EVALUATION OF OBJECTIVE QUALITY MODELS ON NEURAL AUDIO CODECS

Thomas Muller, Stéphane Ragot, Vincent Barriac, Pascal Scalart

EVALUATION OF OBJECTIVE QUALITY MODELS ON NEURAL AUDIO CODECS

Thomas Muller[1,2], *Stéphane Ragot*[1], *Vincent Barriac*[1], *Pascal Scalart*[2]

[1]Orange Innovation, France
[2]IRISA, University of Rennes, France

ABSTRACT

With the emergence of neural audio codecs and new objective quality models based on machine learning, there is a need to clarify which models predict accurately the perceptual quality of coded speech. In this paper, we consider a selected subset of ten objective quality models; we present a correlation analysis based on test results from a P.800 ACR experiment on clean speech, assessing the quality of neural speech/audio codecs – traditional codecs (EVS, Opus) are also included as yardsticks. The evaluation is limited to signal-based models for listening-only quality. Overall results per condition are analyzed in terms of Pearson's correlation, Kendall's Tau and root mean squared error (RMSE); objective scores per codec are also discussed.

Index Terms— speech quality, objective models, neural audio codec, subjective test

1. INTRODUCTION

Audio quality evaluation is necessary when developing speech or audio codecs, or characterizing their performance. Subjective listening tests based on methodologies such as P.800 ACR or DCR [1], or MUSHRA [2] remain the most important and direct way to evaluate codec performance, however these tests require time and resources. Besides codec quality assessment, objective models predicting perceptual speech quality are also essential in many other tasks, such as network monitoring, drive tests, conformance testing or phone quality certification. Different objective models (e.g., PESQ [3], ViSQOL [4], etc.) are used in publications to report and analyze experimental codec evaluations, and one may wonder what objective models are the most reliable or what is their relative performance, especially in the context of recent developments in machine learning applied to audio. With new techniques based on deep learning, the field of audio coding has seen the emergence of new neural audio codecs such as LPCNet [5], SoundStream [6], EnCodec [7], AudioDec [8], or Descript Audio Codec [9]. In parallel, new objective models have often been developed based on machine learning, as seen for instance in challenges (e.g., VoiceMOS [10] and ConferencingSpeech [11]), standardization activities (e.g., P.565.1 [12]), and other works [13, 14].

The aim of the present work is to benchmark objective quality models on speech, especially when evaluating neural speech/audio codecs. We report results from a P.800 ACR test on clean speech comprising anchors, traditional and neural audio codecs, and this experiment is used as "ground truth". A correlation analysis is performed to evaluate a selected subset of objective models. In addition to usual evaluation metrics (Pearson's correlation, Kendall's Tau, RMSE), we also analyze objective scores per codec.

This paper is organized as follows. Section 2 provides a brief review of objective quality models, with a focus on speech quality, and lists models selected for the present evaluation. Section 3 describes the test plan and results from a P.800 ACR subjective experiment that serves as the "ground truth". Section 4 presents the correlation analysis on tested objective models, before concluding in Section 5.

2. REVIEW OF OBJECTIVE QUALITY MODELS

A general review of quality prediction models can be found in [15] and [16] for speech and [17] for audio. We focus here on models predicting overall speech or audio quality in terms of mean opinion scores (MOS) – or an equivalent scale –, especially for codec evaluation purposes. We do not review other quality aspects such as speech naturalness [18] or intelligibility [19], or quality assessment of speech enhancement [20] or synthesized speech [10]. Objective models for speech quality can be classified into different categories: signal-based or parametric; intrusive (full-reference) or non-intrusive (no-reference); conversational quality or quality in one phase (listening, speaking or interaction). In this work, we focus on signal-based models for listening-only quality and consider both intrusive and non-intrusive models.

