**Chapter 01 基于深度学习的说话人识别**

**实验目的**

1. 理解语音特征的原理和应用
2. 掌握常用声学软件包的基本使用方法
3. 了解说话人识别SincNet网络结构的原理和使用方法, 是一种比fbank的特征效果更好的语音特征方法.

**实验内容**

基于Timit数据集进行声纹识别。首先进行数据预处理，将语音特征抽取至数据文件中，然利用深度学习神经网络进行模型训练及测试。

**实验原理**

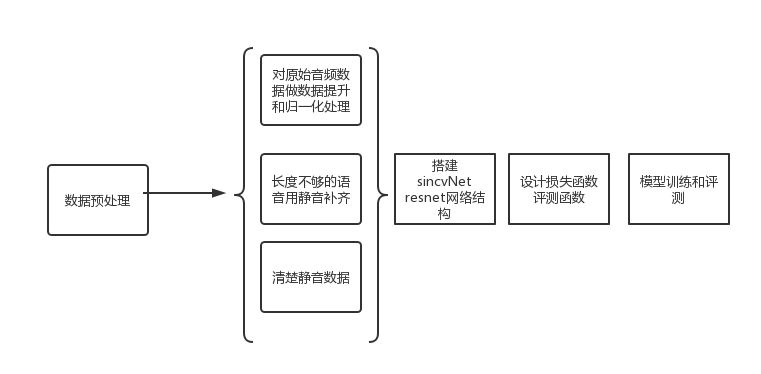
利用pysoundfile软件包进行特征处理。将特征送入由多层卷积SincvNet、池化和全连接层组成的深度学习网络中进行训练。每一个人当成一个训练标签，整个任务作为一个多分类问题。最终得到百分之99.8以上精度的模型。

