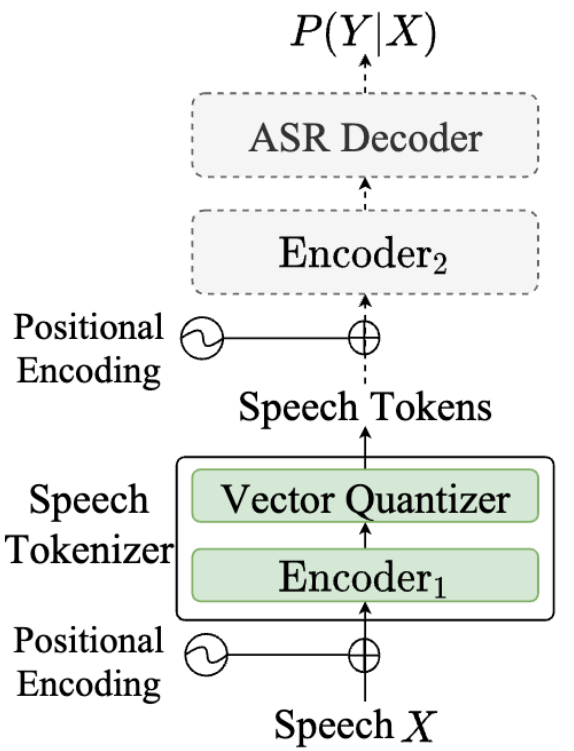
CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens

2024-07-09

CosyVoice: S^3 Tokenizer



(a) Supervised speech tokenizer

- **监督式的 Tokenizer**

$$H = \text{Encoder}_1\left(\text{PosEnc}(X)\right)$$

X: 语音的梅尔特征作为输入

$$\mu_l = \text{VQ}(\mathbf{h}_l, C) = \arg\min_{\mathbf{c}_n \in C} ||\mathbf{h}_l - \mathbf{c}_n||_2$$

EMA 方式更新 codebook

$$\bar{H} = \{\mathbf{c}_{\mu_1}, \mathbf{c}_{\mu_2}, \ldots, \mathbf{c}_{\mu_L}\}$$

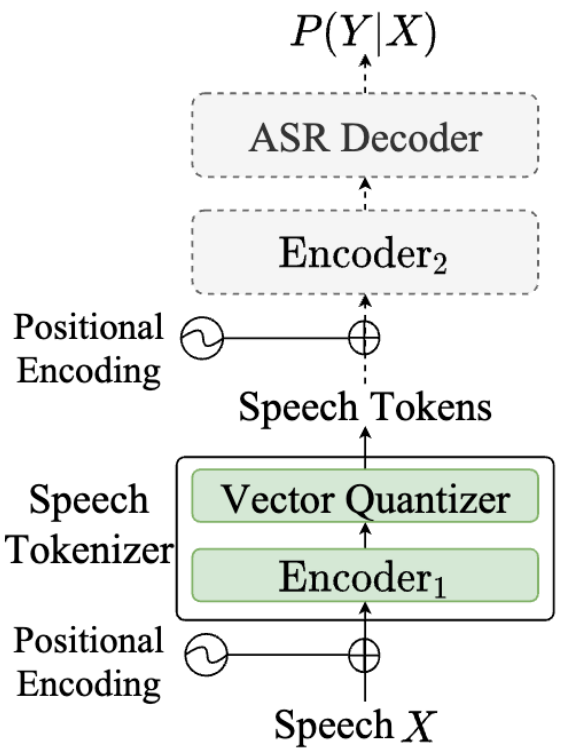
$$\tilde{H} = \text{Encoder}_2\left(\text{PosEnc}(\bar{H})\right)$$

补充位置编码，增强时序信息

$$P(Y|X) = \text{ASRDecoder}\left(\tilde{H}, Y^{Z-1}\right)$$

- **无监督的 Tokenizer**
  - Encodec / DAC / HiFi-Codec …
  - 没有显式建模语义信息，没有与文本模态对齐

CosyVoice: S^3 Tokenizer



(a) Supervised speech tokenizer

- **Backbone**
  - 单语种小数据量：espnet 的 conformer
  - 多语种大数据量：SenseVoice-Large
- **VQ 相关配置**
  - VQ 层均是插入在 6 层 Encoder 之后
  - VQ 的 codebook size = 4096（单层，不做 RVQ）
  - 前 6 层 Encoder + VQ 层，即可作为 Tokenizer

