



HAL
open science

Learning Source Disentanglement in Neural Audio Codec

Xiaoyu Bie, Xubo Liu, Gaël Richard

► **To cite this version:**

Xiaoyu Bie, Xubo Liu, Gaël Richard. Learning Source Disentanglement in Neural Audio Codec. 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, Apr 2025, Hyderabad, India. hal-04902131

HAL Id: hal-04902131

<https://hal.science/hal-04902131v1>

Submitted on 20 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning Source Disentanglement in Neural Audio Codec

Xiaoyu Bie*, Xubo Liu[†], Gaël Richard*

*LTCI, Télécom Paris, Institut Polytechnique de Paris, France

[†]CVSSP, University of Surrey, UK

Abstract—Neural audio codecs have significantly advanced audio compression by efficiently converting continuous audio signals into discrete tokens. These codecs preserve high-quality sound and enable sophisticated sound generation through generative models trained on these tokens. However, existing neural codec models are typically trained on large, undifferentiated audio datasets, neglecting the essential discrepancies between sound domains like speech, music, and environmental sound effects. This oversight complicates data modeling and poses additional challenges to the controllability of sound generation. To tackle these issues, we introduce the Source-Disentangled Neural Audio Codec (SD-Codec), a novel approach that combines audio coding and source separation. By jointly learning audio resynthesis and separation, SD-Codec explicitly assigns audio signals from different domains to distinct codebooks, sets of discrete representations. Experimental results indicate that SD-Codec not only maintains competitive resynthesis quality but also, supported by the separation results, demonstrates successful disentanglement of different sources in the latent space, thereby enhancing interpretability in audio codec and providing potential finer control over the audio generation process.

Index Terms—neural audio codec, source separation, representation learning, quantization.

I. INTRODUCTION

Audio codecs have long played a crucial role in audio data storage and transmission. Traditional methods [1], [2] rely on handcraft audio features and signal coding tools. These approaches compress the raw waveform into codec representations and then decompress it back to the original signal, leading to high-quality reconstruction at an adequate bitrate. Training neural networks to transform signals in codecs is known as neural audio codec (NAC). A typical pipeline consists of an encoder and a decoder for signal transformation, as well as a quantizer between them for information coding. Recent advances in NAC by using residual vector quantization (RVQ) [3]–[5] have achieved higher fidelity and lower bitrates compared to traditional methods. Moreover, the ability to reconstruct high quality audio from discrete latent codes has opened the door for realistic audio generation when combined with language models [6]–[9] or diffusion models [10].

Despite the impressive performance of current NAC models, they are typically trained on a mixture of different audio datasets, compressing audio signals from diverse domains into a single latent space. This unified model neglects the fundamental differences between audio from various source domains. For example, the harmonics in human voice differ significantly from those in musical instruments, and the melodies, which are widely present in musics, are rarely seen in ambient sounds. Although, this discrepancy has been acknowledged in the traditional audio coding field with some solutions such as the Unified Speech Audio Codec (USAC) [11], there is still a lack of investigation into this issue in deep learning based methods. We argue

This work was funded by the European Union (ERC, HI-Audio, 101052978). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2024-AD011015054 on the supercomputer Jean Zay’s V100 and A100 partition.

that this may limit the performance in diverse audio environments, especially the explainability of latent features. Addressing this is crucial for future advancements in audio compression and generation.

In this work, we propose the Source Disentangled Neural Audio Codec (SD-Codec), a novel framework that integrates latent source disentanglement into the neural audio codec framework. Given an audio input that may contain multiple sources, we design multiple domain-specific quantizers, each corresponding to audio sources including speech, music, and sound effects (SFX) [12]. SD-Codec learns to disentangle the latent features and distribute features of source domains to different quantizers. The decoder of SD-Codec can reconstruct either a single source track from a specific quantizer, or reconstruct the mixture using the sum of quantized latent features. Furthermore, inspired from [13], we explore the source disentanglement of SD-Codec by sharing the last several layers of the quantizers. Aiming at a generalized audio codec model, we follow the instructions from [12] and [14] to build mixed training audio clips across different datasets and train our models on a large scale with around 6,000 hours, as shown in Tab. I. We evaluate our approach on an unseen dataset Divide and Remaster (DnR) [12] and compare it with state-of-the-art methods.

