as many as 176 bins per channel, as detailed in Table 55. In each band, the gain (energy) is coded separately from the shape of the spectrum. Coding the gain explicitly makes it easy to preserve the spectral envelope of the signal. The remaining unit-norm shape vector is encoded using a Pyramid Vector Quantizer (PVQ) Section 4.3.4.

<table>
<thead>
<tr>
<th>Band</th>
<th>Bins:</th>
<th>2.5 ms</th>
<th>5 ms</th>
<th>10 ms</th>
<th>20 ms</th>
<th>Start Frequency</th>
<th>Stop Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>0 Hz</td>
<td>200 Hz</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>200 Hz</td>
<td>400 Hz</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>400 Hz</td>
<td>600 Hz</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>600 Hz</td>
<td>800 Hz</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>800 Hz</td>
<td>1000 Hz</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>1000 Hz</td>
<td>1200 Hz</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>1200 Hz</td>
<td>1400 Hz</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>1400 Hz</td>
<td>1600 Hz</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>1600 Hz</td>
<td>2000 Hz</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>2000 Hz</td>
<td>2400 Hz</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>2400 Hz</td>
<td>2800 Hz</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>2800 Hz</td>
<td>3200 Hz</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>3200 Hz</td>
<td>4000 Hz</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>4000 Hz</td>
<td>4800 Hz</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>4800 Hz</td>
<td>5600 Hz</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>12</td>
<td>24</td>
<td>48</td>
<td>5600 Hz</td>
<td>6800 Hz</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>12</td>
<td>24</td>
<td>48</td>
<td>6800 Hz</td>
<td>8000 Hz</td>
<td></td>
</tr>
</tbody>
</table>
Table 55: MDCT Bins Per Channel Per Band for Each Frame Size

Transients are notoriously difficult for transform codecs to code. CELT uses two different strategies for them:

1. Using multiple smaller MDCTs instead of a single large MDCT, and
2. Dynamic time-frequency resolution changes (See Section 4.3.4.5).

To improve quality on highly tonal and periodic signals, CELT includes a prefilter/postfilter combination. The prefilter on the encoder side attenuates the signal’s harmonics. The postfilter on the decoder side restores the original gain of the harmonics, while shaping the coding noise to roughly follow the harmonics. Such noise shaping reduces the perception of the noise.

When coding a stereo signal, three coding methods are available:

- mid-side stereo: encodes the mean and the difference of the left and right channels,
- intensity stereo: only encodes the mean of the left and right channels (discards the difference),
- dual stereo: encodes the left and right channels separately.

An overview of the decoder is given in Figure 17.
Figure 17: Structure of the CELT decoder

The decoder is based on the following symbols and sets of symbols:
<table>
<thead>
<tr>
<th>Symbol(s)</th>
<th>PDF</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>silence</td>
<td>{32767, 1}/32768</td>
<td></td>
</tr>
<tr>
<td>post-filter</td>
<td>{1, 1}/2</td>
<td>post-filter</td>
</tr>
<tr>
<td>octave</td>
<td>uniform (6)</td>
<td>post-filter</td>
</tr>
<tr>
<td>period</td>
<td>raw bits (4+octave)</td>
<td>post-filter</td>
</tr>
<tr>
<td>gain</td>
<td>raw bits (3)</td>
<td>post-filter</td>
</tr>
<tr>
<td>tapset</td>
<td>{2, 1, 1}/4</td>
<td>post-filter</td>
</tr>
<tr>
<td>transient</td>
<td>{7, 1}/8</td>
<td></td>
</tr>
<tr>
<td>intra</td>
<td>{7, 1}/8</td>
<td></td>
</tr>
<tr>
<td>coarse energy</td>
<td>Section 4.3.2</td>
<td></td>
</tr>
<tr>
<td>tf_change</td>
<td>Section 4.3.1</td>
<td></td>
</tr>
<tr>
<td>tf_select</td>
<td>{1, 1}/2</td>
<td>Section 4.3.1</td>
</tr>
<tr>
<td>spread</td>
<td>{7, 2, 21, 2}/32</td>
<td></td>
</tr>
<tr>
<td>dyn. alloc.</td>
<td>Section 4.3.3</td>
<td></td>
</tr>
<tr>
<td>alloc. trim</td>
<td>{2, 2, 5, 10, 22, 46, 22, 10, 5, 2, 2}/128</td>
<td></td>
</tr>
<tr>
<td>skip</td>
<td>{1, 1}/2</td>
<td>Section 4.3.3</td>
</tr>
<tr>
<td>intensity</td>
<td>uniform</td>
<td>Section 4.3.3</td>
</tr>
<tr>
<td>dual</td>
<td>{1, 1}/2</td>
<td></td>
</tr>
<tr>
<td>fine energy</td>
<td>Section 4.3.2</td>
<td></td>
</tr>
<tr>
<td>residual</td>
<td>Section 4.3.4</td>
<td></td>
</tr>
<tr>
<td>anti-collapse</td>
<td>{1, 1}/2</td>
<td>Section 4.3.5</td>
</tr>
<tr>
<td>finalize</td>
<td>Section 4.3.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 56: Order of the Symbols in the CELT Section of the Bitstream
The decoder extracts information from the range-coded bitstream in the order described in Table 56. In some circumstances, it is possible for a decoded value to be out of range due to a very small amount of redundancy in the encoding of large integers by the range coder. In that case, the decoder should assume there has been an error in the coding, decoding, or transmission and SHOULD take measures to conceal the error and/or report to the application that a problem has occurred. Such out of range errors cannot occur in the SILK layer.

4.3.1. Transient Decoding

The "transient" flag indicates whether the frame uses a single long MDCT or several short MDCTs. When it is set, then the MDCT coefficients represent multiple short MDCTs in the frame. When not set, the coefficients represent a single long MDCT for the frame. The flag is encoded in the bitstream with a probability of 1/8. In addition to the global transient flag is a per-band binary flag to change the time-frequency (tf) resolution independently in each band. The change in tf resolution is defined in tf_select_table[] in celt.c and depends on the frame size, whether the transient flag is set, and the value of tf_select. The tf_select flag uses a 1/2 probability, but is only decoded if it can have an impact on the result knowing the value of all per-band tf_change flags.

4.3.2. Energy Envelope Decoding

It is important to quantize the energy with sufficient resolution because any energy quantization error cannot be compensated for at a later stage. Regardless of the resolution used for encoding the spectral shape of a band, it is perceptually important to preserve the energy in each band. CELT uses a three-step coarse-fine-fine strategy for encoding the energy in the base-2 log domain, as implemented in quant_bands.c

4.3.2.1. Coarse energy decoding

Coarse quantization of the energy uses a fixed resolution of 6 dB (integer part of base-2 log). To minimize the bitrate, prediction is applied both in time (using the previous frame) and in frequency (using the previous bands). The part of the prediction that is based on the previous frame can be disabled, creating an "intra" frame where the energy is coded without reference to prior frames. The decoder first reads the intra flag to determine what prediction is used. The 2-D z-transform [z-transform] of the prediction filter is:
\[ A(z_l, z_b) = \frac{-(1 - \alpha z_{l-1} )*(1 - z_b )}{1 - \beta z_b} \]

where \( b \) is the band index and \( l \) is the frame index. The prediction coefficients applied depend on the frame size in use when not using intra energy and are \( \alpha = 0, \beta = 4915/32768 \) when using intra energy. The time-domain prediction is based on the final fine quantization of the previous frame, while the frequency domain (within the current frame) prediction is based on coarse quantization only (because the fine quantization has not been computed yet). The prediction is clamped internally so that fixed point implementations with limited dynamic range always remain in the same state as floating point implementations. We approximate the ideal probability distribution of the prediction error using a Laplace distribution with separate parameters for each frame size in intra- and inter-frame modes. These parameters are held in the e_prob_model table in quant_bands.c. The coarse energy quantization is performed by \texttt{unquant_coarse_energy()} and \texttt{unquant_coarse_energy_impl()} (quant_bands.c). The encoding of the Laplace-distributed values is implemented in \texttt{ec_laplace_decode() (laplace.c)}.

4.3.2.2. Fine energy quantization

The number of bits assigned to fine energy quantization in each band is determined by the bit allocation computation described in Section 4.3.3. Let \( B_i \) be the number of fine energy bits for band \( i \); the refinement is an integer \( f \) in the range \([0, 2^{B_i}-1]\). The mapping between \( f \) and the correction applied to the coarse energy is equal to \((f+1/2)/2^{B_i} - 1/2\). Fine energy quantization is implemented in \texttt{quant_fine_energy()} (quant_bands.c).

When some bits are left "unused" after all other flags have been decoded, these bits are assigned to a "final" step of fine allocation. In effect, these bits are used to add one extra fine energy bit per band per channel. The allocation process determines two "priorities" for the final fine bits. Any remaining bits are first assigned only to bands of priority 0, starting from band 0 and going up. If all bands of priority 0 have received one bit per channel, then bands of priority 1 are assigned an extra bit per channel, starting from band 0. If any bits are left after this, they are left unused. This is implemented in \texttt{unquant_energy_finalise()} (quant_bands.c).
4.3.3. Bit Allocation

Because the bit allocation drives the decoding of the range-coder stream, it MUST be recovered exactly so that identical coding decisions are made in the encoder and decoder. Any deviation from the reference’s resulting bit allocation will result in corrupted output, though implementers are free to implement the procedure in any way which produces identical results.

The per-band gain-shape structure of the CELT layer ensures that using the same number of bits for the spectral shape of a band in every frame will result in a roughly constant signal-to-noise ratio in that band. This results in coding noise that has the same spectral envelope as the signal. The masking curve produced by a standard psychoacoustic model also closely follows the spectral envelope of the signal. This structure means that the ideal allocation is more consistent from frame to frame than it is for other codecs without an equivalent structure, and that a fixed allocation provides fairly consistent perceptual performance [Valin2010].

Many codecs transmit significant amounts of side information to control the bit allocation within a frame. Often this control is only indirect, and must be exercised carefully to achieve the desired rate constraints. The CELT layer, however, can adapt over a very wide range of rates, and thus has a large number of codebook sizes to choose from for each band. Explicitly signaling the size of each of these codebooks would impose considerable overhead, even though the allocation is relatively static from frame to frame. This is because all of the information required to compute these codebook sizes must be derived from a single frame by itself, in order to retain robustness to packet loss, so the signaling cannot take advantage of knowledge of the allocation in neighboring frames. This problem is exacerbated in low-latency (small frame size) applications, which would include this overhead in every frame.

For this reason, in the MDCT mode Opus uses a primarily implicit bit allocation. The available bitstream capacity is known in advance to both the encoder and decoder without additional signaling, ultimately from the packet sizes expressed by a higher-level protocol. Using this information, the codec interpolates an allocation from a hard-coded table.

While the band-energy structure effectively models intra-band masking, it ignores the weaker inter-band masking, band-temporal masking, and other less significant perceptual effects. While these effects can often be ignored, they can become significant for particular samples. One mechanism available to encoders would be to
simply increase the overall rate for these frames, but this is not possible in a constant rate mode and can be fairly inefficient. As a result three explicitly signaled mechanisms are provided to alter the implicit allocation:

- **Band boost**
- **Allocation trim**
- **Band skipping**

The first of these mechanisms, band boost, allows an encoder to boost the allocation in specific bands. The second, allocation trim, works by biasing the overall allocation towards higher or lower frequency bands. The third, band skipping, selects which low-precision high frequency bands will be allocated no shape bits at all.

In stereo mode there are two additional parameters potentially coded as part of the allocation procedure: a parameter to allow the selective elimination of allocation for the 'side' (i.e., intensity stereo) in jointly coded bands, and a flag to deactivate joint coding (i.e., dual stereo). These values are not signaled if they would be meaningless in the overall context of the allocation.

Because every signaled adjustment increases overhead and implementation complexity, none were included speculatively: the reference encoder makes use of all of these mechanisms. While the decision logic in the reference was found to be effective enough to justify the overhead and complexity, further analysis techniques may be discovered which increase the effectiveness of these parameters. As with other signaled parameters, an encoder is free to choose the values in any manner, but unless a technique is known to deliver superior perceptual results the methods used by the reference implementation should be used.

The allocation process consists of the following steps: determining the per-band maximum allocation vector, decoding the boosts, decoding the tilt, determining the remaining capacity of the frame, searching the mode table for the entry nearest but not exceeding the available space (subject to the tilt, boosts, band maximums, and band minimums), linear interpolation, reallocation of unused bits with concurrent skip decoding, determination of the fine-energy vs. shape split, and final reallocation. This process results in a per-band shape allocation (in 1/8th bit units), a per-band fine-energy allocation (in 1 bit per channel units), a set of band priorities for controlling the use of remaining bits at the end of the frame, and a remaining balance of unallocated space, which is usually zero except at very high rates.
The "static" bit allocation (in 1/8 bits) for a quality q, excluding the minimums, maximums, tilt and boosts, is equal to
channels*N*alloc[band][q]<<LM>>2, where alloc[][] is given in
Table 57 and LM=log2(frame_size/120). The allocation is obtained by
linearly interpolating between two values of q (in steps of 1/64) to
find the highest allocation that does not exceed the number of bits
remaining.

Rows indicate the MDCT bands, columns are the different quality (q)
parameters. The units are 1/32 bit per MDCT bin.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>90</td>
<td>110</td>
<td>118</td>
<td>126</td>
<td>134</td>
<td>144</td>
<td>152</td>
<td>162</td>
<td>172</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>80</td>
<td>100</td>
<td>110</td>
<td>119</td>
<td>127</td>
<td>137</td>
<td>145</td>
<td>155</td>
<td>165</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>75</td>
<td>90</td>
<td>103</td>
<td>112</td>
<td>120</td>
<td>130</td>
<td>138</td>
<td>148</td>
<td>158</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>69</td>
<td>84</td>
<td>93</td>
<td>104</td>
<td>114</td>
<td>124</td>
<td>132</td>
<td>142</td>
<td>152</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>63</td>
<td>78</td>
<td>86</td>
<td>95</td>
<td>103</td>
<td>113</td>
<td>123</td>
<td>133</td>
<td>143</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>56</td>
<td>71</td>
<td>80</td>
<td>89</td>
<td>97</td>
<td>107</td>
<td>117</td>
<td>127</td>
<td>137</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>49</td>
<td>65</td>
<td>75</td>
<td>83</td>
<td>91</td>
<td>101</td>
<td>111</td>
<td>121</td>
<td>131</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>40</td>
<td>58</td>
<td>70</td>
<td>78</td>
<td>85</td>
<td>95</td>
<td>105</td>
<td>115</td>
<td>125</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>34</td>
<td>51</td>
<td>65</td>
<td>72</td>
<td>78</td>
<td>88</td>
<td>98</td>
<td>108</td>
<td>118</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>29</td>
<td>45</td>
<td>59</td>
<td>66</td>
<td>72</td>
<td>82</td>
<td>92</td>
<td>102</td>
<td>112</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>20</td>
<td>39</td>
<td>53</td>
<td>60</td>
<td>66</td>
<td>76</td>
<td>86</td>
<td>96</td>
<td>106</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>32</td>
<td>47</td>
<td>54</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>183</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>26</td>
<td>40</td>
<td>47</td>
<td>54</td>
<td>64</td>
<td>74</td>
<td>84</td>
<td>94</td>
<td>178</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>20</td>
<td>31</td>
<td>39</td>
<td>47</td>
<td>57</td>
<td>67</td>
<td>77</td>
<td>87</td>
<td>173</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>12</td>
<td>23</td>
<td>32</td>
<td>41</td>
<td>51</td>
<td>61</td>
<td>71</td>
<td>81</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>25</td>
<td>35</td>
<td>45</td>
<td>55</td>
<td>65</td>
<td>75</td>
<td>163</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>17</td>
<td>29</td>
<td>39</td>
<td>49</td>
<td>59</td>
<td>69</td>
<td>158</td>
<td></td>
</tr>
</tbody>
</table>
Table 57: CELT Static Allocation Table

The maximum allocation vector is an approximation of the maximum space that can be used by each band for a given mode. The value is approximate because the shape encoding is variable rate (due to entropy coding of splitting parameters). Setting the maximum too low reduces the maximum achievable quality in a band while setting it too high may result in waste: bitstream capacity available at the end of the frame which can not be put to any use. The maximums specified by the codec reflect the average maximum. In the reference implementation, the maximums in bits/sample are precomputed in a static table (see cache_caps50[] in static_modes_float.h) for each band, for each value of LM, and for both mono and stereo. Implementations are expected to simply use the same table data, but the procedure for generating this table is included in rate.c as part of compute_pulse_cache().

To convert the values in cache.caps into the actual maximums: first set nbBands to the maximum number of bands for this mode, and stereo to zero if stereo is not in use and one otherwise. For each band set N to the number of MDCT bins covered by the band (for one channel), set LM to the shift value for the frame size, then set i to nbBands*(2*LM+stereo). Then set the maximum for the band to the i-th index of cache.caps + 64 and multiply by the number of channels in the current frame (one or two) and by N, then divide the result by 4 using integer division. The resulting vector will be called cap[]. The elements fit in signed 16-bit integers but do not fit in 8 bits. This procedure is implemented in the reference in the function init_caps() in celt.c.

The band boosts are represented by a series of binary symbols which are entropy coded with very low probability. Each band can potentially be boosted multiple times, subject to the frame actually having enough room to obey the boost and having enough room to code the boost symbol. The default coding cost for a boost starts out at six bits (probability p=1/64), but subsequent boosts in a band cost only a single bit and every time a band is boosted the initial cost is reduced (down to a minimum of two bits, or p=1/4). Since the initial cost of coding a boost is 6 bits, the coding cost of the
boost symbols when completely unused is 0.48 bits/frame for a 21 band mode \((21**-\log_2(1-1/2**6))\).

To decode the band boosts: First set 'dynalloc_logp' to 6, the initial amount of storage required to signal a boost in bits, 'total_bits' to the size of the frame in 8th bits, 'total_boost' to zero, and 'tell' to the total number of 8th bits decoded so far. For each band from the coding start (0 normally, but 17 in Hybrid mode) to the coding end (which changes depending on the signaled bandwidth), the boost quanta in units of 1/8 bit is calculated as quanta = \(\min(8*N, \max(48, N))\). This represents a boost step size of six bits, subject to a lower limit of 1/8th bit/sample and an upper limit of 1 bit/sample. Set 'boost' to zero and 'dynalloc_loop_logp' to dynalloc_logp. While dynalloc_loop_log (the current worst case symbol cost) in 8th bits plus tell is less than total_bits plus total_boost and boost is less than cap[] for this band: Decode a bit from the bitstream with a with dynalloc_loop_logp as the cost of a one, update tell to reflect the current used capacity, if the decoded value is zero break the loop otherwise add quanta to boost and total_boost, subtract quanta from total_bits, and set dynalloc_loop_log to 1. When the while loop finishes boost contains the boost for this band. If boost is non-zero and dynalloc_logp is greater than 2, decrease dynalloc_logp. Once this process has been executed on all bands, the band boosts have been decoded. This procedure is implemented around line 2474 of celt.c.

At very low rates it is possible that there won’t be enough available space to execute the inner loop even once. In these cases band boost is not possible but its overhead is completely eliminated. Because of the high cost of band boost when activated, a reasonable encoder should not be using it at very low rates. The reference implements its dynalloc decision logic around line 1304 of celt.c.

The allocation trim is a integer value from 0-10. The default value of 5 indicates no trim. The trim parameter is entropy coded in order to lower the coding cost of less extreme adjustments. Values lower than 5 bias the allocation towards lower frequencies and values above 5 bias it towards higher frequencies. Like other signaled parameters, signaling of the trim is gated so that it is not included if there is insufficient space available in the bitstream. To decode the trim, first set the trim value to 5, then if and only if the count of decoded 8th bits so far (ec.tell_frac) plus 48 (6 bits) is less than or equal to the total frame size in 8th bits minus total_boost (a product of the above band boost procedure), decode the trim value using the PDF in Table 58.
Table 58: PDF for the Trim

For 10 ms and 20 ms frames using short blocks and that have at least LM+2 bits left prior to the allocation process, then one anti-collapse bit is reserved in the allocation process so it can be decoded later. Following the the anti-collapse reservation, one bit is reserved for skip if available.

For stereo frames, bits are reserved for intensity stereo and for dual stereo. Intensity stereo requires ilog2(end-start) bits. Those bits are reserved if there is enough bits left. Following this, one bit is reserved for dual stereo if available.

The allocation computation begins by setting up some initial conditions. 'total' is set to the remaining available 8th bits, computed by taking the size of the coded frame times 8 and subtracting ec_tell_frac(). From this value, one (8th bit) is subtracted to ensure that the resulting allocation will be conservative. 'antiCollapse_rsv' is set to 8 (8th bits) if and only if the frame is a transient, LM is greater than 1, and total is greater than or equal to (LM+2) * 8. Total is then decremented by antiCollapse_rsv and clamped to be equal to or greater than zero.

'skip_rsv' is set to 8 (8th bits) if total is greater than 8, otherwise it is zero. Total is then decremented by skip_rsv. This reserves space for the final skipping flag.

If the current frame is stereo, intensity_rsv is set to the conservative log2 in 8th bits of the number of coded bands for this frame (given by the table LOG2_FRAC_TABLE in rate.c). If intensity_rsv is greater than total then intensity_rsv is set to zero. Otherwise total is decremented by intensity_rsv, and if total is still greater than 8, dual_stereo_rsv is set to 8 and total is decremented by dual_stereo_rsv.

The allocation process then computes a vector representing the hard minimum amounts allocation any band will receive for shape. This minimum is higher than the technical limit of the PVQ process, but very low rate allocations produce an excessively sparse spectrum and these bands are better served by having no allocation at all. For each coded band, set thresh[band] to twenty-four times the number of MDCT bins in the band and divide by 16. If 8 times the number of channels is greater, use that instead. This sets the minimum
allocation to one bit per channel or 48 128th bits per MDCT bin, whichever is greater. The band-size dependent part of this value is not scaled by the channel count, because at the very low rates where this limit is applicable there will usually be no bits allocated to the side.

The previously decoded allocation trim is used to derive a vector of per-band adjustments, 'trim_offsets[]'. For each coded band take the alloc_trim and subtract 5 and LM. Then multiply the result by the number of channels, the number of MDCT bins in the shortest frame size for this mode, the number of remaining bands, $2^{**LM}$, and 8. Then divide this value by 64. Finally, if the number of MDCT bins in the band per channel is only one, 8 times the number of channels is subtracted in order to diminish the allocation by one bit, because width 1 bands receive greater benefit from the coarse energy coding.

4.3.4. Shape Decoding

In each band, the normalized "shape" is encoded using a vector quantization scheme called a "pyramid vector quantizer".