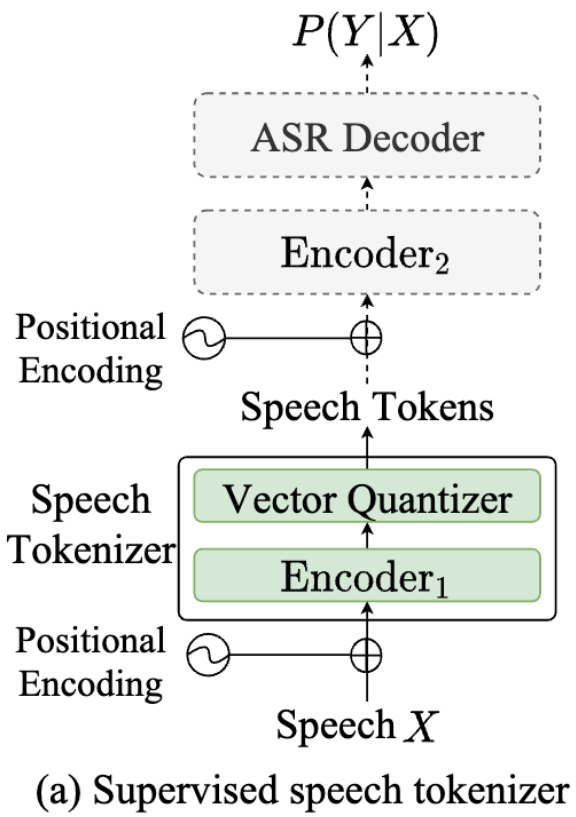
Model	dev_clean	test_clean	test_other
Conformer	2.62	2.89	6.57
Conformer-VQ	3.13	3.18	7.56

Table 5: Impact of inserting vector quantization on speech recognition in terms of word error rate (%).

Test set	Whisper-L-V3 w/o lid	Whisper-L-V3 w/ lid	SenseVoice-L w/o lid	SenseVoice-L w/ lid	S^3 tokens w/o lid	S^3 tokens w/ lid
zh-CN	12.82	12.55	8.76	8.68	12.24	12.06
en	13.55	9.39	9.79	9.77	15.43	15.38

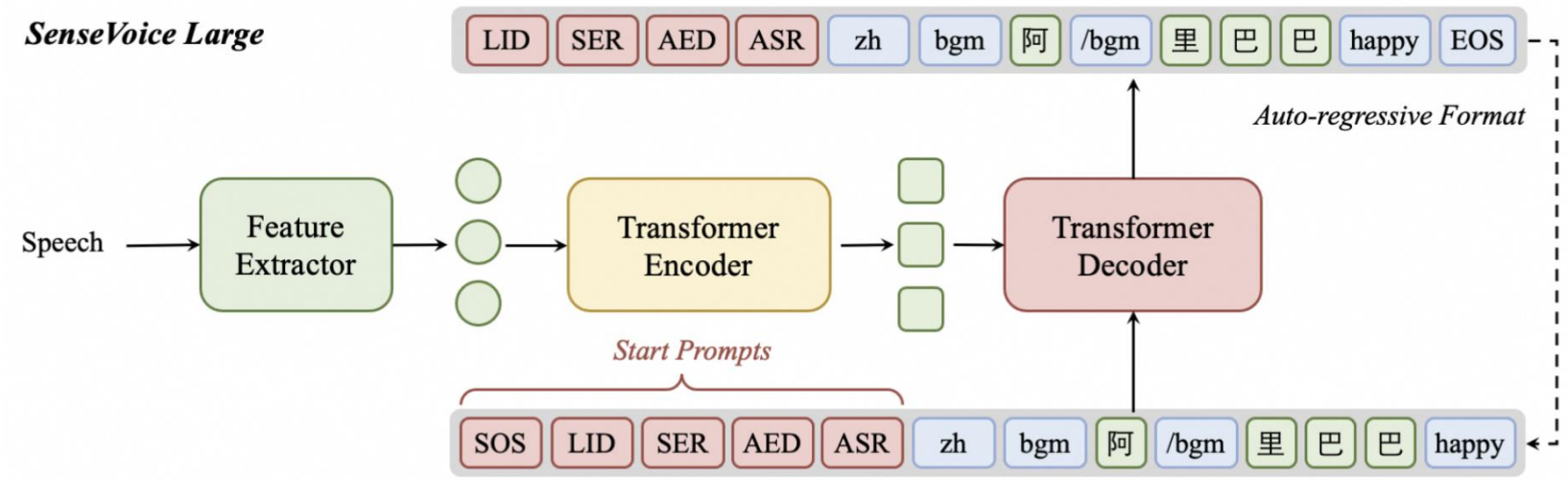
Table 6: The evaluation on S^3 tokens' capability to preserve semantic information. We employ word and character error rates for zh-CN and en languages on the Common Voice benchmarks.