Initial objective tools relied on simple criteria in time, frequency, or parametric domain, such as signal to noise ratio (SNR), segmental SNR, mean squared error (MSE), spectral distortion, etc. It is well known that such objective metrics do not correlate well with subjective tests [21]. Following [22], full-reference models have been developed to predict the perceptual quality of coded speech or audio – see for instance [23, 24, 25]. The PEAQ model [26] was standardized using this approach using both frequency and auditory transforms; PSQM is based on similar principles [27] in the speech do-

Table 1. Tested objective models – f_s used in tests in bold.

Metric	Content	f_s (kHz)	Intrusive	Version
PESQ	Speech	8,**16**	Yes	pesq v0.0.4 [30]
POLQA	Speech	8, **48**	Yes	v3.0 (in MultiDSLA)
ViSQOL-S	Speech	**16**	Yes	v3.3.3 [31]
WARP-Q	Speech	8,**16**	Yes	v1.0.0 [32]
DNSMOS	Speech	**16**	No	commit 591184a [33]
NISQA	Speech	**48**	No	commit ac83137 [34]
NORESQA	Speech	**16**	No*	commit 8d56b95 [35]
UTMOS	Speech	**16**	No	commit 2d6d612 [36]
PEAQ	Audio	**48**	Yes	Basic, AFsp v9r0 [37]
ViSQOL-A	Audio	**48**	Yes	v3.3.3 [31]

$*$ Non-matching references [13] but used here with original input signal as reference

main, using internal representation matching and time alignment by cross-correlation. PESQ was later developed to allow better accuracy in end-to-end measurements (including codec degradations as well as filtering, jitter, etc.), using auditory transforms, improved time-delay estimation, equalization and mapping to subjective tests. POLQA (P.863) [28] covered bandwidths up to fullband, impairments from real network conditions (in particular in Voice over IP) and both electrical and acoustic interfaces. In ITU-T there is ongoing work where NISQA [29] is considered and improved in P.SAMD.

In recent years, many new objective models have been proposed. ViSQOL [4] has been developed as an alternative to POLQA, with two modes: speech and audio (denoted ViSQOL-S and ViSQOL-A, respectively). WARP-Q [38] targets quality prediction for generative neural speech codecs. Different objective models were submitted to challenges (e.g., VoiceMOS [10] and ConferencingSpeech [11]) and we retain here only UTMOS [39] given that it scored best in the VoiceMOS challenge. DNSMOS [40] is known for evaluation in noise suppression tasks and is used here to predict raw_SIG as in [8]. NORESQA-MOS [13] combines neural networks and non-matching references to estimate MOS. There are many other objective models in the literature, e.g. SESQA [41] or InSE-Net [14], however they are not considered here because no public implementation is available.

For practical reasons, we selected a subset of ten models that are often used in publications or reflecting recent proposals – see list in Table 1. Note that input audio was downsampled to f_s when $f_s < 48$ kHz. We did not include STOI [19] in the benchmark, because this metric is meant for intelligibility assessment. We chose models that have an available implementation to maximize reproducibility, with one exception: we included POLQA (P.863), given that this model represents the state of the art of ITU-T objective models and is still widely used in the field. Note that PESQ (P.862) is included here even if this model is considered obsolete and withdrawn in ITU-T; in practice, PESQ is still popular in publications. PEAQ [26] is also included in the present evaluation as an "anchor" even if it is meant for high-quality audio. For PEAQ which output Objective Difference Grade (ODG) scores between -4 and 0, a +5 offset is applied to get a scale from 1 to 5.

