AUDIO AND SPEECH QUALITY SURVEY OF THE OPUS CODEC IN WEB REAL-TIME COMMUNICATION

Oliver Jokisch, Michael Maruschke, Martin Meszaros and Viktor Iaroshenko

Hochschule für Telekommunikation Leipzig (HfTL)
{jokisch; maruschke}@hft-leipzig.de

Abstract: The Opus codec is used in several applications fields of speech and audio communication. This article describes the instrumental quality assessment of Opus-coded speech in a web browser-based real-time communication using POLQA and AQuA method. Furthermore, we tested Opus with mixed vocal and music signals and also performed a perceptual test. WebRTC framework-coded speech achieves a similar MOS assessment compared to standalone coding. The observed degradations depend on signal bandwidth, on variations in speech (e. g. by emotions) or in music (vocal vs. instrumental) and on the assessment method.

1 Introduction

Different kinds of audio and speech codecs for various applications including wired or mobile communication, broadcasting, audio and video entertainment, gaming etc. have been established in the last decades. The research targets include near-to-natural audio or speech quality but also low bandwidth and low calculation complexity requirements.

During the last years, the multifunctional Opus codec for speech and audio communication was integrated in several applications fields such as browsers for the Web Real-Time Communication Web Real-Time Communication (WebRTC). This contribution deals with the instrumental and perceptive assessment of the Opus codec implementation in WebRTC. The main research question is whether the WebRTC operation conditions influence the Opus performance. Beside of this we test the Opus in special conditions which are not defined by the standards such as singing voice or mixed music-voice signals. The implementation issues of the Opus codec in the WebRTC environment of Google browser are surveyed and discussed in a further contribution [1]. The current contribution is focused on the audio and speech quality assessment of the Opus codec within the the real-time communication mode.

We tested two instrumental assessment methods – the Perceptual Objective Listening Quality Assessment (POLQA) with regard to the ITU-T P.863 recommendation and the non-standardized Audio Quality Analyzer (AQuA) from Sevana company. Both methods incorporate a perceptual model to predict human speech quality decisions by means of digital signal analysis. Consequently, there are regarded as objective measures. The prediction of the instrumental measures should come as close as possible to quality scores from a listening test with human probands which are considered as subjective results. With regard to a widely used absolute category rating, usually, the Mean Opinion Score (MOS) is predicted. As test stimuli, both natural reference speech and degraded speech output e. g. of the telephony network is required. For comparison we included a test with conventional G.711 codec. To get a broader view on the audio and speech quality assessment of the Opus codec in real-time communication, we additionally performed a listening test with 26 probands including a representative excerpt of the instrumental test database. Following an overview about our experimental setup, we discuss the instrumental and perceptual assessment results of the Opus codec.

2 Opus Codec Implementation in the Web Real-Time Communication

The Web Real-Time Communication (WebRTC) extends regular web browsers by additional communication functionality whereby browsers can directly interact with each other. This open source technology – initiated by Google Inc. – offers bi-directional High Definition (HD) audio and video transfer without any additional installation requirements for end users. The WebRTC technology is still under development and does not request any particular signaling protocol. The necessary Real-Time-Communication (RTC) function is provided in common web browsers like Google Chrome, Mozilla Firefox and Opera by default, i. e., many systems including desktop and tablet computers, laptops or smartphones can easily use web-based real-time communication.
Our main research questions deals with the potential limitations of the Opus codec performance caused by the WebRTC implementation or potential restrictions in the operation mode of Opus.
The Opus codec characteristics are described in detail in [2]. The WebRTC codec implementation is extensively discussed in [1]. By implementing logging functions, following default Opus codec parameters were detected:

