Measuring and Monitoring Speech Quality for Voice over IP with POLQA, ViSQOL and P.563

Andrew Hines[1,2], *Eoin Gillen*[2], *Naomi Harte*[2]

[1]School of Computing, Dublin Institute of Technology, Ireland
[2] Sigmedia, School of Engineering, Trinity College Dublin, Ireland
andrew.hines@dit.ie

10.21437/Interspeech.2015-171

Abstract

There are many types of degradation which can occur in Voice over IP (VoIP) calls. Of interest in this work are degradations which occur independently of the codec, hardware or network in use. Specifically, their effect on the subjective and objective quality of the speech is examined. Since no dataset suitable for this purpose exists, a new dataset (TCD-VoIP) has been created and has been made publicly available. The dataset contains speech clips suffering from a range of common call quality degradations, as well as a set of subjective opinion scores on the clips from 24 listeners. The performances of three objective quality metrics: POLQA, ViSQOL and P.563, have been evaluated using the dataset. The results show that full reference metrics are capable of accurately predicting a variety of common VoIP degradations. They also highlight the outstanding need for a wideband, single-ended, no-reference metric to monitor accurately speech quality for degradations common in VoIP scenarios.

Index Terms: Speech Quality, VoIP, POLQA, P.563, ViSQOL

1. Introduction

The growth in high speed mobile and fixed broadband has seen Voice over Internet Protocol (VoIP) services adopted by both consumer and business users as a viable alternative and potential replacement to Public Switched Telephone Networks (PSTNs) for domestic and international voice calls [1, 2]. Services such as Google Hangouts and Skype provide a variety of free and paid VoIP services from two-way voice-only calls to multi-party video and voice conferences. Measuring and monitoring speech quality for VoIP applications are different goals and are fulfilled by different objective speech quality models. During development and testing of VoIP systems objective metrics can be used to measure and predict speech quality. For deployed systems, realtime monitoring is essential to provide predictions of the actual speech quality experienced by users of the VoIP applications. While subjective testing is considered the gold standard for evaluating speech quality, objective metrics are essential to the application developers and network system operators to ensure changes to the platform do not have a negative impact on users' quality of experience (QoE).

Metrics can be classified into categories: parametric and signal-based models. The main types of signal based models are full-reference and no-reference. A parametric model estimates quality using rules based analysis of parameters of the network, signal and degradation. They are useful for network planning and can estimate the quality based on the relationship between factors, e.g. speech signal bit-rate and network bandwidth available. Full-reference signal based metrics use a clean, undistorted signal as a reference, and the degraded signal under evaluation is compared to the reference to predict quality. No-reference metrics predict the speech quality based on an evaluation of the degraded signal only and as-such are sometimes referred to as single-ended. Full-reference metrics are generally more accurate and are useful for the measuring speech quality in the development and evaluation of VoIP services, while no-reference metrics can be deployed for realtime monitoring where access to a clean reference signal may not be practical [3]. Whether full-reference or no-reference, signal based metrics are usually designed to predict the quality of a speech signal on using a 5 point rating that mimics the mean opinion scores (MOS) from subjective listener tests. This paper evaluates two full-reference metrics, namely POLQA and ViSQOL and one no-reference metric, P.563. These are described in more detail in section 3.

Many speech quality databases have been developed to evaluate speech quality and to develop and train objective speech metrics (e.g. see [4] for a list of databases benchmarked in the standardisation of POLQA). However, due to their valuable proprietary nature, many databases are not publicly available. As such, a newly developed database of wideband speech with a wide variety of degradations that are common to VoIP scenarios was developed by the authors, called TCD-VoIP [5]. It is used here to evaluate POLQA, ViSQOL and P.563 metrics' speech quality prediction for a variety of VoIP scenarios. There have been some studies evaluating POLQA for packet loss and jitter in VoIP conditions [6, 7] and to evaluate Skype with POLQA [8]. To our knowledge, to date no work has examined the robustness of these metrics within this specific and increasingly important domain. The results also have wide applicability as the degradations used (background noise, competing speaker, echo effects, clipping effects and choppy speech) occur independently of the codec, network or hardware.

