Speech quality evaluation of neural audio codecs

Thomas Muller, Stéphane Ragot, Laetitia Gros, Pierrick Philippe, Pascal Scalart

Speech quality evaluation of neural audio codecs

Thomas Muller[1,2]*, Stéphane Ragot*[1]*, Laetitia Gros*[1]*, Pierrick Philippe*[1]*, Pascal Scalart*[2]

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

thomas.muller@orange.com, stephane.ragot@orange.com, laetitia.gros@orange.com,
pierrick.philippe@orange.com, pascal.scalart@irisa.fr

Abstract

This paper presents speech quality results to characterize the state of the art and technological advance of recent neural audio codecs targeting low bitrates. Audio quality was evaluated in one clean speech experiment (in French). Degradation Mean Opinion Score (DMOS) results are reported and discussed for neural audio codecs (LPCNet, Lyra V2, EnCodec, AudioCraft, AudioDec, Descript Audio Codec) – traditional codecs (Opus, EVS) are also included as performance yardsticks. We also discuss observed codec complexity to complement subjective test results.

Index Terms: speech and audio coding, subjective test, neural audio coding

1. Introduction

The development of generative audio models (e.g., WaveNet [1], SampleRNN [2], WaveRNN [3]) and autoencoders with discretized latent space (VQ-VAE in [4]) has led to a new generation of vocoders or audio codecs based on neural networks [5, 6, 7]. Significant progress has been made thanks to advanced architectures based on generative adversarial networks (GAN) [8, 9] and recent methods also consider the use of diffusion models [10]. Such neural audio codecs address various applications, including voice/video calling [11, 12], text-to-speech synthesis [13], or music generation [14, 15].

It is of great interest to assess the audio quality of this new generation of codecs and characterize the state of the art and technological advance. To the best of our knowledge, there is no comprehensive benchmarking of neural audio codec quality in the literature. Until now, limited quality assessments have been reported as part of model proposals [16, 17, 18], often relying on objective evaluations (e.g., ViSQOL [19] or STOI [20]), or the MUSHRA test methodology [21] which limits the number of test conditions and test samples and requires experienced listeners. In this work we consider the evaluation of neural audio codecs with a test methodology (P.800 DCR [22]) relying on naive subjects and allowing a more comprehensive comparison. To get performance "yardsticks", we include "traditional" speech/audio codecs: Opus [23, 24] and EVS [25].

Subjective quality is evaluated here in terms of listening audio quality. Different input types and channel conditions should be considered, including clean and noisy speech (with different languages and noise types and levels), music and mixed content, noisy channels. In this work, we only report test results for one clean speech experiment (in French); more experiments should be conducted for a complete characterization.

This article is organized as follows. In Section 2, we present the audio codecs selected for the subjective evaluation. In Section 3, the use of DCR is justified, and relevant test plan aspects are described. Experimental results are presented in Section 4, before concluding in Section 5.

2. Speech/audio codecs under test

2.1. Neural speech/audio codecs

For this evaluation, we selected neural codecs which have a publicly available implementation. For this reason we did not consider codecs such as SoundStream [16], MDCTNet [26], CBRC [27] or LMCodec [28]. Codecs considered below operate at constant bitrates (CBR), either at a single bitrate or in a multirate fashion. We used "default" versions that sometimes are not "streamable", i.e., they do not process one input frame at a time in a sequential manner. Note some codecs such as HiFi-Codec [29] could not be considered due to limitations in number of test conditions.

2.1.1. GAN-based autoencoders

Lyra V2 [12], EnCodec [17], AudioDec [30] and Descript Audio Codec (DAC) [18] are based on a VQ-VAE architecture [4], i.e. an autoencoder whose latent vectors are quantized. Audio sampled at a frequency f_s is converted to $f_l = \frac{f_s}{S}$ latent vectors per second, where S is the product of the convolution strides present in the encoder. Note that DAC makes use of non-causal convolution layers. Generally, the same strides are present in reverse order in the decoder to generate audio with the same sampling frequency f_s. The product of strides S also determines the frame length. The quantization technique used is the Residual Vector Quantization (RVQ) proposed in [16], inspired from multistage vector quantization [31]. It consists of a cascade of vector quantizers (VQs), allowing to choose the trade-off between quality of reconstruction and bitrate as a function of the number of quantizer stages. The bitrate of the codec is then $f_l \times N \times B$ where N is the chosen number of VQs in the RVQ and B is the number of bits per VQ. The other interest of RVQ is to enable multi-rate coding. For Lyra V2 and EnCodec we selected all available bitrates. For DAC, we only included the model at 44.1 kHz and evaluated the four bitrates studied in [18], using the value of bitrates computed as $f_l \times N \times B$. For AudioDec, we selected the "default" autoencoder models at 24 and 48 kHz described in [30], and no vocoder variant.