1. Sample rate: 48 kHz,
2. Audio bandwidth: Fullband (FB),
3. Used encoder bit rate: 32 kbit/s,
4. Used channels: 1 (mono),
5. Opus working mode: CELT and Hybrid,
6. Frame duration: 20 ms,
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 end devices and value 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 by the SDP-based session description. The Opus working mode depends on the used encoder bit rate while the bit rate tightly depends on the sample rate. While analyzing the internals of the Opus encoder function, another important encoder_state parameter (*voice_ratio*) – which directly influences the current working mode – can be observed. *Voice_ratio* means the probability (in %) of being a voice signal in the given audio input.
The Figure 1 shows a concatenated voice/music example which consists of the German utterance "Die roten Äpfelchen sind für.." (The little red apples are for..) and a piece of Ska music from "Take one down" (Sounds Like Chicken, 2004). Regarding a dynamically modified working mode, the *voice_ratio* threshold for a switch-over from a voice-optimized mode to a non-voice (music) favored mode is about 25 %. Vice versa, from music to voice mode one can observe a threshold of about 50 %.

3 Audio and Speech Quality Assessment

3.1 Evaluation targets and test database

A potential audio and speech quality degradation can be caused by noise, silent phases, limitations of low bit-rate encoding, network errors (e. g. packet loss), additional effects like echos, delay,

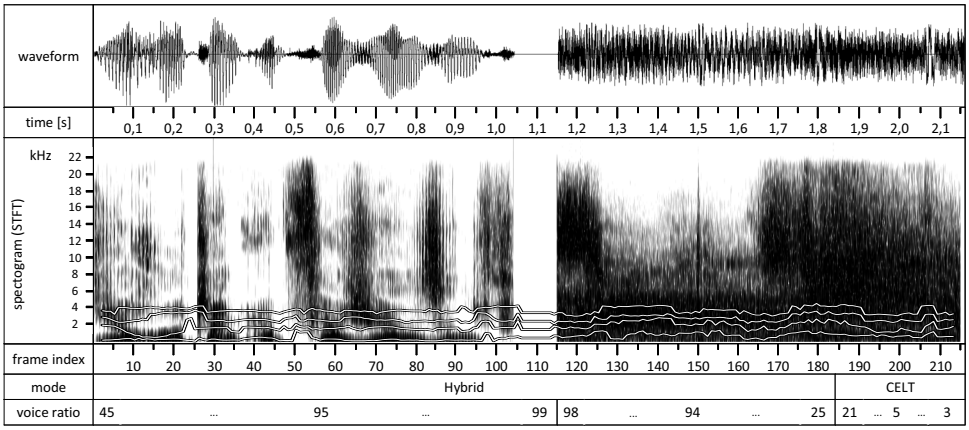

Figure 1 - Seamless switch-over in Opus working mode: audio stream excerpt with voice part left and music part right (frame length 10 ms, spectrogram including formant curves).

jitter or suboptimal communication terminals. An accurate measurement of end-to-end quality is essential to a wide range of communication systems and algorithms – beyond the testing of codecs. A main criterion is the consistency with the mean assessment of human decisions (listening test).

According to quality requirements and system resources there are different bandwidth options in speech signal and music processing (cf. Table 1).

Table 1 - Abbreviations for signal bandwidths

abbreviation	meaning	pass-band	quality expectation
NB	narrowband	0.3 . . . 3.4 kHz	telephone
WB	wideband	50 Hz . . . 7 kHz	AM radio
SWB	superwideband	50 Hz . . . 14 kHz	FM radio
FB	fullband	20 Hz . . . 20 kHz	CD quality

The design of the test database aims at different testing conditions:

- Natural read speech examples (fullband, studio recording) from the EUN-Veith database (part 1a) to serve a best case scenario. In this case, an excellent codec performance is expected. Possible degradation might be network-based.

- Wideband speech involving emotional and neutral speech (part 1b). In this scenario actors simulate different emotions which creates additional challenge to encoder and evaluation methods. In the resulting test setup an additional signal degradation is expected which let us consider the scenario as a kind of ”normal case”.

- Fullband music and singing voice examples (part 2). In this scenario we are going to examine the limitations of the instrumental and perceptual assessment. Whereas Opus specification includes mixed operation on speech and music signals, all assessment methods used in this contribution are restricted to speech signals only. Nevertheless, it seems obvious that special signals – such as singing voice – might be partly covered by the perceptual models of the instrumental and the subjective (human) evaluation.

