WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling

季圣鹏

导师：赵洲

https://novateurjsp.github.io

WavTokenizer的背景动机一，大环境下的Codec模型

1. Codec语音类表征领域的快速发展

SoundStream/Encodec开创了端到端codec模型的范式，DAC极大地提高了acoustic codec模型的重建质量，SpeechTokenizer/SemantiCodec尝试在codec模型中加入semantic信息缩小codec与下游LLM的Gap，HiFi-Codec/SNAC/Language-Codec尝试用四层的量化器减少token数量

2. Codec离散声学表征在生成模型上验证了可行性

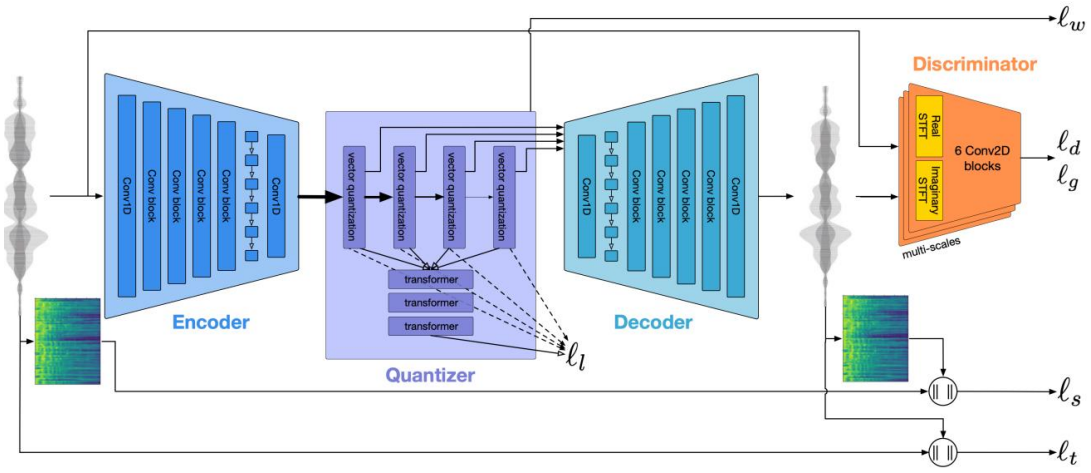
ParlerTTS在风格可控TTS上达到了很好的生成效果，VALL-E2可以生成接近人类主观听觉的音频。

Codec优势：重建的架构训练简单；包含语音特有声学信息；统一建模audio，music，speech；离散化Token适配LLM的CE loss；Token化可以和文本对齐

3. GPT-4o端到端语音对话系统的火热，多模态大模型的一种重要范式，简单的Tokenizer和DeTokenizer的趋势

以Qwen-Audio和Mini-Omni为例，在理解端是Whisper的架构，在生成端是codec架构，历史信息的互动困难

以Codec模型为基础，长远来看完成Tokenizer化和DeTokenizer化



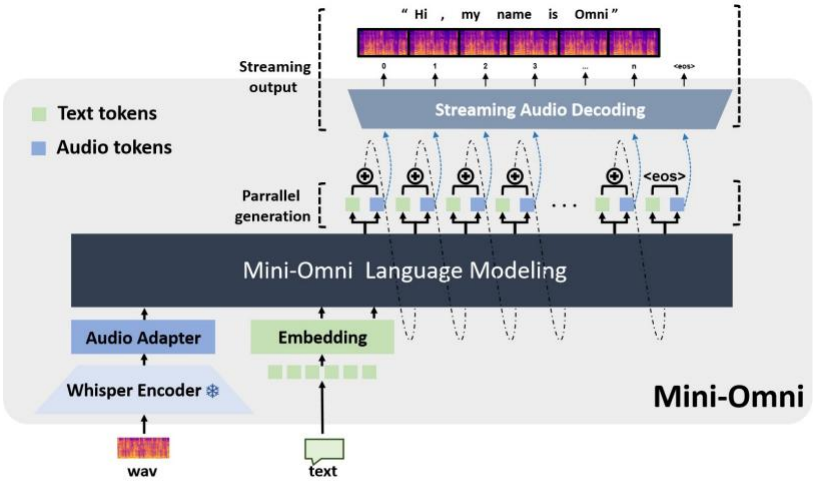


Figure 1: The **Mini-Omni** model architecture.

