DATA-CENTRIC TRAINING OF SPEECH-ENHANCING NEURAL CODECS: A STUDY UNDER LOW-BITRATE CONSTRAINTS

S.A. Raczyński, E. Gómez, K. Główczewski, M. Olszewski, K. Duzinkiewicz, and M. Kraiński

Revoize, Inc.
8 The Green, #10383, Dover, DE, 19901, United States
stanislaw@revoize.com

ABSTRACT

We describe a low-resource neural audio codec system designed for the constraints of the 2025 Low-Resource Audio Codec (LRAC) Challenge, Track 2: a streaming-causal pipeline that jointly performs speech enhancement and low-bitrate coding at 24 kHz, decomposed into a discriminative masking-based front-end (Octantis) and a generative residual-vector-quantised neural codec (Canopus), pre-trained separately and then jointly fine-tuned. While the architecture follows established neural-codec design, the focus of this work is data-centric: we ask which property of the training corpus most strongly determines reconstruction quality at 6 kbps. Across five data configurations sharing the same model and the same input degradation pipeline, we find that within the LRAC-derived family the dominant factor is the fidelity of the supervision target. Replacing the LRAC reference signals with versions re-enhanced by a strong speech-enhancement model produces large and statistically significant gains on every objective metric we considered, while quality-assurance filtering of the input corpus alone has small and mixed effects. A separate proprietary corpus shows a metric-dependent pattern of small additional gains on clean 6 kbps speech but no equivalent advantage at 1 kbps or on noisy inputs. Although our system was not formally entered in the challenge, the LRAC organisers evaluated it post-event using challenge-matched methodology; in that pass it scored substantially above the LRAC baseline at 6 kbps on real-world noisy speech.

***Index Terms*—** neural audio codec, speech enhancement, residual vector quantization, low-bitrate coding, LRAC challenge, data-centric training

1. INTRODUCTION

Recent neural audio codecs combine convolutional encoder–decoder networks with residual vector quantisation (RVQ) to represent speech at very low bitrates, with SoundStream [1], EnCodec [2] and DAC [3] as canonical examples. Beyond compression, the same family of architectures can be trained to perform speech restoration: with degraded inputs and clean references, an RVQ codec learns to denoise, dereverberate and bandwidth-extend speech jointly with coding. The 2025 Low-Resource Audio Codec (LRAC) Challenge [4] formalises this setting under realistic deployment constraints — 24 kHz operation, 6 kbps and 1 kbps bitrate modes, fixed compute and latency budgets, and a closed set of public training data.

Most work on codec-based speech enhancement has emphasised architecture and loss design: discrete-token modelling [5, 6, 7], continuous-embedding modelling [8], adversarial training, and multi-scale spectral losses. Far less attention has been paid to how the training corpus is prepared — which recordings are used as references, how aggressively they are curated, and how they are matched to the inputs the codec must reconstruct. Yet a generative codec is, by construction, biased toward whatever clean signal its targets provide: if the references contain residual noise, low-frequency artefacts, or production peculiarities, the model will reproduce them at deployment.

We investigate this question with a system designed for the LRAC Track 2 setting. The system is a two-stage pipeline (*Hadar*) consisting of a discriminative masking-based enhancer (*Octantis*, in the DeepFilterNet family [9, 10]) followed by a streaming RVQ codec (*Canopus*, in the SoundStream/EnCodec/DAC lineage), trained jointly end-to-end with multi-resolution STFT and multi-period discriminators [11] and a multi-resolution mel reconstruction loss. The architectural choices follow established practice. Our contribution is the systematic study of the training-data side of the problem.

Concretely, we hold the model, the input degradation pipeline, the optimisation recipe, and the validation/checkpoint-selection protocol fixed, and vary only the training corpus. Within the LRAC-derived family we explore three axes: **quality-assurance filtering** of the LRAC speech material; **supervision-target fidelity**, by re-enhancing the LRAC references with a high-quality enhancement model; and the **choice of target enhancer** (Octantis vs. DeepFilterNet 3, to test whether the principle generalises beyond a single enhancer). We additionally report results for an **independent proprietary corpus** as a non-LRAC reference point; the

LRAC challenge restricts training to its released data, so we report this variant as an additional data-centric comparison rather than as a challenge-rule-compliant configuration. We report two sets of results. First, an internal objective evaluation on the LRAC blind test set, where paired statistical tests identify supervision-target fidelity as the dominant factor within the LRAC-derived family: re-enhancing the references yields large, highly significant gains on every metric considered, while QA filtering alone has small and mixed effects. Second, post-event subjective listening-test scores produced by the LRAC organisers using methodology matched to the official challenge evaluation, which place our system substantially above the LRAC baseline at 6 kbps on real-world noisy speech.

