Audio and Speech Coding & Transcoding in Web Real-Time Communication

Oliver Jokisch, Michael Maruschke
Leipzig University of Telecommunications (HfTL), Leipzig, Germany
{jokisch, maruschke}@hft-leipzig.de

Abstract: Highly-efficient technologies or algorithms, including speech & audio coding, play an important role in human-machine interaction, human communication and user interface design. This contribution summarizes selected results from our performance studies on speech and audio codecs – mainly within the internet-oriented WebRTC scenario in comparison to our tests with codecs designed for cellular phone networks. Furthermore, some implications of transcoding are surveyed. Finally, we address the research potential with regard to both, Opus codec and Enhanced Voice Services (EVS) codec.

Keywords: Opus, EVS, RTC, VoLTE, speech quality, latency

1. Introduction

Web browser-based real-time communication is a relevant topic in daily voice and video communication of people. Many electronic devices such as smartphones, tablet PCs or smart laptops and their integrated web browsers support the Web Real-Time Communication (WebRTC) functionality [1]. Hence an intuitive browser design including highly-efficient baseline technologies plays an increasing role for many applications in human-machine interaction and daily communication which we consider as a technological aspect of human life design (HLD).

The RFC 6716 specified the Opus codec as a highly versatile audio codec for interactive voice and music transmission with frequency ranges up to 20 kHz (full band – FB) [2], which requires an adequate quality, compression and processing time performance of the audio, speech or video coding.

In a previous review we surveyed the dynamic functioning of the Opus codec within a WebRTC framework based on the Google Chrome browser [3]. The codec behavior and the effectively utilized features during the active communication process were tested and analyzed under various testing conditions.

Continuing this investigation, we analyzed the audio and speech quality in the mentioned WebRTC framework by different methods [4]. For the instrumental quality assessment we used two methods, the Perceptual Objective Listening Quality Assessment (POLQA), version 2, with regard to the ITU-T P.863 recommendation [5]. For the comparison with human decisions as a ground truth, we performed a perceptual test with 26 probands.

Furthermore, the impact of audio transcoding procedures during an active WebRTC communication session has been reviewed and published [6]. In this study we analyzed the delay time of High-Definition (HD) voice codecs like Opus and G.722 caused by potential transcoding operations between different codecs.

Modern audio and speech codecs are characterized by a low codec processing time (latency) and by a low transfer bitrate. Typically for cellular phone networks of all used generations (from 2G up to 4G), the requirements for a low transfer bitrate are very stringent. Therefore, the application-specific codecs such as AMR-WB can be distinguished from other codecs (e. g. the internet-oriented Opus codec) with regard to the achievable audio quality by a variable low bit rate – depending on the according mobile network side. To survey possible implications of these aspects in coding and transcoding, a further study was focused to the mouth-to-ear transmission delay and the perceived voice transmission delay for Voice over Long-Term Evolution (VoLTE) calls in the Pan-European network of a leading mobile network operator [7]. A second study dealt with the audio quality performance using different generations of media gateways within a public mobile network (2G vs. 3G) [8].

Our extended abstract summarizes selected results from the mentioned studies – also considering converging technologies (here: coding and transcoding methods) in HLD applications.

Finally, we try to give a short perspective to the potential of full band audio and speech codecs in the daily life communication, as well as for the internet browser-focused Opus codec and cellular network-optimized codecs like the Enhanced Voice Services (EVS) specified by the 3rd Generation Partnership Project (3GPP) in Release 12 [9].

2. Opus codec in a WebRTC framework

We targeted on already established WebRTC scenarios. In our reviews, we used the real-time communication (RTC) function of the Google Chrome browser.

2.1 Verifying codec functioning and performance

Figure 1 illustrates how the audio codecs are integrated in the browsers' RTC function framework. In default, the following Opus codec parameters we detected running a WebRTC communication with the Google Chrome browser:

1. Sample rate: 48 kHz,
2. Audio bandwidth: Full-band (FB),
3. Used encoder bit rate: 32 kbps,
4. Used channels: 1 (mono),
5. Opus working mode: CELT and Hybrid,
6. Frame duration: 20ms,
7. Complexity: 9.