2.1.2. Hybrid codec: LPCNet

LPCNet [7] is based on the source–filter model, where the source (excitation) is modeled by WaveRNN [3] and the filter is represented using linear prediction coding (LPC). LPCNet is known to be a lightweight vocoder model that can run in real-time on a CPU. Many variants of LPCNet have been developed, here we only used the original LPCNet at 1.6 kbps.

2.1.3. Neural coding with diffusion-based synthesis

An extension of EnCodec (limited to 1.5, 3, and 6 kbps) has been proposed using multiband diffusion [10]. The audio waveform is generated by the diffusion model using the latent vectors from EnCodec. We will later refer to this extension as "AudioCraft" [32].

2.2. Traditional speech/audio codecs

2.2.1. Opus

Opus [23, 24] is standardized by IETF and used in WebRTC and various Internet applications. It supports different applications (VoIP, audio, restricted low-delay), mono and stereo input signals (as well as multistream frames), a wide range of bitrates, sampling rates from 8 to 48 kHz, frame lengths from 2.5 to 120 ms, different complexity levels, etc. Audio bandwidth depends on bitrate and goes from narrowband (NB) to fullband (FB), including superwideband (SWB) for signals sampled at 24 kHz. Note that recent versions of Opus include blocks using neural networks (e.g. speech/music classification, and even enhancement and redundancy [33]), however the version of Opus available for processing (see Tab. 1) does not qualify as a neural audio codec for this evaluation. For Opus testing, we selected the VoIP application, since clean speech is tested. CBR operation was used from 12 to 24 kbps.

2.2.2. EVS

The EVS codec [25, 34] is standardized by 3GPP and deployed in mobile telephony. It operates with 20 ms frames and it supports several input/output sampling frequencies (8, 16, 32, 48 kHz) with mono input signals. The EVS codec supports 4 types of audio bandwidth: NB, wideband (WB), SWB, FB. For EVS testing, we focus on EVS "Primary modes" with CBR operation. We selected main operation points (EVS-SWB at 13.2 and 24.4 kbps), and we included lower bitrates used in comparisons [16]: EVS-WB at 7.2 and 8 kbps and EVS-SWB at 9.6 kbps. Note that SWB is only supported at bitrates starting at 9.6 kbps. Discontinuous transmission (DTX) was disabled.

2.3. Summary of tested codecs

The selected codec conditions for testing are listed in Tab. 1, where the input sampling frequency (f_s), frame length (in ms), tested bitrates, and codec versions are also specified.

Table 1: *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	[35] Nov. 2023
AudioCraft	24	13.3	1.5, 3, 6	[32] 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
EVS-WB	16	20	7.2, 8	[36] v16.3.0
EVS-SWB	32	20	9.6, 13.2, 24.4	[36] v16.3.0

3. Test methodology

3.1. Justification for DCR

The aim of the test is to compare several codecs operating with different coded bandwidths and a large range of bitrates (from 1.5 to 24.4 kbps). It is important to select a proper test methodology to assess (listening) speech quality. A comparison of test methodologies for speech quality can be found in [37]; general guidance on subjective quality evaluation of audio codecs can be found in [38].

Two methodologies in ITU-T P.800 [22] could be considered: Absolute Category Rating (ACR) [22] based on a five-category scale: Excellent=5, Good=4, Fair=3, Poor=2, Bad=1; Degradation Category Rating (DCR) resulting in a Degradation Mean Opinion Score (DMOS), based on the comparison with the original sample on a 5-category scale defined in Tab. 2. Historically, the ACR method has been extensively used in speech testing for NB and WB quality ranges for clean speech (with and without channel errors) and for music and mixed content; it allows potential comparisons with objective speech quality predictions (e.g., P.863/POLQA [39]). DCR was typically used for testing conditions with background noise where ACR is less applicable. In test exercises such as 3GPP EVS, DCR was selected to test audio quality in higher bandwidths (SWB, FB) and multi-bandwidth scenarios. In particular, DCR is more applicable than ACR to multi-bandwidth testing, because each trial is more independent since the stimuli are presented to listeners by pairs (A-B) where A is the reference and B is the sample processed by the system under evaluation.

