Learning Source Disentanglement in Neural Audio Codec

Xiaoyu Bie

Audio Data Analysis and Signal Processing Group (ADASP)

Télécom Paris, Institut Polytechnique de Paris

xiaoyu.bie@telecom-paris.fr

About Me

- Postdoctoral researcher at ADASP group in Télécom Paris, IP Paris
- Research interests:
  - generative models and representation learning for audio data
  - audio signal processing and generation
- Currently involved in Hi-Audio project
  - an ERC project hosted by Prof. Gaël Richard
  - aims at building efficient and interpretable machine listening models

“Learning Source Disentanglement in Neural Audio Codec”

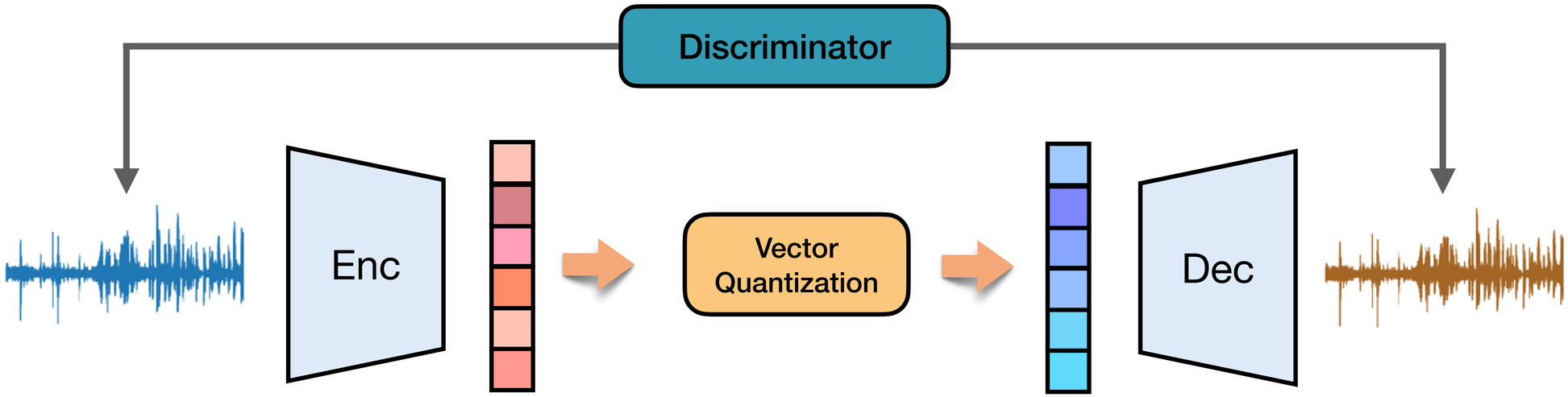
to be presented at ICASSP 2025 in Hyderabad, India

Xiaoyu Bie
Télécom Paris, IP Paris

Xubo Liu
University of Surrey

Gaël Richard
Télécom Paris, IP Paris

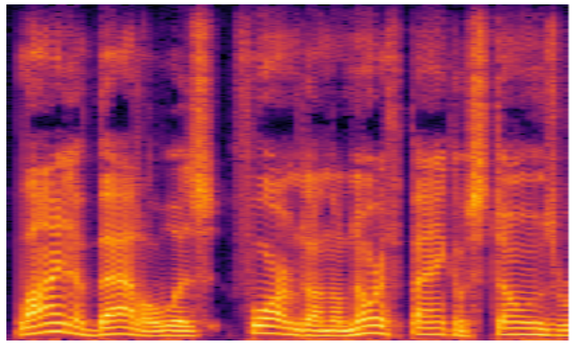
Neural Audio Codec



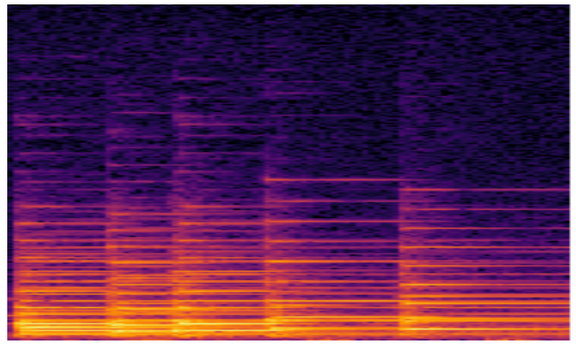
- Neural audio codec typically consists of encoder/decoder, quantization and discriminator
- Impressive quality on reconstructed audio even with very low-bitrate (e.g. Encodec[1], DAC [2] etc)

[1] A. Défossez, J. Copet, G. Synnaeve, et al., “High fidelity neural audio compression,” TMLR, 2023.
[2] R. Kumar, P. Seetharaman, A. Luebs, et al., “High- fidelity audio compression with improved rvqgan,” NeurIPS, 2024.

