2015

ViSQOLAudio: An Objective Audio Quality Metric for Low Bitrate Codecs

Andrew Hines
Technological University Dublin, andrew.hines@tudublin.ie

Eoin Gillen
University of Dublin, Trinity College

Damien Kelly
Google.inc.

See next page for additional authors

Recommended Citation

Hines et al. (2015) ViSQOLAudio: An Objective Audio Quality Metric for Low Bitrate Codecs, *JASA Express Letters*vol.137, no. 6. doi:10.1121/1.4921674

ViSQOLAudio: An objective audio quality metric for low bitrate codecs

Andrew Hines[a] **and Eoin Gillen**
Sigmedia Group, School of Engineering, Trinity College Dublin, Ireland
andrew.hines@dit.ie, ogiollae@tcd.ie

Damien Kelly, Jan Skoglund, and Anil Kokaram
Google, Incorporated, Mountain View, California 94043, USA
damienkelly@google.com, jks@google.com, anilkokaram@google.com

Naomi Harte
Sigmedia Group, School of Engineering, Trinity College Dublin, Ireland
nharte@tcd.ie

Abstract: Streaming services seek to optimise their use of bandwidth across audio and visual channels to maximise the quality of experience for users. This letter evaluates whether objective quality metrics can predict the audio quality for music encoded at low bitrates by comparing objective predictions with results from listener tests. Three objective metrics were benchmarked: PEAQ, POLQA, and VISQOLAudio. The results demonstrate objective metrics designed for speech quality assessment have a strong potential for quality assessment of low bitrate audio codecs.

[DOS]
Date Received: October 31, 2014 **Date Accepted:** May 12, 2015

1. Introduction

Media streaming is now an established method for listening to music and watching movie and TV content. Network bandwidth constraints are variable across the diverse range of devices on which content is consumed (e.g., mobile, desktop, home theatre). As a result, content distributors, such as YouTube, Netflix, or Spotify, must support a range of codecs and bit rates ("treatments") to optimise consumers' Quality of Experience (QoE).[1] Bandwidth allocation for multimedia streaming on mobile devices needs to optimise the QoE across senses and while 256 kb/s treatments can deliver QoE indistinguishable from uncompressed audio,[2] any savings in the audio bandwidth could be used to improve the video stream. To accommodate environments from smartphone to home theatre, both the video and audio content are transmitted in compressed form using lossy compression schemes. Psychoacoustic inspired compression schemes have output signals that are optimised from the perspective of the human auditory system.[3–6] Standard mean square error or segmental signal-to-noise ratios are not well suited to evaluating audio quality for modern codecs due to the optimised bit allocation resulting from psychoacoustic models.[7,8] Suitable objective metrics could help automate the evaluation of changes in QoE as a result of this transcoding process. Treatments that are commonly used for streaming are examined here: AAC-HE and AAC-LC codecs[9] at four bit rates and examples of MP3 and OPUS codecs.[10] Prior work by the authors presented subjective listener test results for these treatments.[2] The treatments were subjectively evaluated in the context of stereo music to investigate whether listeners perceived differences in the audio quality for the codecs tested and whether presentation mode

[a]Current address: School of Computing, Dublin Institute of Technology, Dublin, Ireland.

(headphones/speakers) influenced the results. The results showed that listeners found some treatments were noticeably degraded while others were indistinguishable from the original uncompressed stereo audio samples. This work benchmarks four objective metrics against the subjective listener test results carried out with headphones to evaluate their suitability for measuring audio quality for low bit rate codecs.

