音频分析与处理作业文档

小组成员：向首兴 2014013421　楼昀恺 2014013407
师晓阳 2014013414

一、实验环境

本次作业实验环境是 linux 系统 ubuntu 64 位虚拟机，版本号 16.04.

二、功能

记录对话，在输出框显示结果，可以对音频转文字结果翻译错误的地方人为地进行修改，最后将最终结果导出到文件中。根据输入方式分为离线模式和实时模式：

离线模式：

直接导入一段音频，系统进行音频识别。

实时模式（实时会话记录）：

可以人为设置人数，该模式需要每个人事先录入一段话语，随着谈话的进行而实时的识别出话者，并将其谈话内容实时的转化为文本，谈话结束，将内容反应到输出框。

三、算法分析

主要实现的算法包括音频分段、话者人数估计、话者识别、音频转文本的相关算法，部分算法在离线模式和实时模式中实现方法有所

不同，下面对这些算法进行简单介绍。

1、音频分段

离线模式和实时模式使用的相同的音频分段算法相同，都是基于音频的响度或能量进行静音段的识别，从而对音频分段。

在离线模式中，对音频分帧，求每帧音频的能量，认为能量小于一定阈值的帧是静音帧，当连续出现 20 个静音帧时，认为该段静音帧前后的是不同的段落，即可能出现不同人说话。在软件中，20 个静音帧的时间大致等于 0.5s，而我们认为，在正常交流过程中，说话人交替用时在 0.5s 以上，因此使用 20 个连续静音帧作为分段的依据。

在实时模式中，我们首先尝试与离线模式相同的判断依据，即用音频帧的响度来判断静音帧，但实际效果不好，最后我们选择使用音频帧的能量来作为实时模式中静音帧识别的依据。参数选择方面与离线模式相同，同样用 20 个连续静音帧作为分段的依据。

要注意的是，在实时模式中阈值的选择是设备相关的，不同的设备采集到的声音的能量是有一定区别的，这是会影响分段效果的一个主要因素。

2、话者人数估计

只有离线模式有话者人数估计的相关算法。

在离线模式中，我们使用高斯混合模型进行话者人数的估计。估计是在分段结果上进行的，对分段结果的每段音频抽取其相关的特征，这里我们选择的是音频的去除低两维的 MFCCs 系数、MFCC 系数的一阶差分、信号过零率、音频能量、高频成分作为音频的特征，使用 sklearn 库中的 GaussianMixture 模型对这些特征进行聚类，使用 EM 方法确定参数。若用户制定聚类个数，则按用户制定个数进行聚类得到模型，若未指定个数，则遍历聚类个数 2-9，选择其中 BIC(Bayesian information criterion)值最小的作为模型。这里，聚类个数就是估计的话者人数。

3、话者识别

离线模式和实时模式有不同的话者识别算法。

离线模式的算法是紧接话者人数估计的。获得 GMM 模型后，将各段特征分别输入模型，获得每段所属的类，则若两段属于同一个类，则认为是同一个人说出的话，将紧邻的同类的段落合并，完成话者识别部分。

实时模式中，话者识别使用微软的说话人识别的 API，通过在录音前先对每个说话人录一段话并创建对应的 Profile。随后，程序采用多线程的方式，一个线程使用 pyaudio 库相关功能进行录音，每隔 5s，

将录音得到的音频数组加到数组列表 frames 的末尾，另一个线程判断 frames 是不是空，为空时 sleep2 秒后重新判断，当 frames 不为空时，取出 frames 的第一个录音结果，进行分段，然后对每段分别通过调用 API 进行话者识别。

4、语音识别

离线模式和实时模式使用的语音识别算法大致相同。

离线模式中，在话者识别后，任意相邻的两个音频段都是由不同人说的，因此具有语义的连贯性，将每段音频分别调用讯飞的语音识别库进行相关的识别，获得相应的文本结果。

实时模式中，由于当前录音结果的末尾可能和下一个录音结果的开头是连续的，因此，对分段后的每段进行话者识别后，判断与前一段是否是由同一个人说的，若不是，则对前一段调用语音识别库进行识别，否则，将此段连接到前一段的末尾，这样的操作同样是为了保证一个人说话内容的完整性，提高语音识别的准确度。

四、作业创新

1、使用多线程实现了实时模式的语音记录

2、实现了一个有 GUI 的功能较完善的软件

五、小组分工

师晓阳：文档撰写、PPT、参与语音识别

向首兴：GUI、参与语音识别

楼昀恺：参与语音识别、语音分段、话者识别的相关算法的实现