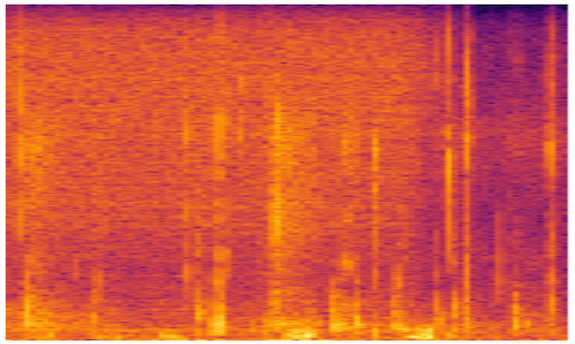
Motivation



Speech



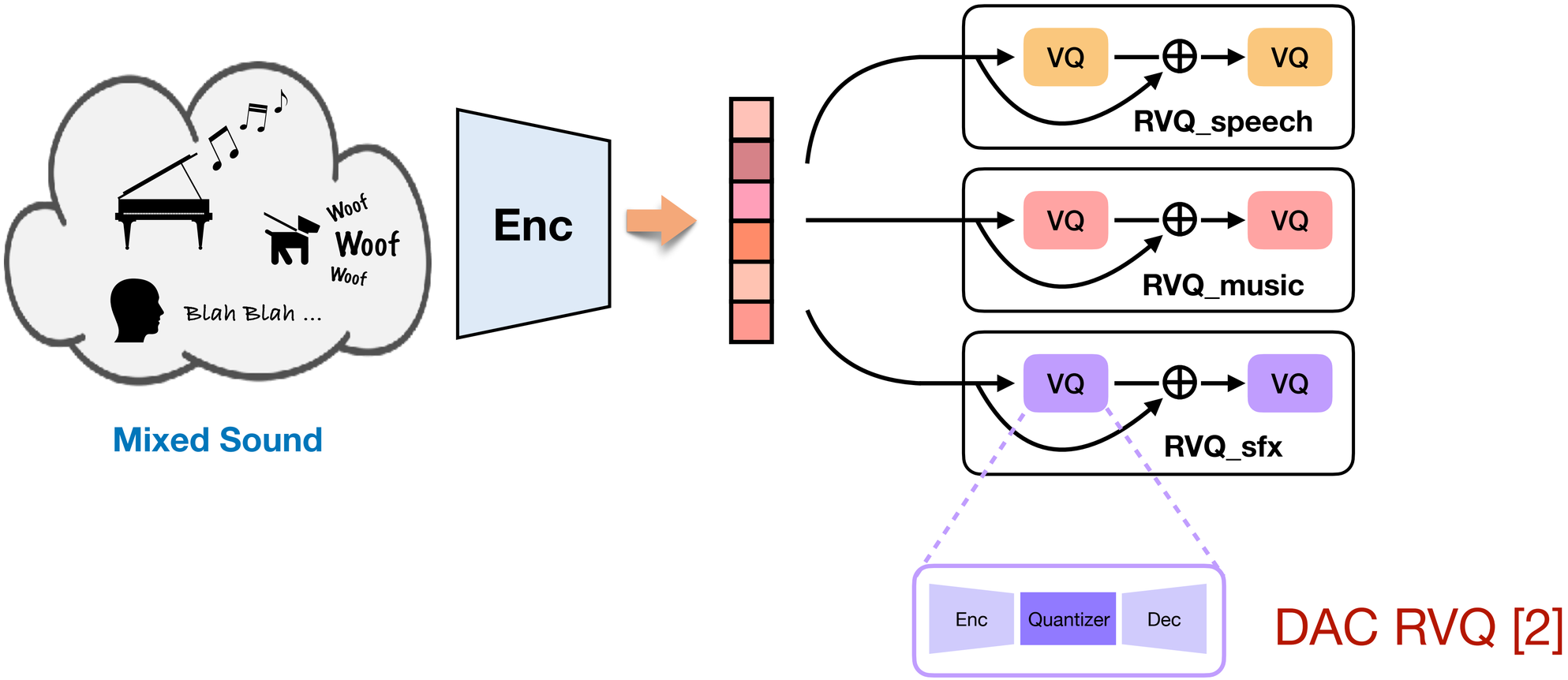
Music



Sound Effect

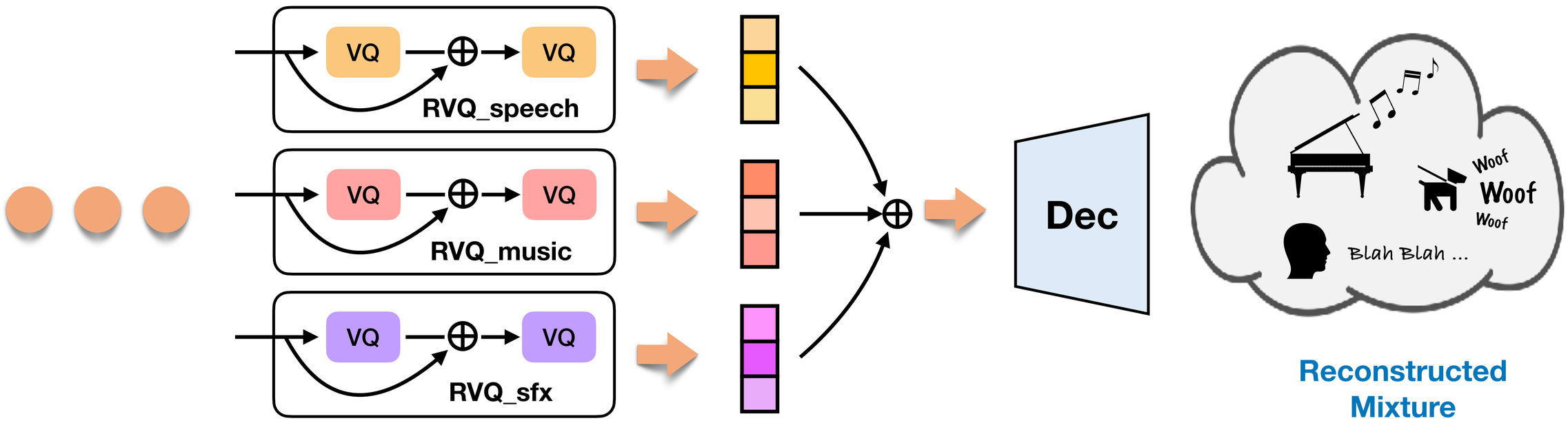
- Discrepancy in audio signals from different source domain
- Mapping different sounds into unified codebook will loose of efficiency and controllability
- It is easy to find data with general tags (speech, music, general sounds)

