Voice Quality Characterization of IETF Opus Codec

Anssi Rämö, Henri Toukomaa

Nokia Research Center, Tampere, Finland

anssi.ramo@nokia.com, henri.toukomaa@nokia.com

Abstract

This paper discusses the voice quality of Opus, IETF driven open source voice and audio codec. Opus is a newly developed hybrid codec based on SILK and CELT codec technologies. Opus construction is described shortly in this paper and more importantly its optimal operating points are found out based on the listening test results. Voice quality was evaluated with two subjective listening tests. Industry standard voice codecs: 3GPP AMR and AMR-WB, and ITU-T G.718B, G.722.1C and G.719 as well as direct signals were used as voice quality references.
Index Terms: speech coding, subjective listening test, open source

1. Introduction

There is an ongoing effort in IETF to standardize a new open source voice codec for internet telephony (VoIP) and other demanding realtime applications. Although there already exist several open and ”closed” source IETF standardized codecs not all performance requirements can be fulfilled with a single codec. Opus standardization effort was initiated in order to develop a codec that can fulfill the wide range of requirements and operation points within a single frame-work. Opus tries to be a codec that can be used for all applications from the low bitrate telephony to stereo full bandwidth realtime teleconferencing.[1]

Opus introduces several new features that are not available yet in any other codec. In future, when it is finalized it will supports bitrates from 6 kbit/s (NB mono) to more than 200 kbit/s (FB stereo) both variable and fixed bitrate. It has several frame length options available from 2.5ms to 60ms. It also support mono as well as stereo. In this paper the latest Opus version (obtained from GIT- repository on February 16th 2011 [2]) is studied in detail. Opus codec is evaluated in several operation points against 3GPP standardized AMR and AMR-WB and ITU-T standardized G.718B, G.719 and G.722.1C codecs. All the codecs were tested in two listening tests: the first listening test was conducted with clean and the second with noisy mono speech signals.

This paper is constructed as follows. First some general information about user experience is given in Section 2. Next Section 3 describes the test methodology and listening test setup in detail. Section 4 tells how Opus codec is constructed to perform with different bitrates and how it affects the signal bandwidth and voice quality. Finally conclusions are drawn in Section 5.

2. Improved user experience

For over a century telephony has relied on narrowband (NB) voice quality, which is barely good enough to transport the most important elements of human speech. Wideband (WB) has finally begun coming to mobile services and devices. For example ”HD Voice” has been launched recently by several operators and mobile device manufacturers. ”HD Voice” uses AMR-WB voice codec and thus provides wideband capability. In the internet world widely used Skype VoIP service uses Silk codec as well as some other proprietary voice codecs for wider bandwidths up to superwideband. Overall, current trend in voice codec development is to use even wider bandwidths and thus higher sampling rates for even further improved user experience. Later on it can be expected that stereo or binaural voice communication makes an appearance in some form. Traditionally sampling rates have increased by doubling (like 8, 16, and 32 kHz). In addition Opus supports internally also 12 and 24 kHz sampling rates. Hybrid and MDCT mode also support 48 kHz sampling rate. Opus codec uses critical sampling and for example at 16 kHz sampling rate all frequencies up to 8 kHz are present. See Table 1 for summary of all bandwidth abbreviations and related signal bandwidths and sampling rates. As can be seen there are several different high-pass frequencies as well as high frequency cut-off frequencies available depending on who you are referring to. Opus codec’s LP and hybrid modes use variable frequency high-pass filter. In high SNR and when there is a low pitched speaker the high pass is set as low as 80 Hz, in low SNR and with high pitched voices up to 150 Hz high pass frequency is used. In this paper ITU-T definitions were used for signal preprocessing.