The contributions of this paper are: (i) a description of the Hadar + Canopus system, designed for the LRAC Track 2 constraints; (ii) a controlled, paired-design ablation of four data-centric training strategies on the LRAC blind test set, plus a comparison to a proprietary corpus as a non-LRAC reference point; (iii) post-event subjective scores for our system, evaluated using LRAC-matched methodology; and (iv) a brief note on *ReMOS v2*, the proprietary fusion metric we used for checkpoint selection.

2. SYSTEM ARCHITECTURE

Our system, *Hadar*, is a streaming-causal pipeline that processes 24 kHz speech in two sequential stages joined by an end-to-end fine-tuning objective:

1. *Octantis* — a discriminative masking-based front-end that suppresses additive noise.
2. *Canopus* — a streaming residual-vector-quantised (RVQ) neural codec that performs joint compression and generative restoration of the remaining, non-additive degradations (reverberation, residual artefacts, bandwidth limitations).

The two stages are first pre-trained independently — Octantis as a denoiser/dereverberator on the LRAC speech, noise and room-impulse-response (RIR) stems, Canopus as an autoencoding codec on clean speech — and then jointly fine-tuned end-to-end under a single GAN objective whose target is *dry, anechoic clean speech*, even when the input is reverberant.

The motivation for this two-stage decomposition is that discriminative denoising is cheap and well-suited to additive noise, while a generative RVQ codec is suited to filling in information a discriminative front-end cannot recover (notably high-frequency content lost to bandwidth limitations or aggressive masking, and direct-to-reverberant ratio inversion). A single neural codec can do both, but allocates one compute budget to two problems with different inductive biases. Pre-training the stages separately and joint-fine-tuning lets each start from an already-competent solution to its own subproblem; the joint stage aligns the two interfaces.

2.1. Octantis

Octantis is a hybrid frequency-domain enhancer in the DeepFilterNet family [9, 10, 12]: a coarse ERB-band envelope mask covers the full spectrum, and a low-frequency complex deep-filter branch refines the spectral region where speech energy is concentrated. The Hadar configuration uses an STFT analysis of 480 samples (20 ms) with 240-sample (10 ms) hop and a single frame of look-ahead. The front-end has approximately 1.3 M parameters.

2.2. Canopus

Canopus is a streaming-causal RVQ codec in the SoundStream/EnCodec/DAC [1, 2, 3] lineage, with Snake1d periodic activations [13]. The encoder is a stack of four convolutional blocks with strides $2 \cdot 3 \cdot 5 \cdot 8 = 240$, yielding a 100 fps frame rate at 24 kHz. The RVQ bottleneck has 6 quantisers of 1024 codes with 8-dimensional codebook entries (L2-normalised, cosine-distance, DAC-style), giving 60 bits per frame and a nominal bitrate of **6 kbps**. The LRAC challenge requires a single system supporting both 1 kbps and 6 kbps operation, including switching between modes within one inference run; we satisfy this with quantiser dropout during training [2], so that at inference the system operates at any integer bitrate from 1 to 6 kbps by truncating the active set of quantisers, with no retraining. Our dropout schedule keeps the full 6-quantiser stack active for half of training samples; the other half draws a random truncation length uniformly in $\{1, \ldots, 6\}$, so the 1 kbps configuration receives substantially less training time than 6 kbps. The decoder mirrors the encoder. The generator has approximately 7–8 M parameters.

2.3. Discriminators and loss

Two discriminator stacks are run jointly: a multi-resolution complex-STFT discriminator (EnCodec-style [2]) and a multi-period discriminator (HiFi-GAN-style [11]). The generator loss combines least-squares adversarial losses on both discriminators with deep feature matching, an L_1 time-domain loss, and a multi-resolution mel-spectrum reconstruction loss; the mel term carries the largest weight. RVQ commitment and codebook losses are added per quantiser. Optimisation uses AdamW ($\beta_1 = 0.8$, $\beta_2 = 0.99$, learning rate 10^{-4}) with one discriminator step followed by one generator step per training iteration.

