MSBFCodec: A Low-Bitrate Neural Speech Codec with Multi-Scale Back-Projection Feature Fusion

Tianyu Cai[1,2], Ye Li[1,2,*], Peng Zhang[1,2], Shuxian Ren[1,2], Jingxiang Wang[1,2]
[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
* Corresponding author: Ye Li (liye@sdas.org)

***Abstract*—With the development of communication technology, narrowband speech coding plays an indispensable role in low-bandwidth or resource-constrained environments. Therefore, research on narrowband speech coding is of great significance. Recently, neural speech coding has rapidly developed, and the most advanced methods show superior compression performance compared to traditional methods. Despite significant progress, existing methods still struggle with reconstructing details, especially at low bitrates. In this study, we introduce MSBFCodec, a neural speech codec that achieves state-of-the-art performance in narrowband low bitrate speech coding. MSBFCodec employs a multi-scale back-projection feature fusion method, effectively extracting multi-scale information and addressing the inconsistency in semantic and hierarchical information caused by multi-scale feature fusion. Furthermore, we propose a convolutional feedforward network module designed to perform efficient local modeling and enhance local speech feature representations. Experimental results show that MSBFCodec outperforms mainstream neural speech codecs in terms of reconstructed speech quality at 0.6 kbps and 0.8 kbps, with even higher quality reconstructed speech at 0.8 kbps than LyraV2 at 3.2 kbps.**

***Index Terms*—Speech coding, Neural codec, Multi-scale back-projection, Feature fusion, Convolutional feedforward network**

I. Introduction

Speech coding aims to compress speech signals to the minimum bit rate while maintaining minimal distortion. Opus [1], Codec 2 [2], and MELP [3] are representative traditional speech codecs widely used in communication scenarios. These methods achieve good speech quality and compression efficiency by parameter extraction, codebook compression, and applying psychoacoustic principles. However, at extremely low bitrates, these methods are impacted by quantization errors, making it difficult to meet the requirements for high-quality speech reconstruction.

In recent years, neural vocoders have made significant progress in narrowband low-bitrate speech coding. One approach combines powerful generative models with traditional vocoders to achieve better synthesis quality. For example, Kleijn et al. combined the WaveNet [4] generative model with Codec 2 [2], generating high-quality speech at 2.4 kbps. Valin et al. proposed a neural vocoder based on LPCNet, achieving real-time speech coding at 1.6 kbps [5]. Lyra [6], based on the autoregressive WaveGRU model, synthesizes speech from quantized mel spectrograms, producing high-quality speech at 3 kbps. The CQNV [7] method combines HiFiGAN [8] as a decoder with Codec 2 and achieves high synthesis quality at 1 kbps. Another approach is end-to-end speech codecs, utilizing the VQ-VAE [9] framework and convolution-based encoder-decoder architectures. For example, SoundStream [10] proposed a fully convolutional end-to-end universal speech codec and introduced a residual vector quantizer (RVQ), achieving high-quality synthesized speech at 3 kbps. Encodec [11] modified the SoundStream architecture and introduced a lightweight Transformer [12] in the quantization phase to further compress latent space, achieving good synthesis quality at 1.5 kbps. Recently, some works have focused on the quantization phase. For example, HiFiCodec [13] improved codebook utilization by combining group vector quantization with residual quantization. Language-Codec [14] adopted a masking strategy to guide the quantizer to learn from local speech frames, effectively enhancing representational capabilities. GBRVQ [15] introduced a grouping mechanism and beam search algorithm to improve quantization efficiency. ERVQ [16] optimized both intra- and inter-codebook parameters to enhance quantization performance. However, existing neural end-to-end speech codecs still encounter limitations in faithfully reconstructing the original speech signal, particularly when the bitrate is below 1 kbps. Two significant drawbacks can be summarized:

- **Insufficient multi-scale information extraction:** Standard convolutional structures fail to capture speech details at different scales, leading to the loss of multi-scale detail information.
- **Insufficient local context modeling:** Existing neural speech codecs tend to focus on capturing global structures or long-term dependencies during the feature modeling process, neglecting the modeling of short-term local contexts in speech signals. This limitation restricts the quality of reconstructed speech.

