Supplemental figures for rebuttal

Below, we show figures that address issues reviewers raised with evaluation, namely fair comparison of our proposed method to existing baselines at the same sample rate, in both quantitative and qualitative metrics. We break down how the proposed method performs on each category of audio (speech, music, environmental sounds), and its performance when trained at 16kbps.

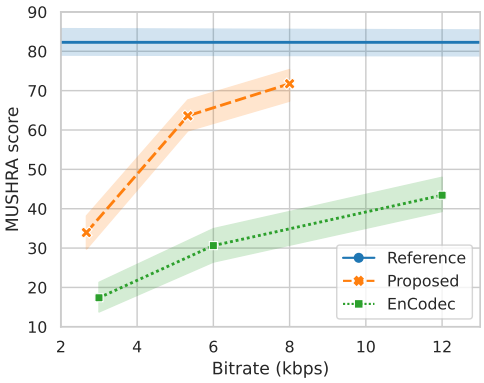

Figure 1: **Listening tests at 24 KHz**: MUSHRA scores with 95% confidence intervals vs bitrate for EnCodec, our proposed approach, and the reference. Here all samples under comparison are resampled to 24 KHz.

Codec	Bitrate (kbps)	Bandwidth (kHz)	Mel distance ↓	STFT distance ↓	ViSQOL ↑	SI-SDR ↑
Proposed@24kHz	1.5	12	1.48	2.24	4.04	0.32
	3	12	1.24	2.01	4.23	4.44
	6	12	1.00	1.78	4.38	8.44
	12	12	0.74	1.54	4.51	12.51
	24	12	0.49	1.33	4.61	16.40
EnCodec	1.5	12	1.63	2.69	3.98	0.02
	3	12	1.46	2.54	4.16	2.99
	6	12	1.30	2.39	4.30	6.06
	12	12	1.15	2.28	4.39	8.44
	24	12	1.05	2.21	4.42	9.69

Table 1: **Encodec Configuration**: Objective evaluation of the proposed model trained with the same exact configuration as EnCodec (24 KHz sampling rate, 24 kbps bitrate, 320 stride, 32 codebooks of 10 bits each) at varying bitrates, along with results from EnCodec.

Codec	Bitrate (kbps)	Bandwidth (kHz)	Mel distance ↓	STFT distance ↓	ViSQOL ↑	SI-SDR ↑
Proposed@8kbps	1.78	22.05	1.39	1.95	3.76	2.16
	2.67	22.05	1.28	1.85	3.90	4.41
	5.33	22.05	1.07	1.69	4.09	8.13
	8	22.05	0.93	1.60	4.18	10.75
Proposed@16kbps	8	22.05	0.95	1.62	4.18	10.00
	16	22.05	0.71	1.46	4.33	14.14

Table 2: **Proposed at 16 kbps**: Objective evaluation of the proposed model with 16 kbps max bitrate at varying bitrates, along with results from the 8 kbps model.

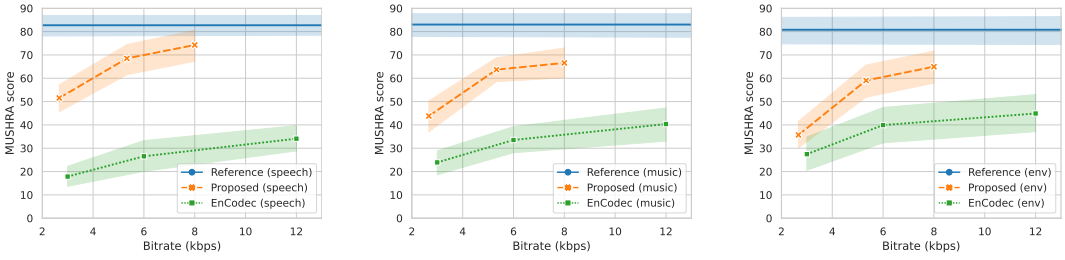

Figure 2: **MUSHRA by category**: MUSHRA scores with 95% confidence intervals vs bitrate for our proposed model, EnCodec and reference.