进度报告

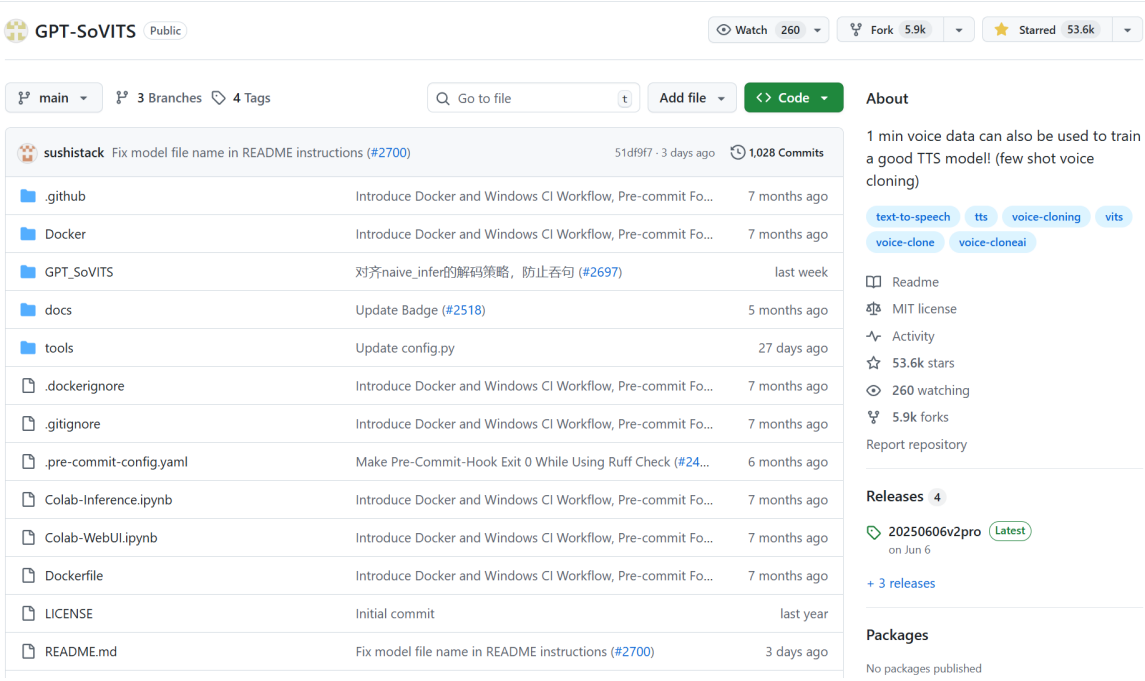
本次主要记录TTS（文字转语音）的数据集收集，训练，以及推理功能实现。

1 GPT-SoVITS

> ### 作品简介模板
>
> **音声来源：**[训练集音声来源]
>
> **免责声明：**本作品仅作为毕业设计发布，可能造成的后果与使用的语音合成项目的作者、贡献者无关。
>
> [attention:] 最好发视频可以带上GPT-SoVITS的Tag

项目地址：https://github.com/RVC-Boss/GPT-SoVITS

GPT-SoVITS是花儿不哭研发的低成本AI音色合成软件。



TTS（Text-To-Speech）这是一种文字转语音的语音合成。类似的还有SVC（歌声转换）、SVS（歌声合成）等。目前GPT-SoVITS实现了：

GPT-SoVITS-V1实现了：

- 由参考音频的情感、音色、语速控制合成音频的情感、音色、语速
- 可以少量语音微调训练，也可不训练直接推理
- 可以跨语种生成，即参考音频（训练集）和推理文本的语种为不同语种

