Investigating the Extent and Impact of Time-Scaling in WebRTC Voice Traffic Under Light, Moderate and Heavily Congested Wi-Fi APs

Mohannad Al-Ahmadi, Yusuf Cinar, Hugh Melvin, Peter Pocta

Discipline of Information Technology, College of Engineering & Informatics, National University of Ireland, Galway, Ireland
Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia

m.alahmadi1@nuigalway.ie, cinar.yusuf@gmail.com, hugh.melvin@nuigalway.ie, peter.pocta@fel.uniza.sk

Abstract

Real-time communication (RTC) applications like VoIP ideally require networks that support the necessary quality of service (QoS) whereas the reality is that network impairments such as latency, jitter and packet loss exist. In order to cope with jitter and delay, some VoIP applications employ time-scale modification or warping in the jitter buffer that adjusts the rate of playout while controlling the pitch to minimize Mouth-to-Ear (M2E) delay whilst preserving speech intelligibility and quality. In this paper, we firstly investigate the extent to which time-scaling occurs using WebRTC [1] VoIP clients over Wi-Fi networks with different levels of congestion. We then assess the impact of such time-scaling, both subjectively via expert listening test and objectively using POLQA, on quality experienced by the end user, and review the correlation between scores.

1. Introduction

Real Time Communication (RTC) applications like voice/video conferencing face a lot of challenges over unmanaged networks in the absence of QoS support for latency, packet loss and jitter.

Jitter is currently one of the main challenges and occurs largely due to intermittent congestion, resulting in packets arriving at various time intervals. Jitter buffers are thus used at the receiver's side in order to optimize the quality for the end user using a range of adaptive strategies to remove delay variation, maximize speech quality whilst minimizing packet loss and overall Mouth-to-Ear (M2E) delay. Such strategies can be classified as per talkspurt or per packet [2]. The former applies adjustments to silent periods within talkspurts thus preserving the integrity of the active speech whereas the latter technique also referred to as time-scale modification acts on active voice signal. As defined by the ITU-T [3], time-scaling is a temporal compression or stretching applied to the actual signal or a small section of it. In this paper we focus on the per-packet strategy deployed by WebRTC. Its jitter buffer uses NetEQ [4] described as "a dynamic jitter buffer and error concealment algorithm used for concealing negative effects of packet loss caused by network or end point. It keeps latency as low as possible while maintaining the highest voice quality." It also deals with clock skew between sender and receiver clients.

It is worth noting here that WebRTC is receiving much attention these days as it is free and available worldwide for both developers and the end users. This gives the community an access to open real-time communication technology. With browser API, WebRTC enables developers to implement audio and video applications into web browsers like Chrome, Firefox and Opera.

Wi-Fi or IEEE 802.11 increasingly represents the last network hop and the consequent contention for access, particularly in the downlink direction via the CSMA/CA mechanism can lead to significant jitter which can impact greatly on RTC applications. As stated in [5], the basic IEEE 802.11 standards do not guarantee an upper limit of packet delay or loss. Moreover, call drop can occur quite frequently at this transmission technology. As a result, significant research has examined ways how to improve QoS via traffic prioritization over WiFi [6] and the new 802.1aa extension will also help in this regard [7].

The main objective of this research is firstly to investigate the extent to which packet-scaling can occur in real-world VoIP traffic over Wi-Fi networks, and secondly, to quantify its impact on quality experienced by the end user using both expert subjective listening tests and objective tests, the latter using POLQA [11-13]. Such research can inform the community about the possible need to adapt objective methods to correctly deal with everyday reality experienced by the end users of RTC applications, particularly WebRTC.

The remainder of this paper is structured as follows. Section 2 describes related research. Section 3 provides an overview of the two testbeds – the first one was used to capture real-world jitter profiles, the second one was used to firstly determine the extent of time-scaling resulting from these profiles on real-world WebRTC VoIP traffic and secondly, to assess the

10.21437/PQS.2016-19

consequent impact on quality perceived by the end user via both expert listening and objective means. Section 4 presents the results. Section 5 concludes the paper and outlines further steps in this research.