In the simplest case, the number of bits allocated in Section 4.3.3 is converted to a number of pulses as described by Section 4.3.4.1. Knowing the number of pulses and the number of samples in the band, the decoder calculates the size of the codebook as detailed in Section 4.3.4.2. The size is used to decode an unsigned integer (uniform probability model), which is the codeword index. This index is converted into the corresponding vector as explained in Section 4.3.4.2. This vector is then scaled to unit norm.

4.3.4.1. Bits to Pulses

Although the allocation is performed in 1/8th bit units, the quantization requires an integer number of pulses K. To do this, the encoder searches for the value of K that produces the number of bits nearest to the allocated value (rounding down if exactly halfway between two values), not to exceed the total number of bits available. For efficiency reasons, the search is performed against a precomputed allocation table which only permits some K values for each N. The number of codebook entries can be computed as explained in Section 4.3.4.2. The difference between the number of bits allocated and the number of bits used is accumulated to a "balance" (initialized to zero) that helps adjust the allocation for the next bands. One third of the balance is applied to the bit allocation of each band to help achieve the target allocation. The only exceptions are the band before the last and the last band, for which half the balance and the whole balance are applied, respectively.
4.3.4.2. PVQ Decoding

Decoding of PVQ vectors is implemented in decode_pulses() (cwrs.c). The unique codeword index is decoded as a uniformly-distributed integer value between 0 and V(N,K)-1, where V(N,K) is the number of possible combinations of K pulses in N samples. The index is then converted to a vector in the same way specified in [PVQ]. The indexing is based on the calculation of V(N,K) (denoted N(L,K) in [PVQ]).

The number of combinations can be computed recursively as V(N,K) = V(N-1,K) + V(N,K-1) + V(N-1,K-1), with V(N,0) = 1 and V(0,K) = 0, K != 0. There are many different ways to compute V(N,K), including precomputed tables and direct use of the recursive formulation. The reference implementation applies the recursive formulation one line (or column) at a time to save on memory use, along with an alternate, univariate recurrence to initialize an arbitrary line, and direct polynomial solutions for small N. All of these methods are equivalent, and have different trade-offs in speed, memory usage, and code size. Implementations MAY use any methods they like, as long as they are equivalent to the mathematical definition.

The decoded vector X is recovered as follows. Let i be the index decoded with the procedure in Section 4.1.5 with ft = V(N,K), so that 0 <= i < V(N,K). Let k = K. Then for j = 0 to (N - 1), inclusive, do:

1. Let p = (V(N-j-1,k) + V(N-j,k))/2.
2. If i < p, then let sgn = 1, else let sgn = -1 and set i = i - p.
3. Let k0 = k and set p = p - V(N-j-1,k).
4. While p > i, set k = k - 1 and p = p - V(N-j-1,k).
5. Set X[j] = sgn*(k0 - k) and i = i - p.

The decoded vector X is then normalized such that its L2-norm equals one.

4.3.4.3. Spreading

The normalized vector decoded in Section 4.3.4.2 is then rotated for the purpose of avoiding tonal artifacts. The rotation gain is equal to

\[ g_r = N / (N + f_r*K) \]
where $N$ is the number of dimensions, $K$ is the number of pulses, and $f_r$ depends on the value of the "spread" parameter in the bit-stream.

<table>
<thead>
<tr>
<th>Spread value</th>
<th>$f_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>infinite (no rotation)</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 59: Spreading Values

The rotation angle is then calculated as

$$
\theta = \frac{\pi}{4} \frac{g_r}{2}
$$

A 2-D rotation $R(i,j)$ between points $x_i$ and $x_j$ is defined as:

$$
x_i' = \cos(\theta) x_i + \sin(\theta) x_j
$$
$$
x_j' = -\sin(\theta) x_i + \cos(\theta) x_j
$$

An N-D rotation is then achieved by applying a series of 2-D rotations back and forth, in the following order: $R(x_1, x_2), R(x_2, x_3), \ldots, R(x_{N-2}, X_{N-1}), R(x_{N-1}, X_N), R(x_N, X_{N-1}), \ldots, R(x_1, x_2)$.

If the decoded vector represents more than one time block, then this spreading process is applied separately on each time block. Also, if each block represents 8 samples or more, then another N-D rotation, by $(\pi/2-\theta)$, is applied _before_ the rotation described above. This extra rotation is applied in an interleaved manner with a stride equal to $\text{round}(\sqrt{N/nb\text{blocks}})$, i.e., it is applied independently for each set of sample $S_k = \{\text{stride}n + k\}, n=0..N/\text{stride}-1$.

4.3.4.4. Split decoding

To avoid the need for multi-precision calculations when decoding PVQ codevectors, the maximum size allowed for codebooks is 32 bits. When larger codebooks are needed, the vector is instead split in two sub-vectors of size $N/2$. A quantized gain parameter with precision
derived from the current allocation is entropy coded to represent the relative gains of each side of the split, and the entire decoding process is recursively applied. Multiple levels of splitting may be applied up to a limit of \( LM+1 \) splits. The same recursive mechanism is applied for the joint coding of stereo audio.

4.3.4.5. Time-Frequency change

The time-frequency (TF) parameters are used to control the time-frequency resolution tradeoff in each coded band. For each band, there are two possible TF choices. For the first band coded, the PDF is \( \{3, 1\}/4 \) for frames marked as transient and \( \{15, 1\}/16 \) for the other frames. For subsequent bands, the TF choice is coded relative to the previous TF choice with probability \( \{15, 1\}/15 \) for transient frames and \( \{31, 1\}/32 \) otherwise. The mapping between the decoded TF choices and the adjustment in TF resolution is shown in the tables below.

| Frame size (ms) | 0  | 1  |
|-----------------+---+----|
| 2.5             | 0  | -1 |
| 5               | 0  | -1 |
| 10              | 0  | -2 |
| 20              | 0  | -2 |

Table 60: TF Adjustments for Non-transient Frames and tf_select=0

| Frame size (ms) | 0  | 1  |
|-----------------+---+----|
| 2.5             | 0  | -2 |
| 5               | 0  | -2 |
| 10              | 0  | -3 |
| 20              | 0  | -3 |

Table 61: TF Adjustments for Non-transient Frames and tf_select=1
A negative TF adjustment means that the temporal resolution is increased, while a positive TF adjustment means that the frequency resolution is increased. Changes in TF resolution are implemented using the Hadamard transform \([\text{Hadamard}]\). To increase the time resolution by \(N\), \(N\) "levels" of the Hadamard transform are applied to the decoded vector for each interleaved MDCT vector. To increase the frequency resolution (assumes a transient frame), then \(N\) levels of the Hadamard transform are applied _across_ the interleaved MDCT vector. In the case of increased time resolution the decoder uses the "sequency order" because the input vector is sorted in time.

### 4.3.5. Anti-Collapse Processing

The anti-collapse feature is designed to avoid the situation where the use of multiple short MDCTs causes the energy in one or more of the MDCTs to be zero for some bands, causing unpleasant artifacts. When the frame has the transient bit set, an anti-collapse bit is decoded. When anti-collapse is set, the energy in each small MDCT is prevented from collapsing to zero. For each band of each MDCT where a collapse is detected, a pseudo-random signal is inserted with an
energy corresponding to the minimum energy over the two previous frames. A renormalization step is then required to ensure that the anti-collapse step did not alter the energy preservation property.

4.3.6. Denormalization

Just as each band was normalized in the encoder, the last step of the decoder before the inverse MDCT is to denormalize the bands. Each decoded normalized band is multiplied by the square root of the decoded energy. This is done by denormalise_bands() (bands.c).

4.3.7. Inverse MDCT

The inverse MDCT implementation has no special characteristics. The input is N frequency-domain samples and the output is 2*N time-domain samples, while scaling by 1/2. A "low-overlap" window reduces the algorithmic delay. It is derived from a basic (full overlap) 240-sample version of the window used by the Vorbis codec:

$$W(n) = \left| \sin \frac{\pi}{2} \cdot \sin \frac{n + 1/2}{\pi} \right| \cdot \sin \frac{2}{2} \cdot \sin \frac{2}{L} \cdot \frac{2}{L}$$

The low-overlap window is created by zero-padding the basic window and inserting ones in the middle, such that the resulting window still satisfies power complementarity [Princen86]. The IMDCT and windowing are performed by mdct_backward (mdct.c).

4.3.7.1. Post-filter

The output of the inverse MDCT (after weighted overlap-add) is sent to the post-filter. Although the post-filter is applied at the end, the post-filter parameters are encoded at the beginning, just after the silence flag. The post-filter can be switched on or off using one bit (logp=1). If the post-filter is enabled, then the octave is decoded as an integer value between 0 and 6 of uniform probability. Once the octave is known, the fine pitch within the octave is decoded using 4+octave raw bits. The final pitch period is equal to $16 \cdot \text{octave} + \text{fine_pitch} - 1$ so it is bounded between 15 and 1022, inclusively. Next, the gain is decoded as three raw bits and is equal to $G = 3 \cdot (\text{int_gain} + 1)/32$. The set of post-filter taps is decoded last, using a pdf equal to $\{2, 1, 1\}/4$. Tapset zero corresponds to the filter coefficients $g_0 = 0.3066406250$, $g_1 = 0.2170410156$, $g_2 = 0.1296386719$. Tapset one corresponds to the filter coefficients $g_0 = 0.4638671875$, $g_1 = 0.2680664062$, $g_2 = 0$, and tapset two uses filter coefficients $g_0 = 0.7998046875$, $g_1 = 0.1000976562$, $g_2 = 0$. 
The post-filter response is thus computed as:

\[
y(n) = x(n) + G*(g0*y(n-T) + g1*(y(n-T+1)+y(n-T+1)) + g2*(y(n-T+2)+y(n-T+2)))
\]

During a transition between different gains, a smooth transition is calculated using the square of the MDCT window. It is important that values of \(y(n)\) be interpolated one at a time such that the past value of \(y(n)\) used is interpolated.

4.3.7.2. De-emphasis

After the post-filter, the signal is de-emphasized using the inverse of the pre-emphasis filter used in the encoder:

\[
\frac{1}{A(z)} = \frac{1}{1 - \alpha_p z^{-1}}
\]

where \(\alpha_p = 0.8500061035\).

4.4. Packet Loss Concealment (PLC)

Packet loss concealment (PLC) is an optional decoder-side feature that SHOULD be included when receiving from an unreliable channel. Because PLC is not part of the bitstream, there are many acceptable ways to implement PLC with different complexity/quality trade-offs.

The PLC in the reference implementation depends on the mode of last packet received. In CELT mode, the PLC finds a periodicity in the decoded signal and repeats the windowed waveform using the pitch offset. The windowed waveform is overlapped in such a way as to preserve the time-domain aliasing cancellation with the previous frame and the next frame. This is implemented in \texttt{celt_decode_lost()} (\texttt{mdct.c}). In SILK mode, the PLC uses LPC extrapolation from the previous frame, implemented in \texttt{silk_PLC()} (\texttt{PLC.c}).

4.4.1. Clock Drift Compensation

Clock drift refers to the gradual desynchronization of two endpoints whose sample clocks run at different frequencies while they are streaming live audio. Differences in clock frequencies are generally attributable to manufacturing variation in the endpoints’ clock hardware. For long-lived streams, the time difference between sender and receiver can grow without bound.
When the sender’s clock runs slower than the receiver’s, the effect is similar to packet loss: too few packets are received. The receiver can distinguish between drift and loss if the transport provides packet timestamps. A receiver for live streams SHOULD conceal the effects of drift, and MAY do so by invoking the PLC.

When the sender’s clock runs faster than the receiver’s, too many packets will be received. The receiver MAY respond by skipping any packet (i.e., not submitting the packet for decoding). This is likely to produce a less severe artifact than if the frame were dropped after decoding.

A decoder MAY employ a more sophisticated drift compensation method. For example, the NetEQ component [Google-NetEQ] of the Google WebRTC codebase [Google-WebRTC] compensates for drift by adding or removing one period when the signal is highly periodic. The reference implementation of Opus allows a caller to learn whether the current frame’s signal is highly periodic, and if so what the period is, using the OPUS_GET_PITCH() request.

4.5. Configuration Switching

Switching between the Opus coding modes, audio bandwidths, and channel counts requires careful consideration to avoid audible glitches. Switching between any two configurations of the CELT-only mode, any two configurations of the Hybrid mode, or from WB SILK to Hybrid mode does not require any special treatment in the decoder, as the MDCT overlap will smooth the transition. Switching from Hybrid mode to WB SILK requires adding in the final contents of the CELT overlap buffer to the first SILK-only packet. This can be done by decoding a 2.5 ms silence frame with the CELT decoder using the channel count of the SILK-only packet (and any choice of audio bandwidth), which will correctly handle the cases when the channel count changes as well.

When changing the channel count for SILK-only or Hybrid packets, the encoder can avoid glitches by smoothly varying the stereo width of the input signal before or after the transition, and SHOULD do so. However, other transitions between SILK-only packets or between NB or MB SILK and Hybrid packets may cause glitches, because neither the LSF coefficients nor the LTP, LPC, stereo unmixing, and resampler buffers are available at the new sample rate. These switches SHOULD be delayed by the encoder until quiet periods or transients, where the inevitable glitches will be less audible. Additionally, the bit-stream MAY include redundant side information ("redundancy"), in the form of additional CELT frames embedded in each of the Opus frames around the transition.
The other transitions that cannot be easily handled are those where
the lower frequencies switch between the SILK LP-based model and the
CELT MDCT model. However, an encoder may not have an opportunity to
delay such a switch to a convenient point. For example, if the
content switches from speech to music, and the encoder does not have
enough latency in its analysis to detect this in advance, there may
be no convenient silence period during which to make the transition
for quite some time. To avoid or reduce glitches during these
problematic mode transitions, and also between audio bandwidth
changes in the SILK-only modes, transitions MAY include redundant
side information ("redundancy"), in the form of an additional CELT
frame embedded in the Opus frame.

A transition between coding the lower frequencies with the LP model
and the MDCT model or a transition that involves changing the SILK
bandwidth is only normatively specified when it includes redundancy.
For those without redundancy, it is RECOMMENDED that the decoder use
a concealment technique (e.g., make use of a PLC algorithm) to "fill in" the gap or discontinuity caused by the mode transition.
Therefore, PLC MUST NOT be applied during any normative transition,
i.e., when

- A packet includes redundancy for this transition (as described
  below),
- The transition is between any WB SILK packet and any Hybrid
  packet, or vice versa,
- The transition is between any two Hybrid mode packets, or
- The transition is between any two CELT mode packets,

unless there is actual packet loss.

4.5.1. Transition Side Information (Redundancy)

Transitions with side information include an extra 5 ms "redundant"
CELT frame within the Opus frame. This frame is designed to fill in
the gap or discontinuity in the different layers without requiring
the decoder to conceal it. For transitions from CELT-only to SILK-
only or Hybrid, the redundant frame is inserted in the first Opus
frame after the transition (i.e., the first SILK-only or Hybrid
frame). For transitions from SILK-only or Hybrid to CELT-only, the
redundant frame is inserted in the last Opus frame before the
transition (i.e., the last SILK-only or Hybrid frame).
4.5.1.1. Redundancy Flag

The presence of redundancy is signaled in all SILK-only and Hybrid frames, not just those involved in a mode transition. This allows the frames to be decoded correctly even if an adjacent frame is lost. For SILK-only frames, this signaling is implicit, based on the size of the of the Opus frame and the number of bits consumed decoding the SILK portion of it. After decoding the SILK portion of the Opus frame, the decoder uses ec_tell() (see Section 4.1.6.1) to check if there are at least 17 bits remaining. If so, then the frame contains redundancy.

For Hybrid frames, this signaling is explicit. After decoding the SILK portion of the Opus frame, the decoder uses ec_tell() (see Section 4.1.6.1) to ensure there are at least 37 bits remaining. If so, it reads a symbol with the PDF in Table 64, and if the value is 1, then the frame contains redundancy. Otherwise (if there were fewer than 37 bits left or the value was 0), the frame does not contain redundancy.


+----------------+
| PDF            |
+----------------+
| {4095, 1}/4096 |
+----------------+

Table 64: Redundancy Flag PDF

4.5.1.2. Redundancy Position Flag

Since the current frame is a SILK-only or a Hybrid frame, it must be at least 10 ms. Therefore, it needs an additional flag to indicate whether the redundant 5 ms CELT frame should be mixed into the beginning of the current frame, or the end. After determining that a frame contains redundancy, the decoder reads a 1 bit symbol with a uniform PDF (Table 65).


+----------+
| PDF      |
+----------+
| {1, 1}/2 |
+----------+

Table 65: Redundancy Position PDF

If the value is zero, this is the first frame in the transition, and the redundancy belongs at the end. If the value is one, this is the second frame in the transition, and the redundancy belongs at the...
beginning. There is no way to specify that an Opus frame contains separate redundant CELT frames at both the beginning and the end.

4.5.1.3. Redundancy Size

Unlike the CELT portion of a Hybrid frame, the redundant CELT frame does not use the same entropy coder state as the rest of the Opus frame, because this would break the CELT bit allocation mechanism in Hybrid frames. Thus, a redundant CELT frame always starts and ends on a byte boundary, even in SILK-only frames, where this is not strictly necessary.

For SILK-only frames, the number of bytes in the redundant CELT frame is simply the number of whole bytes remaining, which must be at least 2, due to the space check in Section 4.5.1.1. For Hybrid frames, the number of bytes is equal to 2, plus a decoded unsigned integer less than 256 (see Section 4.1.5). This may be more than the number of whole bytes remaining in the Opus frame, in which case the frame is invalid. However, a decoder is not required to ignore the entire frame, as this may be the result of a bit error that desynchronized the range coder. There may still be useful data before the error, and a decoder MAY keep any audio decoded so far instead of invoking the PLC, but it is RECOMMENDED that the decoder stop decoding and discard the rest of the current Opus frame.

It would have been possible to avoid these invalid states in the design of Opus by limiting the range of the explicit length decoded from Hybrid frames by the actual number of whole bytes remaining. However, this would require an encoder to determine the rate allocation for the MDCT layer up front, before it began encoding that layer. By allowing some invalid sizes, the encoder is able to defer that decision until much later. When encoding Hybrid frames which do not include redundancy, the encoder must still decide up-front if it wishes to use the minimum 37 bits required to trigger encoding of the redundancy flag, but this is a much looser restriction.

After determining the size of the redundant CELT frame, the decoder reduces the size of the buffer currently in use by the range coder by that amount. The MDCT layer reads any raw bits from the end of this reduced buffer, and all calculations of the number of bits remaining in the buffer must be done using this new, reduced size, rather than the original size of the Opus frame.

4.5.1.4. Decoding the Redundancy

The redundant frame is decoded like any other CELT-only frame, with the exception that it does not contain a TOC byte. The frame size is fixed at 5 ms, the channel count is set to that of the current frame,
and the audio bandwidth is also set to that of the current frame, with the exception that for MB SILK frames, it is set to WB.

If the redundancy belongs at the beginning (in a CELT-only to SILK-only or Hybrid transition), the final reconstructed output uses the first 2.5 ms of audio output by the decoder for the redundant frame as-is, discarding the corresponding output from the SILK-only or Hybrid portion of the frame. The remaining 2.5 ms is cross-lapped with the decoded SILK/Hybrid signal using the CELT’s power-complementary MDCT window to ensure a smooth transition.

If the redundancy belongs at the end (in a SILK-only or Hybrid to CELT-only transition), only the second half (2.5 ms) of the audio output by the decoder for the redundant frame is used. In that case, the second half of the redundant frame is cross-lapped with the end of the SILK/Hybrid signal, again using CELT’s power-complementary MDCT window to ensure a smooth transition.

4.5.2. State Reset

When a transition occurs, the state of the SILK or the CELT decoder (or both) may need to be reset before decoding a frame in the new mode. This avoids reusing "out of date" memory, which may not have been updated in some time or may not be in a well-defined state due to, e.g., PLC. The SILK state is reset before every SILK-only or Hybrid frame where the previous frame was CELT-only. The CELT state is reset every time the operating mode changes and the new mode is either Hybrid or CELT-only, except when the transition uses redundancy as described above. When switching from SILK-only or Hybrid to CELT-only with redundancy, the CELT state is reset before decoding the redundant CELT frame embedded in the SILK-only or Hybrid frame, but it is not reset before decoding the following CELT-only frame. When switching from CELT-only mode to SILK-only or Hybrid mode with redundancy, the CELT decoder is not reset for decoding the redundant CELT frame.

4.5.3. Summary of Transitions

Figure 18 illustrates all of the normative transitions involving a mode change, an audio bandwidth change, or both. Each one uses an S, H, or C to represent an Opus frame in the corresponding mode. In addition, an R indicates the presence of redundancy in the Opus frame it is cross-lapped with. Its location in the first or last 5 ms is assumed to correspond to whether it is the frame before or after the transition. Other uses of redundancy are non-normative. Finally, a c indicates the contents of the CELT overlap buffer after the previously decoded frame (i.e., as extracted by decoding a silence frame).
SILK to SILK with Redundancy:
S → S
&
!R → R
&
;S → S → S

NB or MB SILK to Hybrid with Redundancy:
S → S → S
&
!R →;H → H → H

WB SILK to Hybrid:
S → S → S → !H → H → H

SILK to CELT with Redundancy:
S → S → S
&
!R → C → C → C

Hybrid to NB or MB SILK with Redundancy:
H → H → H
&
!R → R
&
;S → S → S

Hybrid to WB SILK:
H → H → H → c
\+
> S → S → S

Hybrid to CELT with Redundancy:
H → H → H
&
!R → C → C → C

CELT to SILK with Redundancy:
C → C → C → R
&
;S → S → S

CELT to Hybrid with Redundancy:
C → C → C → R
&
|H → H → H

Key:
S  SILK-only frame ;  SILK decoder reset
H  Hybrid frame  |  CELT and SILK decoder resets
C  CELT-only frame !  CELT decoder reset
c  CELT overlap  +  Direct mixing
R  Redundant CELT frame &  Windowed cross-lap

Figure 18: Normative Transitions
illustrative, i.e., there is no requirement that a stream remain in
the same configuration for three consecutive frames before or after a
switch.

The behavior of transitions without redundancy where PLC is allowed
is non-normative. An encoder might still wish to use these
transitions if, for example, it doesn’t want to add the extra bitrate
required for redundancy or if it makes a decision to switch after it
has already transmitted the frame that would have had to contain the
redundancy. Figure 19 illustrates the recommended cross-lapping and
decoder resets for these transitions.

