《视听信息系统导论》课程大作业：

基于视频信息的说话人识别与分离

2021 年 11 月 29 日

一． 问题背景

我们在生活中常常会遇到多个人同时讲话的场景，比如在一个嘈杂的饭店里交谈，或者激烈地争吵等等。对于人类而言，从一段嘈杂的录音中准确捕捉到每个说话人的音频可能有些困难，但当我们面对面进行交谈时，就可以很容易地区分出每个人表达的信息，特别是当我们对说话人的身份很熟悉时。这说明视觉信息对于语音的分离是有益的。本次大作业提供 20 个说话人的若干视频，以及若干多人说话的视频，希望同学们逐步实现基于单人说话视频中视觉和声音信息的说话人识别，以及对多音源视频的逐说话人音源分离。

二． 数据集介绍

本次大作业的全部训练和测试视频来自 20 个说话人，其中男、女各 10 人。每个说话人拥有唯一 ID，依次为 ID1，ID2，...，ID20。本次大作业的数据集文件夹被命名为 dataset，内含 3 个文件夹，分别为训练数据集 train，线下测试集 test_offline，以及线上测试集 test_online（不公开）。每个文件夹的内容如下：

- train 文件夹：包含 20 个命名为 ID1~ID20 的子文件夹，每个子文件夹含有若干段（数量不等）对应 ID 的说话人的单声道单人说话视频（命名为%03d.mp4，分辨率 224*224）。
- test_offline/test_online 文件夹：包含 3 个任务的子文件夹，分别命名为 task1，task2（均对应说话人识别任务）和 task3（对应音源分离任务）。task1 文件夹包含若干段无声音单人说话视频（按顺序命名为%03d.mp4，分辨率 224*224）。task2 文件夹包含若干段单声道单人说话音频（按顺序命名为%03d.wav）。task3 文件夹包含若干段双声道多人说话视频（按顺序命名为 combine%03d.mp4，每段视频左中右三人同时说话，视频分辨率为 672*224）。

数据集的组织方式如下：

```
./dataset                                  根目录
|    ./train                               训练数据
|    |    ./ID1                            说话人编号
|    |    |    036.mp4                     单声道单人说话视频
|    |    |    017.mp4
|    |    |    ⋮
|    |    ./ID2
|    |    ⋮
|    ./test_offline                        线下测试数据
|    |    ./task1                          任务一
|    |    |    001.mp4                     无声音单人说话视频
|    |    |    ⋮
|    |    ./task2                          任务二
|    |    |    001.wav                     单声道单人说话音频
|    |    |    ⋮
```

```
|   |   ./task3                                  任务三
|   |   |   combine001.mp4                       双声道多人说话视频
|   |   |   ⋮
|   |   ./task3_gt
|   |   |   001_left.wav
|   |   |   001_middle.wav
|   |   |   001_right.wav
|   |   |   ⋮
|   |   ./task1_gt.json
|   |   ./task2_gt.json
|   ./test_online（组织形式同 test_offline）      线上测试数据, 不公开
```

另外, 本次大作业提供一份简短的样例代码用于定义接口, 以及待补全的算法核心代码, 详见 test.py

三. 任务描述

1. 任务一：基于视觉信息的说话人识别

单人说话的测试视频中包含说话人的人脸信息, 因此可以考虑仅根据视觉信息对说话人进行识别。**任务一要求对 task1 数据中的每一个无声音单人说话视频，通过与训练集中 20 个说话人视觉信息进行模板匹配或其它匹配方法，判断出对应说话人的 ID。**

为了方便测试，需要同学们参考 test.py 中定义的接口，在函数 **test_task1** 内部添加测试代码。test_task1 函数要求以 task1 的测试数据路径为输入，输出说话人识别的结果，结果以字典的方式保存，比如：{'001.mp4': 'ID3', '002.mp4': 'ID5', …}，以视频文件名为键，值为对应的说话人 ID（详见 test_task1 函数中的注释）。

2. 任务二：基于声音信息的说话人识别

除视觉信息外，说话人的声音信息也具有较高的特异性。与任务一类似，**任务二要求对 task2 数据中的每一个单声道单人说话音频，通过与训练集中 20 个说话人声音信息进行匹配，判断出对应说话人的 ID。**

为了方便测试，需要同学们参考 test.py 中定义的接口，在函数 **test_task2** 内部添加测试代码。test_task2 函数要求以 task2 的测试数据路径为输入，输出说话人识别的结果，结果以字典的方式保存，比如：{'001.wav': 'ID3', '002.wav': 'ID5', …}，以音频文件名为键，值为对应的说话人 ID（详见 test_task2 函数中的注释）。