2. Background & Research Questions

In [2] a comprehensive analysis of per-talkspurt jitter buffer strategies was carried out to investigate the impact of small and frequent silence-period adjustments typical of such strategies on speech quality perceived by the end user. The authors compared subjective results with results obtained from the two most popular objective speech quality prediction models, PESQ (Perceptual Evaluation of Speech Quality) [8-10]and POLQA (Perceptual Objective Listening Quality Assessment) [11-13]. The impact of the playout adjustments on the subjective scores was found to be insignificant, whereas regarding the objective speech quality assessment models, i.e. PESQ and POLQA, the impact was found to be significant, though it is worth noting that the POLQA model performed significantly better than PESQ.

More recent research published in [14] investigated the extent and impact of time scale modification introduced by WebRTC VoIP client under extreme network jitter conditions using a black-box approach. To emulate extreme burstiness, sustained packet arrival rates of 10ms and 15ms were applied to 20 ms packets. The results obtained from the POLQA model showed a very significant impact on MOS scores under such high compression scaling whereas expert listening tests revealed an insignificant impact. It is important to note here that POLQA was designed to cope only with moderate (3%) levels of packet scaling, much less than that introduced in [14]. As such, we revisit the issue in this paper, and extend the study published in [14] by addressing the following three key questions:

1. To what extent will jitter profiles arising from running WebRTC VoIP traffic over Wi-Fi Access Points under different levels of congestion result in packet scaling?
2. What impact on quality perceived by the end user will such scaling have both subjectively and objectively?
3. Is the extent to which POLQA is tuned to accommodate scaling sufficient for current congested networks which are commonplace, especially in Wi-Fi scenarios?

3. Methodology

Two testbeds were designed and implemented to facilitate tests. Firstly, a Wi-Fi testbed is used to capture real-world network scenarios under varying WiFi AP congestion levels. In the second testbed, the network profiles are then applied to the WebRTC testbed to obtain speech quality scores from POLQA model as well as from expert listening tests.

3.1 Wi-Fi Testbed

The testbed shown in Fig. 1 involves two hosts running WebRTC VoIP clients to send/receive 20 ms voice packets using different codecs. Our testbed uses both narrowband (PCMU) and super wideband (OPUS) speech codecs which are the most common codecs used in current WebRTC applications. It deploys the Wireshark packet sniffer [15] to capture timestamps (and thus jitter profiles at the receiver side in both directions). Tests were carried out on the National University of Ireland WiFi campus network at various times and locations to experience different congestion levels. Each test lasted 120 seconds. The captured jitter profiles were then used in the WebRTC testbed to test the jitter buffer scaling strategy deployed by WebRTC clients and quantify it in terms of quality perceived by the end user deploying both subjective listening and objective testing.