WavTokenizer的背景动机二，AcousticCodec模型本身的诸多瓶颈

1. 现有Codec model Token数量生成过多，引出下游CodecLLM的各种生成范式。（需要探索低码率下的重建性能）

9层的44.1khzDAC模型需要900个Token，4层的HiFi-Codec模型需要300个Token，过多的Token限制语言模型的生成能力　（单层和多层具有本质区别）

2. 如何从Codec本身出发，增强Semantic信息，缩小重建范式和LLM之前的gap，统一CodecLLM理解和生成

蒸馏的方式会限制Codec模型的质量上限

不优雅，限制统一建模music，audio和music能力

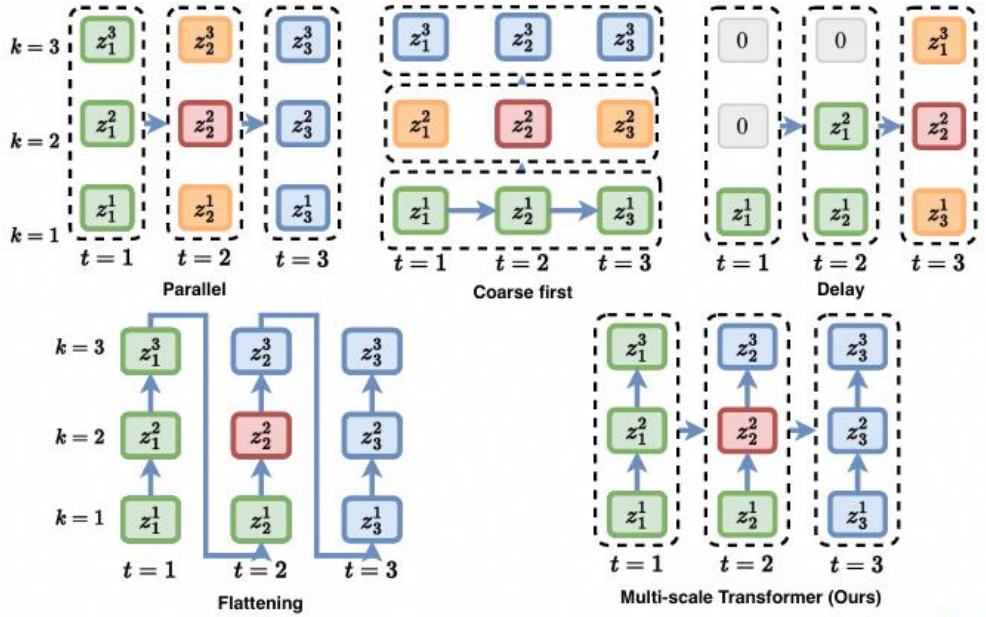
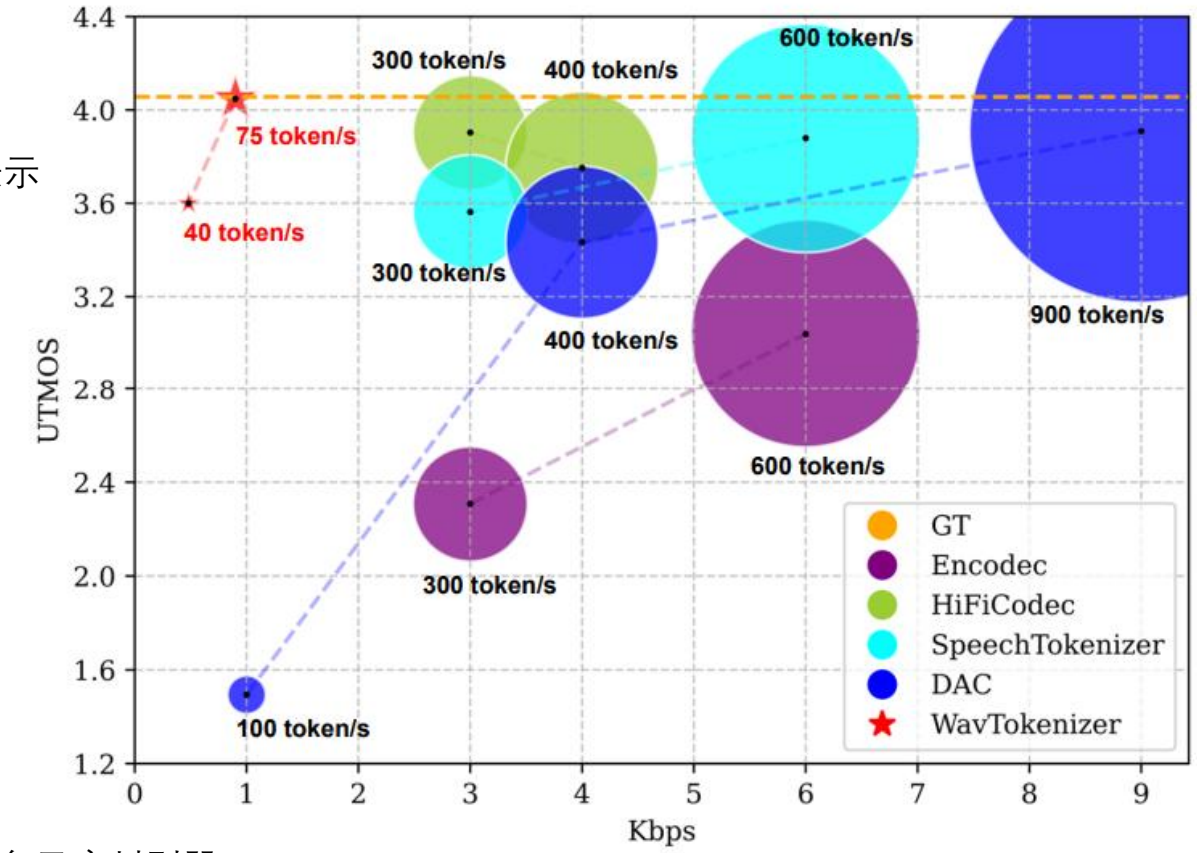