CosyVoice: S^3 Tokenizer



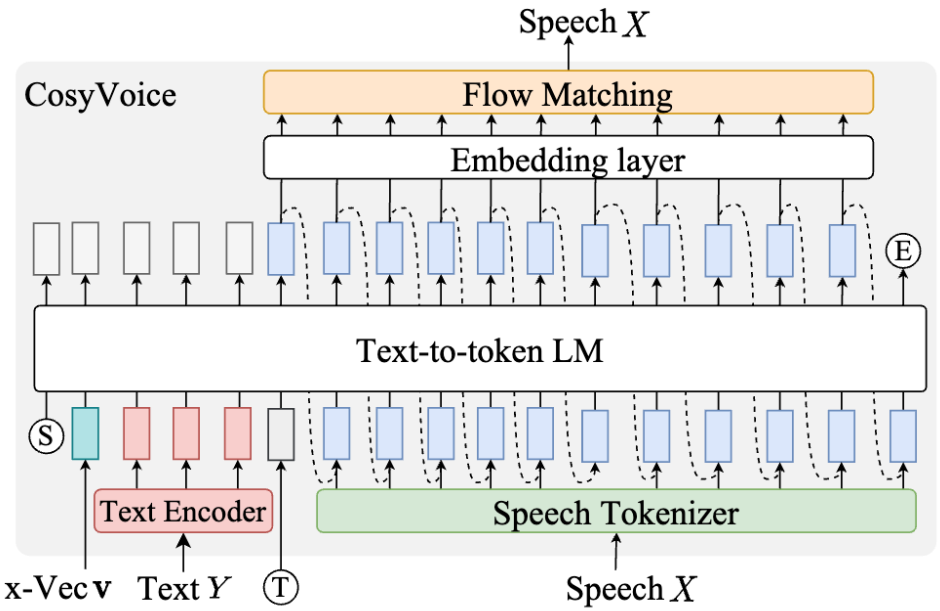
(a) Supervised speech tokenizer

SenseVoice-Large



- 多任务训练：语种识别、情感分类、音频事件分类、是否反正则化
- 支持语种数量：超过 50 种
- 训练音频总数据量：超过 30 万小时

CosyVoice: LLM



(b) An overview of the proposed CosyVoice LM

- 1. **Text Encoder**: 将 text token 与 speech token 对齐到相同空间
- 2. **Speech Tokenizer**: S^3 Tokenizer
- 3. **Text-to-Token LM**: decoder-only LLM
- 4. **CFM**: Conditional Flow Matching（不需要对齐预测模块）
- 5. **HiFiGAN**: 梅尔特征 → 波形

- **训练数据序列** $[\text{Ⓢ}, \mathbf{v}, \{\bar{\mathbf{y}}_u\}_{u\in[1:U]}, \text{Ⓣ}, \{\mu_l\}_{l\in[1:L]}, \text{Ⓔ}]$
- **文本 Tokenizer**

$$\bar{Y} = \{\bar{\mathbf{y}}_u\}_{u\in[1:U]} \qquad \bar{Y} = \text{TextEncoder}(\text{BPE}(Y))$$

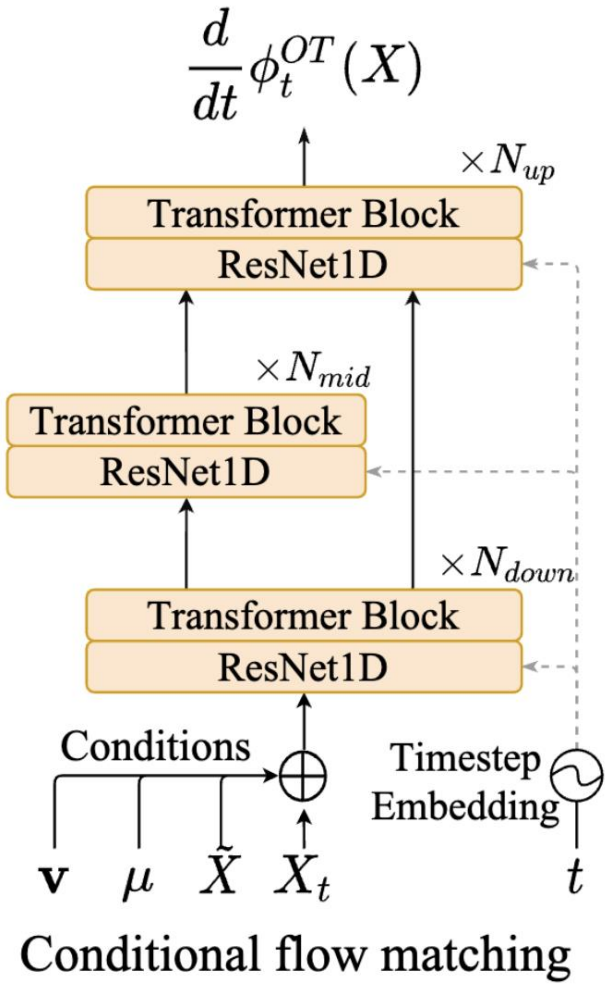
- **训练 Loss 只关注 Speech Token**

$$\mathcal{L}_{LM} = -\frac{1}{L+1}\sum_{l=1}^{L+1}\log q(\mu_l)$$

Settings	Tiny	Normal
Text Encoder		
Layers	6	6
Attention Dim.	512	1,024
Attention Heads	8	16
Linear Units	2,048	4,096
Language Model		
Layers	12	14
Attention Dim.	512	1,024
Attention Heads	8	16
Linear Units	2,048	4,096

Table 4: Details of model architecture settings in the tiny and normal CosyVoice models.