The paper is laid out as follows. Section 2 describes the TCD-VoIP database and section 3 introduces the objective metrics that were benchmarked in this study. Section 4 describes the evaluation method and section 5 presents and discusses the results before concluding remarks are made in section 6.

2. TCD-VoIP

The TCD-VoIP dataset [5] is a publicly available dataset containing degradations common to VoIP [1]. It contains five categories of degradation: background noise, competing speaker, echo effects, clipping effects and choppy speech. The degradations are platform-independent, i.e. they are conditions that occur independently of the codec, network or hardware.

[1]Available for download from: www.sigmedia.tv/resources

Copyright © 2015 ISCA

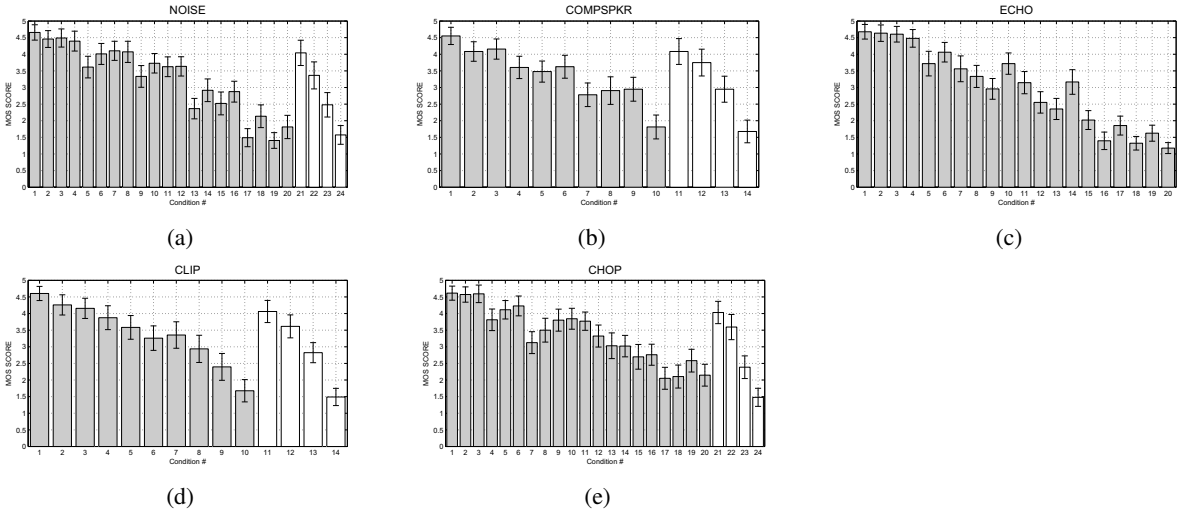

Figure 1: The subjective MOS results from the five degradation types. The MOS value for a condition is the average score given by all listeners to the four speech samples affected by that condition. The error bars are 95% confidence intervals obtained using the method in ITU-T Rec. P.1401 [9]. The 1st condition in each figure represents an the clean reference condition. The MNRUs are highlighted in white. The bar charts highlight that conditions covering the ACR quality range were tested across each degradation type.

The database has been subjectively labelled with listener tests complying with the ACR methodology presented in ITU-T Rec. P.800 [10]. The TSP speech database from McGill University in Canada [11] provided the reference speech material. It was recorded in an anechoic chamber and consists of speakers reading sentences from the Harvard test sentence list. The speech samples in the TSP speech database are 16-bit WAV files sampled at 48 kHz.

Full information on the degradation types in the TCD-VoIP dataset can be found in [5] with a summary presented here in Table 1. The database was designed to have each type of degradation spanning the full MOS range, i.e. from Bad to Excellent. The per-condition results, grouped by degradation class are presented in Figure 1. For chop, the degradation varied according to whether zeros were inserted to replace samples, or whether they were deleted or overwritten with earlier samples. The chop period refers to the length of the chopped sample; the rate to how often the samples got chopped. To create clipped samples, a multiplier was applied to the original signal and values over the maximum value were simply clipped to that level. Competing speakers is treated as a separate issue to large crowd babble noise as the speaker is intelligible and this is a common VoIP call scenario. The gender and SNR level of the competing speaker were varied. For echo, the alpha value is the % amplitude of the first delayed version of the signal relative to the original. The delay parameter was the number of ms before the first delayed version of the signal relative to the original. Four types of noise were used: speech babble noise; car noise; road noise and office noise. The SNR was also varied for these noise degradations.

