



# FreeCodec: A disentangled neural speech codec with fewer tokens

Youqiang Zheng<sup>1,\*</sup>, Weiping Tu<sup>1,†</sup>, Yueteng Kang<sup>2</sup>, Jie Chen<sup>2</sup>, Yike Zhang<sup>2</sup>, Li Xiao<sup>1</sup>, Yuhong Yang<sup>1</sup>,  
Long Ma<sup>2</sup>

<sup>1</sup>NERCMS, School of Computer Science, Wuhan University, China

<sup>2</sup>Tencent Youtu Lab, China

youqiangzheng@whu.edu.cn, tuweiping@whu.edu.cn, yuetengkang@tencent.com

## Abstract

Neural speech codec is a crucial component in generative tasks such as speech resynthesis and zero-shot TTS. However, most works exhibit degraded performance with fewer tokens due to low coding efficiency in modeling complex coupled information. In this paper, we propose a self-supervised disentangled neural speech codec named FreeCodec. It employs distinct frame-level encoders to decompose intrinsic speech properties into separate components and adopts enhanced decoders to reconstruct speech signals. By encoding and quantizing the different frame-level information with dedicated quantizers, FreeCodec gets higher encoding efficiency with 57 tokens. Furthermore, our proposed method can be applied flexibly in reconstruction and disentanglement scenarios with different training strategies. Subjective and objective experimental results demonstrate that our framework outperforms existing methods in both reconstruction and disentanglement tasks.

**Index Terms:** speech generation, speech conversion, speech codec.

## 1. Introduction

Neural speech codecs are widely used to compress speech signals for a limited number of bits with minimal distortion. Compared to traditional parametric algorithms [1, 2], it has progressed significantly in medium- or low-bitrate scenarios. With the development of large language models (LLM), the discrete codes of neural speech codecs play a pivotal role in LLM-driven generative speech models [3–5]. In general, the fewer tokens, the lower the bitrates while remaining high-quality, which is the goal of neural speech codecs.

The existing mainstream neural speech codecs [6–10] rely on the architecture of VQ-VAE [11]. An encoder, vector quantization layers, and a decoder are learned in end-to-end (E2E) by data-driven. These techniques utilize vector quantization layers to discretize the continuous latent features from the encoder. Recently, many studies have explored disentanglement methods to enhance the quality of reconstructed speech. There are two mainstream methods for disentanglement: 1) The supervised method, and 2) The unsupervised method. The supervised method, e.g. FACodec in NaturalSpeech3 [12], considers disentanglement by the amount of data annotation such as Phone, F0, Speaker labels, etc. Although it is efficient in disentanglement scenarios, it operates at higher bitrates.

The unsupervised method [13–19] is an implicit disentangled technique that usually focuses on using disentanglement to enhance coding efficiency. On the one hand, TiCodec [14]

and SingleCodec [16] incorporate an additional global encoder to extract time-invariant information from speech. The quantization of the extracted global embedding is unnecessary in this way. These methods reduce the redundancy of frame-level information to attain improved encoding efficiency and exhibit improved performance using one or two codebooks at reconstruction scenarios. On the other hand, in [15, 17], self-supervised learning models are employed to factorize semantic and acoustic representations within vector quantization layers, achieving more effective compression compared to conventional neural speech codecs such as EnCodec [7] and Descript-audio-codec (DAC) [20].

The aforementioned methods have demonstrated a strong ability to improve the quality of reconstructed speech using disentanglement techniques. However, these methods has failed to get a balance between reconstruction and disentanglement. Moreover, speech includes several attributes(not just global and non-global), and each of them should be modeled using a module [21]. Inspired by this, we explore a self-supervised disentanglement of representations framework, which can be used flexibly in reconstruction and disentanglement scenarios.

In this paper, we propose a self-supervised disentangled neural speech codec - FreeCodec. It models complex speech into intrinsic attributes(speaker, prosody, and content) in the encoder and disentangles speaker information explicitly. We adopt different frame-level representations for different attributes, enabling more effective quantization and higher compression. In addition, we adopt an improved decoder to improve information reconstruction. Our main contributions are as follows:

- We propose FreeCodec, a self-supervised disentangled neural speech codec that encodes intrinsic properties in speech to disentangle speaker information explicitly and adopts enhanced decoders to get better reconstruction.
- We show that our proposed framework can be flexibly used in reconstruction(e.g., zero-shot TTS, speech compression) and disentanglement(e.g., voice conversion) scenarios when using different training strategies.
- Our proposed method using approximately 57 tokens per second, surpasses the existing state-of-the-art models in subjective and objective evaluation.

## 2. Proposed Method

### 2.1. Overall

As illustrated in Fig.1(a), our proposed method consists of three components: encoders, quantizers, and decoders. Unlike existing works, our encoder proposes a more detailed modeling focus on different intrinsic properties in human speech. We introduce three types of encoders to encode content, speaker, and prosody (in addition to the content and the speaker) information,

\*Work is done during the internship at Tencent YouTu Lab

†Corresponding author

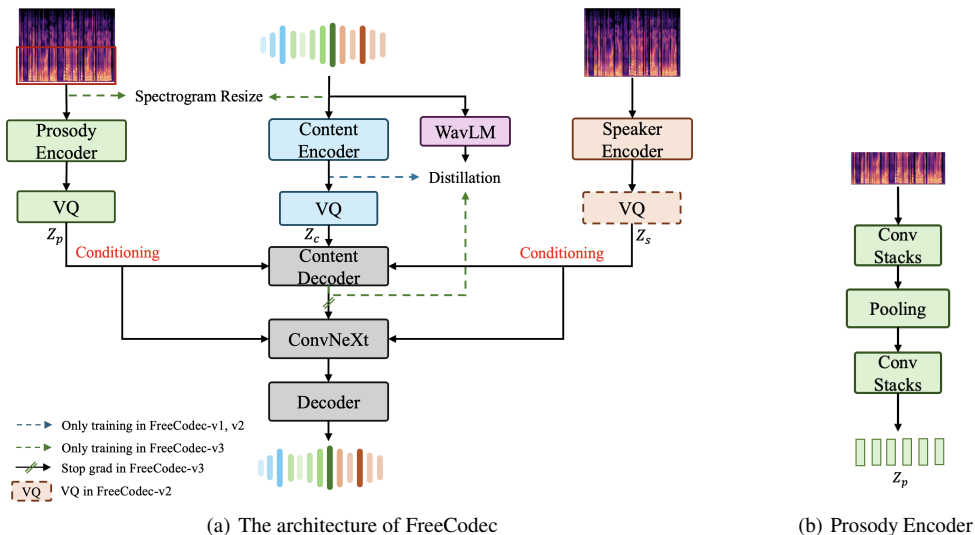


Figure 1: The architecture of FreeCodec and prosody encoder. In Fig.1(a), we show different versions of the FreeCodec training paradigm in different scenarios. In Fig.1(b), the prosody encoder removes the speaker information by a max pooling layer.

respectively. Then the quantization layers produce compressed representations. Finally, the improved decoders, consisting of a content decoder, a backbone module, and an up-sampling decoder, reconstruct the speech signal from the compressed latent representations. In addition, we utilize different training strategies to provide three versions for reconstruction and disentanglement scenarios. The details of training strategies are described in Section 2.3 and 2.5.

## 2.2. Encoders

**Speaker Encoder** Existing approaches assume that a global embedding can represent time-invariant information, such as the characteristics of the speaker and speaking style. Here, we follow this unsupervised manner and further extract the speaker’s information more precisely. We utilize a pre-trained speaker encoder, ECAPA-TDNN [22], a state-of-the-art speaker recognition network based on convolution neural networks with an attentive statistics pooling layer. A mel-spectrogram sampled from the raw speech signal is fed into the speaker encoder to get one global timbre vector.

**Content Encoder** The architecture of the content encoder follows SuperCodec [9] encoder using (2, 4, 5, 8) as strides, a number  $B_{enc} = 4$  of convolution blocks. It indicates a total down-sampling of 320 times and outputs 256-dimensional content features with a frame rate of 50 Hz from 16 kHz speech. In order to reduce the redundancy of the content encoder, we use a self-supervised model to explicitly model the content information, as shown in Fig.1(a).

**Prosody Encoder** The prosody encoder extracts the information apart from the speaker and content information, as shown in Fig.1(b). In [23, 24], the first 20 bins in each mel-spectrogram frame are taken as input to extract prosody because it contains almost complete prosody and much less speaker and content information than the full band. Following the related work [24], we adopt the prosody encoder consisting of two convolution stacks, a max pooling layer with a stride of 8 to remove content and speaker information further. Our proposed method sets the FFT and hop size to 1024 and 320. With these setups, the prosody encoder results in roughly a frame rate of 7 Hz feature embeddings with 256 dimensions.

## 2.3. Quantization

We adopt different methods to quantize different features. For the content and prosody information, we adopt a plain vector quantizer with one codebook, and the codebook size is set to 256. As for the speaker embedding, we use two types: continuous representation for FreeCodec-v1 and FreeCodec-v3 and discrete representation for FreeCodec-v2. Specifically, we compress the speaker embedding by group vector quantization (GVQ) in FreeCodec-v2 for speech coding. It divides the speaker embedding into eight groups that are quantized by one codebook with 1024 codebook size respectively. As for FreeCodec-v1 and FreeCodec-v3, we provide the continuous representation to the decoder for better reconstruction in such as zero-shot TTS and voice conversion scenarios, similar to [12, 25].

## 2.4. Improved Decoders

FreeCodec does not merely rely on a mirrored upsampling decoder. Prior to upsampling, we first employ a content decoder consisting of a 4-layer Transformer encoder to enhance semantic modeling. Then, we use ConvNeXt [26] as a fundamental backbone to condition the prosody and speaker representations further. Finally, a mirrored decoder upsampling structure is employed to reconstruct speech signals. It uses (8, 5, 4, 2) as strides, resulting in a total upsampling of 320 times.

## 2.5. Training Strategy

We incorporate adversarial training to promote perceptual quality, using a multi-scale STFT-based (MS-STFT) discriminator. The training loss of the proposed method comprises five components: reconstruction loss  $\lambda_{rec}$ , VQ commitment loss  $\lambda_{vq}$ , content loss  $\lambda_c$ , feature matching loss  $\lambda_{feat}$ , and adversarial losses  $\lambda_{adv}$ . The reconstruction loss, feature loss, and adversarial losses follow EnCodec [7]. We extract the last layer representation from a pre-trained WavLM-Large model [27] as the semantic learning target and use cosine similarity loss as the content loss.

In FreeCodec-v1 and FreeCodec-v2, we use it to reduce the redundancy of the content encoder. It maximizes the cosine similarity at the level of the dimensions across all timesteps between the outputs of the content encoder and semantic learning target. In FreeCodec-v3, we only use the semantic learning target

at the decoder to prevent additional speaker information from leaking to the content encoder and quantizer. We also utilize spectrogram-resize based data augmentation on the prosody and content encoder in the training [28]. This approach achieves better performance in disentanglement scenarios.

Overall, the generator loss is defined as,

$$\mathcal{L}_G = \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{feat} \cdot \mathcal{L}_{feat} + \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{vq} \cdot \mathcal{L}_{vq} + \lambda_c \cdot \mathcal{L}_c, \quad (1)$$

where these hyper-parameters  $\lambda_{adv} = 3$ ,  $\lambda_{feat} = 3$ ,  $\lambda_{rec} = 1$ ,  $\lambda_{vq} = 1$ , and  $\lambda_c = 10$ .

### 3. Experimental Setup

#### 3.1. Training Details and Baselines

We trained our model on LibriSpeech [29], which consists of approximately 1000 hours of speech at 16kHz. We use train-clean-100, train-clean-360, and train-other-500 subsets for training. For a fair comparison, we adopt two recent neural codecs, TiCodec and Descript-audio-codec (DAC), which have demonstrated success in the domain of neural speech codecs. The baselines are re-trained with 1 and 2 codebooks, indicating 0.5 kbps and 1 kbps at 16 kHz sampling rate. FreeCodec and TiCodec are trained on two V100 GPUs with 400 k iterations and a batchsize of 20 per GPU. DAC is trained on two V100 GPUs with 800 k iterations and a batchsize of 10 per GPU. In addition, we also consider several open-source speech codecs as baselines, Spechtokenizer at 3 kbps, and FACodec without acoustic details at 2.4 kbps, and SemantiCode at 1.3 kbps, and Wavtokenizer-small at 0.9 kbps. For Wavtokenizer-small, we use the 24 kHz pre-trained model to synthesize speech, corresponding to the same compression rate of 0.6 kbps for the 16 kHz sampling rate.

As for the voice conversion, we include disentangled codecs TiCodec at 0.5 kbps and 1 kbps, and FACodec without detail tokens at 2.4 kbps. In addition, three baseline models are selected to be compared with FreeCodec-v3: two text-based models—VQMIVC [30] and YourTTS [31], a self-supervised learning model—Wav2vec-vc [32], which are trained on the VCTK datasets.

#### 3.2. Evaluation

We evaluate FreeCodec from two aspects: **1) Reconstruction Quality.** We conduct it on VCTK [33] and test-clean subset of LibriSpeech. For VCTK, we randomly select data from 8 speakers and 2911 utterances for the test. For LibriSpeech, we use the test-clean subset, 2620 utterances for the test. All audio samples are downsampled to 16 kHz. **2) Disentanglement Ability.** we evaluate it based on the any-to-any voice conversion benchmark. We randomly select 200 utterances from LibriSpeech test-clean subset as source speech and 6 speakers from VCTK as the target speaker. All models are evaluated in LibriSpeech Test-clean to VCTK scenarios.

**Subjective Evaluation.** We follow the established MUSHRA methodology [34] to evaluate the subjective quality of our baselines and FreeCodec-v2. A group of fifteen listeners participate in the subjective tests. Sixteen utterances are randomly selected from our test sets for evaluation. In addition, we also adopt the Speex [35] at 4 kbps as our low anchor.

**Objective Evaluation.** For objective evaluation of reconstruction, we employ the automatic Mean Opinion Score prediction system (UTMOS) [36], and the short-time objective intelligibility (STOI) [37], and the WARP-Q [38], and the Speaker Embed-

ding Cosine Similarity (SECS)<sup>1</sup> to evaluate the overall speech quality. In addition, we use Word error rate (WER), character error rate (CER), and F0-PCC for the objective evaluation of voice conversion. Among them, WER and CER between source and converted speech are calculated by an ASR model<sup>2</sup>. F0-PCC is the Pearson correlation coefficient used to evaluate  $f_0$  consistency between source and converted speech. It can be used to evaluate the preservation of prosody information apart from speaker and content information.

## 4. Results

### 4.1. Reconstruction Quality

Table 1 summarizes the results of objective reconstruction experiments. FreeCodec-v1 performs best or second-best in almost all objective metrics in test sets. Especially in out-of-domain environments, our proposed method achieves superior reconstruction performance using only approximately 57 tokens per second than existing methods, such as Wavtokenizer at 0.9 kbps, and SemantiCodec at 1.3 kbps. Even compared to FACodec at 2.4 kbps and Spechtokenizer at 3 kbps, FreeCodec-v1 gets comparable performance at STOI and speaker similarity. Compared to FreeCodec-v2, FreeCodec-v1 is better especially in speaker similarity. It shows that the continuous global representation is more effective in reconstruction scenarios(e.g., zero-shot TTS).

Although STOI and SECS are slightly lower than DAC at 1 kbps, FreeCodec-v2 has better objective speech quality, according to UTMOS and WARP-Q. The same result can also be concluded in subjective evaluation, as illustrated in Table 2. As illustrated in subjective results, we can observe that FreeCodec-v2 at 0.45 kbps significantly outperforms both FACodec at 2.4 kbps and Spechtokenizer at 3 kbps. Additionally, compared to TiCodec at 1 kbps and DAC at 1 kbps, FreeCodec-v2 gets higher scores with the same experimental configuration. The results of the subjective evaluation show the absolute advantages of FreeCodec-v2 in reconstruction compared to existing methods.

Furthermore, as shown in Table 1, we also conducted an ablation study to validate the explicit effect of content loss in the content encoder. It can be observed that removing the content loss causes the performance drop in all objective metrics, especially the UTMOS and STOI.

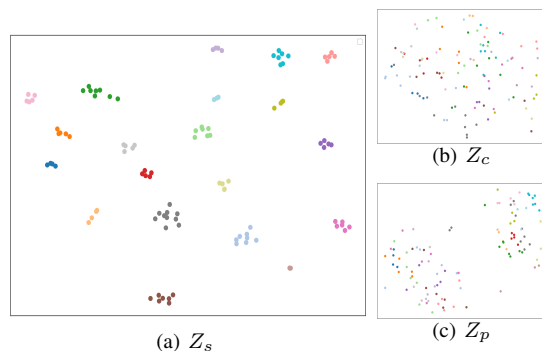


Figure 2: (a)(b)(c) is the  $t$ -SNE visualization of speaker representations  $Z_s$ , content representations  $Z_c$  and prosody representations  $Z_p$  after VQ. Different colors represent different speakers.

<sup>1</sup><https://github.com/resemble-ai/Resemblyzer>

<sup>2</sup><https://huggingface.co/openai/whisper-large>

Table 1: The objective evaluation of reconstruction quality based on VCTK and LibriSpeech Test-Clean corpus.  $\blacklozenge$  represents the reproduced model following the official implementation.  $\clubsuit$  means the results inferred from the official checkpoints. **Bold** is the best result and underline is the second-best result.

Model	Sampling rate	Bandwidth	Token/s	VCTK				Test-clean			
				UTMOS $\uparrow$	STOI $\uparrow$	WARP-Q $\downarrow$	SECS $\uparrow$	UTMOS $\uparrow$	STOI $\uparrow$	WARP-Q $\downarrow$	SECS $\uparrow$
Target	-	-	-	4.085	-	-	-	4.086	-	-	-
SpeechTokenizer $\clubsuit$	16 kHz	3 kbps	300	3.953	0.868	2.048	0.915	3.848	<b>0.908</b>	<b>2.034</b>	<b>0.954</b>
FACodec $\clubsuit$	16 kHz	2.4 kbps	240	<b>4.085</b>	0.855	2.026	<b>0.929</b>	3.616	0.876	2.170	0.943
SemantiCodec $\clubsuit$	16 kHz	1.3 kbps	100	3.334	0.853	2.078	0.868	2.922	0.879	<u>2.049</u>	0.936
TiCodec $\blacklozenge$	16 kHz	1 kbps	100	3.584	0.879	2.333	0.856	3.616	0.881	2.354	0.908
DAC $\blacklozenge$	16 kHz	1 kbps	100	3.780	0.904	2.251	0.883	3.790	<u>0.901</u>	2.274	0.920
WavTokenizer $\clubsuit$	24 kHz	0.9 kbps	75	3.296	0.832	2.192	0.811	3.792	0.897	2.135	0.904
TiCodec $\blacklozenge$	16 kHz	0.5 kbps	50	3.421	0.825	2.578	0.797	3.307	0.821	2.614	0.855
DAC $\blacklozenge$	16 kHz	0.5 kbps	50	3.476	0.852	2.550	0.804	3.543	0.859	2.504	0.883
FreeCodec-v1	16 kHz	0.45 kbps	57	<u>4.034</u>	<b>0.918</b>	<b>1.966</b>	<u>0.919</u>	<b>4.085</b>	0.892	2.195	0.944
w/o. $L_{content}$	16 kHz	0.45 kbps	57	3.805	<u>0.908</u>	<u>1.994</u>	0.893	3.631	0.869	2.308	0.925
FreeCodec-v2	16 kHz	0.45 kbps	57	3.921	0.900	2.190	0.846	<u>3.984</u>	0.893	2.230	0.896
w/o. $L_{content}$	16 kHz	0.45 kbps	57	3.578	0.892	2.175	0.848	3.571	0.885	2.223	0.904

Table 2: The Subjective evaluation of reconstruction quality under unseen speakers from VCTK and LibriSpeech Test-Clean corpus with the 95% confidence interval for each score.

