Machine Learning for Audio Coding
Internet-Draft
Intended status: Informational
Expires: 21 January 2026

L. Lechler
K. Wojcicki
Cisco Systems

20 July 2025

Test Battery for Opus ML Codec Extensions
draft-lechler-mlcodec-test-battery-00

Abstract

This document proposes methodology and data for evaluation of machine learning (ML) codec extensions, such as the deep audio redundancy (DRED), within the Opus codec (RFC6716).

About This Document

This note is to be removed before publishing as an RFC.

Status information for this document may be found at https://datatracker.ietf.org/doc/draft-lechler-mlcodec-test-battery/.

Discussion of this document takes place on the Machine Learning for Audio Coding Working Group mailing list (mailto:mlcodec@ietf.org), which is archived at https://mailarchive.ietf.org/arch/browse/mlcodec/. Subscribe at https://www.ietf.org/mailman/listinfo/mlcodec/.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 21 January 2026.

Table of Contents

1. Introduction

The IETF machine learning for audio coding (mlcodec) working group aims to leverage current and future opportunities presented by ML codecs to enhance the Opus codec [RFC6716] and its extensions, including to improve speech coding quality and robustness to packet loss. Effective evaluation of codec extensions (such as DRED), in both standalone and redundancy settings, is a crucial factor in achieving those objectives. It supports reproducibility for existing extensions (for instance, by enabling validation of whether a retraining pipeline matches baseline model performance) and enables benchmarking of future improvements against previously established baselines.

However, as outlined in subsequent sections, effective evaluation of generative ML models presents numerous challenges and necessitates specialized subjective and objective evaluation methods. This document proposes a crowdsourced subjective test battery, along with associated test datasets, to address the unique requirements for accurate and reproducible evaluations of ML codecs. The proposed test battery covers both speech quality and intelligibility, including tests in clean, noisy, and reverberant conditions, and incorporates real-world audio data. The methodology leverages crowdsourced listeners [CROWDSOURCED-DRT] to enable rapid and scalable assessments, while controlling the variability associated with non-lab-based measurements.

In the era of generative ML models, reference-based objective metrics face additional limitations, while non-intrusive methods struggle with generalization, e.g., [URGENT2025] and [CROWDSOURCED-MUSHRA]. Consequently, the use of human listeners, the gold standard in both quality and intelligibility assessment, is of notable importance. The generative nature of ML codecs also implies that speech intelligibility could be significantly improved and/or degraded by such algorithms. For example, human perception for some phoneme categories could be enhanced, while confusions might be introduced for others, including hallucinations of incorrect phonemes even at high overall perceived quality. Such confusions may not be easily detected in quality tests, highlighting a pressing need for highly diagnostic phoneme-category, or even phoneme-level, intelligibility assessment methods.

The subsequent sections present the methodology, key considerations, and further motivation underlying the proposed test battery, addressing the challenges and requirements discussed above.

1.1. Listening Test Methods

1.1.1. MUSHRA 1S

MUSHRA 1S is variant of the well-established MUSHRA (multiple stimuli with hidden reference and anchor) methodology for assessing quality [ITU-R.BS1534-3] in clean non-reverberant conditions is proposed for testing and benchmarking of ML codecs. MUSHRA is firstly adapted to a crowdsourced, non-expert listener base, as described in [CROWDSOURCED-MUSHRA]. Particularly for generative models, which may cause hallucinations, a reference-based listening test is preferable [URGENT2025]. Secondly, one system under test is assessed at a time, in the context of a fixed reference and anchor. The advantages of testing one system at a time are the unlimited extendability of test conditions within the quality range of anchor and reference, avoiding context effects of other conditions within the same test, avoiding difficulties associated when merging results across multiple tests, and simplifying the task for the participants thereby avoiding listener fatigue, particularly in non-expert listeners. As such, MUSHRA 1S has similarities with to the absolute category ratings (ACR) tests, which can be used to calculate a mean opinion scores (MOS), in that it is simple and easily extendable, while also being more stable than ACR, due to the fixed range of expected audio quality, bound by the anchor and reference. Reference-less MOS scores have been demonstrated to suffer from range-equalizing biases [COOPER2023], with other samples presented within the same test defining the range of expectation of what constitutes "good" or "bad" speech quality. The drawback of the MUSHRA 1S solution, compared to a traditional MUSHRA test, is the slightly decreased sensitivity to very small differences between similar methods, which may only be detectable in direct comparisons.