CosyVoice: Conditional Flow Matching

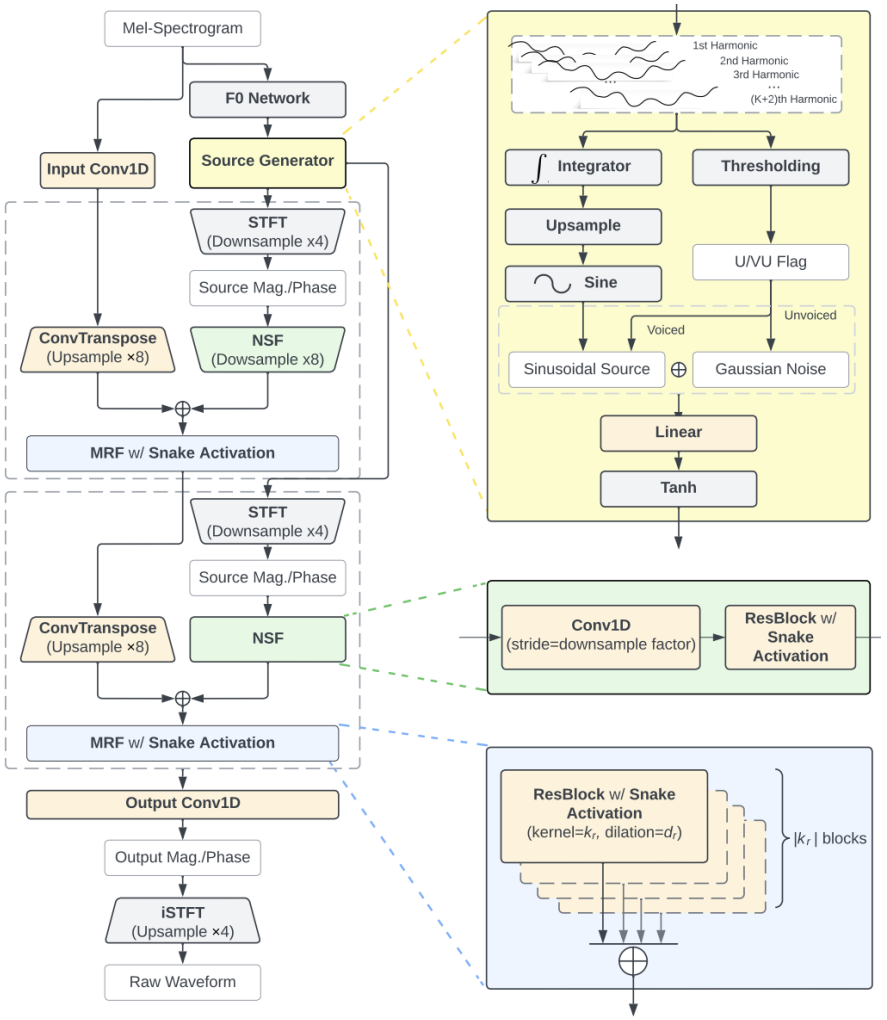


Conditional flow matching

Matcha-TTS：Optimal-Transport Conditional Flow Matching

- 初始先验分布：标准正态分布 $\phi_0(X) \sim p_0(X) = \mathcal{N}(X; 0, I)$
- 数据目标分布： $q(X)$ 用 $p_1(X)$ 近似
- 输入条件：
  1. speech token
  2. speaker embedding **v**
  3. masked 梅尔特征（mask 的含义：随机取一段置为 0）
- 相比于 diffusion：训练更简单、推理更快
- 其他改进
  - cosine scheduler
  - CFG (classifier-free guidance)
    - 训练：20% 的概率，随机将 3 种去除，学习非条件 flow
    - 推理：条件 flow 和非条件 flow 线性插值

CosyVoice: HiFiGAN → HiFTNet (FunAudioLLM)



HiFTNet 的优化

- NSF：Neural Source Filter
- Pitch 估计：使用 Dio 提取的 pitch 训练 JDC 神经网络
  - 任务：从输入梅尔谱预测 pitch 值
  - 数据增广：提高推理时的鲁棒性
- 判别器：Multi-Resolution Discriminator
  - 增加 Snake 函数，但没有 BigVGAN 的抗混叠滤波
- 预测目标：语谱图的幅值和相位，通过 iSTFT 还原波形

Model	Dataset	CMOS (p-value) ↑	MCD ↓	RTF ↓	RAM ↓
Ground Truth	LJSpeech	-0.06 ($p = 0.396$)	—	—	—
HiFTNet	LJSpeech	—	2.567	0.0057	0.90 GB
iSTFTNet	LJSpeech	$+0.64$ ($p < 10^{-7}$)	2.820	0.0031	0.77 GB
HiFi-GAN	LJSpeech	$+0.19$ ($p = 0.028$)	2.816	0.0043	0.75 GB
Ground Truth	*test-clean*	-0.21 ($p = 0.033$)	—	—	—
HiFTNet	*test-clean*	—	2.892	—"—	—"—
BigVGAN-base	*test-clean*	$+0.21$ ($p = 0.001$)	3.079	0.0159	0.90 GB
BigVGAN	*test-clean*	-0.05 ($p = 0.552$)	2.656	0.0243	1.52 GB
Ground Truth	*test-other*	-0.10 ($p = 0.189$)	—	—	—
HiFTNet	*test-other*	—	3.690	—"—	—"—
BigVGAN-base	*test-other*	$+0.17$ ($p = 0.022$)	3.892	—"—	—"—
BigVGAN	*test-other*	$+0.12$ ($p = 0.354$)	3.189	—"—	—"—

