An Intra-BRNN and GB-RVQ Based END-TO-END Neural Audio Codec

Linping Xu[2*], Jiawei Jiang[1], Dejun Zhang[1], Xianjun Xia[1], Li Chen[1],
Yijian Xiao[1], Piao Ding[1], Shenyi Song[1], Sixing Yin[2], Ferdous Sohel[3]

[1] RTC Lab, ByteDance, Beijing, China
[2] Beijing University of Posts and Telecommunications, Beijing, China
[3] School of Information Technology, Murdoch University, Perth, Australia

[1]{jiangjiawei.lahm, zhangdejun, xiaxianjun, chenli.cloud, dingpiao1,
songshenyi1}@bytedance.com [2]{yunzhongxue521, yinsixing}@bupt.edu.cn,
[3]F.Sohel@murdoch.edu.au

Abstract

Recently, neural networks have proven to be effective in performing speech coding task at low bitrates. However, underutilization of intra-frame correlations and the error of quantizer specifically degrade the reconstructed audio quality. To improve the coding quality, we present an end-to-end neural speech codec, namely CBRC (Convolutional and Bidirectional Recurrent neural Codec). An interleaved structure using 1D-CNN and Intra-BRNN is designed to exploit the intra-frame correlations more efficiently. Furthermore, Group-wise and Beamsearch Residual Vector Quantizer (GB-RVQ) is used to reduce the quantization noise. CBRC encodes audio every 20ms with no additional latency, which is suitable for real-time communication. Experimental results demonstrate the superiority of the proposed codec when comparing CBRC at 3kbps with Opus at 12kbps.

Index Terms: Intra-BRNN, Group-wise RVQ, Beam-search RVQ

1. Introduction

Traditional audio codecs utilizing psycho-acoustics and pronunciation models are capable of producing high-quality audio at medium-to-high bitrates, whereas fail to produce satisfactory perceptual quality due to the inefficient coding architecture at low bitrates. With the development of deep learning, data-driven neural audio codecs provide a new direction for high-quality speech coding at low bitrates.

Neural networks used in audio codec can mainly be divided into three categories: post-processors, neural decoders and end-to-end neural codecs. Related works in [1, 2, 3] show that the quality of existing codecs can be improved by using neural network based post-processors without changing the traditional codec's bitstream. Typical neural decoders are dependent on carefully designed audio features and generative models to reconstruct audio waveforms, such as WaveNet [4], Lyra [5], LPCNet [6], SampleRNN [7] and SSMGAN [8]. They perform better than traditional codecs at low bitrates, especially with the use of GAN.

Different from post-processors and neural decoders, the learnable encoder in end-to-end neural codec significantly improves coding efficiency with respect to traditional codec. A learnable residual vector quantizer (RVQ) is adopted in SoundStream [9] which makes it the fully end-to-end codec. Sound-Stream permits to encode audio at bitrates ranging from 3 kbps to 12 kbps with structured dropout applied to RVQ during training. Experimental results in that work demonstrate that Sound-stream is robust under a wide range of real-life coding scenar-

ios. Another end-to-end codec, TFNet [10], takes temporal filtering blocks to explicit audio feature and investigates the joint optimization considering both speech enhancement and packet loss concealment task. More recently, NESC [11] adopts the DPCRNN as the main building block for efficient and reliable encoding and demonstrates its robustness under various noise and reverberation levels.

Although SoundStream and NESC both achieve better quality than traditional codecs at low bitrates, the encoder and decoder in SoundStream adopt fully convolutional network, which pays little attention to intra-frame correlations. However the state-of-the-art traditional codecs, e.g. Opus and EVS, demonstrate the importance of intra-frame correlations. Meanwhile, RVQ used in SoundStream and NESC provides a basis for end-to-end optimization. But there is still a large gap between the potential bitrate achievable by entropy coding and the actual bitrate of RVQ.

To utilize the intra-frame correlations and reduce the quantionzation error, we propose CBRC, an end-to-end neural codec, capable of coding wideband audio at 3kbps. In encoder and decoder of CBRC, each Convolutional and Bidirectional Recurrent Neural Network Block (CBRNBlock) adopts an interleaved network using 1D-CNN and intra-frame bidirectional RNN (Intra-BRNN) to effectively capture intra-frame correlations. Meanwhile, Group-wise and Beam-search RVQ are designed in quantizer to reduce the quantization error. This paper compares the quality of audio generated by our proposed CBRC, Lyra-V2 and Opus between 3kbps and 12kbps. It is worth noting that CBRC at 3kbps outperforms Opus at 12kbps.

