

codec
Internet-Draft
Intended status: Informational
Expires: December 18, 2011

JM. Valin
Octasic Inc.
K. Vos
Skype Technologies S.A.
J. Skoglund
Google
June 16, 2011

Summary of Opus listening test results
draft-valin-codec-results-00

Abstract

This document describes and examines listening test results obtained for the Opus codec and how they relate to the requirements.

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 18, 2011.

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. Pre-Opus listening tests	4
2.1. SILK Dynastat listening test	4
2.2. SILK Deutsche Telekom test	4
2.3. SILK Nokia test	4
2.4. CELT 0.3.2 listening test	5
2.5. CELT 0.5.0 listening test	5
3. Opus listening tests on non-final bit-stream	6
3.1. First hybrid mode test	6
3.2. Broadcom stereo music test	6
4. Opus listening tests on final bit-stream	8
4.1. Google listening tests	8
4.1.1. Google narrowband listening test	8
4.1.2. Google wideband and fullband listening test	9
4.1.3. Google stereo music listening test	10
4.2. HydrogenAudio stereo music listening test	12
4.3. Nokia Interspeech 2011 listening test	12
5. In-the-field testing	13
6. Conclusion on the requirements	14
6.1. Comparison to Speex (narrowband)	14
6.2. Comparison to iLBC	14
6.3. Comparison to Speex (wideband)	14
6.4. Comparison to G.722.1	14
6.5. Comparison to G.722.1C	15
6.6. Comparison to AMR-NB	15
6.7. Comparison to AMR-WB	15
7. Security Considerations	16
8. IANA Considerations	17
9. Acknowledgments	18
10. Informative References	19
Authors' Addresses	20

1. Introduction

This document describes and examines listening test results obtained for the Opus codec. Some of the test results presented are based on older versions of the codec or on older versions of the SILK or CELT components. While they do not necessarily represent the exact quality of the current version, they are nonetheless useful for validating the technology used and as an indication of a lower bound on quality (based on the assumption that the codec has been improved since they were performed).

Throughout this document, all statements about one codec being better than or worse than another codec are based on 95% confidence. When no statistically significant difference can be shown with 95% confidence, then two codecs are said to be "tied".

In addition to the results summarized in this draft, Opus has been subjected to many informal subjective listening tests, as well as objective testing.

2. Pre-Opus listening tests

Several listening tests have been performed on the SILK and CELT codecs prior to them being merged as part of the Opus codec.

2.1. SILK Dynastat listening test

The original (pre-Opus) SILK codec was characterized in a Dynastat listening test [SILK-Dynastat]. The test included 32 conditions with 4 male and 4 female talkers. The test signals were wideband speech with and without office background noise at 15 dB SNR. Packet loss was tested at 2, 5, and 10% loss rates. The bitrates ranged from 8.85 kb/s to 64 kb/s. The codecs included in the test were SILK-WB, AMR-WB, Speex-WB and G.722 (which ran at 64 kb/s).

The results showed that for clean speech (1) SILK out-performs AMR-WB at all bit-rates except 8.85 kb/s (which was a tie); (2) SILK out-performs Speex at all bit-rates; and (3) SILK running at 18.25 kb/s and above out-performs G.722 at 64 kbps. For noisy speech, tested at 18.25 kb/s, SILK is tied with AMR-WB, and out-performs Speex. For 2, 5 and 10% packet loss, tested at 18.25 kb/s, SILK out-performs both AMR-WB and Speex in all conditions.

2.2. SILK Deutsche Telekom test

In 2010 Deutsche Telekom published results [Wustenhagen2010] of their evaluation of super-wideband speech and audio codecs. The test included the version of SILK submitted to the IETF. The results showed that for clean speech (item "speechsample") SILK was tied with AMR-WB and G.718, and out-performed Speex. For noisy speech (item "arbeit") SILK out-performed AMR-WB and G.718 at 12 and 24 kb/s, and Speex at all bitrates. At bitrates above 24 kb/s SILK and G.718 were tied.

2.3. SILK Nokia test

In 2010, Anssi Ramo from Nokia presented [Ramo2010] the results of a listening test focusing on open-source codecs at Interspeech 2010. The methodology used was a 9-scale ACR MOS test with clean and noisy speech samples.