Two parameter values define the complexity (based on language C and Libjingle software part) – value 5 for Android, iOS- or ARM-based devices and 9 for all others (like laptop or desktop PC). All parameters except for Opus working mode and audio bandwidth can directly be modified by the Voice Engine module if requested in the SDP-based session description. The Opus working mode depends on the encoder bit rate while the bit rate tightly depends on the sample rate.

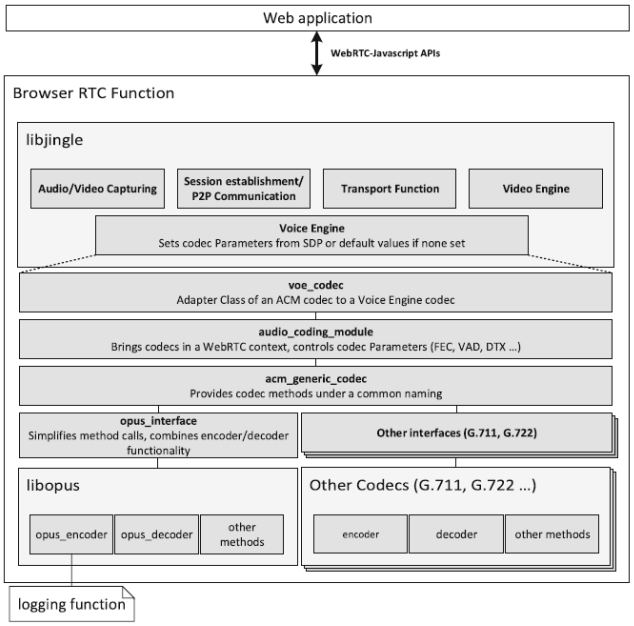

Figure 1. Implementation of audio codecs in the browsers' RTC function [3]

Table 1 illustrates the practical results which were achieved by changing the Opus codec parameters: *effective sample rate* and *channel count*. For various audio bandwidths the shifting of the operation mode is evident. It is obvious that the encoder bit rate is doubled when using stereo instead of mono mode. Using default parameters, the Opus working modes CELT and Hybrid were monitored (cf. last line of Table 1). In this context, Opus operates conform to its definition (cf. [2]).

Table 1. Opus encoder operating mode in dependence of the sample rate [3].

Sample rate [kHz]	**Audio bandwidth**	**Used encoder bitrate [kbps]**		**Operation mode**	**Parameters**
		Mono	**Stereo**		
8	Narrowband	12	24	SILK	Manipulated
12	Mediumband	20	40	SILK	Manipulated
16	Wideband	20	40	SILK	Manipulated
24	Super-WB	32	64	Hybrid & CELT	Manipulated
48	**Fullband**	**32**	**64**	**Hybrid & CELT**	**Default**

2.2 Audio and speech quality assessment

Our test database contained up to 81 audio files:

- 30 full band utterances of 5 male and 5 female speakers (part 1a),
- 36 wide band speech utterances with acted emotional and neutral speech, 1 male and 1 female (part 1b),
- 4 full band music pieces (Jazz and Ska) and 11 different music/singing voice examples from Rock, Blues, Pop, Poprock, Funk, Chanson, acoustic guitar and Ska (part 2).