Our contributions are listed as below: i) The model architecture with Intra-BRNN and ii) The Group-wise RVQ and Beam-search RVQ methods adopted during quantization, which are the main aspect of novelty of our proposed CBRC.

2. Proposed Audio Codec

Figure 1 shows the flowchart of our proposed codec, which consists of an encoder, a quantizer and a decoder. The encoder maps audio inputs to a sequence of embeddings. The quantizer performs to compress the embeddings with a target number of bits. The decoder is adopted to reconstruct audio from quantized embeddings. The codec is trained end-to-end with discriminator and adversarial loss to improve the perceptual quality of the reconstructed audio.

2.1. Encoder and Decoder

The encoder adopts four cascaded CBRNBlocks to perform the feature extraction. Each CBRNBlock is composed of three ResidualUnits for extracting features and a Conv1D layer with stride for down-sampling. Same as SoundStream [9], the num-

---
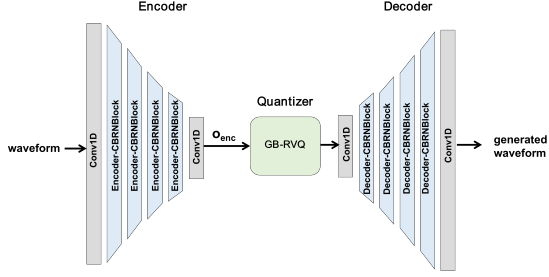
Figure 1: *CBRC model architecture*

Figure 2: *CBRNBlock diagram*

Figure 3: *Group-wise RVQ: The encoded embedding is split into G=2 independent embeddings. They are processed in parallel in quantizer.*
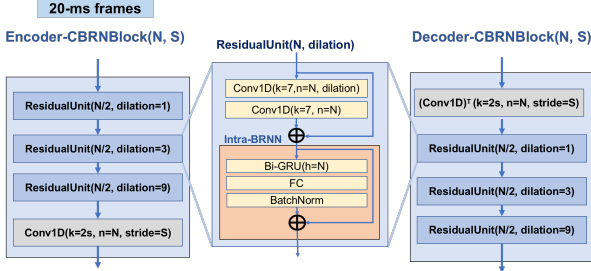
ber of channels is doubled after down-sampling. Supposing the audio waveform $x \in \mathbb{R}^T$ is sampled at f_s with a duration of T, the encoder outputs can be expressed as:

$$\mathbf{o}_{\text{enc}} = \text{Encoder}(\mathbf{x}), \mathbf{o}_{\text{enc}} \in \mathbb{R}^{S \times D} \quad (1)$$
$$S = T/M \quad (2)$$

where D is the encoder embedding dimension, M denotes the sub-frame hopsize and S means the number of embeddings. The decoder mirrors the encoder, using a transposed convolution for up-sampling.

Figure 2 illustrates the architecture of the CBRNBlock. Each ResidualUnit stacks dilated causal 1D-CNN and Intra-BRNN to form an interleaved structure. An internal buffer is inlcuded in each convolution to use information from past frames in inference mode.

Intra-BRNN which consists of Bi-GRU, followed by a linear fully-connected layer and a batchnorm layer, is used in every 20ms frame to capture the intra-frame correlations with no additional delay.

2.2. Quantizer: Group-wise RVQ

The goal of the quantizer is to compress the encoder output $\mathbf{o}_{\text{enc}}$ to a target bitrate R, expressed in bits/second (bps). A codeword is selected from N vectors in Vector Quantizer (VQ) to encode embeddings. Embeddings $\mathbf{o}_{\text{enc}}$ are then mapped to a sequence of one-hot vectors of shape (S, N) , which can be represented using $Slog_2N$ bits. As can be calculated, the codebook size of plain VQ is huge and computation complexity is also very high [9].

