Research Repository UCD

Title	Playout again Sam: Jitter Buffer Playout Adjustments Still an Issue for Speech Quality Prediction Models?
Authors(s)	Cinar, Yusuf, Pocta, Peter, Hines, Andrew
Publication date	2020-06-12
Publication information	Cinar, Yusuf, Peter Pocta, and Andrew Hines. “Playout Again Sam: Jitter Buffer Playout Adjustments Still an Issue for Speech Quality Prediction Models?” IEEE, June 12, 2020. https://doi.org/10.1109/ISSC49989.2020.9180163.
Conference details	The 2020 31st Irish Signals and Systems Conference (ISSC), Letterkenny, Ireland (held online due to Coronavirus outbreak), 11-12 June 2020
Publisher	IEEE
Item record/more information	http://hdl.handle.net/10197/25683
Publisher's statement	© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Publisher's version (DOI)	10.1109/ISSC49989.2020.9180163

Downloaded 2026-05-02 00:25:14

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/342832000

Playout again Sam: Jitter Buffer Playout Adjustments Still an Issue for Speech Quality Prediction Models?

Conference Paper · June 2020

DOI: 10.1109/ISSC49989.2020.9180163

CITATIONS
0

READS
39

3 authors:

Yusuf Cinar
National University of Ireland, Galway
5 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Peter Počta
University of Žilina
64 PUBLICATIONS **191** CITATIONS
SEE PROFILE

Andrew Hines
Technological University Dublin - City Campus
53 PUBLICATIONS **303** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Ambiqual View project

QoEVIS - Study factors influencing the quality of experience of audiovisual streaming View project

All content following this page was uploaded by Peter Počta on 10 July 2020.

The user has requested enhancement of the downloaded file.

Playout again Sam: Jitter Buffer Playout Adjustments Still an Issue for Speech Quality Prediction Models?

Yusuf Cinar[1], Peter Pocta[2], and Andrew Hines[3]
[1] *School of Computer Science, College of Engineering & Informatics, National University of Ireland, Galway, Ireland*
[2] *Department of Multimedia and Information-Communication Technology, FEEIT, University of Zilina, Slovakia*
[3] *School of Computer Science, University College Dublin, Ireland*
cinar.yusuf@gmail.com, peter.pocta@feit.uniza.sk, andrew.hines@ucd.ie

***Abstract*—Objective speech quality assessment techniques, which use the perceptual models to emulate the human listening perception, have seen several revisions in the recent years. This study investigates the evolution of POLQA and ViSQOL models and scrutinise their latest versions. Prior work had identified weaknesses in both prediction models when presented with speech containing imperceptible playout adjustments. This study follows up the experiments to evaluate the progress and report the progress and the current issues, benchmarked against subjective listening quality scores. The assessment is conducted for all published versions of the POLQA and ViSQOL models and the evolution and improvement offered is analysed. We can conclude that the models have been improved in terms of imperceptible jitter buffer adjustments highlighted in prior work. This study also explores the performance of objective quality models and intelligibility (STOI and POLQA Intelligibility) models for a data set produced with realistic but extreme WebRTC scenarios using a standard and novel WebRTC jitter buffer strategy. An expert listening test was conducted to subjectively evaluate the WebRTC data set. It is observed that the standard WebRTC jitter buffer strategy produces more natural speech while the novel approach offers better intelligibility. The subjective and objective quality results suggest that the speech quality for standard jitter buffer were lower but more consistent than for the novel jitter buffer. The objective intelligibility results were conflicting. A followup study will conduct independent subjective evaluations of quality and intelligibility to further explore the relationship between the objective intelligibility and quality results.**

***Index Terms*—Quality of Experience, Playout Delay Adjustments, POLQA, ViSQOL**

I. Introduction

Just as Humphery Bogart is regularly misquoted from the movie Casablanca as saying 'Play it again, Sam' rather than the actual line, 'Play it, Sam', delayed packets can cause words in speech to be missed or corrupted. Jitter buffers use playout adjustment strategies to introduce small, imperceptible delays or silence omissions along with small time-scale adjustments in the speed of speech to improve the quality of speech in delay-sensitive applications such as real-time voice transmission over the internet.