To address these issues, we propose an encoder-decoder architecture called MSBFCodec. MSBFCodec employs a multi-scale back-projection feature fusion method, effectively solv-

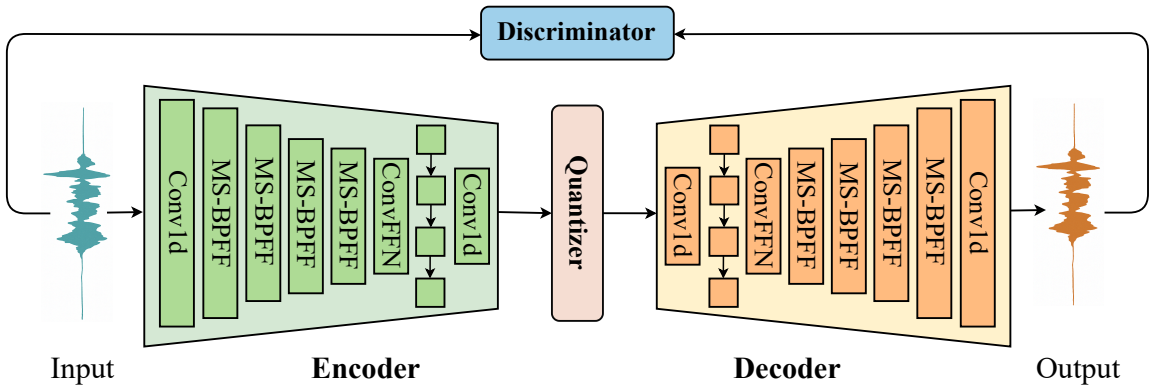

Fig. 1. MSBFCodec framework diagram.

ing the problems of insufficient multi-scale feature extraction and the issue of semantic hierarchy inconsistencies during the fusion process by focusing on multi-scale feature extraction and fusion. Additionally, we introduce a convolutional feed-forward network module to enhance short-term speech details through local context modeling, further improving the quality of the reconstructed speech.

We provide the following contributions:

- We propose MSBFCodec, an end-to-end neural speech codec for narrowband speech, capable of achieving high-quality speech reconstruction at low bitrates.
- We introduce an effective multi-scale back-projection feature fusion module that addresses the issue of semantic hierarchy inconsistency in multi-scale feature fusion while extracting richer multi-scale feature representations.
- We propose a convolutional feedforward network module that effectively captures local context information and speech details, enhancing feature representation.

II. Proposed Method

A. Overall Framework

The framework consists of three components: (1) a feature encoder network that maps the raw speech signal of length L (denoted as $x \in [-1, 1]^L$) into a sequence of latent speech representations (denoted as $e \in \mathbb{R}^{L_e \times D_e}$), where L_e represents the temporal length and D_e is the feature dimension of each frame; (2) the residual vector *quantizer* that quantizes the encoder output by searching for the most matching *codeword* in the *codebook* and transmits the corresponding codeword index to the decoder; and (3) the decoder, which synthesizes the speech signal from the *dequantized* discrete representation.

In our proposed model, both the encoder and decoder apply sequential modeling components to the latent representations. This modeling framework has been shown to perform excellently in various audio-related tasks. Unlike existing works [10], [11], our encoder is primarily composed of four multi-scale back-projection feature fusion modules and a convolutional feedforward network module. Each multi-scale back-projection feature fusion layer is followed by a one-dimensional convolution with downsampling, where the stride is S, and the convolution kernel size is twice the stride S, with the number of channels doubling during the downsampling process. To provide moderate and well-balanced temporal compression between adjacent scales, the values of S are (1, 4, 8, 10). After each downsampling operation, a convolutional feed-forward network is applied to further enhance the locally encoded temporal speech features. The decoder is symmetric to the encoder, as shown in Fig. 1. The encoder outputs 512-dimensional speech features from 8 kHz speech, with a frame rate of 25 Hz.

Regarding the quantizer, we use the residual vector quantizer introduced in [10] to transmit continuous speech features at low bitrates. The entire architecture is trained end-to-end. Below, we provide a detailed explanation of the multi-scale back-projection feature fusion module and the convolutional feedforward network module.