2. Objective metrics

Since the late 1980s objective quality metrics for audio have been actively developed. The Perceptual Evaluation of Audio Quality (PEAQ) was standardised as Recommendation ITU-R BS.1387 from 1998 to 2001.[11] There are two versions: a basic version, which is optimised for speed, and an advanced model that adds a filterbank based ear model to the basic FFT-based model to improve accuracy. Both versions produce a number of model output variables that are mapped to an objective difference grade (ODG) quality score via a multi-layer neural network. The ODG scale is an objective approximation of the subjective difference grade used in Recommendation ITU-R BS.1116 (Ref. 12) to determine small audio impairments. A decade later, ITU-T Recommendation P.863 standardised a new objective metric for measuring speech quality, called POLQA. POLQA was designed for speech quality assessment and can be run in narrowband mode (telephone quality; 300–3400 Hz) or superwideband (SWB) mode (50–14 000 Hz). POLQA has been shown to have potential as an audio quality model[13] and the developers of POLQA are currently working on adapting it for use in audio quality evaluation. An alternative speech quality model called ViSQOL,[14] has been adapted for audio quality testing. This adapted metric is referred to as ViSQOLAudio and is described and evaluated in this letter. For benchmarking purposes, the commercially released POLQA version conforming to the P.863 standard is tested along with the basic and advanced versions of PEAQ (all supplied by Opticom, GmBH).

2.1 ViSQOLAudio metric

ViSQOL was developed as an objective speech quality metric.[14–16] It is a full reference speech quality metric, comparable to POLQA, that uses similarity between spectrograms to measure quality. This paper presents ViSQOLAudio, an adaption of the ViSQOL speech quality metric. A detailed description of ViSQOL for speech quality can be found in a previously published work.[14] A small number of changes to the algorithm were necessary for music and audio evaluation and they can be summarised as follows. The voice activity detection was removed to allow a comparison of the reference sample to the test sample. The basic system of comparing signals over spectrogram "patches" used by ViSQOL was retained and the patch alignment system is still performed between the reference and test patches but all patches from the reference signal are retained, i.e., all patches are considered "active." The number of frequency bands evaluated was increased with standard Bark scale bands used from 50 to 13 500 Hz with an extra band centred at 16 000 Hz to cover the full bandwidth of hearing from 50 Hz to 20 kHz. This adjustment was necessary for audio evaluation as ViSQOL only covers the salient speech frequency bands, i.e., between 50 and 8000 Hz. The mapping from the raw similarity score to a speech MOS score was removed for ViSQOLAudio and the results are quoted on a similarity scale from 0 to 1. A MATLAB implementation of ViSQOLAudio is available to download from the authors' website.[17]

3. Subjective and objective testing

The subjective tests were carried out using the MUSHRA test methodology that is defined in ITU-R Recommendation BS.1534.1.[18] Other audio quality test methodologies exist, but for low bit rate codec testing, MUSHRA is a good compromise between an absolute category rating test (e.g., ITU-R BS.1284-1) and a test for almost undetectable impairments (e.g., ITU-R BS.1116-2). Biases in MUSRHA tests have been reported due to stimulus spacing and range equalising effects[19] but MUSHRA has been used in a variety of tests showing a good ability to rank low bit rate codecs.[20,21]

Following the MUSHRA methodology, listeners were presented with a labeled reference and a number of unlabeled test samples (stimuli). The unlabeled samples were ranked using a numerical continuous scale ranging from 0 to 100 in five descriptive intervals: bad (0–20); poor (20–40); fair (40–60); good (60–80); and excellent (80–100). An unaltered version of the reference and two anchor samples (low-pass filtered versions of the reference) were hidden amongst the treatments under test. Ten listeners ranked 12 music samples for 10 treatments (including the hidden reference and 2 anchors). Tests using Sennheiser HD558, high-quality open-backed headphones showed the least variance between listeners and are used as the ground truth subjective quality for objective metric evaluation. The subjective quality assessments by treatment are reproduced here in Fig. 1. Table 1 details the test material which consisted of stereo music samples of 7–15 s duration covering a variety of musical sounds. The test materials were sourced from CDs and the EBU music database[22] and were all originally sampled at either 48 or 44.1 kHz, 16 bit stereo (so for 44.1 kHz two-channel audio, the bit rate is 1411.2 kb/s). Reference PCM WAV files were created at 48 kHz for all files. These were then coded and resampled using ffmpeg with Fraunhofer AAC encoder for AAC, libmp3lame for MP3, and libopus for Opus 1.1 to produce the range of treatments in Table 2. All samples were formatted as WAV PCM files for presentation and evaluated by the authors to ensure no level difference was perceived between the reference samples and the treatments. Each objective metric was used to compare the 12 reference samples with their 10 treatments. Further information on the subjective testing and a detailed analysis of the results was presented in prior work.[2]