GPT-SoVITS-V2新增特点：

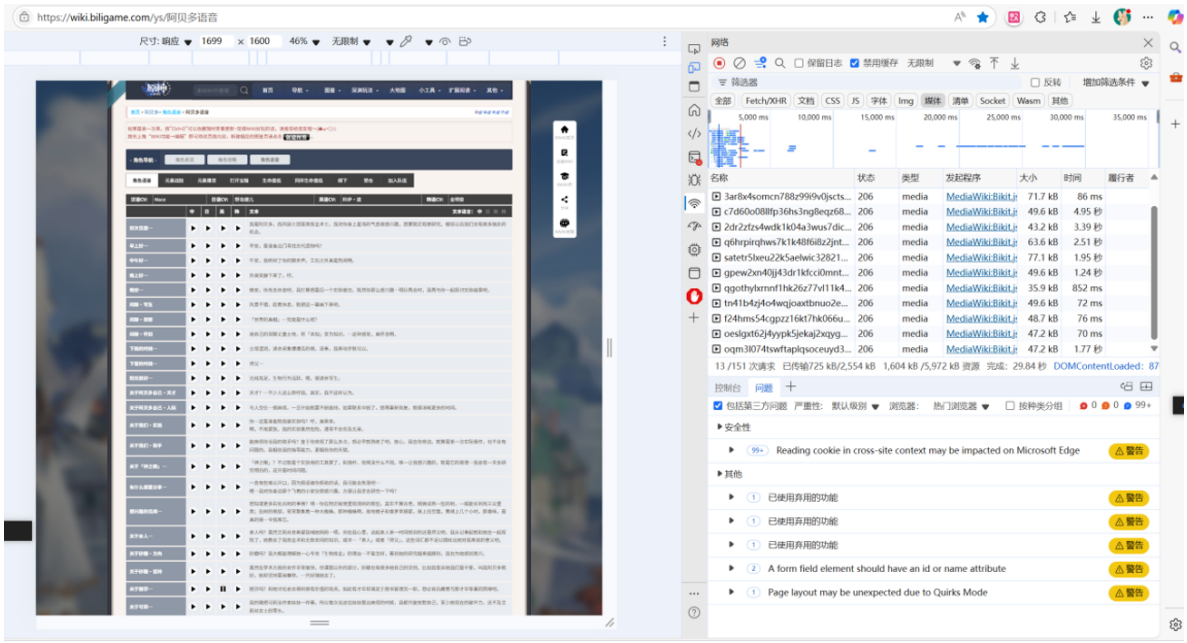
- 对低音质参考音频合成出来音质更好
- 底模训练集增加到5k小时，zero shot性能更好音色更像，所需数据集更少
- 增加韩粤两种语言，中日英韩粤5个语种均可跨语种合成
- 更好的文本前端：持续迭代更新。V2中英文加入多音字优化。

GPT-SoVITS-V3V4新增特点：

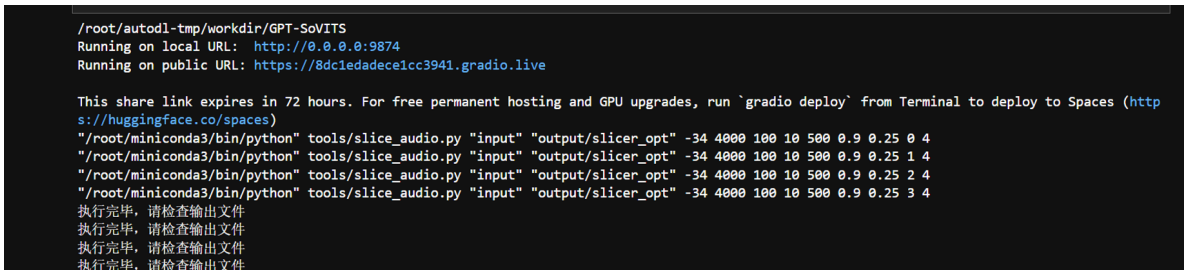
- 音色相似度更像，需要更少训练集来逼近本人（甚至不需要训练SoVITS）
- GPT合成更稳定，重复漏字更少，也更容易跑出丰富情感
- v4修复了v3非整数倍上采样可能导致的电音问题，原生输出48k音频防闷（而v3原生输出只有24k）。作者认为v4是v3的平替，更多还需测试。

2 音源训练集

1.前期训练集收集工作：
收集音源干声。总共收集大约10分钟的音频即可训练出较好的效果。



2.音频切割：
将时长超过 显存数 秒的音频手动切分至 显存数 秒以下。比如显卡是4090 显存是24g，那么就要将超过24秒的音频手动切分至24s以下，音频时长太长的会爆显存。



3. 标注
   fast whisper是目前最好的英语和日语识别，使用整合包里的ASR脚本即可对处理过的音频进行标注，给每个音频配上文字，这样才能让AI学习到每个字该怎么读。

```
output/slicer_opt/12rbi22pgrh6ajt1tsx12eg4d6r6gm6.mp3_0000008320_0000153600.wav|slicer_opt|EN| Citrinitas is the final stage of the alchemical transmutation process.
output/slicer_opt/12rbi22pgrh6ajt1tsx12eg4d6r6gm6.mp3_0000153600_0000289280.wav|slicer_opt|EN| The meaning of the object being transmuted has finally been brought to light.
output/slicer_opt/12rbi22pgrh6ajt1tsx12eg4d6r6gm6.mp3_0000289280_0000455360.wav|slicer_opt|EN| Becoming gold and revealing its true value, I too have found my own meaning.
output/slicer_opt/1e0slujcefosvh1yggomutn44d04iou.mp3_0000006720_0000156160.wav|slicer_opt|EN| Oh, Xingqiu. I find his written works quite interesting.
output/slicer_opt/1e0slujcefosvh1yggomutn44d04iou.mp3_0000173760_0000315520.wav|slicer_opt|EN| The Yai Publishing House in Inazuma has been asking that I cooperate with a-
output/slicer_opt/1e0slujcefosvh1yggomutn44d04iou.mp3_0000315520_0000420160.wav|slicer_opt|EN| different author for greater royalties.
output/slicer_opt/1e0slujcefosvh1yggomutn44d04iou.mp3_0000441920_0000552640.wav|slicer_opt|EN| Do I seem the type to be swayed by a few extra mora?
output/slicer_opt/1rfuvrhx11ssgllxs5xlue5sgzpuu2p.mp3_0000006720_0000208320.wav|slicer_opt|EN| Good night. You go ahead and rest. I will just finish one last experiment before bed.
output/slicer_opt/1rfuvrhx11ssgllxs5xlue5sgzpuu2p.mp3_0000224960_0000361280.wav|slicer_opt|EN| If you are interested in their results, I can discuss them with you tomorrow.
output/slicer_opt/2dr2zfzs4wdk1k04a3wus7dica1e385.mp3_0000012160_0000148800.wav|slicer_opt|EN| You want to accompany me while I experiment?
output/slicer_opt/2dr2zfzs4wdk1k04a3wus7dica1e385.mp3_0000168000_0000368320.wav|slicer_opt|EN| I'm honored. Oh, don't be nervous. My experiments may be dangerous, but no one gets hurt.
output/slicer_opt/2dr2zfzs4wdk1k04a3wus7dica1e385.mp3_0000391680_0000429760.wav|slicer_opt|EN| Most of the time.
output/slicer_opt/3ar8x4somcn788z99i9v0jscts767xt.mp3_0000004160_0000155840.wav|slicer_opt|EN| "'Genius? A number of people call me that.'
output/slicer_opt/3ar8x4somcn788z99i9v0jscts767xt.mp3_0000181760_0000270080.wav|slicer_opt|EN| But I don't think I'm any genius.
output/slicer_opt/4sr4y4vcpgzwethlusffc2e9ovo2lx7.mp3_0000007040_0000131840.wav|slicer_opt|EN| "'Happy birthday. You look especially happy.'
output/slicer_opt/4sr4y4vcpgzwethlusffc2e9ovo2lx7.mp3_0000131840_0000285440.wav|slicer_opt|EN| Would you mind if I sketched you? The capacity of our brains is limited.
output/slicer_opt/4sr4y4vcpgzwethlusffc2e9ovo2lx7.mp3_0000285440_0000486080.wav|slicer_opt|EN| So we are bound to forget things. But when an image is transferred onto paper or canvas,
output/slicer_opt/4sr4y4vcpgzwethlusffc2e9ovo2lx7.mp3_0000489280_0000699200.wav|slicer_opt|EN| The sketch becomes an extension of our memory. We can remember that past feeling when we later look at the sketch.
output/slicer_opt/5q7yy133zm46novvtv3ytu6failiulp.mp3_0000005440_0000138560.wav|slicer_opt|EN| I really do enjoy having dessert. How can I put it?
output/slicer_opt/5q7yy133zm46novvtv3ytu6failiulp.mp3_0000153920_0000345280.wav|slicer_opt|EN| When both physical and mental capacity are spent, high energy materials further provide a kind of
```

4. 训练
   Windows支持 CUDA 的 NVIDIA 显卡，训练要求每张拥有至少 8G 以上显存。因为显存不够，所以到AutoDL上租卡训练。

```
nd UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly.  To access UntypedStorage
directly, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/root/miniconda3/lib/python3.10/site-packages/torch/_utils.py:776: UserWarning: TypedStorage is deprecated. It will be removed in the future a
nd UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly.  To access UntypedStorage
directly, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
"/root/miniconda3/bin/python" GPT_SoVITS/prepare_datasets/3-get-semantic.py
"/root/miniconda3/bin/python" GPT_SoVITS/prepare_datasets/3-get-semantic.py
<All keys matched successfully>
<All keys matched successfully>
"/root/miniconda3/bin/python" GPT_SoVITS/s2_train.py --config "/root/autodl-tmp/workdir/GPT-SoVITS/TEMP/tmp_s2.json"
phoneme_data_len: 116
wav_data_len: 116
100%|██████████| 116/116 [00:00<00:00, 88349.24it/s]
skipped_phone:  0 , skipped_dur:  0
total left:  116
loaded pretrained GPT_SoVITS/pretrained_models/gsv-v2final-pretrained/s2G2333k.pth <All keys matched successfully>
loaded pretrained GPT_SoVITS/pretrained_models/gsv-v2final-pretrained/s2D2333k.pth <All keys matched successfully>
start training from epoch 1
  0%|                                                    | 0/13 [00:00<?, ?it/s][W reducer.cpp:1300] Warning: find_unused_parameters=True was
specified in DDP constructor, but did not find any unused parameters in the forward pass. This flag results in an extra traversal of the autog
rad graph every iteration,  which can adversely affect performance. If your model indeed never has any unused parameters in the forward pass,
consider turning this flag off. Note that this warning may be a false positive if your model has flow control causing later iterations to have
unused parameters. (function operator())
100%|██████████| 13/13 [00:16<00:00,  1.31s/it]
100%|██████████| 13/13 [00:06<00:00,  2.13it/s]
100%|██████████| 13/13 [00:06<00:00,  2.13it/s]
100%|██████████| 13/13 [00:05<00:00,  2.23it/s]
```

训练之后会得到两个权重：

jqr_ZH_e10_s490_l32.pth	2025/5/4 7:59	PTH 文件	73,780 KB
jqr_ZH-e10.ckpt	2025/5/4 8:00	CKPT 文件	151,673 KB

3 实时推理

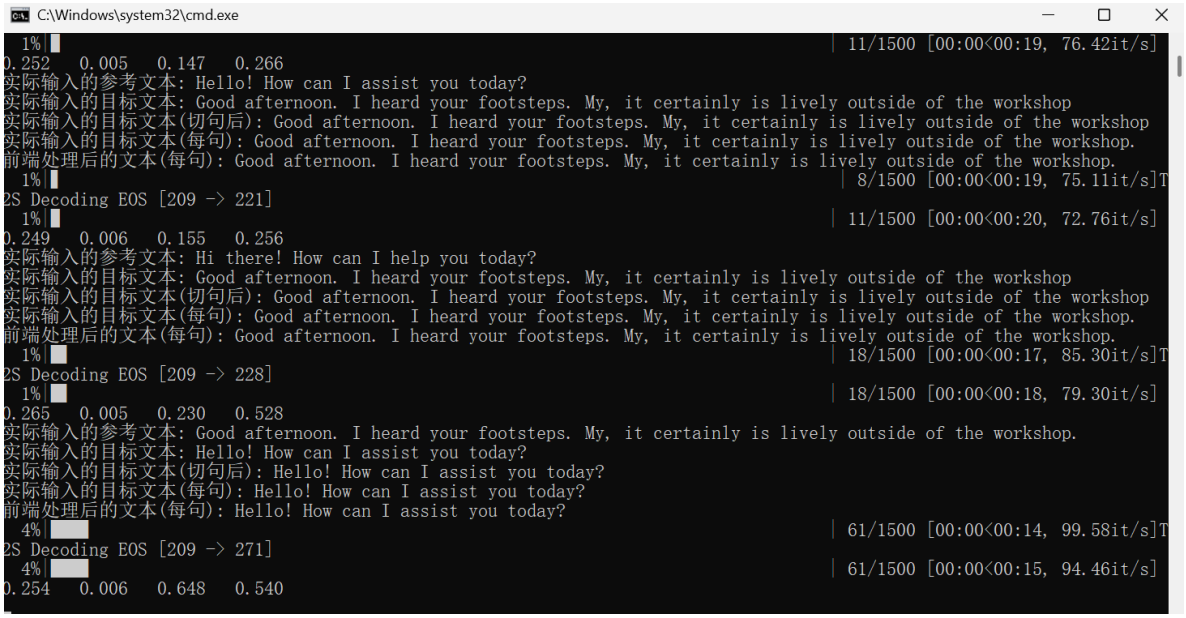
1.放好权重文件：

`GPT_weights_v2`：放.ckpt文件

`SoVITS_weights_v2`:放.pth文件

2.参考音频（一定要准备好）

上传一段参考音频，建议是数据集中的音频。**最好5秒。参考音频很重要！**会学习**语速和语气**。建议选择**有参考文本模式**，如果是无文本参考模式很容易出糟糕的推理结果。



关于top_p,top_k和temperature

这三个值都是用来控制采样的。在推理的时候要挑出一个最好的token，但机器并不知道哪个是最好的。于是先按照top_k挑出前几个token，top_p在top_k的基础上筛选token。最后temperature控制随机性输出。

比如总共有100个token，top_k设置5，top_p设置0.6，temperature设置为0.5。那么就会从100个token中先挑出5个概率最大的token，这五个token的概率分别是（0.3，0.3，0.2，0.2，0.1），那么再挑出累加概率不超过0.6的token（0.3和0.3），再从这两个token中随机挑出一个token输出，其中前一个token被挑选到的几率更大。以此类推。

> 拉满当赌狗，拉低当复读机。

4 效果

> 以下语音纯属TTS合成：

5 下一步工作：

- API接入Unity实现TTS推理
- 实现语音输入-对话功能。