Voice Quality Evaluation of Recent Open Source Codecs

Anssi Rämö, Henri Toukomaa

Nokia Research Center, Tampere, Finland

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

Abstract

This paper introduces Silk, CELT, and BroadVoice that are available in the internet as open source voice codecs. Their voice quality is evaluated with a subjective listening test. AMR, AMR-WB, G.718, G.718B, and G.722.1C were used as standardized references. In addition Skype’s Silk codec’s peculiar bandwidth and bitrate characteristics are examined in more detail.

Index Terms: speech coding, subjective listening test, open source

1. Introduction

There is an effort ongoing in IETF to standardize a new voice codec for internet telephony (VoIP) that is also an open source. In this paper three different relatively recent open source voice codecs Silk, CELT and BroadVoice are evaluated against 3GPP standardized AMR and AMR-WB and ITU-T standardized G.718, G.718B and G.722.1C codecs. All the codecs were tested in a same listening test with both clean and noisy mono speech signals. There exists several other open source codecs, and older codecs such as iLBC and Speex were already tested in our earlier papers ([1] [2]) and they were not included into this listening test. In addition to pure listening test results a more detailed analysis is made from the Silk codec, since it is not a traditional fixed bit-rate codec. It adapts both its bitrate and signal bandwidth on a frame-by-frame basis resulting a highly variable bitrate codec.

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 shows how Silk codec achieves different bitrates and how it affects the signal bandwidth and voice quality. Sections 5 and 6 explain subjective listening test results from CELT and BroadVoice codecs respectively. Finally conclusions are drawn in Section 7.

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 been coming to mobile services and devices for a few years, but the final breakthrough has not yet happened. Current trend in voice codec development is to use even wider bandwidths and higher sampling rates for improved user experience. Traditionally sampling rates have increased by doubling (like 8, 16, and 32 kHz), however Skype’s Silk codec is using some unconventional bandwidths and sampling rates for describing their codecs bandwidth. In addition Silk codec uses critical sampling and for example at 16 kHz sampling rate all frequencies up to 8 kHz are present. See Table 1 for detailed bandwidth ranges and sampling rates.

3. Test Description

Listening test was conducted in Nokia Research Center Listening Test Laboratory [3]. The main research question was: How do naive listeners experience differently encoded NB, WB and SWB signals without any preparatory information. 40 naive listeners took part in the listening test. Each listener evaluated over 30 conditions with 6 voice samples from all scenarios described in section 3.2. Thus each listener scored about 200 individually processed samples each in random order. Since each sample took about 10 seconds to listen and evaluate, and there are mandatory comfort breaks every twenty minutes, the listening took about one hour per listener. Each condition obtained 240 votes. 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. Only the extreme categories were defined with verbal description: 1 “very bad” and 9 “Excellent”. We have experienced that the 9-scale MOS saturates less easily than the standard 5-scale MOS with these kind of high quality samples. In practise this new scale is somewhat between MUSHRA (ITU-R BS.1534) and 5-scale MOS. The assessment is not free sliding, but nine different values still provide listener more ways to discriminate the samples. In practise 9-scale MOS test is also much faster to conduct with naive listeners than MUSHRA.

3.2. Test samples

The test material contained female and male voice samples with clean voice and voice in background noise. Each listened sample contained a sentence pair that lasted a total of 8 seconds. Table 2 shows all six different sample types that were used for the voice quality assessment. Also all different background noise types and relative background noise levels are shown.

In addition to normal bar graphs (Figure 3) codec and ref-
ference conditions are also collected to a X-Y line graph (Figure 5), 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.

4. Skype's Silk voice codec

Silk codec is currently used by the popular VoIP service Skype. It uses technology described in [5]. The main parameters are 20 ms frame size and 5 ms look-ahead, target bitrate can be set to any value between 6 and 40 kbit/s. Silk also supports redundant information and up to five frames can be coded and packetized together. Packetization improves coding efficiency slightly, but increases delay. Redundant information increases bitrate, but improves quality if there are frame losses. Also the complexity can be adjusted. Lighter computation naturally means slightly degraded quality.

Silk codec bitrate and signal bandwidth vary in time. For example Figures 1 and 2 show that when the bitrate is set to either 16 kbit/s or 8 kbit/s the bitrate starts from a much higher value and in couple of ten seconds reaches the required average bitrate. Similarly Silk also uses wider bandwidth at the beginning before it notices that requested bitrate does not allow so wide bandwidth and reduces bandwidth with internal resampling to a lower sampling frequency. Also the lowest frequency (high-pass filtering) frequency is adapted according to the input signal properties. Due to this behaviour, we processed the test samples in such a way that a 30 second voice presignal with low level background music was processed before the actual test sample (24 seconds) was processed. Thus the Silk codec had enough time to iterate to the requested bitrate and select the suitable bandwidth that was possible with that bitrate.

