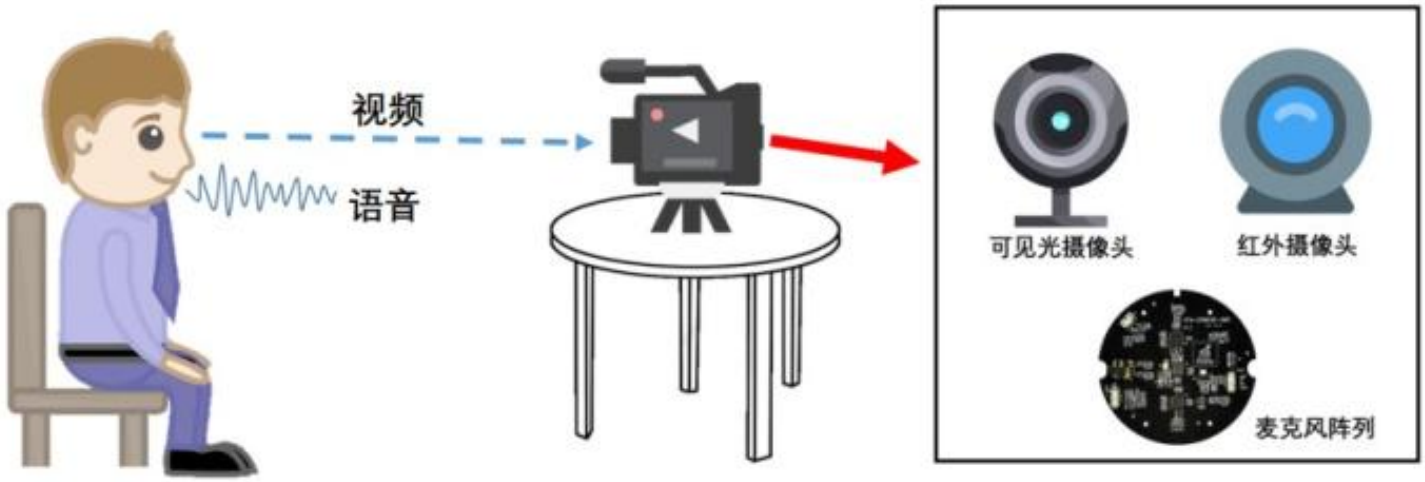
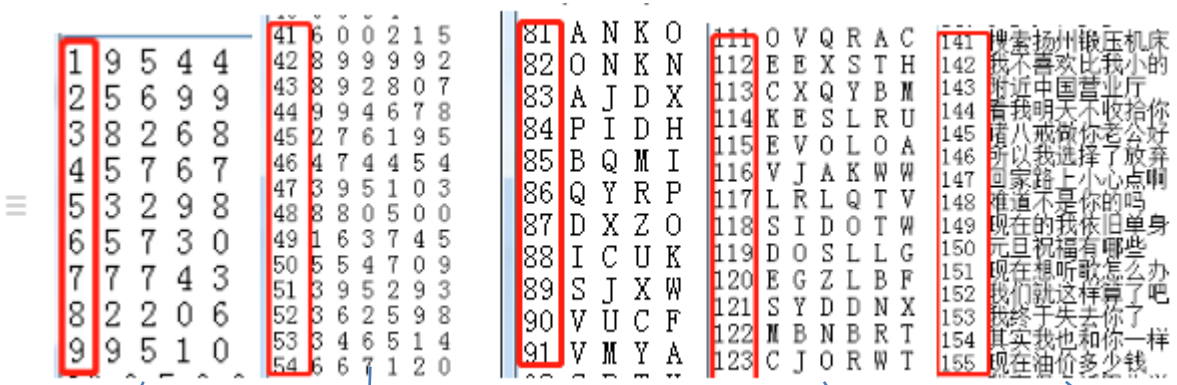
# Thermal_Speaking_2021

