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

Linping Xu∗1, Jiawei Jiang1, Dejun Zhang1, Xianjun Xia1, Li Chen1, Yijian Xiao1, Piao Ding1, Shenyi Song1, Sixing Yin2, Ferdous Sohel3

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

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 Beam-search 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. SoundStream 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 SoundStream is robust under a wide range of real-life coding scenarios.

*This work was done when Linping Xu was an intern at ByteDance.
The number of parameters in RVQ codebook and the computational complexity reduced from 1.57M and 0.158 Macs to 0.79M and 0.079 Macs with Group-wise RVQ (G=2), respectively. Group-wise RVQ: The encoded embedding is split into G = 2 independent embeddings. They are processed in parallel in quantizer.

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:

Number of RVQ parameters = \(N_q \times D \times N\) 
Complexity = \(N_q \times (D + 1) \times N \times S\) 

As can be calculated, the number of RVQ parameters 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 codewords and the input embedding. The selection process can be divided into:

1) Each VQ layer quantizes the potential \(k^q\) candidates leading to \(k^q\) new options, 2) Top \(k\) paths with smaller errors are then selected among the \(k^q\) 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-
Algorithm 1: Beam-search RVQ

Input: Embedding $X$; Number of candidates $k$; Vector Quantizers $Q_{i}$ for $i = 1$ ... $N_{Q}$, where $[z_{1}, ..., z_{k}] = Q_{i}(\text{residual})$; $Res_{topk} = [r_{1}, ..., r_{k}]$ where $r_{i} = X$; $[s_{1}, ..., s_{k}] = \text{find}(\text{Candidates}, k)$ where pick the k values with the smallest value

Output: Quantized output $Y$

1. for $i = 1$ to $N_{Q}$ do
2.   $\text{Res}.tmp[1] = []$
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.     $\text{Res}.tmp$.append($[r_{1}, ..., r_{k}]$)
8.   end
9.   $Res_{topk} = \text{find}(\text{Res}.tmp, k)$
10. end
11. $Y = \text{find}(Res_{topk}, 1)$

pressed as:

$$L = L_{\text{rec}} + \lambda_{\text{feat}} L_{\text{feat}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{pmsqe}} L_{\text{pmsqe}} + \lambda_{\text{feat}} L_{\text{feat}} + \lambda_{\text{adv}} L_{\text{adv}}$$ (5)

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

100. 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.

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

*Note: 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-
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.

<table>
<thead>
<tr>
<th>Encoder Intra-BRNN</th>
<th>Decoder Intra-BRNN</th>
<th>ViSQOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o</td>
<td>w/o</td>
<td>3.41</td>
</tr>
<tr>
<td>w/o</td>
<td>w</td>
<td>3.55</td>
</tr>
<tr>
<td>w</td>
<td>w/o</td>
<td>3.88</td>
</tr>
<tr>
<td>w</td>
<td>w</td>
<td>4.04</td>
</tr>
</tbody>
</table>

Table 3: Comparison of different types of RVQ.

<table>
<thead>
<tr>
<th>Quantizer Type</th>
<th>ViSQOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVQ</td>
<td>3.81</td>
</tr>
<tr>
<td>Group-wise RVQ</td>
<td>3.92</td>
</tr>
<tr>
<td>Beam-search RVQ</td>
<td>4.02</td>
</tr>
<tr>
<td>GB-RVQ</td>
<td>4.04</td>
</tr>
</tbody>
</table>

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 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