<table>
<thead>
<tr>
<th>Transition Type</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SILK to SILK (audio bandwidth change)</td>
<td>S → S → S ; S → S → S</td>
</tr>
<tr>
<td>NB or MB SILK to Hybrid</td>
<td>S → S → S</td>
</tr>
<tr>
<td>SILK to CELT without Redundancy</td>
<td>S → S → S → S → P</td>
</tr>
<tr>
<td></td>
<td>&amp; !C → C → C</td>
</tr>
<tr>
<td>Hybrid to NB or MB SILK</td>
<td>H → H → H → H → c</td>
</tr>
<tr>
<td></td>
<td>+ ; S → S → S</td>
</tr>
<tr>
<td>Hybrid to CELT without Redundancy</td>
<td>H → H → H → H → P</td>
</tr>
<tr>
<td></td>
<td>&amp; !C → C → C</td>
</tr>
<tr>
<td>CELT to SILK without Redundancy</td>
<td>C → C → C → C → P</td>
</tr>
<tr>
<td></td>
<td>&amp; ; S → S → S</td>
</tr>
<tr>
<td>CELT to Hybrid without Redundancy</td>
<td>C → C → C → C → P</td>
</tr>
<tr>
<td></td>
<td>&amp;</td>
</tr>
</tbody>
</table>

Key:
- S: SILK-only frame
- H: Hybrid frame
- C: CELT-only frame
- c: CELT overlap
- P: Packet Loss Concealment
- ;: SILK decoder reset
- |: CELT and SILK decoder resets
- !: CELT decoder reset
- +: Direct mixing
- &: Windowed cross-lap

Figure 19: Recommended Non-Normative Transitions

Encoders SHOULD NOT use other transitions, e.g., those that involve
redundancy in ways not illustrated in Figure 18.
5. Opus Encoder

Just like the decoder, the Opus encoder also normally consists of two main blocks: the SILK encoder and the CELT encoder. However, unlike the case of the decoder, a valid (though potentially suboptimal) Opus encoder is not required to support all modes and may thus only include a SILK encoder module or a CELT encoder module. The output bit-stream of the Opus encoding contains bits from the SILK and CELT encoders, though these are not separable due to the use of a range coder. A block diagram of the encoder is illustrated below.

![Block Diagram of Opus Encoder](image)

Figure 20: Opus Encoder

For a normal encoder where both the SILK and the CELT modules are included, an optimal encoder should select which coding mode to use at run-time depending on the conditions. In the reference implementation, the frame size is selected by the application, but the other configuration parameters (number of channels, bandwidth, mode) are automatically selected (unless explicitly overridden by the application) depend on the following:

- Requested bitrate
- Input sampling rate
- Type of signal (speech vs music)
- Frame size in use

The type of signal currently needs to be provided by the application (though it can be changed in real-time). An Opus encoder implementation could also do automatic detection, but since Opus is an interactive codec, such an implementation would likely have to either delay the signal (for non-interactive applications) or delay
the mode switching decisions (for interactive applications).

When the encoder is configured for voice over IP applications, the input signal is filtered by a high-pass filter to remove the lowest part of the spectrum that contains little speech energy and may contain background noise. This is a second order Auto Regressive Moving Average (i.e., with poles and zeros) filter with a cut-off frequency around 50 Hz. In the future, a music detector may also be used to lower the cut-off frequency when the input signal is detected to be music rather than speech.

5.1. Range Encoder

The range coder acts as the bit-packer for Opus. It is used in three different ways: to encode

- Entropy-coded symbols with a fixed probability model using ec_encode() (entenc.c),
- Integers from 0 to \((2^M - 1)\) using ec_enc_uint() or ec_enc_bits() (entenc.c),
- Integers from 0 to \((f_t - 1)\) (where \(f_t\) is not a power of two) using ec_enc_uint() (entenc.c).

The range encoder maintains an internal state vector composed of the four-tuple \((\text{val}, \text{rng}, \text{rem}, \text{ext})\) representing the low end of the current range, the size of the current range, a single buffered output byte, and a count of additional carry-propagating output bytes. Both \(\text{val}\) and \(\text{rng}\) are 32-bit unsigned integer values, \(\text{rem}\) is a byte value or less than 255 or the special value -1, and \(\text{ext}\) is an unsigned integer with at least 11 bits. This state vector is initialized at the start of each frame to the value \((0, 2^{31}, -1, 0)\). After encoding a sequence of symbols, the value of \(\text{rng}\) in the encoder should exactly match the value of \(\text{rng}\) in the decoder after decoding the same sequence of symbols. This is a powerful tool for detecting errors in either an encoder or decoder implementation. The value of \(\text{val}\), on the other hand, represents different things in the encoder and decoder, and is not expected to match.

The decoder has no analog for \(\text{rem}\) and \(\text{ext}\). These are used to perform carry propagation in the renormalization loop below. Each iteration of this loop produces 9 bits of output, consisting of 8 data bits and a carry flag. The encoder cannot determine the final value of the output bytes until it propagates these carry flags. Therefore the reference implementation buffers a single non-propagating output byte (i.e., one less than 255) in \(\text{rem}\) and keeps a count of additional

propagating (i.e., 255) output bytes in ext. An implementation may choose to use any mathematically equivalent scheme to perform carry propagation.

5.1.1. Encoding Symbols

The main encoding function is ec_encode() (entenc.c), which encodes symbol k in the current context using the same three-tuple (fl[k], fh[k], ft) as the decoder to describe the range of the symbol (see Section 4.1).

ec_encode() updates the state of the encoder as follows. If fl[k] is greater than zero, then

\[
val = val + \frac{rng}{ft} - \left(\frac{ft}{ft} - fl\right) ,
\]

\[
rng = \frac{rng}{ft} - \frac{fh - fl}{ft} .
\]

The divisions here are integer division.

5.1.1.1. Renormalization

After this update, the range is normalized using a procedure very similar to that of Section 4.1.2.1, implemented by ec_enc_normalize() (entenc.c). The following process is repeated until rng > 2**23. First, the top 9 bits of val, (val>>23), are sent to the carry buffer, described in Section 5.1.1.2. Then, the encoder sets

\[
val = (val<<8) & 0x7FFFFFFF ,
\]

\[
rng = rng<<8 .
\]

5.1.1.2. Carry Propagation and Output Buffering

The function ec_enc_carry_out() (entenc.c) implements carry propagation and output buffering. It takes as input a 9-bit unsigned value, c, consisting of 8 data bits and an additional carry bit. If c is equal to the value 255, then ext is simply incremented, and no
other state updates are performed. Otherwise, let \( b = (c >> 8) \) be the carry bit. Then,

- If the buffered byte \( rem \) contains a value other than \(-1\), the encoder outputs the byte \((rem + b)\). Otherwise, if \( rem \) is \(-1\), no byte is output.

- If \( ext \) is non-zero, then the encoder outputs \( ext \) bytes—all with a value of 0 if \( b \) is set, or 255 if \( b \) is unset—and sets \( ext \) to 0.

- \( rem \) is set to the 8 data bits:

\[
rem = c & 255.
\]

### 5.1.2. Alternate Encoding Methods

The reference implementation uses three additional encoding methods that are exactly equivalent to the above, but make assumptions and simplifications that allow for a more efficient implementation.

#### 5.1.2.1. \texttt{ec_encode_bin()}

The first is \texttt{ec_encode_bin()} (entenc.c), defined using the parameter \( ftb \) instead of \( ft \). It is mathematically equivalent to calling \texttt{ec_encode()} with \( ft = (1 << ftb) \), but avoids using division.

#### 5.1.2.2. \texttt{ec_enc_bit_logp()}

The next is \texttt{ec_enc_bit_logp()} (entenc.c), which encodes a single binary symbol. The context is described by a single parameter, \( logp \), which is the absolute value of the base-2 logarithm of the probability of a "1". It is mathematically equivalent to calling \texttt{ec_encode()} with the 3-tuple \((fl[k] = 0, fh[k] = (1 << logp) - 1, ft = (1 << logp))\) if \( k \) is 0 and with \((fl[k] = (1 << logp) - 1, fh[k] = ft = (1 << logp))\) if \( k \) is 1. The implementation requires no multiplications or divisions.

#### 5.1.2.3. \texttt{ec_enc_icdf()}

The last is \texttt{ec_enc_icdf()} (entenc.c), which encodes a single binary symbol with a table-based context of up to 8 bits. This uses the same icdf table as \texttt{ec_dec_icdf()} from Section 4.1.3.3. The function is mathematically equivalent to calling \texttt{ec_encode()} with

\[
fl[k] = (1 << ftb) - icdf[k-1] \quad \text{(or 0 if } k == 0), \quad fh[k] = (1 << ftb) - icdf[k], \quad \text{and } ft = (1 << ftb).
\]

This only saves a few arithmetic operations over \texttt{ec_encode_bin()}, but allows the encoder to use the same icdf tables as the decoder.
5.1.3. Encoding Raw Bits

The raw bits used by the CELT layer are packed at the end of the buffer using ec_enc_bits() (entenc.c). Because the raw bits may continue into the last byte output by the range coder if there is room in the low-order bits, the encoder must be prepared to merge these values into a single byte. The procedure in Section 5.1.5 does this in a way that ensures both the range coded data and the raw bits can be decoded successfully.

5.1.4. Encoding Uniformly Distributed Integers

The function ec_enc_uint() (entenc.c) encodes one of \( f_t \) equiprobable symbols in the range 0 to \((f_t - 1)\), inclusive, each with a frequency of 1, where \( f_t \) may be as large as \((2^{32} - 1)\). Like the decoder (see Section 4.1.5), it splits up the value into a range coded symbol representing up to 8 of the high bits, and, if necessary, raw bits representing the remainder of the value.

ec_enc_uint() takes a two-tuple \((t, f_t)\), where \( t \) is the unsigned integer to be encoded, \( 0 \leq t < f_t \), and \( f_t \) is not necessarily a power of two. Let \( \text{ftb} = \log_2(f_t - 1) \), i.e., the number of bits required to store \((f_t - 1)\) in two’s complement notation. If \( \text{ftb} \) is 8 or less, then \( t \) is encoded directly using ec_encode() with the three-tuple \((t, t + 1, f_t)\).

If \( \text{ftb} \) is greater than 8, then the top 8 bits of \( t \) are encoded using the three-tuple \((t >> (\text{ftb} - 8), (t >> (\text{ftb} - 8)) + 1, ((f_t - 1) >> (\text{ftb} - 8)) + 1)\), and the remaining bits, \((t & ((1 << (\text{ftb} - 8)) - 1))\), are encoded as raw bits with ec_enc_bits().

5.1.5. Finalizing the Stream

After all symbols are encoded, the stream must be finalized by outputting a value inside the current range. Let \( \text{end} \) be the unsigned integer in the interval \([\text{val}, \text{val} + \text{rng})\) with the largest number of trailing zero bits, \( b \), such that \((\text{end} + (1 << b) - 1)\) is also in the interval \([\text{val}, \text{val} + \text{rng})\). This choice of \( \text{end} \) allows the maximum number of trailing bits to be set to arbitrary values while still ensuring the range coded part of the buffer can be decoded correctly. Then, while \( \text{end} \) is not zero, the top 9 bits of \( \text{end} \), i.e., \((\text{end} >> 23)\), are passed to the carry buffer in accordance with the procedure in Section 5.1.1.2, and \( \text{end} \) is updated via

\[
\text{end} = (\text{end} << 8) \ & \ 0x7FFFFFFF.
\]

Finally, if the buffered output byte, \( \text{rem} \), is neither zero nor the
special value -1, or the carry count, ext, is greater than zero, then 9 zero bits are sent to the carry buffer to flush it to the output buffer. When outputting the final byte from the range coder, if it would overlap any raw bits already packed into the end of the output buffer, they should be ORed into the same byte. The bit allocation routines in the CELT layer should ensure that this can be done without corrupting the range coder data so long as end is chosen as described above. If there is any space between the end of the range coder data and the end of the raw bits, it is padded with zero bits. This entire process is implemented by ec_enc_done() (entenc.c).

5.1.6. Current Bit Usage

The bit allocation routines in Opus need to be able to determine a conservative upper bound on the number of bits that have been used to encode the current frame thus far. This drives allocation decisions and ensures that the range coder and raw bits will not overflow the output buffer. This is computed in the reference implementation to whole-bit precision by the function ec_tell() (entcode.h) and to fractional 1/8th bit precision by the function ec_tell_frac() (entcode.c). Like all operations in the range coder, it must be implemented in a bit-exact manner, and must produce exactly the same value returned by the same functions in the decoder after decoding the same symbols.

5.2. SILK Encoder

In many respects the SILK encoder mirrors the SILK decoder described in Section 4.2. Details such as the quantization and range coder tables can be found there, while this section describes the high-level design choices that were made. The diagram below shows the basic modules of the SILK encoder.

```
+----------+        +--------+        +---------+
| Sample   |        | Stereo  |        |  SILK   |
| Rate     |        | Mixing  |        | Encoder |
| Conversion |      |        |        | Bitstream |

Figure 21: SILK Encoder
```

5.2.1. Sample Rate Conversion

The input signal’s sampling rate is adjusted by a sample rate conversion module so that it matches the SILK internal sampling rate. The input to the sample rate converter is delayed by a number of
samples depending on the sample rate ratio, such that the overall delay is constant for all input and output sample rates.

5.2.2. Stereo Mixing

The stereo mixer is only used for stereo input signals. It converts a stereo left/right signal into an adaptive mid/side representation. The first step is to compute non-adaptive mid/side signals as half the sum and difference between left and right signals. The side signal is then minimized in energy by subtracting a prediction of it based on the mid signal. This prediction works well when the left and right signals exhibit linear dependency, for instance for an amplitude-panned input signal. Like in the decoder, the prediction coefficients are linearly interpolated during the first 8 ms of the frame. The mid signal is always encoded, whereas the residual side signal is only encoded if it has sufficient energy compared to the mid signal’s energy. If it has not, the "mid_only_flag" is set without encoding the side signal.

The predictor coefficients are coded regardless of whether the side signal is encoded. For each frame, two predictor coefficients are computed, one that predicts between low-passed mid and side channels, and one that predicts between high-passed mid and side channels. The low-pass filter is a simple three-tap filter and creates a delay of one sample. The high-pass filtered signal is the difference between the mid signal delayed by one sample and the low-passed signal. Instead of explicitly computing the high-passed signal, it is computationally more efficient to transform the prediction coefficients before applying them to the filtered mid signal, as follows

\[
\text{pred}(n) = \text{LP}(n) \times w_0 + \text{HP}(n) \times w_1 \\
= \text{LP}(n) \times w_0 + (\text{mid}(n-1) - \text{LP}(n)) \times w_1 \\
= \text{LP}(n) \times (w_0 - w_1) + \text{mid}(n-1) \times w_1
\]

where \(w_0\) and \(w_1\) are the low-pass and high-pass prediction coefficients, \(\text{mid}(n-1)\) is the mid signal delayed by one sample, \(\text{LP}(n)\) and \(\text{HP}(n)\) are the low-passed and high-passed signals and \(\text{pred}(n)\) is the prediction signal that is subtracted from the side signal.

5.2.3. SILK Core Encoder

What follows is a description of the core encoder and its components. For simplicity, the core encoder is referred to simply as the encoder in the remainder of this section. An overview of the encoder is given in Figure 22.
1: Input speech signal
2: Range encoded bitstream
3: Voice activity estimate
4: Pitch lags (per 5 ms) and voicing decision (per 20 ms)
5: Noise shaping quantization coefficients
   - Short term synthesis and analysis
     noise shaping coefficients (per 5 ms)
   - Long term synthesis and analysis noise
     shaping coefficients (per 5 ms and for voiced speech only)
   - Noise shaping tilt (per 5 ms)
   - Quantizer gain/step size (per 5 ms)
6: Input signal filtered with analysis noise shaping filters
7: Short and long term prediction coefficients
   LTP (per 5 ms) and LPC (per 20 ms)
8: LSF quantization indices
9: LSF coefficients
10: Quantized LSF coefficients
11: Processed gains, and synthesis noise shape coefficients
12: LTP state scaling coefficient. Controlling error propagation
    / prediction gain trade-off
13: Quantized signal

Figure 22: SILK Core Encoder

5.2.3.1. Voice Activity Detection

The input signal is processed by a Voice Activity Detector (VAD) to produce a measure of voice activity, spectral tilt, and signal-to-noise estimates for each frame. The VAD uses a sequence of half-band filterbanks to split the signal into four subbands: 0...Fs/16, Fs/16...Fs/8, Fs/8...Fs/4, and Fs/4...Fs/2, where Fs is the sampling frequency (8, 12, 16, or 24 kHz). The lowest subband, from 0 - Fs/16, is high-pass filtered with a first-order moving average (MA) filter (with transfer function H(z) = 1-z**(-1)) to reduce the energy at the lowest frequencies. For each frame, the signal energy per subband is computed. In each subband, a noise level estimator tracks the background noise level and a Signal-to-Noise Ratio (SNR) value is computed as the logarithm of the ratio of energy to noise level. Using these intermediate variables, the following parameters are calculated for use in other SILK modules:

- Average SNR. The average of the subband SNR values.
- Smoothed subband SNRs. Temporally smoothed subband SNR values.
- Speech activity level. Based on the average SNR and a weighted average of the subband energies.
- Spectral tilt. A weighted average of the subband SNRs, with positive weights for the low subbands and negative weights for the high subbands.

5.2.3.2. Pitch Analysis

The input signal is processed by the open loop pitch estimator shown in Figure 23.
The pitch analysis finds a binary voiced/unvoiced classification, and, for frames classified as voiced, four pitch lags per frame - one for each 5 ms subframe - and a pitch correlation indicating the periodicity of the signal. The input is first whitened using a Linear Prediction (LP) whitening filter, where the coefficients are computed through standard Linear Prediction Coding (LPC) analysis. The order of the whitening filter is 16 for best results, but is reduced to 12 for medium complexity and 8 for low complexity modes. The whitened signal is analyzed to find pitch lags for which the time correlation is high. The analysis consists of three stages for reducing the complexity:

- In the first stage, the whitened signal is downsampled to 4 kHz (from 8 kHz) and the current frame is correlated to a signal delayed by a range of lags, starting from a shortest lag corresponding to 500 Hz, to a longest lag corresponding to 56 Hz.
o The second stage operates on an 8 kHz signal (downsampled from 12, 16, or 24 kHz) and measures time correlations only near the lags corresponding to those that had sufficiently high correlations in the first stage. The resulting correlations are adjusted for a small bias towards short lags to avoid ending up with a multiple of the true pitch lag. The highest adjusted correlation is compared to a threshold depending on:

* Whether the previous frame was classified as voiced

* The speech activity level

* The spectral tilt.

If the threshold is exceeded, the current frame is classified as voiced and the lag with the highest adjusted correlation is stored for a final pitch analysis of the highest precision in the third stage.

o The last stage operates directly on the whitened input signal to compute time correlations for each of the four subframes independently in a narrow range around the lag with highest correlation from the second stage.

5.2.3.3. Noise Shaping Analysis

The noise shaping analysis finds gains and filter coefficients used in the prefilter and noise shaping quantizer. These parameters are chosen such that they will fulfill several requirements:

o Balancing quantization noise and bitrate. The quantization gains determine the step size between reconstruction levels of the excitation signal. Therefore, increasing the quantization gain amplifies quantization noise, but also reduces the bitrate by lowering the entropy of the quantization indices.

o Spectral shaping of the quantization noise; the noise shaping quantizer is capable of reducing quantization noise in some parts of the spectrum at the cost of increased noise in other parts without substantially changing the bitrate. By shaping the noise such that it follows the signal spectrum, it becomes less audible. In practice, best results are obtained by making the shape of the noise spectrum slightly flatter than the signal spectrum.

o De-emphasizing spectral valleys; by using different coefficients in the analysis and synthesis part of the prefilter and noise shaping quantizer, the levels of the spectral valleys can be decreased relative to the levels of the spectral peaks such as
speech formants and harmonics. This reduces the entropy of the signal, which is the difference between the coded signal and the quantization noise, thus lowering the bitrate.

- Matching the levels of the decoded speech formants to the levels of the original speech formants; an adjustment gain and a first order tilt coefficient are computed to compensate for the effect of the noise shaping quantization on the level and spectral tilt.

Figure 24 shows an example of an input signal spectrum (1). After de-emphasis and level matching, the spectrum has deeper valleys (2). The quantization noise spectrum (3) more or less follows the input signal spectrum, while having slightly less pronounced peaks. The entropy, which provides a lower bound on the bitrate for encoding the excitation signal, is proportional to the area between the de-emphasized spectrum (2) and the quantization noise spectrum (3). Without de-emphasis, the entropy is proportional to the area between input spectrum (1) and quantization noise (3) - clearly higher.

The transformation from input signal to de-emphasized signal can be described as a filtering operation with a filter
\[ H(z) = G \times \left( 1 - c_{\text{tilt}} \times z \right) \times \frac{W_{\text{ana}}(z)}{W_{\text{syn}}(z)}, \]

having an adjustment gain \( G \), a first order tilt adjustment filter with tilt coefficient \( c_{\text{tilt}} \), and where

\[
W_{\text{ana}}(z) = \frac{1 - \sum_{k=1}^{d} \left( a_{\text{ana}}(k) \times z^{-k} \right) \times \left( 1 - z^{-L} \times b_{\text{ana}}(k) \times z^{-k} \right)}{\sum_{k=-d}^{d} b_{\text{ana}}(k) \times z^{-k}},
\]

is the analysis part of the de-emphasis filter, consisting of the short-term shaping filter with coefficients \( a_{\text{ana}}(k) \), and the long-term shaping filter with coefficients \( b_{\text{ana}}(k) \) and pitch lag \( L \). The parameter \( d \) determines the number of long-term shaping filter taps.

Similarly, but without the tilt adjustment, the synthesis part can be written as

\[
W_{\text{syn}}(z) = \frac{1 - \sum_{k=1}^{d} \left( a_{\text{syn}}(k) \times z^{-k} \right) \times \left( 1 - z^{-L} \times b_{\text{syn}}(k) \times z^{-k} \right)}{\sum_{k=-d}^{d} b_{\text{syn}}(k) \times z^{-k}},
\]

All noise shaping parameters are computed and applied per subframe of 5 ms. First, an LPC analysis is performed on a windowed signal block of 15 ms. The signal block has a look-ahead of 5 ms relative to the current subframe, and the window is an asymmetric sine window. The LPC analysis is done with the autocorrelation method, with an order of between 8, in lowest-complexity mode, and 16, for best quality. Optionally the LPC analysis and noise shaping filters are warped by replacing the delay elements by first-order allpass filters. This increases the frequency resolution at low frequencies and reduces it at high ones, which better matches the human auditory system and improves quality. The warped analysis and filtering comes at a cost in complexity and is therefore only done in higher complexity modes.

The quantization gain is found by taking the square root of the residual energy from the LPC analysis and multiplying it by a value
inversely proportional to the coding quality control parameter and the pitch correlation.