Abbr.	Meaning	Pass-band	Sampling rate
NB[3]	Narrowband	300- 3 400Hz	8 kHz
NB[1]	Narrowband	80/150- 4 000Hz	8 kHz
NB[2]	Narrowband	20- 4 000Hz	8 kHz
MB[1]	Mediumband	80/150- 6 000 Hz	12 kHz
WB[3]	Wideband	50- 7 000 Hz	16 kHz
WB[1]	Wideband	80/150- 8 000 Hz	16 kHz
WB[2]	Wideband	20- 8 000 Hz	16 kHz
SWB[3]	Superwideband	50- 14 000 Hz	32 kHz
SWB[1]	Superwideband	80/150- 16 000 Hz	32 kHz
SWB[2]	Superwideband	20- 16 000 Hz	32 kHz
FB[2],[3]	Fullband	20- 20 000 Hz	48 kHz

Table 1: Abbreviations used for different signal bandwidths and respective sampling rates. [1] Opus LP and hybrid core bandwidth [2] Opus MDCT core bandwidth [3] ITU-T definition

3. Test Description

Listening test was conducted in Nokia Research Center Listening Test Laboratory [3]. The main research question was: How do naïve listeners experience differently encoded NB, WB and SWB signals without any preparatory information. 24 naïve listeners took part in the listening test. Each listener evaluated over 30 conditions with 8 + 8 voice samples from all scenarios described in section 3.2. Thus each listener scored about 600

individually processed samples in random order. Since each sample took about 5 seconds to listen and evaluate, and there are mandatory comfort breaks every twenty minutes, the listening took about one and half hour per listener. Each condition obtained 200 votes for clean speech and another 200 votes for noisy speech. In order to have some initial scale to the listeners, the test started with 12 introductory (practice) samples, which represented the full scale of the conditions. These preparatory test results were omitted from the final results.

3.1. Extended Range MOS Test Method

A modified version of the traditional ACR (Absolute Category Rating) [4] (MOS) method was used for the listening test. The MOS scale was extended to be 9 categories wide in order to get more accurate results with relatively high quality and wide bandwidth speech signals. Table 2 shows the available categories. Only the extreme categories were defined with verbal description: 1 ”very bad” and 9 ”Excellent”. The assessment is not free sliding, but nine different values still provide listener more ways to discriminate the samples than five. In practice 9-scale MOS test is also much faster to conduct with naïve listeners than for example MUSHRA methodology.

Grading value	Estimated Quality
9	Excellent
8	
...	
2	
1	Very bad

Table 2: 9-step ACR scale without intermediate adjectives

3.2. Test samples

The test material contained female and male voice samples both with clean voice and with voice in background noise. Each listened sample contained a Finnish language sentence that lasted a 2 3 seconds with silence or noise background. Clean speech test contained samples from four males and four females. Noisy speech test had similar distribution of speakers. Table 3 shows all eight different sample types with different background noise types and relative background noise levels that were used in the noisy listening test. For real recordings the background noise level cannot be objectively measured.

Set	Speaker	Background noise	Noise Level
1	Male 1	In car with radio	real recording
2	Female 1	Outside	real recording
3	Male 2	Classical music	-20 dB
4	Female 2	Classical music	-15 dB
5	Male 3	Office	-20 dB
6	Female 3	Pop music	-20 dB
7	Male 4	Street	-15 dB
8	Female 4	Cafeteria	-20 dB

Table 3: Sample sets used for noisy condition listening test

Listening test results are presented in a X-Y line graph (e.g. Figure 1), where bullets point to individual MOS results and interpolated line connects the bullets, when relevant scalable codec or codec family result is shown. On the left side of the table MOS scale is shown. On the bottom bitrate is shown. Confidence intervals were omitted from these graphs for clarity. Final bar-graph Figure 4 includes also 95% confidence intervals.