**实验环境**

硬件：x86\_64 软件：Win10、Python 3.6.10、pysoundfile 0.10.3、torch 1.8等详细查看requirements

**实验步骤**

**整体流程图**

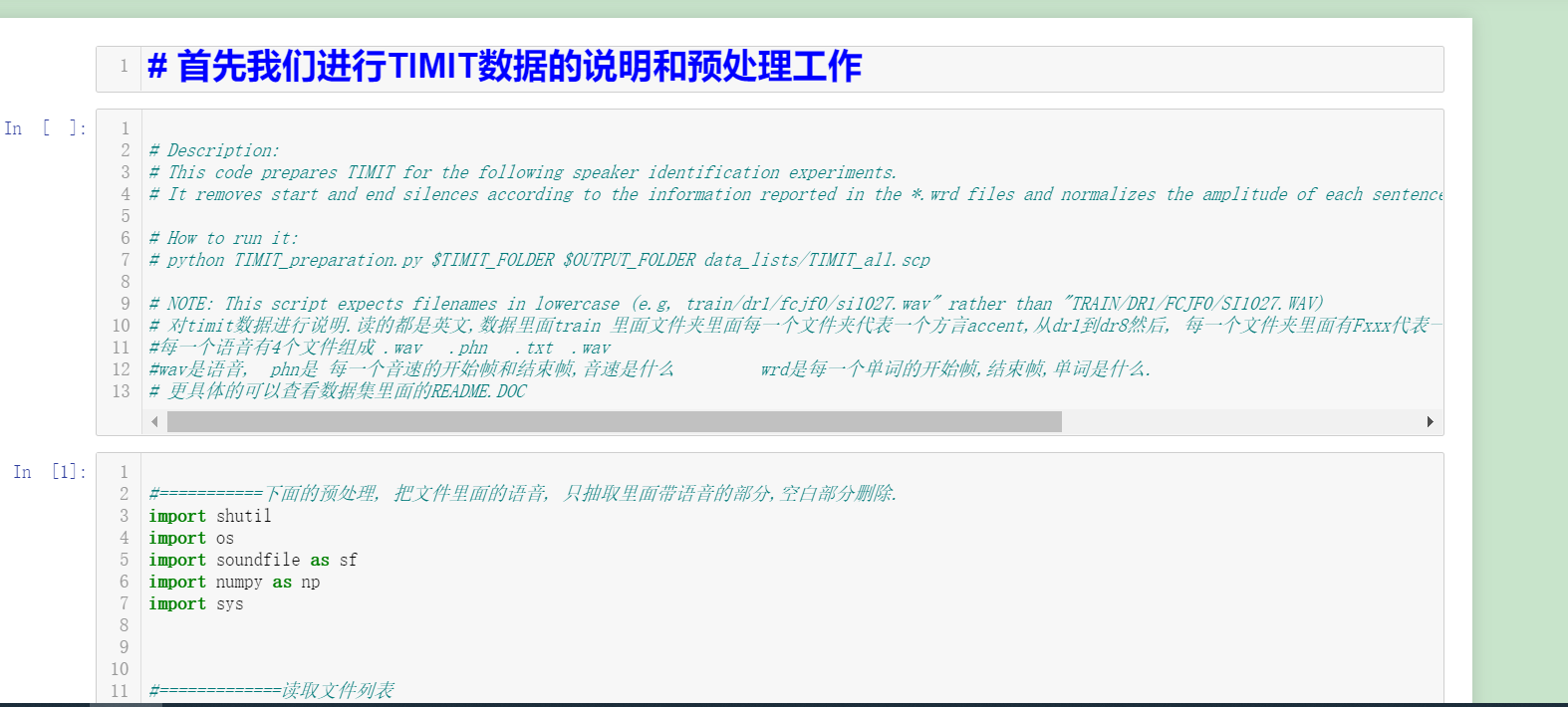
****

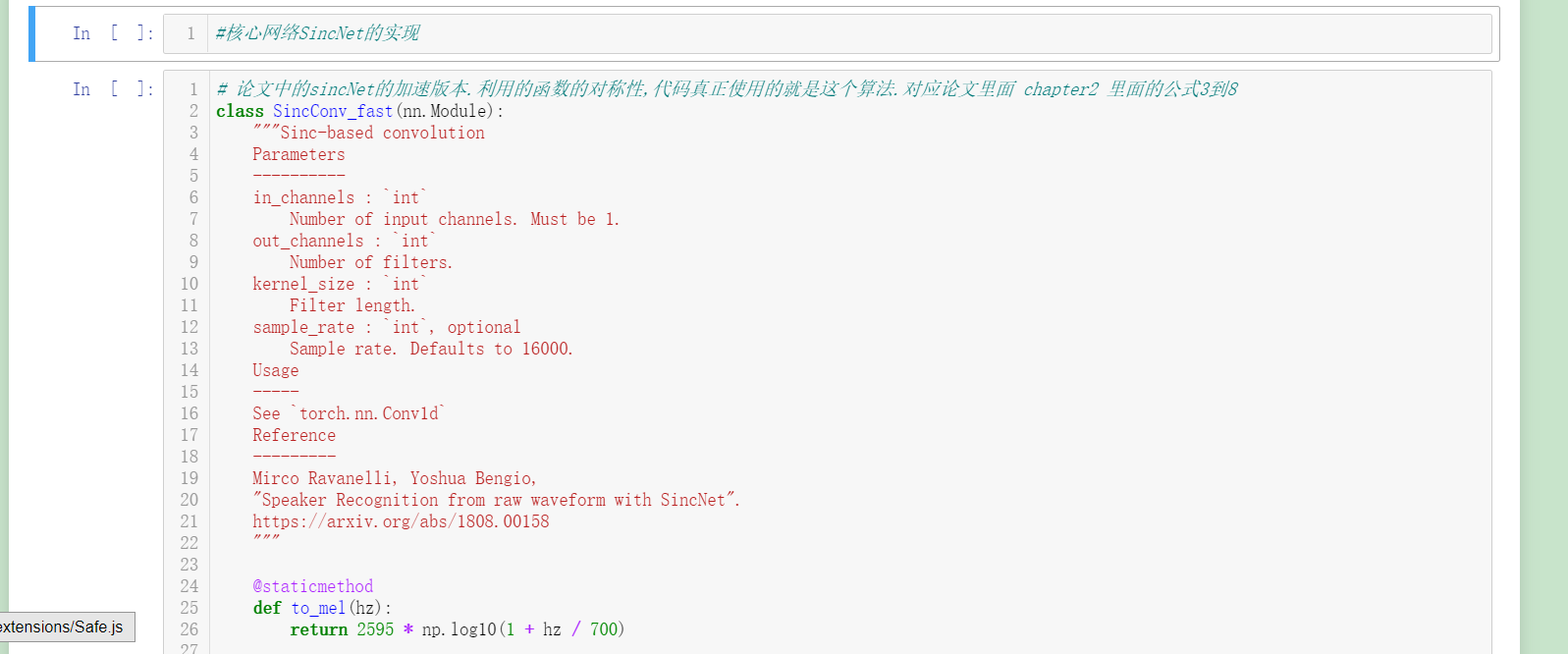
第一种方式 jupyter notebook

Cmd进入项目根目录,然后输入jupyter notebook

推荐按照顺序学习:

1. 数据的预处理工作.ipynb
2. 核心网络sincNet.ipynb
3. 训练和评测代码.ipynb







第二种方式

如果电脑中没有安装jupyter可以进入html版本教程目录

直接观看html版本的jupyter实验记录进行学习，注意浏览器打开文件较大会比较慢。大概30秒左右可以打开。

第三种方式

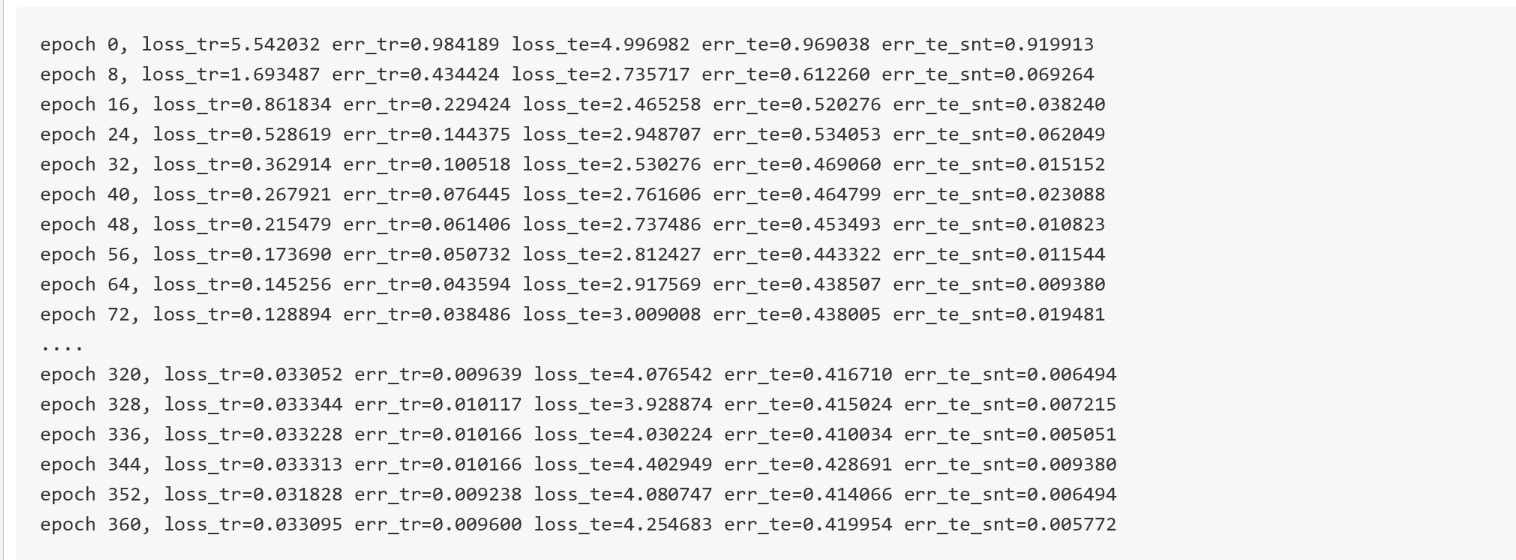
代码方式打开。推荐用pycharm的debug模式运行代码。因为函数调用结构比较复杂，debug的调用栈会让学习清晰很多。

TIMIT\_preparation.py 进行数据预处理

Speaker\_id.py 模型会进行训练和评测

Dnn\_models.py 神经网络的搭建

官方效果展示：



看到在训练集上收敛速度非常快.

官方的评价.

The converge is initially very fast (see the first 30 epochs). After that the performance improvement decreases and oscillations into the sentence error rate performance appear. Despite these oscillations an average improvement trend can be observed for the subsequent epochs. In this experiment, we stopped our training at epoch 360.  
The fields of the res.res file have the following meaning:  
- loss\_tr: is the average training loss (i.e., cross-entropy function) computed at every frame.  
- err\_tr: is the classification error (measured at frame level) of the training data. Note that we split the speech signals into chunks of 200ms with 10ms overlap. The error is averaged for all the chunks of the training dataset.  
- loss\_te is the average test loss (i.e., cross-entropy function) computed at every frame.  
- err\_te: is the classification error (measured at frame level) of the test data.  
- err\_te\_snt: is the classification error (measured at sentence level) of the test data. Note that we split the speech signal into chunks of 200ms with 10ms overlap. For each chunk, our SincNet performs a prediction over the set of speakers. To compute this classification error rate we averaged the predictions and, for each sentence, we voted for the speaker with the highest average probability.

主要关注一下最后一个指标.我们的数据是把每一个测试用例都切成200ms.然后算出每一片他的分类,然后做最高概率分类即可.最终是0.005的错误率.

**实验细节讲解**

核心是sincvNet网络的搭建. 需要理解的点是第一传统卷积网络的计算方式跟自相关的计算之间差一个逆序问题. 第二点是利用了函数的周期性和对称性实现了加速运算这也是最后函数名字里面fast的由来.