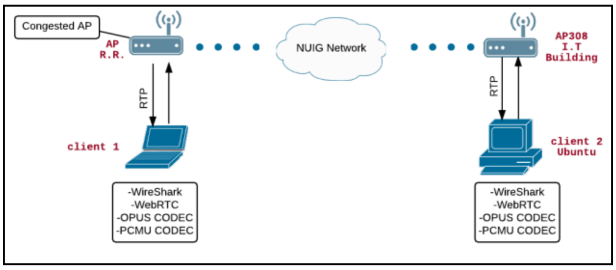

Figure 1. *Testbed setup used to capture jitter profiles*

3.2 WebRTC Testbed

As illustrated in Fig. 2, the second testbed involves a VoIP sender and receiver on a single machine using the WebRTC communication chain with an emulated network channels in between, to introduce network jitter. The sender reads a PCM file containing the ITU-T speech samples, encodes them and sends the data to the network channel. This enables us to focus solely on the impact of time-scaling. The whole testbed is built using Linux container technology, is fully automated and uses Opus codec. The delay/jitter profiles can be preconfigured manually or taken from actual network captured profiles. As the WebRTC jitter buffer is hardcoded to hold up 50 packets, the jitter buffer in this testbed was modified to hold up to 500 packets in order to avoid packet loss under bursty conditions which would otherwise add another variable degradation to the results. This then allowed us to focus on time scaling only. The testbed uses the ITU-T SWB speech samples from ITU-T Rec. P.501 [16], which have passed POLQA transparency test, as input. Each consists of a pair of utterances from male and female speakers.

In order to evaluate the quality perceived by the end user, all the samples are recorded at the receiver side for quality assessment using the POLQA model and are also assessed via expert subjective listening. Our native emulator was shown to replicate jitter profiles with a correlation coefficient greater than 0.97.

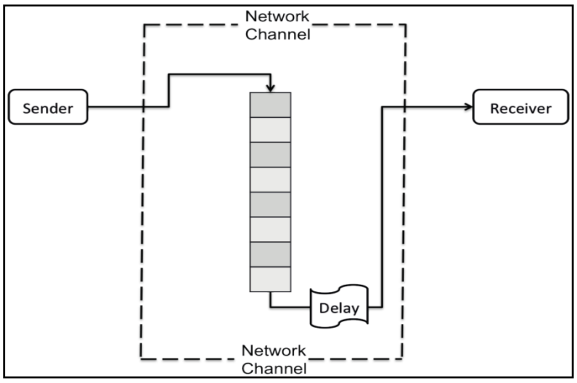

Figure 2. *WebRTC Simulation Testbed* [6]

4. Results

4.1 Jitter profiles over Wi-Fi

As outlined above, we captured various jitter profiles under different congestion situations by transmitting RTP streams over campus WiFi network with both sender and receiver capturing the delta time of the packets as they arrive at the network interface. We present a subset of these results, namely seven jitter profiles as shown in Table 1. Profiles that featured extremely high jitter along with packet loss at the AP buffer in some heavily congested network scenarios were discarded and were not included in the test as jitter behaviour was very inaccurate due to the dropped packets. Figures 3-5 illustrate the jitter profiles, representing light/moderate/heavy congestion situations respectively, with y-axis showing the inter arrival packet time.