We tested prototypical cases of Opus coding and its assessment in the WebRTC framework [5]. In the instrumental assessment via POLQA, the framework-coded speech achieves a MOS* up to 4.73 – compared to 4.39 (G7.11). In the best case scenario – based on read FB speech from database 1a – POLQA predicts MOS* values from 4.64 (female samples) to 4.73 (male) which seems equivalent to standalone assessments of the Opus codec without WebRTC influence. The WB speech results on database 1b are significantly worse showing MOS* values from 4.27 (emotional speech) to 4.60 (neutral speech) but the differences are mainly emotion-based. FB music shows a strong degradation of about 0.80 on MOS scale compared with FB speech whereas the differences between partly vocal and strictly instrumental music are not significant ($< \pm 0.10$) considering the low number of samples (only four in test database 2b). With regard to the previously studied WebRTC operation CELT or Hybrid and competitive codec parameters (bitrate 32 kbps and calculation complexity 9 of 10), the observed degradations are either input-related (emotional vs. neutral speech) or based on limitations in the psycho-acoustic modeling (music vs. speech). The samples of anger score 0.40 higher than the neutral ones in WB speech (1b) and similar to FB neutral speech (1a). The music samples (2a and 2b) achieve better assessment than WB neutral speech (1b).

There was no significant assessment difference between male and female probands whereas an age influence could be observed (for several cases ΔMOS about ± 0.25) – five listeners above 40 years rated most of the samples higher as a rule. The Figure 2 shows the averaged MOS results in the test parts 1a, 1b and 2.

The real (perceptual) MOS values across 30 test samples are generally lower than instrumental (POLQA-predicted) MOS* ones. Nevertheless, selected assessments are within expectations compared to the prediction – e. g. samples for happiness in listener group ≥ 40 (MOS = 4.25 vs. MOS* = 4.29) or vocal music in age group ≥ 40 also (MOS = 3.69 vs. MOS* = 3.81). Both, predicted and perceptual results in FB music support the Opus codec in being a multifunctional and highly-adaptive codec. Beyond, the partial results in vocal music indicate that an assessment via POLQA can be applied to a certain extent for singing voices, too. The very low MOS of 2.96 vs. the predicted MOS* of 3.90 in instrumental music illustrates that POLQA is not appropriate in such test cases.

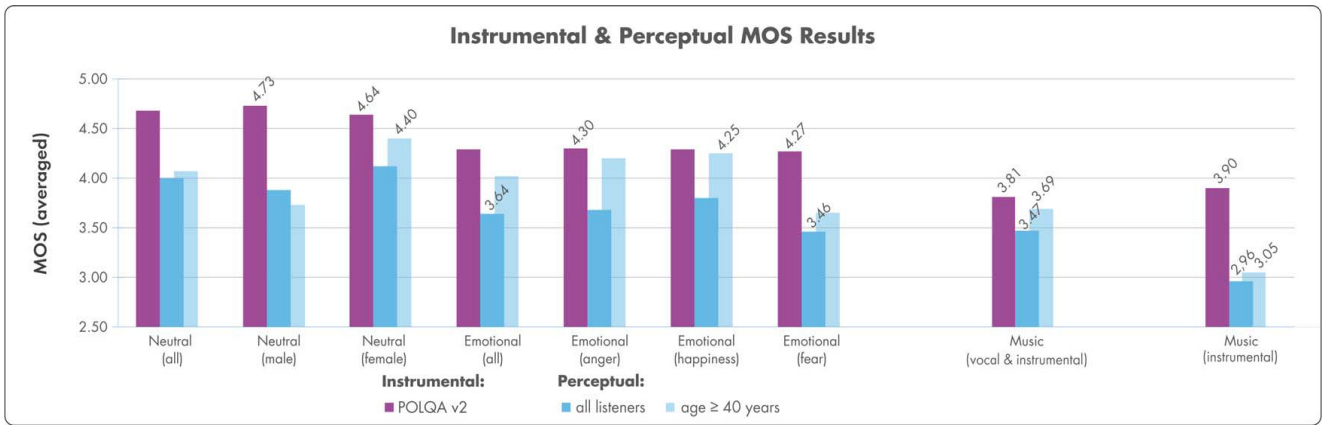

Figure 2. Instrumental versus perceptual MOS results [5]

Some assessments are contradictory – e. g. female voices score with slightly higher MOS than male ones which is reverse in the MOS* values – but this might be not representative as the 30 listening samples incorporate a subset of the 81 samples in the instrumental assessment (for reason of time). Some categories in the listening test are covered by three examples only. Beyond, the data sets were manually selected and potentially biased to noticeable coding examples.

