draft-valin-celt-rtp-profile-02
RTP Payload Format for the CELT Codec

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on January 14, 2010.

Copyright Notice

Copyright (c) 2009 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 in effect on the date of publication of this document (http://trustee.ietf.org/license-info). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.
Abstract

CELT is an open-source voice codec suitable for use in very low delay audio communication applications, including Voice over IP (VoIP). This document describes the payload format for CELT generated bit streams within an RTP packet. Also included here are the necessary details for the use of CELT with the Session Description Protocol (SDP). At the time of this writing, the CELT bit-stream has NOT been finalized yet, and compatibility is usually broken with every new release of the codec.

Table of Contents

1. Conventions used in this document ......................... 3
2. Overview of the CELT Codec .......................... 4
3. RTP payload format for CELT .......................... 5
   3.1. RTP Header .................................. 5
   3.2. CELT payload .................................. 6
   3.3. Multiple CELT frames in a RTP packet .......... 7
   3.4. Multiple channels ............................. 8
4. MIME registration of CELT ............................ 10
5. SDP usage of CELT .................................. 12
   5.1. Multichannel Mapping .......................... 13
   5.2. Low-Overhead Mode ........................... 15
6. Congestion Control .................................. 17
7. Security Considerations ............................... 18
8. Acknowledgments ..................................... 19
9. References ........................................ 20
   9.1. Normative References .......................... 20
   9.2. Informative References ....................... 20
Authors’ Addresses ................................... 21
1. Conventions used in this document

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].
2. Overview of the CELT Codec

CELT stands for "Constrained Energy Lapped Transform". It applies some of the CELP principles, but does everything in the frequency domain, which removes some of the limitations of CELP. CELT is suitable for both speech and music and currently features:

- Ultra-low algorithmic delay (as low as 2 ms)
- Full audio bandwidth (up to 20 kHz audio bandwidth)
- Support for both voice and music
- Stereo support
- Packet loss concealment
- Constant bitrates from under 32 kbps to 128 kbps and above
- Free software/open-source
3. RTP payload format for CELT

For RTP based transportation of CELT encoded audio the standard RTP header [rfc3550] is followed by one or more payload data blocks. An optional padding terminator may also be used.

3.1. RTP Header

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|V=2|P|X|  CC   |M|     PT      |       sequence number         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           timestamp                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           synchronization source (SSRC) identifier            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|            contributing source (CSRC) identifiers             |
|                              ...                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The RTP header is defined in the RTP specification [rfc3550]. This section defines how fields in the RTP header are used.

Padding (P): 1 bit

If the padding bit is set, the packet contains one or more additional padding octets at the end which are not part of the payload. The last octet of the padding contains a count of how many padding octets should be ignored, including itself. Padding may be needed by some encryption algorithms with fixed block sizes or for carrying several RTP packets in a lower-layer protocol data unit.

Extension (X): 1 bit

If the extension, X, bit is set, the fixed header MUST be followed by exactly one header extension, with a format defined in Section 5.3.1. of [rfc3550].

Marker (M): 1 bit

The M bit MUST be set to zero in all packets. The receiver MUST ignore the M bit.

Payload Type (PT): 7 bits

Payload Type (PT): The assignment of an RTP payload type for this
packet format is outside the scope of this document; it is specified by the RTP profile under which this payload format is used, or signaled dynamically out-of-band (e.g., using SDP).

**Timestamp:** 32 bits

A timestamp representing the sampling time of the first sample of the first CELT frame in the RTP payload. The clock frequency MUST be set to the sample rate of the encoded audio data and is conveyed out-of-band (e.g., as an SDP parameter).

### 3.2. CELT payload

For the purposes of packetizing the bit stream in RTP, it is only necessary to consider the sequence of bits as output by the CELT encoder [celt-website], and present the same sequence to the decoder. The payload format described here maintains this sequence.