Method	Bitrate	Supervised Speaker		MUSHRA Score
		Data	Decouple	
Target	-	-	-	95.09 $\pm$ 0.44
Speex $\clubsuit$	4kbps	No	No	24.03 $\pm$ 2.15
SpeechTokenizer $\clubsuit$	3 kbps	No	No	82.0 $\pm$ 1.44
FACodec $\clubsuit$	2.4 kbps	Yes	Explicit	80.44 $\pm$ 1.29
SemantiCodec $\clubsuit$	1.3 kbps	No	No	73.44 $\pm$ 1.3
TiCodec $\blacklozenge$	1 kbps	No	Implicit	83.0 $\pm$ 1.10
DAC $\blacklozenge$	1 kbps	No	No	<u>85.5<math>\pm</math>1.19</u>
Wavtokenizer $\clubsuit$	0.9kbps	No	No	78.56 $\pm$ 2.54
TiCodec $\blacklozenge$	0.5 kbps	No	Implicit	73.06 $\pm$ 2.50
DAC $\blacklozenge$	0.5 kbps	No	No	76.56 $\pm$ 2.35
FreeCodec-v2	0.45 kbps	No	Explicit	<b>87.44<math>\pm</math>0.88</b>

## 4.2. Disentanglement ability

In this section, we describe the disentanglement ability on the voice conversion experiments. FreeCodec-v3 achieves voice conversion by using the speaker information from the target speech. As shown in Table 3, FreeCodec-v3 at 0.45 kbps achieves the best speaker similarity in unseen speaker scenarios than all baseline models, especially supervised FACodec at 2.4 kbps and text-based models. The results show that our proposed method decouples the speaker information well under unseen-speaker scenarios in a self-supervised manner.

Compared to TiCodec at 0.5 kbps and 1 kbps, FreeCodec-v3 exhibits lower WER and higher speaker similarity. This indicates that our method preserves content information at ultra-low bitrates while achieving superior disentanglement compared to methods based solely on implicit bottlenecks. Additionally, FreeCodec-v3 shows comparable performance in F0 PCC, suggesting that our method effectively maintains the prosody of the source speech, achieving a better balance between prosody and speaker information.

We randomly selected 20 unseen speakers from the test sets and generated t-SNE visualizations to examine the distributions of speaker, content, and prosody representations. For content and prosody features, we collapsed the frame-level representations of each sample into a single vector through mean

Table 3: The objective evaluation of disentanglement ability. For WER and CER, the smaller the better. F0-PCC ranges from -1 to 1 and the higher the better.

Method	Bandwidth $\downarrow$	Test-Clean to VCTK			
		WER $\downarrow$	CER $\downarrow$	F0 PCC $\uparrow$	SECS $\uparrow$
FreeCodec-v3	<b>0.45 kbps</b>	<u>8.37</u>	6.14	0.702	<b>0.847</b>
TiCodec $\blacklozenge$	0.5 kbps	35.74	25.19	0.680	0.656
TiCodec $\blacklozenge$	1 kbps	8.83	6.08	<u>0.752</u>	0.607
FACodec $\clubsuit$	2.4 kbps	<b>2.83</b>	<b>2.57</b>	<b>0.755</b>	0.553
YourTTS $\clubsuit$	-	9.20	6.92	0.682	0.815
Wav2vec-vc $\clubsuit$	-	13.23	9.20	-0.037	0.826
VQMIVC $\clubsuit$	-	56.58	39.21	0.611	0.650

pooling. As shown in Fig. 2, the speaker embeddings exhibit clear clustering patterns corresponding to individual speakers. In contrast, content representations appear distributed without discernible patterns across different speakers, while prosody features demonstrate moderate clustering tendencies for certain speakers. These observations collectively indicate that prosodic information maintains partial speaker-specific characteristics while remaining distinct from both speaker and content.

## 5. Conclusion

In this paper, we propose a self-supervised disentanglement speech codec that factorizes speech into its intrinsic attributes. We demonstrate that this framework can be applied to both reconstruction and disentanglement tasks using different training strategies. Compared to existing methods, our approach utilizes fewer tokens and lower bandwidth while achieving high-quality reconstruction and superior disentanglement relative to supervised methods. Our experiments show a significant improvement over existing methods that use more than 2x bitrate, highlighting the effectiveness of our approach in reconstruction quality and disentanglement ability.

## 6. Acknowledgement

This paper is supported by the National Nature Science Foundation of China (No. 62471343).