3. TEST PLAN AND RESULTS FROM A P.800 EXPERIMENT (CLEAN SPEECH)

A good objective metric should predict MOS-LQO scores (objective listening quality) that are as close as possible to MOS-LQS scores (subjective listening quality) from a formal listening test. For this study, we conducted a P.800 ACR [1] test to get "ground truth" MOS-LQS scores for model evaluation. The test is designed to compare traditional and neural audio codecs operating on different bandwidths – wideband (WB), superwideband (SWB) and fullband (FB) – and bitrates (from 1.5 to 24.4 kbps). Four conditions are used as anchors to calibrate the test: the uncoded original audio (or "Direct" condition), and three P.50 Modulated Noise Reference Unit (MNRU) [45] with $Q = 36$, 23 and 10 dB. Two traditional speech/audio codecs are tested: Opus [46] which is standardized by IETF and used in WebRTC and various Internet applications; and EVS [47] which is standardized by 3GPP and used in mobile telephony. For the neural audio codecs, we chose models which have a publicly available implementation, namely Lyra V2 [48], EnCodec [7], LPCNet [5], AudioDec [8], Descript Audio Codec (DAC) [9] and an extension of EnCodec using multiband diffusion that we refer as AudioCraft [49]. All the tested conditions with the details of the version and model used are summed up in Table 2 (including frame length L and sampling rate f_s). For the test, 30 naive listeners were recruited, and split in 5 panels of 6 listeners. Each panel listened and rated all conditions on a specific and different subset of speech samples and in a different order. The speech samples are extracted from an internal French speech database of phonetically balanced sentence pairs, with 3 male and 3 female speakers. The audio samples were filtered by a 20-20,000 Hz bandpass FIR filter [50] and normalized to -26 dB LKFS [51]. The audio fed to the codecs

Table 2. List of tested codec operation points.

Codec	f_s (kHz)	L (ms)	bitrate (kbps)	Version
LPCNet	16	10	1.6	0.1
Lyra V2	16	20	3.2, 6, 9.2	v1.3.2
EnCodec	24	13.3	1.5, 3, 6, 12, 24	[42] Nov. 2023
AudioCraft	24	13.3	1.5, 3, 6	[43] v1.0.0
AudioDec	24	12.5	6.4	libritts_sym
DAC	44.1	11.6	1.7, 2.6, 5.2, 7.8	v1.0.0
AudioDec	48	6.25	12.8	vctk_sym
Opus	48	20	12, 16, 24	v1.4 (–cbr)
EVS-WB	16	20	7.2, 8	[44] v16.3.0
EVS-SWB	32	20	9.6, 13.2, 24.4	[44] v16.3.0

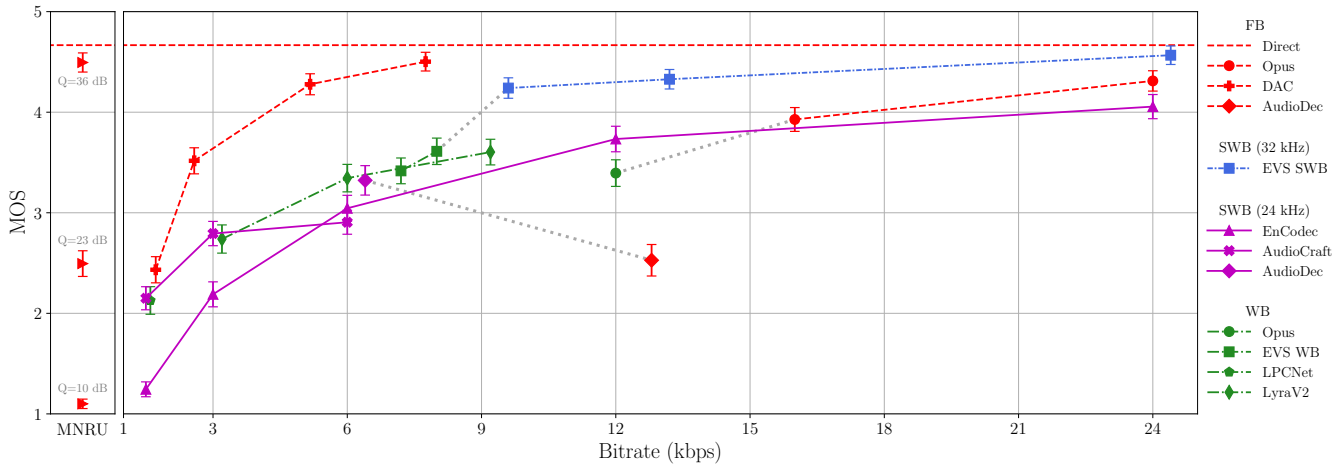

Fig. 1. Codec bitrate vs. MOS-LQS – with 95% confidence intervals; horizontal dotted line corresponds to "Direct".

were resampled to match their input sampling rate, using the ResampAudio routine from the AFsp toolset [37].