1.1.2. DCR

The degradation category rating (DCR) approach is used to produce a degradation mean opinion score (DMOS) [ITU-T.P800]. Although it is typically used with a high-quality reference, the test is also capable of assessing degradation caused by codecs when tested on mild-to-moderately impaired real-world data [MULLER2024]. The approach is more sensitive than absolute category ratings (ACR) [ITU-T.P800]. An implementation of the test procedure for crowdsourced tests is available in [ITU-T.P808].

1.1.3. DRT

The diagnostic rhyme test (DRT) [ITU-T.P807] measures speech intelligibility by presenting minimal pairs where the contrasted phonemes differ in terms of a specific, controlled phonetic category. The linguistic and acoustic insight of the DRT, with test items belonging to classes of distinctive linguistic features which are acoustically interpretable, poses a useful tool for both codec analysis and benchmarking. The test is free from context-effects and memory effects and has a high test sensitivity. It is therefore well-suited for a crowdsourced listener audience. Bearing in mind the principles for crowdsourcing listening tests employed in [ITU-T.P808], the test was adapted for crowdsourced listening tests in [CROWDSOURCED-DRT] and test vectors in five languages were published [DRT-REPO]. The test data was recently adopted by [LESCHANOWSKY2025].

1.1.4. Crowdsourcing Adaptations

Crowdsourced listening tests benefit from rigorous screening and quality control. In addition to the specific implementation of standardized test approaches for crowdsourced listening tests, [ITU-T.P808] has provided useful guiding principles for the adaptation of laboratory-based tests to counteract challenges posed by the comparatively uncontrolled crowdsourcing environment. For instance, steps of qualification and training are added before the actual test stimuli are presented and catch trials are included in the pool of test questions. It is further recommended to assess the quality of participants' responses across different platforms, such as Amazon Mechanical Turk, Prolific, and others [CROWDSOURCED-MUSHRA]. Each platform has a unique set of filters that can be used to recruit a specific participant pool. The platform and any filters used should always be reported along with test results, as absolute results may depend on those settings and may differ considerably between platforms.

2. Proposed Crowdsourced Listening Test Battery

In the literature, evaluations of speech codec quality often focus solely on clean conditions. However, given the wide range of potential applications for modern speech codecs, and the unique ways in which ML codecs may be affected by various types of real-world distortions, it is important to assess their limitations under representative real-world scenarios, including challenging listening conditions.

In addition to clean speech data, the proposed test battery considers performance evaluation on overlapping speech, reverberant and noisy speech, speaker consistency and phoneme-level intelligibility. The current version comprises predominantly English test vectors, but the extension to include multiple languages is desirable. Some of the modules of the test battery outlined below for assessment of standalone ML codec performance can also be used (where applicable), for assessing the performance of redundancy schemes under packet loss conditions (e.g., Opus+DRED).

All proposed test vectors will be made publicly available at the sampling rate of 48 kHz.

2.1. Speech Quality Evaluation

2.1.1. Clean Speech Test Vectors

By employing the MUSHRA 1S approach and utilizing high-quality clean speech data, the system under test is evaluated with respect to the overall quality. The reference allows the listener to assess also the correctness of the linguistic content as well as the preservation of the speaker characteristics. In this test, the quality of each codec or extension is assessed in standalone mode. The diverse test set comprises 100 gender-balanced clean speech files, covering 100 unique speakers, and includes samples from both adult and children’s speech. Furthermore, the set of test vectors covers a diverse range of accents of English.

2.1.2. Real-World Degradation Test Vectors

As speech codecs may be used by a wide variety of applications, it cannot be ensured that the audio to be compressed constitutes clean speech in the sense of dry and noise-free high-quality audio. It is therefore important to assess the codec’s resilience to real-world degradation. For tests where test vectors have impaired quality, DCR offers an effective way to measure the severity of any additional degradation introduced by the codec. The test data consists of 90 crowdsourced speech files in mildly impaired real-world scenarios of noise and reverberation. Of these, 45 files are predominantly focussed on reverberant speech and 45 on speech in noise. The reverberation and noise levels are mild to moderate.

2.1.3. Simultaneous Talker Test Vectors

Most application purposes rely on the codec's capability of preserving simultaneously occurring speech from multiple talkers. However, in practice, this can be a challenging task. A listening test using the DCR methodology offers insights into whether the presence of overlapping speech leads to degradation, which may occur in the form of artifacts or speech suppression. The proposed test set consists of 20 files of conversations between two to three talkers.