Figure 2: Order of token prediction for 4 representative methods in audio generation (Copet et al., 2023) and the proposed multi-scale Transformer. Assume $n_q = 3$ and $T = 3$. Current token prediction (red) is conditioned on prior tokens (in green). Tokens in orange are concurrently predicted with the current token. 0 is a special token indicating empty positions in the delay prediction.

WavTokenizer的贡献

1. 实验效果上

- 极限的Token数量，一秒的语音类表征（24khz）可以仅仅有40个Token表示
- 良好的主观重建效果，UTMOS4.0，并且包含丰富的语义信息

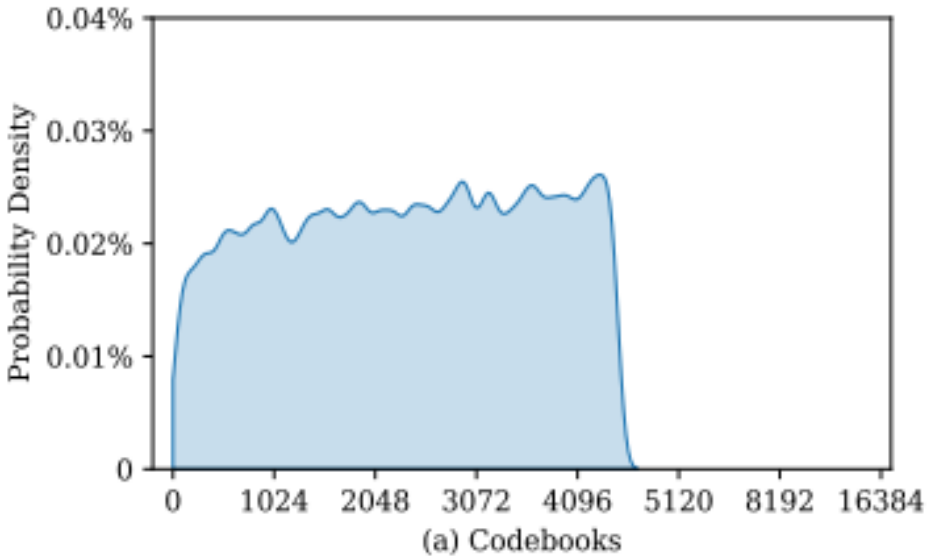


2. 方法设计上

- 验证大码本空间VQ的在极限压缩下的潜力
- 首次提出单层量化器的范式，潜在的语音作为特殊文本语言的对齐能力
- 验证Transformer架构在Codec重建任务上不会拥有长度外推的问题
- 引入更长的上下文建模窗口，引入直接逆傅里叶变换上采样模型，引入多尺度判别器

WavTokenizer的大码本空间

CodeBook的空间大小或者宽度？



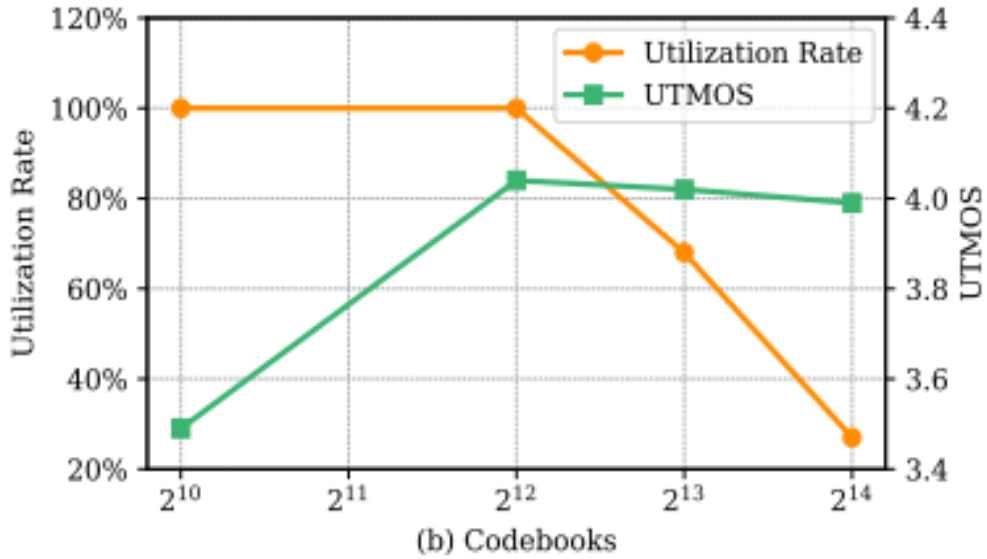


Table 8: The ablation study investigates the impact of dataset size on codebook utilization.

Model	Dataset	Codebooks	Utilization rate	UTMOS ↑	PESQ ↑	STOI ↑
WavTokenizer	585 Hours	16384	27%	3.9989	2.3600	0.8129
WavTokenizer	4000 Hours	16384	26.5%	3.9465	2.3721	0.8217