2.4. Operating point

The deployed system is strictly causal at inference, with an algorithmic look-ahead of approximately 30 ms (one Octantis STFT frame plus a two-frame Canopus pre-roll) and a 10 ms frame hop.

3. TRAINING DATA AND DATA-CENTRIC STRATEGIES

3.1. Data pipeline

All variants share the same on-the-fly degradation generator. For each training example, a 3-second window is constructed

by concatenating speech chunks (with random interstitial silence drawn from $\mathcal{U}[20, 250]$ ms) and is then convolved with a randomly drawn RIR with probability 0.5 and mixed with noise at SNR drawn from $\mathcal{U}[-5, 30]$ dB. The supervision target is always the *dry, anechoic clean speech*, so the system is trained to denoise, dereverberate and code in a single objective. Inputs and targets are jointly peak-normalised to $\mathcal{U}[-12, 0]$ dB.

3.2. Variants

We compare five data variants (D1–D5). Variants D1–D4 differ *only* in which LRAC-derived speech, noise and RIR stems are presented to the training pipeline; the model, optimiser, training schedule, validation set and checkpoint-selection protocol are identical across variants. D5 trains on a different (proprietary) corpus and is included as a non-LRAC reference point; it was *not* eligible for the LRAC challenge, which restricts training to its own released data, and is reported here only as a data-centric comparison.

D1 – LRAC (uncurated). The official LRAC training data [4] as released — approximately 703 h of speech, plus the LRAC noise and RIR stems — with no filtering.

D2 – LRAC (curated). D1 with a quality-assurance filter applied to the speech corpus. Recordings are excluded if any of the following hold: estimated frequency cut-off below 10 kHz; speech-content fraction below 0.75; clipping detected; DC-offset or strong low-frequency rumble detected; short glitches detected; an audible constant tone detected; sound-pressure level below -45 dBFS; or a measurable but moderate noise floor (the noise sRMS estimate falls within an audible-noise range, as opposed to either being inaudibly quiet or being unmeasurable, both of which we accept). The filter removes approximately 14 % of the data, leaving 86 %.

D3 – LRAC (curated, DFN3-enhanced targets). D2 with the speech references re-enhanced by DeepFilterNet 3 [12]. Inputs to the codec are unchanged; only the supervision targets are modified.

D4 – LRAC (curated, Octantis-enhanced targets). D2 with the speech references re-enhanced by Octantis. This is the configuration used by our system in the post-event subjective evaluation reported in Sec. 5. The Octantis instance used as a target enhancer is trained for full-band enhancement and is distinct from the in-pipeline front-end configuration, but they share an architecture; D3 is included specifically to test whether the principle (re-enhanced targets) generalises beyond a single enhancer.

D5 – Proprietary corpus (non-LRAC reference). A larger Revoize-internal speech corpus with extensive manual curation and broader coverage of recording conditions, microphones and acoustic environments than the LRAC training set, used in place of the LRAC speech material. Noise and RIR stems and the rest of the degradation pipeline are unchanged. The LRAC challenge restricts training to its released data, so D5 is not a challenge-rule-compliant configuration; we report it only as a non-LRAC data-centric reference point.

4. EXPERIMENTAL SETUP

4.1. Test sets

We evaluate on the LRAC *blind test set* (190 utterances, 90 clean and 100 noisy, drawn from real-world recordings outside the LRAC training distribution) for the data-strategy ablation, and additionally report post-event subjective scores for our system, produced by the LRAC organisers using methodology matched to the official challenge listening tests (Track 2a, clean speech; Track 2c, noisy speech).

4.2. Objective metrics

We report three families of non-intrusive predicted-MOS metrics: SHEET-SSQA [14]; the four AudioBox-Aesthetics axes [15] — Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE) and Content Usefulness (CU); and *ReMOS v2* (described below). The AudioBox-Aesthetics paper notes that PC does not correlate with speech quality; we omit it from the main table.