Compared to other methodologies such as MUSHRA [21], DCR may not be as sensitive as MUSHRA and it may be subject to saturation for near-transparent quality (typically at very high bitrates). However, DCR allows to expose listeners to a wider range of source material and is more cost efficient while obtaining more votes. Moreover, DCR relies on naive listeners, which makes it more applicable to reflect the assessment of the general population. Note that DCR is not a transparency test, therefore other methodologies such as BS.1116 [40] should be considered for this purpose.

We therefore selected DCR for the test to rely on a well-established procedure, given that this method applies well to the expected quality/bitrate range in this evaluation.

Table 2: *Opinion scale for ITU-T P.800 DCR.*

Scale	Degradation
5	Degradation is inaudible
4	Degradation is audible but not annoying
3	Degradation is slightly annoying
2	Degradation is annoying
1	Degradation is very annoying

3.2. Details on test setup and test plan

For this evaluation, only clean speech was used to compare codecs. The DCR method relies on naive listeners, therefore the dataset needed to be in their native language – in French in our case. A total of 30 listeners were recruited for the subjective test, and split in 5 panels of 6 listeners. The audio samples presented during the test consisted of an internal database of 8s phonetically balanced sentence pairs (in French), with 3 male and 3 female talkers. This speech dataset is "pristine" and proprietary (owned by the test lab) because test samples shall not be used for codec tuning or other related training tasks, however

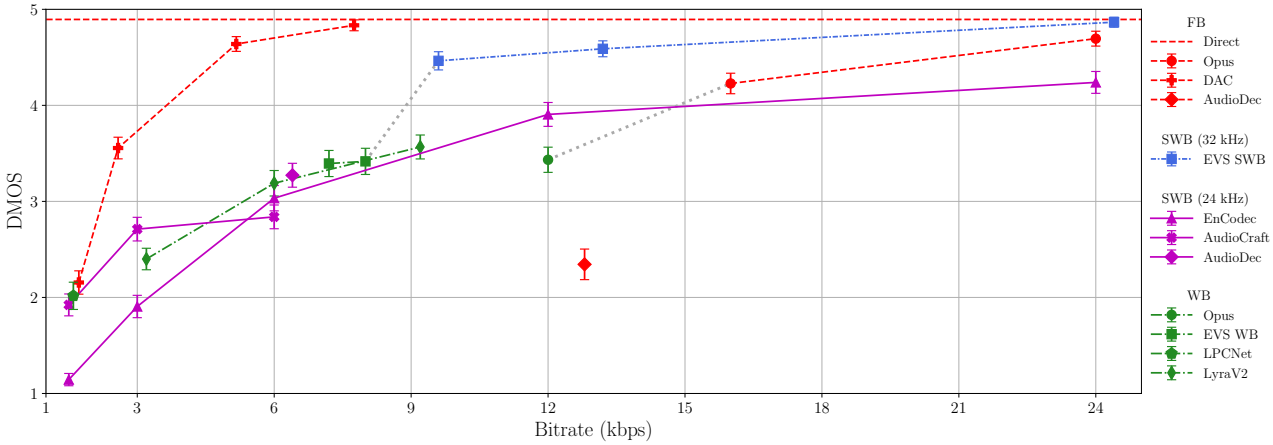

Figure 1: *DCR scores (DMOS) – with 95% confidence intervals – as a function of codec bitrate; the horizontal dotted line indicates the score of the "Direct" condition (uncoded reference).*

this dataset and processing samples can be shared upon request under appropriate legal frameworks. For each talker, 5 double sentences were used for testing and 1 for preliminaries (familiarization). The samples were pre-processed by a 20-20,000 Hz bandpass FIR filter from [41] and then normalized to -26 dB LKFS [42]. The effect of signal level (low, nominal, high) was not considered here. In addition to codec conditions listed in Tab. 1, the (uncoded) original, called "Direct" condition, was included together with calibration conditions (i.e., to make sure listeners will use the entire voting scale), corresponding to P.50 Modulated Noise Reference Unit (MNRU) [43] with $Q = 36$, 23, and 10 dB. Resampling to match the input sampling frequency of codecs was realized using the ResampAudio routine from the AFsp package [44].

Randomizations were constructed under randomized blocks experimental design described in [38]. The test was conducted in dedicated soundproof test rooms over headphones (Sennheiser HD 380 Pro); the (diotic) listening level was set to 73 dB SPL. The overall test duration was about 2 hours, including orientation, instructions, familiarization, test sessions, and rest breaks; there were 30 conditions in the test, with 180 votes per condition.