WavTokenizer的增强版decoder，更长的建模窗口和多尺度判别器

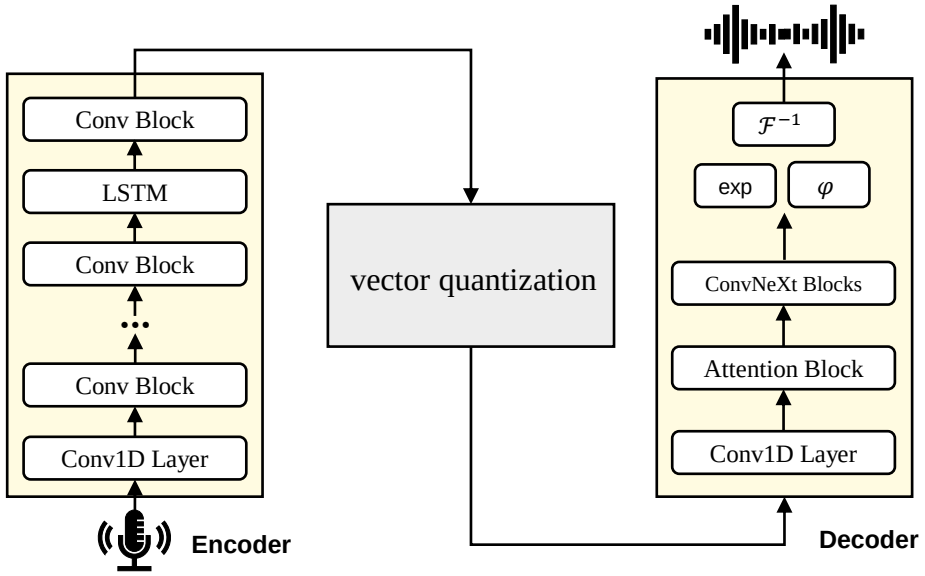


Table 6: Impact of the contextual modeling window size.

Model	Codebooks	windows	UTMOS ↑	PESQ ↑	STOI ↑
WavTokenizer	4096	1	3.7448	2.0112	0.8944
WavTokenizer	4096	3	4.0486	2.3730	0.9139
WavTokenizer	4096	5	4.0448	2.3556	0.9127

Table 7: Ablation on the multi-scale STFT discriminator (MSTFTD), the atention module, and switching from our improved decoder to a mirror decoder, in WavTokenizer.

Model	UTMOS ↑	PESQ ↑	STOI ↑	V/UV F1 ↑
WavTokenizer	**4.0486**	**2.3730**	**0.9139**	**0.9382**
w/ mirror decoder	2.7782	1.5007	0.8249	0.8820
w/o attention module	3.6020	1.9332	0.8734	0.9067
w/o MSTFTD	3.7806	2.1270	0.9008	0.9269

WavTokenizer

WavTokenizer的重建实验结果和TTS实验结果

Table 1: **Objective reconstruction results** of different codec models on LibriTTS ***test-clean*** (clean environment), LibriTTS ***test-other*** (noisy environment), and ***LJSpeech dataset*** (out-of-domain environment). **Nq** denotes the **n**umber of **q**uantizers. **GT** denotes ground truth waveforms. Best results from models with a single quantizer (hence directly comparable to WavTokenizer) are in bold.

