HORCODEC: HORNET BASED NEURAL AUDIO CODEC FOR THE LRAC 2025 CHALLENGE TRACK 1

Qingbo Huang, Weihao Xiong, Congxin Zhang, Xinmin Yan

ByteDance

ABSTRACT

This paper is a description of our team's submission model for LRAC track 1, introducing the HORCODEC based on HORNET, including model structure, training methods, and other details. By introducing Horunit into classic methods such as soundstreaemDAC model and RVQ, our model can consistently improve dense prediction performance with less computation, achieving transparent sound quality as much as possible within the low complexity requirement by LRAC.

***Index Terms*—** neural audio codec, residual vector quantilization

1. INTRODUCTION

High quality and low latency audio encoding algorithms are crucial in real-time communication field. With the rapid development of deep learning technology in recent years, audio codec based on deep neural networks, represented by soundstream[1], DAC[2], have significantly improved compression efficiency compared to traditional audio encoders such as AAC and OPUS. However, the high latency and high complexity of encoding and decoding are fatal flaws of deep neural network-based audio codecs, which prevent them from being widely used in real-time communication. Regarding this issue, LRAC competition track 1 has made clear regulations on the complexity and delay of encoder encoding and decoding. This is extremely challenging for deep neural networks-based audio encoders. To achieve the ultimate goal of low complexity, low latency, and transparent sound quality, we have researched the current mainstream audio encoding methods based on deep neural networks and have referred to the forefront of deep learning in computer vision. Based on the DAC and VOCOS frameworks, we have added the improvement of the basic modules in the transformers described in HORNET[3]. Under the premise of satisfying the requirements of complexity and latency in LRAC, the sound quality of our proposed codec is as close as possible to the original high-complexity DAC.

2. MODEL STRUCTURE

The over view of the proposed codec is shown in Figure 1. The input audio is divided into frames with a frame length of 20ms. The output feature of the encoder network is coded by RVQ, with 0.5kbps for each layer. On the decoder side, the input feature is transformed to the frequency domain, and then audio signals in the time domain are generated by ISTFT.

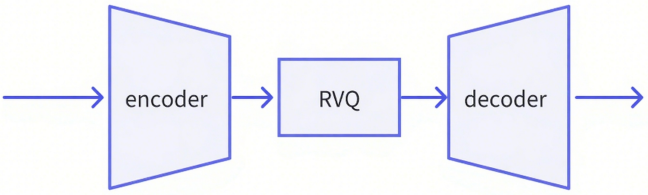

Fig. 1. overview

2.1. Encoder Block

The encoder structure is shown in Figure 2. The first 1D convolution module is set to kernel size k=7. Then four residual convolution modules are applied in sequence with each stride = 3, 4, 5, 8. For each residual convolution module, there are 3 residual units in it and each residual unit's dilation is 1, 3, 9 respectively.

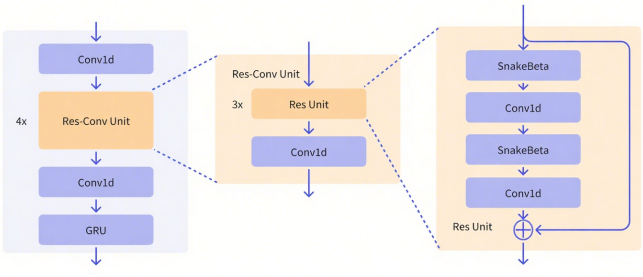

Fig. 2. encoder block

2.2. Decoder Block

Inspired by Hornet in computer vision, we modified the module based on 2D convolution design in Hornet and applied it to audio signal processing. The decoding end receives the quantized feature vectors, which are sequentially processed through one 1D convolution module, 6 horunit modules, and one 1D convolution module before being converted to the frequency domain. The frequency domain signal is then transformed back to the time domain through ISTFT. Each horunit module contains one gConv gating module and one FNN module in sequence, with the gConv gating order set to 3, as shown in Figure 3.