The instrumental assessment requires a combination of two speech samples (minimum length 2.4 s, SNR 45 dB) and pauses (silence) of 2 s before, in between and afterward. We randomly selected according German samples from two different speech databases by considering gender balance.
The music and singing voice examples (English, German and Russian) are constructed in the same way. We selected different rhythmic and tonal styles including percussion parts and loudness variations.
The resulting instrumental test database contains 81 audio files :

- Fullband (FB) 30 combined Veith utterances from the language learning database Euronounce – EUN Veith [3] of 5 male and 5 female speakers (part 1a),
- Wideband (WB) 36 combined speech utterances with acted emotional and neutral speech from the German Database of Emotional Speech – Emo DB [4], 1 male and 1 female (part 1b),
- FB 4 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).

3.2 Instrumental assessment via POLQA method

The Opus codec supports Narrowband (NB), WB, Super-Wideband (SWB) or FB speech and operates in following three modes [5], [6]:

- SILK mode (NB and WB speech),
- CELT mode (music, SWB and FB speech),
- Hybrid mode (SILK and CELT synchronously for SWB and FB speech).

According to ITU-T P.863 [7], Perceptual Objective Listening Quality Assessment (POLQA) is defined as a quality assessment system for NB and SWB speech that does not cover the FB option – provided by the Opus codec. For lack of a more adequate method, we choose the standardized POLQA assessment which predicts, inter alia, the Mean Opinion Score (MOS) of ordinary listeners from 1 ”Bad” till 5 ”Excellent”. POLQA is based on the predecessor method Perceptual Evaluation of Speech Quality (PESQ) [8] and offers two operational modes – optimized either for NB or SWB signals. Both, reference and degraded signal must have the same sample rate. In super-wideband operational mode, the algorithm requires a SWB signal (mono, sampling frequency 48 kHz) with no bandwidth limitation, no additional equalization between 50 Hz and 14 kHz and no significant energy outside the spectral limits. If the sample rate differs from 48 kHz in chosen SWB mode, the signal will be converted internally to 48 kHz before processing it further. The use of reference signals with bandwidth narrower than 14 kHz in SWB mode will lead to wrong results on the Mean Opinion Score (MOS) scale. POLQA is not surveyed or standardized for special configurations like singing voices yet.

3.3 Instrumental assessment via AQuA method

AQuA was introduced 2009 as an alternative for existing audio quality assessment models such as PESQ (ITU-T P.862) [8] by the company Sevana [9]. It can be utilized in VoIP, PSTN, ISDN, GSM, CDMA or LTE/4G networks and has a competitive computational performance. As in PESQ or POLQA, the MOS value can be predicted.

The developers of AQuA claim that the method is more suitable for certain challenges in end-to-end evaluation such as noise, packet losses, variable delays or loudness variation. Furthermore, AQuA Non-intrusive (since 2010) does not require a audio reference. As a matter of a fact, AQuA is not within the ITU standards, and the psychoacoustic model is not published. We did not find a scientific survey which verified those benefits of AQuA. Consequently, we consider the method as a blackbox approach for the comparison of our results of the POLQA method and the perceptual test.

3.4 Perceptual assessment

For restricting the test effort while assuring compatibility with the instrumental tests and keeping the intended variation, we randomly selected samples from the according parts of the test database:

- 6 of 30 combined speech samples from part 1a (FB),
- 11 of 36 combined speech samples from part 1b (WB),
- 13 of 15 mixed music and singing voice samples from part 2 (FB).

We assured a balanced selection for all testing modes (neutral speech and emotions) including gender-balance. The selected examples – originally constructed for the instrumental test, i. e. consisting of two part including 2 s silence – have been presented via multimedia loudspeakers and perceptually evaluated. To create comparability with the instrumental POLQA and AQuA tests, we did not proceed any additional signal manipulation such a loudness normalization. At the beginning, 3 speech and 2 music samples were provided as a quality reference without assessment. The following listening test contained in total 30 examples and was conducted by 26 local students and staff (19 males, 7 females), aged 29.3 ± 11.6 years including 5 probands above 40. In general, our probands were not familiar with speech and language processing or auditive tests. The listeners were presented all samples only once, and they were asked for their spontaneous opinion on a five-point scale from 1 (bad) to 5 (excellent). The overall test lasted about 15 min. Afterwards, the mean of the opinion scores was calculated (MOS test).