Our key contributions can be summarized as follows:

- We introduce SD-Codec, a neural audio codec that can reconstruct audio and separate sources (e.g., speech, music, sound effects) from input audio clips.
- We investigate using a shared RVQ in SD-Codec. The similar performance with and without a shared codebook suggests that RVQ’s shallow layers encode semantic information, while its deeper layers capture local acoustic details.
- We train our model on large scale dataset and demonstrate that SD-Codec achieves strong performance both on audio reconstruction and source separation.

Code and pretrained models will be open-released, the demo page can be found at: <https://xiaoyubie1994.github.io/sdcodec>.

II. RELATED WORK

A. Neural Audio Codec

Early approaches for end-to-end neural audio codecs can be traced back to 1990s [15] and have seen significant advancements with the introduction of deep convolutional neural networks [16]. Concurrently, VQ-VAE [17] introduced a learnable quantizer between the encoder and decoder, optimizing them within a probabilistic modeling framework. This approach has proven effective for neural audio codecs operating at low bit rates [18]–[20]. Subsequently, residual vector quantization and adversarial training were incorporated into VQ-VAE-based neural audio codecs, typical models like SoundStream [3] and EnCodec [4] have demonstrated impressive performance in reconstruction quality. More recently, DAC [5] introduced vector factorization and L2-normalization to enhance codebook usage and achieves state-of-the-art on audio reconstruction quality.

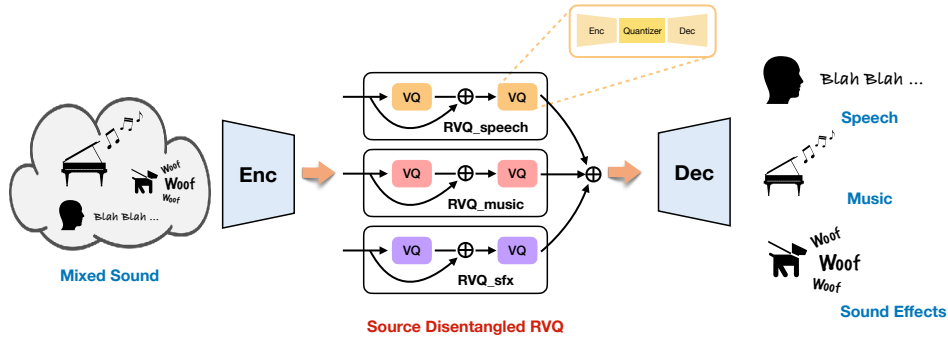


Fig. 1. Source Disentangled Neural Audio Codec (SD-Codec)

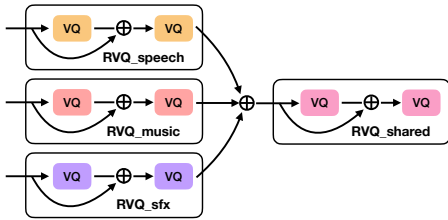


Fig. 2. SD-Codec with shared codebooks ($R = 4$, $S = 2$)

B. Latent Feature Disentanglement

In parallel to improving the quality of reconstruction, some studies dive into learning disentangled latent features for neural audio codec. Multiple encoders were introduced in [21] to extract phonemes, pitch and speaker identity respectively with different training targets. Later in NaturalSpeech3 [22], a unified encoder was shown to be sufficient. In addition to disentangling semantic information, Gull [23] learn to disentangle different bandwidth, whereas [24], [25] learn to separate noisy or reverberated speech. Our work is most related to [25], but we consider a more general case where human speech, music and environment sound effects are all mixed up, and we do not force the latent feature to disentangle each audio source into certain channels, but let the quantizer to handle it.

C. Audio Source Separation

While supervised speech separation has been widely studied over the decade [26], current approaches usually compress the audio signals into a latent space and learn a specific separation network to separate different audio sources, such as ConvTasNet [27] and ResUNet [28]. Latterly, TDANet [29] achieved great advance by using the feature pyramid and repeating the separation module multiple times. Apart from learning a separation model on a single dataset, recent works, such as GASS [14] and AudioSep [30], scale up the size of training data, targeting a single model for universal source separation.