The results of the ACR test are summarized in Fig. 1. A detailed discussion of these results is beyond the scope of this paper, given that the subjective test is only used here as "ground truth" for benchmarking objective quality models. Note that the dataset (processed samples and raw data) can be shared upon request under appropriate legal frameworks.

4. EVALUATION OF OBJECTIVE MODELS

4.1. Results from overall correlation analysis

Objective models listed in Table 1 are compared based on typical evaluation metrics: Pearson correlation coefficient, root mean squared error (RMSE) [52] and Kendall's Tau rank correlation coefficient [17]. Note that these metrics are computed per condition, and there are 30 conditions in the ACR test reported in Section 3. Pearson's correlation is related to the linearity of the relation between subjective and objective scores. RMSE penalizes deviations from the $y = x$ line, where x and y are the subjective and objective scores, while Kendall's Tau measures the similarity between condition rankings. Note that the Spearman rank correlation coefficient was also evaluated, however it is redundant with Kendall's Tau, and the latter is better suited here given the low number of data points (conditions). Moreover, there exists a variant of the RMSE denoted RMSE* that takes the confidence interval on ACR MOS into account [52]; as the confidence intervals in ACR scores are quite similar in our test, RMSE* differs from RMSE by only a slight offset, therefore only RMSE is presented here.

According to ITU-T P.1401 [52], a 3rd order monotonic polynomial mapping $P_{\mathcal{M}}(x) = a_3x^3 + a_2x^2 + a_1x + a_0$, where $\mathcal{M}$ is a given model, should be applied to compare objective models. This mapping corrects a potential offset or gradient, and linearizes a possible "banana shape" of the scatter plot $\{(x_i, y_i)\}$, where i is the condition number ($i = 1, \cdots, 30$ here). In practice, some models already include an internal mapping (e.g., so-called scaling in PESQ). The polynomial mapping $P_{\mathcal{M}}(x)$ was determined here for each model, using the implementation from the ConferencingSpeech Challenge [53]. Fig. 2a shows an example of scatter plot $\{(x_i, y_i)\}$, taking the example of UTMOS; this figure also shows the 3rd order polynomial line $y = P_{\text{UTMOS}}(x)$ used for mapping; Fig. 2b shows the corresponding scatter plot $\{(P_{\text{UTMOS}}(x_i), y_i)\}$ after mapping of UTMOS scores. A 1st order mapping was also determined to verify that the 3rd order mapping did not lead to overly optimistic model performance evaluation.

All evaluation metrics (Pearson, Kendall's Tau, RMSE) are reported for each model before and after mapping in Fig. 3. One can note that Kendall's Tau is not affected by mapping, due to the monotony constraint. Mappings reduced outliers caused by a highly non-linear relationship on the MOS scale, which happens for example with PEAQ and NORESQA models. Overall, results indicate that POLQA, UTMOS, PESQ, and WARP-Q, have here the best performance; for these models, 1st order mapping is sufficient.

