https://doi.org/10.26552/com.C.2016.1.17-22

Yusuf Cinar - Hugh Melvin - Peter Pocta *

A BLACK-BOX ANALYSIS OF THE EXTENT OF TIME-SCALE MODIFICATION INTRODUCED BY WEBRTC ADAPTIVE JITTER BUFFER AND ITS IMPACT ON LISTENING SPEECH QUALITY

WebRTC is an open-source platform for real-time communications over the web and has been experiencing widespread adoption in recent years. WebRTC clients employ the technique of time scaling of packets to cope with the impact of network jitter and/or clock skew. A black-box study presented in this paper focuses on two aspects, namely time scale modification behaviour under different packet arrival interval and its impact on the listening quality perceived by the end user. Specifically, we examine the MOS scores predicted by the POLQA speech quality prediction model. Our tests involve both iSAC and Opus codecs, two of the widely used WebRTC codecs. In the experiment, a speech file played from one client, is directed through a network simulation before reaching the receiving client. Our results surprisingly show that the extent of time scaling is consistently higher for Opus producing shorter speech files. Regarding the consequent impact on quality, we also find that there are many cases where POLQA is reporting MOS predictions that contradict expert listener assessments.

Keywords: *WebRTC, jitter, adaptive playout, time-scale modification.*

1. Introduction

WebRTC is one of the latest developments in the area of multimedia real-time communications (RTC) and it is a set of standards from WC3 [1] and IETF [2] that enables real-time communication on the web. WebRTC has drawn significant interest from not only browser vendors but also application and web developers due to the potential new services it can offer [3]. There is an open source project, with the same name WebRTC [4], which implements the standards and is used by the browser vendors and application developers. Traditional VoIP components, such as audio coding modules, jitter buffer, play-out decision and codec implementations are integral parts of WebRTC project.

There are numerous dynamics that can influence the quality experienced by the end user for a voice call session. The non-deterministic nature of best-effort Internet causes many network impairments, such as network delay variations (jitter) and consequent packet bursts, as well as packet loss. In particular, jitter results in voice packets arriving at irregular intervals to the receiver. In order to maintain the speech intelligibility and quality for the listener, the voice stream must be reconstructed in a similar manner to which it was created. Due to the presence of network jitter and congestion, the voice packets are typically held in the receiver jitter buffer before they are played out in a way that sustains the conversational and listening quality at a certain level. WebRTC has a component called NetEQ for this purpose [4] describes NetEQ as: "A dynamic jitter buffer and error concealment algorithm used for concealing the negative effects of network jitter and packet loss. It aims to keep latency as low as possible while maintaining the highest voice quality." As a by-product, it also deals with clock skew issues between sender and receiver clients.

When playing out the speech from the jitter buffer, the receiver's playout strategy tries to adapt to the changing network conditions. Adaptive playout techniques are grouped in two categories, per packet and per talkspurt, as mostly referenced in the literature [5] and [6]. Per talkspurt techniques apply adjustments only to silence periods between talkspurts. However, the latter technique, which is also referred to as time scale modification, applies compression or expansion to all the packets regardless of silence or voice segment.

Literature to date had focused mostly on the impact of the degradations caused by per-talkspurt playout strategies on the listening quality. A comprehensive study was conducted by some of the authors in [5], and has shown that small playout

* [1]Yusuf Cinar, [1]Hugh Melvin, [2]Peter Pocta
[1]Discipline of Information Technology, College of Engineering & Informatics, National University of Ireland, Galway, Ireland
[2]Dept. of Telecommunications and Multimedia, Faculty of Electrotechnical Engineering, University of Zilina, Slovakia
E-mail: cinar.yusuf@gmail.com

delay adjustments of 30 ms or less introduced to silence periods is negligible according to subjective listening quality scores and POLQA predictions. On the other hand, the PESQ model predicts contradicting scores to the scores obtained from the subjective test; hence they conclude PESQ fails to correctly predict quality scores with those adjustments. It is worth noting here that [5] does not study the impacts of per packet time-scaling modifications.

The fundamental idea of adaptive playout mechanism via time scale modification was first introduced separately by Liu et al. [7] and Liang et al. [8] in 2001, according to [9]. A further study by Liu et al. [9] in 2002 investigated the stretching-ratio transition effect on perceived audio quality by measuring the objective PESQ MOS, and found that PESQ, which was the most up to date ITU standard in the area of objective speech quality assessment at the time, does not provide a good objective quality measure for packet-based time-scale modified speech signals. However, this study was carried out more than a decade ago and doesn't cover more recent methods such as POLQA.