The measured bitrates can be seen in Table 3. In the table there are several bitrates described. The first column tells the bitrate that was requested from the command line. The second column tells the average bitrate during the tested sequence (30 seconds). The third column tells the bitrate during active speech. The difference between average and active bitrate is quite large with increasing bitrate. Typical WB bandwidth can be seen at 16 kbit/s. Full SWB bandwidth is shown when bitrate is set to 16 kbit/s or higher. Even with 24 kbit/s highest frequencies from 10 kHz to 12 kHz are significantly attenuated. At 12 kbit/s, which is claimed to be the first SWB bitrate, the codec is barely reaching 6 kHz bandwidth, pointing that Silk is still internally running at 12 kHz (MB) sampling rate.

4.1. Silk codecs true bandwidth

Silk codec supports internally and externally sampling rates 8, 12, 16, and 24 kHz and bitrates from 6 to 40 kbit/s [6]. External sampling rate however does not tell the actual internally used bandwidth of the codec. Skype claims in [6] that wideband is supported from 7 kbit/s upwards and superwideband from 12 kbit/s. From our evaluation it seems that required bitrates are significantly higher for these bandwidths. For example Figure 2 shows that at 8 kbit/s Silk is providing only NB signal bandwidth. With iterating 500 bit/s steps we found out that with our speech database (including both noisy and clean speech) wideband is reached with 9.5 kbit/s and SWB requires 24.5 kbit/s. See Table 4 for a complete list.

In the Figure 4 we see how the frequency response widens with increasing bitrate. Typical WB bandwidth can be seen at 16 kbit/s. Full SWB 80–12 000 Hz bandwidth is shown when bitrate is set to 16 kbit/s. Even with 24 kbit/s highest frequencies from 10 kHz to 12 kHz are significantly attenuated. At 12 kbit/s, which is claimed to be the first SWB bitrate, the codec is barely reaching 6 kHz bandwidth, pointing that Silk is still internally running at 12 kHz (MB) sampling rate.

4.2. Silk Voice Quality

From Figure 3 Silk voice quality can be compared to all other tested codecs and clean references. As can be seen Silk performs quite well compared to state-of-the-art AMR and AMR-
WB. 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 beneficial 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.

Since Silk bitrate varies from frame to frame, drawing MOS values to X-Y graph in Figure 5 with average bitrate or even active average bitrate (see Table 3) is not fair for other codecs, we chose to set the Silk bitrate to 90% value. 90%-value indicates the bitrate below which 90% of frames have lower than this bitrate, but at the same time it also tells that 10% of the frames have higher bitrate, which may cause frame loss in low bitrate channels.

5. CELT codec voice quality

CELT codec is low delay generic audio codec [7]. It supports sampling rates from 32 kHz upwards. In this test we used only 32 kHz sampling rate with 512 samples long frame (16 ms). Codec also requires half a frame (8 ms) for a look-ahead. LTP was enabled, variable bitrate disabled and complexity was left to default value.

Results in Figures 3 and 5 show, that for useable quality at least 32 kbit/s is needed. At that bitrate it’s voice quality is still behind the ITU-T G.722.1C at 32 kbit/s and about the same as AMR-WB at 12.65 kbit/s or Silk 12 at kbit/s. With 40 kbit/s the voice quality is already very good and similar to G.718B at 32 kbit/s. For truly transparent quality 48 kbit/s or more is needed.
6. BroadVoice codec voice quality

BroadCom has also provided BroadVoice codec for IETF standardization [8]. BroadVoice’s solution consists of two fixed bitrate codecs BV16 and BV32, but they can be considered to be just two different modes of the same codec, since most of the algorithms and source code are the same in both codecs. BroadVoice 16 kbit/s variant is narrowband codec and 32 kbit/s is wideband. They are described in more detail in [9] and [10]. Both codecs have delay of 5 ms without lookahead, making them very low delay. Complexities are 12 and 17 MIPS respectively. From the results in Figures 3 and 5 it can be seen that both BroadVoice codecs provide reasonable NB and WB absolute voice quality. Unfortunately their bitrates are not competitive. AMR 10.2 kbit/s mode provides similar quality to BroadVoice at 16kbit/s. AMR-WB 12.65 kbit/s provides similar quality to BroadVoice 32 kbit/s. However, if some application requires extremely low delay BroadVoice is a good candidate there.

7. Conclusions

Silk codec provides useable voice quality at quite competitive bitrates compared to AMR or AMR-WB. However, there are two things to consider when using the Silk codec. The first one is the highly variable bitrate, that 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 (at average 14.5 kbit/s, with peak bitrate approaching 20 kbit/s). Also Silk’s promises about SWB bandwidth at 12 kbit/s are not valid. In practise 24.5 kbit/s or higher bitrate is needed. CELT codec provides a good alternative to ITU-T G.722.1C or G.718B at slightly higher bitrate. Finally BroadVoice is not competitive bitrate wise compared to AMR, AMR-WB or Silk. However due to its extremely low delay, it may have some special applications.

8. References