It was noted in the test that:

"Especially at around 16 kbit/s or above Silk is better than AMR-WB at comparable bitrates. This is due to the fact that Silk wideband is critically sampled up to 8 kHz instead of ITU- T or 3GPP defined 7 kHz. This added bandwidth (from 7 to 8 kHz) shows up in the results favourable to Silk. It seems that Silk provides quite artifact free

voice quality for the whole 16- 24 kbit/s range with WB signals. At 32 and 40 kbit/s Silk is SWB and competes quite equally against G.718B or G.722.1C although having a slightly narrower bandwidth than the ITU-T standardized codecs."

2.4. CELT 0.3.2 listening test

The first listening tests conducted on CELT version 0.3.2 in 2009 and published in 2010 [valin2010] included AAC-LD (Apple), G.722.1C and MP3 (Lame). Two MUSHRA tests were conducted: a 48 kb/s test and a 64 kb/s test, both at a 44.1 kHz sampling rate. CELT was used with 256-sample frames (5.8 ms). All codecs used constant bit-rate (CBR). The algorithmic delay was 8.7 ms for CELT, 34.8 ms for AAC-LD, 40 ms for G.722.1C and more than 100 ms for MP3.

The 48 kb/s test included two clean speech samples (one male, one female) from the EBU SQAM database, four clean speech files (two male, two female) from the NTT multi-lingual speech database for telephony, and two music samples. In this test, CELT out-performed AAC-LD, G.722.1C and MP3.

The 64 kb/s test included two clean speech samples (one male, one female) from the EBU SQAM database, and six music files. In this test, AAC-LD out-performed CELT, but CELT out-performed both MP3 and G.722.1C (running at its highest rate of 48 kb/s).

2.5. CELT 0.5.0 listening test

Another CELT listening test was conducted in 2009 on version 0.5.0 and presented at EUSIPCO 2009 [valin2009]. In that test, CELT was compared to G.722.1C and to the Fraunhofer Ultra Low-Delay (ULD) codec on 9 audio samples: 2 clean speech samples and 7 music samples. At 64 kb/s with 5.3 ms frames, CELT clearly out-performed G.722.1C running at 48 kb/s with 20 ms frames. Also, at 96 kb/s and equal frame size (2.7 ms), CELT clearly out-performed the ULD codec.

3. Opus listening tests on non-final bit-stream

The following listening tests were conducted on the Opus codec on versions prior to the bit-stream freeze. While Opus has evolved since these tests were conducted, the results should be considered as a lower bound on the quality of the final codec.

3.1. First hybrid mode test

In July 2010, the Opus codec authors conducted a preliminary MUSHRA listening test to evaluate the quality of the recently created "hybrid" mode combining the SILK and CELT codecs. That test was conducted at 32 kb/s and compared the following codecs:

- o Opus hybrid mode (fullband)
- o G.719 (fullband)
- o CELT (fullband)
- o SILK (wideband)
- o BroadVoice32 (wideband)

The test material consisted of two English speech samples from the EBU SQAM (one male, one female) database and six speech samples (three male, three female) from the NTT multi-lingual speech database for telephony. Although only eight listeners participated to the test, the difference between the Opus hybrid mode and all other codecs was large enough to obtain 95% confidence that the Opus hybrid mode provided better quality than all other codecs tested. This test is of interest because it shows that the hybrid clearly out-performs the codecs that it combines (SILK and CELT). It also out-performs G.719, which is the only fullband interactive codec standardized by the ITU-T. These results were presented [Maastricht-78] at the 78th IETF meeting Maastricht.

3.2. Broadcom stereo music test

In December 2010, Broadcom conducted an ITU-R BS.1116-style subjective listening test comparing different configurations of the CELT-only mode of the IETF Opus codec along with MP3 and AAC-LC. The test included stereo 10 audio samples sampled at 44.1 kHz and distributed as follows:

- o 2 pure speech

- o 2 vocal
- o 2 solo instruments
- o 1 rock-and-roll
- o 1 pop
- o 1 classical orchestra
- o 1 jazz

A total of 17 listeners participated to the test. The results of the test are available on the testing slides presented at the Prague meeting [Prague-80]. Although at the time, Opus was not properly optimised for 44.1 kHz audio, the quality of the Opus codec at 96 kb/s with 22 ms frame was significantly better than MP3 and only slightly worse than AAC-LC. Even in ultra low-delay mode (5.4 ms), Opus still outperformed MP3. The test also confirmed the usefulness of the prefilter/postfilter contribution by Raymond Chen, showing that this contribution significantly improves quality for small frames (long frames were not tested with the prefilter/postfilter disabled).

4. Opus listening tests on final bit-stream

The following tests were performed on the Opus codec after the bit-stream was finalized.