In [10], a more recent study has shown via extensive experiments that POLQA can predict the quality with high accuracy for different sampling rate adjustments, (referred to as time scale modification in [11]), if the total range of sample rate deviations is +/-3% of the nominal sample rate. It is also claimed here that POLQA predictions start failing after 6% but no data was presented to support that. There is also no detailed description of how the time scale modification was applied to the speech signals. Firstly, although not explicitly mentioned, it is implicitly indicated that a constant rate between -3% and 2.9% of sampling rate difference is applied to the entire speech which in a real world scenario is not likely. For instance, our experiments show that WebRTC applies a variable rate of time scaling per speech frame. Liu et al. [9] use the term 'stretching dynamic' to describe how flexible the time scale modification ratio can change from one segment to the next, within constraint that audio quality is not sacrificed. Secondly, a simple change in the sampling rate doesn't reflect the true nature of time scale modification that is typically employed in VoIP as can be seen in [7] and [8]. In our experiments, we study simulated WebRTC VoIP calls, where variable rates of time scale modification are applied by the playout algorithm, which better reflects the reality.

It is also worth noting that subjective listening tests [12] or objective methods such as PESQ [13] and POLQA [10] do not consider the effect of mouth-to-ear delay. Hence, the ITU-T E-model [14] is also an important method of validation since it provides a direct link to perceived conversational speech quality by estimating user satisfaction from the combined effect of information loss, delay and echo. This method has been widely used by researchers including one of the authors [11] to evaluate playout buffer schemes.

Based on the above summary, we conclude that regarding per-packet algorithms, aforementioned gaps exist. Therefore, our objective here is to examine the quality impacts of WebRTC's time scaling algorithm, under different codec settings and network conditions.

1.1. Research motivation

The literature to date has not examined the performance of time scale modification of adaptive jitter buffer algorithm employed by the NetEQ component of WebRTC project. This prompted us to undertake this study as WebRTC is available to billions of devices [15].

In advance of the research, there was little published about the internal workings of WebRTC' adaptive jitter strategy. Although several points were made by Hines et al. [16], such as that it accommodates time scale modifications, we believed more characterisation would help further studies.

Furthermore, because WebRTC supports many voice codecs, we wished to see the quality implications of choosing one codec over another when used with same time scale modification technique at the adaptive jitter buffer algorithm. To our knowledge, a comparative performance analysis of playout buffer algorithm coupled with codec has not yet been conducted.

In the context of research motivation above, the following key research questions are identified:

1. What are the core characteristics of the time scaling adaptive jitter buffer algorithm employed in WebRTC?
2. What impact does time scale modification have on the perceived listening quality – both via objective and expert subjective evaluations?

This paper is organised as follows. Section 2 outlines the experiment methodology. Results are presented in Section 3. Section 4 concludes the paper and suggests future work.

2. Methodology

2.1. Testbed Description

The test-bed, illustrated in Fig. 1, involves a sender and receiver side of WebRTC communication chain and an emulated network channel in between, all hosted on a single host machine. Sender represents a speaker of a virtual conversation while receiver represents a listener. The channel emulates the IP network carrying one-way voice packets from the sender to receiver. Our objective was to simulate severe network jitter leading to packets bursts arriving at the receiver, and examine the associated jitter buffer response.

An important note is that the default WebRTC jitter buffer can hold up to 50 packets, i.e. a maximum buffer size at the receiver is hardcoded to 50. Since our simulation results in extreme packet bursts, the buffer reaches its threshold quickly and starts dropping packets. Hence we modified WebRTC's jitter buffer maximum capacity to 500 packets to avoid dropping packets. This allows us to focus on time scaling only.

As illustrated in Fig. 1, the sender reads a PCM file, which contains speech ranging from 8 to 10 seconds including silent periods, and encodes and emits the data to the network channel in the form of RTP packets. The experiment is designed to accumulate all the packets generated in the network channel before dispatching them to the receiver.