2.1.4. Packet Loss Scenarios

Real-world packet loss traces and/or simulated loss patterns (including using the packet loss simulator provided by the working group in Opus) can be utilized to evaluate the overall quality of redundancy codecs, such as Opus and DRED working together.

Details TBD.

2.2. Speech Intelligibility Evaluation

2.2.1. Clean Speech Test Vectors

The DRT for evaluating speech intelligibility, adapted for crowdsourced participants [CROWDSOURCED-DRT], is proposed to be performed on a subset of the stimuli provided in [DRT-REPO]. The subset consists of two test vectors, one male and one female talker sample, for each word pair in the standard DRT word list for English [ITU-T.P807]. Test vectors for four other languages are also available in the same collection. Due to listeners' perceptual sensitivity to the subtle and highly localized cues that distinguish the two target phonemes, this test is primarily applicable in the evaluation of standalone codecs, with limited expected utility when combined with packet losses and redundancy schemes.

2.2.2. Noisy Test Vectors

In order to evaluate a codec's resilience to noise in terms of speech intelligibility, the proposed evaluation battery for ML codecs contains noisy counterparts for the clean speech test vectors described in the previous paragraph. Speech-shaped noise (SSN) is used as a stationary additive masker in which intelligibility can be evaluated. While the presence of noise may lead to particularly severe codec distortion in some models, even the presence of well-preserved noise can help to distinguish the intelligibility of high-quality models that demonstrate a ceiling effect in clean conditions. The use of stationary noise is essential for the DRT to ensure

uniform effects on the short-term localized perceptual cues. For the same reason, the noisy version of the test is also geared towards the evaluation of standalone codecs. The SSN was generated based on long-term-averaged short-term spectra of a publicly available clean speech data set [DEMIRSAHIN2020]. The average spectrum was used as a filter that was convolved with white noise, resulting in SSN.

2.3. Example Results

The results shown in Table 1 below were obtained by using test methodology described above. Subjective tests were run on the Prolific crowdsourcing platform. The participants were required to be native speakers of English, with an approval rate of at least 98% and at least 110 previous submissions. Only participants without any self-reported hearing impairments and without a cochlear implant were invited to participate. Additionally, diagnostic rhyme test studies were only open to participants who self-reported not to have have dyslexia.

Codec	Quality in Clean Speech (MUSHRA) [95% CI]	Intelligibility in Clean Speech (DRT Score) [95% CI]
Clean input	98.3 [+/- 0.2]	94.9 [+/- 1.3]
Opus v1.5.2 9kbps NOLACE	85.4 [+/- 1.7]	90.0 [+/- 2.0]
Opus v1.5.2 9kbps LACE	70.2 [+/- 2.0]	90.6 [+/- 1.8]
Opus v1.5.2 9kbps	56.2 [+/- 2.3]	89.0 [+/- 2.0]
Opus v1.5.2 6kbps	24.0 [+/- 0.7]	86.3 [+/- 2.4]
DRED SA 2kbps	64.9 [+/- 2.3]	88.4 [+/- 2.4]
DRED SA 1kbps	52.0 [+/- 2.4]	84.5 [+/- 2.8]
DRED SA 0.5kbps	20.7 [+/- 2.2]	71.7 [+/- 3.8]

Table 1

3. Objective Evaluation

Objective metrics are often used during the development of speech codecs, with expert evaluations conducted towards the end of the development lifecycle. While effective for traditional DSP-based codecs, traditional well-established reference-based metrics, such as PESQ [ITU-T.P862], often fail to accurately evaluate generative methods. For instance, PESQ has been empirically shown to have an underestimation bias for generative models which may have high output quality but for which the output may also considerably differ from the reference [CROWDSOURCED-MUSHRA].

At present, the research into alternative metrics is flourishing with various innovative methods being proposed,
such as non-intrusive DNN-based metrics (e.g, [UTMOS]), metrics with non-matched references (e.g., [SCOREQ]), or composite score types of

metrics (e.g., [UNI-VERSA]). While recent correlation investigations, e.g., [URGENT2025], are promising, it is too early to include such metrics in this proposal, as it is yet to be seen which metrics can demonstrate both good accuracy and generalization to a variety of generative models and test vectors. Further insights in this area are of potential value for rapid, accessible and inexpensive evaluation of ML codecs. Hence, we propose to investigate which objective metrics are effective predictors of listener responses for the test battery components, and under which conditions.

4. Conventions and Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

5. Security Considerations

TBD

6. IANA Considerations

This document has no IANA actions.