A typical CELT frame, encoded at a high bitrate, is approx. 128 octets and the total size of the CELT frames SHOULD be kept below the path MTU to prevent fragmentation. CELT frames MUST NOT be split across multiple RTP packets,

An RTP packet MAY contain CELT frames of the same bit rate or of varying bit rates, since the bitrate for the frames is explicitly conveyed in band with the signal. The encoding and decoding algorithm can change the bit rate at any frame boundary, with the bit rate change notification provided in-band. No out-of-band notification is required for the decoder to process changes in the bit rate sent by the encoder.

More than one frame may be encoded in the same RTP packet, and for the decoder it is not possible to determine the compressed size (bit-rate) of each encoded frame. Thus the compressed size information MUST be explicitly transmitted. There are two modes for conveying this information: either the compressed size(s) are encoded for each frame at the beginning of the RTP payload Section 3.3, or it is conveyed in the signaling and then fixed for the duration of the session (Low-Overhead Mode, Section 5.2). Note that the compressed frame size information must be present either way even if only single frames are transmitted per RTP packet.

It is RECOMMENDED that sampling rates 32000, 44100, or 48000 Hz be used for most applications, unless a specific reason exists -- such as requirements for a very specific packetization time. For example, 51200 Hz sampling may be useful to obtain a 5 ms packetization time with 256-sample frames. For compatibility reasons, the sender and receiver MUST support 48000 Hz sampling rate.
The CELT codec always produces an integer number of bytes and can produce any integer number of bytes, so no padding is ever required. Bitrate adjustment SHOULD be used instead of padding.

3.3. Multiple CELT frames in a RTP packet

The bitrate used by CELT is implicitly determined by the size of the compressed data. When more than one frame is encoded in the same packet, it is not possible to determine the size of each encoded frame, so the information MUST be explicitly encoded. If N frames are present in a packet, N compressed frame sizes need to be encoded at the beginning of the packet. Each size that is less than 255 bytes is encoded in one byte (unsigned 8-bit integer). For sizes greater or equal to 255, a 0xff byte is encoded, followed by the size-255. Multiple 0xff bytes are allowed if there are more than 510 bytes transmitted. The length is always the size of the CELT frame excluding the length byte itself. The payload MUST NOT be padded, except in accordance with the padding bit definition in the RTP header.