4. IETF Opus voice codec

Opus codec consist of two previously known codecs SILK and CELT [5] [6]. These two codecs are combined into a single framework with three basic operation modes. The operation modes are: LP mode (Silk based), hybrid mode, and MDCT mode (CELT based). The LP mode works best with narrow to wideband bandwidths and with relatively low bitrates and it is optimized for speech signals. MDCT mode on the other hand is a quite traditional audio codec and thus requires significantly higher bitrates for high voice quality and is generally best suited for audio signals. Hybrid mode combines both of these for efficient coding of voice and audio signals with mediocre bitrates. Lower frequencies up to 8 kHz are coded with LP mode and higher frequencies with high frequency coding optimized MDCT mode.

Roughly it can be said that LP mode works best with bitrates below 20 kbit/s and only with narrowband and wideband bandwidths. Hybrid mode is optimal when bitrate range is from 20 to 48 kbit/s while using SWB or FB bandwidth signal. MDCT mode can be used for ultimate voice and music quality with even higher bitrates and it supports all bandwidths from NB upwards. MDCT NB and WB modes are meant to be used with music signals, with speech signals LP mode gives much higher quality with the same bitrate. All optimal bitrate ranges and bandwidths are summarized in Table 5.

Opus codec supports several frame lengths from 2.5 ms (MDCT mode) up to 60 ms (LP mode). In general shorter window sizes require higher bitrate for equivalent voice and audio quality. The listening tests in this were performed with a constant window length of 20 ms. With that frame length there is also 5 ms look ahead. In addition to window length adjustment Opus support multiple frame packetization and redundant information. Packetization improves coding efficiency by reducing the number of packet headers needed per second, but this naturally increases delay. Redundant information increases bitrate, but improves quality by helping the decoder recover faster, if there are frame losses. Also the complexity can be adjusted. Lighter computation naturally means slightly degraded quality. Overall not a complete complexity analysis was performed for Opus codec in this paper, but it definitely more than with G.722.1C or G.719 and quite similar to G.718B codec based on the used runtime during the test material processing.

4.1. LP mode

Silk codec based LP mode is inheritably variable bitrate and also signal bandwidth varies with time, if not forced to a low enough bandwidth relative to the bitrate [5]. More details about this behavior can be found in [7]. We processed the samples in such a way that codec had enough time to achieve steady-state bandwidth and bitrate.

Table 4 shows how Silk’s bandwidth changes with different bitrates. However Opus’s LP mode is restricted to WB bandwidth at maximum, since for medium bitrates new Opus hybrid mode should be used instead. That is why we tested LP mode only up to 20 kbit/s. All shown bitrates with LP mode are the requested bitrate from the command line and in practice this was within 5% of the obtained average bitrate within the test sequence.