Prior to delivering processed samples to the test lab, ViSQOL [19] was applied to test conditions to check informally objective scores; expert inspection of coded waveforms/spectrograms was also done for sanity check. Problems of "warm-up time" were found in Opus (with limited initial coded bandwidths), this was fixed by processing all conditions using concatenated samples with the 5 samples for preliminaries at the beginning of the concatenated sequence.

4. Results and discussion

4.1. Overall results

Fig. 1 shows the results of the subjective evaluation on clean speech. Note that the 95% confidence intervals are from ± 0.05 to ± 0.16. DAC is the codec that "stands out of the crowd". It has the best DMOS score among the codecs operating around 1.5 kbps, and increasing bitrate improves quality, to a quality close to Direct (original quality) at less than 8 kbps. However this comes at the price of extra complexity, and significant codec delay (around 190 ms) due to the use of non-causal convolutional layers. DCR is not a transparency test, and the high score of DAC around 8 kbps is simply in the saturation region of the test, given that there are also "bad" conditions that can explain that the high quality range may be compressed.

Regarding "traditional" codecs, care should be taken in making strong conclusions on the comparison between EVS and Opus, Opus is typically used in VBR, EVS with DTX on, and the test with CBR operation makes a fair comparison but does not reflect typical usage. Here, EVS-SWB is close to Direct at 24.4 kbps – DMOS is in the saturation region. Note that Opus at 12 kbps has a coded bandwidth around 8–9 kHz, this bitrate is categorized as WB in Fig. 1.

Except DAC, tested neural audio codecs suffer from limited coded bandwidth that implies a "quality ceiling" in DCR. For instance, EnCodec does not achieve the quality of "Direct" even at 24 kbps and is below EVS and Opus at such bitrate – the bandwidth being limited to 12 kHz. As presented in [10], AudioCraft improves EnCodec's audio quality at low bitrates of 1.5 and 3 kbps, but the tendency is reversed at 6 kbps. When listening informally to the degradation brought by AudioCraft, we note an annoying residual noise in inactive periods. This suggests that this noise is less annoying at 1.5 and 3 kbps considering the poor quality of EnCodec at such bitrates compared to the gain in quality brought using a diffusion model. However, this noise could be more annoying at 6 kbps.

AudioDec at 6.4 kbps/24 kHz is on par with wideband codecs despite extra coded bandwidth; we observe a clear degradation at 12.8 kbps/48 kHz. After inspecting the processed samples, there is a fullband distortion that is already present in the 24 kHz version but judged as more annoying by listeners. Note that more recent pre-trained versions of AudioDec may improve audio quality, however they were not tested or available when preparing this test.

Lyra V2 is better than EnCodec around 3 kbps and close to AudioDec/24 kHz and EVS-WB in the range 6 to 9 kbps, with a rather low complexity considering Tab. 3. LPCNet has a performance close to some other codecs around 1.5 kbps (DAC, AudioCraft); its much lower model complexity (see Tab. 3) makes it attractive for such low bitrate.

4.2. Talker dependency

To further analyse test results, we plotted scores per test condition by separating male and female talkers (scores averaged over 3 talkers/gender), as shown in Fig. 2. The bars of the