Aside from the VoIP conditions, Modulated Noise Reference Unit (MNRU) conditions were also included in the tests (see ITU-T Rec. P.810 [12] and [5] for further details). 24 listeners were used in all experiments and each condition was tested with 4 speakers (2 male and 2 female).

Table 1: Summary of Degradations and Parameters used in TCD-VoIP

Degradation	Conditions	Parameters	Range
Chop	20	Rate	0-6 chops/s
		Period	0.02-0.04 s
		Mode	Insert, Delete, Overwrite
Clip	10	Multiplier	1-55
Competing Speaker	10	Gender code	1-5
		SNR	10-50 dB
Echo	20	Alpha	0-0.5
		Delay	0-220 ms
Noise	20	Noise Type	Car, Street, Office, Babble
		SNR	5-55 dB
MNRUs	4	SNR (Q)	48, 36, 24, 12

3. Objective Speech Quality Metrics

As mentioned in the introduction, different application scenarios necessitate different speech quality models. The ITU G.107 E-Model [13, 14] is an example of a parametric model and is not investigated here as this paper is focused on evaluating signal-based models. Full-reference models, can produce accurate measurements of speech quality by comparing a reference and test signal. Monitoring models are no-reference and estimate the speech quality from the test signal without access to a clean reference to compare against. With access to more information, full-reference metrics can generally produce more accurate predictions than no-reference metrics but are not suitable for deployment as realtime monitoring metrics in VoIP applications.

The performances of three objective metrics: POLQA, ViSQOL and P.563, have been compared in this study. Prior to the development of POLQA, ITU-T Rec. P.862 [15] presented PESQ, a full-reference metric designed to estimate the quality of narrowband (300 - 3,400Hz) speech and networks. It first aligns the degraded and reference signals, and then compares both signals using a perceptual model. A subsequent revision to P.862 (P.862.2) [16] enabled PESQ to rate wideband (50 - 7,000Hz) signals. PESQ is widely used, however it is inaccurate in some scenarios: suboptimal listening levels, loudness loss, delays in conversational tests, talker echo or sidetone. These limitations are acknowledged in the recommendation.

POLQA, introduced in ITU-T Rec. P.863 [4], is seen as a successor to PESQ, and was designed to conform to newer

industry requirements and address acknowledged shortcomings of PESQ. The extended PESQ revision (P862.2) added wideband support to the model (50 – 7,000 Hz) while POLQA can rate signals with bandwidths of 50 –14,000 Hz (superwideband signals). The basic philosophy used in POLQA is the same as that used in PESQ, i.e. the algorithm first aligns the degraded and reference signals, and then compares both signals using a perceptual model. The POLQA algorithm also contains some additional steps and considerations designed to improve prediction accuracy of the metric. Despite this, some of the limitations of PESQ (specifically in the cases of delays in conversational tests, talker echo or sidetone) are still present in POLQA [17]. In this paper, we only assess the performance of POLQA for the VoIP scenarios, as its performance should be superior to PESQ.

ViSQOL [18] evolved from NSIM (Neurogram Similarity Index Measure), a tool developed by Hines and Harte [19] to predict speech intelligibility for hearing-impaired listeners. An overview of NSIM and ViSQOL is given in Hines et al. [20], and the performance of ViSQOL is compared to that of PESQ for two common VoIP issues (clock drift and jitter). NSIM is a full-reference metric which compares neurograms created from the reference and degraded signals. A simplified algorithm is used in ViSQOL, which compares spectrograms created from Short-term Fourier Transforms (STFT) of both signals. NSIM outputs a similarity score from 0 – 1, which is mapped to a MOS-LQO scale. Metrics such as POLQA and ViSQOL are useful when designing new VoIP systems or measuring and evaluating the performance of systems for particular scenarios. ViSQOL and POLQA have also recently been adapted to be for audio quality evaluation [21]. This work builds on prior work where POLQA, PESQ and ViSQOL were evaluated under a variety of narrowband speech scenarios with different datasets [22, 23, 24].