978-1-7281-9418-9/20/$31.00 ©2020 IEEE

Full reference speech quality models provide good predictions for a range of speech degradations caused by network transmission [1], [2]. The prior work demonstrated that the PESQ model [3] had significant problems when it comes to dealing with playout adjustments caused by jitter buffer strategies [4]. While newer objective quality metrics (POLQA [2] and ViSQOL [1]) were better at this, they still produced poor quality predictions in some scenarios [4], [5]. The POLQA [6] and ViSQOL [7] models were recently enhanced to address the issue of poor quality prediction as a result of small mis-alignments, resulting from the micro-pauses or time-scale adjustments introduced by the jitter buffer strategies, between reference and test signals. The POLQA model has seen two major revisions, POLQA version 2 and 3, since the prior work involving the POLQA version 1 published in [4]. Similarly, the ViSQOL model has seen major revisions since the prior work published in [5]. It is worth noting here that the PESQ model is out of scope of this paper as it has been superseded by the POLQA model.

An updated POLQA specification [6], version 3 released in 2018, attempts to quantify an impact of micro-pauses on the quality perceived by the end user. [6] acknowledges that an extension of a pause between sentences, words or syllables is considered to be less degrading when compared to an interruption during speech activity. The process for POLQA to handle such degradations is two-fold. Firstly, inserted sections are detected and provided to the perceptual model by the temporal alignment component. Secondly, the impact of the inserted sections on the perceived quality is quantified by the perceptual model [6].

Similar to the journey that the POLQA model took, the ViSQOL model has also updated its model with a new algorithm for a delay estimation to compensate for the playout adjustments introduced by the jitter buffer strategies. The issue was identified as a signal sample mis-alignment within a spectrogram time frame that cascaded the error into subsequent frames [1].

[7] presents the proposed revisions to the original ViSQOL model (v131), studied in the previous work [5], and shows that an updated model with a jitter buffer compensation adjustment

(v131adj) produces considerably fewer signal mis-alignments, consequently resulting in better quality predictions.

In this paper, two main contributions, both of which are related to playout adjustments and its quality impacts, are presented. Firstly, we compare and assess the evolution and the current versions of POLQA and ViSQOL for a data set containing imperceptible playout delay adjustment introduced by jitter buffer strategies [4]. The previous studies [4], [5] have already identified the shortcomings of both in this context. Secondly, we explore the performance of the aforementioned objective quality models for a data set produced with realistic but extreme WebRTC scenarios using standard and novel WebRTC jitter buffer strategies, both of which apply playout adjustments in the form of time-scale modifications while reconstructing the speech signal. The main difference between the two strategies is that the first one produces more natural speech while the other one prioritises intelligibility by preventing packet loss but introducing pauses and time-scale modifications to the speech signal.

The remainder of this paper is structured as follows. Section II provides the POLQA and ViSQOL models benchmark in terms of playout delay adjustments introduced by jitter buffer strategies and their evolution since the previous studies. Section III outlines the performance of the POLQA and ViSQOL models under the two different WebRTC scenarios mentioned above. Section IV discusses the results and observations from Section II and III. Section V concludes the paper and summarises some areas for future research arising from this paper.

II. POLQA and VISQOL Benchmark in terms of playoutdelay

A. Experiment description

We evaluated the major versions of the POLQA and ViSQOL models released in the last seven years. We benchmarked their ability to correctly predict speech quality for jitter buffer processed speech with playout adjustments using an existing dataset [4]. It is worth noting here that this dataset was previously used to benchmark the POLQA model version 1 and the PESQ model, see [4] for more detail. The dataset contains 24 test conditions covering a broad range of realistic playout adjustments applied in different locations (variant A and B) and one reference condition representing the ideal transmission conditions without any jitter. Table I details the actual playout adjustments. Each of the conditions involves 2 male and 2 female samples coming from the English subset of ITU-T P Supplement 23 [8]. Each sample consists of two utterances. All the samples were evaluated by 30 naïve listeners in a subjective test following the ACR methodology [9]. More details about the dataset can be found in [4].

