## 信号处理

#### 线性时不变系统：

a\*x(n+b)

a\*y(n+b)

#### **卷积：**



h(n)为冲激响应，X(n)为脉冲信号。卷积系统是一个经典线性时不变系统。

#### **Z变换：**



最重要的性质是能将卷积系统变为相乘系统。

#### **傅里叶变换，频幅，相位（Z变换实例）：**

离散周期傅里叶变换：



离散周期傅里叶逆变换：



时域信号转成频域信号，做的是傅里叶变换。频幅为|X(k)|，其偶对称：





相位表示在输入信号的响应位移：



x(n)为复指数输入信号,y(n)为输出信号。H(w)即为傅里叶变换，|H(w)|为输入信号的频幅变化，为相位变化。

#### **差分方程（滤波器，z变换零极点）：**

差分方程时一个线性时不变系统



做Z变换：



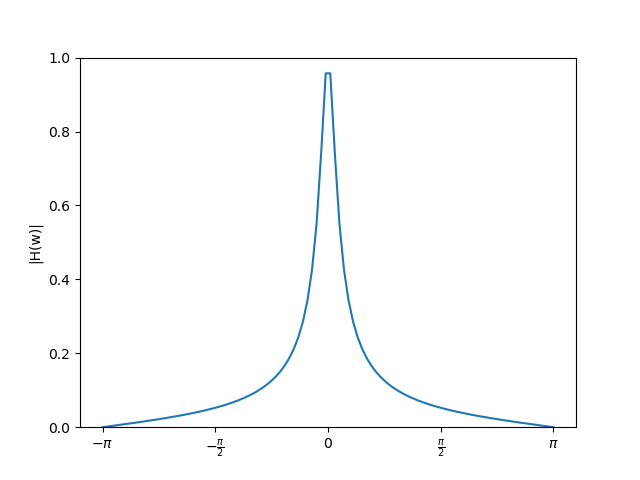
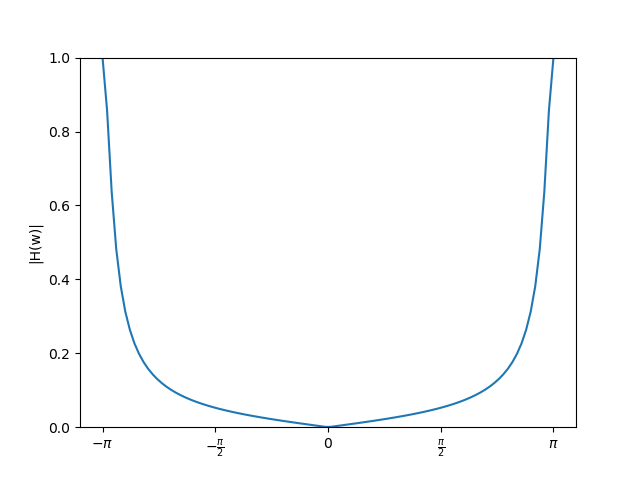
z1...zM是零点，p1...pN是极点。零点是被滤除的频段。极值点对应保留最完整的频段。

一般滤波器，如matlab中，y = filter(b,a,x) 指定参数b, a即为差分方程系数。

画出H(Z)的频幅图可以直观的观察零极点，以及滤波特性，例如：

低通滤波 高通滤波

## MFCC

MFCC(Mel Frequency Cepstral Coefficents), 可以简单理解成，不同音素的声道参数模型(误，个人理解)。具体参照完善资料，这里只对过程进行阐述。

#### **高通滤波：**

预加重的目的是提升高频部分，使信号的频谱变得平坦，保持在低频到高频的整个频带中，能用同样的信噪比求频谱。同时，也是为了消除发生过程中声带和嘴唇的效应，来补偿语音信号受到发音系统所抑制的高频部分，也为了突出高频的共振峰。