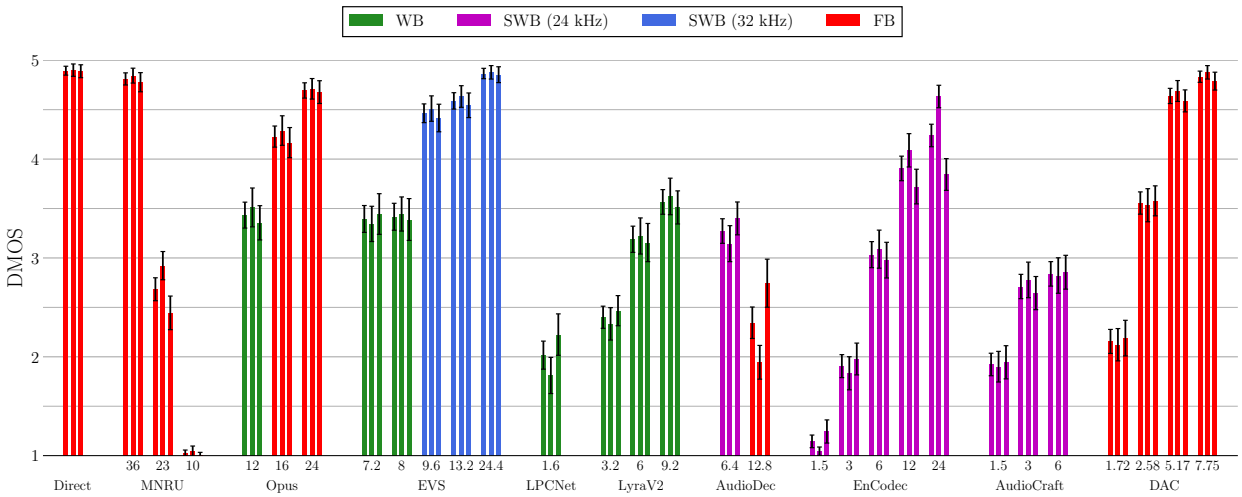

Figure 2: *DCR scores (DMOS) per test condition – in each group of three bars, the mapping, is: first bar = overall average score, second bar = average score for male talkers, third bar = average score for female talkers.*

"Direct" condition show that there is no bias between male and female talkers for the original. It is then possible to find a possible gender dependency for codecs under test. We observe that some codecs show a significant gender dependency: LPCNet and AudioDec (at 48 kHz) seem to better process female samples, while it appears to be the opposite for EnCodec, especially at high bitrates. We verified that these observations on gender dependency are also valid for the dependency on individual talkers. It would be interesting to study root causes for this behaviour.

4.3. Model complexity

While the codecs under test show different performance on speech, we should not forget about codec features (complexity, frame length, delay, etc.) for such capabilities; here we only discuss model complexity. Having a low complexity is necessary when it comes to codec use, for example in a mobile (smartphone) context. A codec needs to operate in real time, on a CPU or dedicated chipset/platform. In order to compare (informally) the computational complexity of codecs under tests, we measured codec execution time on concatenated test samples on a CPU (11th Gen. Intel Core i5-1145G7 @ 2.60 GHz, Windows 10). The minimum execution time was recorded for several processing trials. It should be noted that we are using the implementations described in Tab. 1, which are not all C/C++ implementations. Therefore such measurements are only rough estimates of real execution time on the same platform with the same optimization level, and care should be taken before over-interpreting this extra information.

Tab. 3 presents the real-time factor (RTF) of the tested codecs, i.e. the duration of the processed audio samples divided by the execution time. RTF is computed for encoder and decoder processing only, and for the complete processing (encoder + decoder). A value higher than 1 corresponds to real-time operation (on the computer CPU used here). As the RTF can vary depending on the bitrate used, only the worst-case RTF is presented. Most of the codecs operate in real time, except DAC and AudioCraft. This is not surprising as DAC and AudioCraft have significantly more parameters than other autoencoder-based neural codecs. DAC and AudioCraft have more than 75M and 1150M parameters, respectively, while the other codecs tested here have a maximum of around 20M parameters. For AudioCraft, the limited speed also comes from the diffusion process, as it needs many iterations to achieve a satisfying result.

Table 3: *Real-time factor – underlined values are not real-time.*

Implem.	Codec	Encoder	Decoder	Enc.+Dec.
C/C++	**EVS-WB**	99.6	207.8	67.5
	EVS-SWB	57.1	135.2	44.0
	Opus	32.5	506.8	30.7
	Lyra V2	86.1	108.9	48.2
	LPCNet	111.1	3.8	3.7
Python	**EnCodec**	8.9	10.6	5.1
	AudioDec 24kHz	6.04	7.66	3.38
	AudioDec 48kHz	2.80	2.58	1.34
	DAC	1.23	<u>0.73</u>	<u>0.46</u>
	AudioCraft	8.90	<u>0.04</u>	<u>0.04</u>

5. Conclusion and final remarks

This paper presented an evaluation of speech quality for neural audio codecs. Audio quality was evaluated in one clean speech experiment (in French). We also reported indicative complexity measurements showing that performance gains may come with increased complexity.

Neural audio codecs can achieve much lower bit rates than traditional codecs. In fixed and mobile networks the speech/audio streams at typical bitrates (e.g., 10 to 64 kbps) now represent a tiny portion of network bandwidth and in VoIP the relative overhead of packet headers is significant and the real benefit of very low bitrates (e.g. 1.5 kbps) in terms of general network coverage or capacity may be arguable; still, such new technologies can improve QoE of applications in networks like 2G [45], and they can be used to enhance existing codecs and/or help mitigate packet losses with more efficient low-bitrate redundancy [33]. Note that quantized latent spaces are used to "tokenize" audio frames in other applications [14, 15]. In future work, this study should be extended to get a more complete characterization. A correlation analysis with objective quality methods could also be considered.