TABLE I
Playout adjustments applied to speech samples for experiment 1. Dataset from [4].

	Absolute Adjustments (ms)				Sum of
Test conditions	**1st**	**2nd**	**3rd**	**4th**	**adjustments (ms)**
Ref	0	0	0	0	0
1	2	-2	3	-3	10
2	4	-4	-4	4	16
3	3	-3	-6	6	18
4	5	-5	-5	5	20
5	3	-6	-7	10	26
6	16	-12	-8	4	40
7	10	-17	-6	13	46
8	10	-15	-10	15	50
9	8	-23	-3	18	52
10	5	10	-30	15	60
11	-15	15	-15	15	60
12	-25	22	-8	11	66

Regarding the deployed quality prediction models, the POLQA model versions 1, 1.1, 2 and 3 and the ViSQOL model versions 131, 131adj, 238 and 238adj (where adj is the adjusted ViSQOL algorithm containing micro delay estimation) were benchmarked in this study to thoroughly investigate the progress made by the models in the last years.

B. Experimental Results

The results from dataset of the subjective assessment, which was carried out in the prior work [4], showed that across all variants A and B for all 12 test conditions the effect of test condition was not statistically significant. The analysis of variance (ANOVA) [10] tests showed that neither conditions ($F = 0.82$, $p = 0.635$) nor variants ($F = 1.07$, $p = 0.3032$) were perceptibly different in quality to the reference [4]. The per condition results for POLQA v3 and ViSQOL v131adj are presented in Fig. 1. These per test sample results show that there were some small variations in quality prediction when compared to the reference (e.g. speaker M2 condition 12 for both POLQA and ViSQOL).

For completeness, mean values and 95 percent confidence intervals for each condition were computed for each version of the POLQA and ViSQOL models. Using these metrics, Fig. 2 shows the evolution of the investigated models in the context of the playout adjustments introduced by jitter buffer. It is clear from this figure that the improvements are remarkable for the POLQAv3 and the ViSQOL versions containing the delay estimation adjustment algorithm (denoted adj). As the confidence intervals are rather small, we can conclude that the latest versions of the POLQA model as well as the ViSQOL model are able to accurately predict an impact of the jitter buffer playout adjustments on a quality perceived by the end user.

III. WebRTC Experiments

In the WebRTC experiments, we have worked with two jitter buffer strategies: default WebRTC strategy (Jitter Buffer A) and a modified version (Jitter Buffer B), i.e. burst-aware WebRTC jitter buffer strategy. With the default WebRTC jitter buffer one, all the packets will be dropped if a packet arrives to a full jitter buffer, which is a case that can occur after the arrival of a burst of packets at the same time. The burst-aware WebRTC jitter buffer strategy prevents this excessive late packet loss by removing only a necessary number of