CosyVoice in FunAudioLLM

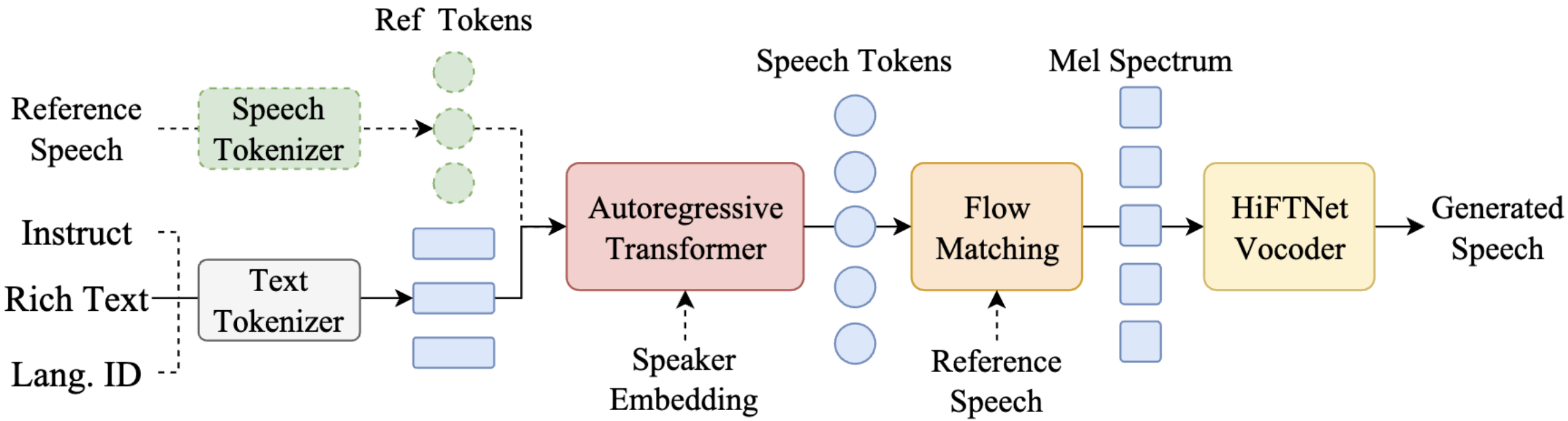


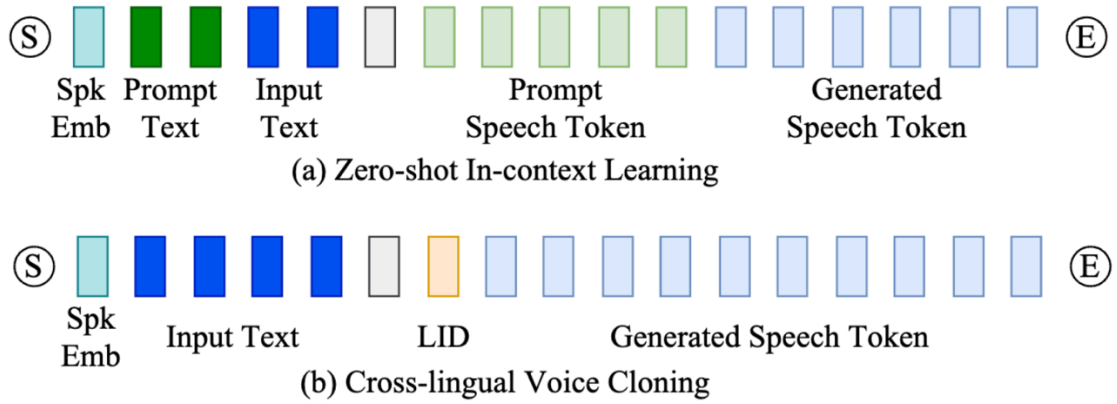
Figure 4: A semantic diagram of CosyVoice models.

CosyVoice: Zero-Shot In-Context Learning

功能一：Zero-Shot In-Context Learning

1. Text-to-Token LM

- 同语种音色克隆
  - 与训练阶段匹配，根据 target 文本，续写 speech tokens
- 跨语种音色克隆
  - 将 prompt 音频的文本和 speech token 忽略
    - 目标：避免不同语种的韵律特性泄露
  - 只使用 speaker embedding 作为 prompt
  - 疑问：speaker embedding 与 speech token 是否相当于对音色和语义信息进行了解耦?



2. Optimal-Transport Conditional Flow Matching

- 输入一：[prompt token, generated speech tokens]
- 输入二：speaker embedding 和 prompt 的梅尔特征
  - 提高音色和音频环境的一致性