Dataset	Model	Bandwidth ↓	Nq ↓	token/s ↓	UTMOS ↑	PESQ ↑	STOI ↑	V/UV F1 ↑
LibriTTS *test-clean*	GT	-	-	-	4.0562	-	-	-
	DAC	9.0kpbs	9	900	3.9097	3.9082	0.9699	0.9781
	Encodec	6.0kbps	8	600	3.0399	2.7202	0.9391	0.9527
	Vocos	6.0kbps	8	600	3.6954	2.8069	0.9426	0.9437
	SpeechTokenizer	6.0kbps	8	600	3.8794	2.6121	0.9165	0.9495
	DAC	4.0kbps	4	400	3.4329	2.7378	0.9280	0.9572
	HiFi-Codec	3.0kbps	4	400	3.7529	2.9611	0.9405	0.9617
	HiFi-Codec	4.0kbps	4	300	3.9035	3.0116	0.9446	0.9576
	Encodec	3.0kbps	4	300	2.3070	2.0517	0.9007	0.9198
	Vocos	3.0kbps	4	300	3.5390	2.4026	0.9231	0.9358
	SpeechTokenizer	3.0kbps	4	300	3.5632	1.9311	0.8778	0.9273
	DAC	1.0kbps	1	100	1.4940	1.2464	0.7706	0.7941
	WavTokenizer	0.5kbps	1	40	3.6016	1.7027	0.8615	0.9173
	WavTokenizer	0.9kbps	1	75	**4.0486**	**2.3730**	**0.9139**	**0.9382**
LibriTTS *test-other*	GT	-	-	-	3.4831	-	-	-
	DAC	9.0kpbs	9	900	3.3566	3.7595	0.9576	0.9696
	Encodec	6.0kbps	8	600	2.6568	2.6818	0.9241	0.9338
	Vocos	6.0kbps	8	600	3.1956	2.5590	0.9209	0.9202
	SpeechTokenizer	6.0kbps	8	600	3.2851	2.3269	0.8811	0.9205
	DAC	4.0kbps	4	400	2.9448	2.5948	0.9083	0.9404
	HiFi-Codec	4.0kbps	4	400	3.0750	2.5536	0.9126	0.9387
	HiFi-Codec	3.0kbps	4	300	3.3034	2.6083	0.9166	0.9318
	Encodec	3.0kbps	4	300	2.0883	2.0520	0.8835	0.8926
	Vocos	3.0kbps	4	300	3.0558	2.1933	0.8967	0.9051
	SpeechTokenizer	3.0kbps	4	300	3.0183	1.7373	0.8371	0.8907
	DAC	1.0kbps	1	100	1.4986	1.2454	0.7505	0.7775