Next the two sets of short-term noise shaping coefficients \( a_{\text{ana}}(k) \) and \( a_{\text{syn}}(k) \) are obtained by applying different amounts of bandwidth expansion to the coefficients found in the LPC analysis. This bandwidth expansion moves the roots of the LPC polynomial towards the origin, using the formulas

\[
a_{\text{ana}}(k) = a(k) \cdot g_{\text{ana}} , \quad \text{and} \quad a_{\text{syn}}(k) = a(k) \cdot g_{\text{syn}} ,
\]

where \( a(k) \) is the \( k \)'th LPC coefficient, and the bandwidth expansion factors \( g_{\text{ana}} \) and \( g_{\text{syn}} \) are calculated as

\[
g_{\text{ana}} = 0.95 - 0.01 \cdot C , \quad \text{and} \quad g_{\text{syn}} = 0.95 + 0.01 \cdot C ,
\]

where \( C \) is the coding quality control parameter between 0 and 1. Applying more bandwidth expansion to the analysis part than to the synthesis part gives the desired de-emphasis of spectral valleys in between formants.

The long-term shaping is applied only during voiced frames. It uses three filter taps, described by

\[
b_{\text{ana}} = F_{\text{ana}} \cdot [0.25, 0.5, 0.25] , \quad \text{and} \quad b_{\text{syn}} = F_{\text{syn}} \cdot [0.25, 0.5, 0.25] .
\]

For unvoiced frames these coefficients are set to 0. The multiplication factors \( F_{\text{ana}} \) and \( F_{\text{syn}} \) are chosen between 0 and 1, depending on the coding quality control parameter, as well as the calculated pitch correlation and smoothed subband SNR of the lowest subband. By having \( F_{\text{ana}} \) less than \( F_{\text{syn}} \), the pitch harmonics are emphasized relative to the valleys in between the harmonics.

The tilt coefficient \( c_{\text{tilt}} \) is for unvoiced frames chosen as
c_tilt = 0.25,

and as

c_tilt = 0.25 + 0.2625 * V

for voiced frames, where V is the voice activity level between 0 and 1.

The adjustment gain G serves to correct any level mismatch between the original and decoded signals that might arise from the noise shaping and de-emphasis. This gain is computed as the ratio of the prediction gain of the short-term analysis and synthesis filter coefficients. The prediction gain of an LPC synthesis filter is the square root of the output energy when the filter is excited by a unit-energy impulse on the input. An efficient way to compute the prediction gain is by first computing the reflection coefficients from the LPC coefficients through the step-down algorithm, and extracting the prediction gain from the reflection coefficients as

\[
predGain = \left( \prod_{k=1}^{K} \left( 1 - r_k \right)^{2 -0.5} \right),
\]

where \( r_k \) is the k’th reflection coefficient.

Initial values for the quantization gains are computed as the square-root of the residual energy of the LPC analysis, adjusted by the coding quality control parameter. These quantization gains are later adjusted based on the results of the prediction analysis.

5.2.3.4. Prediction Analysis

The prediction analysis is performed in one of two ways depending on how the pitch estimator classified the frame. The processing for voiced and unvoiced speech is described in Section 5.2.3.4.1 and Section 5.2.3.4.2, respectively. Inputs to this function include the pre-whitened signal from the pitch estimator (see Section 5.2.3.2).
5.2.3.4.1. Voiced Speech

For a frame of voiced speech the pitch pulses will remain dominant in the pre-whitened input signal. Further whitening is desirable as it leads to higher quality at the same available bitrate. To achieve this, a Long-Term Prediction (LTP) analysis is carried out to estimate the coefficients of a fifth-order LTP filter for each of four subframes. The LTP coefficients are quantized using the method described in Section 5.2.3.6, and the quantized LTP coefficients are used to compute the LTP residual signal. This LTP residual signal is the input to an LPC analysis where the LPC coefficients are estimated using Burg’s method [Burg], such that the residual energy is minimized. The estimated LPC coefficients are converted to a Line Spectral Frequency (LSF) vector and quantized as described in Section 5.2.3.5. After quantization, the quantized LSF vector is converted back to LPC coefficients using the full procedure in Section 4.2.7.5. By using quantized LTP coefficients and LPC coefficients derived from the quantized LSF coefficients, the encoder remains fully synchronized with the decoder. The quantized LPC and LTP coefficients are also used to filter the input signal and measure residual energy for each of the four subframes.

5.2.3.4.2. Unvoiced Speech

For a speech signal that has been classified as unvoiced, there is no need for LTP filtering, as it has already been determined that the pre-whitened input signal is not periodic enough within the allowed pitch period range for LTP analysis to be worth the cost in terms of complexity and bitrate. The pre-whitened input signal is therefore discarded, and instead the input signal is used for LPC analysis using Burg’s method. The resulting LPC coefficients are converted to an LSF vector and quantized as described in the following section. They are then transformed back to obtain quantized LPC coefficients, which are then used to filter the input signal and measure residual energy for each of the four subframes.

5.2.3.4.2.1. Burg’s Method

The main purpose of linear prediction in SILK is to reduce the bitrate by minimizing the residual energy. At least at high bitrates, perceptual aspects are handled independently by the noise shaping filter. Burg’s method is used because it provides higher prediction gain than the autocorrelation method and, unlike the covariance method, produces stable filters (assuming numerical errors don’t spoil that). SILK’s implementation of Burg’s method is also computationally faster than the autocovariance method. The implementation of Burg’s method differs from traditional implementations in two aspects. The first difference is that it
operates on autocorrelations, similar to the Schur algorithm [Schur],
but with a simple update to the autocorrelations after finding each
reflection coefficient to make the result identical to Burg’s method.
This brings down the complexity of Burg’s method to near that of the
autocorrelation method. The second difference is that the signal in
each subframe is scaled by the inverse of the residual quantization
step size. Subframes with a small quantization step size will on
average spend more bits for a given amount of residual energy than
subframes with a large step size. Without scaling, Burg’s method
minimizes the total residual energy in all subframes, which doesn’t
necessarily minimize the total number of bits needed for coding the
quantized residual. The residual energy of the scaled subframes is a
better measure for that number of bits.

5.2.3.5. LSF Quantization

Unlike many other speech codecs, SILK uses variable bitrate coding
for the LSFs. This improves the average rate-distortion (R-D)
tradeoff and reduces outliers. The variable bitrate coding minimizes
a linear combination of the weighted quantization errors and the
bitrate. The weights for the quantization errors are the Inverse
Harmonic Mean Weighting (IHMW) function proposed by Laroia et al.
(see [laroia-icassp]). These weights are referred to here as Laroia
weights.

The LSF quantizer consists of two stages. The first stage is an
(unweighted) vector quantizer (VQ), with a codebook size of 32
vectors. The quantization errors for the codebook vector are sorted,
and for the N best vectors a second stage quantizer is run. By
varying the number N a tradeoff is made between R-D performance and
computational efficiency. For each of the N codebook vectors the
Laroia weights corresponding to that vector (and not to the input
vector) are calculated. Then the residual between the input LSF
vector and the codebook vector is scaled by the square roots of these
Laroia weights. This scaling partially normalizes error sensitivity
for the residual vector, so that a uniform quantizer with fixed step
sizes can be used in the second stage without too much performance
loss. And by scaling with Laroia weights determined from the first-
stage codebook vector, the process can be reversed in the decoder.

The second stage uses predictive delayed decision scalar
quantization. The quantization error is weighted by Laroia weights
determined from the LSF input vector. The predictor multiplies the
previous quantized residual value by a prediction coefficient that
depends on the vector index from the first stage VQ and on the
location in the LSF vector. The prediction is subtracted from the
LSF residual value before quantizing the result, and added back
afterwards. This subtraction can be interpreted as shifting the
quantization levels of the scalar quantizer, and as a result the quantization error of each value depends on the quantization decision of the previous value. This dependency is exploited by the delayed decision mechanism to search for a quantization sequency with best R-D performance with a Viterbi-like algorithm [Viterbi]. The quantizer processes the residual LSF vector in reverse order (i.e., it starts with the highest residual LSF value). This is done because the prediction works slightly better in the reverse direction.

The quantization index of the first stage is entropy coded. The quantization sequence from the second stage is also entropy coded, where for each element the probability table is chosen depending on the vector index from the first stage and the location of that element in the LSF vector.

5.2.3.5.1. LSF Stabilization

If the input is stable, finding the best candidate usually results in a quantized vector that is also stable. Because of the two-stage approach, however, it is possible that the best quantization candidate is unstable. The encoder applies the same stabilization procedure applied by the decoder (see Section 4.2.7.5.4 to ensure the LSF parameters are within their valid range, increasingly sorted, and have minimum distances between each other and the border values.

5.2.3.6. LTP Quantization

For voiced frames, the prediction analysis described in Section 5.2.3.4.1 resulted in four sets (one set per subframe) of five LTP coefficients, plus four weighting matrices. The LTP coefficients for each subframe are quantized using entropy constrained vector quantization. A total of three vector codebooks are available for quantization, with different rate-distortion trade-offs. The three codebooks have 10, 20, and 40 vectors and average rates of about 3, 4, and 5 bits per vector, respectively. Consequently, the first codebook has larger average quantization distortion at a lower rate, whereas the last codebook has smaller average quantization distortion at a higher rate. Given the weighting matrix \( W_{ltp} \) and LTP vector \( b \), the weighted rate-distortion measure for a codebook vector \( cb_i \) with rate \( r_i \) is given by

\[
RD = u \cdot (b - cb_i)' \cdot W_{ltp} \cdot (b - cb_i) + r_i,
\]

where \( u \) is a fixed, heuristically-determined parameter balancing the distortion and rate. Which codebook gives the best performance for a given LTP vector depends on the weighting matrix for that LTP vector.
For example, for a low valued $W_{ltp}$, it is advantageous to use the codebook with 10 vectors as it has a lower average rate. For a large $W_{ltp}$, on the other hand, it is often better to use the codebook with 40 vectors, as it is more likely to contain the best codebook vector. The weighting matrix $W_{ltp}$ depends mostly on two aspects of the input signal. The first is the periodicity of the signal; the more periodic, the larger $W_{ltp}$. The second is the change in signal energy in the current subframe, relative to the signal one pitch lag earlier. A decaying energy leads to a larger $W_{ltp}$ than an increasing energy. Both aspects fluctuate relatively slowly, which causes the $W_{ltp}$ matrices for different subframes of one frame often to be similar. Because of this, one of the three codebooks typically gives good performance for all subframes, and therefore the codebook search for the subframe LTP vectors is constrained to only allow codebook vectors to be chosen from the same codebook, resulting in a rate reduction.

To find the best codebook, each of the three vector codebooks is used to quantize all subframe LTP vectors and produce a combined weighted rate-distortion measure for each vector codebook. The vector codebook with the lowest combined rate-distortion over all subframes is chosen. The quantized LTP vectors are used in the noise shaping quantizer, and the index of the codebook plus the four indices for the four subframe codebook vectors are passed on to the range encoder.

5.2.3.7. Prefilter

In the prefilter the input signal is filtered using the spectral valley de-emphasis filter coefficients from the noise shaping analysis (see Section 5.2.3.3). By applying only the noise shaping analysis filter to the input signal, it provides the input to the noise shaping quantizer.

5.2.3.8. Noise Shaping Quantizer

The noise shaping quantizer independently shapes the signal and coding noise spectra to obtain a perceptually higher quality at the same bitrate.

The prefilter output signal is multiplied with a compensation gain $G$ computed in the noise shaping analysis. Then the output of a synthesis shaping filter is added, and the output of a prediction filter is subtracted to create a residual signal. The residual signal is multiplied by the inverse quantized quantization gain from the noise shaping analysis, and input to a scalar quantizer. The quantization indices of the scalar quantizer represent a signal of pulses that is input to the pyramid range encoder. The scalar
quantizer also outputs a quantization signal, which is multiplied by the quantized quantization gain from the noise shaping analysis to create an excitation signal. The output of the prediction filter is added to the excitation signal to form the quantized output signal \( y(n) \). The quantized output signal \( y(n) \) is input to the synthesis shaping and prediction filters.

Optionally the noise shaping quantizer operates in a delayed decision mode. In this mode it uses a Viterbi algorithm to keep track of multiple rounding choices in the quantizer and select the best one after a delay of 32 samples. This improves the rate/distortion performance of the quantizer.

5.2.3.9. Constant Bitrate Mode

SILK was designed to run in Variable Bitrate (VBR) mode. However the reference implementation also has a Constant Bitrate (CBR) mode for SILK. In CBR mode SILK will attempt to encode each packet with no more than the allowed number of bits. The Opus wrapper code then pads the bitstream if any unused bits are left in SILK mode, or encodes the high band with the remaining number of bits in Hybrid mode. The number of payload bits is adjusted by changing the quantization gains and the rate/distortion tradeoff in the noise shaping quantizer, in an iterative loop around the noise shaping quantizer and entropy coding. Compared to the SILK VBR mode, the CBR mode has lower audio quality at a given average bitrate, and also has higher computational complexity.

5.3. CELT Encoder

Most of the aspects of the CELT encoder can be directly derived from the description of the decoder. For example, the filters and rotations in the encoder are simply the inverse of the operation performed by the decoder. Similarly, the quantizers generally optimize for the mean square error (because noise shaping is part of the bit-stream itself), so no special search is required. For this reason, only the less straightforward aspects of the encoder are described here.

5.3.1. Pitch Prefilter

The pitch prefiler is applied after the pre-emphasis. It is applied in such a way as to be the inverse of the decoder’s post-filter. The main non-obvious aspect of the prefiler is the selection of the pitch period. The pitch search should be optimized for the following criteria:
continuity: it is important that the pitch period does not change abruptly between frames; and

avoidance of pitch multiples: when the period used is a multiple of the real period (lower frequency fundamental), the post-filter loses most of its ability to reduce noise

5.3.2. Bands and Normalization

The MDCT output is divided into bands that are designed to match the ear’s critical bands for the smallest (2.5 ms) frame size. The larger frame sizes use integer multiples of the 2.5 ms layout. For each band, the encoder computes the energy that will later be encoded. Each band is then normalized by the square root of the *unquantized* energy, such that each band now forms a unit vector X. The energy and the normalization are computed by compute_band_energies() and normalise_bands() (bands.c), respectively.

5.3.3. Energy Envelope Quantization

Energy quantization (both coarse and fine) can be easily understood from the decoding process. For all useful bitrates, the coarse quantizer always chooses the quantized log energy value that minimizes the error for each band. Only at very low rate does the encoder allow larger errors to minimize the rate and avoid using more bits than are available. When the available CPU requirements allow it, it is best to try encoding the coarse energy both with and without inter-frame prediction such that the best prediction mode can be selected. The optimal mode depends on the coding rate, the available bitrate, and the current rate of packet loss.

The fine energy quantizer always chooses the quantized log energy value that minimizes the error for each band because the rate of the fine quantization depends only on the bit allocation and not on the values that are coded.

5.3.4. Bit Allocation

The encoder must use exactly the same bit allocation process as used by the decoder and described in Section 4.3.3. The three mechanisms that can be used by the encoder to adjust the bitrate on a frame-by-frame basis are band boost, allocation trim, and band skipping.

5.3.4.1. Band Boost

The reference encoder makes a decision to boost a band when the energy of that band is significantly higher than that of the
neighboring bands. Let $E_j$ be the log-energy of band $j$, we define

$$D_j = 2*E_j - E_{j-1} - E_{j+1}$$

The allocation of band $j$ is boosted once if $D_j > t_1$ and twice if $D_j > t_2$. For $LM=1$, $t_1=2$ and $t_2=4$, while for $LM<1$, $t_1=3$ and $t_2=5$.

5.3.4.2. Allocation Trim

The allocation trim is a value between 0 and 10 (inclusively) that controls the allocation balance between the low and high frequencies. The encoder starts with a safe "default" of 5 and deviates from that default in two different ways. First the trim can deviate by +/- 2 depending on the spectral tilt of the input signal. For signals with more low frequencies, the trim is increased by up to 2, while for signals with more high frequencies, the trim is decreased by up to 2. For stereo inputs, the trim value can be decreased by up to 4 when the inter-channel correlation at low frequency (first 8 bands) is high.

5.3.4.3. Band Skipping

The encoder uses band skipping to ensure that the shape of the bands is only coded if there is at least 1/2 bit per sample available for the PVQ. If not, then no bit is allocated and folding is used instead. To ensure continuity in the allocation, some amount of hysteresis is added to the process, such that a band that received PVQ bits in the previous frame only needs 7/16 bit/sample to be coded for the current frame, while a band that did not receive PVQ bits in the previous frames needs at least 9/16 bit/sample to be coded.

5.3.5. Stereo Decisions

Because CELT applies mid-side stereo coupling in the normalized domain, it does not suffer from important stereo image problems even when the two channels are completely uncorrelated. For this reason it is always safe to use stereo coupling on any audio frame. That being said, there are some frames for which dual (independent) stereo is still more efficient. This decision is made by comparing the estimated entropy with and without coupling over the first 13 bands, taking into account the fact that all bands with more than two MDCT bins require one extra degree of freedom when coded in mid-side. Let $L1_{ms}$ and $L1_{lr}$ be the L1-norm of the mid-side vector and the L1-norm of the left-right vector, respectively. The decision to use mid-side is made if and only if

$$\frac{L1_{ms}}{L1_{lr}} < \frac{Valin, et al. Expires December 30, 2012 [Page 152]}$$
where bins is the number of MDCT bins in the first 13 bands and E is the number of extra degrees of freedom for mid-side coding. For LM>1, E=13, otherwise E=5.

The reference encoder decides on the intensity stereo threshold based on the bitrate alone. After taking into account the frame size by subtracting 80 bits per frame for coarse energy, the first band using intensity coding is as follows:

<table>
<thead>
<tr>
<th>bitrate (kb/s)</th>
<th>start band</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>8</td>
</tr>
<tr>
<td>35-50</td>
<td>12</td>
</tr>
<tr>
<td>50-68</td>
<td>16</td>
</tr>
<tr>
<td>84-84</td>
<td>18</td>
</tr>
<tr>
<td>84-102</td>
<td>19</td>
</tr>
<tr>
<td>102-130</td>
<td>20</td>
</tr>
<tr>
<td>&gt;130</td>
<td>disabled</td>
</tr>
</tbody>
</table>

Table 66: Thresholds for Intensity Stereo

5.3.6. Time-Frequency Decision

The choice of time-frequency resolution used in Section 4.3.4.5 is based on R-D optimization. The distortion is the L1-norm (sum of absolute values) of each band after each TF resolution under consideration. The L1 norm is used because it represents the entropy for a Laplacian source. The number of bits required to code a change in TF resolution between two bands is higher than the cost of having those two bands use the same resolution, which is what requires the R-D optimization. The optimal decision is computed using the Viterbi algorithm. See tf_analysis() in celt/celt.c.

5.3.7. Spreading Values Decision

The choice of the spreading value in Table 59 has an impact on the nature of the coding noise introduced by CELT. The larger the $f_r$ value, the lower the impact of the rotation, and the more tonal the
coding noise. The more tonal the signal, the more tonal the noise should be, so the CELT encoder determines the optimal value for \( f_r \) by estimating how tonal the signal is. The tonality estimate is based on discrete pdf (4-bin histogram) of each band. Bands that have a large number of small values are considered more tonal and a decision is made by combining all bands with more than 8 samples. See spreading_decision() in celt/bands.c.

5.3.8. Spherical Vector Quantization

CELT uses a Pyramid Vector Quantization (PVQ) [PVQ] codebook for quantizing the details of the spectrum in each band that have not been predicted by the pitch predictor. The PVQ codebook consists of all sums of \( K \) signed pulses in a vector of \( N \) samples, where two pulses at the same position are required to have the same sign. Thus the codebook includes all integer codevectors \( y \) of \( N \) dimensions that satisfy \( \sum(\text{abs}(y(j))) = K \).

In bands where there are sufficient bits allocated PVQ is used to encode the unit vector that results from the normalization in Section 5.3.2 directly. Given a PVQ codevector \( y \), the unit vector \( X \) is obtained as \( X = y/\|y\| \), where \( \|\cdot\| \) denotes the L2 norm.

5.3.8.1. PVQ Search

The search for the best codevector \( y \) is performed by alg_quant() (vq.c). There are several possible approaches to the search, with a trade-off between quality and complexity. The method used in the reference implementation computes an initial codeword \( y_1 \) by projecting the normalized spectrum \( X \) onto the codebook pyramid of \( K-1 \) pulses:

\[
y_0 = \text{truncate} \left( \frac{(K-1) \times X}{\sum(\text{abs}(X))} \right)
\]

Depending on \( N, K \) and the input data, the initial codeword \( y_0 \) may contain from 0 to \( K-1 \) non-zero values. All the remaining pulses, with the exception of the last one, are found iteratively with a greedy search that minimizes the normalized correlation between \( y \) and \( X \):

\[
J = -X \times y / \|y\|
\]

The search described above is considered to be a good trade-off between quality and computational cost. However, there are other possible ways to search the PVQ codebook and the implementers MAY use any other search methods. See alg_quant() in celt/vq.c.
5.3.8.2. PVQ Encoding

The vector to encode, \( X \), is converted into an index \( i \) such that \( 0 \leq i < V(N,K) \) as follows. Let \( i = 0 \) and \( k = 0 \). Then for \( j = (N - 1) \) down to 0, inclusive, do:

1. If \( k > 0 \), set \( i = i + \frac{V(N-j-1,k-1) + V(N-j,k-1)}{2} \).
2. Set \( k = k + \text{abs}(X[j]) \).
3. If \( X[j] < 0 \), set \( i = i + \frac{V(N-j-1,k) + V(N-j,k)}{2} \).

The index \( i \) is then encoded using the procedure in Section 5.1.4 with \( ft = V(N,K) \).
6. Conformance

It is our intention to allow the greatest possible choice of freedom in implementing the specification. For this reason, outside of the exceptions noted in this section, conformance is defined through the reference implementation of the decoder provided in Appendix A. Although this document includes an English description of the codec, should the description contradict the source code of the reference implementation, the latter shall take precedence.

Compliance with this specification means that in addition to following the normative keywords in this document, a decoder’s output MUST also be within the thresholds specified by the opus_compare.c tool (included with the code) when compared to the reference implementation for each of the test vectors provided (see Appendix A.4) and for each output sampling rate and channel count supported. In addition, a compliant decoder implementation MUST have the same final range decoder state as that of the reference decoder. It is therefore RECOMMENDED that the decoder implement the same functional behavior as the reference. A decoder implementation is not required to support all output sampling rates or all output channel counts.

6.1. Testing

Using the reference code provided in Appendix A, a test vector can be decoded with

```
opus_demo -d <rate> <channels> testvectorX.bit testX.out
```

where <rate> is the sampling rate and can be 8000, 12000, 16000, 24000, or 48000, and <channels> is 1 for mono or 2 for stereo.

If the range decoder state is incorrect for one of the frames, the decoder will exit with "Error: Range coder state mismatch between encoder and decoder". If the decoder succeeds, then the output can be compared with the "reference" output with

```
opus_compare -s -r <rate> testvectorX.dec testX.out
```

for stereo or

```
opus_compare -r <rate> testvectorX.dec testX.out
```

for mono.