3.5 Experimental setup

Figure 2 illustrates the measurement setup for the instrumental POLQA method as an example compared to the perceptive alternative by a listening test. The instrumental AQuA method is applied analogously. The file-based Opus-encoded reference signal (audio or speech example, step 1) is transmitted via WebRTC network from client A to client B (step 2). For simplification, we only test this unidirectional communication although WebRTC is designed for bidirectional processing. The Opus-decoded signal is then stored in a file at client B (step 3). During encoding, transmission and decoding, the sound is potentially degraded. Afterwards both, reference and degraded sound file are fed to the POLQA evaluation tool (step 4a). This procedure was repeated for all 81 sample files. In Addition, we performed the described listening test with the selected 30 files from the instrumental assessment (step 4b).

4 Experimental Results and Discussion

4.1 Previous studies on Opus codec (standalone)

Figure 3 schematically illustrates the speech/sound quality of various codecs as a function of the bitrate cf. [10]. It summarizes results from diferent listening tests including anecdotal

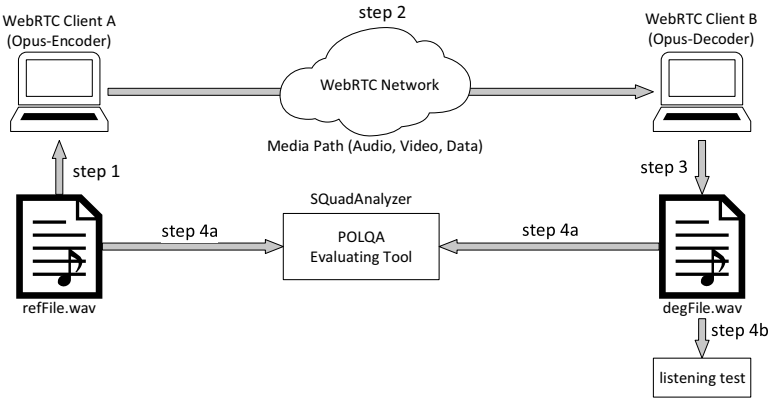

Figure 2 - Experimental setup including the instrumental POLQA method in comparison to the perceptive method (listening test)

evidence and without exact values. One of mentioned studies from Rämö and Toukomaa [11] searched for optimal operating points based on the listening test results and clarified that Opus codec's LP mode provides useable voice quality at quite competitive bitrates compared to the codecs AMR or AMR-WB while facing two issues – the highly variable bitrate which may cause problems depending on the transmission network and the changing signal bandwidth with time. For Opus WB quality significantly higher bitrates than in AMR-WB are needed (around 14.5 kbit/s at minimum). Furthermore, CELT based MDCT mode provides a good alternative to ITU-T G.722.1C or G.719 by providing better quality but with more computational complexity (where that can be supported). Beyond, the Hybrid mode provides excellent voice quality at bitrates from 20 to 40 kbit/s.

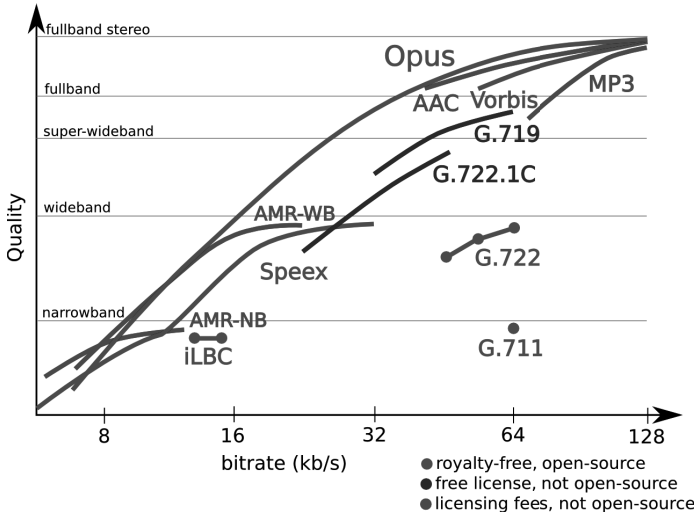

Figure 3 - Schematic representation of codec quality from previous studies from [10]