4.3. ReMOS v2

ReMOS v2 is a proprietary fusion metric trained to predict human MOS on a Revoize-internal corpus of approximately 2 years of subjective ratings, with the corpus deliberately weighted toward *processed* speech — speech that has been passed through enhancement systems, codecs and other algorithms, and that is therefore exposed to the kinds of distortions and artefacts that codec-based restoration systems produce. ReMOS v2 fuses scores from four existing non-intrusive predictors — DNSMOS [16], MOSA-Net+ [17], SHEET [14] and SCOREQ [18] — through a small classical-ML regressor. Note that SHEET also appears as a standalone column in Table 1, so SHEET and ReMOS v2 are not statistically independent. On BVCC, ReMOS v2 achieves Spearman 0.804 with the reference scores; on a held-out combined NISQA + SOMOS + VMC23 set, it achieves Spearman 0.61, in both cases at or above the strongest individual non-intrusive predictor we tested. Because it is calibrated for the processed-speech regime that is most relevant here, we use ReMOS v2 as our primary screening metric.

4.4. Checkpoint selection

All variants were trained to convergence under identical conditions. For each variant we selected the best checkpoint by ranking late-stage checkpoints with a weighted composite of ReMOS v2 scores on a held-out validation set, with weights matching the ranking protocol used in the LRAC challenge.

4.5. Statistical methodology

Because all five data variants are evaluated on identical sets of test utterances (the same 190 inputs are passed through each model and scored by the same predictor), all comparisons among D1–D5 are paired. We report paired t-tests against D1

(uncurated) for D2–D5. The LRAC baseline is evaluated on the same source utterances as our systems, so paired tests involving the baseline are also valid where reported in the text.

5. RESULTS

5.1. Objective results: data-strategy ablation

Fig. 1 summarises the data-strategy comparison on ReMOS v2. Within the four LRAC-derived variants (D1–D4), performance increases nearly monotonically from D1 through D4 across all four conditions, and the D2–D4 gains over D1 are statistically significant in every panel. The *magnitude* of the improvement, however, is uneven: QA filtering alone (D2 vs. D1) delivers small gains, target re-enhancement (D3 and D4 vs. D2) delivers substantially larger gains, and Octantis-enhanced targets (D4) consistently outperform DeepFilterNet-3-enhanced targets (D3).

The proprietary-corpus variant D5, included as a non-LRAC reference point, shows a metric-dependent pattern: on AudioBox-Aesthetics and SHEET it matches or exceeds D4 on clean 6 kbps (Table 1), while on ReMOS v2 (Fig. 1) it matches D4 only on clean 6 kbps and lags D4 on the harder conditions (noisy speech and 1 kbps).

The full ablation across the five reported metrics is given in Table 1. The pattern holds across SHEET, ReMOS v2 and the three speech-relevant AudioBox axes. Counting wins among the five D-variants: D4 is the best D-variant in 11 of 20 rows — almost always at 1 kbps and on noisy speech; D5 is the best in 8 of 20 rows — almost always at 6 kbps clean speech; D3 is best in the remaining row. Comparing to the LRAC baseline, our system in the D4 configuration is below the baseline at 1 kbps but above the baseline on every metric at 6 kbps. This regime split is consistent with the post-event subjective evaluation, reported next.

5.2. Subjective results: post-event evaluation

Our system was not formally entered in the 2025 LRAC Challenge. After the conclusion of the challenge, the LRAC organisers ran two listening tests on it for us, using methodology matched to the official challenge evaluation; the two passes are designated 2a-Revoize and 2c-Revoize by the organisers. *Track 2a (clean speech)* methodology uses MUSHRA-1S; in our pass only our system was rated alongside the shared anchor and reference. The organisers normalised the 2a-Revoize scores using the same procedure as in the official 2025 LRAC Challenge — applying an affine per-file mapping based on precomputed anchor and reference means from the original challenge — placing the scores directly on the official 2a-LRAC scale. After this normalisation, our system scored 80.6 at 6 kbps and 9.1 at 1 kbps on the 0–100 MUSHRA scale. *Track 2c (real-world noisy speech)* methodology uses ACR/MOS; we report raw 2c-Revoize values as primary results and, per the organisers' guidance, do not directly compare them with the original 2c-LRAC pass. The organisers also rated three official LRAC Track 2 entries in the 2c-Revoize pass for context: the LRAC Track 2 baseline [19], `nano_codec` [20], and `nju-aalab` [21]. Fig. 2 reports the 2c-Revoize ACR/MOS scores.

At 6 kbps our system scored 3.03 on noisy speech — substantially above the LRAC baseline (2.15, $+0.88$ MOS) and `nano_codec` (2.93), and below `nju-aalab` (3.31). These pairwise comparisons are within the 2c-Revoize pass and do not constitute an official challenge ranking. At 1 kbps our system scored close to the LRAC baseline and below the other two systems shown; we discuss the likely reasons in Sec. 6. The organisers also provide an auxiliary affine alignment of 2c-Revoize onto the original 2c-LRAC scale (yielding approximately 1.20 and 3.15 for our system), which we report only as context.

