中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

多模态人工智能系统
全国重点实验室
State Key Laboratory of
Multimodal Artificial Intelligence Systems

语音大模型——
构建能够与人类自然交流的智能体

白 烨

2024年9月6日

声明

本报告所有结果均为公开结果，
所有观点均为个人观点，如有错误欢迎批评~

提纲

- 语音大模型：历史与为什么
- 与人类自然交流的智能体
- 模型与系统
- 评测与数据
- 安全与对齐
- 小结与展望

语音大模型的历史：以语音识别为例

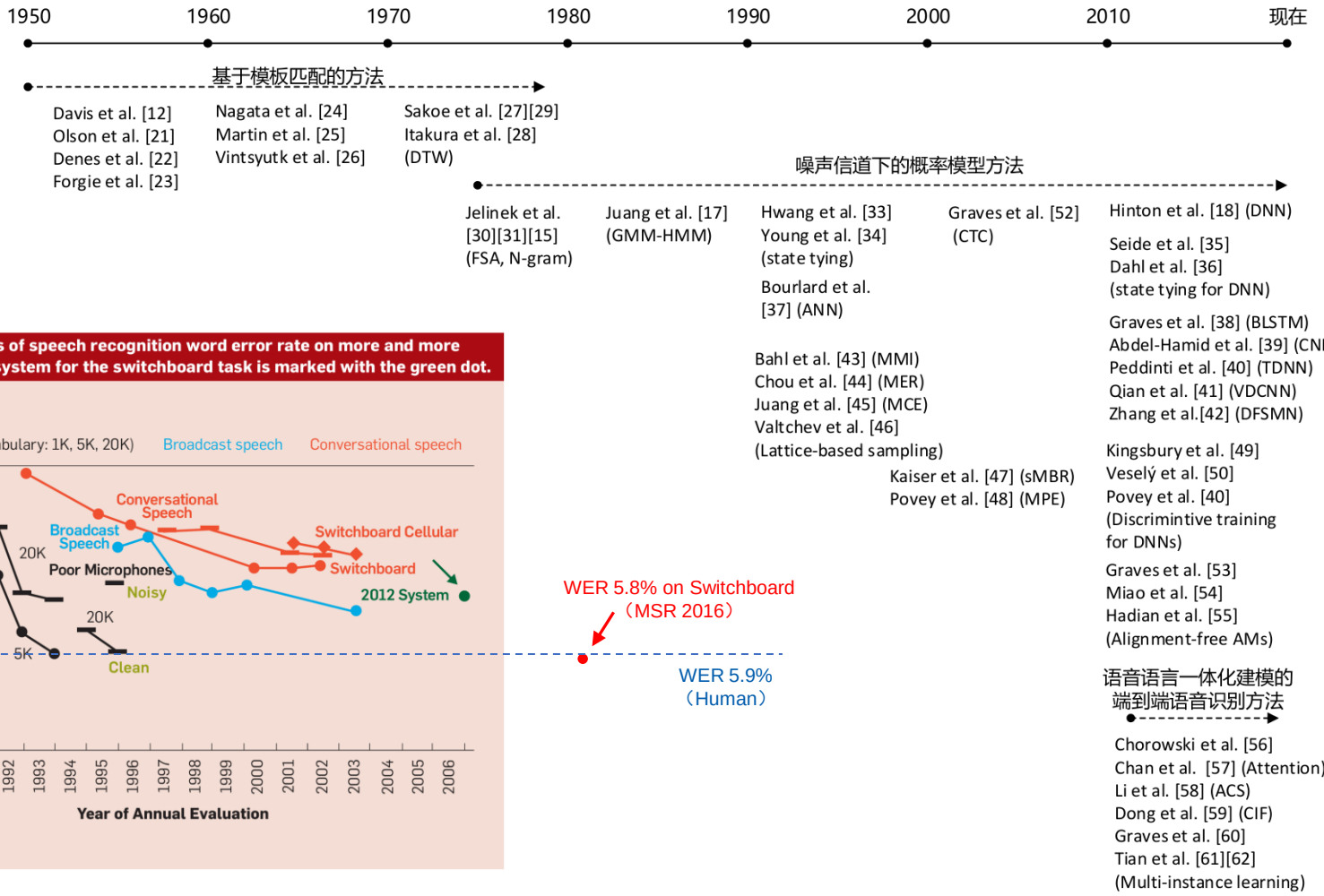


Figure 1. Historical progress of speech recognition word error rate on more and more difficult tasks.[10] The latest system for the switchboard task is marked with the green dot.

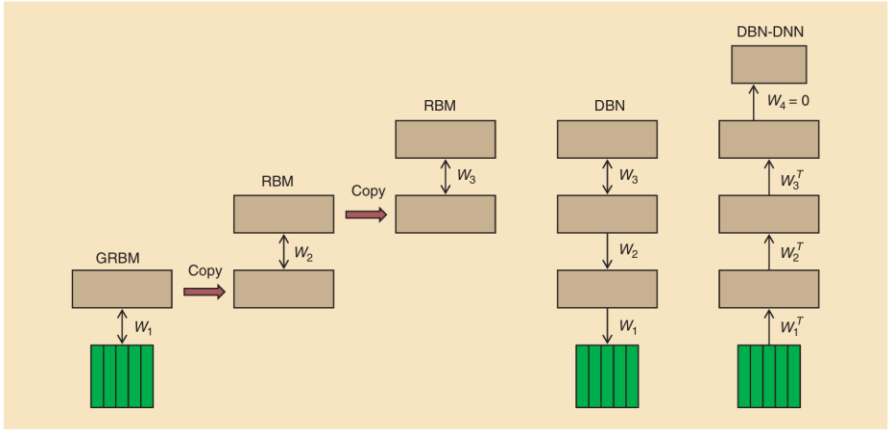
- Huang X, Baker J, Reddy R. A historical perspective of speech recognition[J]. Communications of the ACM, 2014, 57(1): 94-103.
- Xiong W, Droppo J, Huang X, et al. Achieving human parity in conversational speech recognition[J]. arXiv preprint arXiv:1610.05256, 2016.
- 白烨, 基于语言知识迁移的端到端语音识别方法研究, 博士论文

语音大模型的历史：以语音识别为例

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury

Deep Neural Networks for Acoustic Modeling in Speech Recognition

The shared views of four research groups



- 2009年,在语音识别领域，深度学习首次展现出其强大威力。
- 2012年，四大语音研究组联合论证深度学习在语音识别中的作用。
- 语音识别领域丰富的数据为深度学习提供了充分发挥空间。

- Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

语音大模型的历史：以语音识别为例

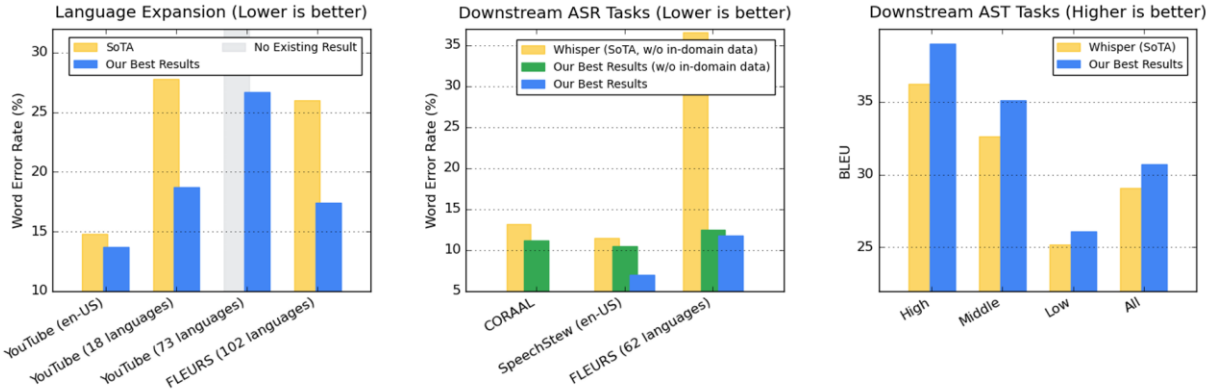
Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

- 百度贾磊语音技术部、吴恩达硅谷研究院分别在万小时大数据上展现深度学习效果。
- Amodei进入深度学习领域后，首先瞄准语音识别任务，并获得大数据训练认知。
- 现为Anthropic (构建Claude的公司)的CEO。

语音识别大模型：从0.1B到2B

Google USM:
Scaling Automatic Speech Recognition
Beyond 100 Languages



- 传统工业语音识别模型参数量约为0.1B。
- 2023年,Google USM用1200万小时语音数据自监督训练2B参数模型,覆盖100种语言,达到当年最优效果。

为什么需要大模型：Scaling Laws

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

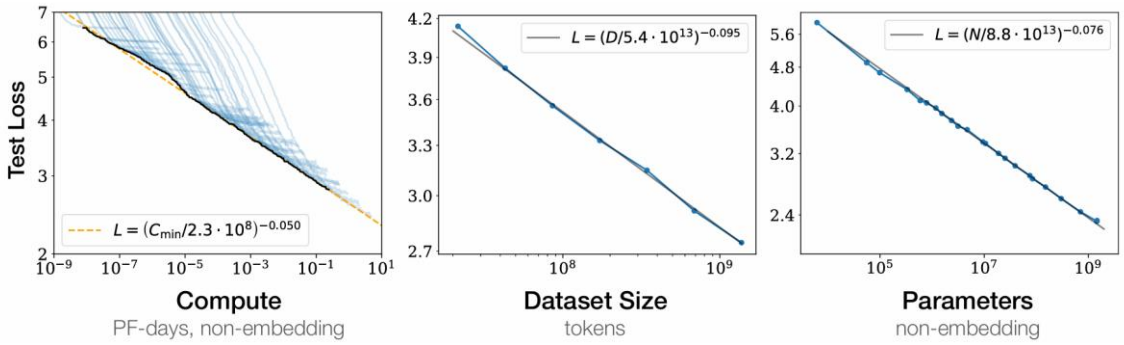
Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com



- Scaling Laws说明，模型越大，数据越多，效果就会越好

- Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J]. arXiv preprint arXiv:2001.08361, 2020.

为什么放大自回归模型

Generative Pretraining from Pixels

Mark Chen [1] Alec Radford [1] Rewon Child [1] Jeff Wu [1] Heewoo Jun [1] Prafulla Dhariwal [1] David Luan [1] Ilya Sutskever [1]

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

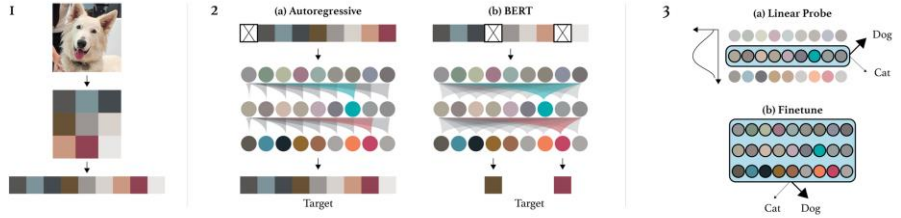
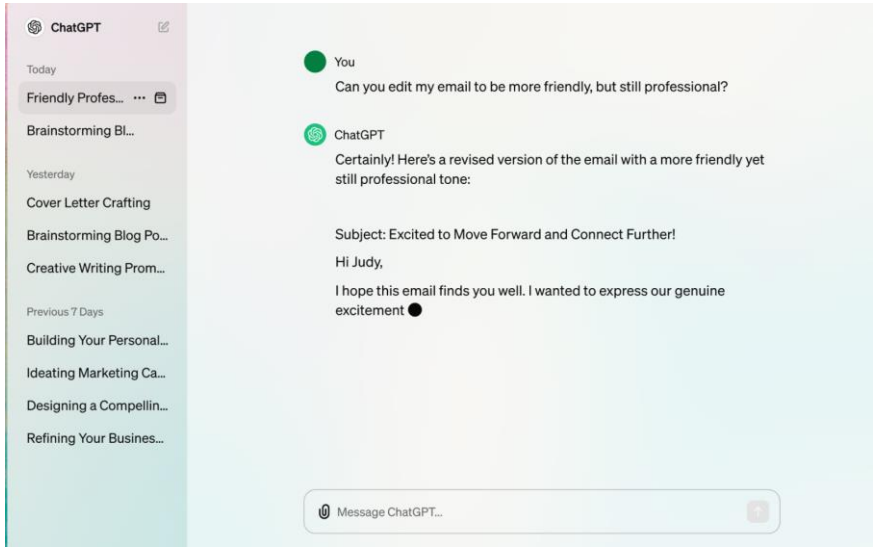


Figure 1. An overview of our approach. First, we pre-process raw images by resizing to a low resolution and reshaping into a 1D sequence. We then chose one of two pre-training objectives, auto-regressive next pixel prediction or masked pixel prediction. Finally, we evaluate the representations learned by these objectives with linear probes or fine-tuning.