	WavTokenizer	0.5kbps	1	40	3.0545	1.6622	0.8336	0.8953
	WavTokenizer	0.9kbps	1	75	**3.4312**	**2.2614**	**0.8907**	**0.9172**
LJSpeech	GT	-	-	-	4.3794	-	-	-
	DAC	9.0kpbs	9	900	4.3007	3.9022	0.9733	0.9757
	Encodec	6.0kbps	8	600	3.2286	2.6633	0.9441	0.9555
	Vocos	6.0kbps	8	600	4.0332	2.9258	0.9497	0.9459
	SpeechTokenizer	6.0kbps	8	600	4.2373	2.6413	0.9316	0.9452
	DAC	4.0kbps	4	400	3.8109	2.7616	0.9338	0.9524
	HiFi-Codec	4.0kbps	4	400	4.1656	2.7629	0.9446	0.9497
	HiFi-Codec	3.0kbps	4	300	4.2692	2.9091	0.9485	0.9469
	Encodec	3.0kbps	4	300	2.3905	2.0194	0.9058	0.9326
	Vocos	3.0kbps	4	300	3.7880	2.5006	0.9310	0.9388
	SpeechTokenizer	3.0kbps	4	300	3.9908	2.0458	0.9021	0.9299
	DAC	1.0kbps	1	100	1.4438	1.2084	0.7822	0.8095
	WavTokenizer	0.5kbps	1	40	4.0186	2.1142	0.9093	**0.9406**
	WavTokenizer	0.9kbps	1	75	**4.2580**	**2.4923**	**0.9312**	0.9397

Table 2: The **subjective reconstruction results** using MUSHRA (comparative scoring of samples) of codec models on speech, music and audio domains. **Nq** denotes the **n**umber of **q**uantizers.

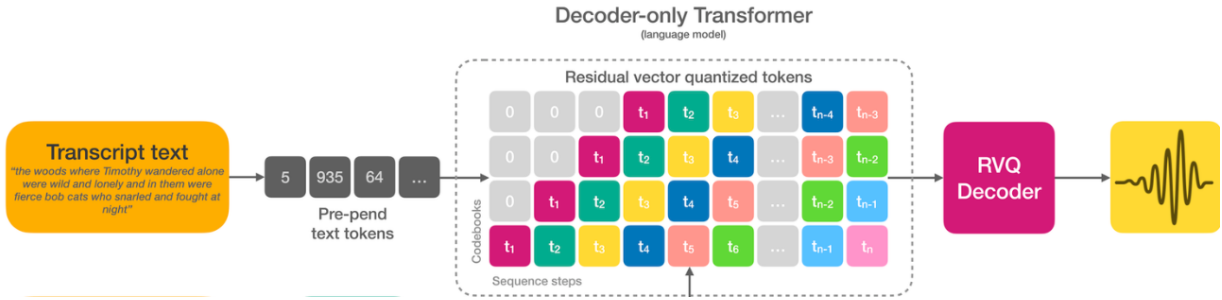
Model	Bandwidth ↓	Nq ↓	token/s ↓	*LibriTTS test-clean* ↑	*MusicDB* ↑	*Audioset* ↑
GT	-	-	-	96.4±1.2	95.3±1.7	95.8±2.1
DAC	9.0kpbs	9	900	92.8±1.8	92.6±2.4	92.7±1.5
Encodec	6.0kbps	8	600	78.6±1.9	76.9±1.6	81.2±1.8
DAC	1.0kbps	1	100	58.4±2.4	57.6±2.1	56.8±1.4
WavTokenizer	0.9kbps	1	75	**96.1±2.3**	**92.9±2.2**	**94.4±1.6**