Determining the influence of codecs and bitrates on audio is difficult. Aside from the inherently challenging task of quantifying subjective listener opinions on quality, measuring audio quality must eliminate the influence of other factors on the listener's quality of experience. Measuring with loudspeakers introduces the potential for room acoustics to influence results, while the choice of samples, rating methodology, number and expertise of listeners can potentially influence results. Music samples sensitive to bitrate reduction and frequency response were carefully chosen to exercise the codecs.[22] The listener tests were carried out in a sound-proofed recording studio and repeated using two types of headphones and again with loudspeakers to evaluate the impact of listening equipment. The results using studio quality Sennheister HD558 headphones were used as the subjective ground truth. While the other listening equipment results exhibited the same trends, there were differences that may have been caused by masking of compression artefacts in the lower quality headphones and due to room reverberation for the loudspeakers. The choice of MUSHRA as a testing methodology allowed for listeners to compare treatments and rank them on a continuous scale. The subjective results across treatments and equipment indicated that listeners were remarkably consistent in their scoring.[2] The data and results provided

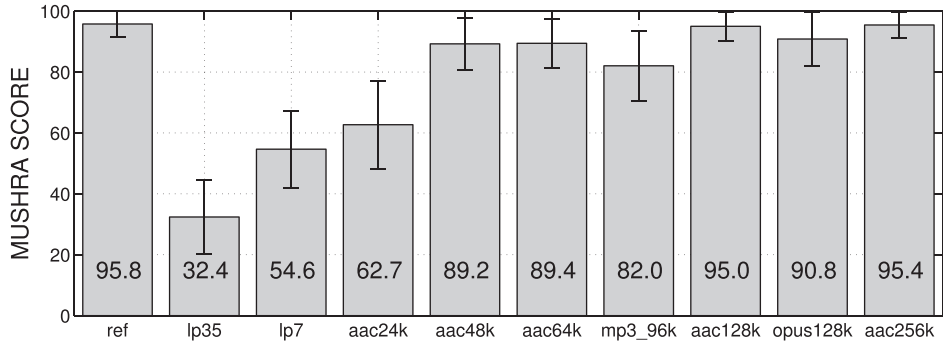

Fig. 1. Mean subjective MUSHRA results by treatment (for 12 music samples and 10 listeners using headphones). Error bars show 95 confidence intervals. lp35 and lp7 are the low pass filtered anchor conditions.

Table 1. Music samples.

Label	Music type	Source	Label	Music type	Source
Boz	Rock/R&B (Boz Scaggs)	CD	Glock	Glockenspiel	EBU
Steely	Soft Rock (Steely Dan)	CD	Contrabassoon	Arpeggio/melodious phrase	EBU
Castanets	Castanets	EBU	Harpsichord	Arpeggio/melodious phrase	EBU
Moonlight	Piano (Moonlight Sonata)	CD	Soprano	Soprano singer	EBU
Vega	Vocals (Suzanne Vega)	CD	Guitar	Larry Coryell	EBU
Ravel	Tzigane	EBU	Strauss	R. Strauss (Orchestra)	EBU

confidence that the listener test experiments would be a useful starting point for assessing the potential of objective metrics to predict low bitrate codecs impact on audio quality.

4. Results

Figure 1 reproduces the results from the subjective listener tests. The higher bitrate codecs had quality scores almost indistinguishable from the reference uncompressed audio. The low pass filtered anchors and AAC-HE 24 kb/s were scored significantly lower.