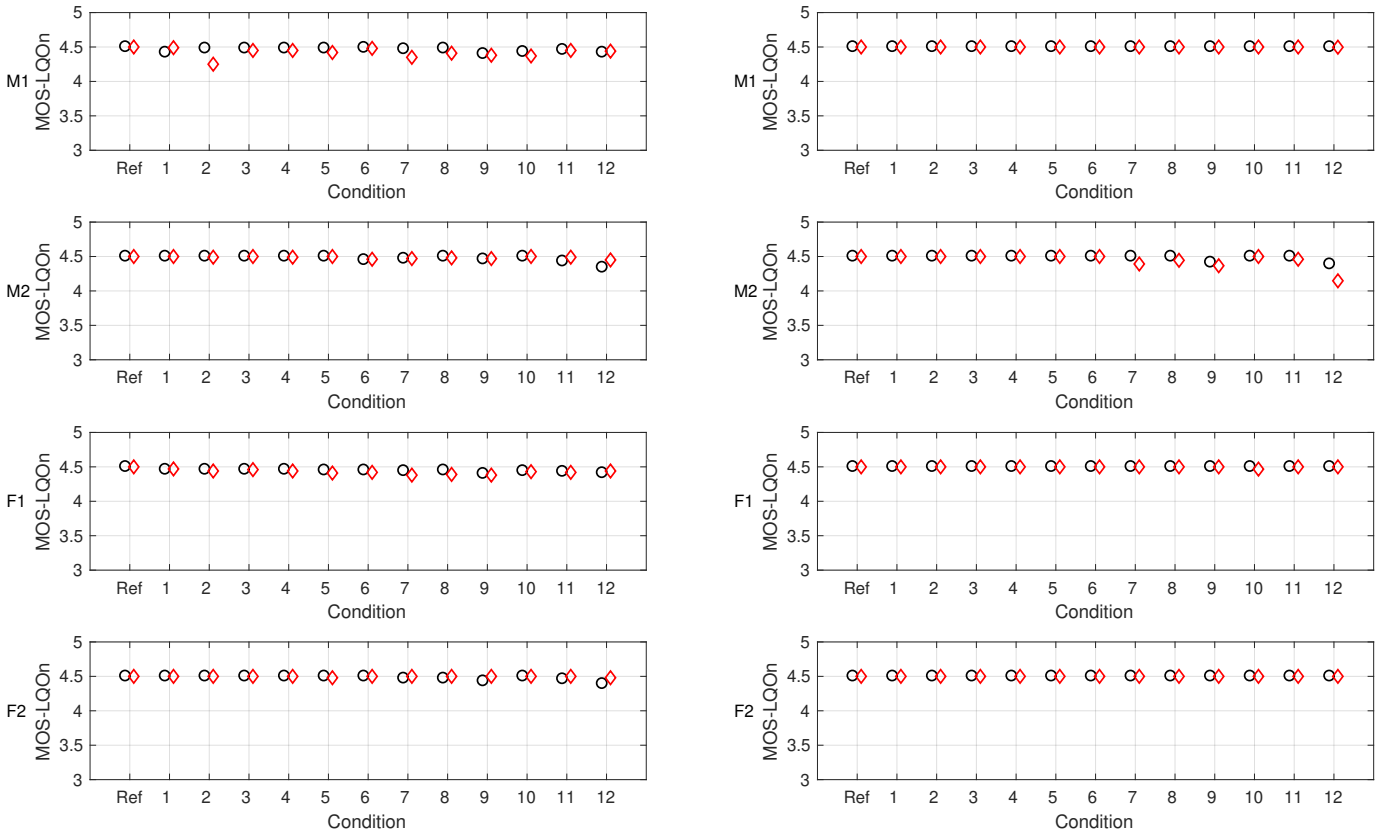

Fig. 1. Quality results of POLQA v3 model (left) and ViSQOL 131adj (right) with micro delay estimation. ○ , ◇ are variants A and B, indicating the adjustments occurred at different locations in the test samples.

packets, i.e. so as to make enough room for the incoming packets, from the jitter buffer. The burst-aware jitter buffer strategy aims to improve intelligibility and quality by reducing the speech loss. The default WebRTC jitter buffer strategy, however, produces more natural speech, although it results in partial speech loss in these circumstances.

A test condition set, presented in Table II, was created containing a variety of severe but realistic network conditions with high jitter, resulting in a number of packets arriving to the destination at the same time. Additionally, jitter buffer capacity was also configured to various sizes, ranging from 15 to 50 packets under each network condition. We have selectively picked 10 test conditions, for which objective speech quality models produced rather interesting results, i.e. conflicting with expert listening results. For instance, for the first test condition, a burst of 43 packets was created and jitter buffer capacity was set to 35 packets. Each test condition was tested with 4 English speech samples, 2 male and 2 female, coming from ITU-T Rec. P501 [11]. Each sample comprised of two utterances.