III. METHOD

Our model is based on DAC [5], designed for 16 kHz. We describe the details of SD-Codec below.

A. Model Architecture

The architecture of our model is shown in Fig. 1. The input is a single-channel audio waveform $\mathbf{x} \in \mathbb{R}^T$, this audio signal can either be a single source from speech, music, sound effects, or any

TABLE I
SUMMARY OF THE DATASETS USED FOR TRAINING.

Dataset	Speech	Music	SFX	# Recordings
DNS 5 [31]	✓	✗	✓	1, 185, 771
MTG-Jamendo [32]	✗	✓	✗	55, 701
MUSAN [33]	✓	✓	✓	1, 983
WHAM! [34]	✗	✗	✓	1, 575
Summary	2, 619h	3, 819h	261h	1, 245, 030

combination of them. The encoder (**Enc**) consists of an input 1D convolution, followed by several residual convolution blocks, and ends up by another 1D convolution. The latent features obtained from the encoder are $\mathbf{z} = \text{Enc}(\mathbf{x}) \in \mathbb{R}^{F \times D}$, where F is the number of frames and D is the latent feature dimension. The decoder (**Dec**) is a mirror structure of the encoder. We use the same hyper-parameters as in DAC-16kHz [5] for the architecture design, the latent features thus have a dimension of 1, 024 and have 50 latent frames per second.

For quantization, We use the residual vector quantizer (**RVQ**). Each VQ layer consists of a projection layer to compress the latent features to lower dimensional space ($d = 8$). The compressed latent features will be assigned to a nearest codebook centroid based on Euclidean distance after L2-normalization, then the quantized features will be projected back to its original dimension D . Different from DAC, SD-Codec have three domain-specific RVQs including $\text{RVQ}_{\text{speech}}$, $\text{RVQ}_{\text{music}}$ and RVQ_{sfx} . Each RVQ has R residual VQ layers and learns domain-disentangled latent features, which can either be used to reconstruct the audio for a specific source domain, or can be added up to reconstruct the mixture. We further design a variant of SD-Codec where all the domain-specific RVQs share the last S VQ layers, i.e. a $\text{RVQ}_{\text{shared}}$ is concatenated after domain-specific RVQs. Fig. 2 gives an example of $R = 4$ and $S = 2$. For example, if we assume the codebook size in each layer is 2^K , then the bits allocated for each frame is $R \times K$. In practice, we use $R = 12$ and $K = 10$ to have a bitrate of 6, 000 bits per second (6 kbps), which is the same as DAC-16kHz for single source resynthesis.

B. Source Disentanglement and Resynthesis

We consider a mixture speech, music and sound effects, the latent features from the encoder can be written as:

$$\mathbf{x}_{\text{mix}} = \mathbf{x}_{\text{speech}} + \mathbf{x}_{\text{music}} + \mathbf{x}_{\text{sfx}} \quad (1)$$

$$\mathbf{z} = \text{Enc}(\mathbf{x}_{\text{mix}}) \quad (2)$$

TABLE II
RESULTS OF ZERO-SHOT AUDIO RESYNTHESIS AND SOURCE SEPARATION.

		Audio Resynthesis								
		Mix		Speech		Music		Sound Effects		
		Method	SI-SDR(↑)	VisQOL(↑)	SI-SDR(↑)	VisQOL(↑)	SI-SDR(↑)	VisQOL(↑)	SI-SDR(↑)	VisQOL(↑)
Val	DAC [5]	4.52±2.21	4.12±0.19	7.60±2.58	4.49±0.12	5.22±4.03	4.17±0.21	0.85±6.02	3.95±0.34	
	SD-Codec	7.02±2.83	4.28±0.17	8.33±3.67	4.45±0.18	7.72±4.91	4.03±0.30	2.32±6.73	3.96±0.37	
Test	DAC [5]	4.57±1.98	4.13±0.17	7.63±2.29	4.49±0.10	5.20±3.77	4.17±0.19	1.25±5.10	3.98±0.31	
	SD-Codec	6.98±2.49	4.29±0.15	8.28±3.26	4.44±0.15	7.65±4.60	4.03±0.28	2.54±5.65	3.98±0.34	

		Source Separation									
		Speech			Music			Sound Effects			
		Method	SI-SDR(↑)	SI-SDRi(↑)	VisQOL(↑)	SI-SDR(↑)	SI-SDRi(↑)	VisQOL(↑)	SI-SDR(↑)	SI-SDRi(↑)	VisQOL(↑)
Val	TDANet [29]	11.95±2.97	9.75±3.16	3.10±0.46	1.94±4.34	8.74±4.05	2.63±0.58	0.88±6.16	8.41±4.29	2.33±0.78	
	SD-Codec	11.26±3.35	9.07±3.13	3.41±0.48	1.73±4.23	8.53±3.45	2.84±0.65	0.91±5.23	8.45±3.88	2.44±0.79	
Test	TDANet [29]	11.86±2.66	9.91±2.94	3.21±0.39	2.12±3.81	8.83±3.68	2.74±0.50	1.87±4.79	8.62±3.45	2.49±0.72	
	SD-Codec	11.31±2.98	9.36±2.80	3.49±0.40	1.85±3.68	8.57±3.04	2.96±0.56	1.77±4.08	8.52±3.23	2.64±0.72	

Then, each source-aware RVQ learns the specific features related to its source, formally:

$$\hat{\mathbf{z}}_s^{(0)} = 0, \quad (3)$$

$$\hat{\mathbf{z}}_s^{(i+1)} = VQ_s^{(i+1)}(\mathbf{z} - \sum_{r=0}^i \hat{\mathbf{z}}_s^{(r)}), \quad (4)$$

where $s \in \{\text{speech, music, sfx}\}$ and $VQ_s^{(i+1)}$ represents the $(i+1)$ th quantization layer for the source s .

To reconstruct a single source such as speech:

$$\hat{\mathbf{z}}_{\text{speech}} = \sum_{i=1}^R \mathbf{z}_{\text{speech}}^{(i)}, \quad (5)$$

$$\hat{\mathbf{x}}_{\text{speech}} = \text{Dec}(\hat{\mathbf{z}}_{\text{speech}}). \quad (6)$$

To reconstruct the mixture, we assume the linear additivity in the latent space:

$$\hat{\mathbf{z}}_{\text{mix}} = \hat{\mathbf{z}}_{\text{speech}} + \hat{\mathbf{z}}_{\text{music}} + \hat{\mathbf{z}}_{\text{sfx}}, \quad (7)$$

$$\hat{\mathbf{x}}_{\text{mix}} = \text{Dec}(\hat{\mathbf{z}}_{\text{mix}}). \quad (8)$$

While this shows the case of three-source mixture, it is also valid for the mixture of two sources. Here we need three times more bitrate for 3-source mixture than single source. To be noted that, if we share all the codebooks, i.e. $S = R$, then the SD-Codec will be the same as the DAC model, whereas if there is no codebook shared, then SD-Codec will consist of three RVQ in parallel.

C. Training Objective

Since SD-Codec is able to reconstruct four audio (three single source and one mixture), the final loss is:

$$\mathcal{L}_{\text{tot}} = \sum_s \mathcal{L}_g(\hat{\mathbf{x}}_s, \mathbf{x}_s), \quad (9)$$

where $s \in \{\text{mix, speech, music, sfx}\}$. The \mathcal{L}_g is the same generator loss used in DAC [5], which consists of the multi-scale mel-spectrograms loss, the feature matching loss, the codebook loss and the commitment loss. Same as DAC, we also use a multi-period waveform discriminator (MPD) and a complex STFT discriminator to compute the feature matching loss and the adversarial loss. Following [25], we randomly shuffle the latent feature combination in the batch to enhance the feature disentanglement. Moreover, to balance the model performance on resynthesis and separation, we randomly select the number of audio tracks used to make up the mix. In our experiments, the probabilities of using 1, 2, and 3 tracks are 0.6, 0.2, and 0.2, respectively.

IV. EXPERIMENTS

A. Dataset

We trained our models on a large dataset across speech, music, and sound effects. As summarized in Tab. I, we used the clean speech corpus and noise corpus from DNS Challenge 5 [31], MTG-Jamendo [32], WHAM! [34] and MUSAN [33]. In general, the training dataset is compiled of 2,619 hours of speech data, 3,819 hours of music data, and 261 hours of sound effects data. To evaluate the generalized performance of our models, we used the validation set and test set from DnR [12]. Since the DnR dataset is composed of artificial mixtures using speech, music, and sound effects which are different from our training sources, the validation and test can be considered as zero-shot, or out-of-distribution evaluation. We trimmed the leading and trailing silence in the mixture audio, and clipped the remaining audio into segments with a length either of 5-second (validation), or of 10-second (test). We also removed those segments where the individual tracks in the mix were too short (less than 50% of the mixture). Finally, the evaluation dataset consisted of 2,853 mixtures (≈ 4 h) from the validation and 1,840 mixtures (≈ 5 h) from the test.

For training, we followed the pre-processing in DnR data composition.¹ We first normalized each source to a specific loudness units full-scale (LUFS) [35]. For speech, music and sound effects, the target LUFS values were -17 , -24 and -21 dB respectively, with additional random perturbation uniformly sampled from ± 2 . Then we normalized the peak value to -0.5 dB if any of the audio sources had a peak value larger than this threshold. Finally, we mixed up all the individual audio sources and normalize the mixture to -27 LUFS with random perturbation.

B. Evaluation

For evaluating the performance on audio resynthesis and source separation tasks, we both use Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [36] for signal-level audio quality evaluation and VisQOL [37] for perceptual audio quality assessment. Since each source has been normalized to different LUFS, we also report SI-SDR improvement (SI-SDRi) in source separation.

C. Training

We train all models for 400,000 iterations with the Adam optimizer with a batch size of 64 examples of 2 seconds each, a learning rate of $1 \cdot 10^{-4}$, $\beta_1 = 0.8$, and $\beta_2 = 0.99$. We linearly warm up

¹<https://github.com/darius522/dnr-utils/>

TABLE III
RESULTS OF ABLATION STUDY ON OUR PROPOSED CODEC.

	Method	Mix	Re-synthesis (SI-SDR \uparrow)			Source Separation (SI-SDR \uparrow)		
			Speech	Music	Sfx	Speech	Music	Sfx
Val	SD-Codec	7.02 \pm 2.83	8.33 \pm 3.67	7.72 \pm 4.91	2.32 \pm 6.73	11.26 \pm 3.35	1.73 \pm 4.23	0.91 \pm 5.23
	+ shared codebook (S=4)	7.19 \pm 2.79	8.65 \pm 3.58	7.74 \pm 4.81	2.44 \pm 6.61	11.18 \pm 3.31	1.60 \pm 4.28	0.66 \pm 5.20
	+ shared codebook (S=8)	7.14 \pm 2.76	8.60 \pm 3.55	7.78 \pm 4.82	2.47 \pm 6.70	11.13 \pm 3.25	1.67 \pm 4.04	0.67 \pm 5.20
	+ separation enhance	7.21 \pm 2.82	7.95 \pm 3.86	6.84 \pm 4.84	1.02 \pm 6.55	11.78 \pm 3.27	2.33 \pm 4.14	1.44 \pm 5.32
	+ initialization from DAC	-5.20 \pm 2.86	-1.64 \pm 3.69	-3.43 \pm 5.34	-14.08 \pm 11.61	10.28 \pm 3.15	0.70 \pm 4.27	-0.12 \pm 5.06
Test	SD-Codec	6.98 \pm 2.49	8.28 \pm 3.26	7.65 \pm 4.60	2.54 \pm 5.65	11.31 \pm 2.98	1.85 \pm 3.68	1.77 \pm 4.08
	+ shared codebook (S=4)	7.15 \pm 2.46	8.60 \pm 3.18	7.67 \pm 4.51	2.68 \pm 5.56	11.21 \pm 3.00	1.71 \pm 3.73	1.52 \pm 3.99
	+ shared codebook (S=8)	7.11 \pm 2.43	8.57 \pm 3.15	7.70 \pm 4.54	2.69 \pm 5.52	11.18 \pm 2.90	1.79 \pm 3.53	1.54 \pm 4.11
	+ separation enhance	7.17 \pm 2.48	7.91 \pm 3.40	6.77 \pm 4.56	1.29 \pm 5.53	11.83 \pm 2.91	2.46 \pm 3.61	2.27 \pm 4.23
	+ initialization from DAC	-5.02 \pm 2.47	-1.54 \pm 3.12	-3.15 \pm 4.83	-12.22 \pm 9.49	10.34 \pm 2.81	0.87 \pm 3.68	0.78 \pm 3.95