In addition to indicating whether the test vector comparison passes, the opus_compare tool outputs an "Opus quality metric" that indicates
how well the tested decoder matches the reference implementation. A quality of 0 corresponds to the passing threshold, while a quality of 100 is the highest possible value and means that the output of the tested decoder is identical to the reference implementation. The passing threshold (quality 0) was calibrated in such a way that it corresponds to additive white noise with a 48 dB SNR (similar to what can be obtained on a cassette deck). It is still possible for an implementation to sound very good with such a low quality measure (e.g. if the deviation is due to inaudible phase distortion), but unless this is verified by listening tests, it is RECOMMENDED that implementations achieve a quality above 90 for 48 kHz decoding. For other sampling rates, it is normal for the quality metric to be lower (typically as low as 50 even for a good implementation) because of harmless mismatch with the delay and phase of the internal sampling rate conversion.

On POSIX environments, the run_vectors.sh script can be used to verify all test vectors. This can be done with

```
run_vectors.sh <exec path> <vector path> <rate>
```

where `<exec path>` is the directory where the opus_demo and opus_compare executables are built and `<vector path>` is the directory containing the test vectors.

### 6.2. Opus Custom

Opus Custom is an OPTIONAL part of the specification that is defined to handle special sample rates and frame rates that are not supported by the main Opus specification. Use of Opus Custom is discouraged for all but very special applications for which a frame size different from 2.5, 5, 10, or 20 ms is needed (for either complexity or latency reasons). Because Opus Custom is optional, streams encoded using Opus Custom cannot be expected to be decodable by all Opus implementations. Also, because no in-band mechanism exists for specifying the sampling rate and frame size of Opus Custom streams, out-of-band signaling is required. In Opus Custom operation, only the CELT layer is available, using the `opus_custom_*` function calls in `opus_custom.h`. 
7. Security Considerations

Like any other audio codec, Opus should not be used with insecure ciphers or cipher-modes that are vulnerable to known-plaintext attacks. In addition to the zeros used in Opus padding, digital silence frames generate predictable compressed results and the TOC byte may have an easily predictable value.

Implementations of the Opus codec need to take appropriate security considerations into account, as outlined in [DOS]. It is extremely important for the decoder to be robust against malicious payloads. Malicious payloads must not cause the decoder to overrun its allocated memory or to take an excessive amount of resources to decode. Although problems in encoders are typically rarer, the same applies to the encoder. Malicious audio streams must not cause the encoder to misbehave because this would allow an attacker to attack transcoding gateways.

The reference implementation contains no known buffer overflow or cases where a specially crafted packet or audio segment could cause a significant increase in CPU load. However, on certain CPU architectures where denormalized floating-point operations are much slower than normal floating-point operations, it is possible for some audio content (e.g., silence or near-silence) to cause an increase in CPU load. Denormals can be introduced by reordering operations in the compiler and depend on the target architecture, so it is difficult to guarantee that an implementation avoids them. For architectures on which denormals are problematic, adding very small floating-point offsets to the affected signals to prevent significant numbers of denormalized operations is RECOMMENDED. Alternatively, it is often possible to configure the hardware to treat denormals as zero (DAZ). No such issue exists for the fixed-point reference implementation.

The reference implementation was validated in the following conditions:

1. Sending the decoder valid packets generated by the reference encoder and verifying that the decoder’s final range coder state matches that of the encoder.
2. Sending the decoder packets generated by the reference encoder and then subjected to random corruption.
3. Sending the decoder random packets.
4. Sending the decoder packets generated by a version of the reference encoder modified to make random coding decisions.
(internal fuzzing), including mode switching, and verifying that the range coder final states match.

In all of the conditions above, both the encoder and the decoder were run inside the Valgrind [Valgrind] memory debugger, which tracks reads and writes to invalid memory regions as well as the use of uninitialized memory. There were no errors reported on any of the tested conditions.
8. IANA Considerations

This document has no actions for IANA.
9. Acknowledgements

Thanks to all other developers, including Raymond Chen, Soeren Skak Jensen, Gregory Maxwell, Christopher Montgomery, and Karsten Vandborg Soerensen. We would also like to thank Igor Dyakonov, Christian Hoene, and Jan Skoglund for their help with subjective testing of the Opus codec. Thanks to Andrew D’Addesio, Elwyn Davies, Ralph Giles, Christian Hoene, John Ridges, Ben Schwartz, Kat Walsh, Keith Yan, and many others on the Opus and CELT mailing lists for their bug reports and feedback.
10. Copying Conditions

The authors agree to grant third parties the irrevocable right to copy, use and distribute the work (excluding Code Components available under the simplified BSD license), with or without modification, in any medium, without royalty, provided that, unless separate permission is granted, redistributed modified works do not contain misleading author, version, name of work, or endorsement information.
11. References

11.1. Normative References

[rfc2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119.

11.2. Informative References


Video and Data Recording, 1979.

[coding-thesis]  

[PVQ]  

[Kaba86]  

[Valgrind]  
"Valgrind website", <http://valgrind.org/>.

[Google-NetEQ]  

[Google-WebRTC]  

[Opus-git]  
"Opus Git Repository", <git://git.xiph.org/opus.git>.

[Opus-website]  
"Opus website", <http://opus-codec.org/>.

[Vorbis-website]  

[Matroska-website]  

[Vectors-website]  
"Opus Testvectors (website)", <http://opus-codec.org/testvectors/>.

[Vectors-proc]  
"Opus Testvectors (proceedings)", <http://www.ietf.org/proceedings/83/slides/slides-83-codec-0.gz>.

[line-spectral-pairs]  
Wikipedia, "Line Spectral Pairs", 


[rangle-coding]  
Wikipedia, "Range Coding",  

[Hadamard]  
Wikipedia, "Hadamard Transform",  

[Viterbi]  
Wikipedia, "Viterbi Algorithm",  

[Whitening]  
Wikipedia, "White Noise",  

[LPC]  
Wikipedia, "Linear Prediction",  

[MDCT]  

[FFT]  
Wikipedia, "Fast Fourier Transform",  

[z-transform]  
Wikipedia, "Z-transform",  

[Burg]  
Burg, JP., "Maximum Entropy Spectral Analysis",  
Proceedings of the 37th Annual International SEG Meeting,  
Vol. 6 1975.

[Schur]  
Le Roux, J. and C. Gueguen, "A fixed point computation of  
Int. Conf. Acoust., Speech, Signal Processing, pp. 257-259,  
October 1977.

[Princen86]  
Princen, J. and A. Bradley, "Analysis/synthesis filter  
bank design based on time domain aliasing cancellation",  

[Valin2010]  
Valin, JM., Terriberry, T., Montgomery, C., and G.  
Maxwell, "A High-Quality Speech and Audio Codec With Less

[Zwicker61]
Appendix A. Reference Implementation

This appendix contains the complete source code for the reference implementation of the Opus codec written in C. By default, this implementation relies on floating-point arithmetic, but it can be compiled to use only fixed-point arithmetic by defining the FIXED_POINT macro. The normative behaviour is defined as the output using the floating-point configuration. Information on building and using the reference implementation is available in the README file.

The implementation can be compiled with either a C89 or a C99 compiler. It is reasonably optimized for most platforms such that only architecture-specific optimizations are likely to be useful. The FFT [FFT] used is a slightly modified version of the KISS-FFT library, but it is easy to substitute any other FFT library.

While the reference implementation does not rely on any _undefined behavior_ as defined by C89 or C99, it relies on common _implementation-defined behavior_ for two’s complement architectures:

- Right shifts of negative values are consistent with two’s complement arithmetic, so that a>>b is equivalent to floor(a/(2**b)),

- For conversion to a signed integer of N bits, the value is reduced modulo 2**N to be within range of the type,

- The result of integer division of a negative value is truncated towards zero, and

- The compiler provides a 64-bit integer type (a C99 requirement which is supported by most C89 compilers).

In its current form, the reference implementation also requires the following architectural characteristics to obtain acceptable performance:

- Two’s complement arithmetic,

- At least a 16 bit by 16 bit integer multiplier (32-bit result), and

- At least a 32-bit adder/accumulator.
A.1. Extracting the source

The complete source code can be extracted from this draft, by running the following command line:

```bash
  o  cat rfcXXXX.txt | grep '^ \ \ ###' | sed -e 's/...###//' | base64 -d > opus-rfcXXXX.tar.gz
  o  tar xzvf opus-rfcXXXX.tar.gz
  o  cd opus-rfcXXXX
  o  make
```

On systems where the provided Makefile does not work, the following command line may be used to compile the source code:

```bash
  o  cc -O2 -g -o opus_demo src/opus_demo.c 'cat *.mk | grep -v fixed | sed -e 's/\s\s\s=//\s\s\s' -e 's/\s\s/\s\s\s' -DOPUS_BUILD -Dinclude -Dcelt -Isilk -Isilk/float -DUSE_ALLOCA -Drestrict= -lm
```

On systems where the base64 utility is not present, the following commands can be used instead:

```bash
  o  cat rfcXXXX.txt | grep '^ \ \ ###' | sed -e 's/...###//' > opus.b64
  o  openssl base64 -d -in opus.b64 > opus-rfcXXXX.tar.gz
```

The SHA1 hash of the opus-rfcXXXX.tar.gz file is 726ef2b721d68b2dd663b67bdf0629eeaa2b5bb34.

A.2. Up-to-date Implementation

As of the time of publication of this memo, an up-to-date implementation conforming to this standard is available in a Git repository [Opus-git]. Releases and other resources are available at [Opus-website]. However, although that implementation is expected to remain conformant with the standard, it is the code in this document that shall remain normative.

A.3. Base64-encoded Source Code

```bash
###H4sIAOoR7U8CA+xde3PiSJLvY/kUud0RM+DD20DX9PTMRGOQsXZ4+BDY3XF7oZVF###
###AdoWEqO327Z++6XWVUSkA0PZ7vhJqs5EtEFSZVZmVYbWF+uhduehv++Nzu/4OXjz###
###Mp9D/JydnhfBv/GS/4e9q9ezsqFY9PDupvTmsHh4d1d7AyZu/0Cer3F/k4ybbv9G7###
###+qx2Wy9hmtPT41Xtf3R0VhPtxs07PTw5wvav4s8380ali/V/803DnD541mQZQff/+###
###eL+GtdgVGvzAwAv9oAyfrPm00vMmZdC+PMwZtK2ZFBpBRGXpmYPiWWS5AzufvzHD2
```
Internet-Draft           Interactive Audio Codec               June 2012

[Internet-Draft] Interactive Audio Codec               June 2012

###WG+ZQHBxrPzfPFzMnOFIZ8r8TfMcKf/rzxTX0QEZdpuRY2eKNIw7VZe2c8Y87bM
###h/9FMRKP7qM505s55g8znJ5oi1zedk0m8JNzcFkuMumubEm8/JNS87Cb7caT5c
##KqcMc6mesR+f9kL8k5U1DFf1ymRprh9N6bV5HSMvmGwSTzuajc7md06mQP+cyn5
###1MxD1nx6kafynfYxy2ZgsSY0inXSVnFmaLmAACC/LFUuM8Tf1ub51ZlpjAdv6X
##y+5+7bLT5h59poY2jW6b7R5yK5r152ucd1Ttto/9KQYu0u1Pm7+j14RTdTP3b1z1
###n+rO1tSz4n6TYU/7QsTnxz3N001KnvUn7u7vSdStDncsKsdgjNq4b1h+30vuyU
###9BqCcnxRvWPAdO+x1DB181cKrWvSuOH5g3M38PkkX/D/a9y8514u6DeEgmiR
##vYjBrpywju2QaBo4TzGrdUqRft6E6e2v3c/jay0TphOsLKSk14Kh4715Epsy+0
###Y+r9UKErBvWn6SmngNMCN1yBovYyjx1Jk/q/cKtE809eR0bW30Uk+A3wUSvEqJhN
###TBF3c5cT5I5lehRCgkUAQ8k4etLPcUz25xpYAPxv58DvHnKh1GxKSSAsh0aSa/fz
###5F1gLxw7W0njpjGZN1L12WxnF5jmLdtDlDhXkOvW1Dzhds+qOqcCvQOBDwu63F9F
##bkDz3CnCebLukxr1CMkAX71SCqDFTCbh1in5B3Ubcch8A/Kb7E4apqYqdk1
###i6Sf31LOGzi1Bztg/tpMy9xd/MmxM7fSdCTepZp1pE3a5qV03kjiwe2++nwE4V4G
###5kD3R8yV6eHjFqVU1K9/kGMBmbc5jrSTKvjrE5b5iEvB2s6o1i8ac90mTRMLTH
###vyd2vX7O1qWChB3s5fVh95x8p+jB/fVPhP3h3AtVv+nuq0t4v88w8+lL9F/N81
###m8R/7e1/1vZ/z0y/g9zK4cHtkRgNjFy16PFGuQk9QaST7jskBtk8OHsaLtnBe
###aaEBAydp1jkYyKCUrQ1Ljdj/EAIwi3nnp1jioa1wugLWvC9L436mohxR2
###x3rs1D6EvMMeE20nkCRCRAn8fcyNQ9j3Y3H9PBxFUXPxDU924bAX72TAxEQ2V1A39
###nt0cm16LmxRmJ8Btd3N9H7bxd1fH8eU6m18bOOGtSfyuhj4P7uTBFB3p
###BC27E7zlAOGD78aN3pMj454nW5+bJ+t6CnKahyfaFR/a/rVqcoaut+orj03hv
###G6xZr/kVprvV99cEBOpiyQgegTXNvzuldrC9vG6t2pBvwKq9BwB7T1L1pF
###+LMNvB8u+NLXe93U67yFmsPrunm4/56f2tpWkt5ry5A1r6b6V03A6t1gnhtqr7
###1rS+Z6j+hdp2ARMFg29zhU29n3W0d3WOCompdFrGtiWsnR8NRM1ONP1OdoY9axA6Cv
###5j+vp+va++/+X4YNfNTu5FvJFKgG8iJ48O0AEZNBfEP1tmgat7f7iLvFr/+iw1FDRd
###cJRui7L7N6kfeEgPys4dwZ4u3h8s0hL642jWjpyHc5p3w3ZQj6Yr15ziozXO6d/
###+j+E14R4YJMVGjx6mSr3X64CNH6dmMnpQtohZda61OPD49XjmstUtPBH7g
####AAnnwkDva2i6sRz3PjDy451KVqIpmHNMNYw3z3m4/7W6vKn5BpqV9qA2B3tZsvGg
###B1cc2z0oAs/aJCECA1SlS07eBpqdd1s vb1H1ShbQ7K17s84muX0UN7g19HJRicvWR
###eFkh2HttZio+oPUOEskUEFpxb1LISwA5qAg1QPXz5wWp7qU13KratZXZX9PQOMXL
###ohZ/2kBfiLBNdp4EvOup6cpQvOwAk92Z3pR8fH3a/4C/Thk8uA1hrsawco3
###h/2Fpm5+xu7FRAb2sz1u+25FisWnXa/a/vr9F1LYa63+p6Y3xhrg6a/ynot
###+v2r+Vdve9h1f1v7a+9Vr7VkfWdpsyBuAWdYFPLHf
###fScDx19iFMYTK9W9Wol/1H4Aqbw7Csdlno7esNqFywDULx1Dx1QbvebV5f0U0Of
###MuyVvVz/Qxi4bOnzj7jwonoFKhXuKbUsu+ojW6x6OluQkWvHzadcey1LQOG7Dx/6Az
###AHeReX3YQ+E46L17ji71blBeEYKJ18LVzdXu5y875BVU77p9P5wOCd+cfpzGj2J5
###+1dl1+NaGI1rXxSYMqTcope/npGEdFrCnHDuzzpsy8uTPumoscI1+mrgn5M0UVCb
###OpFQxRKGR1+wCnDvCj30X7qnm7Y7rp+q7EfDvznsKehjmhpR6Xr+PXR7wX3D7I
###MyYcHb7yV/EFlo41KaPahbH/3EVTK5Y6Hshs886tKXRuDEEFk8AejW5CDCyEpWl
###yMnEnvak/b/9e1fa41b9eW1QV50rFq3oQ7K0aaaaz1Jzt880fQmoRq7W9111Gxas2HAMUj
###+qwo+b30D1/qA0P0jt1eTveG93D0X9gPy152CZfMY1/y+PHbFm727fPmFDouI4GB
###+3pLM2bQx8rsDMLC3omSGeTm0mgxNF8jgWBdA2GSAeBoGcsN4kTDKjvc+PZ
###58czclwxIb11li1911ni1zn1dL+x1TdMclnFyR9kWwkmL5vavelv9CefnMO7QF6jH9LUKD
###NMXKfLIw57QnO3ExwNDExRfHllee2Cd+R8w06hFidTCfpgud87H8+Y3Kg2N8xnt8
###TafDTP+amU9ZvXvuxU9+9X69Z2W/Hdh/FxvaflinyY2ZJH+r+O6Ah2E4We6qBhu
###EG1q7Nq5u/darpvZ400b8sSmx2N3/kfNkFNO48SvD49MEvKxYxM95gFPM2Zz
###n6mxz9NhWhs3wWqozw3w11YLP42BbaaYFwUtB+Vx1klyExww2lW4jAgj0JTW7BN7
###vYve51DqsWSIa3r7v0s5vd6G6E20u/u74og77fzbms1e6l6w04c9sBndN52Ew7VHy
###u2w7s7fEFmhh9XENQ0wqfhU2F4uJ9AT0Jqs4uR+CZ0HvUn8JNvmZx4a4TOGrvJZT

###ORZ2L7F4vcZLmdhb08PFLyeFmxMlFUhrv9H5vw5pX/O8v+rvrb7X7PzPv19d6P8L
###/x+h/y0/0/4+v9D/F/r/Qv/+/E9R/BqV1BC9g2/P/9r/6K373psRuW6N0VmVd/Bf6w
###/VVg56ATI2edlb4qcTCGj9GzzZSFymcrVCizeMBbZTRarrXbffKaW7ZE3tbjFDPW162
###/X+h/y/0/4X+v9D/F/r/Qv//E9R/BqV1BC9gZ/PL9rK6K373psRuW6NOmVd/Bf6w
###Vg56ATI2edlb4qcTCGj9GzzZSFymcrVCizeMBbZTRarrXbffKaW7ZE3tbJFDW162
###LpWThnU4X1b6X0sZd7Jhe4QTRg5H9+7DVcr0e5VGIc5rLmNtCy+2rmEPGY5O4sN0
###/HHQy7mnfFIR40cjoZbpVOp41n6S4bOdwqT6b2acx24HD7N35KKHAGqwa3cT
###/Q43y0e+3zavvWt+uDyDyZ3HdwF8s5aNh9ichzIZydLi1n+1s2ZhtoU6BO+B0CN1+
###/dxdic04ME38M3/lrby1/NoHh/ME7ha30438h8x9v8r8vpr//fMD7Ep7cd6adq9FNN
###/7aeaap+N/6ekW2X9N5aPft5L0ieC10wFkJLsQmBi19+5rYEVufTMC1721LmTnSQt
###/d0oeE4AH7EG+ORIdqtE00mctv1zmSmuiovdFDDWzeeUn145v9+LP2Ux7x7q7aT83s
###/p5bejivzQBv9mt3r5cVZLAuah+LDZsm9k1wDCADW90VDa+a+pWbnM0K/f8j1m5
###/etZGm31vPjkogUy6msCeXPaKt±fsi8W1T4xG8L09MKv5B5RCEmI/F/gQf3q5e5a6
###/DVFVOK3xLrKk4mRc0Gc00RpyxaFrrWAeC10q7as8qyuf+cwLs7j7/OFv0BwpyEF
###/9B90OUC9q98G9naRQtrG7NvSmwAvyymn+/L6+cizv8RLLtc7oaSbn/qVd9x777/
###/gceLr+/mdx/704/c/n/zufx3B39P9F/6w/u8+Dsc1CepKfsf1i4y928x2r
###/Da6it7Bz/F/c3Y374Hoaws2UbYe3LdSs29c18Jn50zcMxfXxkmLgD+PChlW55m
###/3npr1byq4gCmJe6zdpjxmcdqK9+9nc42CA2JQX8egOBQRT884Cyc1jBPCF/P/H
###/VvVdBpWu6b3j9G7b5Pyb7V1kGbyD91f2aqZUX+gV3TPhIdKutesEHOFplD5tci
###/WqF2red9GEWzGdC57bTwT3LXhPABQkx2NvKd2bqCQhIXYUEPWKHSByg5J9H0
###/QRT8qu8U9hdJO10A2Cbu5B501iyimODbbQ0734EE413Hz+wV9Wg9cUPQLBQ+i110
###/BwujN1jv7x3b7u2O/0+65p9FWXe/8uxe7smaZ0RUiAJVFVmSc7noudLjiscHe
###/#3a4xP9uX9qVyUvtdi3vNGmeCrhK6E1Izxt0Q1tAmXrtssiqsgx9etNsbC70q_qx7Pc
###/UCCQJ4NcJntWjWDMwCDVc/ALFjCvnSdCD1AYAI+/zb9MmdpCxeR5M812EeD9wMV8s
###/V+Fh9fP1v7q9s0pVzahbWv4m7u/ac9Nv0H9fRf6cBcCIKGIEmWfuA56a2y2
###/e217U0ytqyYe2l59pcSe4r9PPrdUrFeuBYXWfsoVKE6/EB1rXq9C3EY9v9yq/VFa2
###/#qGR0EPUCU+HloLqDfW02dSAR+g17Mc10H7VYpiZi2jwvHWZDc/tAsLeBd7R3B5
###/xd/7/jY/siihB3rOXTy7/3zwpHmO4FPxGsx1rEjyXPYRUBuBF8V1n4c/1PYYKES
###/dADPK24JaNenBtguZB5Bszw666Ad9ptwzoKDNdNokhCL+r+s3PrDbgBgnaxzOCIsb
###/oAmNa97Jv6xbxliOvSSKO/x/65p9FWXe/8uxe7smaZ0RUiAJVFVmSc7noudLjiscH
###/#2w6n6gzvQepi6T7mu0xmldKtQC2gQK1KDyeowM7x4A1bf9nii6K7ER9Vsvvo1e2M
###/vva1+O1zVCHW7N8mhc+niswAAqjwH7YwgoBnoBtKlZ174Baq+1E1Pa0eCF7vCNK
###/#660S3c8EB076nyo02Z2uj1kkerTwpHyWjJicHDeVY3xNNCC7km7sMY+5a5XeQrQ
###/#55KVKeAkve8Bvexsly6VZLoKzEkvGcKJLzsWm1zGBqgY3ILG+3qOj3pBP6kHoJ
###/#vvn+Vee11U2FzBO3H65x5NKZFXR29x5af/HGFMC6oAK2G1niz06FDFYzGYB1ALTdG8
###/L8C8bE3cRaRqGbPdC/wdGib7Y48zTidUH7iykxqaqygLz1JvKvDFrKceGnfGwvZ
###/#2Y/ki+pLc14E66cdgtNSmZt6pJgGuz42B1zLzg5B++8B+zdxt+4w2QLyEv0114x
###/#xYuiJd92e9+wvVGUGJvXklG0QXQhKJ1uqzL1zTi0JFPQonKnJlYKxeV4tooOfK7
###/#JnJzQuBvBA8kwOoSp8RKnwsEqv02ebqjBv9HyCy6JmmFdnDMv8SwboW
###/ugYOYQ7u6L2LTSX7N7bpdm2zkr5Onu3Qrt2kethZYXBSwr24Rkg2DNn172H88u7w7
###/onE4k3TgTSUBOA7x3blQX0MveCjyPNVngIo+YbnOajvijDin+B/tPxF+S
###/nnOnVCjFPDQSMG2isyvgnjo6p7jklKYAmvoXsIZUwocfC1KQY6G3taW7/yonwyeEHB
###/#HyD0uak3uB5hx+3FttriOs0ef8WVn6o3pl+4q3Ux4skk2v7z3oT0/cY4ufB2uAa/
###/#R7j69C92+G2JVRxmvac/Y+yv/1632aLC.PN5ud5u10C3+/+vVar1qn3/1x1urL7+7/F
###/1/x/v3/f5+4Vv9f3P/PPF8p7uZ4hqdHl7fgy63T78fb29N94keE9p1C8
###/#T18Jxpt4drstCaCe46F2TpaAdfem5Ntrgv2n370Nh2OQQ+ppR2W16u1rAw7T7B
###/hab10SBr5DDeC5WF1QPoBonyu7kaAc7ed5gl06j6F6Ag0wP75m9AgBvJjaOBG
###/#PR9B5a1Q3XxYYIYX/7XDrxtn0z0f+//+uur+zqGrWQW3jxmknkSjRas/obseOryRj
###/#V56EAOqK9ustqnhB0+S/5vFRqH8zwiiF3PgnKiNkAp/F/Q3Ebg7T7Yt+QBSwHuy
###/m4XSK90M1Bp/tx+B4uYfy1jH9niPt1nRLJQFt48maw/aJ0H8NPPNgPksG3xuSJKP

A.4. Test Vectors

Because of size constraints, the Opus test vectors are not distributed in this draft. They are available in the proceedings of the 83th IETF meeting (Paris) [Vectors-proc] and from the Opus codec website at [Vectors-website]. These test vectors were created specifically to exercise all aspects of the decoder and therefore the audio quality of the decoded output is significantly lower than what

Opus can achieve in normal operation.