Below is an example of two CELT frames contained within one RTP packet.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|V=2|P|X|  CC   |M|     PT      |       sequence number         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           timestamp                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         synchronization source (SSRC) identifier              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         contributing source (CSRC) identifiers                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| length frame 1| length frame 2|          CELT frame 1... |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          (frame 1)                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          (frame 2)                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following is an example of C code that interprets the length bytes:

```c
Valin & Maxwell         Expires January 14, 2010                
[Page 7]```
int i, N, pos;
int sizes[MAX_FRAMES][channels];
unsigned int total_size;
total_size=0;
N = 0;
pos = 0;
while (total_size < payload_size) {
    for (i=0;i<channels;i++) {
        int s;
        int sum;
        sum = 0;
        do {
            s = payload[pos++];
            sum += s;
            total_size += s+1;
        } while (s == 255);
        sizes[N][i] = sum;
    }
    N++;
}

3.4. Multiple channels

CELT supports both mono streams and stereo streams. If more than two channels are desired, it is possible to use transmit multiple streams in the same packet. In this case, the number of streams S and the pairing must be agreed with out-of-band negotiation such as SDP. Each stream can be either mono or stereo, depending on whether the channels are assumed to be correlated. For example, a 5.1 surround could have the front-left and front-right channels in a stereo stream, the rear-left and rear-right channels in a separate stereo stream, while the center and low-frequency channels would be in separate mono streams. In that example, the RTP packet would be:
In the case where streams for multiple channels are used with multiple frames of the same streams per packet, then all streams for a certain timestamp are encoded before all streams for the following timestamp. In the case of the 5.1 example above with two frames per packet, the number of compressed length fields would be $S*N = 8$. 
4. MIME registration of CELT

Full definition of the MIME \[rfc2045\] type for CELT will be part of the Ogg Vorbis MIME type definition application \[rfc3534\].

MIME media type name: audio

MIME subtype: celt

Optional parameters:

Required parameters: to be included in the Ogg MIME specification.

Encoding considerations: none

Security Considerations:

See Section 6 of RFC 3047.

Interoperability considerations: none

Published specification:

Applications which use this media type: none

Additional information: none

Person & email address to contact for further information:

Jean-Marc Valin <jean-marc.valin@octasic.com>

Intended usage: COMMON

Author/Change controller:

Author: Jean-Marc Valin <jean-marc.valin@octasic.com>

Change controller: Jean-Marc Valin <jean-marc.valin@octasic.com>

Change controller: IETF AVT Working Group

This transport type signifies that the content is to be interpreted according to this document if the contents are transmitted over RTP. Should this transport type appear over a lossless streaming protocol such as TCP, the content encapsulation should be interpreted as an Ogg Stream in accordance with \[rfc3534\], with the exception that the content of the Ogg Stream may be assumed to be CELT audio and CELT
audio only.
5. SDP usage of CELT

When conveying information by SDP [rfc4566], the encoding name MUST be set to "CELT". The sampling frequency is typically between 32000 and 48000 Hz. Implementations MUST support 48000 Hz and SHOULD also support 44100 Hz.

The SDP parameters have the following interpretation with respect to CELT:

ptime: The desired packetization time. The sender SHOULD choose a number of frames per packet that corresponds to the smallest packetization time greater or equal to the specified ptime for the selected frame size. The default is 20 ms as specified in [rfc3551]

maxptime: The maximum packetization time desired. As specified in [rfc4566], if the maximum is lower than the smallest packetization time determined from the chosen frame size (as described above), then that packetization time SHOULD be used despite the maxptime value. The default is "no maximum".

CELT-specific parameters can be given via the "a=fmtp:" directive. Several parameters can be given in a single a=fmtp line provided that they are separated by a semi-colon. The following parameters are defined for use in this way:

bitrate: The desired bit-rate in kbit/s for the codec only (excluding headers and the length bytes). The value MUST be rounded to an integer number of bytes per frame. The round-to-nearest method is RECOMMENDED. The default bit-rate value is 64 kbit/s per channel.

frame-size: The frame size is the duration of each frame in samples and has to be even. The default is 480.

mapping: Optional string describing the multi-channel mapping.

The selected frame-size values MUST be even. For quality and complexity reasons, they SHOULD also be divisible by 8 and have a prime factorization which consists only of 2, 3, or 5 factors. For example, powers-of-two and values such as 160, 320, 240, and 480 are recommended. Implementations MUST support receiving and sending the default value of 480. Implementations SHOULD also support frame sizes of 256 and 512 since these are the ones that lead to the lowest complexity. When frame sizes that are powers-of-two are supported, they SHOULD be listed first in the offer and chosen over non-powers-of-two in the answer.
Care must be taken when setting the value of ptime: and bitrate: so that the RTP packet size does not exceed the path MTU.

An example of the media representation in SDP for offering a single channel of CELT at 48000 samples per second might be:

```
m=audio 8088 RTP/AVP 97
a=rtpmap:97 CELT/48000/1
```

Note that the RTP payload type code of 97 is defined in this media definition to be ‘mapped’ to the CELT codec at a 48 kHz sampling frequency using the ‘a=rtpmap’ line. Any number from 96 to 127 could have been chosen (the allowed range for dynamic types). If there is more than one channel being encoded the rtpmap MUST specify the channel count. When no channel count is included, the default is one channel.

The following example demonstrates the use of the a=fmtp: parameters:

```
m=audio 8008 RTP/AVP 97
a=ptime: 25
a=rtpmap:97 CELT/44100
a=fmtp:97 frame-size=512;bitrate=48
```

This examples illustrate an offerer that wishes to receive a CELT stream at 44100 Hz, by packing two 512-sample frames in each packet (less than 25 ms) at around 48 kbps (70 bytes per frame).

### 5.1. Multichannel Mapping

When more than two channels are used, a mapping parameter MUST be provided. The mapping parameter is defined as comma separated list of integers which specify the number of channels contained in each CELT stream, OPTIONALLY followed by a ‘/’ and a comma separated list of channel identifiers, then OPTIONALLY another ‘/’ and a string which provides an application specific elaboration on any speaker-feed definitions. The channels per stream entries MUST be either 1 or 2. The total number of channels is indicated by the sum of the channels per stream entries. The sum of the channel counts MUST be equal to the total number of channels.

Channel identifiers are short alphanumeric strings. Each identifier
MUST begin with a letter indicating the type of channel. ‘A’ MUST be used to indicate an ambisonic channel, ‘S’ to indicate a speaker-feed channel, or ‘O’ indicating other usage.

A channel identifier MAY be repeated, but the meaning of such repetition is application specific. Applications SHOULD attempt to utilize channel identifiers such that mixing all identical identifiers would produce a reasonable result.

Non-surround usage such as individual performer tracks, effect send, "order wire", or other administrative channels may be given application specific identifiers which MUST not conflict with the identifiers defined in this draft. These identifiers SHOULD begin with S if it would be sensible to include them in a mono-downmix, or O if it would be most sensible to exclude them from a mono-downmix. An example usage might be mapping=2,1,2,1,1/SLguitar,SRguitar,OheadsetG,SLkeyboard,SRkeyboard,OheadsetK,SMbass,OheadsetB"

Ambisonic channels MUST follow the Furse-Malham naming and weighing conventions for up to third order spherical[Ambisonic]. Higher order ambisonic support is application defined but MUST NOT reuse any of WXYZRSTUVKLMNOPQ for higher order components. For example, second order spherical ambisonics SHOULD use the mapping "mapping=1,1,1,1,1,1,1,1,1/AW,AX,AY,AZ,AR,AS,AT,AU,AV". Any set of Ambisonic channels MUST contain at least one "AW" channel.

Speaker-feed identifiers are named based on the intended speaker locations. "L", "R" for the left and right speakers, respectively, in conventional stereo or the front left and right in 4, 5, 5.1, or 7.1 channel surround. "LR", "RR" for the left and right rear speakers in 4,5 or 5.1 channel surround. "C" is used for a center channel, "MLFE" for a low frequency extension channel. "LS", "RS" for the side channels in 7.1 channel surround. Additional speaker-feeds are application specific but should not reuse the prior identifiers. For 5.1 surround in non-ambisonic form the mapping SHOULD be "mapping=2,2,1,1,L,R,LR,RR,C,MLFE/ITU-RBS.775-1". When only one or two channels are used, the mapping parameter MAY be omitted, in which case the default mapping is used. For one channel, the default is "mapping=1/C", while for two channels, the default is "mapping=2/L,R".

For example a stereo configuration might signal:

```
m=audio 8008 RTP/AVP 97
```
a=ptime: 5
a=rtpmap:97 CELT/44100/2
a=fmtp:97 frame-size=256

Which specifies a single two-channel CELT stream according to the default mapping.

5.2. Low-Overhead Mode

A low-overhead mode is defined to make more efficient use of bandwidth when transmitting CELT frames. In that mode none of the length values need to be transmitted. One the a=fmtp: parameter low-overhead: is defined and contains a single frame size, followed by a ‘/’, followed by a comma-separated list of the number of bytes per frame for each stream defined in the channel mapping. The number of frames per channel can thus be computed as the payload size divided by the sum of the bytes-per-frame values. The frame-size: parameter MUST not be specified and SHOULD be ignored if encountered in an SDP offer or answer. The bitrate: parameter MUST also be ignored since the low-overhead: parameter makes it redundant. When the low-overhead: parameter is specified, the length of each frame MUST NOT be encoded in the payload and the bit-rate MUST NOT be changed during the session.

For example a low-overhead 64 kbit/s mono stream could be signaled as:

m=audio 8008 RTP/AVP 97
a=ptime: 5
a=rtpmap:97 CELT/48000/1
a=fmtp:97 low-overhead=256/43

and a low-overhead 360 kbit/s 5.1 surround configuration could be signaled as:

m=audio 8008 RTP/AVP 97
a=ptime: 5
a=rtpmap:97 CELT/48000/6
a=fmtp:97 low-overhead=256/86,86,43,25;mapping=2,2,1,1/L,R,LR,RR,C,MLFE/ITU-RBS.775-1
In this last example, 4 bytes per packet would be saved. This corresponds to a 6 kbit/s reduction in the overhead, although the 60 kbit/s overhead of the IP, UDP and RTP headers is still present.
6. Congestion Control

CELT allows any bitrate, with a one byte per frame resolution, without any signaling requirement or overhead. Applications SHOULD utilize congestion control to regulate the transmitted bitrate. In some applications it may make sense to increase the packetization interval rather than decreasing the codec bitrate. Congestion control implementations should consider the users differential tolerance for high latency and low quality.
7. Security Considerations

RTP packets using the payload format defined in this specification are subject to the security considerations discussed in the RTP specification [rfc3550], and in any applicable RTP profile. The main security considerations for the RTP packet carrying the RTP payload format defined within this memo are confidentiality, integrity and source authenticity. Confidentiality is achieved by encryption of the RTP payload. Integrity of the RTP packets through suitable cryptographic integrity protection mechanism. Cryptographic system may also allow the authentication of the source of the payload. A suitable security mechanism for this RTP payload format should provide confidentiality, integrity protection and at least source authentication capable of determining if an RTP packet is from a member of the RTP session or not.

Note that the appropriate mechanism to provide security to RTP and payloads following this memo may vary. It is dependent on the application, the transport, and the signalling protocol employed. Therefore a single mechanism is not sufficient, although if suitable the usage of SRTP [rfc3711] is recommended. Other mechanism that may be used are IPsec [rfc4301] and TLS [rfc5246] (RTP over TCP), but also other alternatives may exist.

This RTP payload format and its media decoder do not exhibit any significant non-uniformity in the receiver-side computational complexity for packet processing, and thus are unlikely to pose a denial-of-service threat due to the receipt of pathological data. Nor does the RTP payload format contain any active content.

Because this format supports VBR operation small amounts of information about the transmitted audio may be leaked by a length preserving cryptographic transport. Accordingly, when CELT is used inside a secure transport the sender SHOULD restrict the use of VBR to congestion control purposes.

CELT implementations will typically exhibit tiny content-sensitive encoding time variances. Since transmission is usually triggered by an accurate hardware clock and the encoded data is typically transmitted as soon as encoding is complete this variance may result in a small amount of additional frame to frame jitter which could be measured by a third-party. Encrypted implementations SHOULD transit packets at fixed intervals to avoid the possible information leak.
8. Acknowledgments

The authors would also like to thank the following people for their input: Timothy B. Terriberry, Ben Schwartz, Alexander Carot, Thorvald Natvig, Brian West, Steve Underwood, and Anthony Minessale.
9. References

9.1. Normative References

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


9.2. Informative References


Authors’ Addresses

Jean-Marc Valin
Octasic Semiconductor
4101, Molson Street, suite 300
Montreal, Quebec  H1Y 3L1
Canada

Email: jean-marc.valin@octasic.com

Gregory Maxwell
Juniper Networks
2251 Corporate Park Drive, Suite 100
Herndon, VA  20171-1817
USA

Email: gmaxwell@juniper.net