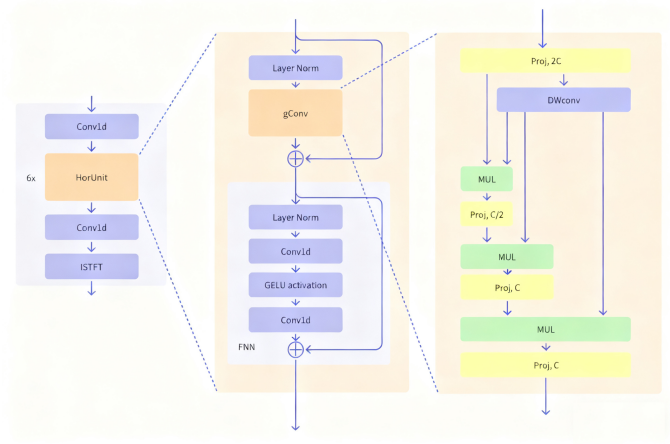

Fig. 3. decoder block

2.3. Quantizer

The features outputted from each frame undergo RVQ (Residual Vector Quantization) hierarchical residual layer coding, consisting of 12 layers. Each layer has 1024 codeword candidates, which requires 10 bits per layer during encoding. Given that the encoding segment is divided into 20ms frames, the bit rate for one layer of RVQ quantization stands at 0.5kbps. If the target bit rate is 6kbps, all 12 layers of RVQ are employed; whereas if the target bit rate is 1kbps, only the first two layers of RVQ are utilized.

2.4. Computational Complexity

The parameter count and computational complexity of each module in the model are shown in Table 1

2.5. System Latency

Since the encoder frame length is 480 points (i.e. 20ms) and there are 240 points (i.e. 10ms) frame overlapping in the decoder, the system latency is 30ms, satisfying the challenge requirement.

	Parameter Count	Model Complexity
Enocder	0.98M	172.97MMACs
Quantizer	0.19M	1.06MMACs
Decoder	3.01M	154.78MMACs
Total	4.18M	328.82MMACs

Table 1. Computational Complexity

3. TRAINING

3.1. Data Processing

On the premise of complying with the competition requirements, we have pre-processed the data provided by the official to achieve data augmentation. When generating noisy frequencies, randomly select SNR within a preset interval. When generating reverberation data, follow the official method and generate it with an appropriate reverberation ratio. For multispeaker data, randomly adjust the volume of a certain speaker.

3.2. Loss Setups

Training is carried out in the form of a generative adversary mode,which is the same as SoundStream and DAC. As described in Equation 1, the total loss of the model is consist of GAN-based loss $\mathcal{L}_g$, RVQ commit loss $\mathcal{L}_c$, RVQ code book loss $\mathcal{L}_r$ related to RVQ to improve the efficiency of code book utilization. The reconstruct loss $\mathcal{L}_{re}$ is set to ensure that reconstructed signal is as consistent as possible with the reference input.

$$\mathcal{L} = \lambda_g * \mathcal{L}_g + \lambda_c * \mathcal{L}_c + \lambda_r * \mathcal{L}_r + \lambda_{re} * \mathcal{L}_{re} \quad (1)$$

Since the reconstruction loss does not occupy the complexity of encoding and decoding, we set a loss function as detailed as possible to evaluate the quality of the reconstructed signal, although this may slow down the training process. The reconstruct loss is set with multiscale STFT loss $\mathcal{L}_{stft}$, multiscale MEL loss $\mathcal{L}_{mel}$, PESQ[4] loss $\mathcal{L}_{peaq}$. The multiscale STFT loss is set with window lengths of 256, 512, 1024 and 2048. The multiscale MEL loss is set with window lengths of 32, 64, 128, 256, 512, 1024 and 2048, corresponding mel bin counts of 5, 10, 20, 40, 80, 160, and 320, respectively. All loss functions are weighted with appropriate coefficients as part of the final loss.

$$\mathcal{L}_{recon} = \lambda_s * \mathcal{L}_{stft} + \lambda_m * \mathcal{L}_{mel} + \lambda_p * \mathcal{L}_{peaq} \quad (2)$$

All the weight coefficients are described in Table 2.

3.3. Network Training Configurations

The learning rate is initialized at 0.0001 and decays by a factor of 0.999996 every epoch, as described in the exponential

loss weight	value
λ_g	1.0
λ_c	0.25
λ_r	1.0
λ_{re}	1.0
λ_s	10.0
λ_m	15.0
λ_p	5.0

Table 2. loss weight configuration

learning rate scheduling technique. The Optimization is performed with Adam, using betas of 0.8 and 0.99.

We train the networks with a batch size of 16 per GPU, and 8 GPUs were used in training progress for Track 1 in total. The model is trained for 500 epochs. We use the checkpoint with the lowest reconstruction loss on the validation set. The reconstruction loss configuration is described in subsection 3.2.

4. REFERENCES

[1] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[2] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser-Nam Lim, and Jiwen Lu, "Hornet: Efficient high-order spatial interactions with recursive gated convolutions," *arXiv preprint arXiv:2207.14284*, 2022.

[4] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2001, vol. 2, pp. 749–752 vol.2.