The SHA1 hash of the files in the test vector package are:

- e49b2862ceec7324790ed8019eb9744596d5be01 testvector01.bit
- b809795ae1bcd606049d76e4ad24236257135e0 testvector02.bit
- e0c4ecaeb44d35a2f5b6575cd996848e5ee2acc testvector03.bit
- a0f870cbef14eb71fa9066ef3ee96e59c9a75187 testvector04.bit
- 9b3d92b48965dfe9edf7b8a85edd4309f8cf7c8 testvector05.bit
- 28e66769ab17e177f2875283c14b19690cbe4e57 testvector06.bit
- bacf467be3215fc7ec288f29e2477de1192947a6 testvector07.bit
- ddbe08b688bf934071f3893cd0030ce48db1af testvector08.bit
- 3932d961944dab1201645b8eeaad595d5705ecb testvector09.bit
- 521eb2a1e0cc9c31bb87406733077c2d3b10c1900 testvector10.bit
- 6bc8f3146fc96450c901b16c3d464ccdf4d5d96 testvector11.bit
- 338c3f1b4b97226bc60b4c10388becbc6de06b28f testvector12.bit
- a20a2122d42de644f94445e2018538559623af testvector01.dec
- 48ac1ff1995250a756e1e17bd32acefa8cd2b820 testvector02.dec
- d15567e919db2d0e818727092c0af8dd9df23c95 testvector03.dec
- 1249dd28f5bde39a66fd699449dca7a8316342 testvector04.dec
- 39eae37e5d26a456d2c24483060132ff7eae2143 testvector05.dec
- a294fcf17e3157768c46c5e0f2116dec02d37ee2 testvector06.dec
- 2bf550e2f072e0941438db3f338fe9944385848 testvector07.dec
- 2695c1f2df1f9748e0bf07249c70fd7b87f61680 testvector08.dec
- 12862add5d3a9d2a707934a542a2f039b992bb testvector09.dec
- a081252bb2ba902fdec500530891f47e2a373d84 testvector10.dec
- dfd0f844f2a42df506934fac2100a3c03beec711 testvector11.dec
- 8c16b2a1fb60e3550ba165068f9d7341357fdb63 testvector12.dec
Appendix B. Self-Delimiting Framing

To use the internal framing described in Section 3, the decoder must know the total length of the Opus packet, in bytes. This section describes a simple variation of that framing which can be used when the total length of the packet is not known. Nothing in the encoding of the packet itself allows a decoder to distinguish between the regular, undelimited framing and the self-delimiting framing described in this appendix. Which one is used and where must be established by context at the transport layer. It is RECOMMENDED that a transport layer choose exactly one framing scheme, rather than allowing an encoder to signal which one it wants to use.

For example, although a regular Opus stream does not support more than two channels, a multi-channel Opus stream may be formed from several one- and two-channel streams. To pack an Opus packet from each of these streams together in a single packet at the transport layer, one could use the self-delimiting framing for all but the last stream, and then the regular, undelimited framing for the last one. Reverting to the undelimited framing for the last stream saves overhead (because the total size of the transport-layer packet will still be known), and ensures that a "multi-channel" stream which only has a single Opus stream uses the same framing as a regular Opus stream does. This avoids the need for signaling to distinguish these two cases.

The self-delimiting framing is identical to the regular, undelimited framing from Section 3, except that each Opus packet contains one extra length field, encoded using the same one- or two-byte scheme from Section 3.2.1. This extra length immediately precedes the compressed data of the first Opus frame in the packet, and is interpreted in the various modes as follows:

- Code 0 packets: It is the length of the single Opus frame (see Figure 25).
- Code 1 packets: It is the length used for both of the Opus frames (see Figure 26).
- Code 2 packets: It is the length of the second Opus frame (see Figure 27).
- CBR Code 3 packets: It is the length used for all of the Opus frames (see Figure 28).
- VBR Code 3 packets: It is the length of the last Opus frame (see Figure 29).
Figure 25: A Self-Delimited Code 0 Packet

Figure 26: A Self-Delimited Code 1 Packet

Figure 27: A Self-Delimited Code 2 Packet
Figure 28: A Self-Delimited CBR Code 3 Packet
Figure 29: A Self-Delimited VBR Code 3 Packet
Authors’ Addresses

Jean-Marc Valin
Mozilla Corporation
650 Castro Street
Mountain View, CA  94041
USA

Phone: +1 650 903-0800
Email: jmvalin@jmvalin.ca

Koen Vos
Skype Technologies S.A.
Soder Malarstrand 43
Stockholm, 11825
SE

Phone: +46 73 085 7619
Email: koen.vos@skype.net

Timothy B. Terriberry
Mozilla Corporation
650 Castro Street
Mountain View, CA  94041
USA

Phone: +1 650 903-0800
Email: tterriberry@mozilla.com
Codec Requirements
draft-ietf-codec-requirements-02

Abstract

This document provides specific requirements for an Internet audio codec. These requirements address quality, sampling rate, bit-rate, and packet loss robustness, as well as other desirable properties.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 15, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
Table of Contents

1. Introduction ................................................. 3
2. Applications ............................................... 4
   2.1. Point to point calls ............................... 4
   2.2. Conferencing ....................................... 4
   2.3. Telepresence ........................................ 5
   2.4. Teleoperation and Remote Software Services .......... 5
   2.5. In-game voice chat ................................ 6
   2.6. Live distributed music performances / Internet music lessons ............................................. 6
   2.7. Other applications ................................... 7
3. Constraints Imposed by the Internet on the Codec ........... 8
   3.1. Security .............................................. 9
4. Detailed Basic Requirements ................................ 10
   4.1. Operating space ..................................... 10
   4.2. Quality and bit-rate ................................ 10
   4.3. Packet loss robustness ................................ 11
   4.4. Computational resources .............................. 11
5. Additional considerations .................................... 14
   5.1. Low-complexity audio mixing ........................ 14
   5.2. Encoder side potential for improvement ............ 14
   5.3. Layered bit-stream ................................... 14
   5.4. Partial redundancy ................................... 15
   5.5. Stereo support ....................................... 15
   5.6. Bit error robustness ................................ 15
   5.7. Time stretching and shortening ..................... 15
   5.8. Input robustness ..................................... 16
   5.9. Legacy compatibility ................................ 16
6. Security Considerations ..................................... 17
7. IANA Considerations ......................................... 18
8. Acknowledgments ............................................. 19
9. Informative References ...................................... 20
Authors’ Addresses ............................................... 21
1. Introduction

This document provides requirements for an audio codec designed specifically for use over the Internet. The requirements attempt to address the needs of the most common Internet interactive audio transmission applications and to ensure good quality when operating in conditions that are typical for the Internet. These requirements address the quality, sampling rate, delay, bit-rate, and packet loss robustness. Other desirable codec properties are considered as well.

Throughout this document, we will use the following conventions when referring to the sampling rate of a signal:

- **Narrowband**: 8 kHz sampling rate
- **Wideband**: 16 kHz sampling rate
- **Super-wideband**: 32 kHz sampling rate
- **Full-band**: 44.1/48 kHz and above

Codec bit-rates in bits per second (b/s) will be considered without counting any overhead (IP/UDP/RTP headers, padding, ...). The codec delay is the total algorithmic delay when one adds the codec frame size to the "look-ahead". It is thus the minimum theoretically achievable end-to-end delay of a transmission system that uses the codec.
2. Applications

The following applications should be considered for Internet audio codecs, along with their requirements:

- Point to point calls
- Conferencing
- Telepresence
- Teleoperation
- In-game voice chat
- Live distributed music performances / Internet music lessons
- Other applications

2.1. Point to point calls

Point to point calls are voice over IP (VoIP) calls from two "standard" (fixed or mobile) phones, and implemented in hardware or software. For these applications, a wideband codec is required, along with narrowband support for compatibility with legacy telephony equipment (PSTN). It is expected for the range of useful bit-rates to be 12 - 32 kb/s for wideband speech and 8 - 16 kb/s for narrowband speech. The codec delay must be less than 40 ms, but no more than 25 ms is desirable. Support for encoding music is not required, but it is desirable for the codec not to make background (on-hold) music excessively unpleasant to hear. Also, the codec should be robust to noise (produce intelligible speech and no annoying artifacts) even at lower bit-rates.

2.2. Conferencing

Conferencing applications (which support multi-party calls) have additional requirements on top of the requirements for point-to-point calls. Conferencing systems often have higher-fidelity audio equipment and have greater network bandwidth available -- especially when video transmission is involved. For that reason, support for super-wideband audio becomes important, with useful bit-rates in the 32 - 64 kb/s range. The ability to vary the bit-rate according to the "difficulty" of the audio signal (VBR) is a desirable feature for the codec. This not only saves bandwidth "on average", but it can also help conference servers make more efficient use of the available bandwidth by using more bandwidth for important audio streams and less bandwidth for less important ones (e.g. background noise).
Conferencing end-points often operate in hands-free conditions, which creates acoustic echo problems. For this reason lower delay is important, as it reduces the quality degradation due to any residual echo after acoustic echo cancellation (AEC). For this reason, the codec delay must be less than 30 ms for this application. An optional low-delay mode with less than 10 ms delay is desirable, but not required.

Most conferencing systems operate with a bridge that mixes some (or all) of the audio streams and sends them back to all the participants. In that case, it is important that the codec not produce annoying artefacts when two voices are present at the same time. Also, this mixing operation should be as easy as possible to perform. To make it easier to determine which streams have to be mixed (and which are noise/silence), it must be possible to measure (or estimate) the voice activity in a packet without having to fully decode the packet (saving most of the complexity when the packet need not be decoded). Also, the ability to save on the computational complexity when mixing is also desirable, but not required. For example, a transform codec may make it possible to mix the streams in the transform domain, without having to go back to time-domain. Low-complexity up-sampling and down-sampling within the codec is also a desirable feature when mixing streams with different sampling rates.

2.3. Telepresence

Most telepresence applications can be considered to be essentially very high-quality video-conferencing environments, so all of the conferencing requirements also apply to telepresence. In addition, telepresence applications require super-wideband and full-band audio capability with useful bit-rates in the 32 - 80 kb/s range. While voice is still the most important signal to be encoded, it must be possible to obtain good quality (even if not transparent) music.

Most telepresence applications require more than one audio channel, so support for stereo and multi-channel is important. While this can always be accomplished by encoding multiple single-channel streams, it is preferable to take advantage of the redundancy that exists between channels.

2.4. Teleoperation and Remote Software Services

Teleoperation applications are similar to telepresence, with the exception that they involve remote physical interactions. For example, the user may be controlling a robot while receiving real-time audio feedback from that robot. For these applications, the delay has to be less than 10 ms. The other requirements of telepresence (quality, bit-rate, multi-channel) apply to...
teleoperation as well. The only exception is that mixing is not an important issue for teleoperation.

The requirements for remote software services are similar to those of teleoperation. These applications include remote desktop applications, remote virtualization, and interactive media application being rendered remotely (e.g. video games rendered on central servers). For all these applications, full-band audio with an algorithmic delay below 10 ms are important.

2.5. In-game voice chat

An increasing number of computer/console games make use of VoIP to allow players to communicate in real-time. The requirements for gaming are similar to those of conferencing, with the main difference being that narrowband compatibility is not necessary. While for most applications a codec delay up to 30 ms is acceptable, a low-delay (<10 ms) option is highly desirable, especially for games with rapid interactions. The ability to use VBR (with a maximum allowed bitrate) is also highly desirable because it can significantly reduce the bandwidth requirement for a game server.

2.6. Live distributed music performances / Internet music lessons

Live music over the Internet requires extremely low end-to-end delay and is one of the most demanding application for interactive audio transmission. It has been observed that for most scenarios, total end-to-end delays up to 25 ms could be tolerated by musicians, with the absolute limit (where none of the scenarios are possible) being around 50 ms [carot09]. In order to achieve this low delay on the Internet -- either in the same city or a nearby city -- the network propagation time must be taken into account. When also subtracting the delay of the audio buffer, jitter buffer, and acoustic path, that leaves around 2 ms to 10 ms for the total delay of the codec. Considering the speed of light in fiber, every 1 ms reduction in the codec delay increases the range over which synchronization is possible by approximately 200 km.

Acoustic echo is expected to be an even more important issue for network music than it is in conferencing, especially considering that the music quality requirements essentially forbid the use of a "nonlinear processor" (NLP) with the AEC. This is another reason why very low delay is essential.

Considering that the application is music, the full audio bandwidth (44.1 or 48 kHz sampling rate) must be transmitted with a bit-rate that is sufficient to provide near-transparent to transparent quality. With the current audio coding technology, this corresponds
to approximately 64 kb/s to 128 kb/s per channel. As for telepresence, support for two or more channels is often desired, so it would be useful for a codec to be able to take advantage of the redundancy that is often present between audio channels.

2.7. Other applications

The above list is by no means a complete list of all applications involving interactive audio transmission on the Internet. However, it is believed that meeting the needs of all these different applications should be sufficient to ensure that most applications not listed will also be met.
3. Constraints Imposed by the Internet on the Codec

Packet losses are inevitable on the Internet and dealing with those is one of the most fundamental requirements for an Internet audio codec. While any audio codec can be combined with a good packet loss concealment (PLC) algorithm, the important aspect is what happens on the first packets received _after_ the loss. More specifically, this means that:

- it should be possible to interpret the contents of any received packet, irrespective of previous losses as specified in BCP 36 [PAYLOADS]; and
- the decoder should re-synchronize as quickly as possible (i.e. the output should quickly converge to the output that would have been obtained if no-loss had occurred).

The constraint of being able to decode any packet implies the following considerations for an audio codec:

- The size of a compressed frame must be kept smaller than the MTU to avoid fragmentation;
- The interpretation of any parameter encoded in the bit-stream must not depend on information contained in other packets. For example, it is not acceptable for a codec to allow signaling a mode change in one packet and assume that subsequent frames will be decoded according to that mode.

Although the interpretation of parameters cannot depend on other packets, it is still reasonable to use some amount of prediction across frames, provided that the predictors can resynchronize quickly in case of a lost packet. In this case, it is important to use the best compromise between the gain in coding efficiency and the loss in packet loss robustness due to the use of inter-frame prediction. It is a desirable property for the codec to allow some real-time control of that trade-off so that it can take advantage of more prediction when the loss rate is small, while being more robust to losses when the loss rate is high.

To improve the robustness to packet loss, it would be desirable for the codec to allow an adaptive (data- and network-dependent) amount of side information to help improve audio quality when losses occur. For example, this side information may include the retransmission of certain parameters encoded in the previous frame(s).

Another important property of the Internet is that it is mostly a best-effort network, with no guaranteed bandwidth. This means that
the codec has to be able to vary its output bit-rate dynamically (in real-time), without requiring an out-of-band signaling mechanism, and without causing audible artifacts at the bit-rate change boundaries. Additional desirable features are:

- Having the possibility to use smooth bit-rate changes with one byte/frame resolution;
- Making it possible for a codec to adapt its bit-rate based on the source signal being encoded (source-controlled VBR) to maximize the quality for a certain average bit-rate.

Because the Internet transmits data in bytes, a codec should produce compressed data in integer numbers of bytes. In general, the codec design should take into consideration explicit congestion notification (ECN) and may include features that would improve the quality of an ECN implementation.

The IETF has defined a set of application-layer protocols to be used for transmitting real-time transport of multimedia data, including voice. It is thus important for the resulting codec to be easy to use with these protocols. For example, it must be possible to create an [RTP] payload format that conforms to BCP 36 [PAYLOADS]. If any codec parameters need to be negotiated between end-points, the negotiation should be as easy as possible to carry over SIP/SDP or alternatively over XMPP/Jingle.

3.1. Security

Just like for any protocol to be used over the Internet, security is a very important aspect to consider. This goes beyond the obvious considerations of preventing buffer overflows and similar attacks that can lead to denial-of-service or remote code execution. One very important security aspect is to make sure that the decoders have a bounded and reasonable worst-case complexity. This prevents an attacker from causing a DoS by sending packets that are specially crafted to take a very long (or infinite) time to decode.

A more subtle aspect is the information leak that can occur when the codec is used over an encrypted channel (e.g. [SRTP]). For example, it was suggested [wright08] that use of source-controlled VBR may reveal some information about a conversation through the size of the compressed packets. For that reason, it should be possible to use the codec at truly constant bit-rate if needed.
4. Detailed Basic Requirements

This section summarizes all the constraints imposed by the target applications and by the Internet into a set of actual requirements for codec development.

4.1. Operating space

The operating space for the target applications can be divided in terms of delay: most applications require a "medium delay" (20-30 ms), while a few require a "very low delay" (< 10 ms). It makes sense to divide the space based on delay because lowering the delay has a cost in terms of quality vs bit-rate.

For medium delay, the resulting codec must be able to efficiently operate within the following range of bit-rates (per channel):

- Narrowband: 8 kb/s to 16 kb/s
- Wideband: 12 to 32 kb/s
- Super-wideband: 24 to 64 kb/s
- Full-band: 32 to 80 kb/s

Obviously, a lower-delay codec that can operate in the above range is also acceptable.

For very low delay, the resulting codec will need to operate within the following range of bit-rates (per channel):

- Super-wideband: 32 to 80 kb/s
- Full-band: 48 to 128 kb/s
- (Narrowband and wideband not required)

4.2. Quality and bit-rate

The quality of a codec is directly linked to the bit-rate, so these two must be considered jointly. When comparing the bit-rate of codecs, the overhead of IP/UDP/RTP headers should not be considered, but any additional bits required in the RTP payload format after the header (e.g. required signalling) should be considered. In terms of quality vs bit-rate, the codec to be developed must be better than the currently available codecs that satisfy the IPR requirements in the guidelines document, which are:
The codecs marked with (*) do not meet all the licensing guidelines, but the codecs to be developed should still not perform significantly worse. Quality should be measured for multiple languages, including tonal languages. The case of multiple simultaneous voices (as sometimes happens in conferencing) should be evaluated as well.

The comparison with the above codecs assumes that the codecs being compared have similar delay characteristics. The bit-rate required for a certain level of quality may be higher than the referenced codecs in cases where a much lower delay is required. In that case, the increase in bit-rate must be less than the ratio between the delays.

It is desirable for the codecs to support source-controlled variable bit-rate (VBR) to take advantage from the fact that different inputs require a different bitrate to achieve the same quality. However, it should still be possible to use the codec at truly constant bit-rate to ensure that no information leak is possible when using an encrypted channel.

4.3. Packet loss robustness

Robustness to packet loss is a very important aspect of any codec to be used on the Internet. Codecs must maintain acceptable quality at loss rates up to 5% and maintain good intelligibility up to 15% loss rate. At any sampling rate, bit-rate, and packet loss rate, the quality must be no less than the quality obtained with the Speex codec or the GSM-FR codec in the same conditions. The actual packet loss "patterns" to be used in testing must be obtained from real packet loss traces collected on the Internet, rather than from loss models. These traces should be representative of the typical environments in which the applications of Section 2 operate. For example, traces related to VoIP calls should consider the loss patterns observed for typical home broadband and corporate connections.

4.4. Computational resources

The resulting codec should be implementable on a wide range of devices, so there should be a fixed-point implementation or at least assurance that a reasonable fixed-point is possible. The computational resources figures listed below are meant to be upper
bounds. Even below these bounds, resources should still be minimized. Any proposed increase in computational resources consumption (e.g. to increase quality) should be carefully evaluated even if the resulting resource consumption is below the upper bound. Having variable complexity would be useful (but not required) in achieving that goal as it would allow trading quality/bit-rate for lower complexity.

The computational requirements for real-time encoding and decoding are:

- Narrowband should require little CPU resources and be implementable on most DSPs with a 16x16 multiplier (e.g. < 40 MIPS).
- Wideband can have a bit more complexity than narrowband, but should still be implementable on a cheap DSP (e.g. < 80 MIPS)
- Super-wideband/full-band may require higher complexity, but should be implementable on higher-end DSP (e.g. < 200 MIPS), and if possible also on cheaper DSPs as well.

The MIPS values are approximate clock frequencies required for real-time encoding+decoding on a DSP capable of single-cycle MAC operations (16x16 multiplication accumulated into 32 bits). Similar computational requirements apply to floating-point processors. For example Narrowband encoding and decoding should be possible using 40 MHz on a modern CPU (e.g. 2% of a 2 GHz x86 CPU). For applications that require mixing (e.g. conferencing), it should be possible to estimate the energy and/or the voice activity status of the decoded signal with less than 10% of the complexity figures listed above.