SD-Codec

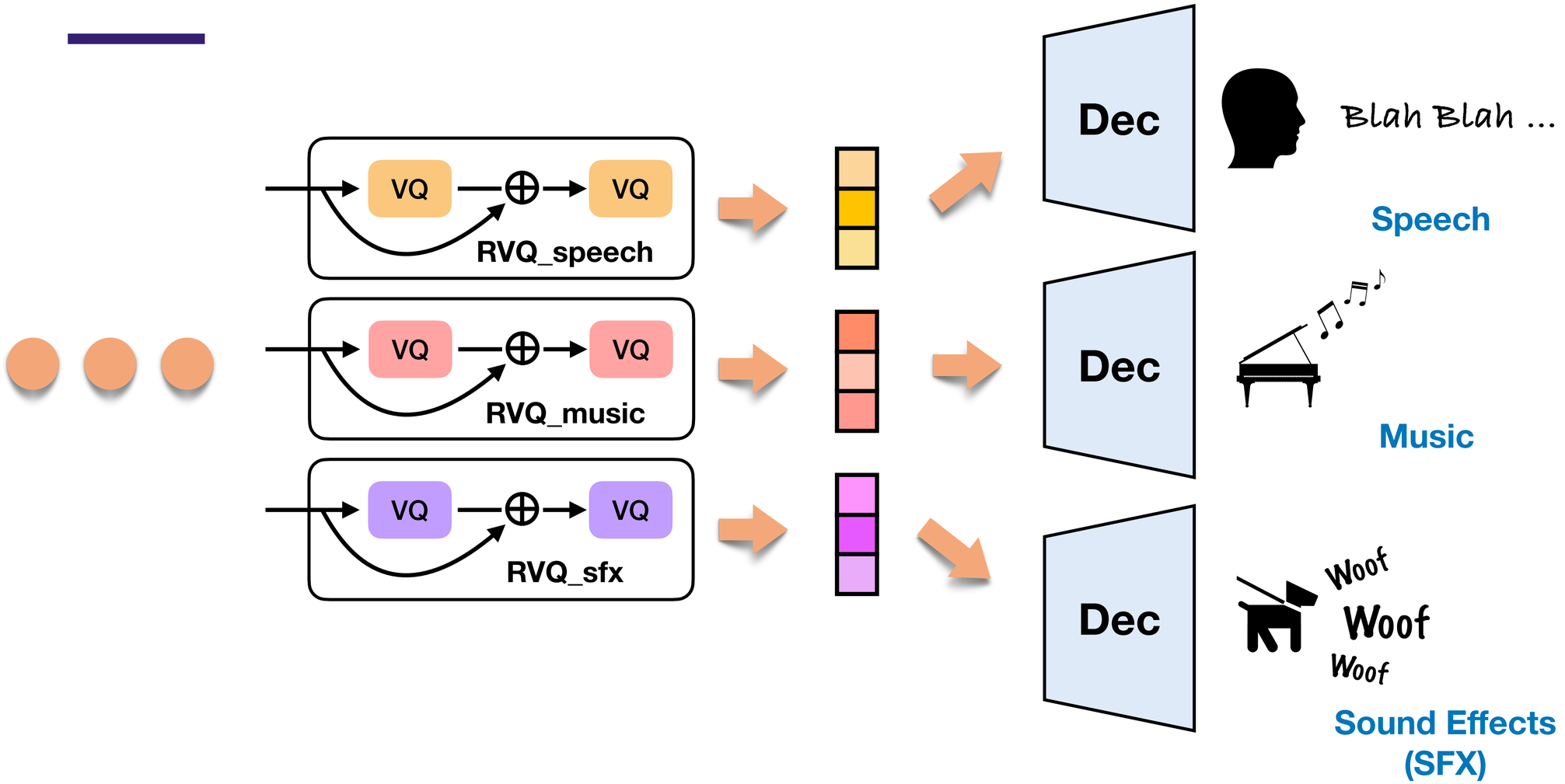


[2] R. Kumar, P. Seetharaman, A. Luebs, et al., “High- fidelity audio compression with improved rvqgan,” NeurIPS, 2024.

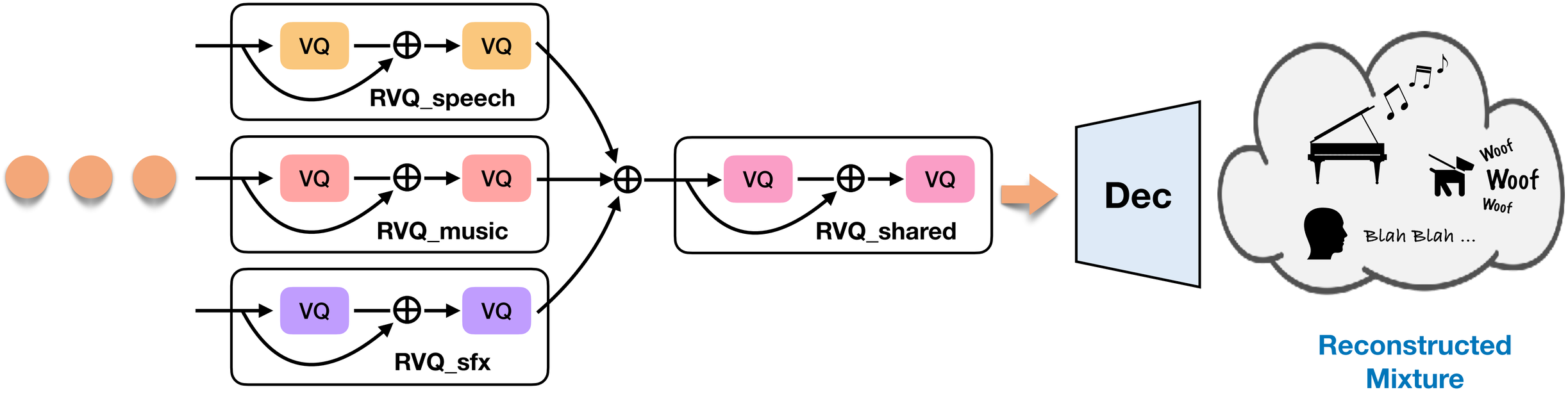
SD-Codec


VQ
VQ
RVQ_speech
VQ
VQ
RVQ_music
VQ
VQ
RVQ_sfx
Dec
Woof
Woof
Woof
Blah Blah ...
Reconstructed
Mixture

SD-Codec


VQ
VQ
RVQ_speech
VQ
VQ
RVQ_music
VQ
VQ
RVQ_sfx
Dec
Dec
Dec
Blah Blah ...
Speech
Music
Woof
Woof
Woof
Sound Effects
(SFX)

Shared Codebook



- Shared codebook doesn’t decrease the bite-rates
- Shallow layers encode semantic, deep layers encode local acoustic details

Training

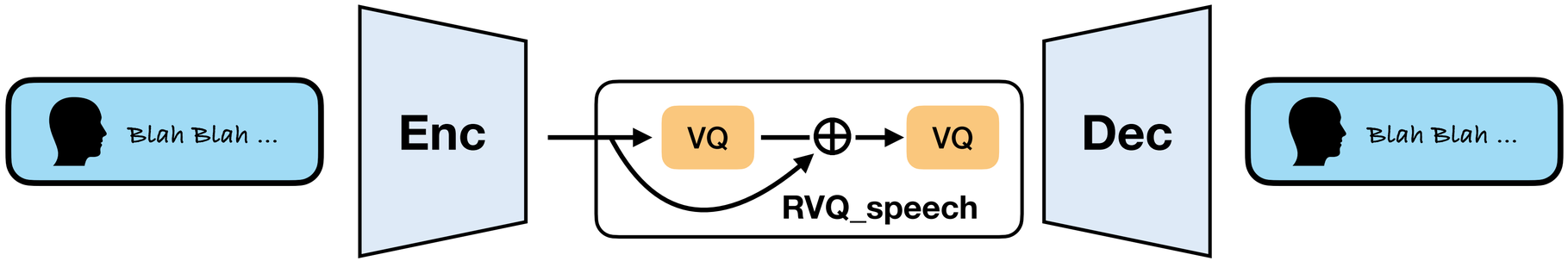
- Training corpus: ~ 6k hours
- Test corpus: Divide and Remaster (DnR) [3]
  - mixture of librispeech, FMA and FSD50k
  - val splits (5s) and test splits (10s)
- Batch size 64 with 2s segments
- 400k iterations for each model

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

Summary of training data

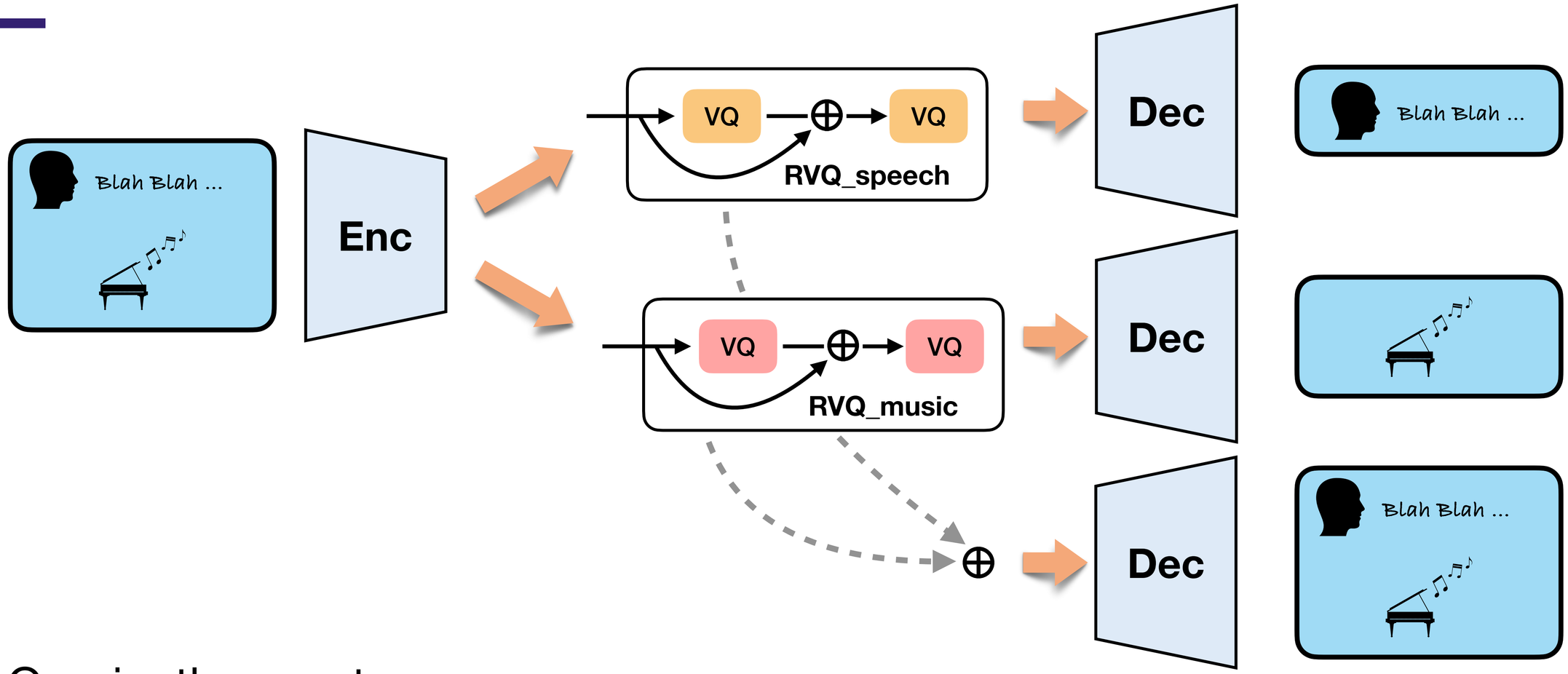
[3] D. Petermann, G. Wichern, Z.-Q. Wang et al., “The cocktail fork problem: Three-stem audio separation for real-world soundtracks,” in ICASSP, 2022.