P.563, introduced in ITU-T Rec. P.563 [25], is a no-reference metric designed to estimate the quality of narrowband (100 – 3,100 Hz) speech. Sometimes referred to as single ended, no-reference metrics like P.563 predict speech quality without access to a clean reference signal. This class of metric is particularly useful in realtime monitoring scenarios where no reference signal is available to compare against. P.563 was designed to account for the full range of degradations present in PSTNs. To rate a speech signal without a reference, P.563 makes use of a large number of characteristic speech parameters, which can be split into 6 categories: basic speech descriptors, vocal tract analysis, speech statistics, static SNR, segmental SNR and interruptions/mutes. The output score is based on these parameters. Output scores have been mapped to MOS values using a set of speech clips and subjective test results. Some of its limitations include talker echo, sidetones and singing voice as well as limited testing with amplitude clipping during standardisation [3].

The metrics chosen for this test are the current recommended full-reference (POLQA) and no-reference (P.563) metrics and the full-reference metric ViSQOL that was developed to specifically target VoIP scenarios. It should be noted there are other speech quality metrics have been developed and are in common use that were not tested here (e.g. [26, 27, 16]). The three chosen metrics, provide a baseline benchmark for objective full-reference and no-reference metrics on this dataset.

4. Metric Evaluation

The subjective listener test mean opinion scores (MOS-LQS) for the database were compared with predictions from the three objective metrics. POLQA was tested using its super-wideband mode and ViSQOL was tested using its wideband speech mode. As P.563 is a narrowband metric the degraded signals were resampled at 8 kHz for testing with the no-reference metric. The test evaluated 384 sample speech files. For each condition, 4 samples (2 male and 2 female speakers) were tested giving 96 conditions. The objective mean opinion score (MOS-LQO) for the given condition was computed. The per-condition MOS-LQS were used to benchmark the metrics.

5. Results and Discussion

The results for the objective metrics POLQA, ViSQOL and P.563 on the dataset (denoted by MOS-LQO) are compared to the subjective results (denoted by MOS-LQS) in Figure 2 broken down by degradation condition. A statistical analysis of the object metric prediction accuracy compared with the subjective listener test results is presented in Table 2. Pearson correlation coefficients ($\rho_{pearson}$), Spearman rank order coefficients ($\rho_{spearman}$) and root mean squared error ($RMSE$) for each metric. The results are further broken down by condition type. As ViSQOL performed poorly with the CHOP data, two aggregated condition totals are displayed including and excluding the CHOP conditions.

POLQA performs well in predicting quality across all the degradation types tested. With careful review of Figure 2, it can been seen that POLQA has a general trend of over-estimating quality for noise, echo and competing speakers. It tends to underestimate for clip, with more consistent performance for chop. Examining the statistics in Table 2 confirms POLQA's ability to predict speech quality accurately across all condition classes in the TCD-VoIP database with high scores both in terms of Pearson correlation coefficients and Spearman rank correlations.

The quality predictions from ViSQOL are well-correlated with the subjective results in the clip, competing speaker, echo and noise tests. There is a general trend visible that ViSQOL underestimates quality for echo conditions. ViSQOL's scores on the choppy speech are not well-matched to the subjective scores. An analysis of the individual conditions found that the insert and delete conditions account for the over-estimated cluster (above the diagonal), while the underestimated cluster of 7 chop conditions seen significantly below the diagonal was composed exclusively of the overwrite chop sub-condition. When portions of the signal are overwritten, ViSQOL's comparison algorithm can be tricked into aligning speech segments with the overwritten repetition segment rather than the original segment. This causes mis-aligned comparisons with the reference spectrogram and results in a low speech quality estimate. Conditions using the other two chop modes do not cause this problem, in fact, ViSQOL marginally overestimated quality for these conditions. Overall, the correlation statistics reveal that the performance of ViSQOL and POLQA is close, particularly if the chop conditions are not taken into account. This is useful for researchers as ViSQOL is a freely available speech quality metric.

