

TTS(Text To Speech)初步调研

TTS(Text To Speech)初步调研

1.fish-speech

1.0 模型背景

1.1模型效果

1.1.1 模型优势

1.1.2 存在的问题

1.3 Further Engineering

总结

2.XTTS

2.0 模型背景

2.1 模型效果

2.1.0 中文案例

2.1.1 英文案例

2.1.3 中英混杂案例

2.1.4 模型性能

2.1.5 存在的问题

总结

3.CosyVoice

3.0 模型背景

3.1 模型效果

3.1.1 中文案例

CosyVoice-300M-SFT

CosyVoice-300M

3.1.2 英文案例

CosyVoice-300M-SFT

CosyVoice-300M

3.1.3 中英混杂案例

CosyVoice-300M-SFT

CosyVoice-300M

3.1.4 模型性能

3.1.5 存在的问题

3.2 Further Engineering

总结

总结

1.fish-speech

1.0 模型背景

[fish-speech-1.2](#) 发布于2024年7月2日，国内Fish Audio团队开发的模型，其中多名开发者参与过 So-VITS-SVC、GPT-SoVITS 等著名TTS项目，使用了 30万小时的中英日数据训练而成，github上收获了4.8kstars。demo网页：<https://fish.audio/zh-CN/>，据开发者所悉，网页上的模型即huggingface上发布的预训练模型sft而成。

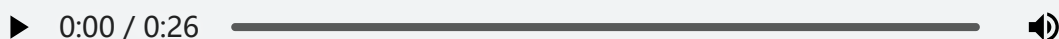


1.1 模型效果

1.1.1 模型优势

1.通过设置一段该人的语音及其对应的标注文件（lab）作为参考音频，可以快速复制某人的发音样式。

利用施一公老师的一段话作为参考，生成的一段语音：



2.模型显存占用低（1704MiB）；显存占用率低（37%；A40）；单推理的生成速度约是3:4（音频长度:生成时间）

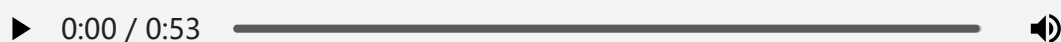
3.模型已有完整的api（含流式传输）服务框架

1.1.2 存在的问题

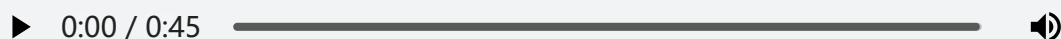
1.参考序列只能设置一个，如果生成的语种与参考序列不同，那么效果适得其反

- 1 Genome-wide association studies (GWAS) refer to the association research between genes and diseases conducted on a genome-wide scale, involving multi-center, large-sample, and repeated verification. GWAS is a research method that involves genotyping of high-density genetic markers (such as SNPs or CNVs) across large population DNA samples to identify genetic factors associated with complex diseases, comprehensively revealing the genetic genes related to the onset, development, and treatment of diseases

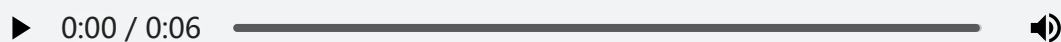
使用施一公中文音频作为参考音频后，对以上英文内容进行生成：



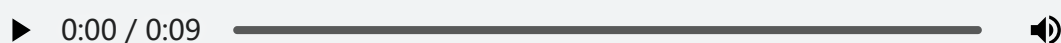
不使用参考序列时生成的效果：



2.大部分时候，模型生成在一些地方尤其是标点处有不自然的停顿：



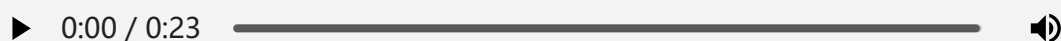
3.在不使用参考序列的情况下，有一定概率出现唱歌现象，可能使用了歌曲音频作为训练数据：



4.在不使用参考序列的情况下，有一定概率音频产生不完整：



5.在不使用参考序列的情况下，有一定概率完全杂音：



6.在不使用参考序列的情况下，无法固定speaker音色：



(源自与开发者的对话)

7.中英文混杂能力有待提升，英文出现大写字母时会单独读出来，合成词无法正确读出来

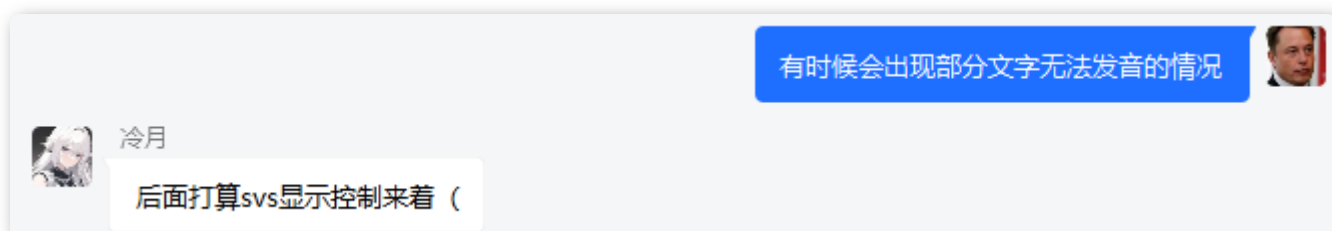
1 我是Futuregene公司开发的westlakechat

▶ 0:00 / 0:04 🔊

1.3 Further Engineering

- 1.开源的是pre-trained model，网页上的是sft后的版本，目前还有提升空间。
- 2.以目前的框架设计来看，如果要上线要分开两个语言环境去设计，且对于中英混杂型的发言不大友好。
- 3.最好等待开发者的迭代更新，这些情况我都已经反馈给开发者了。

总结





1. 其实模型是支持多个参考音频序列的, 目前线上版本实装了, 可以给一个中文一个英文一起推 (只给中文推英文还在优化)
2. 这个主要是预训练数据和自动切割的问题, 目前在线上版本通过 sft 修复了, 我们下一个版本的训练数据修复了该问题
3. 我们确实使用了歌曲数据, 后续考虑添加说话人控制来优化这个问题
- 4/5. 主要是稳定性问题, sft 基本解决
6. 考虑到生成是一个采样过程, 只有锁定种子可以稳定
7. 中英混读在下个版本的数据中得到修复了

我们当前线上的 sft 会在两周后开源, 同时线上会发布下一版预训练模型. 感谢您的支持与关注

(源自与开发者的对话, 开发者对以上问题的解释)

该模型在部分案例上表现较好, 但大部分案例表现的不是很稳定, 只有使用参考序列且生成语种一致时, 稳定性较高, 如果语种不一致则适得其反, 在运行性能上有较好的表现, 同时还存在很多时不时出现的问题隐患, 模型技术还是不大成熟, 开发者们目前正在积极调整模型框架, 后续将会有更先进的模型上线, 目前情况不适合直接上线。

模型性能: ★★★★★

模型效果: ★★★★★

模型稳定性: ★★☆☆☆

模型成长性: ★★★★★

2.XTTS

2.0 模型背景

[XTTS-v2](#) 发布于2023年11月, 国外Coqui公司开发的模型, 母公司为火狐公司。目前支持 17 种语言, 在github上收获了31.9k stars, 在huggingface上所有TTS模型中下载量排行第一、likes排行第二。demo页面: <https://huggingface.co/spaces/coqui/xtts>

2.1 模型效果

2.1.0 中文案例

- 1 全基因组关联研究是指在全基因组层面上, 开展多中心、大样本、反复验证的基因与疾病的关联研究。

▶ 0:00 / 0:09



2.1.1 英文案例

- 1 Genome-wide association studies (GWAS) refer to the association research between genes and diseases conducted on a genome-wide scale, involving multi-center, large-sample, and repeated verification. GWAS is a research method that involves genotyping of high-density genetic markers (such as SNPs or CNVs) across large population DNA samples to identify genetic factors associated with complex diseases, comprehensively revealing the genetic genes related to the onset, development, and treatment of diseases

▶ 0:00 / 0:29



2.1.3 中英混杂案例

- 1 我是Futuregene公司开发的westlakechat

以en为目标语言：

▶ 0:00 / 0:04



以zh为目标语言：

▶ 0:00 / 0:06



2.1.4 模型性能

显存占用约3.7GB；显存利用率约为40%

单推理的生成速度约是 1:1.23（音频长度:生成时间）

2.1.5 存在的问题

1.对输入文本长度有限制，文本过长会导致截断，如果上线可能需要拼接处理。

[!] Warning: The text length exceeds the character limit of 82 for language 'zh', this might cause truncated audio.

[!] Warning: The text length exceeds the character limit of 250 for language 'en', this might cause truncated audio.