Training: Single Track



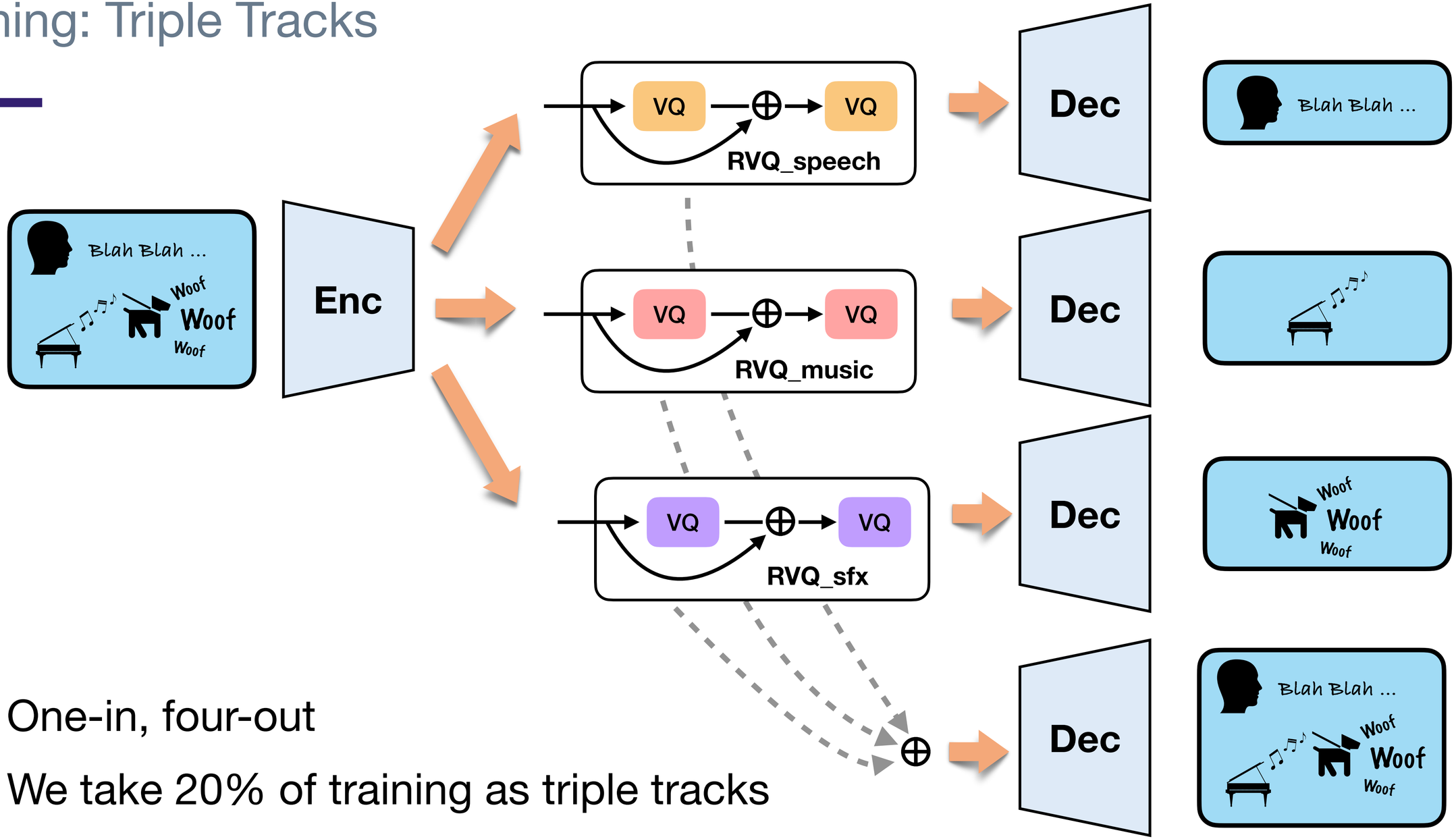
- One-in, one-out
- We take 60% of training as single track

Training: Double Tracks



- One-in, three-out
- We take 20% of training as double tracks

Training: Triple Tracks



- One-in, four-out
- We take 20% of training as triple tracks

Results

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

Comparable synthesis performance compared to DAC [2]

[2] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High- fidelity audio compression with improved rvqgan,” NeurIPS, 2024.

Results

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

Comparable separation performance compared to TDANet [4]

[4] K. Li, R. Yang, and X. Hu, “An efficient encoder-decoder architecture with top-down attention for speech separation,” ICLR, 2023.