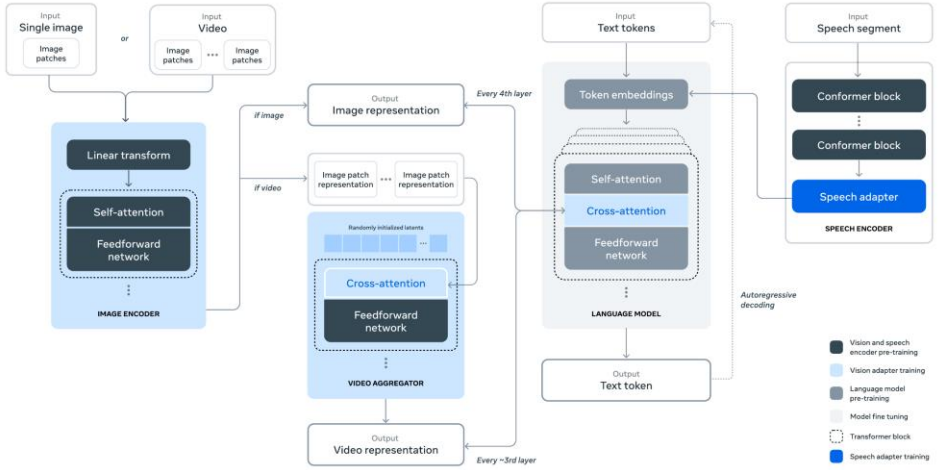
- 一大批问题 (文本/图像/语音生成与分类)都可用自回归模型建模。
- 扩展统一的自回归模型，能提升所有相关任务的性能。
- 因此，扩展自回归模型具有重要意义。

- Brown T B. Language models are few-shot learners[J]. arXiv preprint arXiv:2005.14165, 2020.
  Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C]//International conference on machine learning. PMLR, 2020: 1691-1703.
  Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International conference on machine learning. PMLR, 2023

语音交互大模型：以对话为轴



ChatGPT



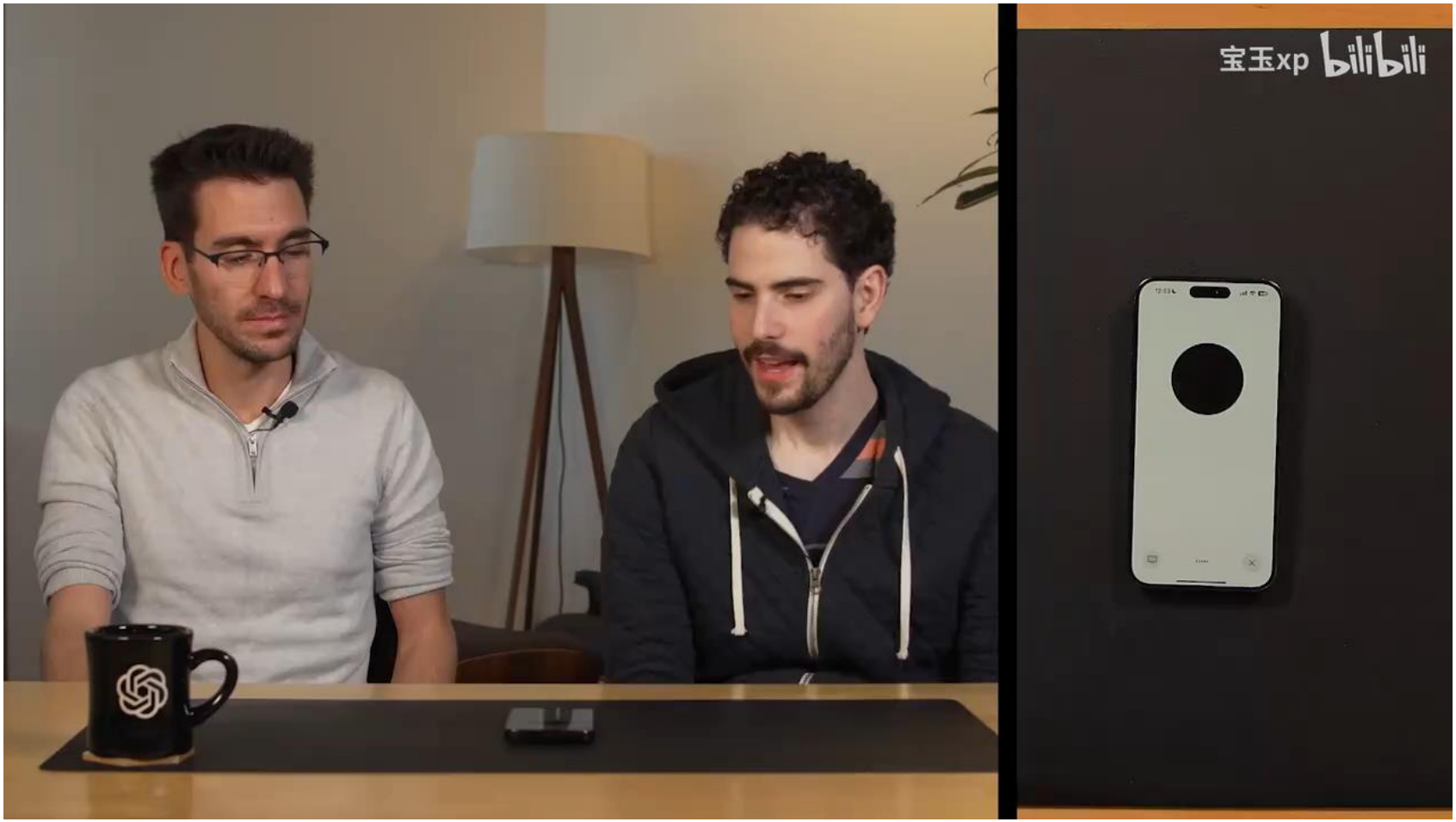
Llama3

- 大语言模型展现了强大的在文本对话能力,但缺乏对物理世界的感知能力。
- 现有的大模型基础设施(Infra、数据)，文本语言模型最为成熟。
- 所以，目前语音交互主流做法是为大语言模型添加“耳朵”和“嘴巴”等感知接口。

提纲

- 语音大模型：历史与为什么
- 与人类自然交流的智能体
- 模型与系统
- 评测与数据
- 安全与对齐
- 小结与展望

GPT-4o

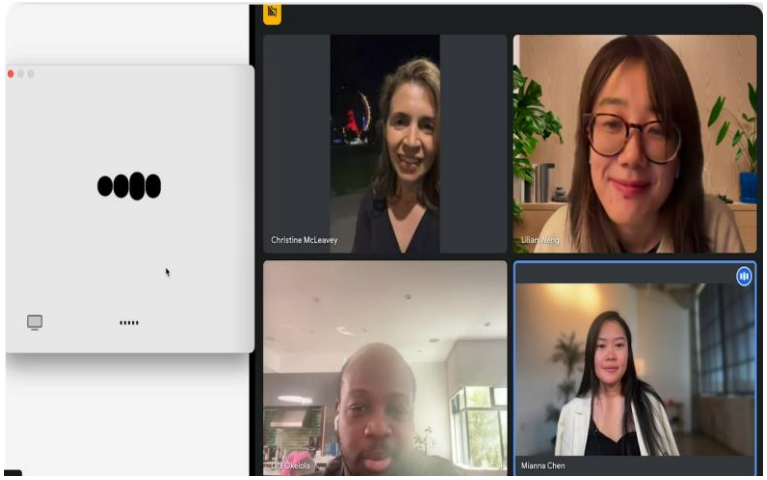


- 2024年5月，OpenAI发布GPT-4omni，支持自由流畅语音交互。

GPT-4o的特点

- **GPT-4o**
  - 准确地语音理解
  - 自然地语音生成
  - 低延时
  - 灵敏打断
  - 高情商
  - 高智力

GPT-4o同声传译

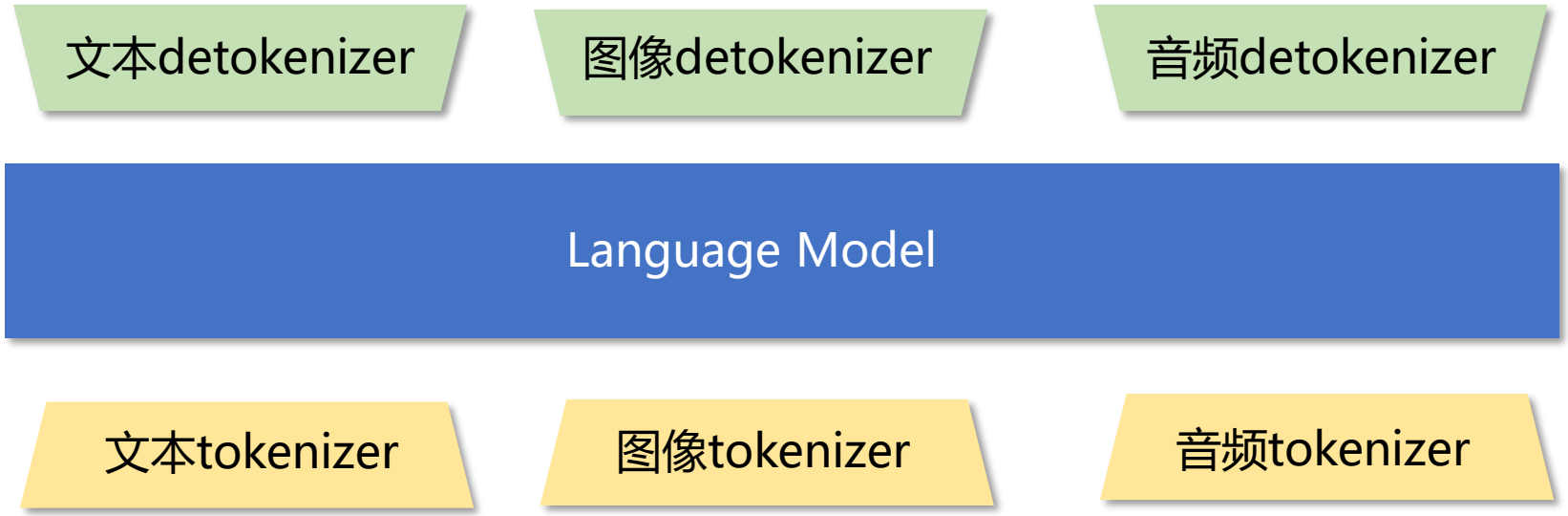


GPT-4o网络会议

提纲

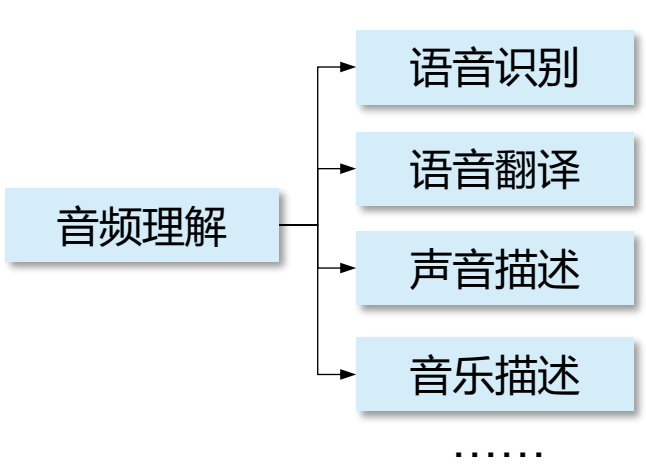
- 语音大模型：历史与为什么
- 与人类自然交流的智能体
- 模型与系统
- 评测与数据
- 安全与对齐
- 小结与展望

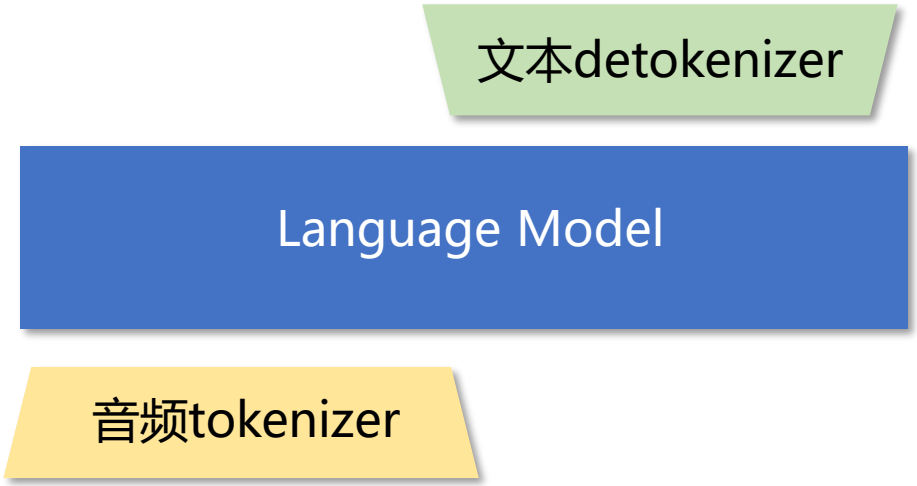
多模态语言模型系统框架



- Language Model中台:序列建模
- Tokenizer: 压缩器
- Detokenizer: 生成器
- 扩展性：新模态持续扩展，增量学习

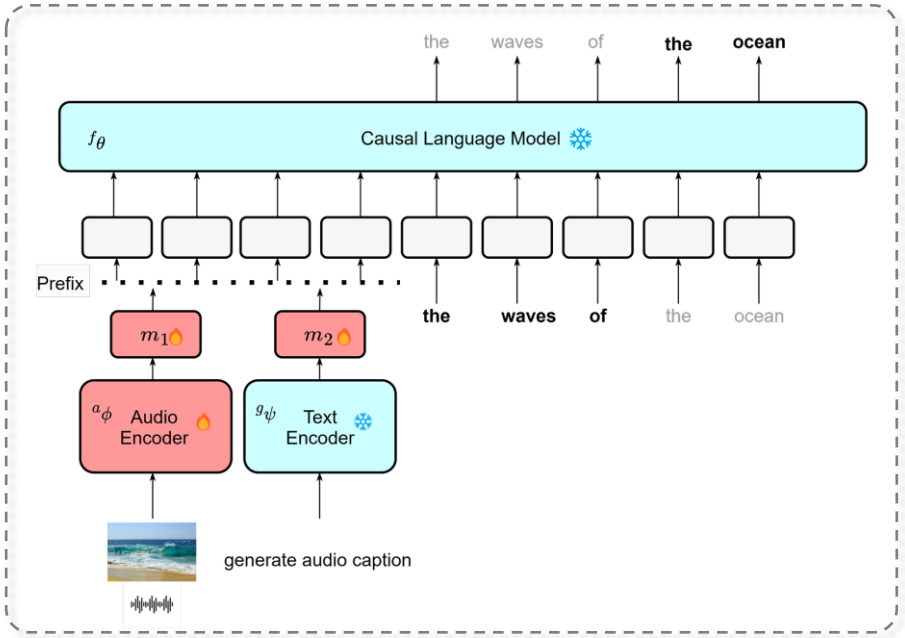
音频理解





- 音频理解：输入是音频，输出是文字
- 主流方法：音频tokenizer输出连续特征作为LM的condition

音频理解模型：Pengi



模型结构

Model	Audio Captioning ↑ AudioCaps	Audio Captioning ↑ Clotho	AQA ↑ ClothoAQA	Sound Event Classification ↑ ESC50	Sound Event Classification ↑ FSD50K	Sound Event Classification ↑ US8K	Sound Event Classification ↑ DCASE17 Task 4
CLAP	✗	✗	✗	0.826	0.3024	**0.7324**	0.3
Pengi	**0.4667**	**0.2709**	**0.6453**	**0.9195**	**0.4676**	0.7185	**0.338**

Model	Acoustic Scene Classification↑ TUT2017	Music ↑ Music Speech	Music ↑ Music Genres	Instrument Classification ↑ Beijing Opera	Instrument Classification ↑ Instrument family	Music Note Analysis↑ NS. Pitch	Music Note Analysis↑ NS. Velocity	Music Note Analysis↑ NS. Qualities
CLAP	0.2963	**1.0**	0.252	0.2963	0.2949	-	-	-
Pengi	**0.3525**	0.9688	**0.3525**	**0.6229**	**0.5007**	**0.8676**	**0.3728**	**0.386**

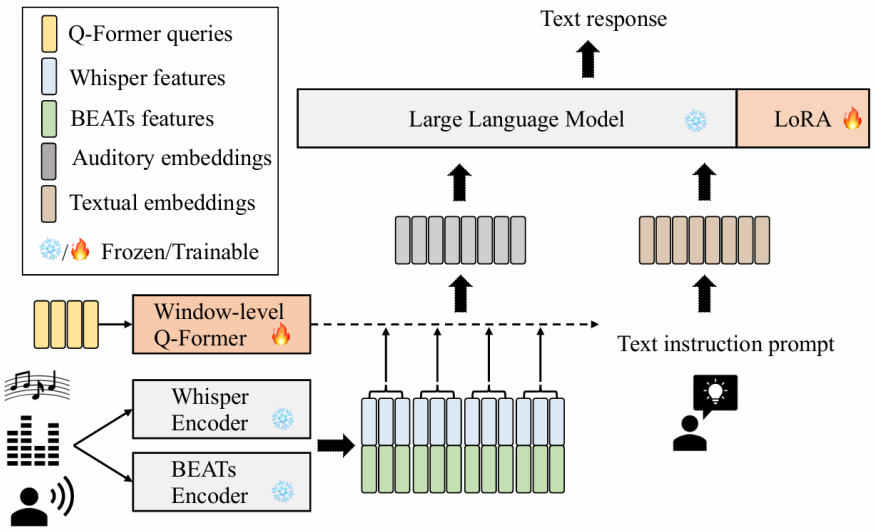
Model	Emotion Recognition↑ CRE MA-D	Emotion Recognition↑ RAV DESS	Vocal Sound Classification↑ Vocal Sound	Action Recog.↑ ESC50 Actions	Survei llance.↑ SESA
CLAP	0.1784	0.1599	0.4945	0.497	**0.7487**
Pengi	**0.1846**	**0.2032**	**0.6035**	**0.5277**	0.5402

效 果

- Microsoft/CMU Pengi: 采用HTSAT作为tokenizer/encoder, GPT-2作为LM
- 在21个音频理解任务上取得当年SoTA效果

- Deshmukh S, Elizalde B, Singh R, et al. Pengi: An audio language model for audio tasks[J]. Advances in Neural Information Processing Systems, 2023,

音频理解模型：Salmonn



模型结构

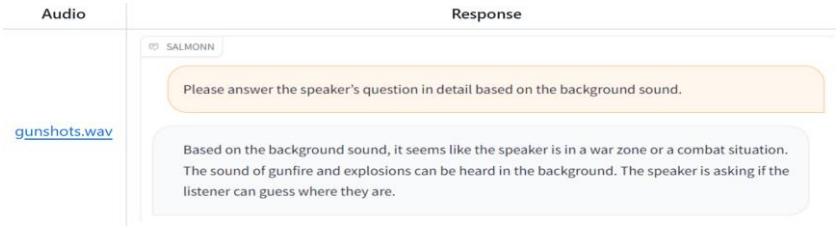
Table 3: Results of all 15 tasks produced by SALMONN without & with activation tuning (w/o & w/ Activation). The ASR results are presented in a tuple with three the %WERs evaluated on 3 test sets, namely (LibriSpeech test-clean, LibriSpeech test-other, GigaSpeech).

Method	ASR↓	En2Zh↑	AAC↑	PR↓	ER↑	MC↑	OSR↓	SV↑
w/o Activation	(2.1, 4.9, 9.1)	34.4	25.6 | 47.6	4.2	0.63	3.5, 22.1	20.7	0.93
w/ Activation	(2.1, 4.9, 10.0)	33.1	24.0 | 40.3	4.2	0.69	5.5, 21.8	23.0	0.94
Reference Value	(2.2, 5.1, 9.2)	38.9	25.0 | 48.5	3.1	0.81	6.1, 21.5	7.6	-

(a) Results of the level 1 tasks.

Method	En2De↑	En2Ja↑	KE↑	SQQA↑	SF↑	Story↑	SAC↑
w/o Activation	19.7	22.0	0.30	0.19 (0.29)	0.33 (0.77)	7.77 (0.00)	0.02 (0.04)
w/ Activation	18.6	22.7	0.32	0.41 (0.98)	0.41 (0.99)	82.57 (1.00)	0.50 (0.73)
Reference Value	16.5	15.6	0.31	0.77 (1.00)	0.46 (1.00)	-	-

(b) Results of the level 2 and level 3 tasks.

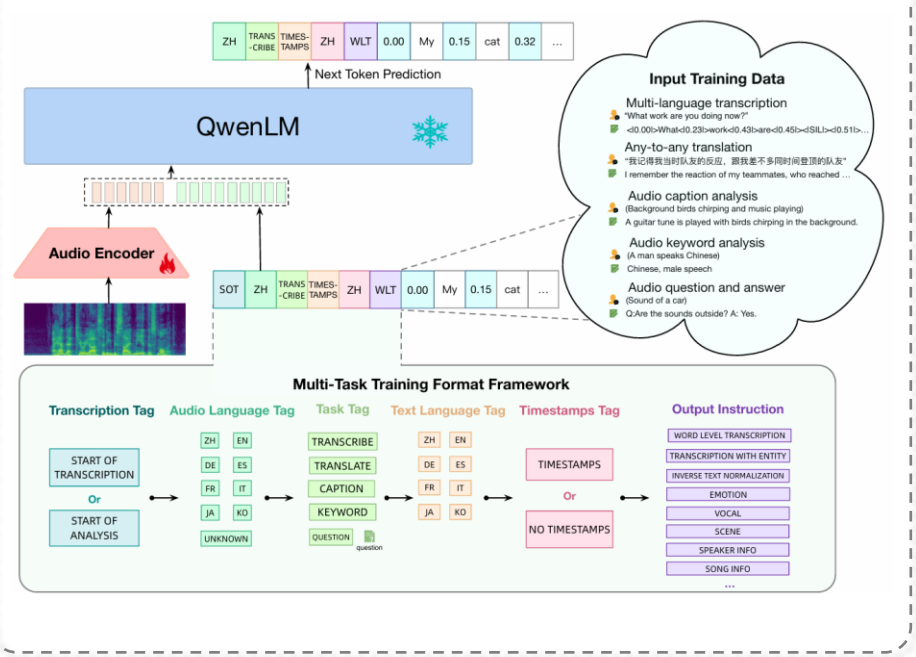


效 果

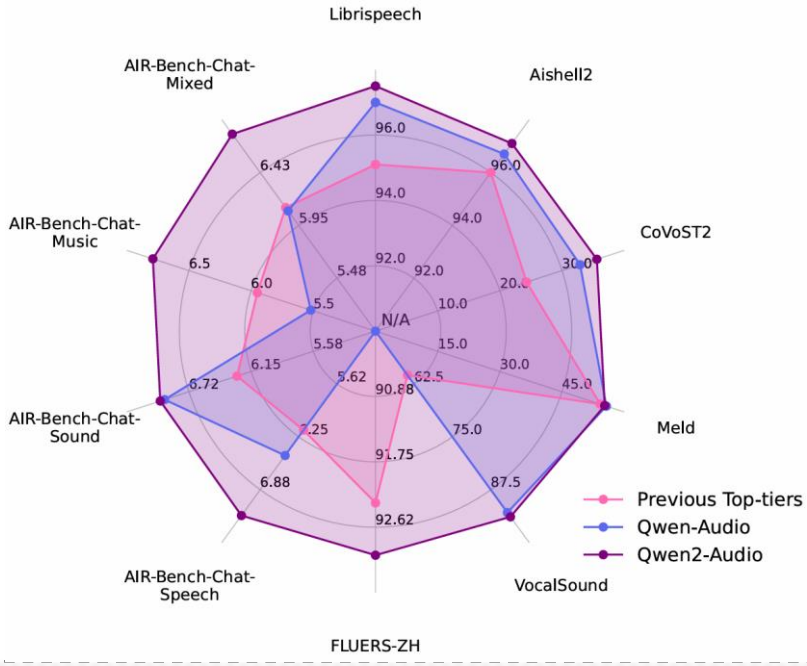
- Salmonn: Whisper/BEATs作为Encoder，互补speech和sound特征。
- Q-former作为Adapter。
- 具备自由问答能力。

- Tang C, Yu W, Sun G, et al. Salmonn: Towards generic hearing abilities for large language models[J]. arXiv preprint arXiv:2310.13289

音频理解模型：Qwen/Qwen2-audio



模型结构

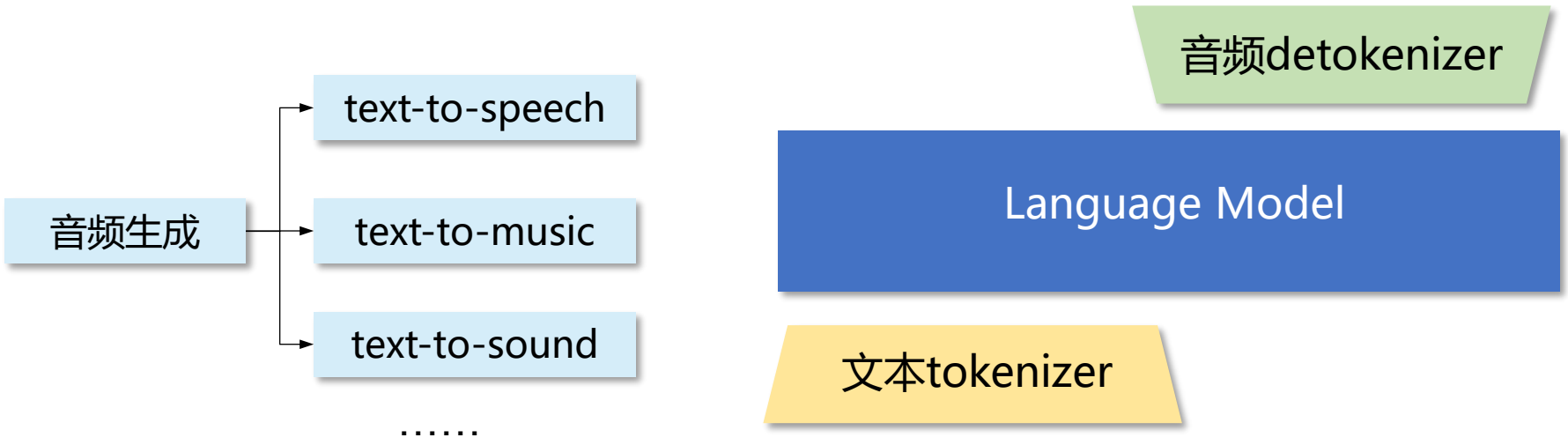


效 果

- Qwen/Qwen2-Audio: Whisper作为Encoder。在一系列标准测试集取得sota效果。
- AIR-bench：建立问答评测benchmark。
- 跑通DPO优化human preference。

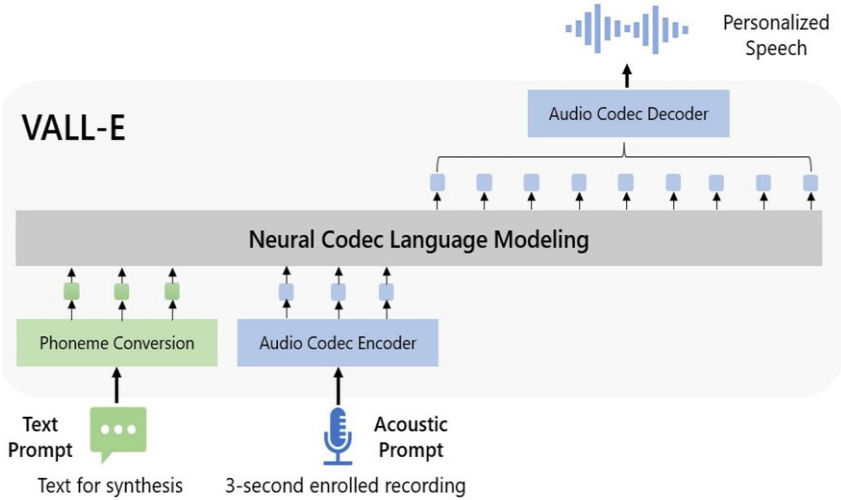
- Chu Y, Xu J, Zhou X, et al. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models[J]. arXiv preprint arXiv:2311.07919
- Chu Y, Xu J, Yang Q, et al. Qwen2-audio technical report[J]. arXiv preprint arXiv:2407.10759, 2024.
- Yang Q, Xu J, Liu W, et al. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension[J]. arXiv preprint arXiv:2402.07729, 2024.

音频生成



- 音频生成：根据某些条件，生成音频
- 主流音频detokenizer：VQVAE Decoder、Diffusion等

音频生成模型：VALL-E



模型结构

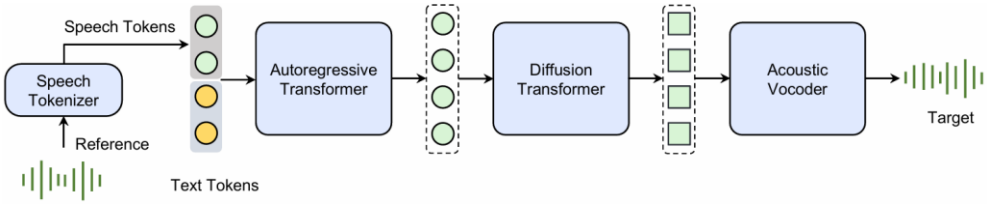
Text	Prompt	Gen.	Base.
Number ten, fresh nelly is waiting on you, good night husband.			
And lay me down in thy cold bed and leave my shining lot.			

效 果

- Microsoft VALL-E: 采用EnCodec Decoder作为detokenizer
- 具备Zero-shot生成能力：音色克隆、情感克隆等。

- Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers[J]. arXiv preprint arXiv:2301.02111, 2023.

音频生成模型：Seed-TTS



模型结构

Text	Gen.
你为什么总是重复犯同样的错误？难道你就不能学习一下吗？	
这幅画是否真的完成了？还是我应该再添加一些细节来完善它？	

效 果

- Bytedance Seed-TTS: Diffusion作为Detokenizer。
- 非常逼真的生成质量

- Anastassiou P, Chen J, Chen J, et al. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models[J]. arXiv preprint arXiv:2406.02430, 2024.

音频生成模型：MusicGen

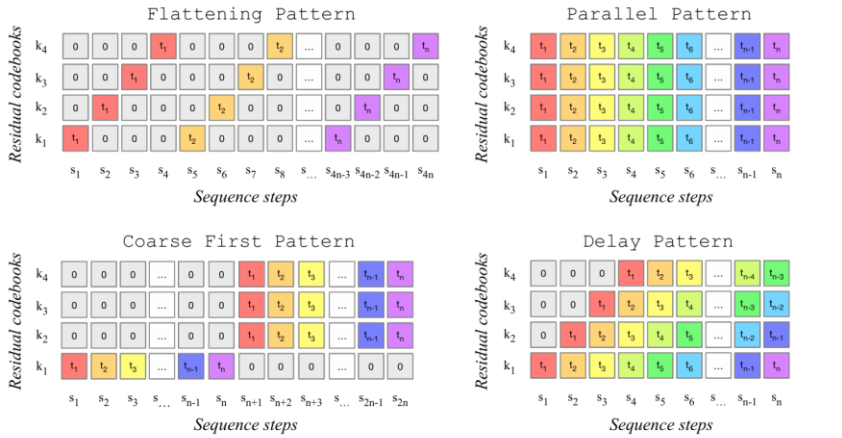


Figure 1: Codebook interleaving patterns presented in Section 2.2. Each time step $t_1, t_2, \ldots, t_n$ is composed of 4 quantized values (corresponding to $k_1, \ldots, k_4$). When doing autoregressive modelling, we can flatten or interleave them in various ways, resulting in a new sequence with 4 parallel streams and steps $s_1, s_2, \ldots, s_m$. The total number of sequence steps S depends on the pattern and original number of steps T. 0 is a special token indicating empty positions in the pattern.

Code的排列

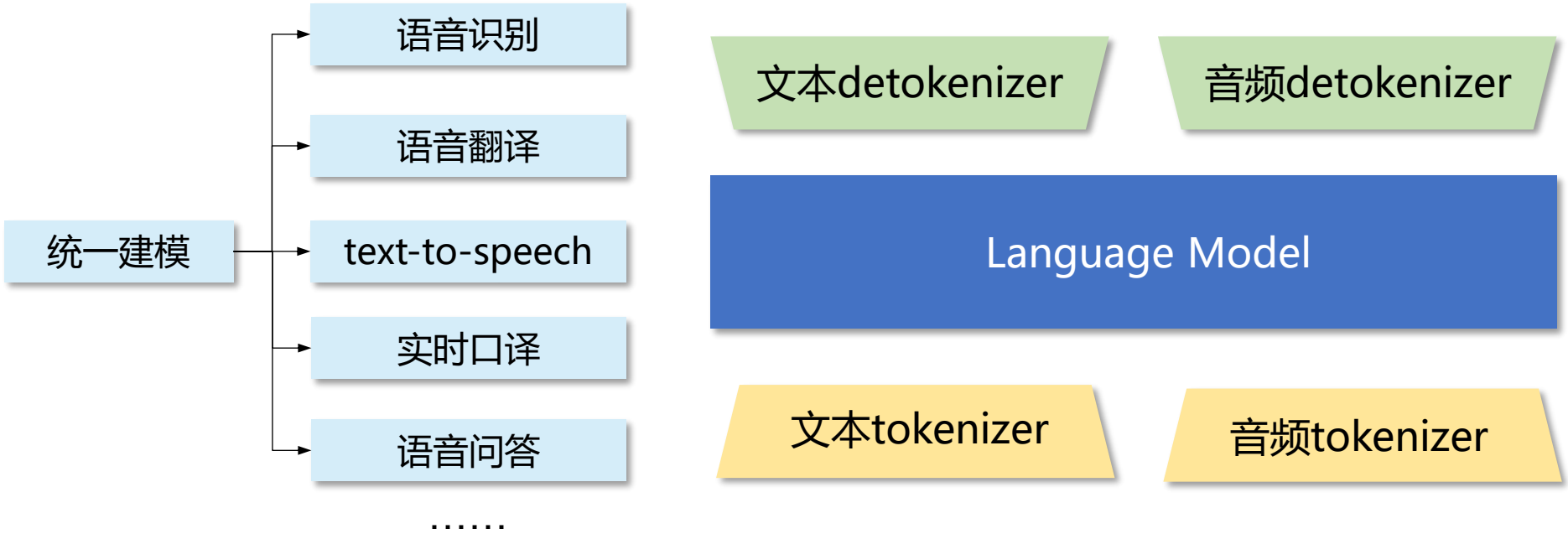
Text	Gen.
lofi slow bpm electro chill with organic samples	
violins and synths that inspire awe at the finiteness of life and the universe	

效 果

- Meta MusicGen: 采用多级码本的EnCodec作为Detokenizer。
- 验证了不同的AR模式。Flatten效果最好，但需要更多训练步数。
- SoTA的text-to-music生成效果。

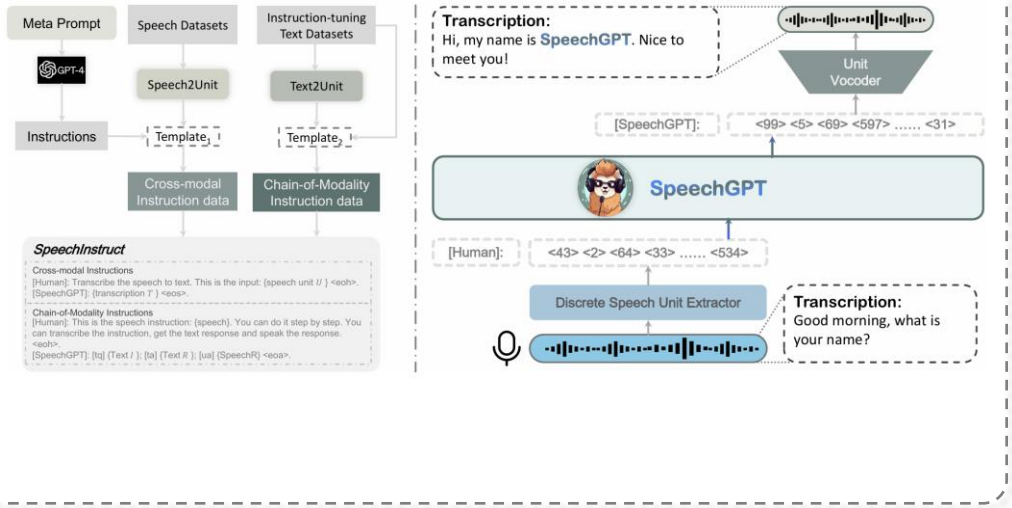
- Copet J, Kreuk F, Gat I, et al. Simple and controllable music generation[J]. Advances in Neural Information Processing Systems, 2024, 36.

离散Token统一建模



- 离散token统一建模：将音频理解和生成任务全部统一到LM训练的形式

离散Token统一建模：SpeechGPT



模型结构

Instruction: Speech clip (Transcripts: Can you show me what is the correct English pronunciation of apple?)
Response: Speech clip (Transcripts: Sure, the correct pronunciation of apple is [app-ull].)

Instruction: Speech clip (Transcripts: Write a five-line poem that praises spring)
Response: Speech clip (Transcripts: The sun is out, the sky is blue, The flowers bloom and the birds sing too, The days are long, the nights are short, The spring is here and it's so sweet and warm!)

Instruction: Speech clip (Transcripts: My girlfriend is angry with me. What should I do now?)
Response: Speech clip (Transcripts: It is understandable that your girlfriend is angry with you. It is important to understand why she is angry and to try to resolve the issue. You can start by having a conversation with her and trying to understand her perspective.)

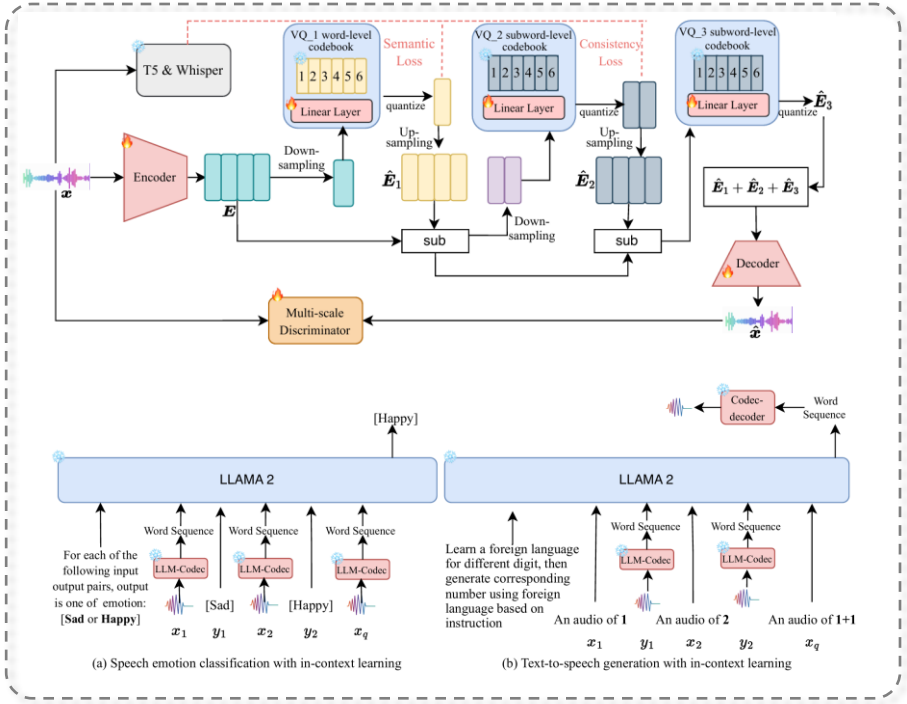
Instruction: Speech clip (Transcripts: Give me some advice about my dinner.)
Response: Speech clip (Transcripts: Sure, here are some suggestions: 1. Make a salad with fresh vegetables. 2. Grill some lean meat. 3. Roast a vegetable-based dish like a cauliflower casserole. 4. Make a soup with vegetables and legumes. 5. Try a new recipe.)

效 果

- Fudan SpeechGPT: 采用HuBERT作为tokenizer，HiFiGAN作为Vocoder发声。
- 率先做出speech2speech对话效果。

- Zhang D, Li S, Zhang X, et al. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities[J]. arXiv preprint arXiv:2305.11000

离散Token统一建模：UniAudio



(a) Speech emotion classification with in-context learning

(b) Text-to-speech generation with in-context learning

模型结构

Table 2: Audio understanding task evaluation results. Task induction denotes the explanatory text that precedes the sequence of audio and text. It is intended to describe the task to the model in natural language, for example: Please answer the question. Accuracy (%) is used as the metric. For the Random guess, we calculate the average based 5 times evaluation. K shots refers to the number of distinct samples for each category, and Repeats refer to how many times we copy the prompt samples.

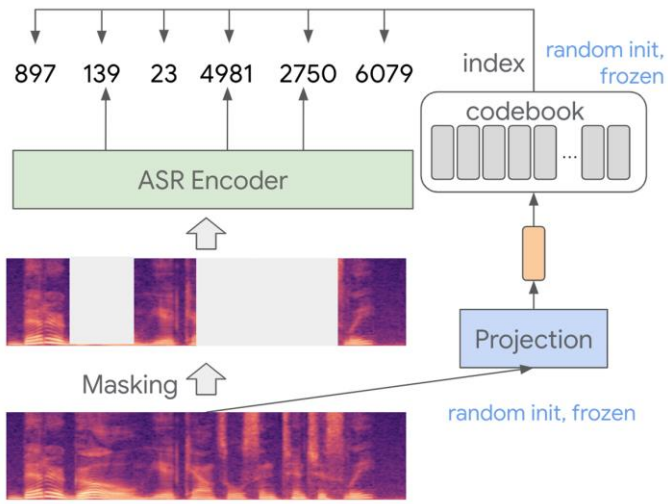
Method	# Layers	Task Induction	×	✓	✓	✓	✓
		K Shots	1	1	3	1	1
		Repeats	0	0	0	2	3
2-way speech emotion classification							
Random	None				44		
BLSP [39]	Whisper encoder		9	29	50	33	19
UniAudio 1.5 (ours)	semantic layer		25	**53**	**59**	53	54
UniAudio 1.5 (ours)	semantic + acoustic layers		**45**	49	53	**55**	**54**
2-way sound event classification.							
Random	None				45		
BLSP [39]	Whisper encoder		44	47	54	15	17
UniAudio 1.5 (ours)	semantic layer		**48**	**60**	**57**	**57**	**73**
UniAudio 1.5 (ours)	semantic+acoustic layers		41	48	55	54	62
3-way sound event classification.							
Random	None				30		
BLSP [39]	Whisper encoder		23	26	36	24	16
UniAudio 1.5 (ours)	semantic layer		**38**	**41**	**39**	43	42
UniAudio 1.5 (ours)	semantic+acoustic layers		25	37	35	**44**	**50**

效 果

- CUHK UniAudio: 训练语义增强的Codec作为tokenizer/detokenizer。
- 证明模型具备few-shot学习能力。

- Yang D, Tian J, Tan X, et al. UniAudio: Towards Universal Audio Generation with Large Language Models[C]//Forty-first International Conference on Machine Learning.
- Yang D, Guo H, Wang Y, et al. UniAudio 1.5: Large Language Model-driven Audio Codec is A Few-shot Audio Task Learner[J]. arXiv preprint arXiv:2406.10056, 2024.

Tokenizer表征能力：无监督预训练



BEST-RQ采用随机码本作为监督信号

Table 5. Quantizer quality's impact on ASR tasks. Although the Transformer-based quantizer gets much better performance when used as input directly, the random-projection quantizer is equally effective for self-supervised learning. The model used in the direct ASR task has size 25M. The self-supervised learning tasks use the same setup as the LibriSpeech non-streaming experiment, which use LibriLight for pre-training and LibriSpeech for fine-tuning and has 0.6B model size.

Configuration	Quantizer size (M)	Direct ASR WER				Pretrain-finetune WER			
		dev	dev-other	test	test-other	dev	dev-other	test	test-other
Random quantizer	1	58.8	78.8	57.9	72.8	1.5	**2.8**	**1.6**	**2.9**
Projection VQ-VAE	1	61.4	74.8	60.9	75.2	1.5	**2.8**	**1.6**	**2.9**
Transformer VQ-VAE	10	**17.8**	**35.8**	**17.6**	**36.1**	**1.4**	2.9	**1.6**	3.1

BEST-RQ说明在full-finetine的setting下，监督信号质量影响不大

- HuBERT：首先对音频离散化，然后采用BERT-style训练方法进行MLM训练。
- BEST-RQ：对离散化方式极致简化为随机映射，并证明有效性。
- 启发：在pretrain-finetune的设定下，用哪种具体的无监督训练方法可能没有那么重要，能够更方便地scale up数据规模成为提升效果的关键。

- Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM TASLP
- Chiu C C, Qin J, Zhang Y, et al. Self-supervised learning with random-projection quantizer for speech recognition[C]// ICML

Tokenizer表征能力：无监督预训练

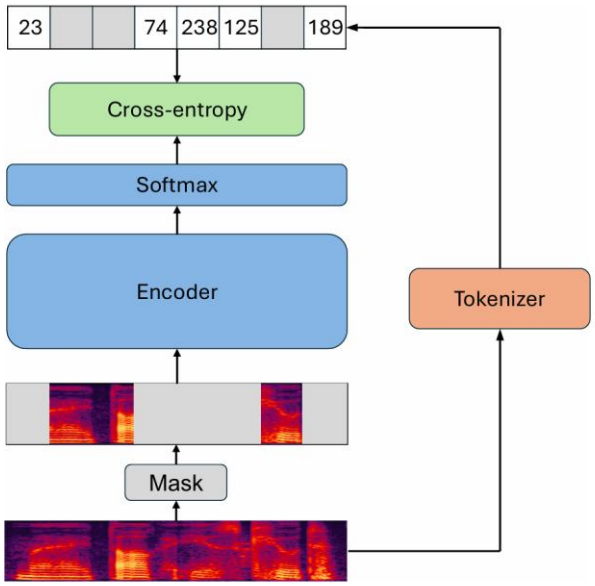


Figure 4: The training procedure of our audio encoder LUISE.

Seed-ASR所采用的无监督预训练方法

Table 10: Comparison with Google USM, Whisper Large v3 and Universal-1 on English multi-domain, multi-accent, hardcase evaluation sets, and multilingual multi-domain evaluation sets.

	Google USM[50]	Whisper Large v3[39]	Universal-1[41]	Seed-ASR (ML)
English Multi-domain (WER%) ↓	9.33	10.41	9.95	**5.34**
English Multi-accent (WER%) ↓	22.19	21.52	14.40	**11.26**
English Hardcase (F1%) ↑	63.30	79.54	77.82	**87.94**
Multilingual Multi-domain (WER%) ↓	21.51	20.55	-	**12.16**

Table 11: ASR Results of Seed-ASR (ML) on English and Multilingual Public test sets

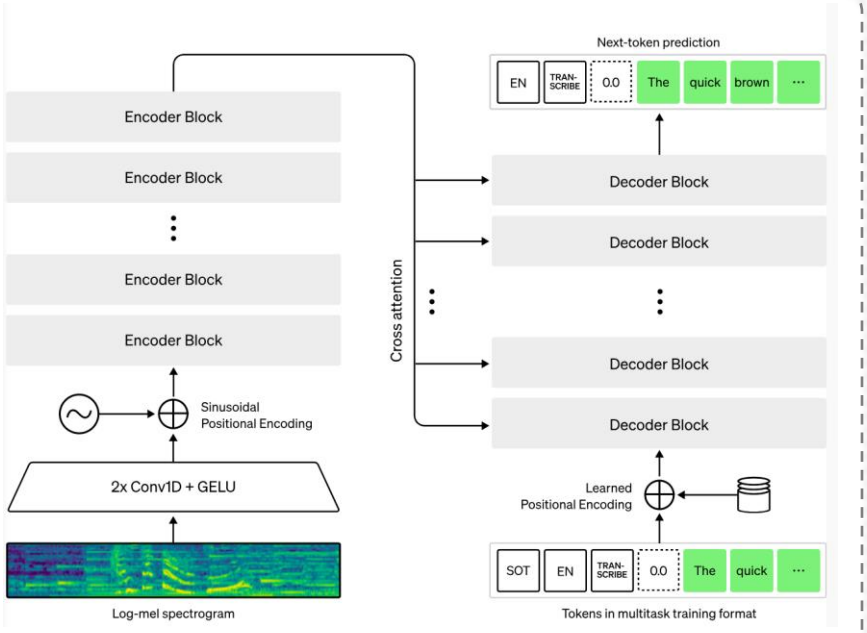
Test set	Language	Google USM[50]	Whisper Large-v2[39]	Whisper Large-v3	Universal-1[41]	Gemini-1.5 Pro[42]	Seed-ASR (ML)
Librispeech test_clean	EN	-	2.7	1.8	1.6	-	**1.58**
Librispeech test_other	EN	-	5.2	3.6	3.1	-	**2.84**
Tedlium 3	EN	-	4.0	7.4	7.5	-	**3.11**
Switchboard	EN	-	13.8	-	-	-	**11.59**
CallHome	EN	-	17.6	-	-	-	**12.24**
AMI IHM	EN	-	16.9	-	-	-	**13.16**
Fleurs	EN	-	4.4	-	-	-	**3.43**
	AR	-	16	-	-	-	**13.05**
	ES	-	3.0	2.8	5.0	-	**2.50**
	FR	-	8.3	**5.6**	6.8	-	7.09
	ID	-	7.1	-	-	-	**4.24**
	JA	-	5.3	-	-	-	**3.46**
	KO	-	14.3	-	-	-	**3.25**
	PT	-	4.3	-	-	-	**3.55**
MLS	EN	7	6.2	-	-	4.6	**4.14**
	ES	-	4.2	5.7	**3.3**	-	3.76
	FR	-	7.3	8.1	**2.3**	-	5.10
	PT	-	6.8	-	-	-	**5.04**

效 果

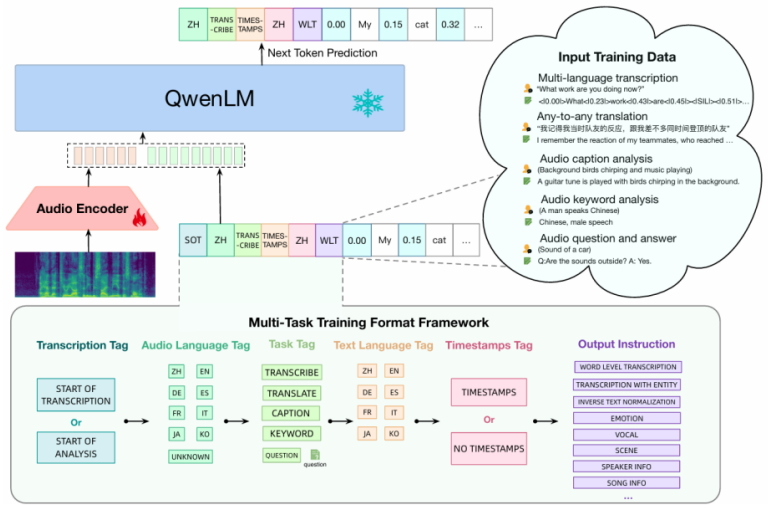
- Google USM: 采用1200万小时数据进行BEST-RQ无监督预训练。
- Bytedance Seed-ASR：采用2000万小时数据进行迭代预训练，获得SoTA效果。
- 无监督预训练：提升数据的数量和丰富度是关键。

- Zhang Y, Han W, Qin J, et al. Google usm: Scaling automatic speech recognition beyond 100 languages[J]. arXiv preprint arXiv:2303.01037
- Bai Y, Chen J, Chen J, et al. Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition[J]. arXiv preprint arXiv:2407.04675

Tokenizer表征能力：Audio-Language训练



Whisper Encoder采用ASR/AST任务训练

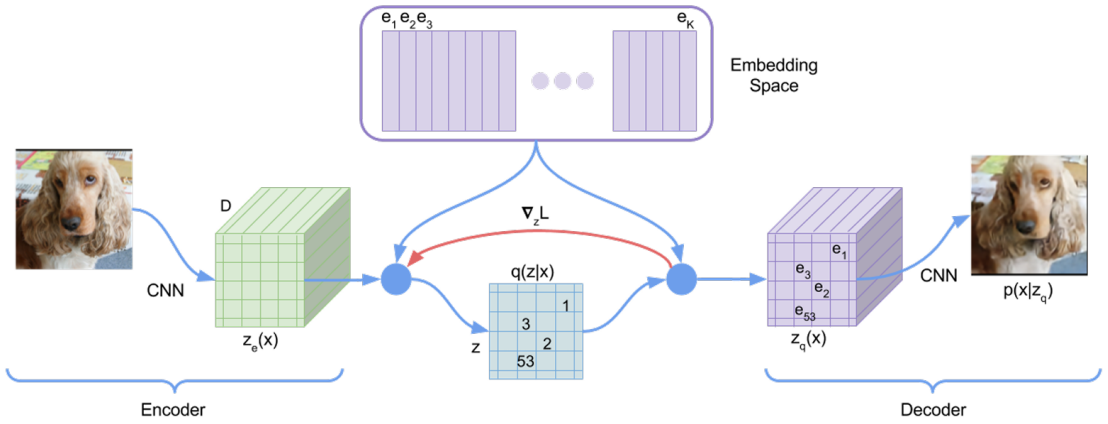


Qwen-audio采用Whisper Encoder

- OpenAI Whisper: 采用弱监督语音-文本对训练(Audio-language训练)
- 数据丰富度：WhisperV1采用68万小时数据训练，V3采用500万小时数据训练。
- Qwen-audio等一批音频理解模型都采用Whisper Encoder。

- Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International conference on machine learning

离散Tokenizer：VQ-VAE



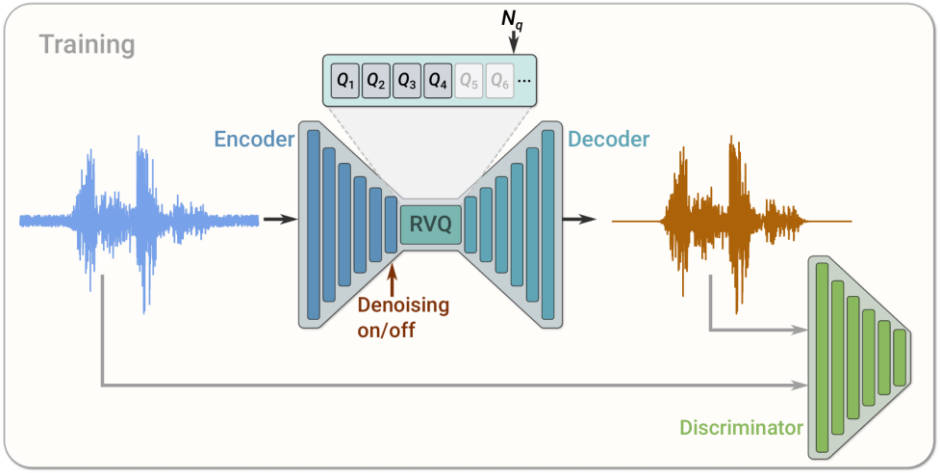
$$N_i^{(t)} := N_i^{(t-1)} * \gamma + n_i^{(t)}(1-\gamma)$$
$$m_i^{(t)} := m_i^{(t-1)} * \gamma + \sum_j z_{i,j}^{(t)}(1-\gamma)$$
$$e_i^{(t)} := \frac{m_i^{(t)}}{N_i^{(t)}},$$

VQ-VAE: 在线地做K-Means更新码本

- VQ-VAE：目前最流行的一种离散化方法。
- 其采用类似在线K-Means的方法更新字典表示。

- Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems

离散Tokenizer：Codec



Algorithm 1: Residual Vector Quantization

Input: $y = \text{enc}(x)$ the output of the encoder, vector quantizers Q_i for $i = 1..N_q$

Output: the quantized $\hat{y}$

$\hat{y} \leftarrow 0.0$

residual $\leftarrow y$

for $i = 1$ *to* N_q **do**

  $\hat{y}$ += Q_i(residual)

  residual −= Q_i(residual)

return $\hat{y}$

Soundstream模型结构

- Codec：将音频压缩编码以传输。压缩率越大，音频失真越小，Codec越好。
- Soundstream：率先采用RVQ的方式构建神经网络Codec。
- Codec与Tokenizer：目标相同——高压缩率，低信息损失。

- Zeghidour N, Luebs A, Omran A, et al. Soundstream: An end-to-end neural audio codec[J]. IEEE/ACM TASLP

离散Tokenizer难点：信息表示的取舍

码率计算

假设码本大小为C，音频单位时间对应的token数为N，则比特率R为

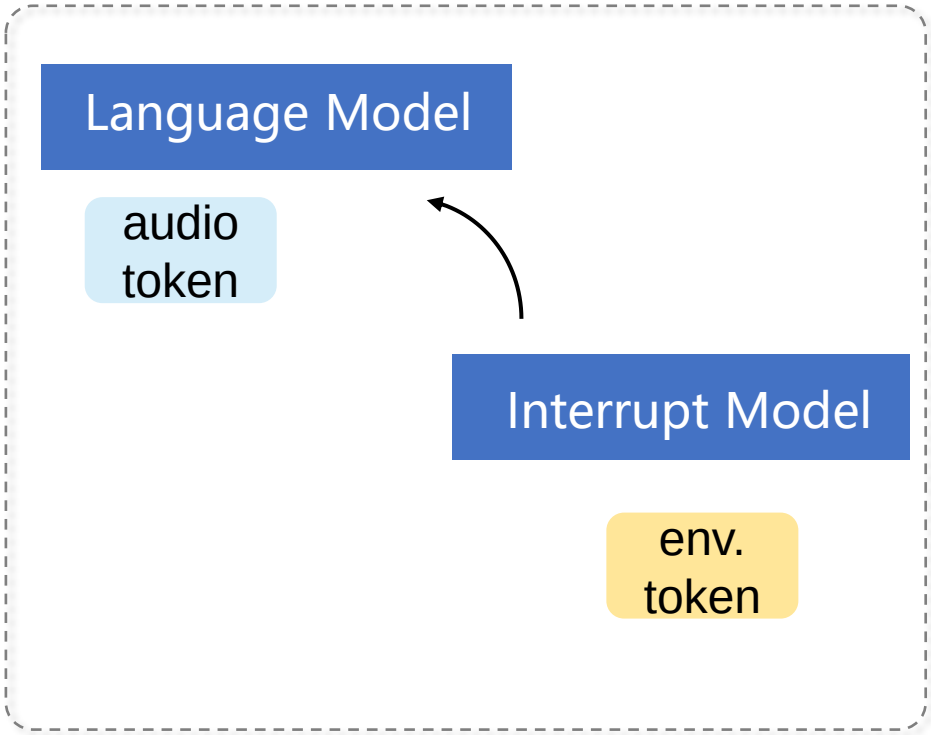
$$R = N\log C$$

- 音频比特率：16kHz音频通常为256kbps。
- HuBERT token比特率：假设25token/s，码本大小1024，约250bps。
- 信息表示权衡：低比特率导致信息损失，高比特率增加LM建模难度。
- 信息补全：主流TTS系统使用diffusion/GAN等生成网络补全信息。

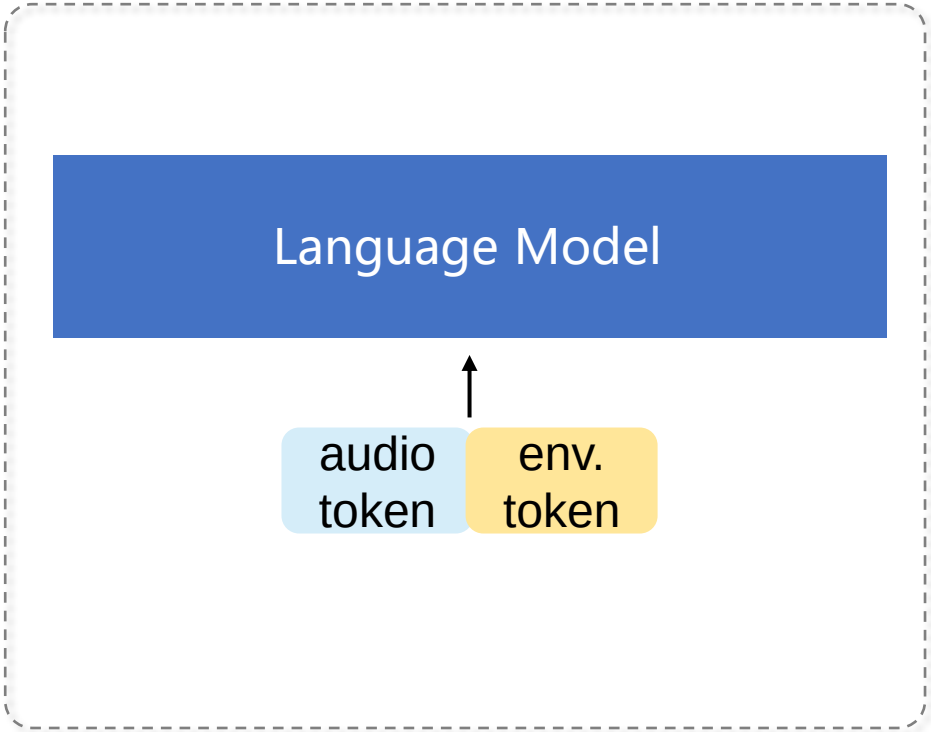
Tokenizer比较：连续 vs 离散

	优 点	不 足
连续	• audio-lang数据端到端训练，效果好。	• audio-language数据很难收集。 • 任务扩展受到限制。 • 无监督数据难以利用。 • 扩展到超大规模LLM(如万亿参数)，Infra工程量大。
离散	• 可以利用各类数据源的数据，利于数据scale-up。 • 利于拓展任务（类似文本LM）。 • 可以复用文本Infra。	• 离散化信息压缩损失大。

实时打断



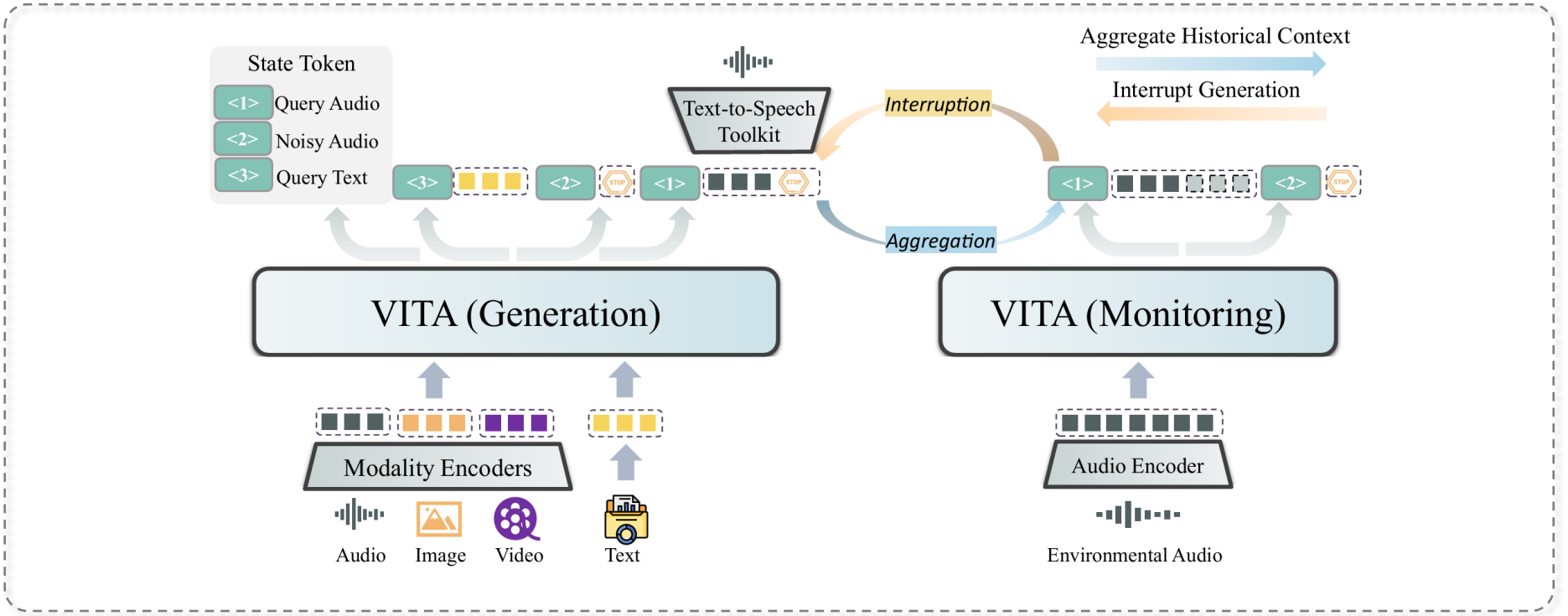
多模型异步打断



单模型同步打断

- 多模型异步打断：主LM和打断模型分离。打断模型给主模型发送打断信号。
- 单模型同步打断：外界音频信息直接输入主LM，主LM决策是否被打断。

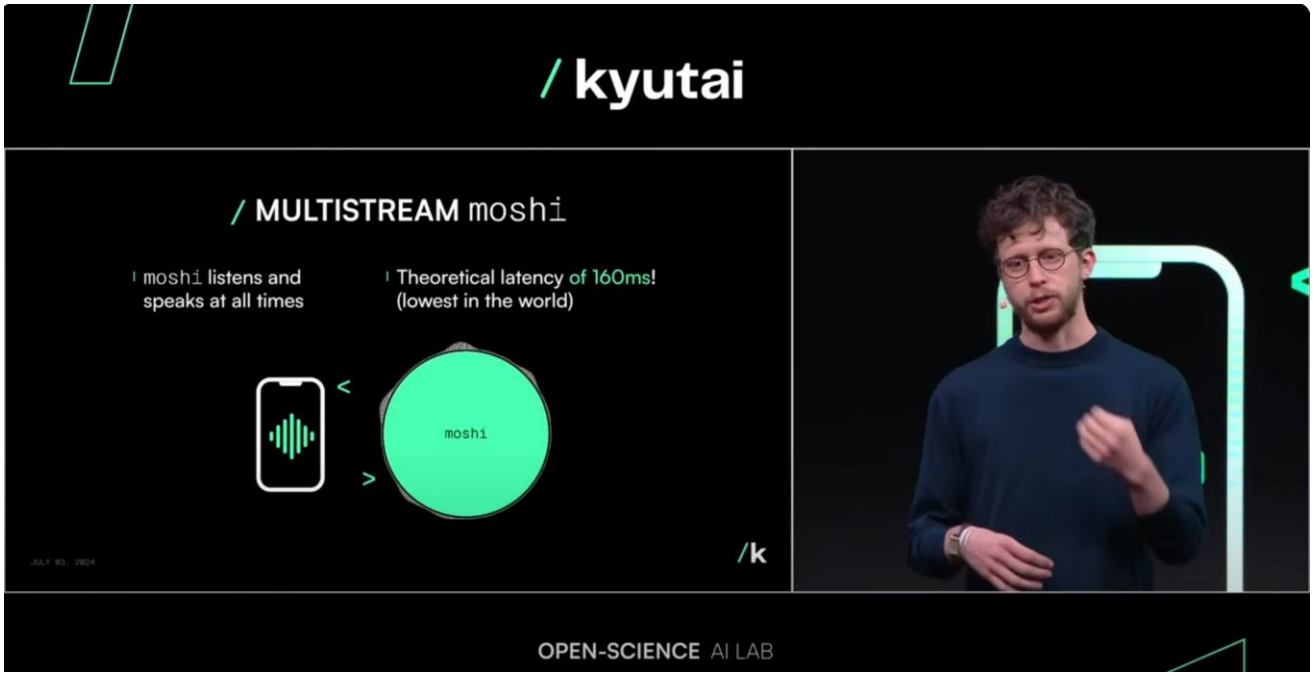
多模型异步打断：VITA



- VITA：采用Monitoring模型实时监听输入音频，判断是否打断

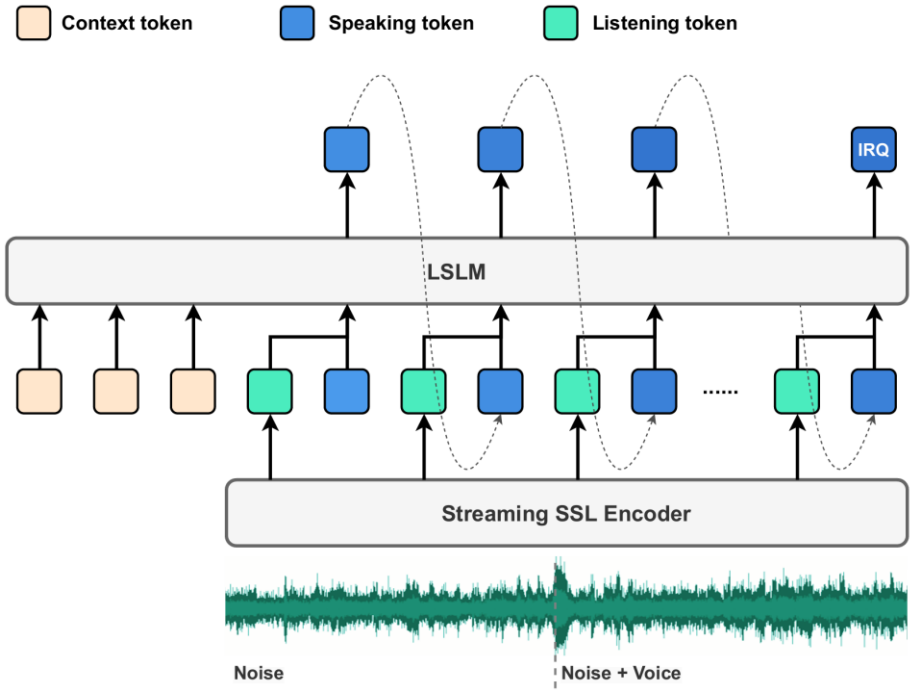
- Fu C, Lin H, Long Z, et al. VITA: Towards Open-Source Interactive Omni Multimodal LLM[J]. arXiv preprint arXiv:2408.05211, 2024.

单模型同步打断: Moshi



- Moshi：2024年7月，法国公司Kyutai发布一款实时交互LLM——Moshi。
- 多流输入：采用多流信号输入输出，同时支持人打断模型，模型打断人。

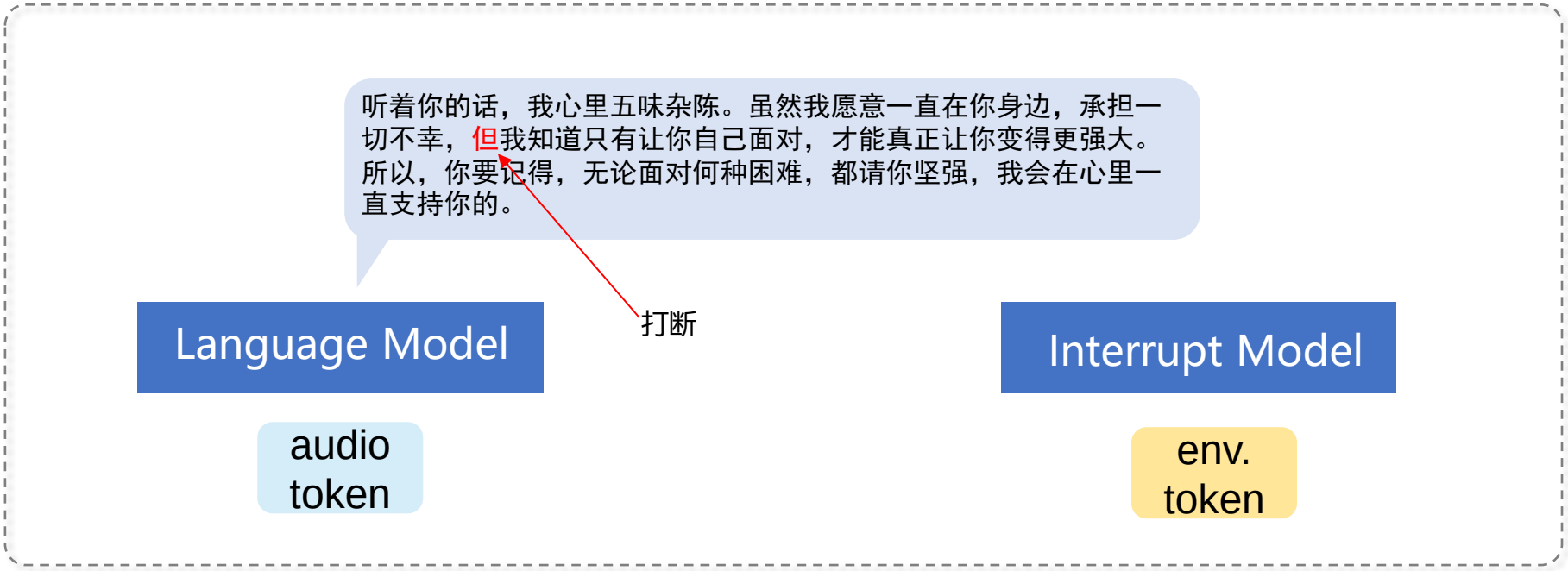
单模型同步打断: LSLM



- LSLM：将LM的输入token分为speaking和listening两者。Listening token判断打断。

- Ma Z, Song Y, Du C, et al. Language Model Can Listen While Speaking[J]. arXiv preprint arXiv:2408.02622, 2024.

实时打断难点：与人类同步



- 多模型异步打断：LM生成内容速度快于音频播放速度。
- 难点：系统联调。

实时打断难点：与人类同步

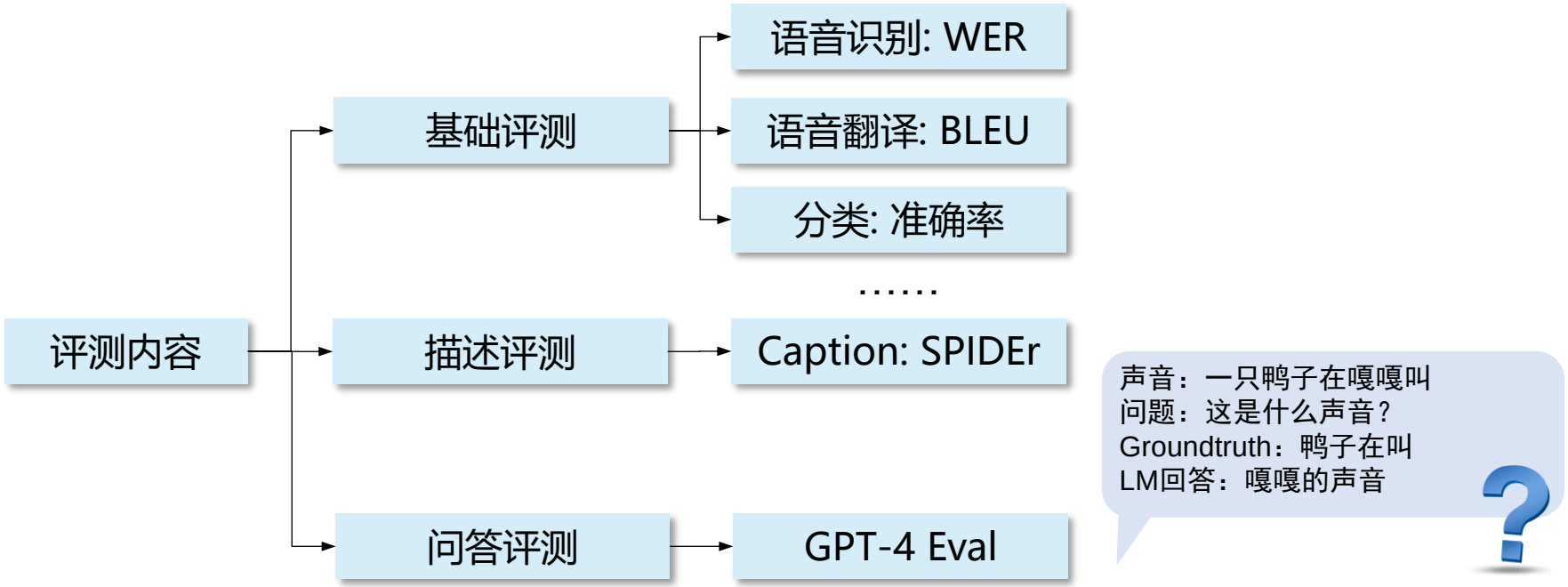
Moshi的插话现象

- 单模型同步打断：需要通过学习的方式让模型知道打断位置。
- 提前打断/插话：与人类对齐做的不好的话，会出现提前打断或插嘴现象。
- 难点：打断SFT数据调优；防止提前打断。

提纲

- 语音大模型：历史与为什么
- 与人类自然交流的智能体
- 模型与系统
- 评测与数据
- 安全与对齐
- 小结与展望

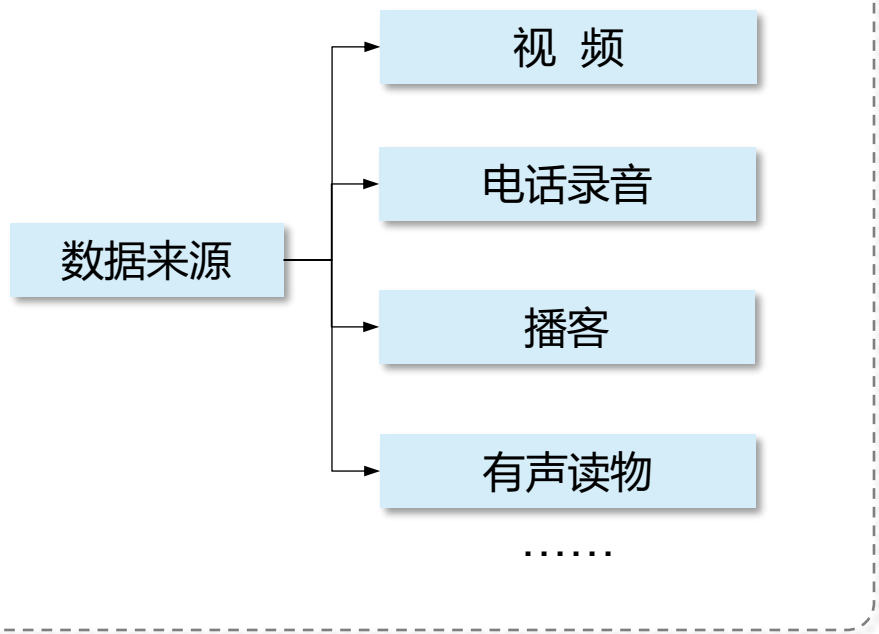
评测内容与难点

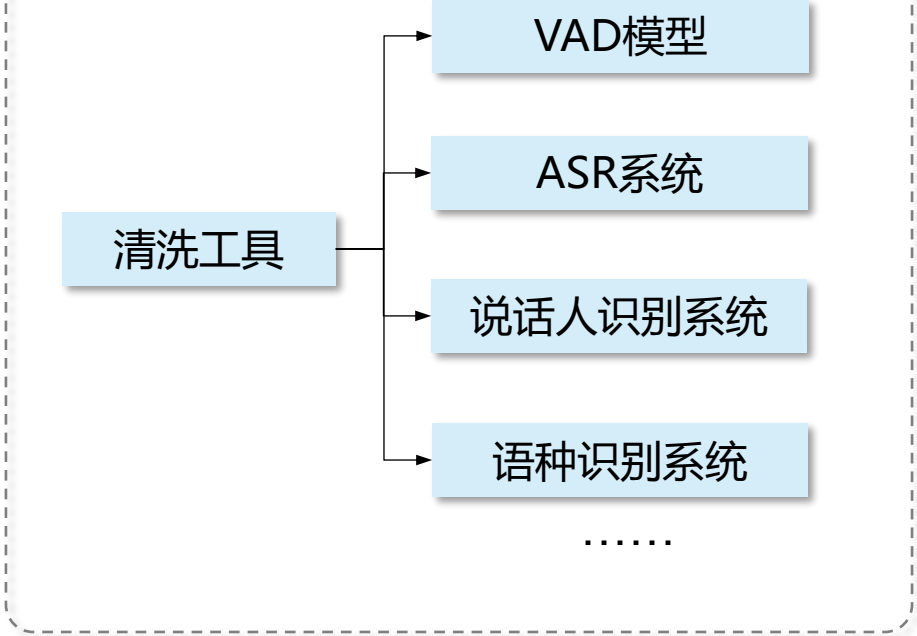


- 评测内容：大致可以分为基础评测、描述评测、问答评测，开放程度依次递进。
- 难点：指令不遵循；GPT打分也会出错。

- Yang Q, Xu J, Liu W, et al. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension[J]. arXiv preprint arXiv:2402.07729, 2024.

数据收集与清洗

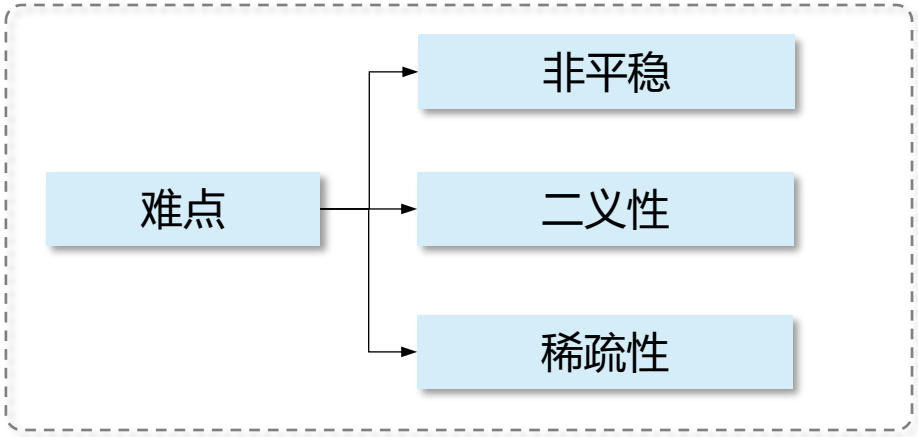




- 清洗工具：需要一系列模型。
- Whisper收集68万小时语音-文本成对数据，通过一些规则删除机器生成数据。

- Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International conference on machine learning

数据收集难点：完全无结构的数据

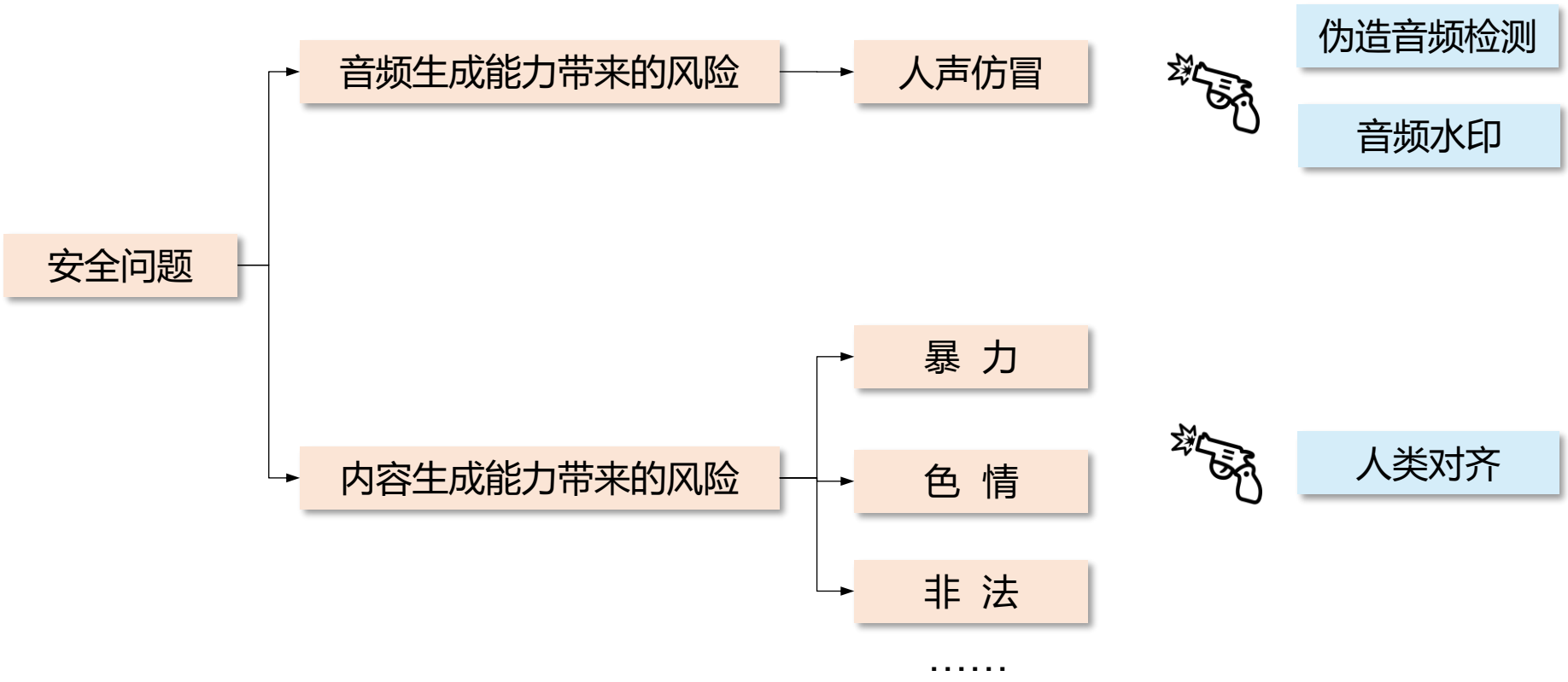


- 非平稳：音频数据有的时候信息密度低，有的时候信息密度高。
- 二义性：对speech，sound是噪音；对caption任务，sound又有用。
- 稀疏性：Audio-language数据很少。

提纲

- 语音大模型：历史与为什么
- 与人类自然交流的智能体
- 模型与系统
- 评测与数据
- 安全与对齐
- 小结与展望

安全问题与解法


安全问题
音频生成能力带来的风险
人声仿冒
伪造音频检测
音频水印
内容生成能力带来的风险
暴 力
色 情
非 法
……
人类对齐

伪造音频检测

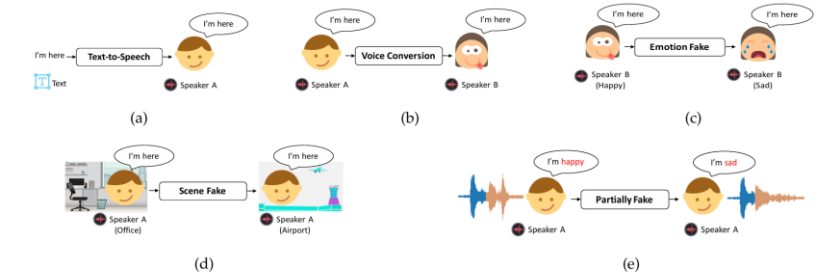


Fig. 2. Five kinds of deepfake audio: (a) text-to-speech, (b) voice conversion, (c) emotion fake, (d) scene fake, (e) partially fake.