The voice activity detector (VAD) from webRTC [12] was used to compute the VAD masks for the reference and test samples. Fig. 3 illustrates the VAD masks computed for the 10 conditions using the Female 1 speaker and the built-in webRTC VAD with an aggressiveness level set to 1. Each row shows the reference followed by the corresponding mask for the conditions listed in Table II produced with Jitter Buffer A and B. Masks for the other speakers followed a similar trend. The conditions 5, 7 and 8 highlight scenarios where the first utterance has lost some speech with Jitter Buffer A but less with Jitter Buffer B.

TABLE II
TEST CONDITIONS APPLIED FOR EXPERIMENT 2.

Test conditions	Number of packets in the burst	Jitter buffer capacity (packets)
1	43	35
2	49	15
3	56	15
4	56	25
5	57	45
6	58	15
7	58	45
8	60	50
9	61	25
10	63	25

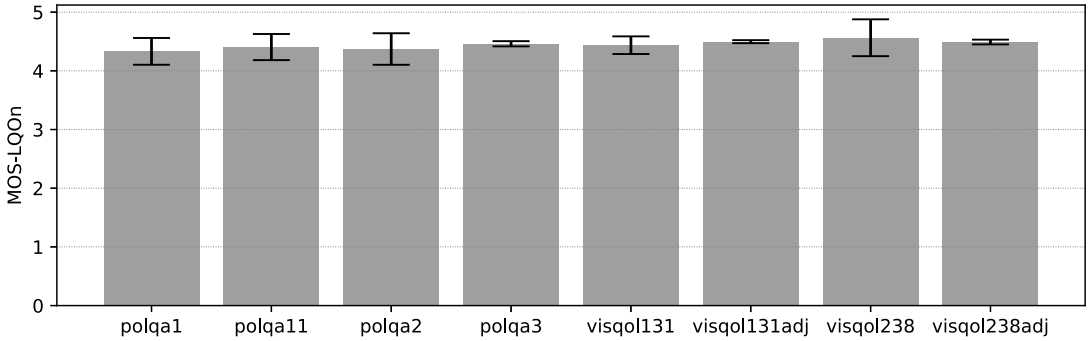

Fig. 2. Objective quality predictions averaged for all conditions from the playout adjustment data [4]. The improvements for POLQA can be clearly seen in v3 and in the ViSQOL versions containing the delay estimation adjustment algorithm (denoted adj).

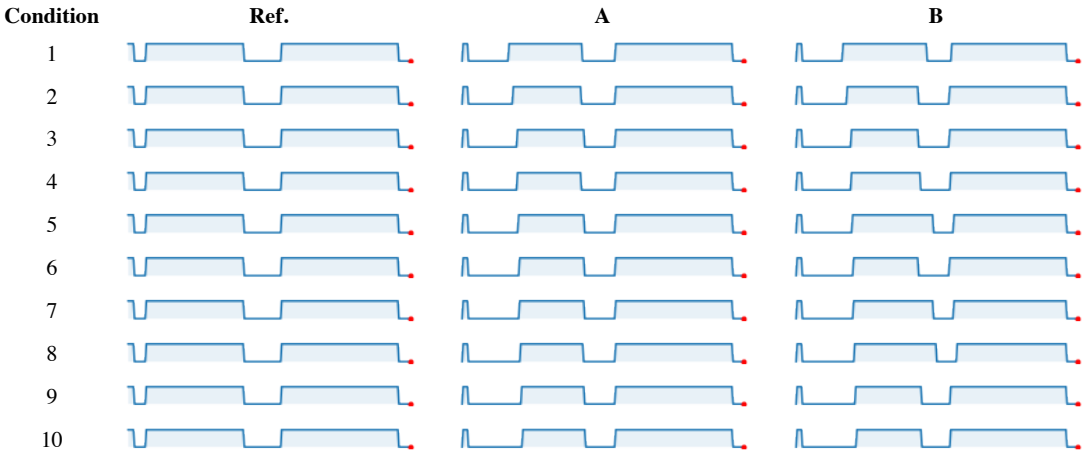

Fig. 3. WebRTC VAD for Reference files (left) and Jitter Buffer A and B

TABLE III
TEST RESULTS FOR EXPERIMENT 2.