B. Multi-Scale Back-Projection Feature Fusion Module

In existing neural codecs, networks typically use standard convolutional and deconvolutional layers for downsampling and upsampling operations to extract speech features and reconstruct speech signals. However, this standard upsampling and downsampling architecture has limitations in modeling multi-scale contextual features and recovering speech details. On one hand, there is the issue of fine-grained information loss during multi-scale feature extraction; on the other hand, there is a problem of semantic hierarchy inconsistency during multi-scale feature fusion. To address this, this paper focuses on multi-scale feature modeling. Inspired by [17]–[19], we propose a multi-scale back-projection feature fusion (MS-BPFF) module to resolve these issues.

As shown in Fig. 2, the module consists of three components: the multi-scale feature extraction module, the back-projection module (BP), and the selective feature fusion module (SFF). The multi-scale feature extraction module is responsible for capturing rich speech features from different scales, and the back-projection module compensates for the details of the multi-scale features. The selective feature fusion module dynamically integrates features from different scale

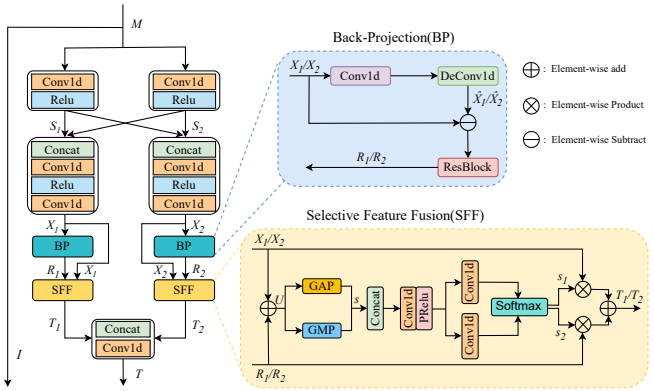

Fig. 2. Multi-scale back-projection feature fusion module.

branches, addressing the issue of semantic inconsistency in multi-scale feature fusion.

Specifically, given an intermediate feature $M \in \mathbb{R}^{C_M \times L_M}$, with channel C_M and feature size L_M, the intermediate feature first passes through two convolutional branches with different kernel sizes to extract multi-scale features S_1 and S_2. Then, the two features are concatenated and sent through separate convolutional layers to enable information cross-sharing and non-linear mapping between the features. Finally, one-dimensional convolutions are applied to the obtained fused features to reduce dimensionality, yielding the final multi-scale feature representations X_1 and X_2. The entire multi-scale feature extraction process is defined by:

$$S_1 = \delta(\omega_1 * M + b_1) \tag{1}$$
$$S_2 = \delta(\omega_1 * M + b_1) \tag{2}$$
$$X_1 = \omega_3 * \delta(\omega_2 * [S_1, S_2] + b_2) + b_3 \tag{3}$$
$$X_2 = \omega_3 * \delta(\omega_2 * [S_2, S_1] + b_2) + b_3 \tag{4}$$

where M, $*$, $\delta(\cdot)$ and $[\cdot]$ represent the input features, convolution operation, ReLU activation function, and concatenation operation, respectively. ω and $\mathbf{b}$ denote the weights and bias parameters of the corresponding convolutional layers, with the subscript indicating the respective layer.

Next, each scale feature X_1 and X_2 is passed through the back-projection module, where a compression-reconstruction operation is applied to enhance the detail representation capability. The specific process is described as:

$$\hat{X}_1 = \omega_5 * (\omega_4 * X_1 + b_4) + b_5 \tag{5}$$
$$\hat{X}_2 = \omega_5 * (\omega_4 * X_2 + b_4) + b_5 \tag{6}$$
$$R_1 = \text{ResBlock}\left(\hat{X}_1 - X_1; \theta^{r1}\right) \tag{7}$$
$$R_2 = \text{ResBlock}\left(\hat{X}_2 - X_2; \theta^{r2}\right) \tag{8}$$

Here, θ^{r1} and θ^{r2} represent all learnable parameters within the ResBlock. Specifically, $\hat{X}_1$ and $\hat{X}_2$ represent the reconstructed outputs obtained by performing one-dimensional convolution and deconvolution operations on X_1 and X_2, respectively. These reconstructed features are then subtracted from the original features to form residuals, which are subsequently passed through a residual block (ResBlock) for feature refinement. This residual block is composed of convolutional layers with a kernel size of 3 and dilation rates of [1, 3], ultimately yielding more accurate residual feature representations, R_1 and R_2.

Next, the selective feature fusion module receives inputs from the two parallel feature branches. Taking R_1 and its corresponding original feature X_1 as an example, element-wise addition is applied to achieve preliminary feature fusion. Subsequently, the fused feature $U \in \mathbb{R}^{D_e \times L_e}$ along the temporal dimension is processed using global average pooling (GAP) and global max pooling (GMP) operations to obtain a statistical representation, also denoted as $s \in \mathbb{R}^{D_e \times 1}$. Then, two parallel convolutional layers followed by a softmax function are applied to generate attention weights s_1 and s_2. These attention weights are then used to perform adaptive weighted summation of the fused features R_1 and X_1. The entire process of feature reweighting and aggregation is defined as:

$$T_1 = s_1 * X_1 + s_2 * R_1 \tag{9}$$

The other feature fusion block follows the same procedure to obtain feature T_2. The two are then fused to obtain the feature T. This fused feature T is combined with the identity mapping feature I from the backbone network via residual connection to produce the final output Z, thereby enhancing information flow and ensuring stable gradient propagation.

C. Convolutional Feedforward Network Module

In neural speech codec tasks, modeling the temporal dependencies of speech features is critical for improving reconstruction quality. Current mainstream approaches often employ architectures such as LSTM or Transformer to enhance the contextual modeling capacity of latent representations. However, these models are computationally intensive and struggle to balance efficiency and performance in low-bitrate or resource-constrained environments. Moreover, they exhibit limited capability in capturing local details in speech signals, particularly when processing short-time utterances or fine-grained features. To address these limitations, this study draws on the dual feedforward network architecture of Conformer [20] and the feedforward network design of Multi-Scale VMamba [21], incorporating depthwise separable convolution to propose a lightweight convolutional feedforward network module (ConvFFN).

As illustrated in Fig. 3, the input feature B first passes through a one-dimensional convolution layer for channel expansion, resulting in an intermediate representation B_1. This is then processed by a depthwise separable convolution to perform local modeling, followed by a residual connection with B_1 to produce an enhanced representation B^*. Subsequently, a GELU activation function and another one-dimensional convolution project the feature back to its original dimensionality.

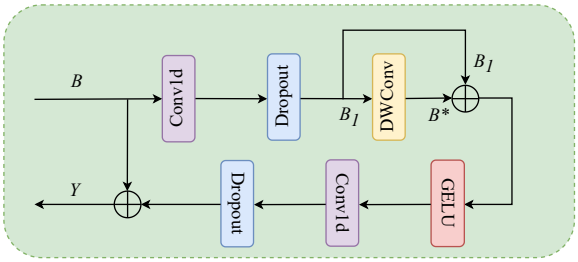

Fig. 3. Architecture of the convolutional feedforward network.

Finally, the output feature sequence Y is obtained via residual connection. The specific process is described as:

$$B_1 = \text{Dropout}(\text{Conv}_{\text{expand}}(B)) \tag{10}$$

$$B^* = B_1 + \text{DWConv}(B_1) \tag{11}$$

$$Y = B + \text{Dropout}(\text{Conv}_{\text{proj}}(\text{GELU}(B^*))) \tag{12}$$

This module effectively enhances the local feature representation capability of speech while maintaining stable gradient propagation.

D. Training Loss

To improve the quality of synthesized speech, we adopt an adversarial loss-based training framework. This adversarial training setup incorporates the MS-STFT discriminator [11], which aims to ensure that the spectral-level reconstruction closely resembles the original, as well as the multi-period discriminator (MPD) [8], which enforces similarity at the waveform level. Following the configuration in [11], the MS-BFCodec model is jointly trained using standard adversarial loss and feature matching loss, ensuring that the generated speech is as close as possible to natural speech.

III. Experimental Setup

In this set of experiments, our objective is to validate the effectiveness of the proposed method under narrowband and low-bitrate conditions. We focus on evaluating the performance of the codec model at two different bitrates.

A. Dataset

LibriSpeech ASR is a large-scale public English speech corpus, widely used for training and evaluating neural speech processing systems, as described in [22]. The original sampling rate of the dataset is 16 kHz, with a total duration of approximately 982 hours. To align with narrowband communication scenarios, all speech data were downsampled to 8 kHz. During training, we used the entire 960 hours of speech segments from the training set. For testing, 10 hours of speech were selected from the standard test set for evaluation. Additionally, to further assess the generalization capability of the model, we randomly selected 5 hours of speech data from two public datasets—LJ Speech [23] and THCHS-30 [24]—to evaluate model performance.

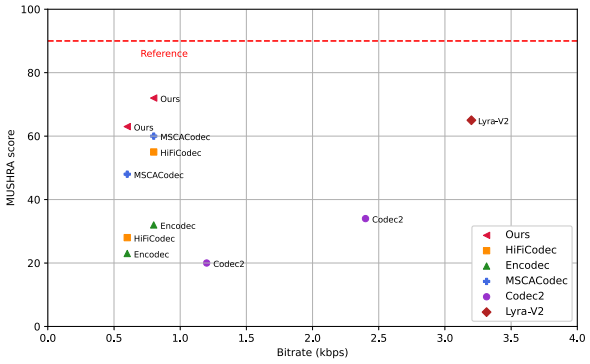

Fig. 4. Comparison of subjective evaluation results between the proposed method and reference methods.

B. Parameter Configuration

In our proposed model, the codebook is updated using an exponentially moving average k-means algorithm [25] with a decay factor set to 0.99. The codebook size is 256, and the dimensionality is 512, corresponding to 3 and 4 vector quantizers for bitrates of 0.6 kbps and 0.8 kbps, respectively. During training, we employed the AdamW optimizer [26] and an exponential learning rate scheduler to optimize the generator and discriminator, with the initial learning rate set to 5×10^{-5}. The model was trained for 200 epochs, with 1780 updates per epoch, on eight NVIDIA 3090 GPUs.

C. Evaluation Metrics

Both subjective and objective evaluations were conducted to assess the performance of MSBFCodec. For subjective evaluation, the MUSHRA method with hidden reference was employed to assess perceptual speech quality. A group of 20 listeners (10 female and 10 male), aged between 22 and 32 years, participated in the listening tests. Twenty utterances were randomly selected from the LibriSpeech ASR test set for evaluation. As for objective metrics, we adopted the perceptual evaluation of speech quality (PESQ) [27], short-time objective intelligibility (STOI) [28], and log-spectral distance (LSD) [29] to measure the objective quality of the proposed method. All data in the experiments were sampled at 8 kHz.

1) Baselines: The proposed method was compared against the following five baseline models:

1) **Codec2** [2]: An open-source traditional low-bitrate speech codec designed for low-bandwidth applications.
2) **LyraV2** [6], [10], **Encodec** [11], **HiFiCodec** [13], and **MSCACodec** [30]: among them, [13] and [30] are recently proposed advanced neural speech codecs that support narrowband speech reconstruction and are capable of producing high-quality audio.

To ensure a fair comparison, LyraV2, Encodec, HiFiCodec, and MSCACodec were all retrained on the LibriSpeech ASR corpus downsampled to 8 kHz. For Encodec, HiFiCodec, and MSCACodec, the codebook size was set to 256, supporting bitrates of 0.6 kbps and 0.8 kbps.

TABLE I
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↓
0.6 kbps	Encodec	2.41	0.82	0.95	2.33	0.79	0.99
	HiFiCodec	2.43	0.83	0.94	2.34	0.80	0.99
	MSCACodec	2.68	0.86	0.90	2.56	0.83	0.96
	Ours	**2.78**	**0.87**	**0.87**	**2.63**	**0.84**	**0.94**
0.8 kbps	Encodec	2.52	0.83	0.93	2.42	0.80	0.98
	HiFiCodec	2.73	0.87	0.88	2.63	0.84	0.93
	MSCACodec	2.81	0.87	0.88	2.67	0.85	0.95
	Ours	**2.90**	**0.89**	**0.86**	**2.72**	**0.85**	**0.92**
1.2 kbps	Codec2	2.47	0.64	1.15	2.35	0.57	1.18
2.4 kbps	Codec2	2.65	0.69	1.10	2.56	0.64	1.14
3.2 kbps	LyraV2	2.85	0.88	0.91	2.79	0.85	0.94

TABLE II
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↓
0.6 kbps	Encodec	2.49	0.84	0.97	2.24	0.78	0.96
	HiFiCodec	2.40	0.84	0.97	2.26	0.78	0.97
	MSCACodec	2.61	0.87	0.95	2.62	0.83	0.91
	Ours	**2.74**	**0.88**	**0.91**	**2.64**	**0.84**	**0.88**
0.8 kbps	Encodec	2.59	0.86	0.95	2.38	0.80	0.94
	HiFiCodec	2.76	0.88	0.89	2.54	0.84	0.88
	MSCACodec	2.69	0.88	0.93	2.76	0.85	0.89
	Ours	**2.85**	**0.90**	**0.88**	**2.80**	**0.86**	**0.85**
1.2 kbps	Codec2	2.51	0.65	1.13	2.43	0.61	1.17
2.4 kbps	Codec2	2.73	0.70	1.05	2.63	0.68	1.10
3.2 kbps	LyraV2	2.83	0.86	0.91	2.75	0.85	0.92

IV. RESULTS

A. Subjective Evaluation

As illustrated in Fig. 4, our proposed codec, MSBFCodec, was compared with baseline codecs under various bitrates. We can see that the performance of MSBFCodec at 0.6 kbps and 0.8 kbps is superior to that of retrained Encodec, HiFiCodec, and MSCACodec at the same bitrates, and significantly better than Codec2 at 1.2 kbps and 2.4 kbps. Moreover, MSBFCodec at 0.8 kbps achieves superior performance compared to LyraV2 while operating at only one-quarter of the bitrate, underscoring the effectiveness of our approach.

B. Objective Evaluation

Tables I and II present the objective evaluation results of the proposed method across multiple public speech datasets and under various bitrate conditions. Table I reports results on the test set of the LibriSpeech ASR corpus, where Test-clean and Test-other refer to 5-hour clean and noisy speech subsets, respectively. Table II includes evaluations on the LJ Speech and THCHS-30 datasets.

As shown in Table I, MSBFCodec achieves the best performance on all three metrics—PESQ, STOI, and LSD—at ultra-low bitrates of 0.6 kbps and 0.8 kbps. The superiority is particularly pronounced on the clean dataset, where MSBFCodec clearly outperforms mainstream methods such as Encodec, HiFiCodec, and MSCACodec. Even under noisy conditions, MSBFCodec maintains its leading performance, demonstrating strong robustness.

Results in Table II indicate a slight performance degradation on the LJ Speech and THCHS-30 datasets; however, MSBFCodec still consistently outperforms the baseline methods. Overall, the objective evaluation results clearly demonstrate the consistent advantages of our approach across diverse speech scenarios, further validating the effectiveness of MSBFCodec at low bitrates.

C. Ablation Study

To evaluate the contributions of the MS-BPFF and ConvFFN modules to the overall performance of our model, we conducted two sets of ablation experiments. In the first configuration, the MS-BPFF module in MSBFCodec was removed and replaced with the standard residual network used in Encodec; this variant is referred to as MSBFCodec-1. In the second configuration, the ConvFFN module was removed from MSBFCodec, denoted as MSBFCodec-2. All experiments were conducted under identical training conditions at a bitrate of 0.8 kbps, using the test-clean subset for evaluation. The results, shown in Table III, indicate that both the MS-BPFF and ConvFFN modules play a critical role in enhancing model performance, confirming their effectiveness in low-bitrate speech coding.

TABLE III
PERFORMANCE EVALUATION OF MSBFCODEC COMPONENT ABLATION EXPERIMENTS AT 0.8 KBPS.

Model	PESQ↑	STOI↑	LSD↓
MSBFCodec-1	2.78	0.86	0.89
MSBFCodec-2	2.83	0.87	0.88
MSBFCodec	**2.90**	**0.89**	**0.86**

D. Bandwidth Expansion

To further demonstrate the effectiveness of MSBFCodec, we set the codebook size of MSBFCodec 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 16 kHz and evaluate it on the Test-clean and Test-other subsets, respectively. Table IV presents the results, where MSBFCodec demonstrates significantly higher performance compared to the benchmark codec, confirming the superior effectiveness of our method on wideband datasets.

