Evaluation of Objective Quality Models on Neural Audio Codecs

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

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

1. Objective Quality and Neural Audio Coding

Context and motivations: A new generation of audio codecs has emerged based on deep learning. Neural audio codecs such as SoundStream, EnCodec or Descript Audio Codec (DAC) demonstrate promising audio quality at low bitrates at the cost of higher computational complexity compared to traditional audio codecs. Several objective quality models have been developed, sometimes using neural networks. There is no clear guidance on which objective model to use in audio coding, especially in the context of neural audio coding.

Objective of this study: Ten models are compared against subjective scores. The selected models are mainly developed to assess speech quality and can be intrusive (i.e. rely on the uncoded reference speech) or not, and target two different bandwidths: wideband (16 kHz sampling) and fullband (48 kHz sampling).

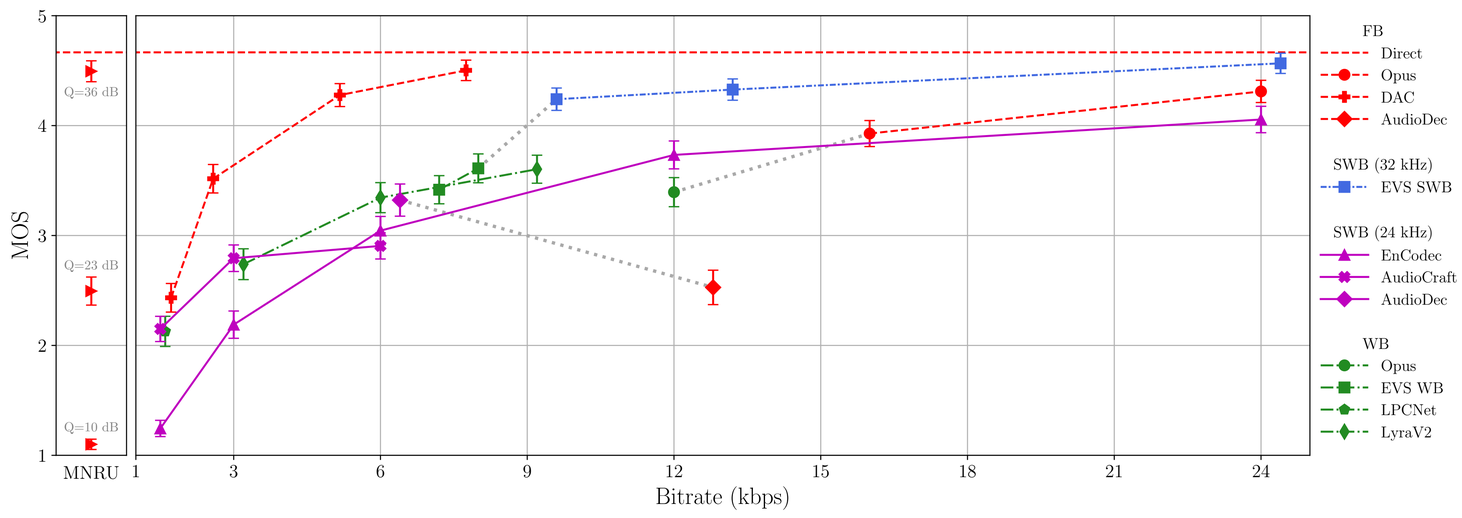
Metric	Content	f_s (kHz)	Intrusive
PESQ	Speech	8,**16**	Yes
POLQA	Speech	8, **48**	Yes
ViSQOL-S	Speech	**16**	Yes
WARP-Q	Speech	8,**16**	Yes
DNSMOS	Speech	**16**	No
NISQA	Speech	**48**	No
NORESQA	Speech	**16**	No*
UTMOS	Speech	**16**	No
PEAQ	Audio	**48**	Yes
ViSQOL-A	Audio	**48**	Yes

∗ Non-matching references (NMRs)