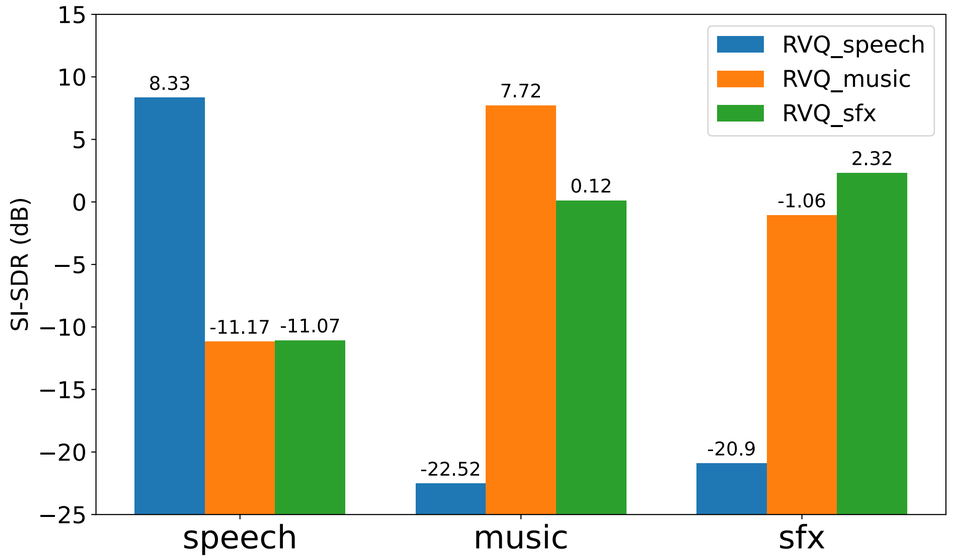
Results

	Method	Re-synthesis (SI-SDR ↑) Mix	Speech	Music	Sfx	Source Separation (SI-SDR ↑) 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	$7.21^{\pm 2.82}$	$7.95^{\pm 3.86}$	$6.84^{\pm 4.84}$	$1.02^{\pm 6.55}$	$11.78^{\pm 3.27}$	$2.33^{\pm 4.14}$	$1.44^{\pm 5.32}$
	+ initization 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	$7.17^{\pm 2.48}$	$7.91^{\pm 3.40}$	$6.77^{\pm 4.56}$	$1.29^{\pm 5.53}$	$11.83^{\pm 2.91}$	$2.46^{\pm 3.61}$	$2.27^{\pm 4.23}$
	+ initization from DAC	$-5.02^{\pm 2.47}$	$-1.54^{\pm 3.12}$	$-3.15^{\pm 4.83}$	$-12.22^{\pm 9.49}$	$10.34^{\pm 2.81}$	$0.87^{\pm 3.68}$	$0.78^{\pm 3.95}$

- Shared codebook won’t affect the performance
- Separation enhance (60% triple tracks and 20% single tracks): trade-off training
- Initialized from DAC pre-train doesn’t work

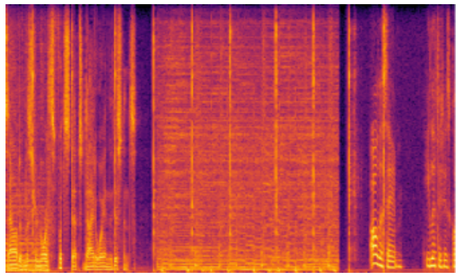
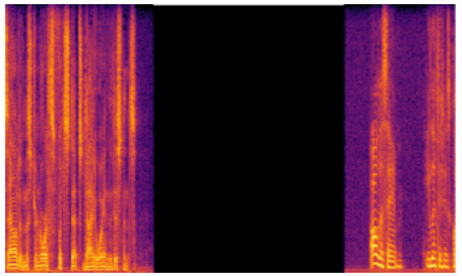
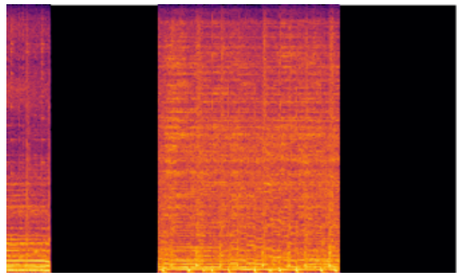
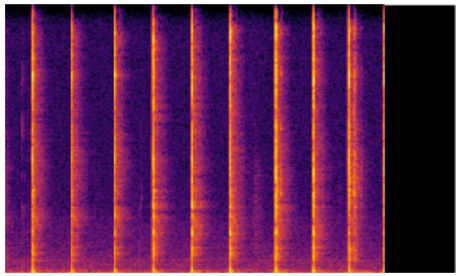
Results

- Results on out-domain coding
- Disentangled VQ branch only works for the corresponding source domain