It is the intent to maximize the range of devices on which a codec can be implemented. For this reasons, the reference implementation must not depend on special hardware features or instructions to be present in order to meet the complexity requirement. However, it may be desirable to take advantage of such hardware when available, (e.g., hardware accelerators for operations like FFTs and convolutions). A codec should also minimize the use of saturating arithmetic so as to be implementable on architectures that do not provide hardware saturation (e.g. ARMv4).

The combined codec size and data ROM should be small enough not to cause significant implementation problems on typical embedded devices. The codec context/state size required should be no more than \(2^R*C\) bytes in floating-point, where \(R\) is the sampling rate and \(C\) is the number of channels. For fixed-point, that size should be less than \(R*C\). The scratch space required should also be less than
2*R*C bytes for floating point or less than R*C bytes for fixed-point.
5. Additional considerations

There are additional features or characteristics that may be desirable under some circumstances, but should not be part of the strict requirements. The benefit of meeting these considerations should be weighted against the associated cost.

5.1. Low-complexity audio mixing

In many applications that require a mixing server (e.g. conferencing, games), it is important to minimize the computational cost of the mixing. As much as possible, it should be possible to perform the mixing with fewer computations than it would take to decode all the streams, mix them, and re-encode the result. Properties that reduce the complexity of the mixing process include:

- the ability to derive sufficient parameters, such as loudness and/or spectral envelope, for estimating voice activity of a compressed frame without fully decoding that frame;
- the ability to mix the streams in an intermediate representation (e.g. transform domain), rather than having to fully decode the signals before the mixing;
- the use of bit-stream layers (Section 5.3) by aggregating a small number of active streams at lower quality.

For conferencing applications, the total complexity of the decoding, VAD and mixing should be considered when evaluating proposals.

5.2. Encoder side potential for improvement

In many codecs, it is possible to improve the quality by improving the encoder without breaking compatibility (i.e. without changing the decoder). Potential for improvement varies from one codec to another. It is generally low for PCM or ADPCM codecs and higher for perceptual transform codecs. All things being equal, being able to improve a codec after the bit-stream is a desirable property. However, this should not be done at the expense of quality in the reference encoder.

5.3. Layered bit-stream

A layered codec makes it possible to transmit only a certain subset of the bits and still obtain a valid bit-stream with a quality that is equivalent to the quality that would be obtained from encoding at the corresponding rate. While this is not a necessary feature for most applications, it can be desirable for cases where a "mixing
server" needs to handle a large number of streams with limited computational resources.

5.4. Partial redundancy

One possible way of increasing robustness to packet loss is to include partial redundancy within packets. This can be achieved either by including the base layer of the previous frame (for a layered codec) or by transmitting other parameters from the previous frame(s) to assist the PLC algorithm in case of loss. The ability to include partial redundancy for high-loss scenarios is desirable, provided that the feature can be dynamically turned on or off (so that no bandwidth is wasted in case of loss-free transmission).

5.5. Stereo support

It is highly desirable for the codec to have stereo support. At a minimum, the codec should be able to encode two channels independently without causing significant stereo image artefacts. It is also desirable for the codec to take advantage of the inter-channel redundancy in stereo audio to reduce the bitrate (for an equivalent quality) of stereo audio compared to coding channels independently.

5.6. Bit error robustness

The vast majority of Internet-based applications do not need to be robust to bit errors because packets either arrive unaltered, or do not arrive at all. Considering that, the emphasis should be on packet loss robustness and packet loss concealment. That being said, it is often the case that extra robustness to bit errors can be achieved at no cost at all (i.e. no increase in size, complexity or bit-rate, no decrease in quality or packet loss robustness, ...). In those cases then it is useful to make a change that increases the robustness to bit errors. This can be useful for applications that use UDP Lite transmission (e.g. over a wireless LAN). Robustness to packet loss should *never* be sacrificed to achieve higher bit error robustness.

5.7. Time stretching and shortening

When adaptive jitter buffers are used it is often necessary to stretch or shorten the audio signal to allow changes in buffering. While this operation can be performed directly on the decoder’s output, it is often more computationally efficient to stretch or shorten the signal directly within the decoder. It is desirable for the reference implementation to provide a time stretching/shortening implementation, although it should not be normative.
5.8. Input robustness

The systems providing input to the encoder and receiving output from the decoder may be far from ideal in actual use. Input and output audio streams may be corrupted by compounding non-linear artifacts from analog hardware and digital processing. The codecs to be developed should be tested to ensure that they degrade gracefully under adverse audio input conditions. Types of digital corruption that may be tested include tandeming, transcoding, low-quality resampling, and digital clipping. Types of analog corruption that may be tested include microphones with substantial background noise, analog clipping, and loudspeaker distortion. No specific end-to-end quality requirements are mandated for use with the proposed codec. It is advisable, however, that several typical in-situ environments/processing chains be specified for the purpose of benchmarking end-to-end quality with the proposed codec.

5.9. Legacy compatibility

In order to create the best possible codec for the Internet, there is no requirement for compatibility with legacy Internet codecs.
6. Security Considerations

The codec requirements themselves do not have security considerations. However, codec security issues are discussed in Section 3.1.
7. IANA Considerations

This document has no actions for IANA.
8. Acknowledgments

The original authors of this document are: Jean-Marc Valin, Slava Borilin, Koen Vos, Christopher Montgomery and Raymond (Juin-Hwey) Chen. We would like to thank all the other people who contributed directly or indirectly to this document, including Jason Fischl, Gregory Maxwell, Alan Duric, Jonathan Christensen, Julian Spittka, Michael Knappe, Christian Hoene, and Henry Sinnreich. We also like to thank Cullen Jennings and Gregory Lebovitz for their advice.
9. Informative References


Authors’ Addresses

Jean-Marc Valin
Octasic Inc.
4101, Molson Street
Montreal, Quebec
Canada

Email: jean-marc.valin@octasic.com

Koen Vos
Skype Technologies S.A.
Stadsgarden 6
Stockholm, 11645
Sweden

Email: koen.vos@skype.net
Abstract

This document provides specific requirements for an Internet audio codec. These requirements address quality, sampling rate, bit-rate, and packet loss robustness, as well as other desirable properties.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 28, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
# Table of Contents

1. Introduction .................................................. 3
2. Definitions .................................................... 4
3. Applications ................................................... 5
   3.1. Point to point calls ...................................... 5
   3.2. Conferencing .............................................. 5
   3.3. Telepresence ............................................... 6
   3.4. Teleoperation and Remote Software Services ............... 6
   3.5. In-game voice chat ........................................ 7
   3.6. Live distributed music performances / Internet music
        lessons .................................................... 7
   3.7. Delay Tolerant Networking or Push-to-Talk Services ....... 8
   3.8. Other applications ........................................ 8
4. Constraints Imposed by the Internet on the Codec ............... 9
5. Detailed Basic Requirements .................................... 11
   5.1. Operating space ............................................ 11
   5.2. Quality and bit-rate ...................................... 11
   5.3. Packet loss robustness ................................... 12
   5.4. Computational resources ................................... 13
6. Additional considerations ....................................... 15
   6.1. Low-complexity audio mixing ............................... 15
   6.2. Encoder side potential for improvement ................... 15
   6.3. Layered bit-stream ........................................ 15
   6.4. Partial redundancy ....................................... 16
   6.5. Stereo support ............................................ 16
   6.6. Bit error robustness ...................................... 16
   6.7. Time stretching and shortening ............................ 16
   6.8. Input robustness .......................................... 17
   6.9. Support of Audio forensics ............................... 17
   6.10. Legacy compatibility ..................................... 17
7. Security Considerations ....................................... 18
8. IANA Considerations ........................................... 19
9. Acknowledgments ............................................... 20
10. Informative References ....................................... 21
Authors’ Addresses ............................................... 23
1. Introduction

This document provides requirements for an audio codec designed specifically for use over the Internet. The requirements attempt to address the needs of the most common Internet interactive audio transmission applications and to ensure good quality when operating in conditions that are typical for the Internet. These requirements address the quality, sampling rate, delay, bit-rate, and packet loss robustness. Other desirable codec properties are considered as well.
2. Definitions

Throughout this document, we will use the following conventions when referring to the sampling rate of a signal:

Narrowband: 8 kHz
Wideband: 16 kHz
Super-wideband: 24/32 kHz
Full-band: 44.1/48 kHz

Codec bit-rates in bits per second (b/s) will be considered without counting any overhead (IP/UDP/RTP headers, padding, ...). The codec delay is the total algorithmic delay when one adds the codec frame size to the "look-ahead". It is thus the minimum theoretically achievable end-to-end delay of a transmission system that uses the codec.
3. Applications

The following applications should be considered for Internet audio codecs, along with their requirements:

- Point to point calls
- Conferencing
- Telepresence
- Teleoperation
- In-game voice chat
- Live distributed music performances / Internet music lessons
- Delay Tolerant Networking or Push-to-Talk Services
- Other applications

3.1. Point to point calls

Point to point calls are voice over IP (VoIP) calls from two "standard" (fixed or mobile) phones, and implemented in hardware or software. For these applications, a wideband codec is required, along with narrowband support for compatibility with legacy telephony equipment (PSTN). It is expected for the range of useful bit-rates to be 12 – 32 kb/s for wideband speech and 8 – 16 kb/s for narrowband speech. The codec delay must be less than 40 ms, but no more than 25 ms is desirable. Support for encoding music is not required, but it is desirable for the codec not to make background (on-hold) music excessively unpleasant to hear. Also, the codec should be robust to noise (produce intelligible speech and no annoying artifacts) even at lower bit-rates.

3.2. Conferencing

Conferencing applications (which support multi-party calls) have additional requirements on top of the requirements for point-to-point calls. Conferencing systems often have higher-fidelity audio equipment and have greater network bandwidth available -- especially when video transmission is involved. For that reason, support for super-wideband audio becomes important, with useful bit-rates in the 32 – 64 kb/s range. The ability to vary the bit-rate (VBR) according to the "difficulty" of the audio signal is a desirable feature for the codec. This not only saves bandwidth "on average", but it can also help conference servers make more efficient use of the available
bandwidth by using more bandwidth for important audio streams and less bandwidth for less important ones (e.g. background noise).

Conferencing end-points often operate in hands-free conditions, which creates acoustic echo problems. For this reason lower delay is important, as it reduces the quality degradation due to any residual echo after acoustic echo cancellation (AEC). For this reason, the codec delay must be less than 30 ms for this application. An optional low-delay mode with less than 10 ms delay is desirable, but not required.

Most conferencing systems operate with a bridge that mixes some (or all) of the audio streams and sends them back to all the participants. In that case, it is important that the codec not produce annoying artefacts when two voices are present at the same time. Also, this mixing operation should be as easy as possible to perform. To make it easier to determine which streams have to be mixed (and which are noise/silence), it must be possible to measure (or estimate) the voice activity in a packet without having to fully decode the packet (saving most of the complexity when the packet need not be decoded). Also, the ability to save on the computational complexity when mixing is also desirable, but not required. For example, a transform codec may make it possible to mix the streams in the transform domain, without having to go back to time-domain. Low-complexity up-sampling and down-sampling within the codec is also a desirable feature when mixing streams with different sampling rates.

3.3. Telepresence

Most telepresence applications can be considered to be essentially very high-quality video-conferencing environments, so all of the conferencing requirements also apply to telepresence. In addition, telepresence applications require super-wideband and full-band audio capability with useful bit-rates in the 32 - 80 kb/s range. While voice is still the most important signal to be encoded, it must be possible to obtain good quality (even if not transparent) music.

Most telepresence applications require more than one audio channel, so support for stereo and multi-channel is important. While this can always be accomplished by encoding multiple single-channel streams, it is preferable to take advantage of the redundancy that exists between channels.

3.4. Teleoperation and Remote Software Services

Teleoperation applications are similar to telepresence, with the exception that they involve remote physical interactions. For example, the user may be controlling a robot while receiving real-
time audio feedback from that robot. For these applications, the delay has to be less than 10 ms. The other requirements of telepresence (quality, bit-rate, multi-channel) apply to teleoperation as well. The only exception is that mixing is not an important issue for teleoperation.

The requirements for remote software services are similar to those of teleoperation. These applications include remote desktop applications, remote virtualization, and interactive media application being rendered remotely (e.g. video games rendered on central servers). For all these applications, full-band audio with an algorithmic delay below 10 ms are important.

3.5. In-game voice chat

An increasing number of computer/console games make use of VoIP to allow players to communicate in real-time. The requirements for gaming are similar to those of conferencing, with the main difference being that narrowband compatibility is not necessary. While for most applications a codec delay up to 30 ms is acceptable, a low-delay (<10 ms) option is highly desirable, especially for games with rapid interactions. The ability to use VBR (with a maximum allowed bitrate) is also highly desirable because it can significantly reduce the bandwidth requirement for a game server.

3.6. Live distributed music performances / Internet music lessons

Live music over the Internet requires extremely low end-to-end delay and is one of the most demanding application for interactive audio transmission. It has been observed that for most scenarios, total end-to-end delays up to 25 ms could be tolerated by musicians, with the absolute limit (where none of the scenarios are possible) being around 50 ms [carot09]. In order to achieve this low delay on the Internet -- either in the same city or a nearby city -- the network propagation time must be taken into account. When also subtracting the delay of the audio buffer, jitter buffer, and acoustic path, that leaves around 2 ms to 10 ms for the total delay of the codec. Considering the speed of light in fiber, every 1 ms reduction in the codec delay increases the range over which synchronization is possible by approximately 200 km.

Acoustic echo is expected to be an even more important issue for network music than it is in conferencing, especially considering that the music quality requirements essentially forbid the use of a "nonlinear processor" (NLP) with the AEC. This is another reason why very low delay is essential.

Considering that the application is music, the full audio bandwidth
(44.1 or 48 kHz sampling rate) must be transmitted with a bit-rate that is sufficient to provide near-transparent to transparent quality. With the current audio coding technology, this corresponds to approximately 64 kb/s to 128 kb/s per channel. As for telepresence, support for two or more channels is often desired, so it would be useful for a codec to be able to take advantage of the redundancy that is often present between audio channels.

3.7. Delay Tolerant Networking or Push-to-Talk Services

Internet transmissions are subjected to interruptions of connectivity that severely disturb a phone call. This may happen in cases of route changes, handovers, slow fading, or device failures. To overcome this distortion, the phone call can be halted and resumed after the connectivity has been reestablished again.

Also, if transmission capacity is lower than the minimal coding rate, switching to a push-to-talk mode still allows for effective communication. In that situation, voice is transmitted at slower-than-real-time bitrate and conversations are interrupted until the speech has been transmitted.

These modes require interrupting the audio playout and continuing after a pause of arbitrary duration.

3.8. Other applications

The above list is by no means a complete list of all applications involving interactive audio transmission on the Internet. However, it is believed that meeting the needs of all these different applications should be sufficient to ensure that most applications not listed will also be met.
4. Constraints Imposed by the Internet on the Codec

Packet losses are inevitable on the Internet and dealing with those is one of the most fundamental requirements for an Internet audio codec. While any audio codec can be combined with a good packet loss concealment (PLC) algorithm, the important aspect is what happens on the first packets received _after_ the loss. More specifically, this means that:

- it should be possible to interpret the contents of any received packet, irrespective of previous losses as specified in BCP 36 [PAYLOADS]; and
- the decoder should re-synchronize as quickly as possible (i.e. the output should quickly converge to the output that would have been obtained if no-loss had occurred).

The constraint of being able to decode any packet implies the following considerations for an audio codec:

- The size of a compressed frame must be kept smaller than the MTU to avoid fragmentation;
- The interpretation of any parameter encoded in the bit-stream must not depend on information contained in other packets. For example, it is not acceptable for a codec to allow signaling a mode change in one packet and assume that subsequent frames will be decoded according to that mode.

Although the interpretation of parameters cannot depend on other packets, it is still reasonable to use some amount of prediction across frames, provided that the predictors can resynchronize quickly in case of a lost packet. In this case, it is important to use the best compromise between the gain in coding efficiency and the loss in packet loss robustness due to the use of inter-frame prediction. It is a desirable property for the codec to allow some real-time control of that trade-off so that it can take advantage of more prediction when the loss rate is small, while being more robust to losses when the loss rate is high.

To improve the robustness to packet loss, it would be desirable for the codec to allow an adaptive (data- and network-dependent) amount of side information to help improve audio quality when losses occur. For example, this side information may include the retransmission of certain parameters encoded in the previous frame(s).

To ensure freedom of implementation, decoder-side only error concealment does not need to be specified, although a functional PLC
algorithm is desirable as part of the codec reference implementation. Obviously, any information signaled in the bitstream intended to aid PLC needs to be specified.

Another important property of the Internet is that it is mostly a best-effort network, with no guaranteed bandwidth. This means that the codec has to be able to vary its output bit-rate dynamically (in real-time), without requiring an out-of-band signaling mechanism, and without causing audible artifacts at the bit-rate change boundaries. Additional desirable features are:

- Having the possibility to use smooth bit-rate changes with one byte/frame resolution;

- Making it possible for a codec to adapt its bit-rate based on the source signal being encoded (source-controlled VBR) to maximize the quality for a certain _average_ bit-rate.

Because the Internet transmits data in bytes, a codec should produce compressed data in integer numbers of bytes. In general, the codec design should take into consideration explicit congestion notification (ECN) and may include features that would improve the quality of an ECN implementation.

The IETF has defined a set of application-layer protocols to be used for transmitting real-time transport of multimedia data, including voice. It is thus important for the resulting codec to be easy to use with these protocols. For example, it must be possible to create an [RTP] payload format that conforms to BCP 36 [PAYLOADS]. If any codec parameters need to be negotiated between end-points, the negotiation should be as easy as possible to carry over SIP [RFC3261]/SDP [RFC4566] or alternatively over XMPP [RFC6120]/Jingle [XEP-0167].
5. Detailed Basic Requirements

This section summarizes all the constraints imposed by the target applications and by the Internet into a set of actual requirements for codec development.

5.1. Operating space

The operating space for the target applications can be divided in terms of delay: most applications require a "medium delay" (20-30 ms), while a few require a "very low delay" (< 10 ms). It makes sense to divide the space based on delay because lowering the delay has a cost in terms of quality vs bit-rate.

For medium delay, the resulting codec must be able to efficiently operate within the following range of bit-rates (per channel):

- Narrowband: 8 kb/s to 16 kb/s
- Wideband: 12 to 32 kb/s
- Super-wideband: 24 to 64 kb/s
- Full-band: 32 to 80 kb/s

Obviously, a lower-delay codec that can operate in the above range is also acceptable.

For very low delay, the resulting codec will need to operate within the following range of bit-rates (per channel):

- Super-wideband: 32 to 80 kb/s
- Full-band: 48 to 128 kb/s
- (Narrowband and wideband not required)

5.2. Quality and bit-rate

The quality of a codec is directly linked to the bit-rate, so these two must be considered jointly. When comparing the bit-rate of codecs, the overhead of IP/UDP/RTP headers should not be considered, but any additional bits required in the RTP payload format after the header (e.g. required signalling) should be considered. In terms of quality vs bit-rate, the codec to be developed must be better than the following codecs, that are generally considered as royalty-free:
o For narrowband: Speex (NB) [Speex], and iLBC(*) [RFC3951]

o For wideband: Speex (WB) [Speex], G.722.1(*) [ITU.G722.1]

o For super-wideband/fullband: G.722.1C(*) [ITU.G722.1]

The codecs marked with (*) have additional licensing restrictions, but the codec to be developed should still not perform significantly worse. In addition to the quality targets listed above, a desirable objective is for the codec quality to be no worse than AMB-NB and AMR-WB, for narrowband and wideband, respectively. Quality should be measured for multiple languages, including tonal languages. The case of multiple simultaneous voices (as sometimes happens in conferencing) should be evaluated as well.

The comparison with the above codecs assumes that the codecs being compared have similar delay characteristics. The bit-rate required for a certain level of quality may be higher than the referenced codecs in cases where a much lower delay is required. In that case, the increase in bit-rate must be less than the ratio between the delays.

It is desirable for the codecs to support source-controlled variable bit-rate (VBR) to take advantage from the fact that different inputs require a different bitrate to achieve the same quality. However, it should still be possible to use the codec at truly constant bit-rate to ensure that no information leak is possible when using an encrypted channel.

5.3. Packet loss robustness

Robustness to packet loss is a very important aspect of any codec to be used on the Internet. Codecs must maintain acceptable quality at loss rates up to 5% and maintain good intelligibility up to 15% loss rate. At any sampling rate, bit-rate, and packet loss rate, the quality must be no less than the quality obtained with the Speex codec or the GSM-FR codec in the same conditions. The actual packet loss "patterns" to be used in testing must be obtained from real packet loss traces collected on the Internet, rather than from loss models. These traces should be representative of the typical environments in which the applications of Section 3 operate. For example, traces related to VoIP calls should consider the loss patterns observed for typical home broadband and corporate connections.
5.4. Computational resources

The resulting codec should be implementable on a wide range of devices, so there should be a fixed-point implementation or at least assurance that a reasonable fixed-point is possible. The computational resources figures listed below are meant to be upper bounds. Even below these bounds, resources should still be minimized. Any proposed increase in computational resources consumption (e.g. to increase quality) should be carefully evaluated even if the resulting resource consumption is below the upper bound. Having variable complexity would be useful (but not required) in achieving that goal as it would allow trading quality/bit-rate for lower complexity.

The computational requirements for real-time encoding and decoding of a mono signal on one core of a recent x86 CPU (as measured with the unix "time" utility or equivalent) are as follows:

- Narrowband: 40 MHz (2% of a 2 GHz CPU core)
- Wideband: 80 MHz (4% of a 2 GHz CPU core)
- Superwideband/fullband: 200 MHz (10% of a 2 GHz CPU core)

It is a desirable objective that the MHz values listed above also be achievable on fixed-point digital signal processors that are capable of single-cycle multiply-accumulate operations (16x16 multiplication accumulated into 32 bits).

For applications that require mixing (e.g. conferencing), it should be possible to estimate the energy and/or the voice activity status of the decoded signal with less than 10% of the complexity figures listed above.

It is the intent to maximize the range of devices on which a codec can be implemented. For this reasons, the reference implementation must not depend on special hardware features or instructions to be present in order to meet the complexity requirement. However, it may be desirable to take advantage of such hardware when available, (e.g., hardware accelerators for operations like fast Fourier transforms and convolutions). A codec should also minimize the use of saturating arithmetic so as to be implementable on architectures that do not provide hardware saturation (e.g. ARMv4).