Shenjingxin
2021.09

# Collection process



# Randomly generated text(200utt)



utt_1-40(40)
随机的4位数字

utt_41-80(40)
随机的6位数字

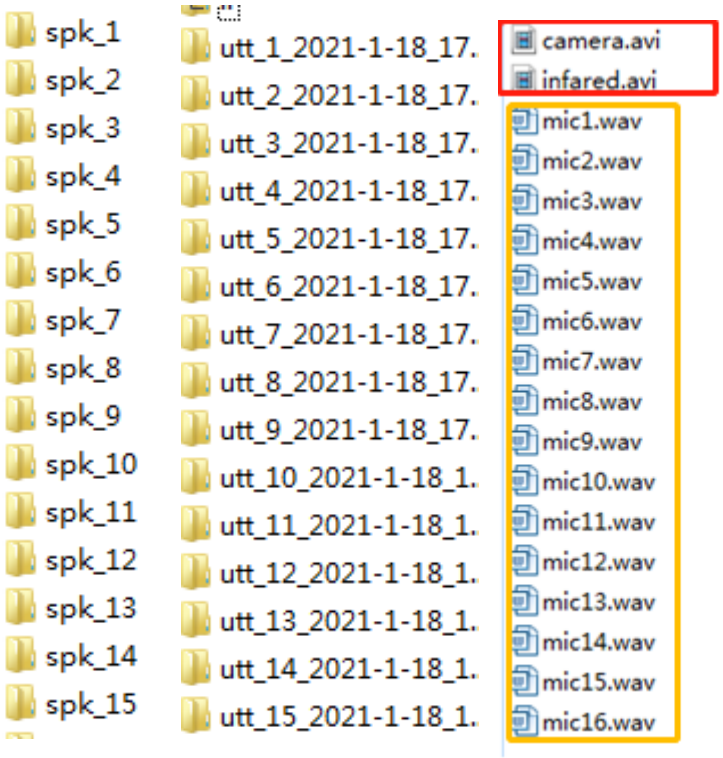
utt_81-110(30)
4位随机的英文字母

utt_111-140(30)
6位随机的英文字母

Utt_141-200(60)
随机的一句中文句子

# Structure of dataset

一共包含247个id：spk_1,spk_2,spk_3…….spk_252,其中缺少spk_17、spk_73、spk_96、spk_242、spk_251文件夹



# Preliminary inspection work

红外视频里人脸过大或者过小的问题
1 spk_8 utt_195-200
2 spk_55 utt_19
3 spk_55 utt_45

问题一：音频没有录制成功
1 spk_55 utt_78
2 spk_18 utt_1-200
3 spk_19 utt_1-200
4 spk_21 utt_135
5 spk_30 utt_80
6 spk_38 utt_2

问题二：出现了其他人的说话声
1 spk_2 utt_200
2 spk_3 utt_200
3 spk_4 utt_200
3 spk_5 utt_200
4 spk_6 utt_12
5 spk_6 utt_200
6 spk_7 utt_200
7 spk_8 utt_200
1 spk_9 utt_200
2 spk_10 utt_200
3 spk_12 utt_200
4 spk_13 utt_200
5 spk_14 utt_200
6 spk_15 utt_200
7 spk_16 utt_200
8 spk_20 utt_200
9 spk_21 utt_14
10 spk_21 utt_145-147
11 spk_22 utt_149
12 spk_22 utt_150
13 spk_22 utt_200
14 spk_23 utt_7
15 spk_23 utt_134-140
16 spk_23 utt_147

问题三：非人声的背景噪声过大
1 spk_2 utt_174
1 spk_13 utt_108
2 spk_21 utt_126
3 spk_21 utt_127
4 spk_22 utt_81
5 spk_22 utt_92
6 spk_22 utt_112
7 spk_22 utt_168
8 spk_23 utt_6
9 spk_23 utt_9-12
10 spk_23 utt_14-15
11 spk_23 utt_59
12 spk_23 utt_64
13 spk_23 utt_66-70
14 spk_23 utt_94
15 spk_23 utt_96-98
16 spk_23 utt_107
17 spk_23 utt_125-126
18 spk_23 utt_133
19 spk_23 utt_141-145
20 spk_23 utt_151
21 spk_23 utt_155-157
22 spk_23 utt_175-182
23 spk_23 utt_186
24 spk_23 utt_190-191
25 spk_23 utt_198
26 spk_24 utt_60-62
27 spk_24 utt_76-79
28 spk_24 utt_149-150
29 spk_24 utt_159
30 spk_24 utt_165
31 spk_24 utt_167
32 spk_25 utt_16-19

问题四：红外数据中出现了其他非当前说话人的图像

1 spk_12 utt_200
2 spk_15 utt_1
3 spk_18 utt_198
4 spk_18 utt_199
5 spk_19 utt_28-32
6 spk_19 utt_35-44
7 spk_28 utt_61-64
8 spk_29 utt_28-30