6. DISCUSSION

6.1. Supervision-target fidelity is the dominant lever

Across every objective metric, every condition and both bitrates, replacing the LRAC reference signals with versions re-enhanced by a strong speech enhancement model (D3 or D4 vs. D2) yields larger and more consistent gains than QA filtering alone (D2 vs. D1). This is intuitive in retrospect: a generative codec inherits the statistical signature of its supervision targets, including any residual noise, low-frequency rumble, or production peculiarities. QA filtering removes the worst examples but leaves untouched the typical-quality signal that dominates the loss; re-enhancing the references shifts the entire distribution the codec is asked to reproduce. The fact that DFN3-enhanced targets (D3) also help, and that the Octantis advantage over DFN3 is smaller than the D3-over-D2 advantage, supports the generality of the principle: the gains are not a self-reference effect of training the codec to match the front-end's idiosyncrasies.

6.2. Why our system underperforms at 1 kbps

Our system underperforms the LRAC baseline at 1 kbps on every metric. Since all systems including ours and the baseline obtain 1 kbps by truncating the active RVQ set at inference (Sec. 2), the differences lie in training, not in the runtime mechanism. Two factors are likely responsible. First, our quantiser-dropout schedule keeps the full 6-quantiser stack active for half of training and samples shorter truncations the rest of the time, so the 1 kbps configuration sees substantially less training signal than 6 kbps; the LRAC baseline, by contrast, trains with a 50/50 mix of 1 and 6 kbps [19]. Second, supervision-target re-enhancement helps the codec reproduce nuances that a 1 kbps bottleneck cannot represent — the data-centric gains we see at 6 kbps depend on capacity that 1 kbps does not have.

6.3. What does D5 add?

The proprietary-corpus variant D5 is not a challenge-rule-compliant configuration (the LRAC challenge restricts training to its released data) and serves only as a data-centric reference: what does an independent, high-quality speech cor-

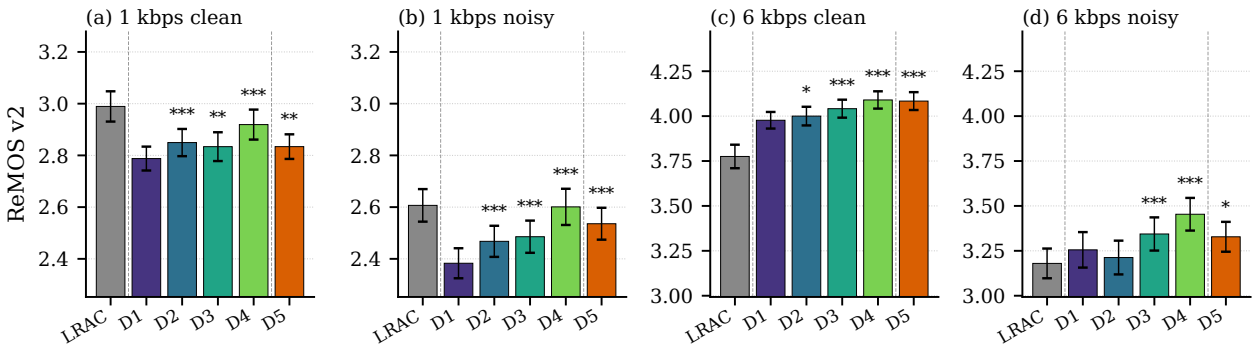

Fig. 1. ReMOS v2 (predicted MOS, higher is better) on the LRAC blind test set, broken down by bitrate and speech condition. Bars show the LRAC baseline (grey) and five data-centric variants of our system: four LRAC-derived variants in the viridis palette (D1: uncurated; D2: curated; D3: curated with DeepFilterNet-3-enhanced targets; D4: curated with Octantis-enhanced targets) and the proprietary-corpus variant D5 in orange. D5 is not a challenge-rule-compliant configuration and is reported as a non-LRAC reference point. Error bars are 95 % confidence intervals on the mean. Significance markers above D2–D5 indicate paired t-tests against D1 ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