4.1. Google listening tests

The tests followed the MUSHRA test methodology. Two anchors were used, one lowpass-filtered at 3.5 kHz and one lowpass-filtered at 7.0 kHz. Both trained and untrained listeners participated in the tests. The reference signals were manually normalized to the same subjective levels according to the experimenters' opinion. Experiments with automatic normalization with respect to both level and loudness (in Adobe Audition) did not result in signals having equal subjective loudness. The sample magnitude levels were kept lower than 2^{14} to provide headroom for possible amplification through the codecs. However, the normalization exercise was not repeated with the processed sequences as neither the experimenters nor any of the subjects (which included expert listeners) noticed any significant level differences between the conditions in the tests. The only post-processing performed was to remove noticeable delays in the MP3 files, as one could identify the MP3 samples when switching between conditions when the MP3 had the longer delay. The testing tool Step from ARL was used for tests and all listeners were instructed to carefully listen through the conditions before starting the grading. The results of the tests are available on the testing slides presented at the Prague meeting [Prague-80].

4.1.1. Google narrowband listening test

The test sequences in Test 1 were mono recordings (between 2 and 6 seconds long) of 4 different male and 4 different female speakers sampled at 48 kHz in low background noise. 17 listeners were presented with 6 stimuli according to Table 1 for each test sequence. The corresponding bit rate for the reference is $48000 \text{ (sampling frequency in Hz)} \times 16 \text{ (bits/sample)} = 768 \text{ kbps}$. Since the anchors are low-pass filtered they can also be downsampled for transmission which corresponds to lower bit rates. Three narrowband codecs were compared in this test: Opus NB, the royalty-free iLBC, and the royalty-free Speex. The codecs all have an encoder frame length of 20 ms. Both Opus and Speex had variable rate whereas iLBC operated at a fixed bit rate.

Type	Signal bandwidth	Bitrate
Reference	24 kHz (Fullband)	
Anchor 1	3.5 kHz (Narrowband)	
Anchor 2	7 kHz (Wideband)	
iLBC	4 kHz (Narrowband)	15.2 kbps, CBR
Opus NB	4 kHz (Narrowband)	11 kbps, VBR
Speex NB	3.5 kHz (Narrowband)	11 kbps, VBR

Test 1 stimuli

Table 1

The overall results of the narrowband test, i.e., averaged over all listeners for all sequences, are presented in the Prague meeting slides [Prague-80]. The results suggest that Opus at 11 kbps is superior to both iLBC at 15 kbps and Speex at 11 kbps. T-tests performed by Greg Maxwell confirm that there is indeed a statistically significant difference. Note also that Opus has a slightly higher average score than the 3.5 kHz anchor, likely due to the higher bandwidth of Opus.

4.1.2. Google wideband and fullband listening test

The eight test sequences for Test 1 were also used in Test 2. 16 listeners rated the stimuli listed in Table 2. In this test comparisons were made between four wideband codecs: Opus WB, the royalty-free Speex, the royalty-free ITU-T G.722.1, AMR-WB (ITU-T G.722.2), and two fullband codecs: Opus FB and the royalty-free ITU-T G.719. All six codecs utilize 20 ms encoding frames. Opus used variable bitrate, while other codecs used constant bit rate.

Type	Signal bandwidth	Bitrate
Reference	24 kHz (Fullband)	
Anchor 1	3.5 kHz (Narrowband)	
Anchor 2	7 kHz (Wideband)	
G.722.1	7 kHz (Wideband)	24 kbps, CBR
Speex WB	7 kHz (Wideband)	23.8 kbps, CBR
AMR-WB	7 kHz (Wideband)	19.85 kbps, CBR
Opus WB	8 kHz (Wideband)	19.85 kbps, VBR
G.719	~20 kHz (Fullband)	32 kbps, CBR
Opus FB	~20 kHz (Fullband)	32 kbps, CBR

Test 2 stimuli

Table 2

The results from Test 2 are depicted in the Prague meeting slides [Prague-80]. Opus at 32 kbps is almost transparent, although there is a small, but statistically significant, difference from the fullband reference material. Opus at 20 kbps is significantly better than all the other codecs, including AMR-WB and the fullband G.719, and both low-pass anchors.