问题五：少读漏读或重复
1 spk_1 utt_42
2 spk_4 utt_167
3 spk_4 utt_177
1 spk_10 utt_134
2 spk_10 utt_150
3 spk_13 utt_3
4 spk_15 utt_160
5 spk_16 utt_164
6 spk_16 utt_165
7 spk_16 utt_168
8 spk_16 utt_170
9 spk_16 utt_179
10 spk_20 utt_150
11 spk_20 utt_152
12 spk_20 utt_180
13 spk_21 utt_159
14 spk_21 utt_126
15 spk_21 utt_173
16 spk_22 utt_128
17 spk_22 utt_133
18 spk_22 utt_160
19 spk_23 utt_48
20 spk_23 utt_62

# First step

1.查找缺失的spk　缺少5个spk_id

2.音频缺失　无声、打不开、文件缺失

| id | 缺失utt | 缺失utt条数 |
|---|---|---|
| spk_88 | utt_200 | 1 |
| spk_98 | utt_200 | 1 |
| spk_118 | utt_187 | 1 |
| spk_135 | utt_102(没声) | 1 |
| spk_209 | utt_169-200 | 32 |
| spk_218 | utt_96-200 | 105 |
| spk_220 | utt_186-200 | 15 |
| spk_223 | utt_87(没声)-200 | 114 |
| spk_227 | utt_140-200 | 61 |
| spk_229 | utt_121-200 | 80 |
| spk_233 | utt_151-200 | 50 |
| spk_243 | utt_121-200 | 80 |

3.视频缺失

Q1:文件打不开或者不存在

| 视频有问题的id | infrared ( 问题红外视频 ) | camera ( 问题可见光视频 ) |
|---|---|---|
| spk_22 | utt_31-111 | |
| spk_23 | utt_154-200 | |
| spk_26 | utt_100-169 | |
| spk_35 | utt_140-184 | |
| spk_41 | utt_43-109 | |
| spk_47 | utt_68-69 ; utt_132-139 | |
| spk_57 | utt_22-104 | |
| spk_58 | utt_16-26 | |
| spk_63 | utt_116-139 | |
| spk_86 | | utt_1-129 |
| spk_88 | utt_200 | utt_200 |
| spk_112 | utt_74-86 | |
| spk_118 | utt_187 | utt_187 |

