MSCACodec: A Low-rate Neural Speech Codec With Multi-scale Residual Channel Attention

Xingye Yu, Ye Li(✉), Peng Zhang, Lingxia Lin, and Tianyu Cai

[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China.
[2] Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan, China.
yxy138646@gmail.com , liye@sdas.org

Abstract. In the development of modern communication technology, although wideband speech coding provides high-fidelity speech transmission, its high bandwidth requirement limits its application in resource-constrained environments. Thus, narrowband speech coding is still of great significance. Recently, end-to-end neural speech coding has made significant progress and demonstrated superior compression performance over traditional methods. However, existing methods are limited in reconstructing details, especially in low birate environments. To address this, we introduce MSCACodec, a narrowband-based neural speech codec that achieves advanced performance at low bitrates. MSCACodec adopts a multi-scale residual and channel attention feature fusion method to selectively focus on multi-scale information to enhance feature representation, solving the problem of inconsistent hierarchical information caused by multi-scale feature fusion. In addition, we also propose a Temporal Convolutional Gated Recurrent Unit (TCGRU) module, which combines temporal convolutional networks and gated recurrent units to enhance the reconstruction quality using global context and gating mechanisms. The experimental results show that, whether in subjective or objective evaluation, MSCACodec achieves higher quality reconstructed speech than Encodec and HiFiCodec at bitrates of 1.2kbps and 2.4kbps, and is even better than LyraV2 and Opus at 6kbps.

Keywords: Speech coding · Multi-scale residual channel attention · Temporal convolutional gated recurrent unit

1 Introduction

Narrowband low-rate speech coding aims to further compress the speech coding rate while maintaining acceptable speech quality to meet the transmission requirements under limited bandwidth resources. Traditional low-rate narrowband speech coding usually adopts parameter coding methods to produce intelligible speech, however, this often sacrifices the naturalness of speech and leads to the

limited quality of synthesized speech. For some application scenarios with high requirements on speech quality, this quality loss is unacceptable. Narrowband-based low-rate speech coding is still a challenging task.

Recently, neural vocoders have achieved remarkable results in the field of narrowband low-rate speech coding. By combining generative models with traditional vocoders, they have shown excellent performance in speech coding tasks. For example, Kleijn combined WaveNet [1] as a decoder with Codec2 [2], and achieved high synthesis quality at 2.4kbps [3]. Lyra [4] encodes the quantized Mel-spectrogram features of speech and then decodes them using WaveGRU, achieving high synthesis quality at 3kbps. CQNV [5] combined HiFiGAN [6] as a decoder with Codec2, and achieved high synthesis quality at 1kbps. With the development of generative models, end-to-end neural speech codecs have been proposed. For example, Gârbacea combined VQVAE [7] with WaveNet, achieving higher reconstruction quality at 1.6kbps [8]. SoundStream [9] proposed a fully convolutional end-to-end universal speech codec, which extended the VQ-VAE vector quantizer to a residual vector quantizer (RVQ), and achieved high-quality synthesized speech at 3kbps. Encodec [10] modified the SoundStream architecture and introduced Transformer [11] to further compress the latent space, achieving good synthesis quality at 1.5kbps. Recently, some works have focused on the problem of coarse quantization. For example, LMCodec [12] trains an AudioLM [13] language model to generate the fine tokens from the coarse, enabling the transmission of a reduced number of codes. HiFiCodec [14] addressed the issue of insufficient codebook utilization in RVQ by combining group vector quantization with residual vector quantization. Language-Codec [15] utilizes a masking mechanism to restrict the quantizer to learn information from specific spatial speech frames, effectively integrating the information in the codebook. GBRVQ [16] applied the Group-wise and Beam-search algorithm to group residual vector quantization, improving quantization efficiency.

In experiments based on neural speech coding, we observe that although improving quantization methods can enhance the quality of reconstructed speech, the inherent limitations of the encoder-decoder architecture still constrain the quality of reconstructed speech. We have analyzed existing encoder-decoder architectures and summarized them into three main drawbacks:

- **Multi-scale information loss**: Speech signals inherently possess multi-scale features, covering speech information in different time and frequency ranges. However, current methods often fail to fully capture these complex multi-scale features, resulting in loss of speech detail information.
- **Hierarchical information inconsistency**. The residual network alleviates the gradient vanishing phenomenon through its skip connection, but this design also introduces the problem of hierarchical information inconsistency. This inconsistency may make it difficult for simple addition operations to fully integrate signals from different levels, thus affecting the model's learning of high-level abstract features.
- **Contextual information loss**. Current neural encoder-decoder architectures typically employ two-layer LSTM to learn contextual features. How-

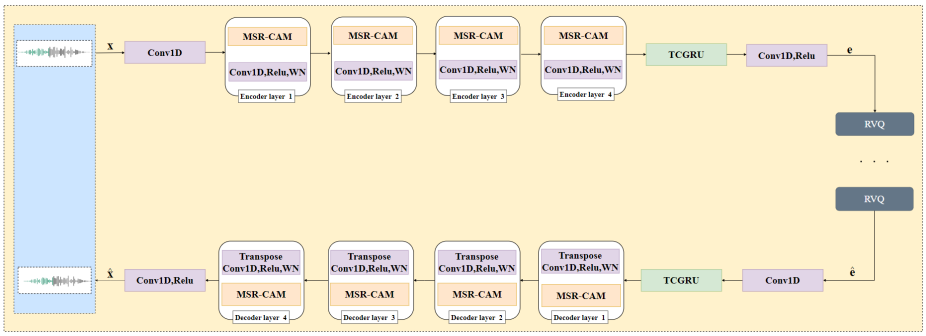

Fig. 1. The architecture of MSCACodec.

ever, since speech is a long sequence problem, the fixed-length memory units of LSTM may cause the gradient to disappear, making it difficult to capture long-range contextual information.

Recently, multi-scale channel attention has achieved significant progress as an effective feature extraction method in the field of speech processing. For instance, TDANET [17] proposed a top-down multi-scale attention encoder-decoder architecture, which significantly improves the ability to capture speech features at different scales. FullNET++ [18] introduced a multi-scale attention architecture for frequency bands, which can effectively capture speech features in different frequency bands, thereby enhancing the ability to model complex speech signals. These studies provide a solid framework for applying multi-scale channel attention in the field of speech coding. To address the above issues, we propose an encoder-decoder architecture called MSCACodec. MSCACodec employs a multi-scale residual and channel attention feature fusion method, selectively focusing on multi-scale information to enhance feature representation, and addressing the hierarchical information inconsistency problem brought by multi-scale feature fusion. In addition, we also propose a Temporal Convolution Gated Recurrent Unit (TCGRU) module, which combines temporal convolutional networks and gated recurrent units to enhance reconstruction quality by leveraging global context and gating mechanisms. We offer the following contributions:

- We propose MSCACodec, a narrowband end-to-end neural speech codec capable of achieving high-quality speech reconstruction at low bitrates.
- We present an effective multi-scale residual channel attention module, which extracts richer and more refined features to represent the input feature maps, while addressing the problem of inconsistent hierarchical information caused by multi-scale residual fusion.
- We propose a TCGRU module that effectively captures both global contextual information and local details.

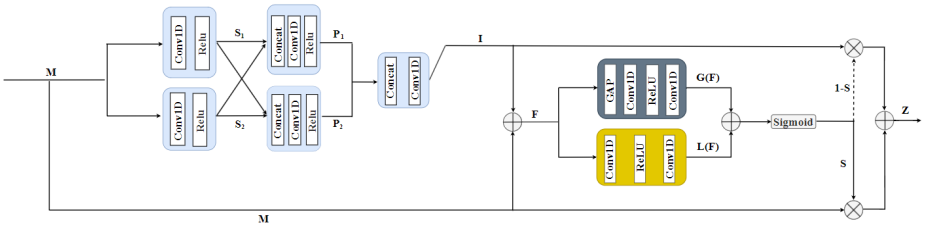

Fig. 2. The network architecture of MSR-CAM.

2 Methods

2.1 Overall Framework

Figure 1 illustrates our architecture. In this architecture, the encoder first maps the input speech features $x \in [-1, 1]^L$ of length L into latent representations $e \in \mathbb{R}^{L_e \times D_e}$, where L_e and D_e denote the length and dimension, respectively. Next, the residual vector quantizer searches the codebook for the codeword that best matches e and transmits the codeword index to the decoder. At the decoder, the residual vector quantizer obtains the dequantized latent speech features $\hat{e}$ according to the index. Finally, the decoder then reconstructs the speech features $\hat{x}$ from $\hat{e}$.

Distinguishing from existing works [9, 10], The encoder mainly consists of 4 multi-scale residual channel attention modules with downsampling and a TCGRU module, and the decoder is symmetric to the encoder. Below, we will introduce the multi-scale residual channel attention module and the TCGRU module in detail.

2.2 Multi-Scale Residual Channel Attention Module

Existing neural codecs typically employ standard residual networks for downsampling and upsampling to reconstruct speech signals. However, this standard residual network architecture cannot guarantee the reconstruction of multi-scale speech details, and the inconsistent hierarchical information caused by residual fusion makes it difficult for the model to learn higher-level features. This paper focuses on multi-scale features, inspired by [19, 20], we propose a multi-scale residual channel attention module(MSR-CAM) to solve the problem of existing neural speech codec.

As shown in Figure 2, this module consists of a multi-scale residual module and a channel attention fusion module, where the multi-scale residual module is responsible for capturing multi-scale feature information, and the channel attention fusion module is responsible for dynamically adjusting multi-scale information to solve the problem of inconsistent hierarchical information caused by multi-scale residual fusion. Specifically, given an intermediate feature $M \in \mathbb{R}^{C_M \times L_M}$ with a channel of C_M and a feature size of L_M. First, the intermediate feature passes through the multi-scale residual module, which is composed

of convolutional kernels of different sizes. The information between these convolution kernels can be shared with each other to learn speech information of different scales, thereby obtaining the fused multi-scale features $F \in \mathbb{R}^{C_F \times L_F}$. The overall process of the multi-scale residual module is defined in Equations (1) to (6).

$$S_1 = \delta(w^1 * M + b^1) \tag{1}$$
$$S_2 = \delta(w^1 * M + b^1) \tag{2}$$
$$P_1 = \delta(w^2 * [S_1, S_2] + b^2) \tag{3}$$
$$P_2 = \delta(w^2 * [S_2, S_1] + b^2) \tag{4}$$
$$I = w^3 * [P_1, P_2] + b^3 \tag{5}$$
$$F = I + M \tag{6}$$

Here, w and b denote the weights and biases, respectively, with superscripts indicating the corresponding layer. The dimension of $w^1 \in \mathbb{R}^{C \times C}$ remains unchanged, and the dimensions of $w^2 \in \mathbb{R}^{2C \times 2C}$ and $w^3 \in \mathbb{R}^{4C \times C}$ are reduced. M, $\delta(\cdot)$ and $[\cdot]$ are respectively represented as the input feature, ReLU activation function, and connection operation. Subsequently, the attention fusion module aggregates the global channel context $G(\cdot)$ and local channel context $L(\cdot)$ of the multi-scale feature F to obtain the attention score S after the fusion of global and local features, and applies it to the skip connection M and multi-scale residual I respectively, perform weighted averaging between the multi-scale residual and the identity mapping, and finally obtain the output feature $Z \in \mathbb{R}^{C_Z \times L_Z}$. The overall process of the channel attention fusion module is defined in Equations (7) to (10).

$$G(F) = B(PWConv_1(\delta(PWConv_1(g(F))))) \tag{7}$$
$$L(F) = B(PWConv_1(\delta(PWConv_1(F)))) \tag{8}$$
$$S = \sigma(G(F) + L(F)) \tag{9}$$
$$Z = M * S + I * (1 - S) \tag{10}$$

where $g(\cdot)$, $\delta(\cdot)$, $B(\cdot)$ and $\sigma(\cdot)$ represent global average pooling, ReLU activation function, batch normalization and Sigmoid activation function, respectively. The number of channels of $PWConv_1$ and $PWConv_2$ are respectively $C \times \frac{C}{r}$ and $\frac{C}{r} \times C$, and r represents the channel reduction ratio.

2.3 Temporal Convolutional Gated Recurrent Unit Module

Existing neural speech codecs usually use a two-layer LSTM to capture global and local information of speech. However, due to the fixed-length memory cell problem, LSTM may not be able to effectively capture features with long-term dependencies and have high complexity. To address this issue, the Temporal Convolutional Networks (TCN) was proposed [21]. As shown in Figure 3(a), TCN mainly consists of input 1D convolutions, depthwise dilated convolutions, and

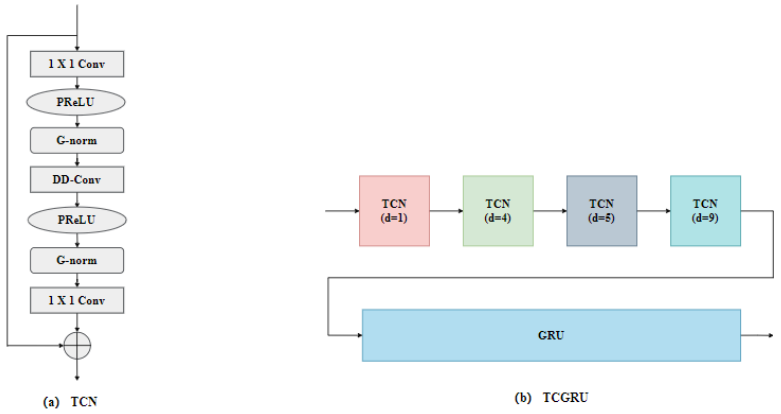

Fig. 3. The network structure of TCGRU, where (a) represents the network structure of TCN and (b) represents the network structure of TCGRU.

output 1D convolutions, and alleviates the problem of gradient vanishing through residual connections. TCN has the advantages of parallel computation and an extended receptive field, and performs well in processing long-term dependencies tasks [18, 22].

To address the issue of long-term dependency loss in the two-layer LSTM, we propose the Temporal Convolutional Gated Recurrent Unit (TCGRU) module to replace the two-layer LSTM. As shown in Figure 3(b), the TCGRU module structure consists of four stacked groups of TCN and GRU units. The dilation factors for the four stacked groups of TCN are 1, 4, 5, and 9 respectively, ensuring a sufficiently large temporal context window to fully utilize the long-range context information of speech signals. We added a more computationally efficient GRU after the stacked TCN and used its gating mechanism to further optimize the integration and transmission of information, ensuring a balance between local and global information.

2.4 Training Loss

We jointly train the model with discriminators for adversarial training to enhance perceptual quality. We use two types of discriminators: the MS-STFT discriminator [10], which attempts to make the spectrogram-level reconstruction similar to the original reconstruction, and the MPD [6], which aims to make the waveform-level reconstruction similar to the original waveform. We use the standard adversarial loss and feature matching loss from [10] to train the MSCA-Codec [3] model.

[3] `https://github.com/GitYesm/MSCACodec`

3 Experimental Setup

3.1 Datasets

We first use the publicly available LibriSpeech ASR corpus [23], which has a sampling rate of 16kHz and contains 982 hours of speech segments. To adapt to narrowband scenarios, we downsampled the speech data to 8kHz. Specifically, the training data comes from 800 hours of speech in the LibriSpeech ASR corpus training set, and the test data comes from 10 hours of speech in the LibriSpeech ASR corpus test set. Additionally, to further validate the generalization ability of the model, we randomly selected 5 hours of speech from the LJ Speech [24] and THCHS-30 [25] datasets for testing.

3.2 Parameter design

In our proposed model, the codebook uses the k-means algorithm with exponential moving average update [26], the decay factor is set to 0.99, the codebook size is 256, the dimension is 512, and 1.2kbps and 2.4kbps corresponding to 3 and 6 vector quantizers respectively. The entire architecture is based on causal 1D convolutions and trained end-to-end. During training, we use the AdamW [27] optimizer and ExponentialLR to optimize the generator and discriminator, respectively, and set the initial learning rate to 1e-4. The model was trained for 100 epochs with 1680 updates per epoch and trained on 8 NVIDIA 3090 GPUs.

3.3 Evaluation Metrics

Evaluation metrics. We evaluate the proposed system using both subjective and objective metrics. To assess the subjective quality of speech, we employed the MUSHRA method. A group of 10 listeners (5 female and 5 male), aged between 23 and 32, evaluated 20 randomly selected utterances from the LibriSpeech ASR corpus test set. As for objective metrics, we conducted an evaluation based on the following metrics:

1) **Perceptual Evaluation of Speech Quality (PESQ)** [28]. PESQ predicts the MOS by comparing the original speech with the target speech. The PESQ score ranges from -0.5 to 4.5, with higher scores indicating better speech quality.
2) **Short-Time Objective Intelligibility (STOI)** [29]. STOI measures the intelligibility of speech signals by comparing the spectral correlation between the original and target speech. The STOI score ranges from 0 to 1, with higher values indicating better speech clarity.
3) **Log-Spectral Distance (LSD)** [30]. LSD is computed by comparing the logarithmic power spectra of the original and target speech. Lower LSD values indicate better speech quality.
4) **Virtual Speech Quality Objective Listener(VISQOL)** [31].VISQOL measures the similarity between the reference and test speech signals using a spectro-temporal measurement method. The VISQOL score ranges from 1 to 5, with higher scores indicating better speech quality.

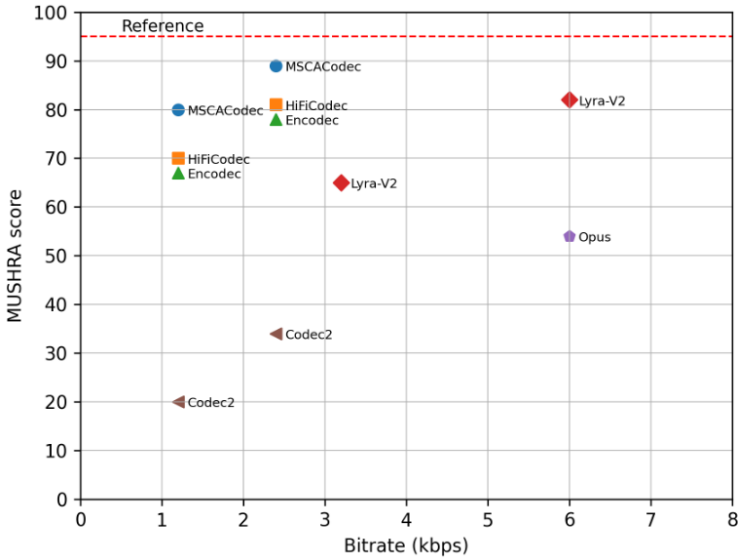

Fig. 4. Comparison of subjective evaluation results between our proposed method and the reference methods.

Baselines. The propose method is compared with the following five baselines:

1) **Codec2** [2]. An open-source traditional low bitrates speech codec designed for low bandwidth applications.
2) **Opus** [32]. A versatile codec widely used for real-time communication, supporting narrowband and wideband.
3) **LyraV2** [3, 4, 9], **Encodec** [10], and **HiFiCodec** [14]. A recently proposed neural speech codec that supports narrowband and wideband, capable of reconstructing high-quality speech

To ensure a fair comparison, we trained Encodec and HiFiCodec on the narrowband LibriSpeech ASR corpus. The codebook size is set to 256, supporting two rates of 1.2 kbps and 2.4 kbps.

4 Results

4.1 Subjective Evaluation

As shown in Figure 4, we compare MSCACodec with the baseline codec at different bitrates. We can see that MSCACodec at 1.2kbps and 2.4kbps is significantly better than Codec2 at the same bitrates, and MSCACodec achieves better performance than Encodec and HiFiCodec. Additionally, MSCACodec achieves better performance than LyraV2 and Opus at 6.0 kbps using only half the bitrate.

4.2 Objective Evaluation

Table 1 shows the objective evaluation results of our proposed method on the test set of the LibriSpeech ASR corpus, where Test-clean and Test-other represent 5 hours of clean and noisy speech segments, respectively. Firstly, we demonstrate that MSCACodec at 1.2 kbps and 2.4 kbps performs the best in terms of PESQ, STOI, and LSD on the clean dataset. Secondly, although there is a performance decline on the noisy dataset, MSCACodec still exhibits superior performance. Table 2 shows the objective evaluation results of our proposed method on the LJ Speech and THCHS-30 . The results show that the performance of MSCACodec on LJ Speech and THCHS-30 datasets slightly degrades, but is still better than the baseline methods. In summary, the objective evaluation results further confirm the subjective evaluation results and consistently highlight the superiority of MSCACodec.

Table 1. Objective testing of the benchmark codecs on the narrowband Test-clean and Test-other datasets.

Bitrate	Model	Test-clean			Test-other		
		PESQ↑	STOI↑	LSD↓	PESQ↑	STOI↑	LSD↓
1.2kbps	Codec2	2.47	0.64	1.15	2.35	0.57	1.18
	Encodec	2.86	0.87	0.89	2.75	0.84	0.91
	HiFiCodec	2.91	0.89	0.85	2.77	0.85	0.91
	MSCACodec(proposed)	**3.07**	**0.92**	**0.81**	**2.95**	**0.90**	**0.86**
2.4kbps	Codec2	2.65	0.69	1.10	2.56	0.64	1.14
	Encodec	3.07	0.91	0.81	2.89	0.89	0.85
	HiFiCodec	3.09	0.91	0.81	2.94	0.90	0.84
	MSCACodec(proposed)	**3.24**	**0.93**	**0.78**	**3.14**	**0.92**	**0.80**
3.2kbps	LyraV2	2.85	0.88	0.91	2.79	0.85	0.94
6kbps	LyraV2	3.19	0.92	0.83	3.01	0.91	0.85

4.3 Ablation Experiment

To evaluate the contribution of the proposed components to the model performance, we set up two groups of ablation experiments. The first group removed the MSR-CAM in MSCACodec and replaced it with the standard residual network in Encodec, denoted as MSCACodec-v1; the second group removed the TC-GRU module in MSCACodec and replaced it with a two-layer LSTM, denoted as MSCACodec-v2. Ablation experiments were performed on the clean-test set

Table 2. Objective testing of the benchmark codecs on the narrowband LJ Speech and THCHS-30 datasets.

Bitrate	Model	LJ Speech			THCHS-30		
		PESQ↑	STOI↑	LSD↓	PESQ↑	STOI↑	LSD↓
1.2kbps	Codec2	2.51	0.65	1.13	2.43	0.61	1.17
	Encodec	2.82	0.86	0.90	2.76	0.85	0.91
	HiFiCodec	2.85	0.86	0.88	2.80	0.86	0.90
	MSCACodec(proposed)	**3.05**	**0.92**	**0.83**	**2.97**	**0.87**	**0.85**
2.4kbps	Codec2	2.73	0.70	1.05	2.63	0.68	1.10
	Encodec	3.02	0.90	0.85	2.97	0.86	0.87
	HiFiCodec	3.06	0.91	0.83	3.02	0.91	0.84
	MSCACodec(proposed)	**3.21**	**0.93**	**0.80**	**3.18**	**0.92**	**0.80**
3.2kbps	LyraV2	2.83	0.86	0.91	2.75	0.85	0.92
6kbps	LyraV2	3.17	0.92	0.82	3.15	0.91	0.85

at a bitrate of 1.2 kbps, with the results presented in Table 3. Experimental results show that the MSR-CAM and TCGRU module in MSCACodec play a key role in the overall performance of the model, especially the multi-scale residual channel attention module.

Table 3. Performance evaluation of ablation experiments on MSCACodec components at 1.2kbps

Model	PESQ↑	STOI↑	LSD↓
MSCACodec -v1	2.93	0.89	0.84
MSCACodec -v2	3.01	0.91	0.83
MSCACodec	**3.07**	**0.92**	**0.81**

4.4 Bandwidth Expansion

To further demonstrate the effectiveness of MSCACodec, we set the codebook size of MSCACodec to 1024, which is consistent with the codebook size of the wideband codec in the benchmark. We train the model on the LibriSpeech ASR dataset with a sampling rate of 16kHz and evaluate it on the test-clean and test-other respectively. Table 4 presents the results, where MSCACodec demonstrates a significantly higher performance compared to the benchmark codec, confirming the superior effectiveness of our method on wideband datasets.

Table 4. Objective testing of the benchmark codecs on the wideband Test-clean and Test-other datasets

Bitrate	Model	Test-clean			Test-other		
		PESQ↑	STOI↑	VISQOL↑	PESQ↑	STOI↑	VISQOL↑
1.5kbps	Encodec	2.59	0.86	3.06	2.44	0.82	2.91
	HiFiCodec	2.77	0.87	3.19	2.56	0.85	3.01
	MSCACodec(proposed)	**3.02**	**0.91**	**3.38**	**2.86**	**0.88**	**3.21**
3kbps	Encodec	2.76	0.87	3.21	2.62	0.86	3.03
	HiFiCodec	2.94	0.90	3.34	2.78	0.87	3.11
	MSCACodec(proposed)	**3.16**	**0.93**	**3.52**	**2.97**	**0.90**	**3.39**
3.2kbps	LyraV2	2.65	0.85	2.98	2.53	0.85	2.81
6kbps	LyraV2	2.98	0.90	3.14	2.84	0.87	3.02

5 Conclusion

This paper introduces an end-to-end neural speech codec that achieves high-quality speech reconstruction at low bitrates. By extending multi-scale residual channel attention to the field of speech coding, we effectively solve the problem of multi-scale information loss and inconsistency of hierarchical information in the feature fusion process. In addition, we propose the TCGRU block, which effectively captures both global contextual and local information in speech. Experimental results further verify the effectiveness of this method. Compared with existing methods, our method shows obvious advantages in reconstruction quality, especially in narrowband and low bit rate conditions.