P.563 was the only no-reference metric tested in this work. As a no-reference metric, its scores were not expected to be as well-matched as those of POLQA or ViSQOL. However, as can been seen in Figure 2, P.563's predictions bear almost no relation to the subjective results. It appears that the lowest (MOS ≤ 2) subjective results also obtain the lowest P.563 results, but no further relationship can be discerned. Almost all of P.563's results lie between MOS values of 2.5 and 3.5. Looking at Ta-

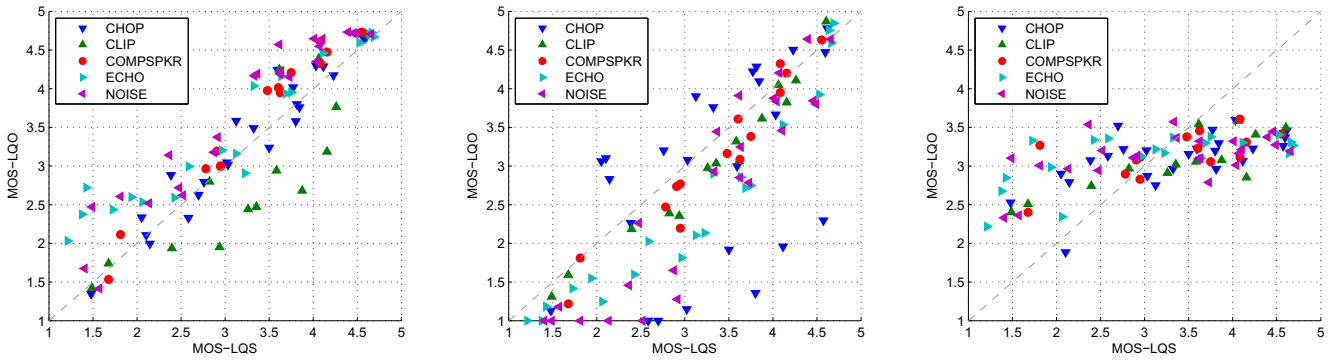

Figure 2: Scattter plots for POLQA, ViSQOL and P.563

Table 2: Pearson correlation coefficients, Spearman rank correlations and RMSE per condition. The results are broken down by degradation class with a grouped result for all conditions and a final grouping for all conditions excluding the chop condition.

		CHOP	CLIP	COMPSPKR	ECHO	NOISE	ALL	ALL (excluding CHOP)
$\rho_{pearson}$	ViSQOL	0.485	**0.978**	0.968	0.950	0.927	0.834	**0.939**
	POLQA	**0.968**	0.843	**0.986**	**0.957**	**0.960**	**0.900**	0.885
	P.563	0.628	0.823	0.661	0.640	0.547	0.630	0.638
$\rho_{spearman}$	ViSQOL	0.543	**0.987**	0.952	**0.979**	0.886	0.836	**0.939**
	POLQA	**0.955**	0.890	**0.965**	0.953	**0.952**	**0.903**	0.891
	P.563	0.617	0.719	0.626	0.520	0.469	0.561	0.552
$RMSE$	ViSQOL	0.381	**0.244**	**0.204**	0.557	0.533	0.772	0.593
	POLQA	**0.074**	0.387	0.263	**0.385**	**0.466**	**0.480**	**0.537**
	P.563	0.732	0.694	0.685	0.961	0.885	0.813	0.838

ble 2, the only test in which a a positive trend can be seen is the clipped speech test. This was somewhat of a surprise as it was noted earlier that amplitude clipping was a condition with limited testing during the metrics's development.

These results show that POLQA is capable of predicting the subjective MOS value of any condition in the TCD-VoIP dataset, although their predictions for some clipped speech may be slightly low. ViSQOL is capable of predicting the subjective MOS values for speech with clipping, noise, competing speakers or echo, but struggles with choppy speech, specifically in cases where portions of the signal have been overwritten.

The results for P.563 show that it is incapable of predicting subjective MOS values for conditions in the dataset. P.563's use cases (listed by Möller et al. [3]) are mostly for detecting signal warping or network effects. Also, two of its limitations are echo and clipping. From this, it can be concluded that P.563 is unsuitable for the task of rating clips in TCD-VoIP. A gap exists for a wideband speech quality metric capable of monitoring VoIP applications as none of the no-reference objective quality metrics currently available were specifically developed with this task in mind.