4.1.3. Google stereo music listening test

The sequences in this test were excerpts from 10 different stereo music files:

- o Rock/RnB (Boz Scaggs)
- o Soft Rock (Steely Dan)
- o Rock (Queen)
- o Jazz (Harry James)
- o Classical (Purcell)

- o Electronica (Matmos)
- o Piano (Moonlight Sonata)
- o Vocals (Suzanne Vega)
- o Glockenspiel
- o Castanets

These sequences were originally recorded at a sampling frequency of 44.1 kHz and were upsampled to 48 kHz prior to processing. Test 3 included comparisons between six codecs (c.f., Table 3): Opus at three rates, G.719, AAC-LC [9] (Nero 1.5.1 [10]), and MP3 [11,12] (Lame 3.98.4 [13]). G.719 is a mono codec, so the two channels were each coded independently at 32 kbps. 9 listeners participated in Test 3, and the results are depicted in the Prague meeting slides [Prague-80]. The codecs operated at constant (or comparable) bit rate.

Type	Signal bandwidth	Frame size (ms)	Bitrate
Reference	22 kHz (Fullband)	-	(1536 kbps)
Anchor 1	3.5 kHz (Narrowband)	-	(256 kbps)
Anchor 2	7 kHz (Wideband)	-	(512 kbps)
MP3	16 kHz (Super wideband)	>100	96 kbps, CBR
AAC-LC	~20 kHz (Fullband)	21	64 kbps, CBR (bit reservoir)
G.719	~20 kHz (Fullband)	20	64 kbps (2x32), CBR
Opus FB	~20 kHz (Fullband)	20	64 kbps, constrained VBR
Opus FB	~20 kHz (Fullband)	10	80 kbps, constrained VBR

Opus FB	~20 kHz (Fullband)	5	128 kbps, constrained VBR
+-----+	+-----+	+-----+	+-----+

Test 3 stimuli

Table 3

The results indicate that all codecs had comparable performance, except for G.719, which had a considerably lower score. T-tests by Greg Maxwell verified that the low-delay Opus at 128 kbps had a significantly higher performance and that G.719 had a significantly lower performance than the other four.

4.2. HydrogenAudio stereo music listening test

In March 2011, the HydrogenAudio community conducted a listening test comparing codec performance on stereo audio at 64 kb/s [ha-test]. The Opus codec was compared to the Apple and Nero implementations of HE-AAC, as well as to the Vorbis codec. The test included 30 audio samples, including known "hard to code" samples from previous HydrogenAudio listening tests.

A total of 33 listeners participated in the test, 10 of which provided results for all the audio samples. The results of test showed that Opus out-performed both HE-AAC implementations as well as Vorbis.

4.3. Nokia Interspeech 2011 listening test

In 2011, Anssi Ramo from Nokia submitted [Ramo2011] the results of a second listening test, focusing specifically on the Opus codec, to Interspeech 2011. As in the previous test, the methodology used was a 9-scale ACR MOS test with clean and noisy speech samples.

The results show Opus clearly out-performing both G.722.1C and G.719 on clean speech at 24 kb/s and above, while on noisy speech all codecs and bit-rates above 24 kb/s are very close. It is also found that the Opus hybrid mode at 28 kb/s has quality that is very close to the recent G.718B standard at the same rate. At 20 kb/s, the Opus wideband mode also out-performs AMR-WB, while the situation is reversed for 12 kb/s and below. The only narrowband rate tested is 6 kb/s, which is below what Opus targets and unsurprisingly shows poorer quality than AMR-NB at 5.9 kb/s.

5. In-the-field testing

Various versions of Opus (or SILK/CELT components) are currently in use in production in the following applications:

- o Skype: VoIP client used by hundreds of millions of people
- o Steam: Gaming distribution and communications platform with over 30 million users
- o Mumble: Gaming VoIP client with more than 200 thousand users
- o Soundjack: Client for live network music performances
- o Freeswitch: Open-source telephony platform
- o Ekiga: Open-source VoIP client
- o CHNC: Radio station using CELT for its studio-transmitter link

6. Conclusion on the requirements

The requirements call for the Opus codec to be better than Speex and iLBC in narrowband mode, better than Speex and G.722.1 in wideband mode, and better than G.722.1C in super-wideband/fullband mode.

6.1. Comparison to Speex (narrowband)

The Opus codec was compared to Speex in narrowband mode in the Google narrowband test (Section 4.1.1). This test showed that Opus at 11 kb/s was significantly better than Speex at the same rate. In fact, Opus at 11 kb/s was tied with the 3.5 low-pass of the original. Considering the results, we conclude that the Opus codec is better than the Speex codec.