4.2. Analysis of objective scores per codec

We summarize below observations when comparing MOS-LQS scores (see Fig. 1) with MOS-LQO scores *per codec conditions* – this summary is limited to the best performing models identified in Section 4.1 due to space constraints:

- POLQA was only trained on traditional codecs including EVS and Opus. We observe that POLQA underpredicts DAC, AudioDec, EnCodec (higher bitrates), Lyra V2, while POLQA overpredicts EnCodec (lower bitrates), AudioCraft, LPCNet, compared to ACR tests results in Fig. 1.
- UTMOS scores without mapping were in the $[1.3, 3.7]$ range (see Fig. 2a); the linear mapping really helps to adjust to the full ACR scale. UTMOS with no mapping underpredicts most codec conditions, but overpredicts AudioCraft, possibly because there is residual coding noise

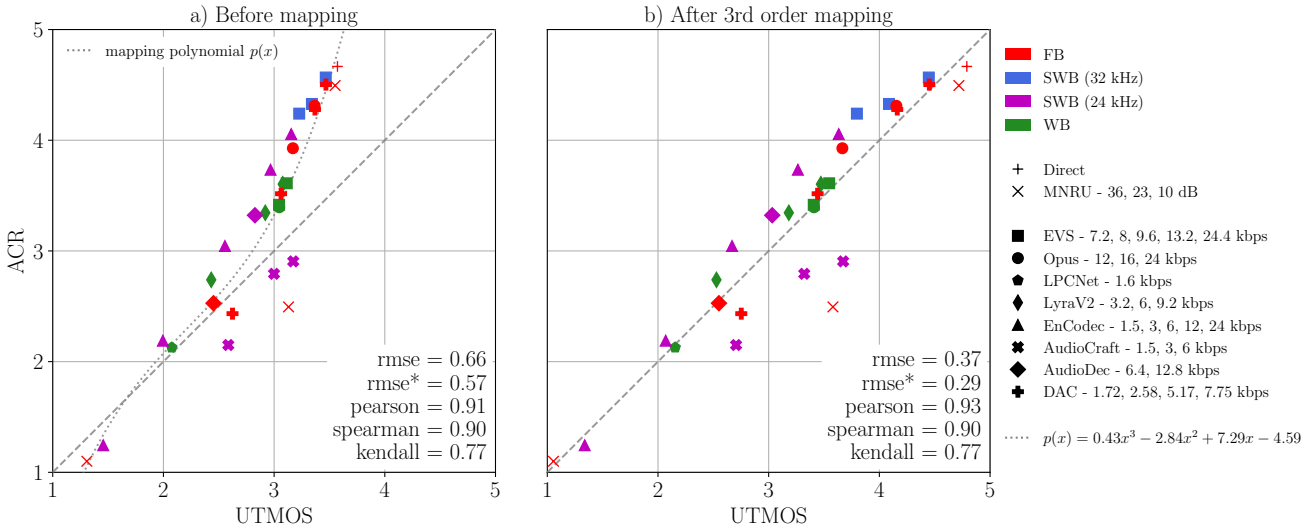

Fig. 2. ACR vs UTMOS scores, before and after monotonic polynomial mapping.

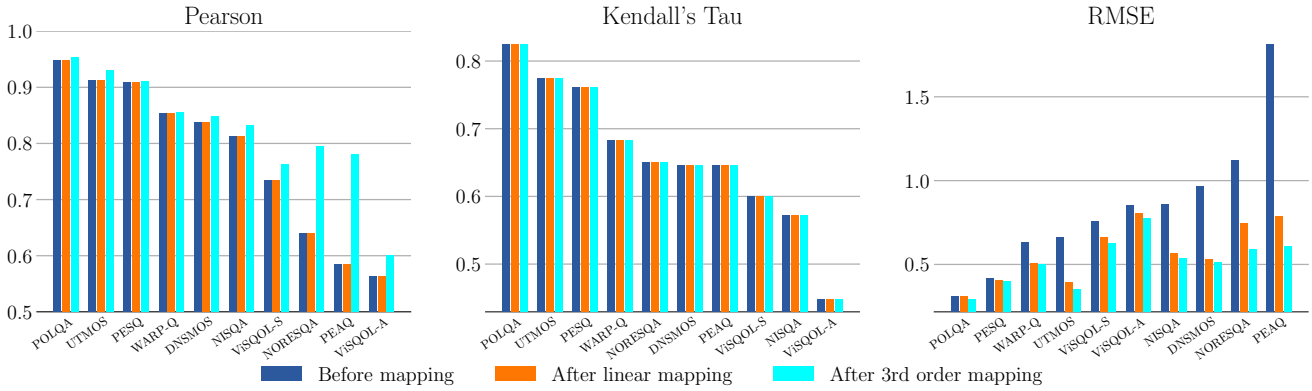

Fig. 3. Pearson, Kendall's Tau and RMSE metrics, before and after linear and 3rd order polynomial mapping.