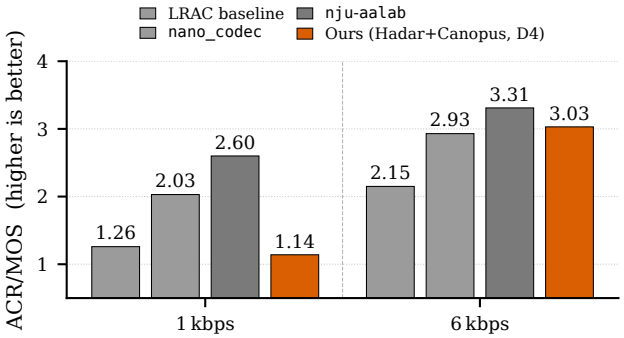

Fig. 2. ACR/MOS scores on real-world noisy speech (2c-Revoize), from a post-event evaluation by the LRAC organisers using methodology matched to the official LRAC Track 2c listening test. Higher is better; scale is 1–5. Our system was not a formal LRAC entrant; the three other systems were entrants in the official challenge and are shown here for context.

pus look like compared to the best LRAC-derived variant? The picture is mixed and metric-dependent. On ReMOS v2 (Fig. 1), our own metric calibrated for processed speech, D5 matches D4 on clean 6 kbps and lags D4 on the other three conditions; on AudioBox-Aesthetics and SHEET (Table 1), D5 matches or exceeds D4 on clean 6 kbps but lags D4 on noisy speech and on 1 kbps operation. A consistent reading is that a corpus that already contains cleaner reference speech raises the ceiling on what the codec can reproduce when capacity allows (clean 6 kbps), but provides no equivalent advantage when the bottleneck dominates (1 kbps), and may underperform when its noise or RIR coverage diverges from the test distribution (real-world noisy speech). We do not attempt to attribute the metric-by-metric disagreement between ReMOS v2 on the one hand and AudioBox-Aesthetics/SHEET on the other to any specific cause. Within the LRAC-derived family, the same ceiling-raising effect on clean speech can be obtained at lower data-cost by re-enhancing the LRAC references, which is what D4 does.

6.4. Limitations

Our objective ablation uses a single test set (the LRAC blind set, 190 utterances) and a fixed set of non-intrusive predictors; PESQ and UTMOS were not available for our systems and are absent from Table 1. The QA filter we describe is one specific operationalisation of "data quality"; other filters might allocate the QA-versus-enhancement trade-off differently. Finally, D5 differs from D1–D4 along multiple axes simultaneously (corpus, scale, curation procedure), so we draw conclusions from D5 only at the level of overall ceiling, not at the level of any one factor.

7. CONCLUSION

We presented the Hadar/Canopus system, designed for the constraints of the 2025 LRAC Challenge Track 2, framed as a study of which property of the training corpus most strongly determines codec performance at 6 kbps. With the model, degradation pipeline and selection protocol fixed, we found that within the LRAC-derived family supervision-target fidelity is the dominant data-centric lever: re-enhancing the LRAC references with a strong external enhancement model produces large and statistically significant objective gains, QA filtering alone produces small ones, and the principle holds across two target enhancers. An independent proprietary corpus shows small additional gains on clean 6 kbps speech on some metrics but no equivalent advantage at 1 kbps or on noisy inputs. In a post-event evaluation by the LRAC

Metric	Cond.	Bitrate	LRAC	D1	D2	D3	D4	D5
SHEET	clean	1 kbps	2.16	1.57	1.58	1.60	**1.66****	1.65***
	clean	6 kbps	3.84	4.15	4.23	4.39***	4.42***	**4.48*****
	noisy	1 kbps	2.22	1.41	1.40	1.50***	**1.60*****	1.49***
	noisy	6 kbps	2.89	3.07	2.97	3.39***	**3.48*****	3.21*
ReMOS v2	clean	1 kbps	2.99	2.79	2.85***	2.83**	**2.92*****	2.83**
	clean	6 kbps	3.78	3.98	4.00*	4.04***	**4.09*****	4.08***
	noisy	1 kbps	2.61	2.38	2.47***	2.49***	**2.60*****	2.54***
	noisy	6 kbps	3.18	3.26	3.21	3.34***	**3.45*****	3.33*
AB-PQ	clean	1 kbps	4.91	3.62	3.87***	4.03***	**4.38*****	3.79***
	clean	6 kbps	6.48	6.98	7.04*	7.26***	7.23***	**7.36*****
	noisy	1 kbps	4.92	3.72	4.12***	4.23***	**4.61*****	3.96***
	noisy	6 kbps	5.40	6.05	6.18*	**6.48*****	6.36***	6.41***
AB-CE	clean	1 kbps	4.00	2.91	2.88	3.07***	**3.37*****	3.12***
	clean	6 kbps	5.29	5.66	5.64	5.71**	5.75***	**5.75*****
	noisy	1 kbps	3.33	2.66	2.59**	2.81***	**2.93*****	2.82***
	noisy	6 kbps	4.00	4.34	4.21**	4.36	4.36	**4.40**
AB-CU	clean	1 kbps	4.74	2.97	3.32***	3.82***	**4.21*****	3.65***
	clean	6 kbps	6.03	6.46	6.50	6.70***	6.65***	**6.83*****
	noisy	1 kbps	4.12	2.97	3.33***	3.75***	**4.02*****	3.59***
	noisy	6 kbps	4.51	4.94	4.96	5.28***	5.19***	**5.32*****