TODO：确认代码

CosyVoice: Rich Generation with Instruction

功能二：Rich Generation with Instruction

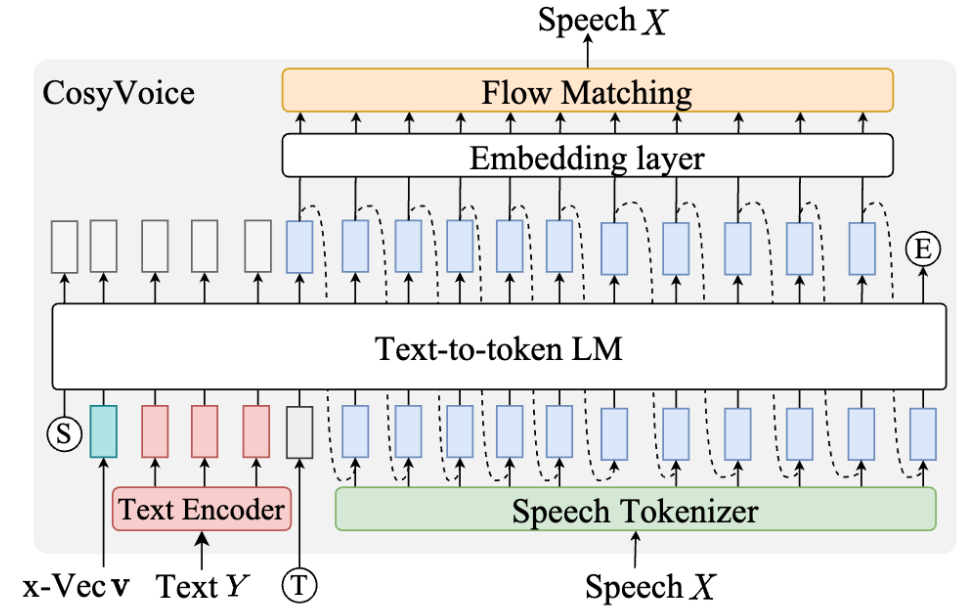
功能：通过自然文本描述，控制合成语音的更多特性

- speaker identity、speaking style (emotion/gender/pitch/speed)
- laughter, breaths, speaking while laughing、指定词语的 emphasis

实现方案

- 使用 instruct 类型的数据 finetune CosyVoice 基础模型
- finetune 时，不加入 speaker embedding

Speaker Identity 1. Selene 'Moonshade', is a **mysterious**, **elegant dancer** with a connection to the night. Her movements are both **mesmerizing** and **deadly**.<endofprompt>Hope is a good thing. 2. Theo 'Crimson', is a **fiery**, **passionate** rebel leader. Fights with fervor for justice, but struggles with **impulsiveness**.<endofprompt>You don't know about real loss.
Speaking Style 1. A **happy girl** with **high tone** and **quick speech**.<endofprompt>The sun is shining brightly today. 2. A **sad woman** with **normal tone** and **slow speaking speed**.<endofprompt>I failed my important exam.
Fine-grained Paralinguistics 1. Well that's kind of scary **[laughter]**. 2. I don't think I over eat yeah **[breath]** and um I do exercise regularly. 3. Well that pretty much covers <**laughter**>**the subject**</**laughter**> well thanks for calling me. 4. The team's <**strong**>**unity**</**strong**> and <**strong**>**resilience**</**strong**> helped them win the championship.



(b) An overview of the proposed CosyVoice LM

CosyVoice: Datasets

训练数据信息

- 小数据量：LibriTTS
- 大数据量：大规模多语言数据集
  - 5 种语言，中文为主
- 数据清洗关键流程
  - Speech detection
  - SNR estimation
  - Speaker diarization
  - Speech separation
  - Pseudo Label Generation
  - Force Alignment
  - …

Language	Duration (hr)
ZH	130,000
EN	30,000
Yue	5,000
JP	4,600
KO	2,200

Table 2: Hours of CosyVoice training data across languages in the large-scale experiments.

Type	Duration (hr)
Speaker Identity	101
Speaking Style	407
Fine-grained Paralinguistics	48

Table 3: Duration statistics of instruction training data by type.

CosyVoice: Evaluation

音色克隆评测

- **WER**：中文用 Paraformer-CN，英文用 Whisper-Large v3
- **音色相似度（SS）**：ERes2Net 计算 prompt 与合成音频之间的余弦相似度
  - 现象：合成音频与 prompt 音频之间的音色相似度，大于真实音频与 prompt 音频之间的音色相似度
- **Re-ranking**：生成 5 条，使用 WER 最低的一条作为合成语音进行指标评价