6. References

[1] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," in *arXiv:1609.03499*, 2016.

[2] S. Mehri *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, 2017.

[3] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," in *Proc. ICML*, 2018.

[4] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Proc. NeurIPS*, 2018.

[5] W. B. Kleijn *et al.*, "Wavenet based low rate speech coding," in *Proc. ICASSP*, 2018.

[6] C. Garbacea, A. van den Oord, Y. Li, F. S. C. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low Bit-rate Speech Coding with VQ-VAE and a WaveNet Decoder," in *Proc. ICASSP*, 2019.

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

[8] K. Kumar *et al.*, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Proc. NeurIPS*, 2019.

[9] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," arXiv:2010.05646, 2020.

[10] R. S. Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve, and A. Défossez, "From discrete tokens to high-fidelity audio using multi-band diffusion," arXiv:2308.02560, 2023.

[11] Google GitHub repository, "Lyra, version 0.0.2," https://github.com/google/lyra/releases/tag/v0.0.2.

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

[13] R. Vipperla *et al.*, "Bunched LPCNet : Vocoder for Low-cost Neural Text-To-Speech Systems," arXiv:2008.04574, 2020.

[14] Z. Borsos *et al.*, "AudioLM: a Language Modeling Approach to Audio Generation," arXiv:2209.03143, 2023.

[15] J. Copet *et al.*, "Simple and controllable music generation," arXiv:2306.05284, 2024.

[16] N. Zeghidour *et al.*, "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Trans. TASLP*, vol. 30, 2021.

[17] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," in *arXiv:2210.13438*, 2022.

[18] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," in *Advances in Neural Information Processing Systems*, 2023.

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

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010.

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

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

[23] IETF RFC 6716, "Definition of the Opus Audio Codec," Sep. 2012.

[24] IETF RFC 8251, "Updates to the Opus Audio Codec," Oct. 2017.

[25] M. Dietz *et al.*, "Overview of the EVS codec architecture," in *Proc. ICASSP*, 2015.

[26] G. Davidson, M. Vinton, P. Ekstrand, C. Zhou, L. Villemoes, and L. Lu, "High Quality Audio Coding with MDCTNet," in *Proc. ICASSP*, 2023.

[27] L. Xu *et al.*, "An intra-BRNN and GB-RVQ based end-to-end neural audio codec," in *Proc. Interspeech*, 2023.

[28] T. Jenrungrot, M. Chinen, W. B. Kleijn, J. Skoglund, Z. Borsos, N. Zeghidour, and M. Tagliasacchi, "LMCodec: A Low Bitrate Speech Codec with Causal Transformer Models," in *Proc. ICASSP*, 2023.

[29] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," arXiv:2305.02765, 2023.

[30] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "Audiodec: An open-source streaming high-fidelity neural audio codec," in *Proc. ICASSP*, 2023.

[31] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1991.

[32] Meta, "Audiocraft," https://github.com/facebookresearch/audiocraft.

[33] Opus, "libopus 1.5.1," https://opus-codec.org/release/stable/2024/03/04/libopus-1_5_1.html.

[34] 3GPP TS 26.441, "Codec for Enhanced Voice Services (EVS); General overview."

[35] Défossez and others, "EnCodec's GitHub repository," https://github.com/facebookresearch/encodec.

[36] 3GPP TS 26.443, "Codec for Enhanced Voice Services (EVS); ANSI C code (floating-point)."

[37] I. L. Panzer, A. D. Sharpley, and W. D. Voiers, *A Comparison of Subjective Methods for Evaluating Speech Quality*. Boston, MA: Springer US, 1993, pp. 59–65.

[38] ITU-T Handbook, "Handbook of subjective testing practical procedures," 2011.

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

[40] ITU-R Rec. BS.1116–3, "Methods for the subjective assessment of small impairments in audio systems," Feb. 2015.

[41] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," https://github.com/openitu/STL, 2023.

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

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

[44] P. Kabal, "AFsp Package: Audio files programs and routines," https://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/AFsp/AFsp.html.

[45] Android, "Google Duo's new audio technology Lyra gives people high-quality and reliable audio, even on a 2G network," https://twitter.com/Android/status/1366801139547660289, Mar. 2021.