4.2 Reference test with G.711 codec

Beside Opus codec, the conventional NB audio codec G.711 (PCM a/μ law) is implemented in the browsers' RTC function and serves as a reference for our Opus codec assessment. As a baseline test, we performed the WebRTC measurement setup (steps 1 ... 4a, cf. setup Figure 2) by using G.711 codec instead. On the FB speech data (EUN Veith database, part 1a, 30 samples) and with POLQA (v2) method the average assessment resulted to MOS*=4.39.

4.3 Instrumental results

Table 2 summarizes the MOS*[1] results of POLQA (v2) and AQuA across different test setups.

Table 2 - Predicted MOS* – average (tested samples)

no. (part)	band	assessment method		test data
		POLQA v2	AQuA	
30 (1a)	FB	**4.68**	4.98	EUN (all)
15 (1a)	FB	4.73	4.97	EUN (male)
15 (1a)	FB	4.64	4.99	EUN (female)
36 (1b)	WB	4.44	4.63	Emo (all)
18 (1b)	WB	**4.60**	4.55	Emo (neutral)
18 (1b)	WB	**4.29**	4.71	Emo (emotional)
6 (1b)	WB	4.30	4.95	Emo (anger)
6 (1b)	WB	4.29	4.66	Emo (happiness)
6 (1b)	WB	4.27	4.52	Emo (fear)
11 (2a)	FB	**3.81**	4.60	vocal & instr.
4 (2b)	FB	**3.90**	4.61	instrum. music

In the best case scenario – based on read FB speech from database 1a – POLQA predicts MOS* values of 4.64 (female samples) ... 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 with MOS* values of 4.27 (emotional speech) ... 4.60 (neutral speech) but the differences are mainly emotion-based. FB music shows a strong degradation of about 0.8 on MOS scale compared with FB speech whereas the differences between partly vocal and strictly instrumental music are not significant ($< \pm 0.1$) considering the low number of samples (only 4 in database 2b). With regard to the previously studied WebRTC operation CELT/Hybrid and competitive codec parameters (bitrate 32 kbit/s 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 AQuA-generated MOS* values are generally higher than the POLQA ones (except for neutral WB speech) and the low degradations in different test setups is not plausible – in particular in comparison to the perceptual measures in the next section. The samples of anger score 0.4 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).

4.4 Perceptual results

There was no significant assessment difference between native and second language speakers ($\Delta MOS < \pm 0.1$ across our samples) or between male and female probands whereas an age

[1] The marker * refers to the fact that this Mean Opinion Score is based on instrumental prediction.

influence could be observed (for several cases ΔMOS about ± 0.25) – the 5 listeners above 40 years rated most of the samples higher as a rule. Table 3 shows the averaged MOS results in the test parts 1a, 1b and 2. The assure the reliability of a perceptual test, minimum of 15 ... 40 listeners is normally required. Therefore, the fourth column for the listener group 40+ is given for illustration only.