## 7. References

- [1] 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.
- [2] L. M. Supplee, R. P. Cohn *et al.*, “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.
- [3] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.
- [4] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov *et al.*, “Audiolm: a language modeling approach to audio generation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.
- [5] A. Défossez, L. Mazaré *et al.*, “Moshi: a speech-text foundation model for real-time dialogue,” *arXiv preprint arXiv:2410.00037*, 2024.
- [6] N. Zeghidour, A. Luebs *et al.*, “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2022.
- [7] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *arXiv preprint arXiv:2210.13438*, 2022.
- [8] X. Jiang, X. Peng, H. Xue, Y. Zhang, and Y. Lu, “Latent-domain predictive neural speech coding,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2111–2123, 2023.
- [9] Y. Zheng, W. Tu *et al.*, “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.
- [10] S. Ji, Z. Jiang *et al.*, “Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=yB1VIS2Fd9>
- [11] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6309–6318.
- [12] Z. Ju, Y. Wang *et al.*, “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: <https://openreview.net/forum?id=dVhrnjZJad>
- [13] A. Polyak, Y. Adi *et al.*, “Speech Resynthesis from Discrete Disentangled Self-Supervised Representations,” in *Proc. Interspeech 2021*, 2021, pp. 3615–3619.
- [14] Y. Ren, T. Wang, J. Yi *et al.*, “Fewer-token neural speech codec with time-invariant codes,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 737–12 741.
- [15] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Spechtokenizer: Unified speech tokenizer for speech language models,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [16] H. Li, L. Xue *et al.*, “Single-codec: Single-codebook speech codec towards high-performance speech generation,” in *Interspeech 2024*, 2024, pp. 3390–3394.
- [17] H. Liu, X. Xu *et al.*, “Semanticcodec: An ultra low bitrate semantic audio codec for general sound,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 8, pp. 1448–1461, 2024.
- [18] H.-H. Guo, K. Liu *et al.*, “Firereditts: A foundation text-to-speech framework for industry-level generative speech applications,” *arXiv preprint arXiv:2409.03283*, 2024.
- [19] Y. Guo *et al.*, “Lscodec: Low-bitrate and speaker-decoupled discrete speech codec,” *arXiv preprint arXiv:2410.15764*, 2024.
- [20] R. Kumar, P. Seetharaman *et al.*, “High-fidelity audio compression with improved rvqgan,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [21] Z. Jiang, Y. Ren *et al.*, “Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias,” *arXiv preprint arXiv:2306.03509*, 2023.
- [22] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tddn: Emphasized channel attention, propagation and aggregation in tddn based speaker verification,” in *Interspeech 2020*, 2020, pp. 3830–3834.
- [23] Y. Ren, M. Lei, Z. Huang *et al.*, “Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7577–7581.
- [24] Z. Jiang, J. Liu, Y. Ren *et al.*, “Mega-TTS 2: Boosting prompting mechanisms for zero-shot speech synthesis,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=mvMI3N4AvD>
- [25] Y. Pan, L. Ma, and J. Zhao, “Promptcodec: High-fidelity neural speech codec using disentangled representation learning based adaptive feature-aware prompt encoders,” *arXiv preprint arXiv:2404.02702*, 2024.
- [26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.
- [27] S. Chen, C. Wang *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [28] J. Li, W. Tu, and L. Xiao, “Freevc: Towards high-quality text-free one-shot voice conversion,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [29] 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.
- [30] D. Wang, L. Deng *et al.*, “Vqmvic: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion,” in *Interspeech 2021*, 2021, pp. 1344–1348.
- [31] E. Casanova, J. Weber *et al.*, “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.
- [32] J. Lim and K. Kim, “Wav2vec-vc: Voice conversion via hidden representations of wav2vec 2.0,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 326–10 330.
- [33] J. Yamagishi, C. Veaux *et al.*, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” 2019.
- [34] R. ITU-R, “1534-1, method for the subjective assessment of intermediate quality levels of coding systems (mushra),” *International Telecommunication Union*, 2003.
- [35] J.-M. Valin, “Speex: A free codec for free speech,” *arXiv preprint arXiv:1602.08668*, 2016.
- [36] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “Utmos: Utokyo-sarulab system for voicemos challenge 2022,” in *Interspeech 2022*, 2022, pp. 4521–4525.
- [37] C. H. Taal, Hendriks *et al.*, “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.
- [38] W. A. Jassim, J. Skoglund *et al.*, “Warp-q: Quality prediction for generative neural speech codecs,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 401–405.