Jitter Buffer	POLQA		ViSQOL		POLQA Int		STOI		VAD Delta %		WER		Expert MOS	
	mean	std.	mean	std.	mean	std.	mean	std.	mean	std.	mean	std.	mean	std.
A	1.31	0.48	1.22	0.47	67.85	4.92	**0.68**	0.08	17.67	2.57	0.32	0.14	2.22	0.19
B	**1.69**	0.77	**1.46**	0.62	**71.99**	7.81	0.62	0.09	**14.44**	3.46	**0.31**	0.13	**2.65**	0.64

A. Experimental Evaluation

The 40 test samples were evaluated comparing the reference to the test samples produced using the default WebRTC strategy (labelled A in the results) and the burst-aware jitter buffer strategy (labelled B in the results). The POLQA version 3 [6] and ViSQOL version 238adj [7] were used to predict the speech quality on a Mean Opinion Score-Listening Quality Objective (MOS-LQO) scale of 1–5. STOI [13] and POLQA Intelligibility [14], [15] were used to predict an intelligibility as the samples contained a loss of active speech segments. A percentage loss of speech was computed using the VAD masks as in Equation 1:

$$\%VAD_{Delta} = \frac{\sum_{t=1}^{N} V_{Ref}[t] - \sum_{t=1}^{N} V_{Test}[t]}{\sum_{t=1}^{N} (V_{Ref}[t]} \quad (1)$$

where t is the time frame from 1 to length N and V_{Ref} and V_{test} are the reference and test VAD masks. Under this test scenario, computing a $\%VAD_{Delta}$ score like this will estimate the loss of active speech in the test sample compared to the reference, with a lower score being preferred.

Google's cloud Automatic Speech Recognition (ASR) API [16] was used to predict the utterance annotations for the reference and test samples. A Word Error Rate (WER) was computed using the Wagner-Fisher method [17] implementation contained in the Jiwer python library [18].

Finally, a cursory expert evaluation was carried out with the authors conducting a subjective evaluation and rating the quality of each sample on a 5 point MOS scale. As there was significant loss of speech in some samples the evaluation was carried out as a Degradation Category Rating [9] by listening to the reference and then the test signal.

B. Results

A summary of the results is presented in Table III where the mean and standard deviation of results across the test samples for each metric are presented.

Computing a mean score across a range of conditions will not give us an indication of a given objective metrics prediction per condition. However it will allow us to explore the general trends in severe but realistic bursty network conditions.

Starting with the expert evaluation, overall the quality was rated higher for outputs from Jitter Buffer B. For Jitter Buffer B the voice quality and naturalness of the speech content were compromised in an effort to maximise playout speech play and 'catch up' rather than dropping delayed packets. Jitter Buffer A is not burst aware and the intelligibility was lower. These factors led to the mean MOS being higher but also to the higher standard deviation.

For the objective quality metrics, both the POLQA and ViSQOL performed similarly in their ratings. Both scored Jitter Buffer B above Jitter Buffer A and as with the Expert MOS evaluation, there was also a higher standard deviation for Jitter Buffer B.

Intelligibility estimates using the STOI and POLQA Intelligibility produced conflicting results. The POLQA Intelligibility has a trend that is very similar to that of the POLQA Speech Quality model in both, i.e. for the relative distance between the rating of Jitter Buffer B over A and the high range of the standard deviation for Jitter Buffer B in particular.

The STOI was the only metric to rank Jitter Buffer A over B by 6%. The difference between the standard deviations show that the STOI did not see as large a difference in the variation of scores that the other metrics and subjective listening highlighted.

The VAD Delta % and WER are also presented in the table for completeness. The WER was almost identical, and the high standard deviation highlighted that the ASR struggled more with some sentences than others. The low level of variety in both network conditions simulated and speech content unfortunately limited the value of this test.

IV. Discussion

The growing popularity of WebRTC based voice communication platforms has made the WebRTC ubiquitous in web applications. Thus, the ability of objective speech quality models to deal with jitter buffer adjustments in a perceptually sound manner remains important today.

