数字媒体(2)：多媒体

音频分析及处理作业

实验报告

软件 41 唐人杰 2014013425

软件 41 罗皓天 2014013435

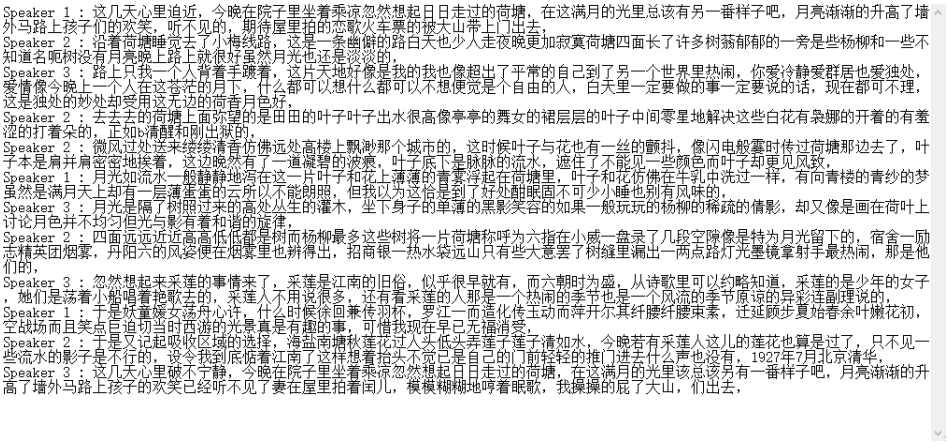
软件 42 卫国扬 2014013436

一、 实验成果概述及环境说明

在实现本次实验要求的过程中，我们小组最终实现了音频分段、话者估计、音频转文本等功能。能够简单地实现离线会话 wav 文件转换成文本的功能。

提交的代码中包含了一段测试音频和结果，在代码项目的 wav 目录下。同时目录下还保存有音频分段的结果。

下图即为话者的语段的标记结果与文本结果，准确率 75%：



实验进行的环境及依赖库列出如下：

- 操作系统：windows 10
- 编程语言：python3.5
- 依赖库：pyaudio、matplotlib、scipy、sklearn、numpy、pillow、tensorflow、pycurl、python_speech_features
- 需要良好网络环境

二、 使用说明

- **录制音频**

  在项目根目录下运行例如”python ./record.py 1000 output.wav”，看到提示信息后，单击回车即可开始录制时长为 1000s 的 wav 音频，并保存为”output.wav”，录制结束后会有新的提示信息。

- **话者识别**

  在项目根目录下运行例如”python ./analyzer.py test.wav output.txt”即可。会读取 test.wav 最终输出到 output.txt。在 TensorFlow 初始化完成后会输出标注好的话者识别 label。

三、 实验算法实现简要说明

- **音频录制实现**

  音频的录制主要使用了 pyaudio 来进行录制。我们采用 16000 个采样点（即 1s）的最小空白长度，17600 个采样点（即 1.1s）的卷积

核大小，2000 的最小音量进行分段，分段效果较好，在测试的《荷塘月色》文本中分段正确率 100%。

- **音频分段实现**

  音频的分段的主要思路为确定出现持续一段时间的低音量片段为分割点，进行分隔。在具体实现的过程中，我们发现直接对音高进行判断会受到环境音以及字词间隔的影响。我们借鉴了均值滤波的思想，对于每个采样点，我们计算这个点与时间序列前后采样点的均值作为这个点的新值，利用新值作为判断，更为有效。

- **话者估计实现**

  我们将话者估计分解成话者特征提取和点聚类两个问题。我们假定上一步音频分段后的每一段都来自于同一个话者，通过话者特征提取可以将可判别话者的特征提取出来，再通过聚类就可以看出有多少个话者。我们借鉴了《SPEAKER IDENTIFICATION AND CLUSTERING USING CONVOLUTIONAL NEURAL NETWORKS》这篇论文的思想。我们将语音段进行傅立叶变换操作，绘制时间和频率为坐标的灰度频谱图。然后利用 CNN 进行特征抽取。每一段音频会得到数个向量，对这些向量进行聚类，点最集中的质心就认为是这段音频的话者代表向量。之后根据话者向量间的距离就可以判断两段话是否为同一个话者