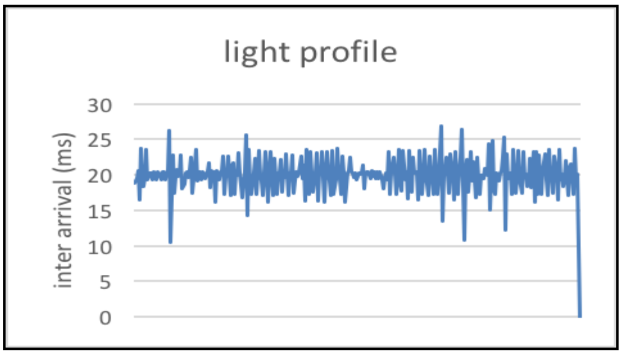

Figure 3. *Light jitter profile*

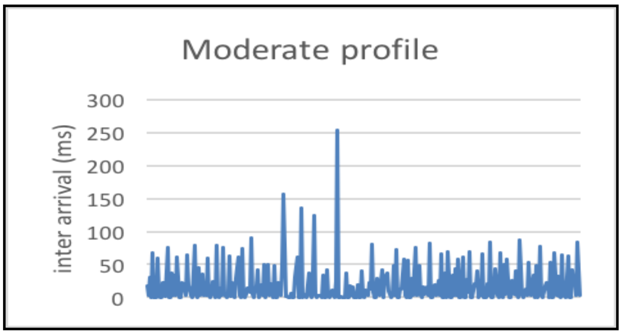

Figure 4. *Moderate jitter profile*

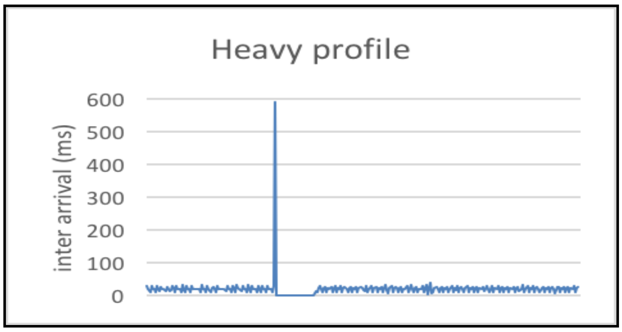

Figure 5. *Heavy jitter profile*

As can be seen in Fig. 6, in some cases, RTP packets are queued up in the AP buffer for up to 800 ms awaiting transmission to the receiver and then released in form of bursts where packet delta is close to 0ms. Fig. 7 outlines jitter profile arising from a moderately congested AP, using packets/sec as metric, bearing in mind that the default rate of packets is 50 RTP packets per second. Intermittent congestion periods resulted in buffer starvation followed by packet bursts of up to 90 packet per second.

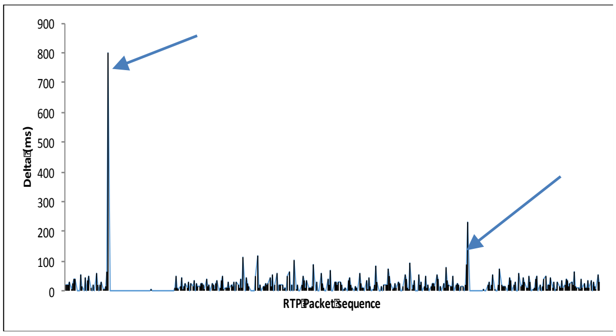

Figure 6. *RTP packet delta of 800ms followed by packets burst*

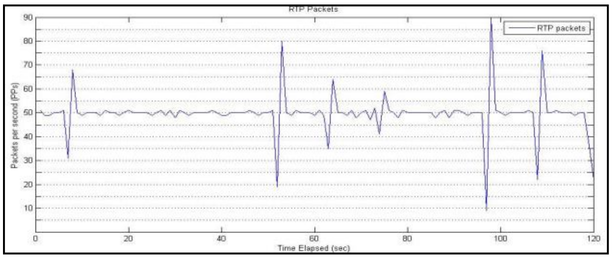

Figure 7. *Packets/sec arriving on network interface*

As shown clearly in Fig.5-7, when packets are queued at the access point due to the congestion in the network, they are then released in form of bursts at a later stage. When packets are received at the receiving interface, the jitter buffer reacts to a gradually decreasing (starvation) and a rapidly increasing (burst) jitter by scaling the packets to adjust the rate of playout, as illustrated in Fig.8. This specifically causes stretch followed by rapid compression in the voice segments. In such scenarios, we are investigating whether the 3% scaling mentioned in ITU-T Rec. P.863.1 is exceeded for significant periods and if so what is its impact on quality perceived by the end user.

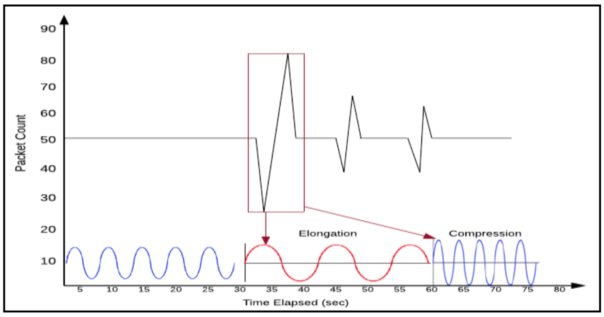

Figure 8. *impact of bursts on speech signal in the jitter buffer*

4.2 Test Results of expert listening test

In this paper, we express the acceleration rate or scaling defined in [11] as:

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

where *Ar* is the acceleration rate, *T*d is the duration of degraded speech file and *T*o is the duration of the original speech file. An expansion is thus expressed by a negative value of the acceleration rate parameter whereas a positive value is used when the sample is accelerated.