- 2.停顿不自然（见中文案例）
- 3.英文词汇合成词，或者某些简写，无法正确发音（见英文案例）
- 4.多语言混杂时，无法支持。（见中英混杂案例）
- 5.最新的模型是去年发布的，产品迭代很慢，模型上限不高

总结

模型性能：★★★★☆

模型效果：★★☆☆☆

模型稳定性：★★☆☆☆

模型成长性：★★☆☆☆

3. [CosyVoice](#)

3.0 模型背景

[CosyVoice](#)发布于2024年7月4日，国外国内阿里巴巴发布的模型，目前开源了三种类型：[CosyVoice-300M](#)（base模型，用于零样本生成），[CosyVoice-300M-Instruct](#)（经过指令微调的模型，能够生成带有情感和特定风格的语音），[CosyVoice-300M-SFT](#)（经过监督微调的模型，提供更高的语音生成质量）

sft模型支持的speaker： '中文女', '中文男', '日语男', '粤语女', '英文女', '英文男', '韩语女'

instruct模型支持的特殊功能： <laughter></laughter>
[laughter] [breath]

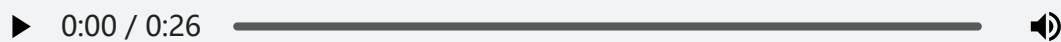
3.1 模型效果

3.1.1 中文案例

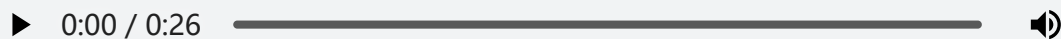
- 1 全基因组关联研究是指在全基因组层面上,开展多中心、大样本、反复验证的基因与疾病的关联研究,是通过对大规模的群体DNA样本进行全基因组高密度遗传标记（如SNP或CNV等）分型,从而寻找与复杂疾病相关的遗传因素的研究方法,全面揭示疾病发生、发展与治疗相关的遗传基因。

CosyVoice-300M-SFT

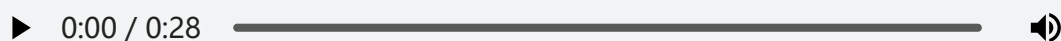
CosyVoice-300M--SFT-中文女



CosyVoice-300M--SFT-中文男

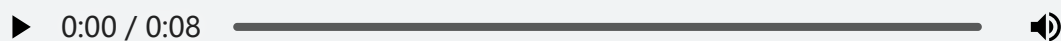


CosyVoice-300M--SFT-粤语女



CosyVoice-300M

CosyVoice-300M--施一公参考语音



3.1.2 英文案例

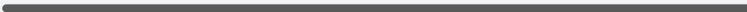

- 1 Genome-wide association studies (GWAS) refer to the association research between genes and diseases conducted on a genome-wide scale, involving multi-center, large-sample, and repeated verification. GWAS is a research method that involves genotyping of high-density genetic markers (such as SNPs or CNVs) across large population DNA samples to identify genetic factors associated with complex diseases, comprehensively revealing the genetic genes related to the onset, development, and treatment of diseases

CosyVoice-300M-SFT

CosyVoice-300M-英文女

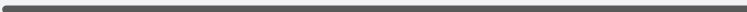



CosyVoice-300M-英文男

▶ 0:00 / 0:33  

CosyVoice-300M

CosyVoice-300M--施一公参考语音

▶ 0:00 / 0:14  



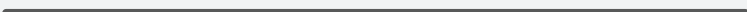

遇到 . 可能会停下,句子长度过长也可能停下

3.1.3 中英混杂案例

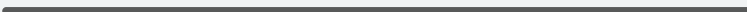

1 我是Futuregene公司开发的westlakechat

CosyVoice-300M-SFT

CosyVoice-300M-英文女

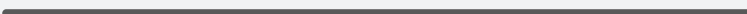

▶ 0:00 / 0:03  

CosyVoice-300M-中文女

▶ 0:00 / 0:02  

CosyVoice-300M

CosyVoice-300M--施一公参考语音

▶ 0:00 / 0:03  

3.1.4 模型性能

1.7: 1 (音频长度:生成时间) (仅为参考,不同场景不同语种推理速度不一样)

显存占用约17GB,显存利用率可达100%

3.1.5 存在的问题

1.当输入较长句子或者几个句子合并时，可能会导致自动分割或者语速加快。（见英文案例）

3.2 Further Engineering

1.流式api服务

2.对句子正确分割服务

总结

模型性能： ★★★★★

模型效果： ★★★★★☆

模型稳定性： ★★★★★☆

模型成长性： ★★★★★

总结

	模型性能	模型效果	模型稳定性	模型成长性
fish-speech	★★★★☆☆	★★★★☆☆	★★☆☆☆☆	★★★★★★
XTTS	★★★★★☆☆	★★☆☆☆☆	★★☆☆☆☆	★★☆☆☆☆
CosyVoice	★★★★★★	★★★★★☆☆	★★★★★☆☆	★★★★★★

目前CosyVoice模型最适合上线服务，表现最好最稳定，但还需完成流式服务api以及对句子分割的工程，而fish-speech模型存在部分案例音色音质表现特别好的情况，2周后会开源sft模型解决之前提出的问题，后续将会持续关注。