Residual Vector Quantizer (RVQ) [9] cascades several VQ layers to reduce the codebook size. Each RVQ layer takes the quantization residual from the previous layer as the layer input. The codebook size of each quantizer can be calculated as with N_q denoting the number of VQ layers:
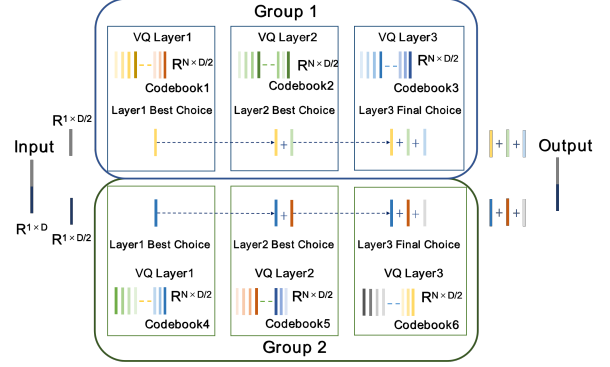
$$N = 2^{R/N_q/S} \quad (3)$$

Figure 3 shows a more promising quantization architecture, namely the Group-wise RVQ. Group-wise RVQ splits the D-dimensional encoder output into G sub-groups (G is set to 2 in Fig. 3 for demonstration). Each sub-group uses a smaller (D/G) codebook size and quantizes the split embedding independently. Afterwards, G parallel optimal codewords from sub-groups are concatenated to form the final quantization result. The number of parameters in RVQ codebook and the computational complexity of RVQ can be expressed as:

$$\text{Number of RVQ parameters} = N_q \times D \times N$$
$$\text{Complexity} = N_q \times (D+1) \times N \times S \quad (4)$$

As can be calculated, the number of RVQ parameter and quantization complexity reduced from 1.57M and 0.158 Macs to 0.79M and 0.079 Macs with Group-wise RVQ (G=2), in which R is set to 6kpbs, D is set to 256 and S equals 100 when 10ms sub-frame hopsize is used within a 1-second audio segment. Although the number of parameters and complexity are halved, Group-wise RVQ outperforms the typical RVQ by VISQOL score 0.11(see Section 3.4).

2.3. Quantizer: Beam-search RVQ

To reduce the quantization error, Beam-search algorithm based RVQ is proposed to select codebook vectors in a larger search space. Algorithm 1 shows Beam-search implementation details. As can be seen in Algorithm 1, Beam-search RVQ selects vectors based on the minimum quantization path error measured by the mean square error (MSE) between the sum of selected vectors and the input embedding. The selection process can be divided into: **1)** Each VQ layer quantizes the potential k candidates leading to k^2 new options, **2)** Top k paths with smaller errors are then selected among the k^2 options as optimal outputs for current VQ layer and **3)**The best path is determined in the last VQ layer.

Figure 4 shows an example when k is set to 2. A larger k in Beam-search RVQ is associated with a larger search space and a smaller quantization error. Beam-search RVQ slightly increases by about $(1 + k(N_q - 1))/N_q$ times with N_q VQ layers and k candidates.

2.4. Loss function

The discriminator of the experiments is the same as the discriminator of SMGAN [12]. The overall function $\mathcal{L}$ can be ex-
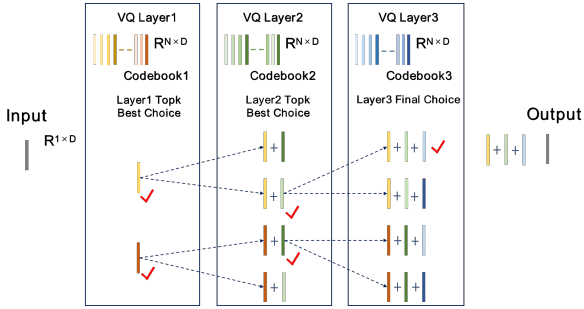
Figure 4: *Beam-search RVQ: The figure shows the process of quantization in three layers of VQ when the number of candidates k = 2. Each layer of VQ except the last reserves 2 quantization paths.*

Algorithm 1: Beam-search RVQ

Input: Embedding X; Number of candidates k; Vector Quantizers Q_i for i = 1 ... N_q, where $[z_1, ..., z_k] = Q_i(residual)$; $Res_{topk} = [r_1, ..., r_k]$ where $r_i = X$; $[s_1, ..., s_k] = find(Candidates, k)$ where pick the k values with the smallest value

Output: Quantized output Y

1 **for** *i= 1 to N_q* **do**
2   $Res\_tmp = [\quad]$;
3   **for** *j = 1 to k* **do**
4     $r = Res_{topk}[j]$;
5     $[z_1, ..., z_k] = Q_i(r)$;
6     $[r_1, ..., r_k] = [r - z_1, ..., r - z_k]$;
7     $Res\_tmp.append([r_1, ..., r_k])$;
8   **end**
9   $Res_{topk} = find(Res\_tmp, k)$
10 **end**
11 $Y = find(Res_{topk}, 1)$