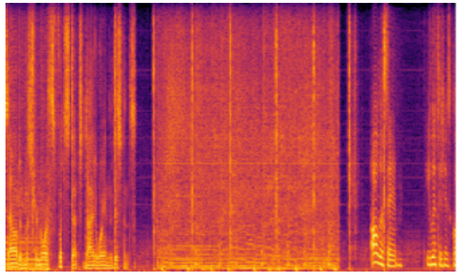
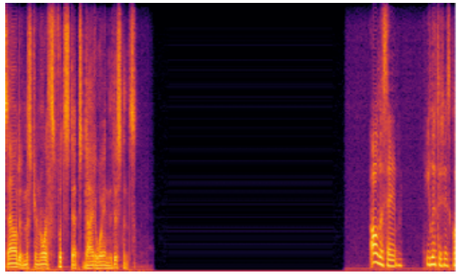
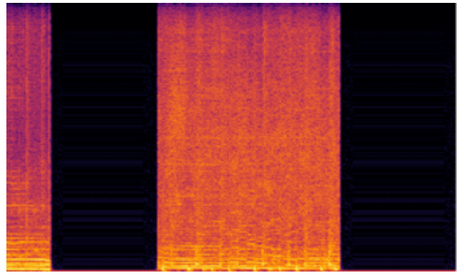
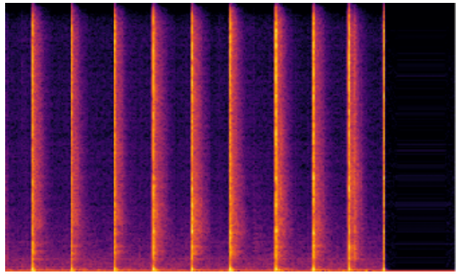


Results

GT

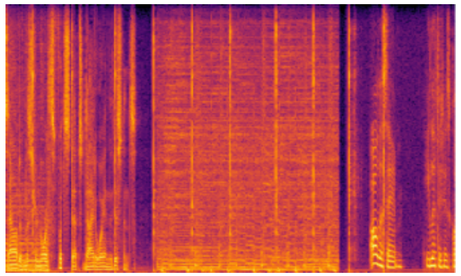
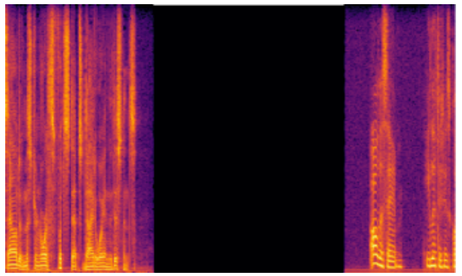
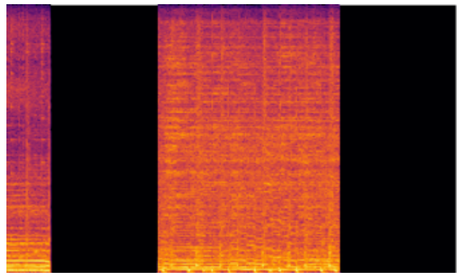
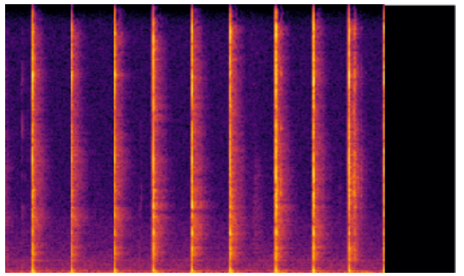
Recon

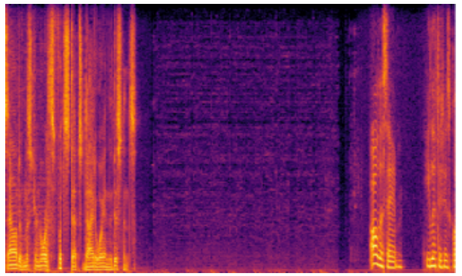
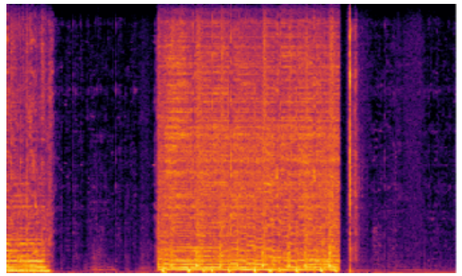
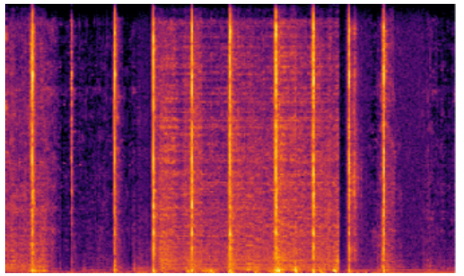
Mixture Speech Music SFX

Results

GT

Separation

Mixture Speech Music SFX

Summary

To conclude:

- SD-Codec successfully learns disentangled latent codes via source separation
- SD-Codec achieves comparable results to SOTA codec model (DAC) and source separation model (TDANet)
- Code and weights available https://github.com/XiaoyuBIE1994/SDCodec

Future work:

- Scale up to more diverse datasets
- Incorporate specific architectures for source separation
- Apply LLMs on disentangled codes for generation

Demo Page

More results on our demo page:

https://xiaoyubie1994.github.io/sdcodec/

Questions ?