伪造音频分类

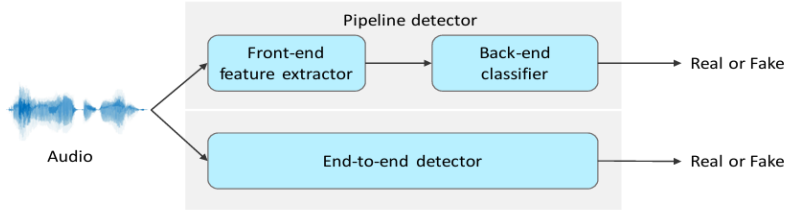


Fig. 1. Mainstream solutions on audio deepfake detection: pipeline and end-to-end detector.

鉴别流程

- 伪造音频检测：通过训练分类器判别是否是生成的音频

- Yi J, Wang C, Tao J, et al. Audio deepfake detection: A survey[J]. arXiv preprint arXiv:2308.14970

音频水印

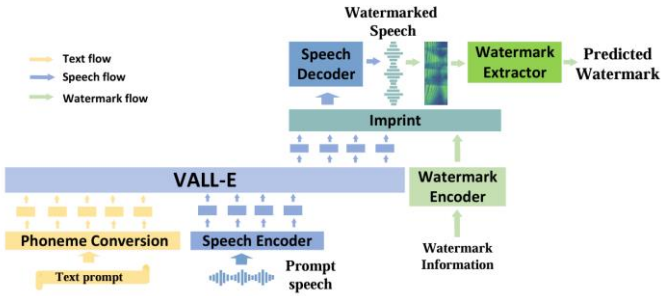


Figure 2: *The second stage: Watermarking mechanism integrate into language model of VALL-E.*

TraceableSpeech：将水印集成进LM-based TTS

Table 1: *Watermark Imperceptibility Metrics in Speech Reconstruction*

Model	PESQ ↑	STOI ↑	ViSQOL ↑
HiFicodec + WavMark(16bit)	3.197	0.947	3.880
TraceableSpeech(4@10)	**3.641**	**0.950**	**4.060**
TraceableSpeech(4@16)	3.569	0.948	3.985

[1] @ denotes the watermarking capacity. For example, 4@16 indicates 4-digit base-16, equivalent to the 16-bit capacity of WavMark used in the baseline. This annotation is applicable to other tables as well.

效 果

- 水印：在生成的音频里打上标记，方便辨别/溯源
- TraceableSpeech: 在LM-based TTS系统中端到端集成水印。

- Zhou J, Yi J, Wang T, et al. TraceableSpeech: Towards Proactively Traceable Text-to-Speech with Watermarking[J]. arXiv preprint arXiv:2406.04840, 2024.

人类对齐

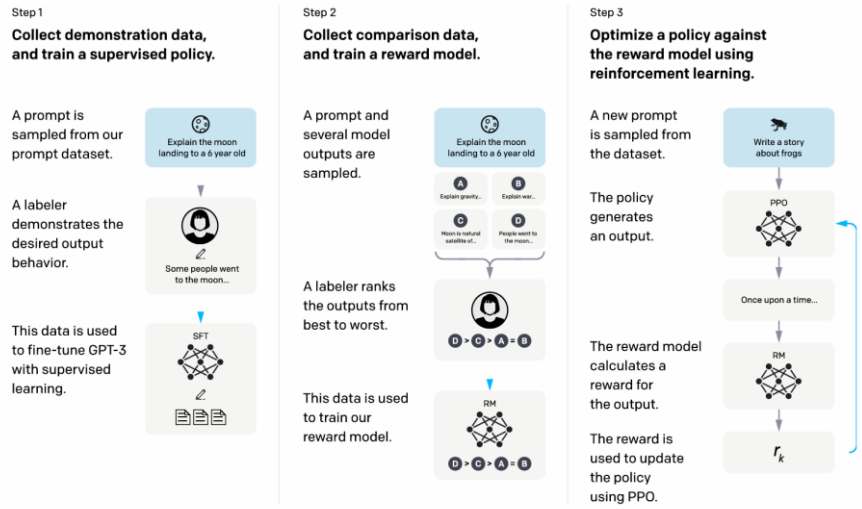
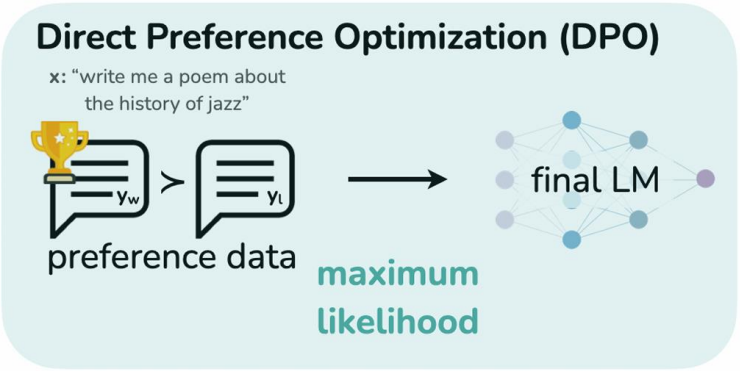


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

RLHF

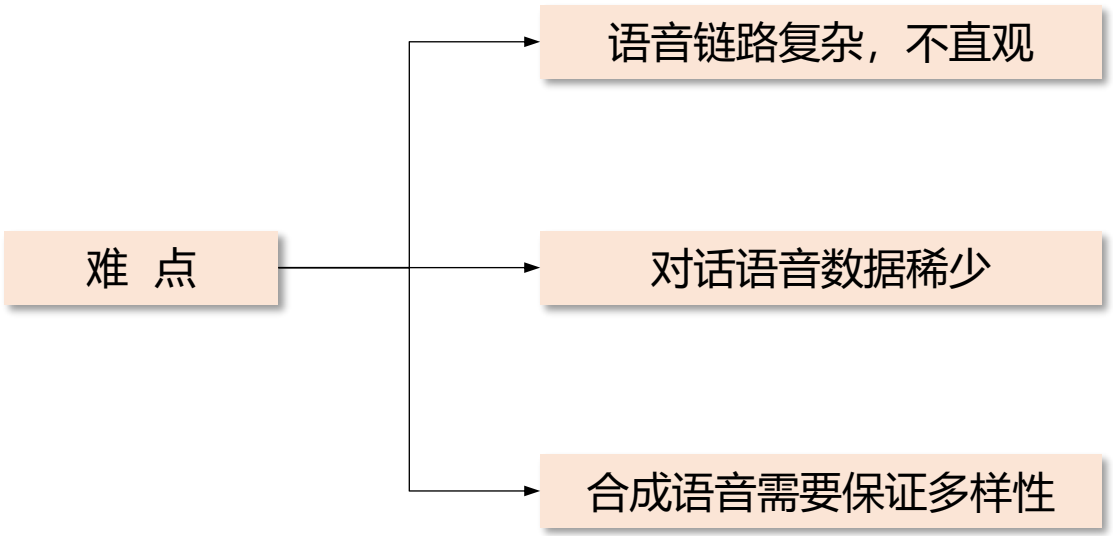


DPO

- 人类对齐：最常用的两个方法是RLHF和DPO。
- Human preference：人类对LM生成的样本进行打分选择。

- Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. NIPS2022
- Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model[J]. NIPS2022

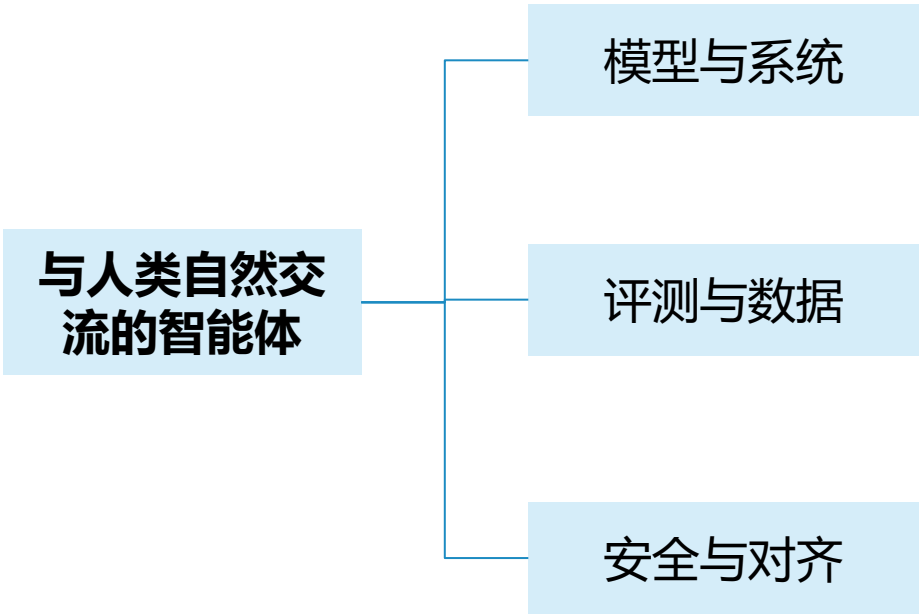
人类对齐难点：数据标注



提纲

- 语音大模型：历史与为什么
- 与人类自然交流的智能体
- 模型与系统
- 评测与数据
- 安全与对齐
- 小结与展望

小结



展望：待解决的问题

- 科学问题
  - Tokenizer的设计
  - 跨模态对齐
  - 打断的人类同步
  - 数据的收集
  - 音频生成安全
  - 人类对齐

- 社会问题
  - 人类是否真的准备好接纳完全拟人的机器？

语音大模型难点

- 说话人、口音、环境噪声等因素的多样性。
- 语音特有的韵律、情感等信息的建模。
- 语音和文本模态之间的对齐问题。
- 多语言、多口音支持的挑战。
- 实时性要求高，计算复杂度大。

展望：未来发展趋势

- 自监督学习: 减少标注数据依赖
- 强化学习: 实现主动试错学习
- 可解释性: 揭示模型有效机理
- 个性化: 提升用户适应能力
- 可控性: 解决边界情况
- 大型化: 触摸scaling law的上限
- 小型化: 提升模型经济效益

参考文献

1. Deshmukh S, Elizalde B, Singh R, et al. Pengi: An audio language model for audio tasks[J]. Advances in Neural Information Processing Systems, 2023, 36: 18090-18108.
2. Tang C, Yu W, Sun G, et al. Salmonn: Towards generic hearing abilities for large language models[J]. arXiv preprint arXiv:2310.13289, 2023.
3. Chu Y, Xu J, Zhou X, et al. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models[J]. arXiv preprint arXiv:2311.07919, 2023.
4. Chu Y, Xu J, Yang Q, et al. Qwen2-audio technical report[J]. arXiv preprint arXiv:2407.10759, 2024.
5. Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers[J]. arXiv preprint arXiv:2301.02111, 2023.
6. Betker J. Better speech synthesis through scaling[J]. arXiv preprint arXiv:2305.07243, 2023.
7. Anastassiou P, Chen J, Chen J, et al. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models[J]. arXiv preprint arXiv:2406.02430, 2024.
8. Kreuk F, Synnaeve G, Polyak A, et al. Audiogen: Textually guided audio generation[J]. arXiv preprint arXiv:2209.15352, 2022.
9. Copet J, Kreuk F, Gat I, et al. Simple and controllable music generation[J]. Advances in Neural Information Processing Systems, 2024, 36.
10. Zhang D, Li S, Zhang X, et al. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities[J]. arXiv preprint arXiv:2305.11000, 2023.
11. Yang D, Tian J, Tan X, et al. UniAudio: Towards Universal Audio Generation with Large Language Models[C]//Forty-first International Conference on Machine Learning.
12. Yang D, Guo H, Wang Y, et al. UniAudio 1.5: Large Language Model-driven Audio Codec is A Few-shot Audio Task Learner[J]. arXiv preprint arXiv:2406.10056, 2024.
13. Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 3451-3460.
14. Chiu C C, Qin J, Zhang Y, et al. Self-supervised learning with random-projection quantizer for speech recognition[C]//International Conference on Machine Learning. PMLR, 2022: 3915-3924.
15. Zhang Y, Han W, Qin J, et al. Google usm: Scaling automatic speech recognition beyond 100 languages[J]. arXiv preprint arXiv:2303.01037, 2023.
16. Bai Y, Chen J, Chen J, et al. Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition[J]. arXiv preprint arXiv:2407.04675, 2024.
17. Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International conference on machine learning. PMLR, 2023: 28492-28518.
18. Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.
19. Borsos Z, Sharifi M, Vincent D, et al. Soundstorm: Efficient parallel audio generation[J]. arXiv preprint arXiv:2305.09636, 2023.
20. Défossez A, Copet J, Synnaeve G, et al. High fidelity neural audio compression[J]. arXiv preprint arXiv:2210.13438, 2022.

参考文献

21. Zeghidour N, Luebs A, Omran A, et al. Soundstream: An end-to-end neural audio codec[J]. IEEE/ACM TASLP
22. Fu C, Lin H, Long Z, et al. VITA: Towards Open-Source Interactive Omni Multimodal LLM[J]. arXiv preprint arXiv:2408.05211, 2024.
23. Ma Z, Song Y, Du C, et al. Language Model Can Listen While Speaking[J]. arXiv preprint arXiv:2408.02622, 2024.
24. Yi J, Wang C, Tao J, et al. Audio deepfake detection: A survey[J]. arXiv preprint arXiv:2308.14970, 2023.
25. Chen G, Wu Y, Liu S, et al. Wavmark: Watermarking for audio generation[J]. arXiv preprint arXiv:2308.12770, 2023.
26. Zhou J, Yi J, Wang T, et al. TraceableSpeech: Towards Proactively Traceable Text-to-Speech with Watermarking[J]. arXiv preprint arXiv:2406.04840, 2024.
27. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
28. Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model[J]. Advances in Neural Information Processing Systems, 2024, 36.

谢谢！
请大家批评指教。