Figure 2 presents the objective results grouped per treatment type. The AAC-LC 128 kb/s and AAC-LC 256 kb/s were ranked highest by all of the objective metrics in terms of their mean values. However, the error bars show that the results for a given treatment varied significantly across samples. For example, observe the standard deviation for the AAC-LC 256 kb/s as measured by PEAQ. An analysis of the per sample results that caused this large standard deviation showed that both PEAQ models had estimates that clustered bi-modally with four samples between 0 and −1 (boz, castanets, steely, vega) and another six samples between −3 and −4 (sopr, ravel, guitar, harpsichord, contrabassoon, strauss). This is illustrated in Fig. 2 by the per-sample red circles on the PEAQ-Advanced plot for AAC-LC 256 kb/s. This indicates that for this particular codec, PEAQ exhibits a sensitivity to sample type that was not experienced by the subjective listeners.[2] No obvious pattern with respect to sample type was observed for these clusters.

Accurate estimation of the quality of the low pass anchors was a problem for all of the metrics. While the subjective tests show that listeners rank the low pass treatments very poorly in comparison with even the very low bitrate AAC-HE 24 kb/s codec, this is not captured in the objective metric predictions.

Table 3 contains correlation coefficients for all metrics against the mean subjective scores grouped by treatment. Figure 2 gives an indication of the performance and also highlights the poor estimation of the anchors by all of the objective metrics tested. The Spearman and Kendall correlation results were very promising for VISQOLAudio, pointing to a strong capacity to rank the treatments correctly. However, tight clustering of a number of the treatments and the confidence intervals of the subjective ground truth MUSHRA scores should be taken into account before over interpreting the meaning of these statistics. POLQA and ViSQOLAudio are

Table 2. Treatments.

Type	Bandwidth	Bit rate (kb/s)	Type	Bandwidth	Bit rate (kb/s)
reference	22 kHz/Raw-PCM	1536	aac-he	20 kHz (fullband)	48
anchor 1	3.5 kHz narrowband	256	aac-he	20 kHz (fullband)	64
anchor 2	7 kHz wideband	512	aac-lc	20 kHz (fullband)	128
mp3	16 kHz (SWB)	96 (CBR)	opus	20 kHz (fullband)	128
aac-he	20 kHz (fullband)	24	aac-lc	20 kHz (fullband)	265

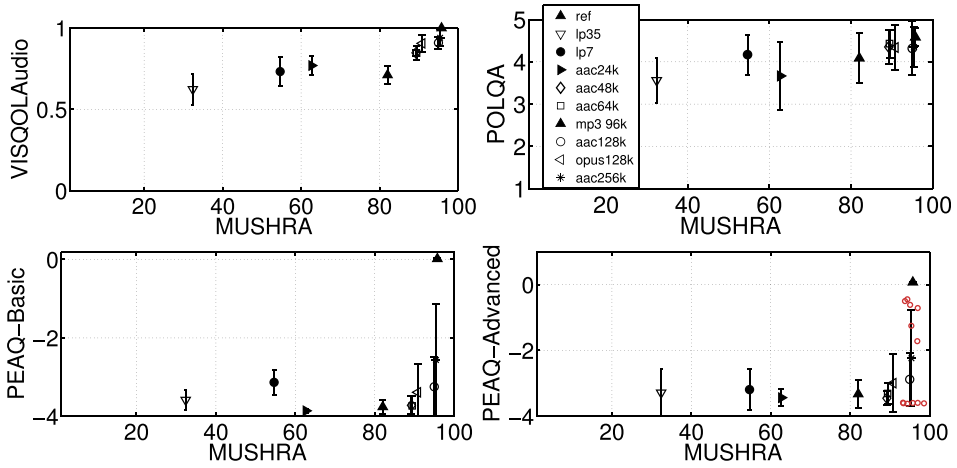

Fig. 2. (Color online) Mean objective metric results plotted against subjective MUSHRA quality scores per treatment for 12 samples. Error bars are 1 standard deviation. Red circles (unfilled) in PEAQ-Advanced plot are the individual sample results for aac256k.