Table 3 - Perceptual test: MOS – average (listened samples)

no. (part)	**band**	**listener group (age)**			**test data**
		all	< 40	≥ 40	
6 (1a)	FB	**4.00**	3.98	4.07	EUN (all)
3 (1a)	FB	3.88	3.91	3.73	EUN (male)
3 (1a)	FB	4.12	4.05	4.40	EUN (female)
11 (1b)	WB	**3.64**	3.55	4.02	Emo (emotional)
3 (1b)	WB	3.68	3.55	4.20	Emo (anger)
4 (1b)	WB	3.80	3.69	4.25	Emo (happiness)
4 (1b)	WB	3.46	3.42	3.65	Emo (fear)
9 (2a)	FB	**3.47**	3.42	3.69	vocal & instr.
4 (2b)	FB	**2.96**	2.94	3.05	instrum. music

The perceptual MOS values in Table 3 are generally lower than POLQA-predicted MOS* ones in Table 2. 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 Opus in being a multifunctional and highly adaptive codec. Furthermore, the partial results in vocal music indicate that an assessment via POLQA can be applied to a certain extent for singing voices too. The low MOS of 2.96 vs. predicted MOS* of 3.90 in instrumental music illustrates that POLQA is not appropriate in such test cases. 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 3 examples only cf. Table 3. Beyond, the data sets were manually selected and potentially biased to noticeable coding examples.

5 Conclusion

We tested prototypical cases of Opus coding and its assessment in a WebRTC framework. In the instrumental assessment via POLQA, the framework-coded speech achieves a similar MOS* up to 4.73 as in standalone coding – compared to 4.39 (G7.11). 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 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. The AQuA results (generated in analogous manner) are widely not plausible and will not be followed up.

Acknowledgment

We would like to thank SwissQual AG, Switzerland (a Rhode & Schwarz Company) for supplying the POLQA testbed and for the fruitful discussion – in particular Dr. Jens Berger. Further thanks goes to Valeri Sitnikov from Sevana Oy Ltd., Finland for providing the AQuA tools.

References

[1] M. MARUSCHKE, O. JOKISCH, M. MESZAROS, and V. IAROSHENKO, "Review of the opus codec in a webrtc scenario for audio and speech communication," in *Speech and Computer, 17th International Conference, SPECOM, Proceedings on*, Springer International Publishing, vol. 9319, 2015, pp. 348–355.

[2] J. VALIN, K.VOS, and T. TERRIBERRY, *Definition of the opus audio codec*, RFC 6716 (Proposed Standard), Internet Engineering Task Force, Sep. 2012. [Online]. Available: http://www.ietf.org/rfc/rfc6716.txt.

[3] O. JOKISCH, A. WAGNER, R. SABO, R. JAECKEL, N. CYLWIK, M. RUSKO, A. RONZHIN, and R. HOFFMANN, "Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system," *Proc. of 13th Intern. Conf. SPECOM, St. Petersburg*, pp. 515–520, 2009.

[4] F. BURKHARDT, A. PAESCHKE, M. ROLFES, W. F. SENDLMEIER, and B. WEISS, "A database of german emotional speech.," *Proc. of Interspeech 2005, Lisboa*, pp. 1517–1520, 2005.

[5] J.-M. VALIN, G. MAXWELL, T. TERRIBERRY, and K. VOS, "Voice coding with opus," *AES Convention*, Oct. 2013.

[6] ——, "High-quality, low-delay music coding in the opus codec," *AES Convention*, Oct. 2013.

[7] ITU-T, "Methods for objective and subjective assessment of speech quality (polqa): perceptual objective listening quality assessment," International Telecommunication Union (Telecommunication Standardization Sector), REC P.863, Sep. 2014. [Online]. Available: http://www.itu.int/rec/T-REC-P.863-201409-I/en.

[8] ——, "Methods for objective and subjective assessment of quality perceptual evaluation of speech quality (pesq): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union (Telecommunication Standardization Sector), REC P.862, Feb. 2001. [Online]. Available: http://www.itu.int/rec/T-REC-P.862-200102-I/en.

[9] SEVANA OY LTD. (Jan. 2016). AQuA: An Audio Quality Assessment tool, [Online]. Available: http://sevana.biz/products/aqua/call-quality-monitoring/.

[10] J. VALIN, K.VOS, and T. TERRIBERRY. (Dec. 2015). Codec landscape, [Online]. Available: http://www.opus-codec.org/comparison/.

[11] A. RÄMÖ and H. TOUKOMAA, "Voice quality characterization of ietf opus codec," *Interspeech 2011; Florence; Italy*, pp. 2541–2544, Aug. 2011.