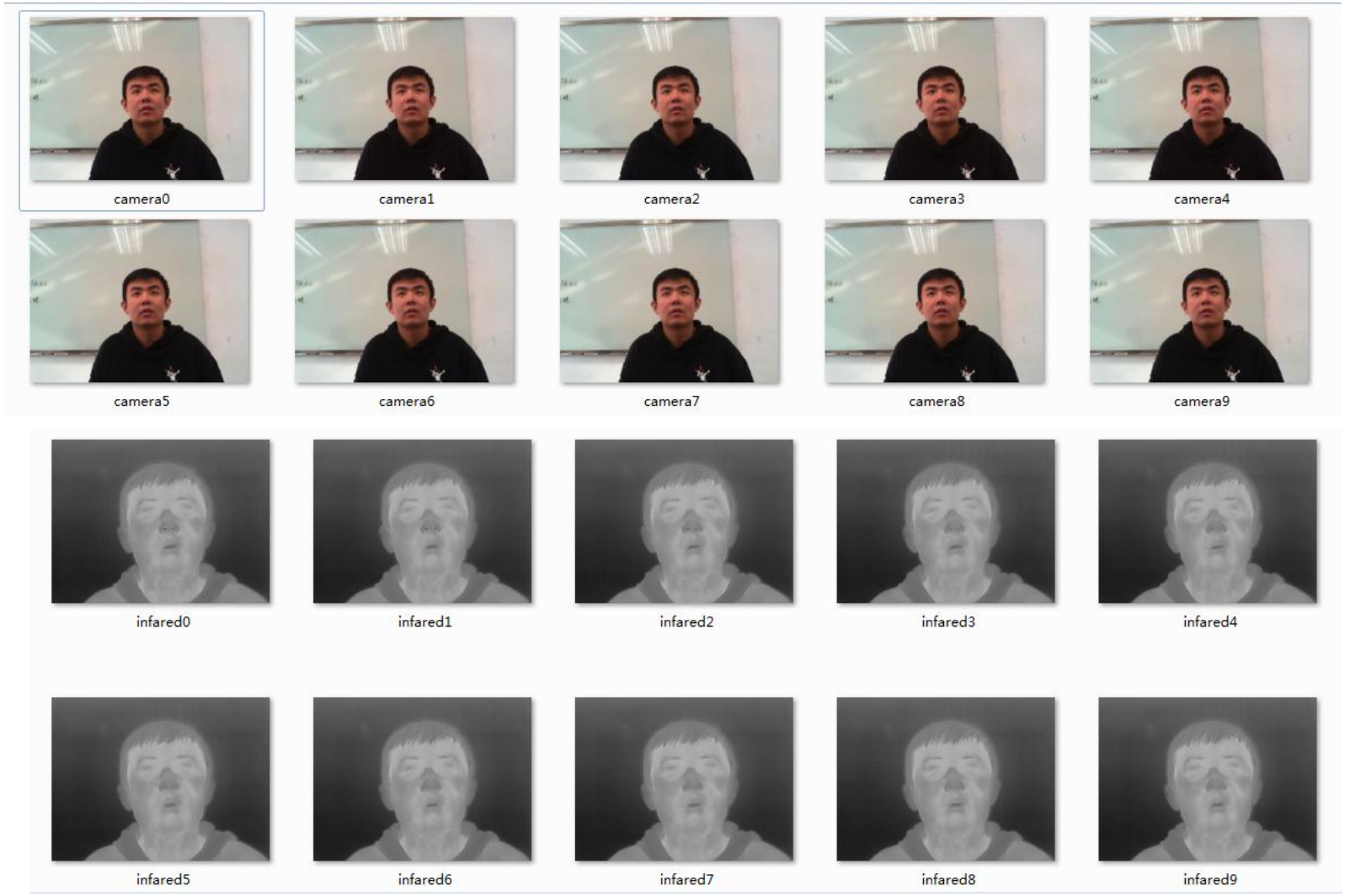
Q2：视频过短不足一秒

| id | infrared | camera | 备注 |
|---|---|---|---|
| spk_193 | utt_26 | | 转图片只有3张 |
| spk_234 | utt_664 | | 转了3张 |
| spk_150 | utt_16;utt_67; | | 只转了2张图片 |
| spk_178 | utt_8 | | 转了2张 |

问题音频：1117utt

问题视频：1030utt

# Check the video



# Scene classification

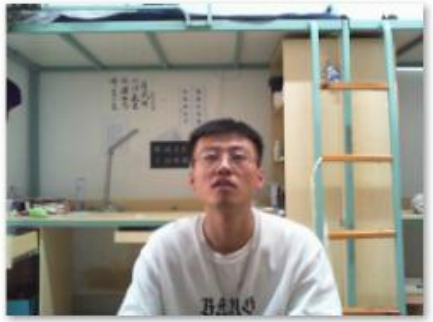
scene1

scene2

scene3

scene4

scene5



scene6

scene7

scene8

scene9

# Problem of video

1.视频文件无法打开

2.视频文件缺失

3.视频过短与音频对不上（转成图片仅仅几帧）

4.视频中出现其他人

# Check the audio-GMM-HMM

247spk语音总的 utt ： 48652，其中中文（数字和句子）34105utt；英文：14178utt

1 英文的语音识别模型

英文模型词错误率（wer）：2.9%
英文模型句错误率（ser）：4.78%

英文字母合计：233896个wav文件，解码的有233892个wav（14960个utt），针对能够解码的wav，解码后与参考文本不一致的有11191个wav文件（1021个utt）。
解码文本不一致率：6.8%

2 中文的语音识别模型

中文数字模型词错误率（wer）：2.01%
中文数字模型句错误率（ser）：4.03%

中文数字音频合计315904个wav，其中解码的有315904个wav（19744个utt），解码后与参考文本不一致的有9824个wav文件（614个utt）。
解码文本不一致率：3.1%

中文汉字模型词错误率（wer）：19.17%
中文汉字模型句错误率（ser）：27.90%

中文汉字音频合计229764个wav，其中解码的有229744个wav（14361个utt），解码后与参考文本不一致的有62576个wav文件（3913个utt）。
解码文本不一致率：27.2%

结论：

| 汇总 | 总数utt | 读错的utt | 正确的utt | 读错率 |
|---|---|---|---|---|
| 英文部分 | 14547 | 291 | 14256 | 2% |
| 中文部分 | 34105 | 1297 | 32808 | 3.8% |

# Problem of the audio

1.16路音频不全缺失的(部分缺失；全部缺失)

2.音频存在但打不开或者没声音

3.发音与文本是否相符，是否读错，重复，漏读等问题

4. 因噪音过大导致识别错

# Conclusion

可见光视频：247人，48590个视频
红外视频：247人，47863个视频

英文字母音频：244人，14256个utt
中文音频：238人，13661个utt
中文数字：245人，19175个utt

# Experiments of THS2021(DNF)

Dataset : 每个spk的每个utt中取5张图片，每个人200*5=1000张，随机选50个spk作为测试集，197个spk作为训练集

# Tsne:训练过程：上可见光下红外