comparable in their performance. Figure 3 presents a close-up view for scores above 60 on the MUSHRA scale for ViSQOLAudio and POLQA, i.e., excluding the two anchor treatments. POLQA had a good fit to all treatments with the exception of the anchors, while VISQOLAudio overestimates degradation in quality for the MP3 96 kb/s treatment. The robustness of metrics to the type of sample content was better with VISQOLAudio than with POLQA or PEAQ, highlighted by the size of the error bars.

Table 3 also presents statistics excluding the two anchor points. ViSQOLAudio is the most consistent across both scenarios although the correlation scores dropped without the anchor treatments. The RMSE has dropped for all measures having removed the anchor treatments. Without them, the correlation statistics for PEAQ have noticeably improved but the high RMSE still highlights a lack of robustness between audio samples for a given treatment. The Pearson correlation for POLQA is also noticeably higher with these exclusions but the Spearman ranking warns that listeners' treatment preferences are not captured.

5. Discussion and conclusions

The results for all metrics are promising, despite the fact that PEAQ was not designed with low bitrate codecs in mind and POLQA and ViSQOL were conceived as speech

Table 3. Correlation statistics for raw and regression fitted mean treatment scores. Top ranked measure for each statistic in bold.

	Raw			Fitted			
Measure	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	RMSE
POLQA	0.859	0.77	0.77	0.86	0.77	0.6	14.06
VISQOLAudio	**0.866**	**0.952**	**0.867**	**0.895**	**0.952**	**0.867**	**9.94**
PEAQ-Advanced	0.673	0.556	0.397	0.4	0.673	0.556	18.93
PEAQ-Basic	0.337	0.564	0.511	0.293	0.273	0.156	20.08
(Without anchor 1 and anchor 2 treatments)							
POLQA	**0.943**	0.619	0.619	**0.948**	0.619	0.5	6.724
VISQOLAudio	0.74	0.952	0.857	0.8	0.929	0.786	**6.262**
PEAQ-Advanced	0.929	0.857	0.474	0.422	0.833	0.714	9.452
PEAQ-Basic	0.479	**0.976**	**0.929**	0.718	**0.976**	**0.929**	7.892

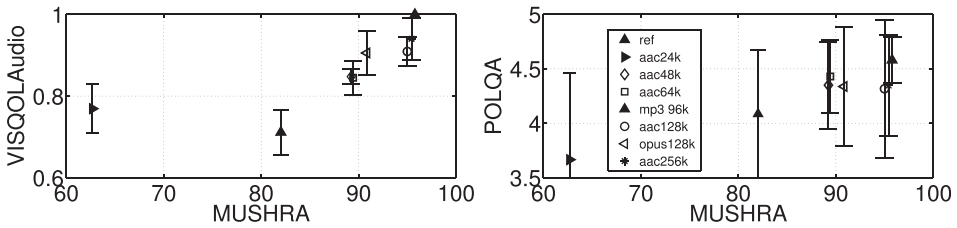

Fig. 3. Close-up of Fig. 2 for ViSQOLAudio and POLQA. Anchor treatments are excluded and range of axes has been reduced.

quality metrics. The adaptations to create ViSQOLAudio were essential as the speech version had no superwideband mode and was optimised for speech evaluation using a voice activity detector. The low standard deviation between samples for treatments is the most advantageous feature when comparing results with the other objective metrics.