the learning rate for the first 10,000 iterations and then apply an exponential scheduler with $\gamma = 0.999996$. All the models are trained using 8 NVIDIA A100 GPUs.

D. Results

To evaluate the performance of SD-Codec on audio resynthesis and source separation, we compare it with DAC [5] and TDANet the state-of-the-art models on audio codec model and source separation. [29],² For fair comparison, we use their official implementations.^{3 4} and re-train their models on the same training data corpus as SD-Codec. The results are shown in Tab. II. In general, we observe that all the models achieve decent performance on zero-shot evaluation with different data corpus and different data lengths, which confirms the benefit of generalization from scale-up training.

In audio resynthesis, since SD-Codec has triple bitrates than DAC on mixed audio, there is no surprise that SD-Codec achieves better reconstruction quality. When it comes to single source resynthesis, SD-Codec has the same bitrate as DAC. The reconstructed audio clips from SD-Codec have higher SI-SDR of around 1 dB than those from DAC, and achieve comparable VisQOL values.

In source separation experiments, we do not use the direct output of the decoder. Instead, we use a mask augmentation method where the decoder output is used to compute a magnitude mask. This mask is then applied to the magnitude of the mixture input, and each audio track is reconstructed using the noisy phase from the mixture. We observed that this operation will improve the SI-SDR, and has slight positive impact on VisQOL values. Compared to TDANet, although SD-Codec is not specifically designed for source separation and experiences quantization loss, it achieves comparable results to TDANet in terms of SI-SDR, SI-SDR_i, and VisQOL.

From the above results in Tab. II, we can conclude that our proposed SD-Codec, while explicitly assigning latent features from different sources to different codebooks, can be applied both on audio resynthesis and source separation, and achieve promising results.

To investigate whether SD-Codec learns a domain-specific quantizer, we present the results of reconstructing different audio sources with various RVQ configurations in Fig. 3. The results clearly demonstrate that only the corresponding RVQ module can reconstruct the audio signal with high quality.

E. Ablation Study

In ablation experiments shown in Tab. III, we mainly studied the changes brought about by different designs of SD-Codec:

²While TDANet was designed for speech separation, it also showed state-of-the-art performance in general source separation in [14]

³<https://github.com/descriptinc/descript-audio-codec>

⁴<https://github.com/JusperLee/TDANet>

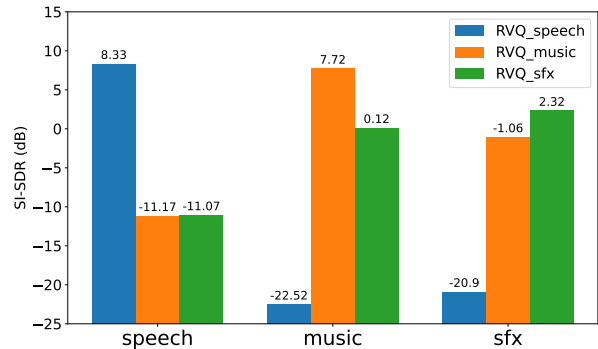


Fig. 3. Single source audio resynthesis using different RVQ modules.