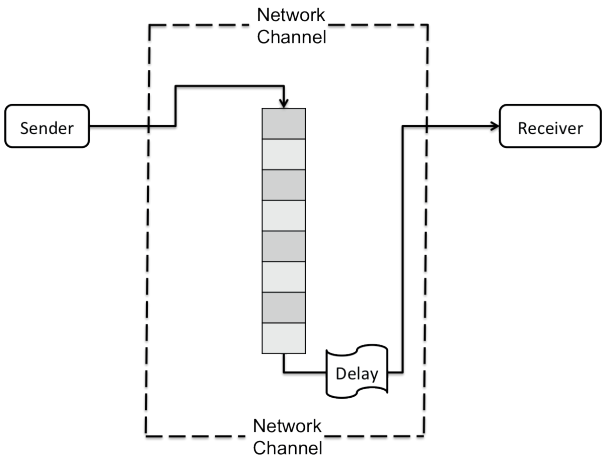

Fig. 1 Simulation Design

Once all the packets are in the channel, i.e. the representation of network in our application, we then dispatch each packet according to the delay profiles of the experiment. This is done by a Delay component displayed in Fig. 1. For instance, one delay profile is to dispatch each packet to the receiver at an interval of 15 ms.

The test is automated end to end and there are several configurations, i.e. test conditions, which can be controlled. As shown in Table 1, we control the packet arrival interval to the receiver buffer. Also, a codec can be automatically switched between Opus and iSAC. The WebRTC project includes many codecs including but not limited to Opus, iSAC, G.711, G.722. In our experiment, iSAC and Opus were only deployed as they represent currently the mostly deployed codecs in WebRTC. Regarding the packet arrival to the receiver, the following values, 10, 15, and 20 ms, were used in our experiments. However, in real world scenarios, these values would not be constant as each packet delay may vary. Therefore, we plan to run a further study involving real world measurements. It is worth noting here that iSAC uses speech frames of length of 30ms and the WebRTC Opus implementation deploys frames of length 20 ms by default. The values 10 and 15 ms were chosen as they represent 50% of the default packetisation for Opus and iSAC respectively, whereas 20 ms was chosen to see the impact for Opus under ideal conditions.

Test Conditions Table 1

Configuration	Values
Arrival interval	10 | 15 | 20 (ms)
Codec	Opus | iSAC

2.2. Speech Quality Assessment

In order to assess how time scale modification impacts on listening quality experienced by the end user of WebRTC, we used both an objective method based on perceptual modelling and expert subjective listening. The objective method employed is POLQA version 1.1 (Perceptual Objective Listening Quality Assessment, ITU-T Recommendation P.863) [17]. We then compared and contrasted the POLQA MOS scores with an expert listener.

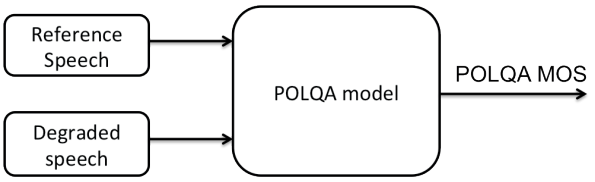

Fig. 2 Objective Listening Quality Tests

As the incoming packets are played out at the receiver, the output is recorded into a file for quality assessment. The reference speech file and recorded one (degraded speech file) are together passed into executable that represent POLQA prediction model to generate MOS scores for the corresponding speech sample as depicted in Fig. 2.

2.3. Speech Samples

We used 25 certified speech samples in the experiments. The samples are taken from SWB conformance databases included in the ITU Recommendation P.863 [17]. Speech files consist of a pair of utterances from 14 male and 11 female speakers with a pause in between. Durations of the speech files are between 8 and 10 seconds and stored in 16 bits, 48000 Hz linear PCM.

3. Results

In this section, we present results and analysis of the black box experiments carried out as well as observations discovered related to the voice aspects of WebRTC. Results can also be found on a website [18] with further details including interactive charts and waveforms.

3.1. General Analysis

We use the term acceleration rate to refer to the duration difference between original speech and degraded speech. We calculate the acceleration rate according to:

$$Ar = \left(1 + \frac{Td}{To}\right) * 100$$

where *Ar* is the acceleration rate, *Td* is duration of degraded speech file and *To* is the duration of original speech file.

Fig. 3 and Fig. 4 separately show the acceleration rates for two of the samples used in experiments. Most of the acceleration rates are close to the results in Fig. 3, lying between 15% and 25% for a packet arrival of 10ms and 15ms. However, there are some samples, which produced extremely low acceleration rates, as shown in Fig. 4, i.e. between 3% and 5% for the same packet arrival configurations.

