Audio_driven 总结

一、小样本概况

BV号	类型	UP	备注
BV1FT411n79E	单人vlog	东尼ookii	
BV17b4y1m79y	多人vlog（其他人少）	东尼ookii	
BV1tJ411P7Cr	两人唱歌	老番茄	
BV1DK411H7Cm	多人玩游戏	小潮院长	
BV1BK411L7DJ	讲课	罗翔说刑法	几乎整个视频可用
BV1hU4y1g7sp	美食评测	真探唐仁杰	
BV1KK411V7MK	自拍剧	自来卷三木	

1. 多人视频要保证目标说话人说话至少50%以上
2. 感觉5秒可能有点过长，复杂场景目标说话人可能不会有连续这么长的说话时间
3. 有些唱歌、演讲的视频几乎整个都可用

二、数据观察结果

BV1FT411n79E（送外卖vlog）

176 204是大部分是其他说话人的声音，目标说话人说话很少
214 232是目标说话人小声说话
230 只有背景音
203 只有目标说话人的声音但是很少
总共256

BV1BK411L7DJ（罗翔讲课）

只有最后一个视频不可用，**共201条**

BV1DK411H7Cm（玩游戏）

31个视频开始说话人的声音很少了 **第50个视频几乎没有目标说话人声音 61没有目标说话人声音** 再之后不一定有没有目标说话人声音 90开始几乎没有说话人声音（要么一起说话要么很短） **共164**

BV1hU4y1g7sp（美食评测）

前40一直在说话 只有最后一个视频完全没人声 **共72**

BV1KK411V7MK（剧）

同一个人 不同口音（视频3 12 19 34）**36后面没说话人声音 共57**

BV1tJ411P7Cr（多人唱歌）

（没有人脸不适合我们的数据集，主要用来检测是否能在有背景音乐且在唱歌的情况下区分说话人）**第11个开始错了 共55**

BV17b4y1m79y（挑战-vlog）

前56全是纯目标说话人，**107**个视频不到一秒钟 **共294**

三、初步设想

1. 场景设定（是否需要划分场景存疑）

两个模态都好

讲课

声音模态较好

评测
VLOG

人脸模态较好

近景唱歌

两个模态都不一定很好

多人vlog
剧
远景唱歌

2. audio-drived 采集流程

分为**预筛**和**精筛**

预筛

1. 每个人寻找自己熟悉的UP主（**每个人不重复**），负责采集该UP主的视频，最好不同场景不同风格，人脸和声音都需要出现，视频中该up主所占部分至少为50%，给出选择视频的**BV号**
2. 给出音频的**置信度**
   高置信度 讲课等up主人脸和声音质量都很高的情况
   中置信度 如果是复杂场景vlog或多人拍摄该up占主导的情况
   低置信度 唱歌等声音环境较复杂的情况
3. 给出每个视频中**该说话人连续说话10秒**的一段音频用于注册

精筛

程序将根据不同置信度选择保留的段数并得到最后的结果，每个人负责检查自己选择的UP视频结果的情况