The observed quality degradations are mainly influenced by variations in emotional or neutral speech and by vocal or instrumental parts in the evaluated music samples. The tests also indicate that POLQA can be used in the assessment of vocal music although this is not standardized yet.

The selected perceptual assessments support the predicted tendencies whereas the absolute MOS values are generally lower. We need to carry out additional experiments with our coding framework to consolidate possible differences between POLQA and perceptual assessment.

2.3 Transcoding impact

The additional research question concerns the impact of voice transcoding along a Voice-over-IP user communication path – in particular aspects of the end-to-end delay time.

In the investigation, we tested WebRTC-based communication scenarios with both, the HD-voice codecs Opus and G.722 but also the legacy narrowband codec G.711 [6]. As a transcoding unit we used the single-board computer Raspberry Pi with embedded voice transcoding functionalities – taken from an open source project of Doubango Telecom [10]. Subsequently, we determined the speech quality using the POLQA assessment method. The overall transcoding time and the resulting MOS* values are depicted in Table 2.

The results indicate that the transcoding delay budget reached values between 21 and 27 ms. The measured MOS* values confirm the theoretical approach: While narrowband codecs like G.711 achieve MOS values < 4.50, the wideband codecs exceed MOS values of 4.60.

Table 2. Transcoding delay and resulting MOS* values [6]

Transcoding type	Total transcoding time (ms)	POLQA v2 (MOS*)
Opus → G.722	21.71	4.63
Opus → G.711	22.03	3.86
G.722 → Opus	25.67	4.75
G.711 → Opus	26.16	4.00
G.722 → G.722	21.11	4.71
Opus → Opus	27.27	4.75

To consolidate the results, the tests have been repeated several times. We carried out the tests with various transcoding use cases (up to 4 calls at the same time).

3 Audio quality aspects in public mobile communication networks (4G)

The study in [7] deals with various influencing variables for a conceptual design of a VoLTE-based network (e. g. the used audio codec, end-to-end delay, acceptable packet loss rate). The author proposed different end-to-end-delay budgets, depending on the used network elements (only 4G or mixed with 2/3G) or on the transcoding impact. Figure 3 presents the relation between the E-model rating R value to the mouth-to-ear (m2e) delay for the AMR-NB codec modes used in VoLTE calls (no transcoding).

Assuming the worst case codec scenario – i. e. usage of the poorest quality codec (AMR-NB 4.75 kbps) – the available delay budget amounts to approx. 220 ms (cf. dashed line in Figure 3).

The second frequently used voice codec in VoLTE call scenarios, AMR-WB, achieved better m2e-delay values (< 200 ms) by the minimal target R value of 70.

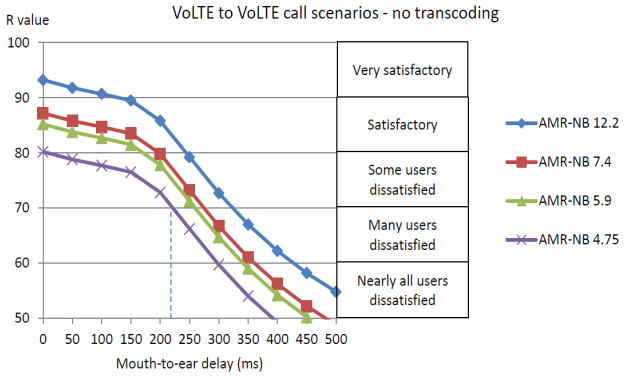

Figure 3. Correlation between R value and delay [7]

4. Future research scope

Currently, two common full band audio codecs coexist – the 3GPP-standardized voice codec EVS beside the IETF-driven open source codec Opus. A quality evaluation study from Anssi Rämö et al. shows that both codecs can achieve a similar performance [11]. Solely in the lower bitrate range (< 24 kbps), EVS provides better results, which is reasonable for coding technics adopted from cellular networks. Otherwise, the Opus codec comes up with a lower processing delay than the EVS (Opus delay can be up to 8 ms shorter in some constellations) but from the current viewpoint it is unclear whether this aspect has a verifiable impact on the overall (end-to-end) delay in communication networks – respectively, on the measures of the quality of service (QoS) or quality of experience (QoE).