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC6716] Valin, JM., Vos, K., and T. Terriberry, "Definition of the Opus Audio Codec", RFC 6716, DOI 10.17487/RFC6716, September 2012, <https://www.rfc-editor.org/rfc/rfc6716>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

7.2. Informative References

[COOPER2023]
Cooper, E. and J. Yamagishi, "Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech", INTERSPEECH 2023, pages 1104--1108, n.d., <https://www.isca-archive.org/interspeech_2023/cooper23_interspeech.pdf>.

[CROWDSOURCED-DRT]
Lechler, L. and K. Wojcicki, "Crowdsourced Multilingual Speech Intelligibility Testing", ICASSP 2024, DOI 10.1109/ICASSP48485.2024.10447869, n.d., <https://ieeexplore.ieee.org/document/10447869>.

[CROWDSOURCED-MUSHRA]
Lechler, L., Moradi, C., and I. Balic, "Crowdsourcing MUSHRA Tests in the Age of Generative Speech Technologies: A Comparative Analysis of Subjective and Objective Testing Methods", INTERSPEECH 2025, n.d., <https://arxiv.org/abs/2506.00950>.

[DEMIRSAHIN2020]
Demirsahin, I., Kjartansson, O., Gutkin, A., and C. Rivera, "Crowdsourced high-quality UK and Ireland English Dialect speech data set.", LREC 2020, pages 6532--6541, ISBN 979-10-95546-34-4, n.d., <https://www.aclweb.org/anthology/2020.lrec-1.804>.

[DRT-REPO] Cisco Systems, "Multilingual Speech Testing - Speech Intelligibility DRT", n.d., <https://github.com/cisco/multilingual-speech-testing/tree/main/speech-intelligibility-DRT>.

[ITU-R.BS1534-3]
ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems", ITU-R Recommendation BS.1534-3, October 2015.

[ITU-T.P800]
ITU-T, "Methods for subjective determination of transmission quality", ITU-T Recommendation P.800, August 1996.

[ITU-T.P807]
ITU-T, "Subjective test methodology for assessing speech intelligibility", ITU-T Recommendation P.807, February 2016.

[ITU-T.P808]
ITU-T, "Subjective evaluation of speech quality with a crowdsourcing approach", ITU-T Recommendation P.808, June 2021.

[ITU-T.P862]
ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", February 2001, <https://www.itu.int/rec/T-REC-P.862>.

[LESCHANOWSKY2025]
Leschanowsky, A., Lakshminarayana, K.K., Rajasekhar, A., Behringer, L., Kilinc, I., Fuchs, G., and E.A.P. Habets, "Benchmarking Neural Speech Codec Intelligibility with SITool", INTERSPEECH 2025, DOI 10.48550/arXiv.2506.01731, n.d., <https://arxiv.org/abs/2506.01731v1>.

[MULLER2024]
Muller, T., Ragot, S., Gros, L., Philippe, P., and P. Scalart, "Speech quality evaluation of neural audio codecs", INTERSPEECH 2024, pages 1760--1764, n.d., <https://www.isca-archive.org/interspeech_2024/muller24c_interspeech.pdf>.

[SCOREQ]
Ragano, A., Skoglund, J., and A. Hines, "SCOREQ: Speech Quality Assessment with Contrastive Regression", NeurIPS 2024, pages 105702--105729, n.d., <https://proceedings.neurips.cc/paper_files/paper/2024/file/bece7e02455a628b770e49fcfa791147-Paper-Conference.pdf>.

[UNI-VERSA]
Shi, J., Shim, H.J., and S. Watanabe, "Uni-VERSA: Versatile Speech Assessment with a Unified Network", DOI 10.48550/arXiv.2505.20741, target https://arxiv.org/abs/2505.20741, 2025, <https://doi.org/10.48550/arXiv.2505.20741>.

[URGENT2025]
Saijo, K., Zhang, W., Cornell, S., Scheibler, R., Li, C., Ni, Z., Kumar, A., Sach, M., Fu, Y., Wang, W., Fingscheidt, T., and S. Watanabe, "Interspeech 2025 URGENT Speech Enhancement Challenge", INTERSPEECH 2025, target https://arxiv.org/abs/2505.23212, n.d..

[UTMOS] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022", INTERSPEECH 2022, pages 4521--4525, n.d., <https://www.isca-archive.org/interspeech_2022/saeki22c_interspeech.pdf>.

Authors' Addresses

Laura Lechler
Cisco Systems
United Kingdom
Email: llechler@cisco.com

Kamil Wojcicki
Cisco Systems
Australia
Email: kamilwoj@cisco.com