Acknowledgments. This research was funded by the Taishan Scholars Special Funding (No. tsqn202306253).

Disclosure of Interests. The authors have no competing interests to declare thatare relevant to the content of this article.

References

1. Van Den Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016, 12.
2. Rowe D. Codec 2-open source speech coding at 2400 bits/s and below[C]//TAPR and ARRL 30th Digital Communications Conference. 2011: 80-84.
3. Kleijn W B, Lim F S C, Luebs A, et al. Wavenet based low-rate speech coding[C]//2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 676-680.

4. Kleijn W B, Storus A, Chinen M, et al. Generative speech coding with predictive variance regularization[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6478-6482.
5. Zheng Y, Xiao L, Tu W, et al. CQNV: A combination of coarsely quantized bitstream and neural vocoder for low-rate speech coding[J]. arXiv preprint arXiv:2307.13295, 2023.
6. Kong J, Kim J, Bae J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in neural information processing systems, 2020, 33: 17022-17033.
7. Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.
8. Gârbacea C, van den Oord A, Li Y, et al. Low bit-rate speech coding with VQ-VAE and a WaveNet decoder[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 735-739.
9. Zeghidour N, Luebs A, Omran A, et al. Soundstream: An end-to-end neural audio codec[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 495-507.
10. Défossez A, Copet J, Synnaeve G, et al. High fidelity neural audio compression[J]. arXiv preprint arXiv:2210.13438, 2022.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
12. Jenrungrot T, Chinen M, Kleijn W B, et al. Lmcodec: A low bitrate speech codec with causal transformer models[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.
13. Borsos Z, Marinier R, Vincent D, et al. Audiolm: a language modeling approach to audio generation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.
14. Yang D, Liu S, Huang R, et al. Hifi-codec: Group-residual vector quantization for high fidelity audio codec[J]. arXiv preprint arXiv:2305.02765, 2023.
15. Ji S, Fang M, Jiang Z, et al. Language-Codec: Reducing the Gaps Between Discrete Codec Representation and Speech Language Models[J]. arXiv preprint arXiv:2402.12208, 2024.
16. Xu L, Jiang J, Zhang D, et al. An intra-BRNN and GB-RVQ based end-to-end neural audio codec[J]. arXiv preprint arXiv:2402.01271, 2024.
17. Li K, Yang R, Hu X. An efficient encoder-decoder architecture with top-down attention for speech separation[J]. arXiv preprint arXiv:2209.15200, 2022.
18. Chen J, Wang Z, Tuo D, et al. Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7857-7861.
19. J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale Residual Network for Image Super-Resolution," in Computer Vision – ECCV 2018,Lecture Notes in Computer Science, 2018, pp. 527–542. doi: 10.1007/978-3-030-01237-332.
20. Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional Feature Fusion," in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, Jan. 2021. doi: 10.1109/wacv48630.2021.00360.
21. S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv: Learning,arXiv: Learning, Mar. 2018.

22. Luo Y, Mesgarani N. Tasnet: time-domain audio separation network for real-time, single-channel speech separation[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 696-700.
23. Panayotov V, Chen G, Povey D, et al. Librispeech: an asr corpus based on public domain audio books[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015: 5206-5210.
24. Keith Ito and Linda Johnson. "The LJ Speech Dataset." 2017. Available at: https://keithito.com/LJ-Speech-Dataset/.
25. Wang D, Zhang X. Thchs-30: A free chinese speech corpus[J]. arXiv preprint arXiv:1512.01882, 2015.
26. MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967, 1(14): 281-297.
27. Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.
28. Recommendation I T U T. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs[J]. Rec. ITU-T P. 862, 2001.
29. Taal C H, Hendriks R C, Heusdens R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]//2010 IEEE international conference on acoustics, speech and signal processing. IEEE, 2010: 4214-4217.
30. Erell A, Weintraub M. Estimation using log-spectral-distance criterion for noise-robust speech recognition[C]//International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1990: 853-856.
31. Chinen M, Lim F S C, Skoglund J, et al. ViSQOL v3: An open source production ready objective speech and audio metric[C]//2020 twelfth international conference on quality of multimedia experience (QoMEX). IEEE, 2020: 1-6.
32. Valin J M, Vos K, Terriberry T. Definition of the opus audio codec[R]. 2012.