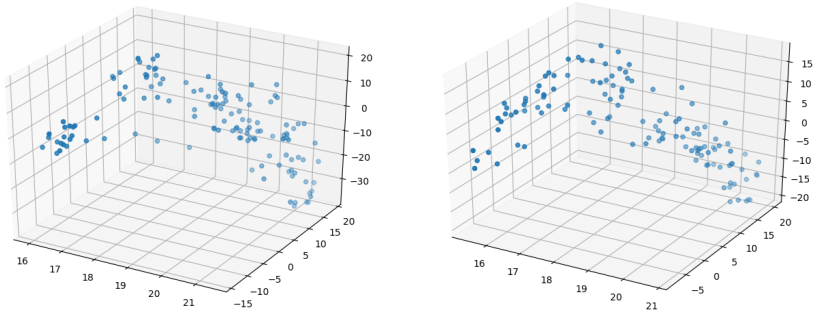
- **音频转文本实现**

  在决定音频转文本的方法上，我们使用百度语音识别 REST API 和 pycurl 模块。百度语音识别通过 REST API 的方式给开发者提供一个通用的 HTTP 接口，适用语种包括普通话、粤语、英文。支持 cm（不压缩）、wav（不压缩，pcm 编码）、amr（压缩格式）等格式，适用于 16K 的采样率与 16bit 位深的单声道。

四、 实验创新点

- **使用 CNN 辅助进行特征提取**

  在提取话者特征的过程中，我们尝试了直接采用 MFCC 再进行聚类，发现效果并不理想，特征向量较为分散不易聚类。下图为两个不同话者的单点图。

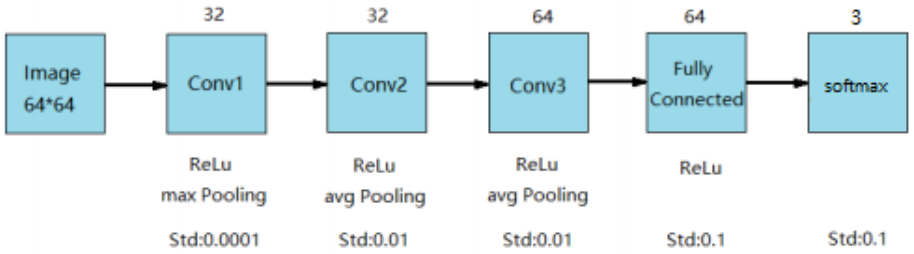


查找了相关资料后我们认为《SPEAKER IDENTIFICATION AND CLUSTERING USING CONVOLUTIONAL NEURAL NETWORKS》这篇论文的思想可以应用在我们的实验中。这篇论文的主要思想为提取音频段经过傅立叶变换得到频谱图，然后通过 CNN 以频谱图为输入进行训练，得到一层可以表示话者特征的输出再进行聚类。

由于论文原作者没有给出源代码及网络模型结构。我们采用了他

的思路，每隔一定的采样点就做一次傅立叶变换，然后绘制时间序列为横轴，频率为纵轴，振幅作为像素灰度的频谱图。具体实现时我们对音频每隔 16384 个采样点进行一次快速傅里叶变换（fft），频谱图片上点的灰度值表示该时刻 fft 结果中相应频率的振幅大小

随后借鉴经典 CNN 分类模型 CIFAR-10，最后得出 3 维的输出，认为这个输出可以代表一个人说话的特征。下图是训练集中的一张频谱图与网络结构示意图。



我们自己录制了累计 30 分钟的素材作为训练集训练模型，其中随机 80%作为训练集，另外 20%作为测试集。我们的 CNN 训练在 800 次训练后趋于收敛，测试集上正确率达到 90%。实验结果证明可以对话者区分有一定的准确率。

实验结果的调整与得出采用由组内 3 名同学朗读的《荷塘月色》文本共 12 个片段进行。

首次测试结果一般，正确率 60%，分析其原因可能为训练集文本内容过于单一，且与《荷塘月色》文本的朗读情况差异较大，故我们在训练集中加入了部分其他文本的朗读音频如《背影》的节选，再次进行训练，第二次训练结果的正确率达到了 75%

第一次识别结果：[1, 1, 2, 1, 1, 1, 1, 3, 1, 1, 2, 2]，正确率 60%

第二次识别结果：[1, 2, 3, 2, 2, 1, 3, 2, 3, 1, 2, 3]，正确率 75%

Ground Truth：[1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3]

- **静音段识别**

  在实现音频分段的过程中，我们在提高去噪效果上做了一部分创新。在具体实现的过程中，我们发现直接对音高进行判断会受到环境噪音以及字词间隔的影响。为了解决这一问题，我们借鉴了均值滤波的思想，使用一个长度为 11 个采样点的一维卷积核在音频的时间轴上滑动，对滑过的音量计算均值作为新采样点的音量，实验发现这种声音去噪方法较为有效。

五、 小组分工说明

在完成实验的前期主要以分工调研资料为主，初步分工为唐人杰同学负责调研实现音频分段部分、罗皓天同学负责调研音频转文本部分、卫国扬同学负责话者估计部分。在完成实验中后期，小组以集体讨论开发为主。小组各位同学工作量相当。

六、 总结与收获

通过此次音频分析及处理实验，我们对语音识别这个领域的相关知识有了更加深入的了解，对课上所学也有了更深入的认识。在寻找解决方案的过程中，我们查阅了一定的资料和相关的论文，对于目前这个领域的发展和遇到的难点都或多或少有了一点认识，收获颇丰。

最后感谢老师和助教的指导与帮助。