6. Conclusion

This paper reports benchmarking results of three speech quality metrics on a VoIP speech database. Two full-reference signal-based metrics were evaluated to establish their accuracy as measurement tools for speech quality. The results showed that for the classes of VoIP degradation tested, full-reference speech quality metrics can provide accurate predictions of speech quality and could be used in developing and testing. The results for the ITU recommended POLQA metric were consistent across all degradation classes, further validating its capabilities in a wide-range of speech scenarios. The tests also highlighted alignment problems for ViSQOL when choppy data uses a overwrite strategy repeating a previous segment of the speech. This will be investigated further in future development of the ViSQOL metric. For monitoring applications, the results showed that the no-reference metric tested, P.563, could not accurately predict quality to a usable level of accuracy. This highlights an important unaddressed requirement for VoIP applications, namely the need for a no-reference wideband speech quality metric capable of monitoring VoIP applications. The authors are currently using the findings presented to address the challenge of monitoring VoIP quality with a realtime, wideband no-reference metric.

7. Acknowledgements

Thanks to Google, Inc. and Enterprise Ireland for funding.

8. References

[1] Alcatel-Lucent, "PSTN industry analysis and service provider strategies: Synopsis," http://goo.gl/tTPFes, Alcatel-Lucent, Paris, France, Tech. Rep. Bell Labs Analysis for BT, 2013.

[2] L. K. Vanston and R. L. Hodges, "Forecasts for the us telecommunications network," *Telektronnik*, vol. 104, no. 3/4, pp. 18–28, 2008.

[3] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, 2011.

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

[5] N. Harte, E. Gillen, and A. Hines, "TCD-VoIP, a research database of degraded speech for assessing quality in voip applications," in *Quality of Multimedia Experience (QoMEX)*, Costa Navarino, Greece, 2015.

[6] M. Soloducha and A. Raake, "Speech quality of VoIP: bursty packet loss revisited," in *Speech Communication; 11. ITG Symposium; Proceedings of*, Sept 2014, pp. 1–4.

[7] J. Holub and O. Slavata, "Impact of IP channel parameters on the final quality of the transferred voice," in *Wireless Telecommunications Symposium (WTS), 2012*, April 2012, pp. 1–5.

[8] J. Zhu, R. Vannithamby, C. Rodbro, M. Chen, and S. Vang Andersen, "Improving QoE for Skype video call in mobile broadband network," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, Dec 2012, pp. 1938–1943.

[9] ITU, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," Int. Telecomm. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1401, 2012.

[10] ——, "Methods for subjective determination of transmission quality," Int. Telecomm. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.800, 1996.

[11] P. Kabal, "Tsp speech database," McGill University, Quebec, Canada, Tech. Rep. Database Version 1.0, 2002.

[12] ITU, "Modulated noise reference unit (mnru)," Int. Telecomm. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.810, 1996.

[13] ——, "The E-model, a computational model for use in transmission planning," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.107, 2009.

[14] ——, "Wideband E-model," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.107.1, 2011.

[15] ——, "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.

[16] ——, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862.2, 2005.

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

[18] 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. 2015:13, May 2015.

[19] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, no. 2, pp. 306 – 320, 2012.

[20] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The virtual speech quality objective listener," in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, Sept 2012, pp. 1–4.

[21] A. Hines, E. Gillen, J. Skoglund, A. Kokaram, and N. Harte, "Visqolaudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137:6, June 2015.

[22] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[23] A. Hines, P. Pocta, and H. Melvin, "Detailed analysis of PESQ and VISQOL behaviour in the context of playout delay adjustments introduced by VoIP jitter buffer algorithms," in *Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Wörthersee, Austria, 2013.

[24] P. Pocta, H. Melvin, and A. Hines, "An analysis of the impact of playout delay adjustments introduced by VoIP jitter buffers on speech quality," *Acta Acoustica united with Acustica*, vol. 101, no. 2, May–June 2015.

[25] ITU, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.563, 2011.

[26] ANSI ATIS, "0100005-2006: Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality," 2006.

[27] V. Grancharov, D. Zhao, J. Lindblom, and W. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1948–1956, Nov 2006.