(from the diffusion model [49]) only in inactive regions.

- PESQ underpredicts EVS, DAC (higher rates), AudioDec/24 kHz at 6.4 kbps, EnCodec (higher rates), AudioCraft, and LPCNet, and overpredicts Opus (scored above EVS) and EnCodec (lower rates).
- WARP-Q scores were in the $[2.3, 4.2]$ range, which makes it hard to discriminate codecs with no mapping – for instance, Opus, EVS, and EnCodec around 24 kbps are ranked as near-equivalent (numerically in this descending order).

The best performing models (POLQA, UTMOS, PESQ, WARP-Q) predicted accurately the monotonic bitrate/quality behavior of tested multirate codecs – which was not always the case for other tested models. It is remarkable that models operating at 16 kHz (PESQ without mapping, UTMOS and WARP-Q with mapping) had relatively good performance, even for fullband codecs. Except for few models (e.g., PESQ or POLQA), mapping helps improving accuracy (RMSE).

5. CONCLUSION

A P.800 ACR experiment on clean speech was conducted; this test included several neural audio codecs, as well as two traditional codecs. Ten objective models were compared against subjective scores. Results showed that the tested objective models do not perform equivalently in terms of overall evaluation metrics before and after mapping; a subset of tested models could predict quality increase as a function of codec bitrate, and no model could predict quality ranks for all pair comparisons between codecs, with issues of codec under- or over-prediction. This study has some limitations as it relies on a single P.800 ACR experiment in clean speech and clean channel conditions. More listening test databases (with different languages) and more test conditions (e.g., packet losses, jitter, etc.) should be included. This study would motivate further work considering neural audio codecs in objective model design.

ACKNOWLEDGEMENTS

The authors thank Laetitia Gros and Caroll Ratazzi (Orange test lab in Lannion, France) for conducting the ACR test.

6. REFERENCES

[1] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Aug. 1996.

[2] ITU-R Rec. BS.1534–3, "Method for the subjective assessment of intermediate quality level of audio systems," Oct. 2015.

[3] A.W. Rix et al., "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752 vol.2.

[4] M. Chinen et al., "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," in *Proc. QoMEX*, 2020.

[5] J.-M. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis through Linear Prediction," in *Proc. ICASSP*, 2019.

[6] N. Zeghidour et al., "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Trans. TASLP*, 2021.

[7] A. Défossez et al., "High Fidelity Neural Audio Compression," in *arXiv:2210.13438*, 2022.

[8] Y.-C. Wu et al., "Audiodec: An open-source streaming high-fidelity neural audio codec," in *Proc. ICASSP*, 2023, pp. 1–5.

[9] R. Kumar et al., "High-Fidelity Audio Compression with Improved RVQGAN," in *Advances in NIPS*, 2023, vol. 36, pp. 27980–27993.

[10] W.C. Huang et al., "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[11] G. Yi et al., "ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications," in *Proc. Interspeech*, 2022.

[12] ITU-R Rec. P.565.1, "Machine learning model for the assessment of transmission network impact on speech quality for mobile packet-switched voice services," Nov. 2021.

[13] P. Manocha and A. Kumar, "Speech Quality Assessment through MOS using Non-Matching References," in *Proc. Interspeech*, 2022.

[14] G. Jiang et al., "InSE-NET: A Perceptually Coded Audio Quality Model based on CNN," in *AES Convention 151*, Oct 2021.