POLQA performed well using the SWB speech mode. The statistics at a treatment level showed POLQA's potential for this application but the ranking ability and sample level variation pointed towards difficulty in distinguishing between the smaller differences in treatments. This could well be addressed by the as yet unpublished audio-modified version of POLQA where alignment changes similar to those applied to ViSQOLAudio will likely be addressed. The results suggest only small adaptations are required. PEAQ showed much variation across different samples for the same treatment where the listener assessed quality was consistent. This letter highlights the need for continued development of objective metrics that can deal with the variety of new low bitrate codecs that have been developed. Further work and testing with a wider range of data is ongoing. ViSQOLAudio has promising potential as an objective audio quality metric and compares favourably with PEAQ for the lower bitrate codecs assessed.

Acknowledgments

Supported by Google, Inc. and Enterprise Ireland, Innovation Partnership Project No. IP-2013 0232.

References and links

[1]European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Qualinet white paper on definitions of quality of experience, 2012.

[2]A. Hines, E. Gillen, J. Skoglund, D. Kelly, A. Kokaram, and N. Harte, *Perceived Audio Quality for Streaming Stereo Music* (ACM Multimedia, Orlando, 2014).

[3]R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May, 2002), Vol. 2, pp. II-1809–II-1812.

[4]B. C. J. Moore, "Masking in the human auditory system," in *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction* (May, 1996).

[5]T. Painter and A. Spanias, "Perceptual coding of digital audio," Proc. IEEE **88**(4), 451–515 (2000).

[6]E. Zwicker and U. T. Zwicker, "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system," J. Audio Eng. Soc. **39**(3), 115–126 (1991).

[7]S. Kandadai, J. Hardin, and C. D. Creusere, "Audio quality assessment using the mean structural similarity measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008* (2008), pp. 221–224.

[8]J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis, "Perceptual-based quality assessment for audio–visual services: A survey," Sign. Process.: Image Commun. **25**(7), 482–501 (2010).

[9]ISO/IEC 13818-7:2006: *Information Technology—Generic Coding of Moving Pictures and Associated Audio Information—Part 7: Advanced Audio Coding (AAC)* (International Organization for Standardization, Geneva, Switzerland, 2006).

[10]J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus audio codec," IETF (2012), URL RFC6716, http://www.ietf.org/rfc/rfc6716.txt (Last viewed 21 May 2015).

[11]Rec.ITU-R.BS.1387: *Perceptual Evaluation of Audio Quality (PEAQ)* (International Telecommunication Union, Geneva, 1998).
[12]Rec.ITU-R.BS.1116-2: *Methods for the Subjective Assessment of Small Impairments in Audio Systems* (International Telecommunication Union, Geneva, 2014).
[13]M. H. Pinson, C. Schmidmer, L. Janowski, R. Pepion, Q. Huynh-Thu, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "Subjective and objective evaluation of an audiovisual subjective dataset for research and development," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)* (2013), pp. 30–31.
[14]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 *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013).
[15]A. Hines, P. Pocta, 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)* (July, 2013), pp. 18–23.
[16]A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The virtual speech quality objective listener," in *2012 International Workshop on Acoustic Signal Enhancement (IWAENC)* (September, 2012), pp. 1–4.
[17]ViSQOLAudio, "MATLAB [computer program]," http://www.sigmedia.tv/tools (Last viewed 21 May 2015).
[18]Rec.ITU-R.BS.1534-1: *Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)* (International Telecommunication Union, Geneva, 2003).
[19]S. Zielinski, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential biases in MUSHRA listening tests," in *Audio Engineering Society Convention 123* (2007).
[20]C. Hoene, J.-M. Valin, K. Vos, and J. Skoglund, "Summary of Opus listening test results—Internet-draft," IETF (2012), http://tools.ietf.org/html/draft-ietf-codec-results (Last viewed 21 May 2015).
[21]D. Marston and A. Mason, "Cascaded audio coding," EBU technical review (2005), https://tech.ebu.ch/publications/trev_304-cascading (Last viewed 21 May 2015).
[22]EBU Tech.3253-E: *Sound Quality Assessment Material [SQUAM CD (Handbook)]* (EBU Technical Centre, Brussels, 1988).