V. CONCLUSION

This paper proposes a narrowband end-to-end neural speech codec that achieves state-of-the-art performance at low bitrates. We design an efficient multi-scale back-projection feature fusion module, which effectively extracts rich multi-scale features and addresses the semantic hierarchy inconsistency issue in multi-scale feature fusion. In addition, we extend the ap-

TABLE IV
OBJECTIVE TESTING OF THE BENCHMARK CODECS ON THE WIDEBAND TEST-CLEAN AND TEST-OTHER DATASETS.

Bitrate	Model	Test-clean			Test-other		
		PESQ↑	STOI↑	LSD↓	PESQ↑	STOI↑	LSD↓
1.5 kbps	Encodec	2.59	0.86	0.95	2.45	0.82	0.99
	HiFiCodec	2.77	0.87	0.91	2.56	0.85	0.96
	Ours	**3.04**	**0.91**	**0.82**	**2.85**	**0.88**	**0.89**
3 kbps	Encodec	2.76	0.88	0.92	2.62	0.80	0.95
	HiFiCodec	2.94	0.90	0.84	2.78	0.84	0.91
	Ours	**3.11**	**0.92**	**0.81**	**2.97**	**0.89**	**0.85**
3.2 kbps	LyraV2	2.65	0.86	0.95	2.53	0.85	0.98

plication of convolutional feedforward networks to the speech encoding domain, achieving more efficient local modeling and enhancing the representational capacity of speech features. Experimental results demonstrate that the proposed method consistently outperforms existing mainstream approaches in terms of speech reconstruction quality under narrowband and ultra-low bitrate conditions, highlighting its effectiveness in preserving and reconstructing speech quality.

ACKNOWLEDGMENT

This work was supported by the Taishan Scholars Program of Shandong Province (No. tsqn202306253) and the National Key R&D Program of China (No. 2021QY2312).

REFERENCES

[1] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," Tech. Rep., 2012.

[2] D. Rowe, "Codec 2-open source speech coding at 2400 bits/s and below," in *TAPR and ARRL 30th Digital Communications Conference*, 2011, pp. 80–84.

[3] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "Melp: the new federal standard at 2400 bps," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1591–1594.

[4] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.

[5] J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6 kb/s using lpcnet," *arXiv preprint arXiv:1903.12087*, 2019.

[6] 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 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6478–6482.

[7] Y. Zheng, L. Xiao, W. Tu, Y. Yang, and X. Xu, "Cqnv: A combination of coarsely quantized bitstream and neural vocoder for low rate speech coding," *arXiv preprint arXiv:2307.13295*, 2023.

[8] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[9] C. Gârbacea, A. van den Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with vq-vae and a wavenet decoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 735–739.

[10] 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.

[11] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[13] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[14] S. Ji, M. Fang, Z. Jiang, S. Zheng, Q. Chen, R. Huang, J. Zuo, S. Wang, and Z. Zhao, "Language-codec: Reducing the gaps between discrete codec representation and speech language models," *arXiv preprint arXiv:2402.12208*, 2024.

[15] L. Xu, J. Jiang, D. Zhang, X. Xia, L. Chen, Y. Xiao, P. Ding, S. Song, S. Yin, and F. Sohel, "An intra-brnn and gb-rvq based end-to-end neural audio codec," *arXiv preprint arXiv:2402.01271*, 2024.

[16] R.-C. Zheng, H.-P. Du, X.-H. Jiang, Y. Ai, and Z.-H. Ling, "Ervq: Enhanced residual vector quantization with intra-and-inter-codebook optimization for neural audio codecs," *arXiv preprint arXiv:2410.12359*, 2024.

[17] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," *arXiv preprint arXiv:2209.15200*, 2022.

[18] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[19] Y. Zheng, W. Tu, L. Xiao, and X. Xu, "Supercodec: A neural speech codec with selective back-projection network," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 566–570.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[21] Y. Shi, M. Dong, and C. Xu, "Multi-scale vmamba: Hierarchy in hierarchy visual state space model," *arXiv preprint arXiv:2405.14174*, 2024.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[24] D. Wang and X. Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[25] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5. University of California press, 1967, pp. 281–298.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[27] 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.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[29] A. Erell and M. Weintraub, "Estimation using log-spectral-distance criterion for noise-robust speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990, pp. 853–856.

[30] X. Yu, Y. Li, P. Zhang, L. Lin, and T. Cai, "Mscacodec: A low-rate neural speech codec with multi-scale residual channel attention," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2024, pp. 346–358.