One key finding is that the Opus codec consistently accelerates (scales) more than iSAC for packet arrival intervals tested of 10 ms or 15 ms as can be seen in Fig. 3. With a packet arrival interval of 20 ms, we expected not to see any acceleration as the default Opus packets in WebRTC are 20 ms. Interestingly, results show that Opus applies acceleration, which shortens the playout duration by about 2% to 3%, for all of the 25 samples used in the experiments. Figure 4 shows an example for one of the samples. It can be clearly seen from Fig. 4 that the acceleration rate is 2% at 20 ms arrival rate.

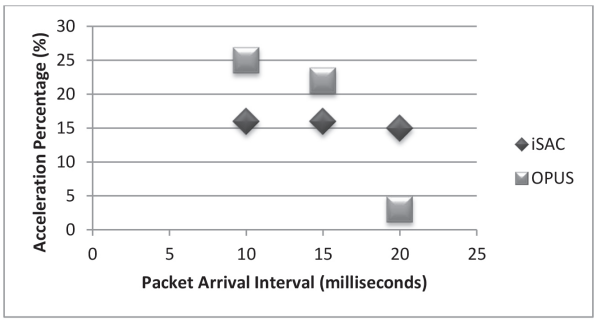

Fig. 3 Acceleration Percentage vs Packet arrival interval - file: CH_m1_s4_file_18.48k.pcm

An investigation of the time scale modification code of NetEQ algorithm in WebRTC showed us that it is codec agnostic. Therefore, it is an interesting outcome that Opus accelerating % is higher than iSAC for 10 and 15 ms arrival interval. This is especially interesting as iSAC has a larger packet size of 30 ms and thus 10/15 ms arrival interval represents a more extreme network burst scenario. Further investigation is needed to explain this behaviour in more detail.

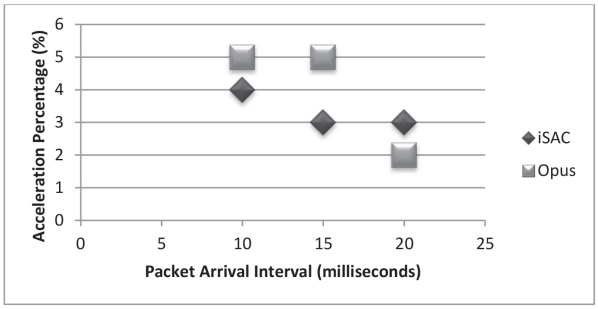

Fig. 4 Acceleration Percentage vs Packet arrival interval - file: CH_m2_s4_file_25.48k.pcm

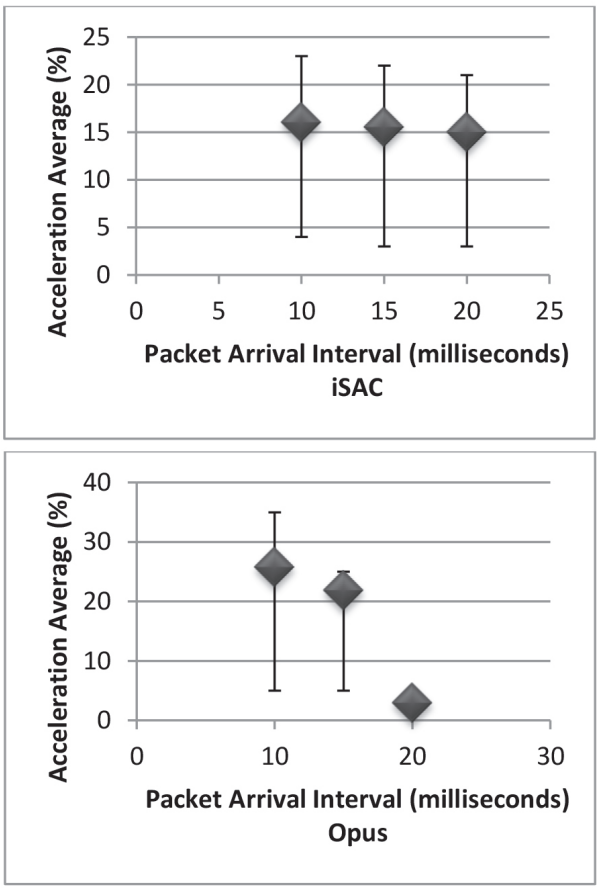

Fig. 5 Average Acceleration for each Packet Arrival

Figure 5 presents the average of the accelerations obtained for all the involved files applied by each codec per packet arrival interval. It is clearly depicted here that Opus applies more acceleration than iSAC, except unsurprisingly, for 20ms packet arrival. It is also quite interesting that whilst for Opus, there is a trend of decreased acceleration as interval increases from 10 to 15, for iSAC there is little variation in acceleration % across all 3 arrival settings.