---

pressed as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{adv}\mathcal{L}_{adv} \\ + \lambda_{VQ}\mathcal{L}_{VQ} + \lambda_{pmsqe}\mathcal{L}_{pmsqe} \quad (5)$$

where $\mathcal{L}_{rec}$ is the multi-scale reconstruction loss [9], $\mathcal{L}_{feat}$ is computed in the feature space defined by the discriminators [9], $\mathcal{L}_{adv}$ is the adversarial loss for the generator, $\mathcal{L}_{VQ}$ is the commitment loss [13], which constraints vector quantization and $\mathcal{L}_{pmsqe}$ is the perceptual loss [14], which is designed to be inversely proportional to PESQ [15]. In this work, λ_{feat}, λ_{adv}, λ_{VQ} and λ_{pmsqe} are set to 100, 1, 1, 1, respectively.

3. Experiments

3.1. Datasets and evaluation metrics

CBRC was trained on 245 hours of speech from the LibriTTS dataset [16] at 16 kHz. To ensure that the model is robust with different audio amplitudes, each segment was normalized with a maximum peak value of 0.95 and multiplied by a random gain from 0.3 to 1. The model is trained on A100-SXM-80GB with a batch size of 60. The computational complexity of the CBRC is approximately 4.57G Macs and the total number of model parameters is approximately 4.38M. In this paper, ViSQOL V3 [17] and PESQ [15] are adopted to evaluate the objective quality of audio reconstructed from CBRC. For subjective test, we

---

Table 1: *Objective scores of different codecs.*

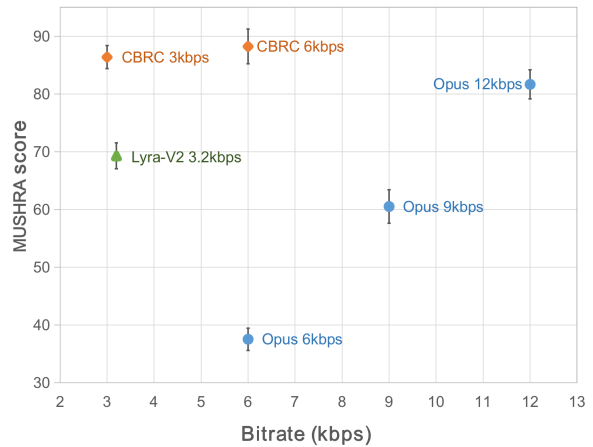| Codec	bitrate	Complexity (Macs)	ViSQOL	PESQ
Opus	6kbps		2.57	2.18
Opus	9kbps		3.51	3.08
Opus	12kbps		3.91	3.81
Lyra-V2	3.2kbps	~344M	3.18	2.34
Lyra-V2	6kbps	~344M	3.55	2.92
Lyra-V2	9.2kbps	~344M	3.70	3.11
CBRC	3kbps	4.5G	**3.71**	**2.88**
CBRC	6kbps	4.5G	**4.04**	**3.57**
CBRC-lite	3kbps	379.3M	**3.56**	**2.67**
CBRC-lite	6kbps	379.3M	**3.84**	**3.09**
CBRC-lite	9kbps	379.3M	**4.01**	**3.22**

Figure 5: *Subjective scores for different codecs. Error bars denote 95% confidence intervals.*

used unseen audio samples with a MUSHRA-inspired crowd-sourced method [18].

3.2. Comparison with other codecs

To evaluate the speech quality of CBRC, different codecs were evaluated on a set of 10 multilingual speech sequences. The subjective and objective scores are shown in Figure 5 and Table 1. As can be seen from the results, our proposed codec at 3kbps outperforms Opus* at 12kbps, despite using a quarter of the bitrate.

We also see that CBRC at 3kbps significantly outperforms Lyra-V2† at 3.2kbps, which demonstrates the superiority of the CBRC architecture. And in order to minimize the influence of complexity, we propose the lite version of CBRC with Intra-BRNN and Beam-search residual quantizer, which slightly increases complexity by 10% but improves PESQ performance from 3.11 to 3.22 at 9kbps. The audio samples‡ are available.