Table 3: The **semantic representation (speech, music, audio)** evaluation of different codec models on ARCH Benchmark in terms of **classification accuracy**. **Nq** represents the **n**umber of **q**uantizers.

Model	Nq ↓	token/s ↓	RAVDESS ↑	SLURP ↑	EMOVO ↑	AM ↑	FMA ↑	MTT ↑	IRMAS ↑	MS-DB ↑	ESC50 ↑	US8K ↑	FSD50K ↑	VIVAE ↑
DAC	9	900	0.3750	0.0779	0.2363	0.6926	0.3504	0.2805	0.4023	0.6014	0.2594	0.4032	0.1297	0.3440
Encodec	8	600	0.2881	0.0636	0.2261	0.4388	0.2790	0.1993	0.3671	0.3917	0.1925	0.3055	0.1091	0.3005
DAC	4	400	0.3194	0.0782	0.2346	0.6838	0.3379	0.2784	0.3833	0.5942	0.2580	0.3824	0.1293	0.3342
Encodec	4	300	0.2951	0.0660	0.2193	0.4301	0.2728	0.1934	0.3684	0.3656	0.1790	0.3097	0.1099	0.2710
Encodec	2	150	0.2743	0.0627	0.2193	0.3649	0.2816	0.1900	0.3704	0.3245	0.1699	0.2960	0.1065	0.2630
DAC	1	100	0.2500	0.0713	0.2278	0.6287	0.3304	0.2502	0.3572	0.5137	0.2065	0.3350	0.1295	0.2991
WavTokenizer	1	75	**0.3255**	**0.0802**	**0.3163**	**0.6957**	**0.3417**	**0.2835**	**0.4117**	**0.5764**	**0.2550**	**0.3975**	**0.1392**	**0.3563**

Table 4: The **subjective evaluations** of various acoustic codec models for downstream speech synthesis models. **GT** denotes ground truth waveforms.

Model	Bandwidth ↓	Nq ↓	CMOS-Q↑	CMOS-P↑
GT	-	-	0.22	0.26
DAC	9.0kbps	9	-0.35	-0.29
WavTokenizer	0.9kbps	1	**0.00**	**0.00**



Token 讨论1

最近有很多工作尝试进一步增加decoder参数量，更好的encoder-decoder结构，流式的codec，更好的VQ利用率策略，更好的统一acoustic和semantic，期待获得更好的重建效果，更好的平衡CodecLLM的重建和生成，以及更好的统一理解和生成等等。

摘自《Recent Advances in Discrete Speech Tokens: A Review》

离散和连续，AR和Diffusion (Transfusion)

- Low-Bitrate Tokens（在保证参数量和RTF的前提下，语音的压缩上限是多少，是否能实现和文本的时序对齐）
- Streaming Ability and Efficiency
- Disentanglement in Acoustic Tokens
- Variable Frame Rate Tokens
- Combining Acoustic and Semantic Tokens （Acoustic能否突破蒸馏上限/（对比模型层次）在Tokenizer的层次统一）
- Paralinguistics in Semantic Tokens
- Noise Preservation vs. Noise Robustness
- Timbre Control in Token Vocoders
- Adaptivity

show-o

Thank you

季圣鹏

导师：赵洲

微信：18943038195