A. Playout delay adjustments

The results of the first experiment confirmed that the significant issues that had been identified and measured in [4], [5] for objective speech quality metrics PESQ, POLQA and ViSQOL have been resolved with the recent algorithm updates to POLQA and ViSQOL.

B. WebRTC jitter buffer playout modifications

Jitter buffers are employed in packet-switched networks in order to smooth the inter-arrival jitter of incoming packets for an uninterrupted playout [19]. IETF Working Group for WebRTC [20] does not identify any specifics on the jitter buffer strategy for WebRTC. WebRTC open source project [21], however, includes a component called NetEq which has an adaptive jitter buffer implementation [22] and its specifics and design principles are not documented by the project, except the code base which has been in existence since Google open sourced it in 2011. Therefore, the evolution, principles, and prospects of the jitter buffer strategy for WebRTC is yet to be further explored. Unlike WebRTC, 3GPP has a history of standardised jitter buffer specifications [23], last updated in 2019, for Enhanced Voice Services used in LTE. While jitter buffer strategies in WebRTC and 3GPP have been the most prevalent ones seen in industry and the recent literature, there are a large number of proposals and patents for jitter buffer strategies.

Fig. 4. WebRTC VAD for Reference files 7 and 8 overlayed with Jitter Buffer B VAD. The loss of speech from utterance 1 and time-scale modification of utterance 2 are visible.

Jitter buffer strategy for WebRTC is studied in [22], [24], [25]. While the WebRTC jitter buffer strategy works well under normal network conditions, it suffers when there is high jitter in the network, such as the LTE scenarios as it is shown in [22]. It should be noted that the test conditions in the second experiment in this paper were prepared based on the real world measurements from [22].

Playout adjustments by jitter buffers can mean more than just insertions or deletions of short silence periods and can also result in time-scale modifications on the speech, in the form of lengthening or shortening of the decoded signals, in order to maintain speech intelligibility. Furthermore, the jitter buffer strategy has an impact on the late packet loss depending on the algorithm and configuration parameters. The two jitter buffer strategies tested in the second experiment adopted different approaches in this regard. While Jitter Buffer A has a naïve jitter buffer management and time-scale modification strategy, Jitter Buffer B adopts a more aggressive one in order to maintain an intelligibility under network conditions with extreme jitter. This results in scenarios where voice quality (naturalness, pitch etc.) are compromised in favour of intelligibility.

While the POLQA and ViSQOL objective models have algorithms to counteract the small and imperceptible time-scale modification applied by the traditional jitter buffers, they are designed to penalise large time-scale modification instances that are perceptible as unnatural by listeners [26].

The results of the second experiment show that the objective quality metrics predicted poor speech quality resulting from the speech loss and the unnatural voice quality alterations. The models also agreed with the Expert MOS test that voice quality issues are preferred over intelligibility issues.

The VAD examples depicted in Fig. 3 illustrates this as there was greater than 3% difference in the Jitter Buffer speech activity but the first utterance is delayed and the second utterance is compressed as shown for the samples 7 and 8 in Fig. 4.

Disagreement between the STOI and POLQA Intelligibility and their predictions relative to the speech quality models point to the need for a better understanding of the relationship between listener preferences regarding an intelligibility and a voice quality. This could yield combined models that would unify the results from objective metrics regarding an intelligibility and a quality.

A subjective study to explore these aspects in more detail is currently being developed. Additionally, the set of test conditions in the second experiment included a limited number of conditions. A future study with extensive test network conditions will explore different types of jitter buffer adjustments. Such adjustments could be one, or a combination, of time-scale modifications, packet discard (late loss), silence insertion or deletion, or other manipulations that may be occurring under a variety of test conditions.