3.2. Listening Quality Assessment and Time Scaling

We carefully listened to the each degraded speech file in order to see how perceptible are the degradations introduced by the time scaling algorithm of WebRTC. We found that acceleration higher than 15% is evident and perceptible. Moreover, we found that it is only marginally impacting the intelligibility and listening quality. On the other hand, a lower acceleration, i.e. less than 15%, is hardly perceptible. Overall, we think that if the samples undergo a formal subjective listening test, most, if not all, of the samples would be rated as 4 MOS, or close to it.

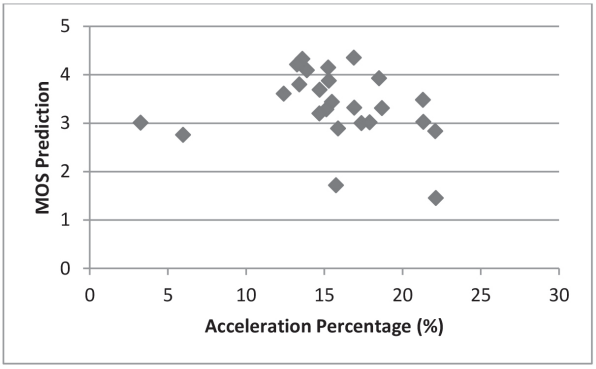

Fig. 6 POLQA MOS vs Acceleration rate - iSAC

We performed an objective quality assessment on the produced speech files and obtained POLQA predictions for all the files involved in this study. We obtained wide-ranging results and they mostly contradict the results from expert listening and thus the results we would expect from formal subjective listening tests according to [16].

Figure 6 shows the POLQA results for iSAC and Fig. 7 presents the results for Opus for a packet arrival interval of 15ms. We can clearly see from both figures that POLQA does not correctly predict quality for the samples degraded by higher acceleration. It is again interesting to note that despite fixed arrival rates set to half the packetisation rate for iSAC (15 ms vs. 30 ms) and 3/4 for Opus (15 ms vs. 20 ms) for all the samples, the acceleration % ranged from 5 to more than 20 depending on sample, with a significant cluster around 12-22 for iSAC and 15-25 for Opus.

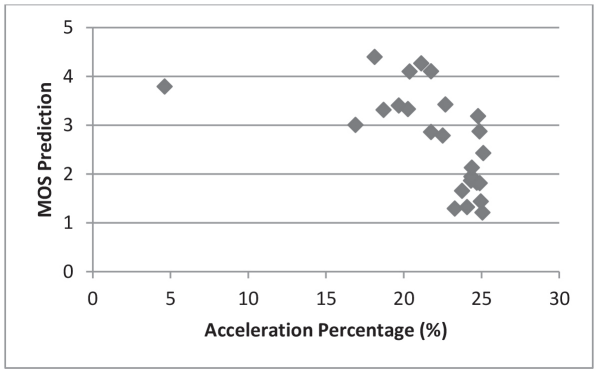

Fig. 7 POLQA MOS vs Acceleration rate - Opus

4. Conclusion and Future Work

In this paper, we have investigated through black-box testing, time scale modification deployed in the adaptive jitter buffer algorithm of WebRTC with a focus on two aspects. The first is the extent of acceleration/scaling for two codecs, Opus and iSAC, under different packet arrival intervals and secondly, the consequent impact on listening quality experienced by the end user. We examined the MOS scores predicted by POLQA and also executed expert listening tests. Two questions, as outlined in Section 1.1, are addressed in this study.

Regarding the first question, we observed that the default WebRTC employs a time scale modification algorithm to provide an adaptive jitter buffer mechanism. Its maximum buffer size is set to 50 packets and if it receives a new packet while it is full, it is designed to drop all of the 50 packets. The extent of scaling/acceleration was seen to differ greatly depending on codec and speech signal. We saw a higher acceleration ratio for Opus even though the simulated network bursts were more extreme for iSAC with a packetisation interval of 30ms. Also, the acceleration varied hugely across the samples even when network burst ratio and codec were kept constant. The reasons behind this behaviour require further research.