We firstly ran an expert ACR (Absolute Category Rating) listening test according to ITU-T Rec. P.800 [17]. In the test, we investigated each level of degraded speech sample (light-moderate-heavy WiFi congestion) to evaluate the impact of degradation introduced by the time scaling algorithm. The average percentage (%) scaling for speech samples for light - moderate - heavy congestion was -0.55%, -1.45% and -1.22%. However, these average values hide the reality as within each sample there are periods where the absolute value of scaling % is much higher – as discussed in the next section. Having listening to all the degraded speech samples obtained from the WebRTC testbed, we concluded as expert listeners that the impact of degradation was inaudible, even for the heavily congested degraded samples.

Congestion Level	Jitter Profile	Average Scaling%	Ref. MOS-LOQ	POLQA MOS-LOQ
Light	1	-1.20%	4.583	2.7171
	2	-1.22%		2.1085
	3	-0.72%		2.0452
	4	0.93%		3.7004
Moderate	5	-1.53%		3.2354
	6	-1.36%		3.388
Heavy	7	-1.22%		2.0377

Table 1. *Average scaling and POLQA MOS-LOQ*

As such, we strongly believe that if we would run a real ACR listening test, all the samples would be rated as 4 MOS, or close to it, and no remarkable difference between MOS scores over the test conditions would be reported.

4.3 Objective Listening Assessment

We then evaluated the quality of the produced speech files using POLQA (version 2.400). Table 1 outlines results for all seven profiles. It firstly includes the reference conditions whereby no jitter profile was applied resulting in score of 4.583. Interestingly, we then found that for each of the seven jitter profiles, the predicted results from POLQA strongly contradicts the results from the expert listening. We postulate that this is likely caused by the intermittent periods within the speech samples when the scaling is applied to the waveform under the congestion scenarios, and which is masked by average scaling value. Figure 9 demonstrates how the scaling was applied to an active speech segment in the heavy congested scenario which resulted in scaling (expansion) of -70% in the waveform. Interestingly, there seems to be no obvious relationship between the average scaling, i.e. profile number and POLQA quality predictions. We plan to investigate this issue further, and in particular we will examine the frequency and magnitude of intermittent scaling periods resulting from each of the profiles, as well as the algorithms deployed by POLQA to aggregate quality over speech segments.

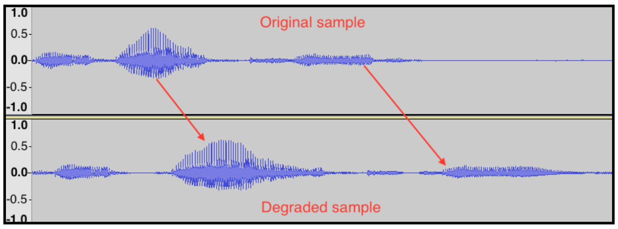

Figure 9. *Demonstration of how scaling was applied to voice segments*

5. Conclusion

In this paper, we have investigated the extent and impact of time-scaling introduced by WebRTC jitter buffer strategy on the quality experienced by the end user, using both subjective listening test and objective assessment via POLQA (version 2.400). By using the WiFi testbed, we captured a number of jitter profiles over congested networks. The captured network traces were analyzed and then applied to the WebRTC testbed to obtain the objective MOS scores.