The combined codec size and data ROM should be small enough not to cause significant implementation problems on typical embedded devices. The codec context/state size required should be no more than 2*R*C bytes in floating-point, where R is the sampling rate and
C is the number of channels. For fixed-point, that size should be less than R*C. The scratch space required should also be less than 2*R*C bytes for floating point or less than R*C bytes for fixed-point.
6. Additional considerations

There are additional features or characteristics that may be desirable under some circumstances, but should not be part of the strict requirements. The benefit of meeting these considerations should be weighted against the associated cost.

6.1. Low-complexity audio mixing

In many applications that require a mixing server (e.g. conferencing, games), it is important to minimize the computational cost of the mixing. As much as possible, it should be possible to perform the mixing with fewer computations than it would take to decode all the streams, mix them, and re-encode the result. Properties that reduce the complexity of the mixing process include:

- the ability to derive sufficient parameters, such as loudness and/or spectral envelope, for estimating voice activity of a compressed frame without fully decoding that frame;

- the ability to mix the streams in an intermediate representation (e.g. transform domain), rather than having to fully decode the signals before the mixing;

- the use of bit-stream layers (Section 6.3) by aggregating a small number of active streams at lower quality.

For conferencing applications, the total complexity of the decoding, VAD and mixing should be considered when evaluating proposals.

6.2. Encoder side potential for improvement

In many codecs, it is possible to improve the quality by improving the encoder without breaking compatibility (i.e. without changing the decoder). Potential for improvement varies from one codec to another. It is generally low for PCM or ADPCM codecs and higher for perceptual transform codecs. All things being equal, being able to improve a codec after the bit-stream is a desirable property. However, this should not be done at the expense of quality in the reference encoder. Other potential improvements include signal-adaptive frame size selection and improved discontinuous transmission (DTX) algorithms that take advantage of predicting the decoder sides packet loss concealment (PLC) algorithms.

6.3. Layered bit-stream

A layered codec makes it possible to transmit only a certain subset of the bits and still obtain a valid bit-stream with a quality that
is equivalent to the quality that would be obtained from encoding at the corresponding rate. While this is not a necessary feature for most applications, it can be desirable for cases where a "mixing server" needs to handle a large number of streams with limited computational resources.

6.4. Partial redundancy

One possible way of increasing robustness to packet loss is to include partial redundancy within packets. This can be achieved either by including the base layer of the previous frame (for a layered codec) or by transmitting other parameters from the previous frame(s) to assist the PLC algorithm in case of loss. The ability to include partial redundancy for high-loss scenarios is desirable, provided that the feature can be dynamically turned on or off (so that no bandwidth is wasted in case of loss-free transmission).

6.5. Stereo support

It is highly desirable for the codec to have stereo support. At a minimum, the codec should be able to encode two channels independently without causing significant stereo image artefacts. It is also desirable for the codec to take advantage of the inter-channel redundancy in stereo audio to reduce the bitrate (for an equivalent quality) of stereo audio compared to coding channels independently.

6.6. Bit error robustness

The vast majority of Internet-based applications do not need to be robust to bit errors because packets either arrive unaltered, or do not arrive at all. Considering that, the emphasis should be on packet loss robustness and packet loss concealment. That being said, it is often the case that extra robustness to bit errors can be achieved at no cost at all (i.e. no increase in size, complexity or bit-rate, no decrease in quality or packet loss robustness, ...). In those cases then it is useful to make a change that increases the robustness to bit errors. This can be useful for applications that use UDP Lite transmission (e.g. over a wireless LAN). Robustness to packet loss should *never* be sacrificed to achieve higher bit error robustness.

6.7. Time stretching and shortening

When adaptive jitter buffers are used it is often necessary to stretch or shorten the audio signal to allow changes in buffering. While this operation can be performed directly on the decoder's output, it is often more computationally efficient to stretch or
shorten the signal directly within the decoder. It is desirable for the reference implementation to provide a time stretching/shortening implementation, although it should not be normative.

6.8. Input robustness

The systems providing input to the encoder and receiving output from the decoder may be far from ideal in actual use. Input and output audio streams may be corrupted by compounding non-linear artifacts from analog hardware and digital processing. The codecs to be developed should be tested to ensure that they degrade gracefully under adverse audio input conditions. Types of digital corruption that may be tested include tandeming, transcoding, low-quality resampling, and digital clipping. Types of analog corruption that may be tested include microphones with substantial background noise, analog clipping, and loudspeaker distortion. No specific end-to-end quality requirements are mandated for use with the proposed codec. It is advisable, however, that several typical in-situ environments/processing chains be specified for the purpose of benchmarking end-to-end quality with the proposed codec.

6.9. Support of Audio forensics

Emergency calls can be analyzed using audio forensics if the context and situation of the caller has to be identified. Thus, it is important to transmit not only the voice of the callees well but also to transmit background noise at high quality. In these situations, sounds or noises of low volume should also not be compressed or dropped. For this reason, the encoder must allow DTX to be disabled when required (e.g. for emergency calls).

6.10. Legacy compatibility

In order to create the best possible codec for the Internet, there is no requirement for compatibility with legacy Internet codecs.
7. Security Considerations

Although this document itself does not have security considerations, this section describes the security requirements for the codec.

Just like for any protocol to be used over the Internet, security is a very important aspect to consider. This goes beyond the obvious considerations of preventing buffer overflows and similar attacks that can lead to denial-of-service or remote code execution. One very important security aspect is to make sure that the decoders have a bounded and reasonable worst-case complexity. This prevents an attacker from causing a DoS by sending packets that are specially crafted to take a very long (or infinite) time to decode.

A more subtle aspect is the information leak that can occur when the codec is used over an encrypted channel (e.g. [SRTP]). For example, it was suggested [wright08] [white11] that use of source-controlled VBR may reveal some information about a conversation through the size of the compressed packets. For that reason, it should be possible to use the codec at truly constant bit-rate if needed.
8. IANA Considerations

This document has no actions for IANA.
9. Acknowledgments

The original authors of this document are: Jean-Marc Valin, Slava Borilin, Koen Vos, Christopher Montgomery and Raymond (Juin-Hwey) Chen. We would like to thank all the other people who contributed directly or indirectly to this document, including Jason Fischl, Gregory Maxwell, Alan Duric, Jonathan Christensen, Julian Spittka, Michael Knappe, Christian Hoene, and Henry Sinnreich. We also like to thank Cullen Jennings and Gregory Lebovitz for their advice.
10. Informative References


[RFC3951] Andersen, S., Duric, A., Astrom, H., Hagen, R., Kleijn, W., and J. Linden, "Internet Low Bit Rate Codec (iLBC)", RFC 3951, December 2004.


Authors’ Addresses

Jean-Marc Valin
Mozilla
650 Castro Street
Mountain View, CA  94041
USA

Email: jmvalin@jmvalin.ca

Koen Vos
Skype Technologies S.A.
Stadsgarden 6
Stockholm,  11645
Sweden

Email: koen.vos@skype.net
Guidelines for the Codec Development Within the IETF
draft-valin-codec-guidelines-08

Abstract

This document provides general guidelines for work on developing and specifying a codec within the IETF. These guidelines cover the development process, evaluation, requirements conformance, and intellectual property issues.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 28, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction ............................................. 3
2. Development Process ..................................... 4
4. Requirements Conformance ................................. 8
5. Intellectual Property ..................................... 10
6. Relationship with Other SDOs ............................... 12
7. Security Considerations .................................. 14
8. IANA Considerations ...................................... 15
9. Acknowledgments .......................................... 16
10. References ................................................ 17
    10.1. Normative References ............................... 17
    10.2. Informative References ......................... 17
Authors’ Addresses ........................................... 19
1. Introduction

This document describes a suggested process for work at the IETF on standardization of a codec that is optimized for use in interactive Internet applications and that can be widely implemented and easily distributed among application developers, service operators, and end users.
2. Development Process

The process outlined here is intended to make the work on an audio codec within the IETF transparent, predictable, and well organized. Such work might involve development of a completely new codec, adaptation of an existing codec to meet the requirements, or integration between two or more existing codecs that results in an improved codec combining the best aspects of each codec. To enable such procedural transparency, the contributor of an existing codec must be willing to cede change control to the IETF and should have sufficient knowledge of the codec to assist in the work of adapting it or applying some of its technology to the development or improvement of other codecs. Furthermore, contributors need to be aware that any codec that results from work within the IETF is likely to be different from any existing codec that was contributed to the Internet Standards Process.

Work on codec development is expected to proceed as follows:

1. IETF participants will identify the requirements to be met by an Internet codec, in the form of an Internet-Draft.

2. Interested parties are encouraged to make contributions proposing existing or new codecs, or elements thereof, to the codec WG as long as these contributions are within the scope of the WG. Ideally, these contributions should be in the form of Internet Drafts, although other forms of contributions are also possible as discussed in [PROCESS] and in the IETF’s Note Well. As always, contributions to the IETF are subject, among other process oriented RFCs, to [PROCESS], [TRUST], and [IPR]. Considering the field of technology, IPR transparency may be particularly high on the priority list of many codec WG participants. Accordingly, contributors are specifically reminded of their IPR disclosure requirement, and all participants are reminded of the solicitation of the disclosure of third party IPR, both as codified in [IPR].

3. As contributions are received and discussed within the working group, the group should gain a clearer understanding of what is achievable within the design space. As a result, the authors of the requirements document should iteratively clarify and improve their document to reflect the emerging working group consensus. This is likely to involve collaboration with IETF working groups in other areas, such as collaboration with working groups in the Transport area to identify important aspects of packet transmission over the Internet and to understand the degree of rate adaptation desirable, and with working groups in the RAI area to ensure that information about and negotiation of the
codec can be easily represented at the signalling layer. In parallel with this work, interested parties should evaluate the contributions at a higher level to see which requirements might be met by each codec.

4. Once a sufficient number of proposals has been received, the interested parties will identify the strengths, weaknesses, and innovative aspects of the contributed codecs. This step will consider not only the codecs as a whole, but also key features of the individual algorithms (predictors, quantizers, transforms, etc.).

5. It is expected that none of the contributed codecs will meet all of the defined requirements. Therefore, it is expected that IETF participants will accept a _baseline codec_ as a WG item to facilitate the development process. This baseline codec will meet as many of the requirements as possible, but probably will need to be adjusted through an iterative development process in order to meet all of the requirements (or as many requirements as possible). The baseline codec might be one of the contributed codecs (especially if it is the only codec that meets most of the requirements), a combination of two or more of the contributed codecs, or an entirely new codec. None of the decisions taken at this step will be definitive. In particular, IETF participants will not provide a "rubber stamp" for any contributed codec.

6. IETF participants should then attempt to iteratively improve each component of the baseline codec reference implementation, where by "component" we mean individual algorithms such as predictors, transforms, quantizers, and entropy coders. The participants should proceed by trying new designs, applying ideas from the contributed codecs, evaluating "proof of concept" ideas, and using their expertise in codec development to improve the baseline codec. Any aspect of the baseline codec might be changed (even the fundamental principles of the codec) or the participants might start over entirely by scrapping the baseline codec and designing a completely new one. The overriding goal shall be to design a codec that will meet the requirements defined in the requirements document. Given the IETF's open standards process, any interested party will be able to contribute to this work, whether or not they submitted an Internet-Draft for one of the contributed codecs. The codec itself should be normatively specified with code in an Internet-Draft.

7. In parallel with work on the codec reference implementation, developers and other interested parties should perform evaluation of the codec as described under Section 3, IETF participants
should define (within the AVT Working Group) the codec’s payload format for use with the Real-time Transport Protocol [RTP], and application developers should start testing the codec by implementing it in code and deploying it in actual Internet applications to identify any potential problems.

8. Once IETF participants agree that the codec being developed meets the requirements, IETF participants can begin the task of characterizing the codec. The characterization process is described under Section 3.
3. Evaluation, Testing, and Characterization

Lab evaluation of the codec being developed should happen throughout the development process because it will help ensure that progress is being made toward fulfillment of the requirements. There are many ways in which continuous evaluation can be performed. For minor, uncontroversial changes to the codec it should usually be sufficient to use objective measurements (e.g., PESQ, PEAQ, and SegSNR) validated by informal subjective evaluation. For more complex changes (e.g., when psychoacoustic aspects are involved) or for controversial issues, internal testing should be performed. An example of internal testing would be to have individual participants rate the decoded samples using one of the established testing methodologies, such as ITU-R BS.1534 (MUSHRA).

Throughout the process, it will be important to make use of the Internet community at large for real-world distributed testing. This will enable many different people with different equipment and use cases to test the codec and report any problems they experience. In the same way, third-party software developers will be encouraged to integrate the codec (with a warning about the bit-stream not being final) and provide feedback on its performance in real-world use cases.

Characterization of the final codec must be based on the reference implementation only (and not on any "private implementation"). This can be performed by independent testing labs or, if this is not possible, using the testing labs of the organizations that contribute to the Internet Standards Process. Packet loss robustness should be evaluated using actual loss patterns collected from use over the Internet, rather than theoretical models. The goals of the characterization phase are to:

- ensure that the requirements have been fulfilled
- guide the IESG in its evaluation of the resulting work
- assist application developers in understanding whether the codec is suitable for a particular application

The exact methodology for the characterization phase is still subject to discussion within the working group.
4. Requirements Conformance

It is the responsibility of the working group to define criteria for evaluating conformance, including but not limited to comparison tools and test vectors. The following text provides suggestions for consideration by the working group:

1. Any codec specified by the IETF must include source code for a normative C89 implementation, documented in an Internet Draft destined for standards track RFC. This implementation will be used to verify conformance of an implementation. Although a text description of the algorithm should be provided, its use should be limited to helping the reader in understanding the source code. Should the description contradict the source code, the latter shall take precedence. For convenience, the source code may be provided in compressed form, with base64 encoding.

2. Because of the size of the codec’s source code, it is possible that even after publishing the RFC, bugs would be found from time to time. An errata of the RFC and its software description should be maintained, along with a public software repository containing the current reference implementation.

3. It is the intention of the group to allow the greatest possible choice of freedom in implementing the specification. Accordingly, the number of binding RFC2119 keywords is going to be the minimum still allowing for interoperable implementations. In practice this generally means that only the decoder needs to be normative, so that the encoder can improve over time. This also enables different tradeoffs between quality and complexity.

4. To reduce the risk of bias towards certain CPU/DSP architectures, ideally the decoder specification should not require "bit-exact" conformance with the reference implementation. The output of a decoder implementation should only be "close enough" to the output of the reference decoder. A comparison tool should be provided along with the codec to verify objectively that the output of a decoder is likely to be perceptually indistinguishable from that of the reference decoder. However, an implementation may still wish to produce an output that is bit-exact with the reference implementation to simplify the testing procedure.

5. To ensure freedom of implementation, decoder-side only error concealment does not need to be specified, although the reference implementation should include the same PLC algorithm as used in the testing phase. Is it up to the working group to decide whether minimum requirements on PLC quality will be required for
compliance with the specification. Obviously, any information signaled in the bitstream intended to aid PLC needs to be specified.

6. An encoder implementation should not be required to make use of all the "features" (tools) in the bit-stream definition. However, the codec specification may require that an encoder implementation be able to generate any possible bit-rate. Unless a particular "profile" is defined in the specification, the decoder must be able to decode all features of the bit-stream. The decoder must also be able to handle any combination of bits, even combinations that cannot be generated by the reference encoder. It is recommended that the decoder specification shall define exactly how the decoder should react to "impossible" packets. However, an encoder must never generate such packets that do not conform to the bit-stream definition.

7. Compressed test vectors should be provided as a means to verify conformance with the decoder specification. These test vectors should exercise all paths in the decoder (100% code coverage).

8. While the exact encoder will not be specified, it is recommended to specify objective measurement targets for an encoder, below which use of a particular encoder implementation is not recommended. For example, one such specification could be: "the use of an encoder whose PESQ MOS is less than 0.1 below the reference encoder in the following conditions is not recommended".
5. Intellectual Property

Producing an unencumbered codec is desirable for the following reasons:

- It is the experience of a wide variety of application developers and service providers that encumbrances such as licensing and royalties make it difficult to implement, deploy, and distribute audio applications for use by the Internet community.

- It is beneficial to have low-cost options whenever possible because standalone voice services are being commoditized and small, innovative development teams often cannot afford to pay per-channel licensing fees and royalties.

- Many market segments are moving away from selling hard-coded hardware devices and toward freely distributing end-user software; this is true of numerous large application providers and even telcos themselves.

- Compatibility with the licensing of typical open source applications implies the need to avoid encumbrances, including even the requirement to obtain a license for implementation, deployment, or use (even if the license does not require the payment of a fee).

Therefore, a codec that can be widely implemented and easily distributed among application developers, service operators, and end users is preferred. Many existing codecs that might fulfill some or most of the technical attributes listed above are encumbered in various ways. For example, patent holders might require that those wishing to implement the codec in software, deploy the codec in a service, or distribute the codec in software or hardware need to request a license, enter into a business agreement, pay licensing fees or royalties, or adhere to other special conditions or restrictions. Because such encumbrances have made it difficult to widely implement and easily distribute high-quality audio codecs across the entire Internet community, the working group prefers unencumbered technologies in a way that is consistent with BCP 78 and BCP 79. In particular, the working group shall heed the preference stated in BCP 79: "In general, IETF working groups prefer technologies with no known IPR claims or, for technologies with claims against them, an offer of royalty-free licensing." Although this preference cannot guarantee that the working group will produce an unencumbered codec, the working group shall follow BCP 79, and adhere to the spirit of BCP 79. The working group cannot explicitly rule out the possibility of adopting encumbered technologies; however, the working group will try to avoid encumbered technologies.
that require royalties or other encumbrances that would prevent such technologies from being easy to redistribute and use.

The following guidelines will help to maximize the odds that the codec will be unencumbered:

1. In accordance with BCP 79 [IPR], contributed codecs should preferably use technologies with no known IPR claims or technologies with an offer of royalty-free (RF) licensing.

2. Whenever possible, the working group should use technologies that are perceived by the participants to be safer with regard to IPR issues.

3. Contributors must disclose IPR as specified in BCP 79.

4. In cases where no RF license can be obtained regarding a patent, the group should consider alternative algorithms or methods, even if they result in lower quality, higher complexity, or otherwise less desirable characteristics (in most cases, the degradation will likely be small once the best alternative has been identified).

5. In accordance with BCP 78 [TRUST], the source code for the reference implementation must be made available under a BSD-style license (or whatever license is defined as acceptable by the IETF Trust when the Internet-Draft defining the reference implementation is published).

IETF participants should be aware that, given the way patents work in most countries, the resulting codec can never be guaranteed to be free of patent claims because some patents may not be known to the contributors, some patent applications may not be disclosed at the time the codec is developed, and only courts of law can determine the validity and breadth of patent claims. However, these observations are no different within the Internet Standards Process than they are for standardization of codecs within other SDOs (or development of codecs outside the context of any SDO), and furthermore are no different for codecs than for other technologies worked on within the IETF. In all these cases, the best approach is to minimize the risk of unknowingly incurring encumbrance on existing patents. Despite these precautions, participants need to understand that, practically speaking, it is nearly impossible to _guarantee_ that implementors will not incur encumbrance on existing patents.
6. Relationship with Other SDOs

It is understood that other SDOs are also involved in the codec development and standardization, including but not necessarily limited to:

- The Telecommunication Standardization Sector (ITU-T) of the International Telecommunication Union (ITU), in particular Study Group 16
- The Moving Picture Experts Group (MPEG)
- The European Telecommunications Standards Institute (ETSI)
- The 3rd Generation Partnership Project (3GPP)
- The 3rd Generation Partnership Project 2 (3GPP2)

It is important to ensure that such work does not constitute uncoordinated protocol development, of the kind described in [UNCOORD] in the following principle:

> The IAB considers an essential principle of the protocol development process that only one SDO maintains design authority for a given protocol, with that SDO having ultimate authority over the allocation of protocol parameter code-points; defining the intended semantics, interpretation, and actions associated with those code-points.

The work envisioned by this guidelines document is not "uncoordinated" in the sense described in the foregoing quote, for the following reasons:

- Internet signalling technologies are designed to enable the negotiation of any codecs that are supported in a particular application (such signalling technologies include the Session Initiation Protocol [SIP], Session Description Protocol [SDP], and the Extensible Messaging and Presence Protocol [XMPP] extensions for media negotiation as specified in [Jingle]).
- Internet transport technologies such as the Real-time Transport Protocol [RTP] (including secure transport as described in [SRTP]) are designed to support any codec for which RTP packetization rules have been defined.
- The IETF codec working group will focus on issues that are specific to the Internet, including robustness to packet loss and other aspects of packet transmission over the Internet. Issues
that are specific to non-Internet transports (e.g., radio communication and circuit-switched networks) are specifically out of scope.

Although there is already sufficient codec expertise available among IETF participants to complete the envisioned work, additional contributions are welcome within the framework of the Internet Standards Process, in the following ways:

- Individuals who are technical contributors to codec work within other SDOs can participate directly in codec work within the IETF.
- Other SDOs can contribute their expertise (e.g., codec characterization and evaluation techniques) and thus facilitate the testing of a codec produced by the IETF.
- Any SDO can provide input to IETF work through liaison statements.

However, it is important to note that final responsibility for the development process and the resulting codec will remain with the IETF as governed by BCP 9 [PROCESS].

Finally, there is precedent for the contribution of codecs developed elsewhere to the ITU-T (e.g., AMR Wideband was standardized originally within 3GPP). This is a model to explore as the IETF coordinates further with the ITU-T in accordance with the collaboration guidelines defined in [COLLAB].
7. Security Considerations

The procedural guidelines for codec development do not have security considerations. However, the resulting codec needs to take appropriate security considerations into account, for example as outlined in [DOS] and [SECGUIDE].
8. IANA Considerations

This document has no actions for IANA.
9. Acknowledgments

We would like to thank all the other people who contributed directly or indirectly to this document, including Jason Fischl, Gregory Maxwell, Alan Duric, Jonathan Christensen, Julian Spittka, Michael Knappe, Timothy Terriberry, Christian Hoene, Stephan Wenger and Henry Sinnreich. We also like to thank Cullen Jennings and Gregory Lebovitz for their advice. Special thanks to Peter Saint-Andre, who originally co-authored this document.
10. References

10.1. Normative References


10.2. Informative References


[UNCOORD] Bryant, S. and M. Morrow, "Uncoordinated Protocol
Development Considered Harmful", RFC 5704, November 2009.

Authors’ Addresses

Jean-Marc Valin
Octasic Inc.
4101, Molson Street
Montreal, Quebec
Canada
Email: jean-marc.valin@octasic.com

Slava Borilin
SPIRIT DSP
Email: borilin@spiritdsp.net

Koen Vos
Skype Technologies S.A.
Stadsgarden 6
Stockholm, 11645
Sweden
Email: koen.vos@skype.net

Christopher Montgomery
Xiph.Org Foundation
Email: xiphmont@xiph.org

Raymond (Juin-Hwey) Chen
Broadcom Corporation
Email: rchen@broadcom.com