[15] P.C. Loizou, "Speech quality assessment," in *Multimedia analysis, processing and communications*, W. Lin and al., Eds., pp. 623–654. Springer, 2011.

[16] S. Möller, *Quality Engineering – Quality of Communication Technology Systems*, Springer, 2023.

[17] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM TALSP*, vol. 29, pp. 1530–1541, 2021.

[18] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech*, 2020, pp. 1748–1752.

[19] C.H. Taal and al., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010.

[20] S. Leglaive et al., "Objective and subjective evaluation of speech enhancement methods in the UDASE task of the 7th CHiME challenge," arXiv:2402.01413, 2024.

[21] T. Barnwell, "Correlation analysis of subjective and objective measures for speech quality," in *Proc. ICASSP*, 1980, vol. 5, pp. 706–709.

[22] M.R. Schroeder, B.S. Atal, and J. L. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in *Frontiers of Speech Communication Research*, B. Lindblom and S. Ohman, Eds., pp. 217–229. Academic, 1979.

[23] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," in *Proc. ICASSP*, 1985, vol. 10.

[24] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.

[25] B. Paillard et al., "Perceval: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc*, vol. 40, no. 1/2, pp. 21–31, 1992.

[26] T. Thiede et al., "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc*, vol. 48, no. 1/2, 2000.

[27] J.G. Beerends and J.A. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc*, vol. 42, no. 3, pp. 115–123, 1994.

[28] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," Mar. 2018.

[29] G. Mittag et al., "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131.

[30] "Pesq wrapper for python users, version 0.0.4," `https://github.com/ludlows/PESQ/releases/tag/v0.0.4`.

[31] "ViSQOl, version 3.3.3," `https://github.com/google/visqol/releases/tag/v3.3.3`.

[32] "WARP-Q," `https://github.com/wjassim/WARP-Q`.

[33] "DNSMOS," `https://github.com/microsoft/DNS-Challenge/tree/master/DNSMOS`.

[34] "NISQA," `https://github.com/gabrielmittag/NISQA`.

[35] "NORESQA," `https://github.com/facebookresearch/Noresqa`.

[36] "UTMOS," `https://github.com/sarulab-speech/UTMOS22`.

[37] "AFsp," `https://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/AFsp/AFsp.html`.

[38] W.A. Jassim et al., "Warp-Q: Quality prediction for generative neural speech codecs," in *Proc. ICASSP*, 2021, pp. 401–405.

[39] T. Saeki et al., "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[40] C.K. Reddy et al., "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2022.

[41] J. Serrà, J. Pons, and S. Pascual, "SESQA: Semi-Supervised Learning for Speech Quality Assessment," in *Proc. ICASSP*, 2021, pp. 381–385.

[42] "EnCodec," `https://github.com/facebookresearch/encodec`.

[43] "Audiocraft," `https://github.com/facebookresearch/audiocraft`.

[44] 3GPP TS 26.443, "Codec for Enhanced Voice Services (EVS); ANSI C code (floating-point)," v16.3.0.

[45] ITU-T Rec. P.810, "Modulated noise reference unit (MNRU)," Mar. 2023.

[46] IETF RFC 6716, "Definition of the Opus Audio Codec," Sept. 2012.

[47] M. Dietz et al., "Overview of the EVS codec architecture," in *Proc. ICASSP*, 2015, pp. 5698–5702.

[48] "Lyra V2, version 1.3.2," `https://github.com/google/lyra/releases/tag/v1.3.2`.

[49] R. San Roman et al., "From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion," arXiv:2308.02560, 2023.

[50] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," Mar. 2023.

[51] ITU-R Rec. BS.1770–4, "Algorithms to measure audio programme loudness and true-peak audio level," Oct. 2015.

[52] ITU-R Rec. P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," Jan. 2020.

[53] "Conferencevoice2022," `https://github.com/ConferencingSpeech/ConferencingSpeech2022/blob/main/eval/eval_result.py`.