V. Conclusions and Future Work

This work investigated whether the latest revisions to the POLQA and ViSQOL objective speech quality metrics have addressed the prediction issues seen with the earlier versions of these metrics. The results of the first experiment show that for the imperceptible playout adjustments the latest models are able to accurately predict the quality perceived by the end user. The second experiment explored how the models dealt with more extreme scenarios containing severe playout delays, speech loss and time-scale modifications of speech. The comparison of the results for the two jitter buffer algorithms showed that the objective quality models presented parallel behaviour with the subjective evaluation but the objective intelligibility results provided by the POLQA Intelligibility and STOI models were conflicting. A followup study will conduct two independent evaluations for a subjective quality and intelligibility respectively, to explore the relationship between the objective intelligibility and quality results.

Acknowledgement

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2289_P2 and 13/RC/2077.

References

[1] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 13, 2015.

[2] ITU, "Perceptual objective listening quality assessment." Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863, 2011.

[3] ITU, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862, 2001.

[4] P. Počta, H. Melvin, and A. Hines, "An analysis of the impact of playout delay adjustments introduced by voip jitter buffers on listening speech quality," *Acta Acustica united with Acustica*, vol. 101, no. 3, pp. 616–631, 2015.

[5] A. Hines, P. Počta, and H. Melvin, "Detailed comparative analysis of PESQ and VISQOL behaviour in the context of playout delay adjustments introduced by VOIP jitter buffer algorithms," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 18–23, IEEE, 2013.

[6] ITU, "Perceptual objective listening quality assessment." Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863, 2018.

[7] T. Mo and A. Hines, "Jitter buffer compensation in voice over ip quality estimation," in *2019 30th Irish Signals and Systems Conference (ISSC)*, pp. 1–6, IEEE, 2019.

[8] ITU, "ITU-T coded-speech database." Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.Sup23, 1998.

[9] ITU, "ITU-T methods for subjective determination of transmission quality." Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.800, 1996.

[10] Laerd statistics, "One-way anova." https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php, 2020.

[11] ITU, "Test signals for use in telephonometry." Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P501, 2018.

[12] "Python interface to the webrtc voice activity detector." https://github.com/wiseman/py-webrtcvad.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[14] P. Počta and J. G. Beerends, "Subjective and objective measurement of synthesized speech intelligibility in modern telephone conditions," *Speech Communication*, vol. 71, pp. 1–9, 2015.

[15] P. Počta and J. G. Beerends, "Subjective and objective measurement of the intelligibility of synthesized speech impaired by the very low bit rate stanag 4591 codec including packet loss," *Acta Acustica united with Acustica*, vol. 103, no. 2, pp. 311–316, 2017.

[16] "Google cloud speech api." https://cloud.google.com/speech-to-text/.

[17] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, p. 31–88, Mar. 2001.

[18] N. Vaessen, "Word error rate for automatic speech recognition." https://github.com/jitsi/asr-wer/, Feb 2019.

[19] B. Oklander and M. Sidi, "Jitter buffer analysis," in *2008 Proceedings of 17th International Conference on Computer Communications and Networks*, pp. 1–6, Aug 2008.

[20] H. T. Alvestrand, "Overview: Real Time Protocols for Browser-based Applications," Internet-Draft draft-ietf-rtcweb-overview-19, Internet Engineering Task Force, Nov. 2017.

[21] WebRTC, "Webrtc." https://webrtc.org, Jan 2020.

[22] N. Majed, S. Ragot, X. Lagrange, A. Blanc, J. Dufour, and G. Grao, "Experimental evaluation of webrtc voice quality in lte coverage tests," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, 2017.

[23] 3GPP, "Codec for Enhanced Voice Services (EVS); Jitter Buffer Management," Technical Specification (TS) 26.448, 3rd Generation Partnership Project (3GPP), 2019. Version 16.0.0.

[24] Y. Cinar, H. Melvin, and P. 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," *Communications-Scientific letters of the University of Zilina*, vol. 18, no. 1, pp. 17–22, 2016.

[25] N. Majed, *Measuring and improving the quality of experience of mobile voice over IP*. PhD thesis, Ecole nationale supérieure Mines-Télécom Atlantique, 2018.

[26] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "VISQOL: The Virtual Speech Quality Objective Listener," in *IWAENC*, 2012.

View publication stats