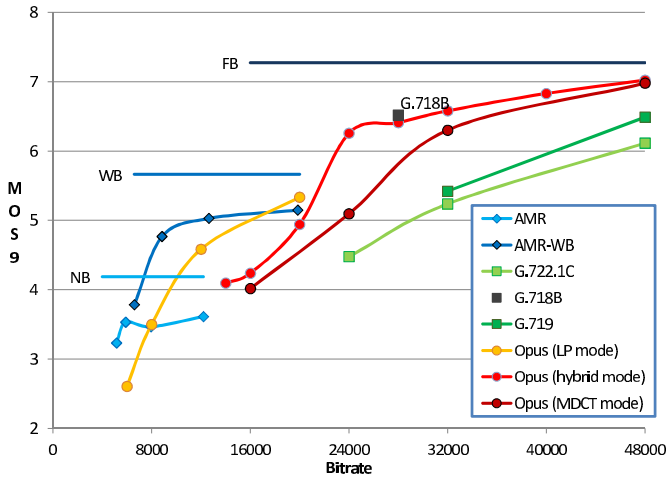

Figure 1: *Voice quality evaluation results in clean speech*

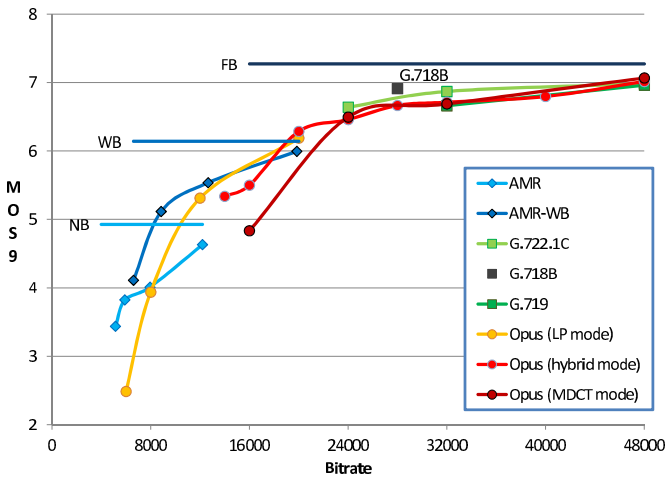

Figure 2: *Voice quality evaluation results in noisy speech*

Bandwidth	Internal sampling rate	Bitrate range in steady state
Narrowband (NB)	8 kHz	5.0- 9.0 kbit/s
Mediumband (MB)	12 kHz	9.5- 14.0 kbit/s
Wideband (WB)	16 kHz	14.5- 24.0 kbit/s
Superwideband (SWB) [1]	24 kHz	$\geq$ 24.5 kbit/s

Table 4: Silk internal sampling rates and supported signal bandwidths with different bitrates. [1] not supported in Opus's LP mode

From Figures 1 to 3 LP mode voice quality can be compared to all other tested codecs with clean and noisy signals and finally all results combined. As can be seen LP mode performs slightly worse than AMR or AMR-WB especially at the lowest bitrates (below 6 kbit/s for NB and below 12 kbit/s for WB). However at 20 kbit/s LP mode is better than AMR-WB and very close to WB direct. This is due to the fact that LP mode's wideband is critically sampled and transmits frequencies 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 giving good results.

4.2. Hybrid mode

Hybrid mode subjective performance is of the greatest interest in the Opus codec. It is meant to fill the void between the good low bitrate performance of Silk based LP mode and high bitrate optimized CELT based MDCT mode. In practice hybrid mode uses filter bank to divide the spectrum to lower (below 8 kHz) and higher parts and code them independently with LP and MDCT mode respectively. The arithmetic lossless coding part of the codec is shared between the modes. This means that fixed bitrate is also supported with hybrid mode, since variable bitrate LP mode is hidden by the common bit-pool. In this paper constant bitrate was used.

The lowest operating bitrate for the hybrid mode was found out to be around 14 kbit/s with our test material (with lower bitrates the codec crashes). At that bitrate the voice quality is however still significantly worse than with LP mode. At around 20 kbit/s hybrid mode achieves about the same quality than LP mode. Already at 24 kbit/s quite respectable voice quality is reached. Hybrid mode at this bitrate produces especially with clean speech much better voice quality than for example G.722.1C or MDCT mode (Figure 1). With increasing bitrate the quality improves gradually. The only standardized codec that is slightly better with SWB speech signals is the recent voice optimized ITU-T G.718B codec operating at 28 kbit/s. With noisy speech the differences between G.722.1C, G.719, MDCT mode or hybrid mode are very marginal. This again proves that clean speech is very hard signal to code for generic audio codecs and for optimal efficiency also a speech optimized mode is mandatory.