Three research questions outlined in Section 2 were addressed in this paper. Regarding the first question, we show that WiFi can add very significant jitter to WebRTC traffic with packet bursts of up to 90 packet per second. Regarding the extent of scaling arising from these real-world profiles, our results show that whilst the average scaling is low across all jitter profiles, frequent and significant (>3%) scaling within the degraded samples is quite prevalent. Question 2 deals with the impact of scaling on the quality perceived by the end user. In this context, we firstly investigated the impact using ACR expert listening test according to ITU-T Rec. P.800 and we have found the impact of degradation inaudible. However, the objective MOS scores provided by POLQA have not reflected the results of the subjective listening assessment. In particular, POLQA provided much lower scores for all the degraded speech samples compared to expert listening. Finally, and regard to Question 3, we postulate that the frequent short periods of significant scaling (>3%) are resulting in low POLQA scores and outline that more work needs to be done to precisely investigate why this was the case. In conclusion, it is important to ensure that objective metrics can better deal with real world jitter scenarios arising especially those coming from WiFi networks so that their output more closely match the subjective experience of the end user.

6. 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:www.w3.org/, 2012.

[2] 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.

[3] Recommendation P.863.1: Methods for objective and subjective assessment of speech quality, International Telecommunications Union (ITU-T), (09/2014)

[4] WebRTC glossary. Retrieved from: webrtcglossary.com/neteq/.

[5] M. A. R. Siddique, J. Kamruzzaman, and M. J. Hossain, "An Analytical Approach for Voice Capacity Estimation Over WiFi Network Using ITU-T E-Model," *IEEE Trans. Multimedia.*, vol. 16, no. 2, pp. 360–372, Feb. 2014.

[6] P. O Flaithearta, H. Melvin, and M. Schukat, "A QoS enabled multimedia WiFi access point," *Int. J. Netw. Manag.*, vol. 25, no. 4, pp. 205–222, Jul. 2015.

[7] IEEE 802.11aa Task Group, Robust streaming of Audio Video Transport Streams, www.ieee802.org: Retrieved: 2016.

[8] Recommendation P.862: 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 Telecommunications Union (ITU-T), 2001

[9] A. W. Rix, M. P. Hollier,A.P.Hekstra, J. G. Beerends: Per- ceptual evaluation of speech quality (PESQ)- The newITU standard for objective measurement of perceivedspeech quality.Part I: Time-delay compensation. J. Audio Eng. Soc. 50 (2002)755–764.

[10] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: Perceptual evaluation of speech quality (PESQ)- The newITU standard for objective measurement of perceivedspeech quality.Part II; psychoacoustic model. J. Audio Eng. Soc. 50 (2002)765–778.

[11] Recommendation P.863: Perceptual Objective Listening Quality Assessment(POLQA), International Telecommunications Union (ITU-T), 09/2014.

[12] J. G. Beerends, C. Schmidmer,J.Berger,M.Obermann, R. Ullman, J. Pomy, M. Keyhl: Perceptual objective listening quality assessment (POLQA). The third generation ITU-T standard for end-to-end speech quality measurement. Part I: Temporal alignment. J. Audio Eng. Soc. 61 (2013)366– 384.

[13] J. G. Beerends, C. Schmidmer,J.Berger,M.Obermann, R. Ullman, J. Pomy, M. Keyhl: Perceptual objective listening quality assessment (POLQA). The third generation ITU-T standard for end-to-end speech quality measurement. Part II: Perceptual model. J. Audio Eng. Soc. 61 (2013)385– 402.

[14] Cinar, Y., Melvin, H., Počta, P.: A Black-Box Analysis of the extent of time-scale modification introduced by WEBRTC adaptive jitter buffer and its impact on listening speech quality. Special Issue of Communications journal (Komunikacie – Scientific Letters of the University of Zilina), on Telecommunications Beyond 2016, vol.18, No.1, pp. 17-22, ISSN 1335-4205.

[15] Wireshark, The Wireshark Foundation, www.wireshark.com, retrieved: 2016

[16] International Telecommunications Union:ITU-T Rec. P.501 : Test signals for use in telephonometry(01/12).

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