A WebRTC-integrated test scenario to compare both, EVS and Opus codec, is still missing in international publications. Beyond, there is a test lack between both codecs in telecommunication carrier networks considering real-world conditions.

For future telecommunication networks of the Fifth Generation (5G) we expect that the source bit rate requirements for full band voice calls are practicable [12], and we rather see a challenge in the processing delay time of codecs.

Acknowledgment

We would like to thank SwissQual AG (a Rhode & Schwarz Company), in particular our colleague Jens Berger, for supplying the POLQA testbed and for fruitful discussions.

References

[1] C. Jennings, A. Narayanan, D. Burnett, and A. Bergkvist, "WebRTC 1.0: Real-time communication between browsers", W3C, W3C Working Draft 28 January 2016. https://www.w3.org/TR/webrtc/

[2] J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus audio codec", RFC 6716, IETF, September 2012. http://www.ietf.org/rfc/rfc6716.txt

[3] M. Maruschke, O. Jokisch, M. Meszaros, and V. Iaroshenko, "Review of the Opus codec in a WebRTC scenario for audio and speech communication", Proc. of 17th Intern. SPECOM Conference, Sept. 20–24, 2015 Athens, Greece, pp. 348–355, Springer Lecture Notes in Artificial Intelligence LNAI 9319. (ISBN: 978-3-319-23131) http://link.springer.com/chapter/10.1007%2F978-3-319-23132-7_43

[4] ITU-T, "Methods for objective and subjective assessment of speech quality (POLQA): Perceptual objective listening quality assessment," International Telecommunication Union (Telecommunication Standardization Body), REC P.863, September 2014. http://www.itu.int/rec/T-REC-P.863-201409-I/en

[5] O. Jokisch, M. Maruschke, M. Meszaros, and V. Iaroshenko, "Audio and Speech Quality Survey of the Opus Codec in Web Real-time Communication", Proc. 27th Conference on Electronic Speech Signal Processing (ESSV), March 2–4, 2016, Leipzig, pp. 254–262. (ISBN 978-3-95908-040-8) http://www1.hft-leipzig.de/ice/essv2016/files/31%20-%20JokischMaruschke-S.254-262%20(Abstract).pdf

[6] M. Meszaros, and M. Maruschke, "Verhaltensanalyse von Einplatinencomputern beim Transcoding von Echtzeit-Audio-daten", Proc. 27th Conference on Electronic Speech Signal Processing (ESSV), March 2–4, 2016, Leipzig, pp. 237–245 (in German, ISBN 978-3-95908-040-8) http://www1.hft-leipzig.de/ice/essv2016/files/20%20-%20MeszarosMaruschke-S.237-245%20(Abstract).pdf

[7] T. Sekulski, "VoLTE solution concept for an international mobile network operator", Bachelor thesis, HfT Leipzig, Germany, July 2015.

[8] T. Fakhouri, „Vergleichende Sprachqualitätsuntersuchungen an ATM/TDM- basierten und IP-basierten Media Gateways", Bachelor thesis, HfT Leipzig, November 2015 (in German).

[9] 3GPP TS 26.441, EVS Codec General Overview, 3GPP, August 2014. http://www.3gpp.org/DynaReport/26441.htm

[10] Doubango Telecom, "Webrtc2sip – Smart SIP and media gateway to connect WebRTC endpoints", 2015. http://webrtc2sip.org/

[11] A. Rämö, and H. Toukomaa, "Subjective quality evaluation of the 3GPP EVS codec", Proc. ICASSP Conf., April 19-24, 2015, Brisbane, Australia, pp. 5157–5161.

[12] R. El Hattachi (Ed.), "5G White Paper", February 2015. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf