AI分离与变音

# 前言

这个算是一本复习笔记喵，也是工具书喵，所以略过了很多基础喵，且衔接其他内容喵

由于我并非专家，并且技术有时效性，如有错误等请多多包涵喵。。。

-by古神喵姆

# 音源与音质

首先就是你要去哪里获取原始素材（音源），以及原始素材的音质，和它的分离大概难度

遇到想做的歌，首先肯定是下载它

最好的是下载源文件，有时候运气好可以在浏览器搜索到

例如部分国外音乐平台允许下载源文件，以及部分网盘有源文件

b站可以用[jijidown](https://www.jijidown.com/)，但是我个人还是喜欢用[Bilibili Downloader 命令行下载器](https://github.com/nilaoda/BBDown)

油管等国外平台可以用[cobalt流媒体下载](https://cobalt.tools/)

还有[网易云VIP解析](https://tools.qzxdp.cn/wyy_vip)之类的音乐平台下载工具

可以先试试网络搜索有没有这个平台的下载工具，或者可下载的平台有没有

实在没有办法下载再考虑录音录屏，例如[OCAM录屏](https://ohsoft.net/eng/)

不同的方式，下载的音质也分高低

通常

源文件>音乐平台>视频平台>录音录屏

国外平台，在都不付费的情况下，通常音质高于国内平台

有时候甚至付费了也是国外平台更高

而且要注意的是国内平台有很多乱七八糟的音质标准，一般是经过自动混音/AI超分后的，反而音质更低

所以下“无损”即可，最好看看关于音质的讲解，避免踩坑

例如网易云音乐：

与常见的双声道音频相比，5.1声道除了保留原信号和方向，还伴随产生围绕感和扩展感，即声音离开时的扩散和混响。

沉浸环绕声通过网易自研创新技术，运用AI分离母带音频的乐器、人声、伴奏等，模拟5.1声道，一个正面声道、左前方声道、右前方声道、左环绕声道、右环绕声道、一个低音声道，并且无损导出音频，实现身临其境的听感体验。

音质高低如何判断？

音频分成两种类型

有损压缩，例如视频文件内的音频（例如mp4里的就是AAC，也叫ACC），还有大名鼎鼎的mp3

这种主要看音频码率来辨别音质

无损压缩，例如wav、flac

|  |  |  |
| --- | --- | --- |
| 参数 | 用途 | 说明 |
| 采样率 | 在将模拟音频信号转换为数字音频时，每秒对信号进行采样并记录其振幅值的次数 | 采样率直接决定赫兹数，低了音质就差  例如高频缺失、声音模糊  常见44.1khz、48khz |
| 位深 | 数字音频中每个样本所使用的比特数，位深度直接影响音频的动态范围 | 越高声音大小越准确  影响信号噪声比，低了音质糊  常见16、24、32 |
| 声道数 | 声音录制时的音源数量或回放时相应的扬声器数量 | 单声道只能出一个声音  双声道声音会有左右的区别  多声道可以上下左右前后 |
| 码率 | 音频码率是指在单位时间内传输的音频数据量，通常以kbps（千比特每秒）为单位 | 采样率x位深x声道数量=无损音频码率  有损音频常见从96~320kbps |

总之就是，一般大小越大越好，在数据有效的情况下

例如原本只有16位，但是扩到了32位，这种是虚假的扩大

还有转码不同的情况下导致音频大小增加了

那这些说完了，如何判断音频分离的难度呢？

总而言之，人声越多越杂，人声音质越差，难度就越高

有些难驾驭的歌也许可以放弃掉，或学学虚拟歌姬再来尝试

# 分离与推理

由于分开讲解不能理解操作的必要性

所以我在这里选择合并讲解两个章节的内容

首先一开始你肯定要先把伴奏分离开，随便选个模型即可（后面再讲细节）

## 人声区分

分开伴奏后，主唱便是听起来最明显的部分，大伙听AI翻唱最主要就是听个主唱，其他像什么和声全是锦上添花

.

一般来说，没有“干扰元素”的主唱直接就按照默认参数推理即可

至于“干扰元素”，我个人是按照与主唱的一致程度进行区分类型的，因为这会对后续处理产生影响

|  |  |  |  |
| --- | --- | --- | --- |
|  | 音高一致 | 节奏一致 | 咬字一致 |
| 合唱 | √ | √ | √ |
| 平行和声 | X | √ | √ |
| 对位和声 | ？ | ？ | ？ |
| 人声采样 | ？ | ？ | ？ |

音高一致：发音高度一致

节奏一致：歌唱节奏一致

咬字一致：歌词一致

## 合唱

首先就是合唱，这里指狭义的合唱，如上表格所述就是多个人唱同样的歌声，这个怎么处理要看人数如何，以及他们所有人的音准

AI变音需要识别基频来确定该唱多高的音，需要识别共振峰来确定发什么音，所以人数越少越好，歌声越一致越好

如果比较清晰，可以直接推理尝试一下

也可以尝试使用分离和声的模型

可能扣多了，也可能扣少了，试着推理一下，用耳朵听效果如何

如果是合唱效果器、镶边效果器，也是同理操作，但是都有模糊的可能性，分离和声可能不起作用

过于模糊的话，分离和声后再试试降噪模型，如果还不行，除了虚拟歌姬调一个人声出来再转换（进阶内容），就只能放弃了

## 平行和声

平行和声，如上表格就是类似合唱的东西，区别是它使用了不同的音高

这个对比合唱，虽然普遍人数更少，甚至一般就一个人唱和声，但是音高是完全错开的

所以必须使用分离和声的模型，如果效果不佳再尝试一下分离和声后降噪

如上所述，因为AI变音识别到了发出的音不止一个，所以懵圈了，只能多选1，导致了哑音

## 对位和声

虽然对位和声唱的歌词、旋律和主旋律有着不同，但与平行和声共同点是，听起来和整体歌曲是和谐的、相同走向的，所以也是先试试分离和声，不行再降噪

两种和声都有可能有时会是合唱，或者是说合唱可能会分出一部分人来唱和声进行，这样推理和声的可行性大大减少了，结果音质普遍较差

这种情况一般是大合唱，例如天主教那种小男孩唱诗班

还有另一种情况，我特地在这里做个区分，那就是多重（平行）和声

我这里再重新讲一次，区分狭义的“合唱”与“多重（平行）和声”关键点在于“音高”是否一致

至于对位和声也是同理，不属于“合唱”（例如对位和声有它的平行和声）

如果同时包含两重人声以上时，以当前技术，对除了主唱外的人声进行再次分离，是很困难的

切换和声分离模型也没什么效果，所以一般只能选择直接放进去

如果硬是要推理音色，要么使用虚拟歌姬，要么强行修音（本笔记内容均提过这些东西）

## 采样

而采样则是切进去的各种人声，例如什么“3 2 1 go”，又或者是主唱切的碎碎的放进去，在电子乐里面很常见

这种也是尝试一下分离和声，不行再降噪

## 其他

部分人声会使用各种各样的效果器，或音源本身就比较“奇怪”，一般多见于“采样”之中

比如

失真效果，整个人声都模糊了的那种，用降噪模型

降噪模型对于人声素材过老，所以音质差，或者是推理后结果很电却又不得不用的情况也是有效果，如果不行就换换模型

高频缺失的素材，也是常见一起的毛病，因为音频采样率低导致的，使用频谱修复，例如RX11，光谱层11都有此功能

声码器/机器人声，这种暂时没有什么好的处理方法，直接推理试一下，不行就上虚拟歌姬

活字印刷/老式虚拟歌姬（例如utau）（这类的），推理后咬字通常奇怪，因为原本就咬字风格独特，可尝试特征检索/扩散模型提高咬字还原度，或者分离伴奏时受阻，尝试可切换模型，如果有需求最好上现代点的虚拟歌姬

还有些杂音去不掉

可以试试SpectraLayers（光谱层），可以拆人群噪音，拍手啊背景的欢呼声什么的，浅浅提一嘴

光谱层还有一个功能，人声有不干净就用光谱层手动抹除上面的杂音

以及比如msst有气声分离模型，分离谐波和不和谐泛音的部分

在混音阶段，齿音/电音导致实在救不回来了的情况下，可以尝试此方法挽回

## 回音

以及你们是否留意到分离后的人声可能包含的“回音”？或多或少有点对吧？

这种声音分成两种类型，一种叫混响，一种叫延迟

混响是再现自然空间中声音的回声，延迟的目的是延迟声音，产生声音重叠的效果（摘自网络）

有些去混响模型不会去延迟，有些则会，延迟也会导致转换后有杂音，混响如果较大也是一样的

延迟导致的杂音几乎和哑音同理，混响导致的杂音几乎和合唱同理

以及部分的去混响模型，会或多或少的把和声扣走（尤其对扒谱时造成影响）（重点点名MSST）

所以如果你要求获取更干净的主唱，可以先扣主唱的混响，然后再扣主唱的和声

反之调换一下步骤把和声先扣走，然后用其他模型（例如VR模型）扣走和声的混响（扣不掉就算了）

最后把降噪（如果你尝试了）、和声两个音轨合并在一起，估计就是完整的和声了

## 伴奏

以及一个重要点，部分伴奏分离模型是会把和声分离到伴奏里的（重点点名MSST）

如果你要保留和声到人声，然后分出来转换，又或者扒谱（虚拟歌姬的必须技能点之一），那么要使用更不会这样的模型

如果你不在乎和声，或者试过了更适合放送原和声，这样的模型可能会让主唱音质有些微提升（真的很少，因为MSST的分离和声已经够强了）

为什么要干净的和声？有时候为了追求音色的统一性，较为方便的一个方法就是直接转换和声的音色，然后放进去

如果实在分不干净，如上所述，去学学虚拟歌姬吧

## 对唱

哪怕没有和声了，那多人分开唱不同的部分呢？比如男女对唱，这种就按主唱看就好了

在只有一个人在唱时，这个声音就是主唱，如果有多个人同时在唱的话，按照上面的标注来分类就可以

例如

1：00~1：10秒仅男声，那么男声算“主唱”

1：15~1：25秒仅女声，那么女声算“主唱”

2：00~2：30秒有男声女声同时在唱，按照上面讲的内容分辨是合唱/平行/对位和声？然后进行处理

## 音域范围

女声的音域高，男声的音域低

男声转换男声、女声转换女声一般是没有问题的

但是男转女、女转男可能会不够

最先应该考虑的是音高增强器之类的参数，让模型适应更高/更低的音域

然后是考虑加减key，也就是变调

不熟悉乐理、乐感差的小伙伴们一般都会使用非12整倍数的变调（12key=一个八度）

但是这会导致歌声与伴奏严重的不和谐，也就是跑调（但是如果是说话、采样什么的就当我没说，你高兴就好，唯独rap可能要注意会不和谐）

## 和声修音

以及变调的高级玩法，如何不更改原本的音高的情况下让声音适应到如此（本小章是超纲内容，关于修音的，就提一两嘴）

例如常见的唱不上高音，首先降调推理，然后再使用麦乐迪类的修音软件改回原位

首先要注意的是这类软件会有一定的电音，例如麦乐迪最好不要超过3key

然后就是要微调一下共振峰，听感觉微调，和修改前的版本比对一下自然度，例如麦乐迪里是提升就微降一点共振峰，（调多了就会超好玩的）

如果你要加减12key，那么一定要注意和其他人声的契合度，最好是全部人声一起升降

如果你的和声比较高或者低，那么其实这种哑音了不明显，尤其是存在感比较淡的情况下

还有一个更进阶的玩法，使用修音修改主旋律为平行和声（这就是为什么我前面要求你分清楚和声之间的区别）

也是使用麦乐迪，然后利用乐理知识、参考原音频、乐感判断和谐

把分离完的主唱丢进去修到跟和声一样的音高

## 预测器

这里着重讲解一下AI变音的运行步骤，这个是很重要的点

1.丢进去AI翻唱的音频会被自动或者强制切片成一段一段的大小，免得把你的运行内存挤爆

2.然后预测器逐字识别它们的咬字，以及它们的音高

3.由AI变音模型开始逐个推理

4.合并剪辑成一整段，保存

上面之所以要讲如何最可能干净的分出主唱的声音，但是有时还是可能不够

预测器也是可以切换的，不同的预测器有不同的识别倾向，它们的识别结果是有可能不同的

比如减八度的和声有时去和声的模型不太能识别到，但是预测器认为发的应该是减八度和声的音，导致了突然音高下坠

但是切换了一个预测器这段就不会如此了，但是其他段还是原本的那个预测器好点

## 音频剪辑

并不是只能选一个版本来用的，你可以同时使用多个版本的推理结果，使用它们之中你认为最好的段落拼凑成一整首歌

并且一首歌中有时会重复一个片段多次，也就意味着这一段的也许不行，从另一段的切过来即可

一般我推荐你使用专业的音频编辑软件来做这种事情，比如FL、cubase之类的，我个人就是从FL入门的

但是你实在不想要下的话那么视频剪辑的软件也未尝不可，只是更不方便而已

对新手来说其中最重要的是把主唱和伴奏对齐、歌声也要记得对齐，剩下的没了，按照感觉来就好了

## 文件管理

我个人喜欢先把东西堆在桌面，然后按照它们的处理步骤排序，确定不需要了才删除

当然你也可以单独开个文件夹放分离后的内容，以及再开个文件夹放推理后的内容（同理，虚拟歌姬扒人声完的内容，如果有的话）

我一般在进行中/打包时这么做，方便分辨文件/返工修改

## 模型挑选

模型（技术）挑选一般都是挑选最新的，最新通常是最好的

比如msst就比uvr强一截，sovits4.1就比4.0强一大截

然后就是大小最大的，这意味着训练资料最多

比如msst的BS-Roformer（705mb）大小就没有big\_beta5e（1.37gb）大，BS会把伴奏的和声扣漏，beta5e就不会

以及训练最针对性的，多茎分离天生不如同期技术的单茎分离

### 分离

我首先推荐各位使用MSST

然后还有些工具例如ripx、RX、SpectraLayers（光谱层）、UVR，也可以放在备选名单里就是

我只能说各有特色，比如ripx可以同时分析midi，RX有很多功能模块、SpectraLayers可以直接编辑频谱

但是记得留意一些参数，例如msst的TTA、重叠数、或者是SpectraLayers的分离精度（可选快速~极端）

还有性能设置，别傻乎乎有显卡却用CPU推理

以及有N卡的不要忘了cuda驱动要装一下，其他显卡同理

### 推理

截至我补充此部分时，现在SVC里面最热门的无非是RVC、so-vits4.1、DDSP6.1这几个

RVC特征检索让声线更还原、可实时（抗噪差）

so-vits抗噪好，上限高（音域低、素材量/配置门槛高）

ddsp训练门槛低（素材量/配置），音域高、也可实时（抗噪一般，上限比sovits略低）

## 扒带辅助

人声的拆分教完了教教拆乐器的

大多数情况下，一首歌里面有鼓和贝斯对吧

而且鼓和贝斯可识别性也强

所以我如果要扒带的话

用AI分离先去掉目前分离技术最成熟的人声（msst，人声分离，big\_beta5e）

然后再分离一份鼓和贝斯（msst，单茎分离，HTDemucs4\_FT\_bass / HTDemucs4\_FT\_drums）

这时候关键点来了，音频是震动产生的对吧，如果有一个完全相反的波形，那么声音就会被抵消

降噪耳机就利用了这种原理，智能分析边上的噪音，然后再播放一个相反的动静

你只需要输出一个音轨，例如drums，然后拖入DAW（数字音频工作站），比如FL里面开启相位反转，然后跟原曲放一起，这样原曲的drums就没了

这个方法对比把其他的内容轨合并起来更节约空间

以及在这里推荐一个拆分鼓组的模型msst 多茎分离 aufr33-jarredou\_DrumSep\_model\_mdx23c\_ep\_141\_sdr\_10.8059

可以把鼓组拆分来方便扒谱，针对架子鼓组效果好

现在拆分最稳定的三个都拆开了，剩下的相比之前肯定好扒了

剩下的也可以用msst，据我所知可以用多茎分离HTDemucs4\_6stems，可以多拆一个钢琴一个吉他

或者光谱层也可以拆吉他和钢琴，以及一个铜管乐器（萨克斯、小号等）

质量略高一点但是半斤八两，毕竟是专门做乐器起家的，训练集绝对够

以及ripx，可以拆的乐器也挺多，好像是最多的一个，但是质量我不做评价了

光谱层还有一个很逆天的功能，先把需要的东西分离出来，如果有不干净就用光谱层手动抹除上面的杂音

一般是一些高阶的母带/分轨混音/音编会用到这个

或者是扒带提取音色

## 模型训练

### 素材准备

在下载你所需的素材后，模型训练的素材一般分成以下几种

什么？你问我需要准备什么素材？当然是包含这个角色声音的音频文件啊

至于AI分离模型？这个训练门槛太高了

有背景音乐的：

这种按照上面的乐器分离，仅保留主唱即可，然后切出ta唱歌的部分，但是音质通常不如其他情况

有其他人说话的部分：

这种一般分离和声都分不出来，暂时自动识别音色的工具也不太好用，只能靠手动切，只留下这个角色的声音

如果要在发音的中间切断最好是在音量降至最低的那么一瞬间切，或者是拉一个渐入/渐出

有其他杂音的：一般先用分离伴奏，然后看情况给降噪/分离混响等，如果还是不干净就上光谱层一类的软件自己清理/切除这段

最好的情况：从游戏里解包的源文件，但是要注意不要混进去其他角色的声音，可能会出问题，尤其是抗噪弱的模型种类

如果你是准备录音可控的训练集，比如自己、家人、朋友的训练集（也适用于其他AI变音模型的要求）

最好高音低音都有，满足模型的音域需求，而且咬字发音要丰富，各种情绪都要有，这样的模型短板才会少

录音环境尽量安静无干扰，麦克风质量要好点，注意保持和麦克风的距离，不要太远太近，导致声音大小忽高忽低，更不要喷麦

在剪辑裁切的时候，一定要使用无损的方法，例如专业的音频编辑软件，例如FL、cubase

然后再考虑剪辑软件，因为视频导出、转码过程一般会导致音质下降很多

一定要把 除了你需要角色的语音 以外的所有声音 都清理干净

最后把音频文件重命名后丢进自动切片，素材准备就完成了喵！

如果你自动切片后要批量重命名

音频名称不是xxx\_0001这样的排序会导致强迫症患者爆死，或者甚至无法训练

**$i=1**

**Get-ChildItem | Sort-Object LastWriteTime | ForEach-Object {**

**$newName = "xxx\_{0:D4}{1}" -f $i, $\_.Extension**

**Rename-Item $\_ -NewName $newName**

**$i++**

**}**

这个脚本把其中xxx的位置改了，改成你需要的英文字母

然后把你的音频文件丢到同一个文件夹，然后右键文件夹打开终端（Windows PowerShell）

右键粘贴该脚本然后回车确定，这样就批量命名完了

### 训练和成品挑选

上面我们讲完了训练素材，直接下一步

跟着软件自带的教程（或者网络自行查找）走完，参数也是看着它的说明、你电脑的性能来调

如果性能不够请使用云端/去网吧

现在很多模型自带一些增强音质的辅助模型，例如扩散模型/特征检索模型

如果你找的模型没有，而这个软件有，就用通用的配置文件

如果你训练集素材太少，少练一点，越少练的越少，但是不要太少了，免得拟合不彻底，可以跟着自带的教程看它的loss值等参数，但是仅参考

可以把它保存后再试着推理听听看，至于推理使用的歌声素材可以参考上面的内容

模型有几个需要留意的点是不同的

咬字，长的短的什么的咬字是否会识别错误，这个跟预测器设置有关系，也跟当初音频的清晰程度、受损程度也有关系

音域，这个模型最高/最低能唱多少高低

电音哑音等其他问题，这个就是整个流程不能落下了，水桶效益，不能有短板

从原本的“音质差”（未拟合），到“音质好”（拟合中），最后再回到“音质差”（过拟合），挑选一个你最喜欢的“拟合中”阶段的模型即可

当你练过之后就能大概估算出自己需要练多少步再停止，然后使用二分法（步数对半切）来逐个测试挑选

如果你训练集质量差点，可以再少练许多，依靠底模的质量来挽救你堪忧的模型

如果你训练集素材太多，甚至爆显存，你可以试着跟着教程里设置来降低占用，或者上云端什么的，最后再考虑忍痛割爱

素材多一定完全可以多练一会，才能发挥它的极限

素材的准备要精，不要多，而且也不要准备完全重复的复制素材来敷衍AI，例如复制粘贴文件、TTS喂出来二次训练等

这样和你增加训练时长没有本质差别

最后当你挑完，决定封丹大吉，不要忘记打包备份

原始素材、处理中工程、训练集、模型文件、模型配置文件（部分模型可以压缩，节约空间）

补充一句

如果你要最高的上限，可以试试微调底模（还有预测器）！

但是这个门槛很高，需要你有很大量的极其优质素材，以及较高的配置，还需要编程基础

不然效果不一定好

祝大家都能做出心目中的音乐，甚至达到发行级别水平！