4.3. MDCT mode

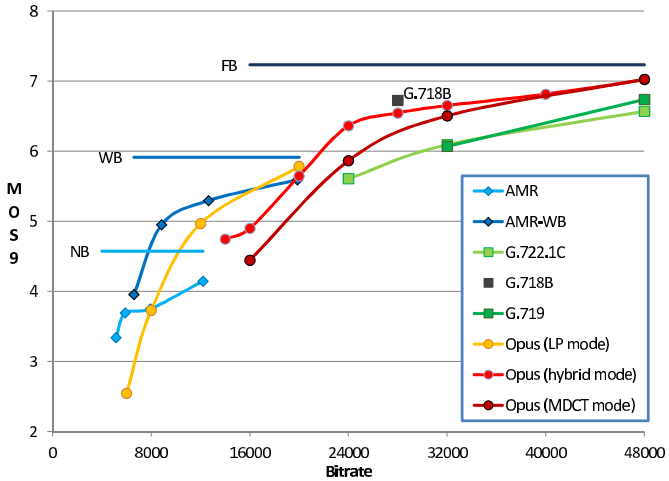

Figure 3: *Overall voice quality. Combined results for both clean and noisy voice.*

CELT codec based MDCT mode is basically a low delay generic audio codec [8]. It supports sampling rates from 8 kHz upwards. In this test we used 16 kHz sampling rate for the lowest bitrate of 16 kbit/s with 20 ms frame size. MDCT mode with 8 kHz sampling rate was not tested, since we know that with speech signals it is anyways much inferior to LP mode at these low bitrates. Fullband 48 kHz sampling rate was used for

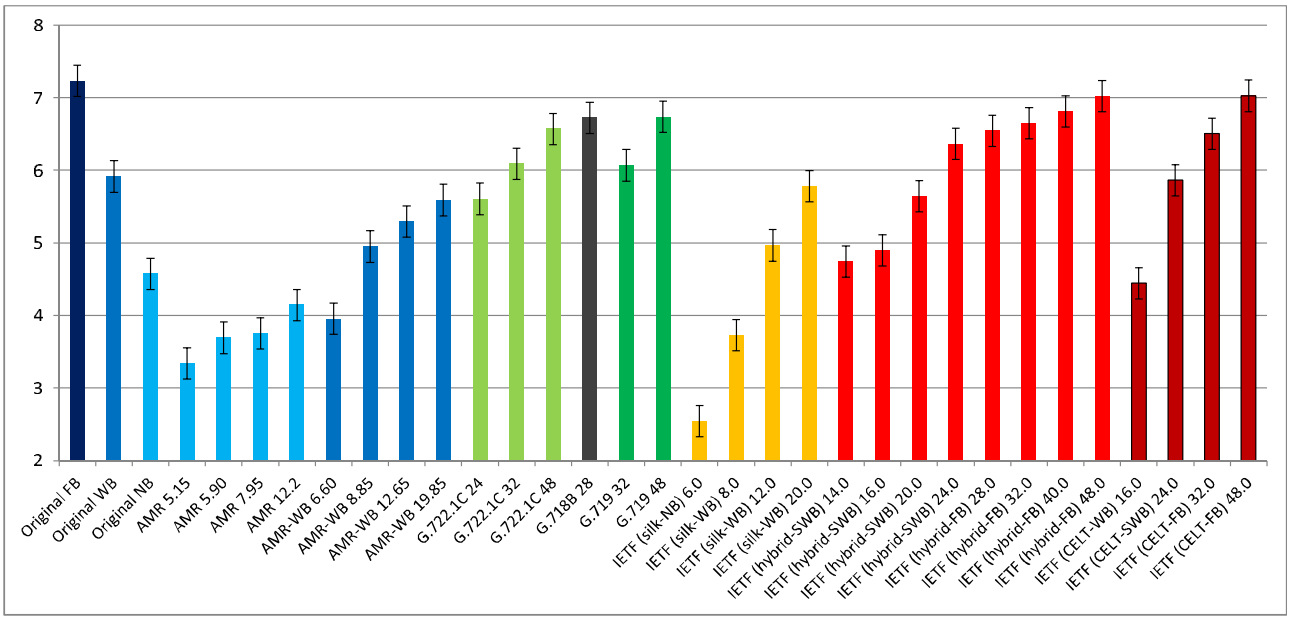

Figure 4: *Combined MOS scores with bars and confidence intervals*

the rest of the modes. 32 kHz sampling rate was omitted from the listening test since the tested codec version did not support it (codec crashed).