3. 任务三：双声道多音源视频的逐说话人音源分离

结合音视频信息可以实现多人同时讲话时的音源分离, 任务三中所给出的双声道多音源视频, 其图像部分为各音源视频图像部分的拼接, 双声道音频部分为各音源视频中音频通过不同随机系数的加权组合。可以通过测试视频每个说话人对应区域进行说话人识别, 进而借助说话人信息对测试视频进行音源分离。**任务三要求对 task3 数据中的每一段双声道多音源视频，根据左中右 3 个说话人的信息，按顺序分离出每个说话人的音频。**

为了方便测试，需要同学们参考 test.py 中定义的接口，在函数 **test_task3** 内部添加测试代码。test_task3 函数要求以 task3 的测试数据路径和结果输出路径为输入，输出每个说话人的音频, 依照格式保存在结果输出路径中, 比如 001_left.wav, 001_middle.wav, 001_right.wav, 分别代表 combine001.mp4 中左中右三个说话人的音频。

同学们须严格按照函数 **test_task1**、**test_task2**、**test_task3** 定义好的输入、输出格式来组织代码，不得改写输入、输出结构。

四. 作业要求:

1. 设计报告：

每小组提交一份设计报告，报告篇幅不得超过 4 页 A4 纸，报告应至少包含以下内容：

- 小组成员名单及分工情况：小组成员评分可能会因分工及完成情况产生差异。
- 提交文件清单。
- 工作开展及研究情况：应至少包含原理、实现方法、结果展示、结果分析、问题与不足，也可以包含其他任何对于解决问题有益的思考和讨论。

2. 提交清单：

每小组提交一份以“提交同学学号_提交同学姓名.zip/rar”命名的压缩文件，压缩文件内至少包含：

- 设计报告（.pdf/docx/doc）。
- 环境说明文件（requirements.txt），包含主要依赖库的版本。
- 补充后的函数 test_task1、test_task2、test_task3。
- 其他代码文件和依赖库文件。

3. 编程语言：

本次作业要求使用 python3.6，请确保使用正确的版本以免测试失败。

五． 评分标准

1. 评价指标：

本次大作业的任务一和任务二使用说话人识别的准确率 ACC（识别正确的视频/音频占全部数据的比例）作为评价指标。参考得分计算方法为

$$\text{clip}(\text{ACC} \times 1.1) \times 15$$

其中 clip(x)将 x 裁剪到 0-1 范围内，即 min(max(x,0),1)

任务三使用 SISDR 作为音源分离的测试指标，根据分离结果以及合成测试视频的每个说话人的音频，对音频的 ground truth 和分解结果的偏差进行一定的建模，最终得到一定条件下的信噪比(越大越好)，单位为 dB，其计算公式为：

$$SISDR(s,\hat{s}) = 10log_{10}\frac{\| e_{target} \|^2}{\| e_{res} \|^2} = 10log_{10}\frac{\| \frac{\hat{s}^T s}{\| s \|} s \|^2}{\| \frac{\hat{s}^T s}{\| s \|} s - \hat{s} \|^2}$$

其中，s 为 ground truth 的音频，$\hat{s}$为输出估计结果。具体建模方法详见参考文献“SDR – HALF-BAKED OR WELL DONE?”[1]。

分离测试指标分别有盲分离指标（$\text{SISDR}_{\text{blind}}$）和定位分离指标（$\text{SISDR}_{\text{match}}$）两种。盲分离指标与顺序无关，即输出的”left” ,”middle”,”right”不需要严格和 ground truth 的左中右音频对应，可以是乱序的。取对应指标最大化的排序方式作为结果。定位分离指标与顺序有关，需要左中右输出音频与 ground truth 音频一一对应。最终评价时会综合这两种指标进行评价。同时，同一段视频中三个说话人的音量大小被用于对指标进行加权平均，得到总的加权平均 SISDR(详见代码 utils.calc_SISDR)。参考得分计算方法为：

$$[0.5 \times \text{clip}(\frac{\text{SISDR}_{\text{blind}} + 5}{10}) + 0.5 \times \text{clip}(\frac{\text{SISDR}_{\text{match}} + 5}{10})] \times 20$$

以上评价系数将根据同学们总体完成情况是否偏低来进行整体性评分调整。

2. 具体评分标准

本次大作业满分 100 分，占期末总评 40 分。一般情况下，组内成员得分相同。满分 100

分由报告和结果两部分组成，其中报告占 50 分，结果占 50 分。根据三中的任务描述，结果分由三部分组成：基于视觉信息的说话人识别（15 分），基于声音信息的说话人识别（15 分），双声道多音源视频的逐说话人音源分离（20 分）。此外，在期末总评满分 40 分的基础上，结果排名前三的小组，期末总评加 2 分（不超过 100 分）。

如有设计文件延期提交，设计报告、程序实现中存在抄袭行为等，将根据情节程度，扣除课程设计的部分或全部分数。

[1] Le Roux J, Wisdom S, Erdogan H, et al. SDR–half-baked or well done? [C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 626-630.