Addressing the second question, our results show that POLQA (version 1.1) produces very varied results at the level of acceleration used in this research. We did not observe any correlation in the results achieved. Related research shows that POLQA can deal with scaling represented by a simple resampling of the speech signal as long as it is less than 3%. Our research applies extreme packet bursts and results in much more significant time scaling, up to around 25%. On the other hand, our informal subjective tests performed by the authors, did not reveal such degradations in the perceived quality. Therefore we conclude that the MOS-LOQ scores predicted by POLQA model do not reflect subjective results. Moreover, it is worth noting here that inability of POLQA to provide reliable results at the acceleration levels

used in this research is not caused at all by the modified buffer size deployed in this study.

Finally, whilst extreme, it is theoretically possible to see situations where there is a very high arrival rate of packets into the jitter buffer under certain network conditions for a period of time. The approach taken in this study, which is to increase the buffer size and sending packets faster than packet spacing at the origin, is an attempt to simulate those conditions. This simulation is a first attempt and we plan to emulate a broader range of real world network conditions in future work.

Acknowledgement

This work has been partially supported by the ICT COST Action IC1304 - Autonomous Control for a Reliable Internet of Services (ACROSS), November 14, 2013 – November 13, 2017, funded by European Union.

References

[1] BERGKVIST, A., BURNETT, D., JENNINGS, C., NARAYANAN, A.: *WebRTC 1.0: Real-time Communication Between Browsers. W3C Editor's Draft, W3C*. Retrieved from W3C: http://www.w3.org/TR/webrtc/, 2012.

[2] IETF. (n.d.). *Real-Time Communication in WEB-browsers*. Retrieved from IETF: https://tools.ietf.org/wg/rtcweb/

[3] CINAR, Y.: *An Objective Black-box Evaluation of Voice Quality within the WebRTC Project in Presence of Network Jitter,* 2013.

[4] WebRTC. (n.d.). *WebRTC*. Retrieved from www.webrtc.org

[5] POCTA, P., MELVIN, H., HINES, A.: An Analysis of the Impact of Playout Delay Adjustments introduced by VoIP Jitter Buffers on Listening Speech Quality. *Acta Acustica united with Acustica , 101* (3), 616-631, 2015.

[6] MOON, S. B., KUROSE, J., TOWSLEY, D.: Packet Audio Playout Delay Adjustment: Performance Bounds and Algorithms. *Multimedia systems* , 2/1, 17-28, 1998.

[7] LIU, F., KIM, J., KUO, C-C. J.: *Adaptive Delay Concealment for Internet Voice Applications with Packet-based Time-scale Modification. IEEE ICASSP'2001.* IEEE, 2001.

[8] LIANG, Y. J., FARBER, N., GIROD, B.: *Adaptive Playout Scheduling Using Time-scale Modification in Packet Voice Communications.* Acoustics, Speech, and Signal Processing, 2001. Proc. of (ICASSP'01), 2001 IEEE International Conference on. 3, pp. 1445-1448, IEEE.

[9] LIU, F., KIM, J., KUO, C-C. J.: *Quality Enhancement of Packet Audio with Time-scale Modification.* ITCom 2002: The Convergence of Information Technologies and Communications, pp. 163-173. International Society for Optics and Photonics, 2002.

[10] SCHMIDMER, C.: *POLQA Characterization for Time Scaling Conditions,* ITU-T, 2011.

[11] MELVIN, H.: *The Use of Synchronised Time in Voice over IP (VoIP) Applications*. PhD Thesis, University College Dublin, October 2004.

[12] International Telecommunications Union: *ITU-T Rec. P.800: Methods for subjective determination of transmission quality.* Geneva, 1996.

[13] International Telecommunications Union: *ITU-T Rec. P.862: Perceptual evaluation of speech qaulity (PESQ).* Geneva, 2001.

[14] International Telecommunication Union: *ITU-T Rec. G.107: The E-Model: a computational model for use in transmission planning.* Geneva, 2009.

[15] TechCrunch. (n.d.): *TechCrunch*. Retrieved from The WebRTC Race Begins Today: http://techcrunch.com/2015/02/28/1123773/

[16] HINES, A., SKOGLUND, J., KOKARAM, A. C., HARTE, N.: ViSQOL: An Objective Speech Quality Model. *EURASIP J. on Audio, Speech, and Music Processing,* 5/17, 2015. [17] International Telecommunications Union: *ITU-T Rec. 863: Perceptual objective listening quality assessment.* Geneva, 2011.

[18] *WebRTC Quality of Experience*. Retrieved from webrtcquality.cloudapp.net.