3.3. Ablation of Intra-BRNN

Several ablation experiments were carried out to evaluate the benefit of Intra-BRNN applied in CBRC. All the models were operated at 6kbps. Intra-BRNN can be applied in encoder, de-

---

*https://opus-codec.org
†https://github.com/google/lyra
‡https://bytedance.feishu.cn/docx/OqtjdQNhZoAbNoxMuntcErcInmb

Table 2: *Ablation of Intra-BRNN in encoder and decoder. w/o indicates that the fully convolutional network, and w indicates that the interleaved network of 1D-CNN and Intra-BRNN.*

	Encoder Intra-BRNN	Decoder Intra-BRNN	ViSQOL
1	w/o	w/o	3.41
2	w/o	w	3.55
3	w	w/o	3.88
4	w	w	**4.04**

Table 3: *Comparison of different types of RVQ.*

Quantizer Type	ViSQOL
RVQ	3.81
Group-wise RVQ	3.92
Beam-search RVQ	4.02
GB-RVQ	**4.04**

coder or both. As can be seen in Table 2, the Intra-BRNN significantly improves ViSQOL from 3.41 to 4.04 when adopted in both encoder and decoder. To demonstrate the efficiency of codebook utilization in the proposed RVQ, the codeword frequency is counted in each VQ layer and the potential bitrate is calculated by entropy coding. The bitrates of the fully convolutional model and CBRC are 4.94kbps and 5.13kbps, which means the codebook utilization efficiency increased from 82.3% to 85.5% when target bitrate R is to 6kbps.

3.4. Ablation of Group-wise RVQ and Beam-search RVQ

To demonstrate the effectiveness of the proposed Group-wise RVQ and Beam-search RVQ, different RVQs were evaluated on ViSQOL at 6kbps. Table 3 shows that the ViSQOL score improves from 3.81 to 3.92 and 4.02 when the Group-wise and Beam-search algorithm are used, respectively. The Group-wise RVQ (number of sub-groups G=2) improves the performance and reduces the computational complexity. Beam-search RVQ (number of candidates k=4) slightly increases complexity of the quantization but significantly improves coding quality. With Group-wise and Beam-search algorithm, the performance significantly increased from 3.81 to 4.04 and the computational complexity of quantizer slightly increases from 0.158G Macs to 0.277G Macs, which matters little compared to the whole model structure, complexity of which is 4.57G Macs.

4. Conclusion

In this paper, we propose an end-to-end neural audio codec, namely CBRC. An interleaved network of 1D-CNN and intra-BRNN were designed to fully capture temporal information. Furthermore, RVQ with Group-wise and Beam-search algorithms were developed to make the model perform better in neural codecs. The subjective and objective results show that CBRC achieves the state-of-the-art coding quality. How to further reduce the coding redundancy and the computational complexity for real-time communication will be our future direction.

5. References

[1] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2018.

[2] S. Korse, K. Gupta, and G. Fuchs, "Enhancement of coded speech using a mask-based post-filter," in *ICASSP*. IEEE, 2020, pp. 6764–6768.

[3] S. Korse, N. Pia, K. Gupta, and G. Fuchs, "Postgan: A gan-based post-processor to enhance the quality of coded speech," in *ICASSP*. IEEE, 2022, pp. 831–835.

[4] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "Wavenet based low rate speech coding," in *ICASSP*. IEEE, 2018, pp. 676–680.

[5] W. B. Kleijn, A. Storus, M. Chinen, T. Denton, F. S. Lim, A. Luebs, J. Skoglund, and H. Yeh, "Generative speech coding with predictive variance regularization," in *ICASSP*. IEEE, 2021, pp. 6478–6482.

[6] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP*. IEEE, 2019, pp. 5891–5895.

[7] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with sample rnn," in *ICASSP*. IEEE, 2019, pp. 7155–7159.

[8] A. Mustafa, J. Büthe, S. Korse, K. Gupta, G. Fuchs, and N. Pia, "A streamwise gan vocoder for wideband speech coding at very low bit rate," in *WASPAA*. IEEE, 2021, pp. 66–70.

[9] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[10] X. Jiang, X. Peng, C. Zheng, H. Xue, Y. Zhang, and Y. Lu, "End-to-end neural speech coding for real-time communications," in *ICASSP*. IEEE, 2022, pp. 866–870.

[11] N. Pia, K. Gupta, S. Korse, M. Multrus, and G. Fuchs, "Nesc: Robust neural end-2-end speech coding with gans," *arXiv preprint arXiv:2207.03282*, 2022.

[12] A. Mustafa, N. Pia, and G. Fuchs, "Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *ICASSP*. IEEE, 2021, pp. 6034–6038.

[13] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[15] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[16] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[17] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *QoMEX*. IEEE, 2020, pp. 1–6.

[18] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.