- *Shared codebook*: As shown in Fig. 2, we share the last layers for each source. We observe that by either sharing the last 4 layers or the last 8 layers, SD-Codec can maintain similar performance. This shows that in RVQ modules, the shallow layers encode the majority of source-aware information, while the deeper layers encode the local acoustic details.
- *Separation enhance*: When training the original SD-Codec, we randomly used 1, 2, or 3 tracks as input with probabilities of 0.6, 0.2, and 0.2, respectively. We also experimented with different probabilities of 0.2, 0.2, and 0.6, which increased the likelihood of the model encountering mixtures of 3 tracks, thereby enhancing its ability to learn separation. As shown in the ablation results, this will bring higher reconstruction quality on the mixture and higher separation quality, with the expense of losing the quality of individual signal reconstruction.
- *Initialization from DAC*: We also studied the potential to leverage from pre-trained DAC on SD-Codec training. However, due to the huge difference in objective, i.e. learning a unified latent space for all audio sources and learning source disentangled latent space, SD-Codec suffers from the DAC pretraining.

V. CONCLUSION

In this work, we designed a source-aware audio codec, SD-Codec. By joint learning audio coding and source separation, SD-Codec obtain excellent results, comparable to the state-of-the-art both on audio reconstruction and separation, with a unified model. Meanwhile, SD-Codec explicitly assigns latent features from different audio domain to different codebook, which brings more explainability and interpretability on the latent features. Furthermore, due to the source disentanglement in the latent space, future audio generation models can have more fine-grained manipulations on the generated audio, paving the way to the research of controllable audio generation.

REFERENCES

- [1] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache *et al.*, "Overview of the evs codec architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015.
- [2] J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus Audio Codec," IETF RFC 6716, 2012, <https://tools.ietf.org/html/rfc6716>.
- [3] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 495–507, 2021.
- [4] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, 2023.
- [5] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
- [6] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.
- [7] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 1336–1354, 2021.
- [8] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2523–2533, 2023.
- [9] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [10] R. San Roman, Y. Adi, A. Deleforge, R. Serizel, G. Synnaeve, and A. Défossez, "From discrete tokens to high-fidelity audio using multi-band diffusion," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.
- [11] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach *et al.*, "Unified speech and audio coding scheme for high quality at low bitrates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2009.
- [12] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, "The cocktail fork problem: Three-stem audio separation for real-world soundtracks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022.
- [13] B. Ginies, X. Bie, O. Fercoq, and G. Richard, "Using random codebooks for audio neural autoencoders," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2024.
- [14] J. Pons, X. Liu, S. Pascual, and J. Serrà, "Gass: Generalizing audio source separation with large-scale data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024.
- [15] S. Morishima, H. Harashima, and Y. Katayama, "Speech coding based on a multi-layer neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1990.
- [16] S. Kankanahalli, "End-to-end optimized speech coding with deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018.
- [17] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
- [18] C. Gărbacea, 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019.
- [19] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021.
- [20] H. Yang, I. Jang, and M. Kim, "Generative de-quantization for neural speech codec via latent diffusion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024.
- [21] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *Proc. Interspeech Conf.*, 2021.
- [22] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.
- [23] Y. Luo, J. Yu, H. Chen, R. Gu, and C. Weng, "Gull: A generative multifunctional audio codec," *arXiv preprint arXiv:2404.04947*, 2024.
- [24] H. Yang, K. Zhen, S. Beack, and M. Kim, "Source-aware neural speech coding for noisy speech compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021.
- [25] A. Omran, N. Zeghidour, Z. Borsos, F. de Chaumont Quitry, M. Slaney, and M. Tagliasacchi, "Disentangling speech from surroundings with neural embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.
- [26] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [27] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [28] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resnet for music source separation," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2021.
- [29] K. Li, R. Yang, and X. Hu, "An efficient encoder-decoder architecture with top-down attention for speech separation," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [30] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *arXiv preprint arXiv:2308.05037*, 2023.
- [31] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, "Icassp 2023 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023.
- [32] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, 2019.
- [33] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [34] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *Proc. Interspeech Conf.*, 2019.
- [35] E. Grimm, R. Van Everdingen, and M. J. L. C. Schöpping, "Toward a recommendation for a european standard of peak and LKFS loudness levels," *SMPTE Motion Imaging Journal*, vol. 119, no. 3, pp. 28–34, 2010.
- [36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019.
- [37] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: an objective speech quality model," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, pp. 1–18, 2015.