2. Subjective Test (Ground Truth)

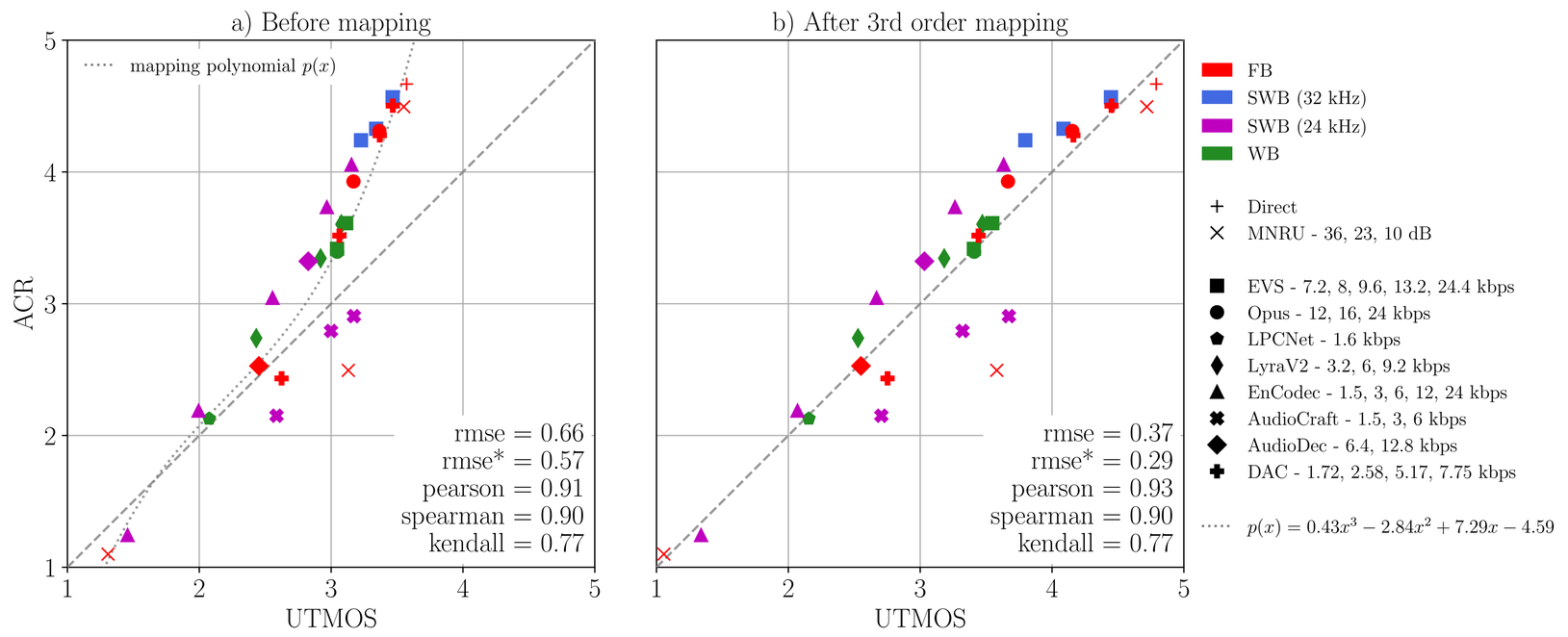
An Absolute Category Rating (ACR) test was conducted with six neural audio codecs and two traditional codecs.
Processing of audio samples is the same as in Muller et al., "*Speech Quality Evaluation of Neural Audio Codecs*," Proc. *Interspeech*, 2024

Codec	f_s (kHz)	L (ms)	bitrate (kbps)
LPCNet	16	10	1.6
Lyra V2	16	20	3.2, 6, 9.2
EnCodec	24	13.3	1.5, 3, 6, 12, 24
AudioCraft	24	13.3	1.5, 3, 6
AudioDec	24	12.5	6.4
DAC	44.1	11.6	1.7, 2.6, 5.2, 7.8
AudioDec	48	6.25	12.8
Opus	48	20	12, 16, 24
EVS-WB	16	20	7.2, 8
EVS-SWB	32	20	9.6, 13.2, 24.4



3. Experimental Results

Comparison with ground truth: All audio samples from the ACR test are processed by objective models. Scores are plotted against ACR test scores. A linear mapping and a 3rd order monotonic polynomial mapping are computed for potential offset, gradient or non-linear relationship correction.



Model evaluation: According to the three chosen evaluation metrics – Peason's correlation coefficient, Kendall's Tau rank correlation coefficient and Root Mean Squared Error (RMSE) – POLQA, UTMOS, PESQ and WARP-Q are best at predicting scores from the ACR test.