Model	Text token	Speech Token	WER (%)	#INS+DEL	#SUB	SS
Original	-	-	3.01	66	200	69.67
VALL-E (Wang et al., 2023)	Phone	Encodec	18.70	342	1312	53.19
UniAudio (Yang et al., 2023)	Phone	Encodec	8.74	254	519	47.56
SpearTTS (Kharitonov et al., 2023)	Phone	Hubert	6.14	133	410	51.71
Exp-1-LibriTTS	Phone	Hubert	7.41	325	409	67.85
Exp-2-LibriTTS	Phone	S^3_{en}	5.05	122	325	67.85
Exp-3-LibriTTS	BPE_{en}	S^3_{en}	3.93	108	239	67.85
Exp-4-LibriTTS	BPE	S^3	4.76	134	287	65.94
Exp-4-Large-scale	BPE	S^3	**3.17**	**96**	**184**	**69.49**

Table 7: Comparison with other TTS models on the LibriTTS test-clean set in terms of content consistency and speaker similarity (SS). Non-autoregressive ASR model, Paraformer-en, is employed for fast evaluation.

Model	WER (%)	#Ins.&Del.	SS
Original	2.66	92	69.67
ChatTTS	8.32	441	-
CosyVoice	2.89±0.18	88.60±3.88	74.30±0.15
+ 5× re-ranking	1.51	47	74.30

Table 8: The comparison of original and CosyVoice generated speeches on the LibriTTS test-clean set in terms of word error rate (WER) and speaker similarity (SS). "±" joins the mean and standard deviation for each evaluation metric. Whisper-Large V3 is employed as the ASR model.

Model	CER (%)	#Ins.&Del.	SS
Original	2.52	25	74.15
ChatTTS	3.87	111	-
CosyVoice	3.82±0.24	24.4±2.24	81.58±0.16
+ 5× re-ranking	1.84	11	81.58

Table 9: The comparison of original and CosyVoice generated speeches on the AISHELL-3 test set in terms of character error rate (CER) and speaker similarity (SS). Paraformer-zh is employed as the ASR model.

CosyVoice: Evaluation

情绪控制性评测

- 测试集：100 条与情感相匹配的文本；6 种情感：happy, angry, sad, surprised, fearful, disgusted
- 合成的方式： Happy.<endofprompt>Content Text
- 评价指标：emotion2vec 开源模型的情感分类准确性

Model	Happy	Sad	Angry	Surprised	Fearful	Disgusted
CosyVoice-base	1.00±0.00	0.45±0.05	0.59±0.03	0.26±0.02	0.88±0.01	0.46±0.06
CosyVoice-instruct	1.00±0.00	0.98±0.02	0.83±0.04	0.64±0.03	0.87±0.03	0.93±0.02
w/o instruction	0.98±0.01	0.77±0.04	0.49±0.12	0.28±0.06	0.83±0.04	0.45±0.16

Table 10: Comparison of emotion control accuracy between CosyVoice-base-300M and CosyVoice-instruct-300M. "±" joins the mean and standard deviation for each evaluation metric.

TTS 造数据的可行性

Training Data	dev_clean	dev_other	test_clean	test_other
Librispeech	2.77	5.84	2.79	5.97
Syn on LS text	2.79	6.37	3.00	6.59
Librispeech + Syn on LS text	2.44	5.52	2.56	5.68
Librispeech + Syn on LS text ×2	2.51	5.23	2.68	5.26
Librispeech + Syn on LS, MLS text	**1.93**	**4.43**	**2.04**	**4.53**

Comparison with Similar Models

名称	论文	时间	spk-emb 拼接策略	Tokenizer	Decoder	Vocoder	Vocoder 输入
Tortoise-TTS	2305.07243	2023.05	不拼接	Mel-VQ	DDPM/DDIM	Univnet	mel
GPT-soVITS	-	2024.01	不拼接	Hubert	VITS	HiFiGAN	embedding
Base-TTS	2402.08093	2024.02	拼接单个 spk-emb	解耦方案	HiFiGAN 结构	BigVGAN	embedding
Chat-TTS	-	2024.05	不拼接	Mel-VQ	Decoder	Vocos	mel
Fish-speech	-	2024.05	不拼接	Mel VQ	VITS	HiFiGAN	embedding
Seed-TTS	2406.02430	2024.06	不拼接	未说明细节	DiT	未说明细节	embedding
XTTS-v2	2406.04904	2024.06	拼接：perceiver	Mel-VQ	HiFiGAN		embedding
CosyVoice	CosyVoice	2024.07	拼接单个 spk-emb	S3-Tokenizer	Flow-Matching	HiFTNet	mel