Table 1. Full ablation on the LRAC blind test set across five non-intrusive predictors. Higher is better for all metrics shown. Bold marks the best D-variant per row. Significance markers (vs. D1, paired t-test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. AB-PQ, AB-CE, AB-CU are the Production Quality, Content Enjoyment and Content Usefulness axes of Audiobox-Aesthetics [15]; AB-PC is omitted because Audiobox-PC does not correlate with speech quality. D5 (proprietary corpus) is not a challenge-rule-compliant configuration and is shown only as a non-LRAC reference point.

organisers using challenge-matched methodology, our system scored substantially above the LRAC baseline at 6 kbps on real-world noisy speech.

8. REFERENCES

[1] N. Zeghidour *et al.*, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 495–507, 2022.

[2] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. on Machine Learning Research*, 2023.

[3] R. Kumar *et al.*, "High-fidelity audio compression with improved RVQGAN," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[4] K. Wojcicki *et al.*, "Low-resource audio codec (LRAC): 2025 challenge description," *arXiv preprint arXiv:2510.23312*, 2025.

[5] H. Yang, J. Su, M. Kim, and Z. Jin, "Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens," in *Proc. Interspeech*, pp. 1170–1174, 2024.

[6] Z. Wang *et al.*, "Selm: Speech enhancement using discrete tokens and language models," in *Proc. ICASSP*, pp. 11561–11565, 2024.

[7] H. Xue, X. Peng, and Y. Lu, "Low-latency speech enhancement via speech token generation," in *Proc. ICASSP*, pp. 661–665, 2024.

[8] H. Li, J. Q. Yip, T. Fan, and E. S. Chng, "Speech enhancement using continuous embeddings of neural audio codec," in *Proc. ICASSP*, 2025.

[9] H. Schröter, A. N. Escalante-B., T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. ICASSP*, pp. 7407–7411, 2022.

[10] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[11] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "DeepFilterNet: Perceptually motivated real-time speech enhancement," in *Proc. Interspeech*, 2023.

[13] S.-g. Lee *et al.*, "BigVGAN: A universal neural vocoder with large-scale training," in *International Conference on Learning Representations (ICLR)*, 2023.

[14] W.-C. Huang, E. Cooper, and T. Toda, "SHEET: A multi-purpose open-source speech human evaluation estimation toolkit," in *Proc. Interspeech*, pp. 2355–2359, 2025.

[15] A. Tjandra *et al.*, "Meta Audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv preprint arXiv:2502.05139*, 2025.

[16] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, pp. 6493–6497, 2021.

[17] R. E. Zezario *et al.*, "A study on incorporating Whisper for robust speech assessment," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2024.

[18] A. Ragano, J. Skoglund, and A. Hines, "SCOREQ: Speech quality assessment with contrastive regression," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[19] Y. Z. Isik and R. Łaganowski, "Baseline systems for the 2025 low-resource audio codec challenge," *arXiv preprint arXiv:2510.00264*, 2025.

[20] A. Li *et al.*, "Enhance-NanoCodec: Enhancement neural audio codec for LRAC 2025," LRAC 2025 Sys. Desc. Report, Inst. of Acoustics, CAS, 2025.

[21] R. Hu *et al.*, "Progressive refinement training for low-resource neural speech coding and enhancement," LRAC 2025 Sys. Desc. Report, Nanjing Univ., 2025.