Results in Figure 1 show that MDCT mode requires significantly higher bitrates in clean speech than either LP or hybrid mode. However at 32 kbit/s it provides significantly better voice quality than G.722.1C or G.719, which is very respectable. Actually it seems that Opus MDCT mode has improved quite much from the tested CELT version last year. E.g. compared to G.722.1C at 32 kbit/s MDCT mode is now much better and last year it was significantly worse [7].

At bitrates of 40 kbit/s or more MDCT mode is already about the same quality as hybrid mode and it is very likely that with even higher bitrates MDCT performs better than hybrid mode, which becomes quality limited by the LP mode's high-pass filtering and filter bank notch at 8 kHz. With noisy speech even 24 kbit/s is useable for MDCT mode as can be seen in Figure 2. However if the used content is known to contain speech at bitrates 20- 40 kbit/s, hybrid mode provides more stable and better overall voice quality (see Figure 3).

5. Conclusions

Opus codec's LP mode provides useable voice quality at quite competitive bitrates compared to AMR or AMR-WB. However, there are two things to consider when using the LP mode. The first one is the highly variable bitrate, which may cause problems depending on the transmission network and channel. The second one is that the provided signal bandwidth is also changing with time and for wideband quality significantly higher bitrates than AMR-WB are needed (around 14.5 kbit/s at minimum). CELT based MDCT mode provides a good alternative to ITU-T G.722.1C or G.719 by providing better quality with more computational complexity, where that can supported. Finally hybrid mode provides excellent voice quality at bitrates from 20 to 40 kbit/s. Table 5 shows how to configure Opus internal modes in order to get the best possible voice quality with limited bit budget.

Codec mode	frame lengths (ms)	pass-band (Hz)	optimal bitrates (kbit/s)
LP NB	10, 20, 40, 60	150- 4 000	speech 5.0- 12
LP MB	10, 20, 40, 60	80- 6 000	speech 9.5- 16
LP WB	10, 20, 40, 60	80- 8 000	speech 14.5- 24
Hybrid SWB	10, 20	80- 16 000	all 20- 40
Hybrid FB	10, 20	80- 20 000	all 28- 48
MDCT NB	2.5, 5, 10, 20	20- 4 000	music -16
MDCT WB	2.5, 5, 10, 20	20- 8 000	music 16- 24
MDCT SWB	2.5, 5, 10, 20	20- 16 000	music 20- 36
MDCT FB	2.5, 5, 10, 20	20- 20 000	all 32- 128

Table 5: Opus core coding options showing all supported frame lengths, pass-bands and approximate optimal bitrate ranges (with 20ms frame size) for different content types

6. References

[1] Jean-Marc Valin and Koen Vos, "Definition of the opus audio codec," in *IETF draft*, March 2011.

[2] GIT repository, "Ietf opus codec," `http://git.xiph.org/?p=users/jm/ietfcodec.git`.

[3] M. Kylliäinen et al., "Compact high performance listening spaces," in *Proc. of Euronoise*, Italy, 2003.

[4] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.

[5] Koen Vos, Soeren Jensen, and Karsten Soerensen, "Silk speech codec," in *IETF draft*, September 2010.

[6] Jean-Marc Valin, T. Terriberry, G. Maxwell, and C. Montgomery, "Constrained-energy lapped transform (celt) codec," in *IETF draft*, July 2010.

[7] A. Rämö and H. Toukomaa, "Voice quality evaluation of recent open source codecs," in *Proc. of Interspeech*, Tokyo, Japan, 2010.

[8] Jean-Marc Valin et. al, "A high-quality speech and audio codec with less than 10 ms delay," *IEEE Transactions on audio, speech and language processing*, vol. 18, no. 1, pp. 58–67, January 2010.