6.2. Comparison to iLBC

The Opus codec was compared to iLBC in the Google narrowband test (Section 4.1.1). This test showed that Opus at 11 kb/s was significantly better than iLBC running at 15 kb/s. Considering the results, we conclude that the Opus codec is better than the iLBC codec.

6.3. Comparison to Speex (wideband)

The Opus codec was compared to Speex in wideband mode in the Google wideband and fullband test (Section 4.1.2). This test showed that Opus at 20 kb/s was significantly better than Speex at 24 kb/s. In fact, Opus at 20 kb/s was better than the 7 kHz low-pass of the original. These results are consistent with an earlier Dynastat test (Section 2.1) that also concluded that SILK had significantly higher quality than Speex in wideband mode at the same bit-rate. Considering the results, we conclude that the Opus codec is better than the Speex codec for wideband.

6.4. Comparison to G.722.1

In the Google wideband and fullband test (Section 4.1.2), Opus at 20 kb/s was shown to significantly out-perform G.722.1 operating at 24 kb/s. An indirect comparison point also comes from the Nokia Interspeech 2011 listening test (Section 4.3) that shows Opus out-performing AMR-WB at 20 kb/s, while AMR-WB is known to out-perform G.722.1. Considering these results, we conclude that the Opus codec is better than the G.722.1 codec for wideband.

6.5. Comparison to G.722.1C

Opus has been compared to G.722.1C in multiple listening tests. As early as 2008, an old version of the CELT codec (Section 2.4) using very short frames was found to have higher quality than G.722.1C at 48 kb/s. More recently, the Nokia Interspeech 2011 listening test (Section 4.3) showed that Opus out-performed G.722.1C at 24 kb/s, 32 kb/s, and 48 kb/s. We thus conclude that the Opus codec is better than the G.722.1C codec for superwideband/fullband audio.

6.6. Comparison to AMR-NB

In the Google narrowband test (Section 4.1.1), Opus was shown to out-perform AMR-NB at 12 kb/s. On the other hand, in the Nokia Interspeech 2011 listening test (Section 4.3), AMB-NB was found to have better quality than Opus at 6 kb/s. This indicates that Opus is better than AMR-NB at higher rates and worse at lower rates, which is to be expected given Opus' emphasis on higher quality and higher rates.

6.7. Comparison to AMR-WB

In the Google wideband and fullband test (Section 4.1.2), Opus at 20 kb/s was shown to out-perform AMR-WB at the same rate. This was also confirmed by the Nokia Interspeech 2011 listening test (Section 4.3), with also found AMR-WB to out-perform Opus at 12 kb/s and below. As with AMR-NB, we conclude that Opus is better than AMR-WB at higher rates and worse at lower rates.

7. Security Considerations

No security considerations.

8. IANA Considerations

This document has no actions for IANA.

9. Acknowledgments

The authors would like to thank Anssi Ramo and the HydrogenAudio community, who conducted some of the Opus listening test cited in this draft.

10. Informative References

- [valin2010]
Valin, J., Terriberry, T., Montgomery, C., and G. Maxwell,
"A High-Quality Speech and Audio Codec With Less Than 10
ms delay", 2010.
- [valin2009]
Valin, J., Terriberry, T., and G. Maxwell, "A High-Quality
Speech and Audio Codec With Less Than 10 ms delay", 2010.
- [Wustenhagen2010]
Wuestenhagen, U., Feiten, B., Kroll, J., Raake, A., and M.
Waeltermann, "Evaluation of Super-Wideband Speech and
Audio Codecs", 2010.
- [Ramo2010]
Ramo, A. and H. Toukomaa, "Voice Quality Evaluation of
Recent Open Source Codecs", 2010.
- [Ramo2011]
Ramo, A. and H. Toukomaa, "Voice Quality Characterization
of IETF Opus Codec", 2011.
- [Maastricht-78]
Valin, J. and K. Vos, "Codec Prototype", 2010.
- [Prague-80]
Chen, R., Terriberry, T., Maxwell, G., Skoglund, J., and
H. Nguyet, "Testing results", 2011.
- [SILK-Dynastat]
Skype, "SILK Datasheet", 2009.
- [ha-test] Dyakonov, "Results of the public multiformat listening
test @ 64 kbps", 2011.

Authors' Addresses

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

Email: jmvalin@jmvalin.ca

Koen Vos
Skype Technologies S.A.
Stadsgarden 6
Stockholm, 11645
Sweden

Email: koen.vos@skype.net

Jan Skoglund
Google

Email: jks@google.com