#### **分帧：**

一般每帧取25ms，交叉10ms左右。例如，采样为44100Hz，可以取1024个样本，交叉取512，每秒有87帧。

#### **加hamming窗：**

对每帧加hamming窗，目的是为了是小段帧信号具有周期型。窗函数为：



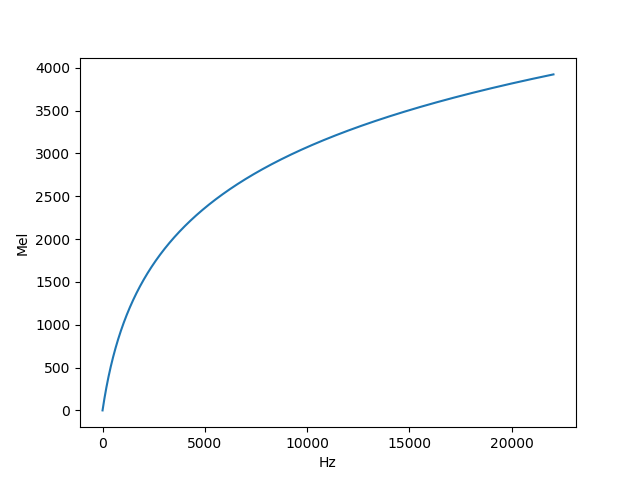
#### **傅里叶变换:**

把时域信号转为频域能量谱(频谱的模的平方)，用于后续分析语音在频域上的特性。

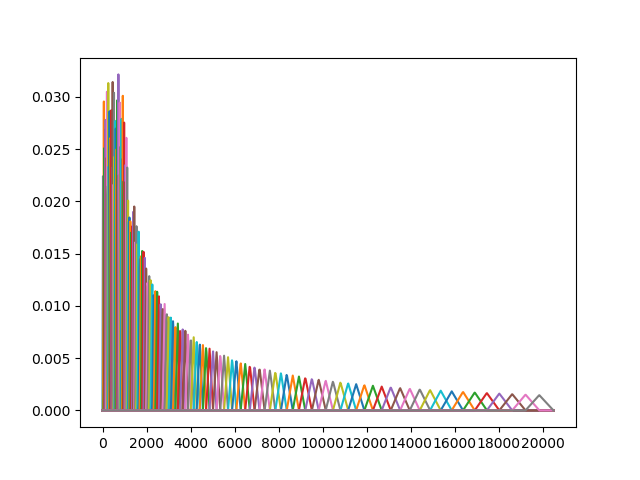
#### **梅尔能量谱：**

[梅尔刻度](https://zh.wikipedia.org/wiki/%E6%A2%85%E5%B0%94%E5%88%BB%E5%BA%A6" \t "https://blog.csdn.net/xmdxcsj/article/details/_blank)是一种基于人耳对等距的音高(pitch)变化的感官判断而定的非线性频率刻度。普通频率和梅尔频率关系如下：





以采样频率为44100Hz，采样样本为2048，三角带通滤波器为128个在梅尔刻度上均匀分布（上图纵坐标）。对应正常频率的三角带通滤波为（下图三角顶点对应上图横坐标）：



#### **倒谱分许：**

对梅尔能量谱取对数后做DCT变换得到倒频谱，。

## 哭声识别

#### **数据：**

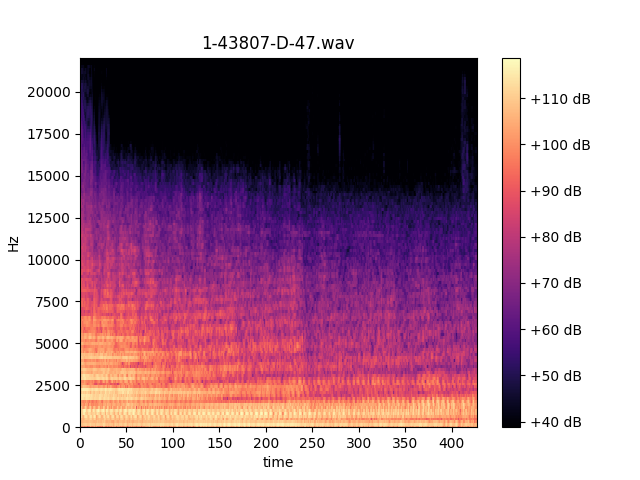
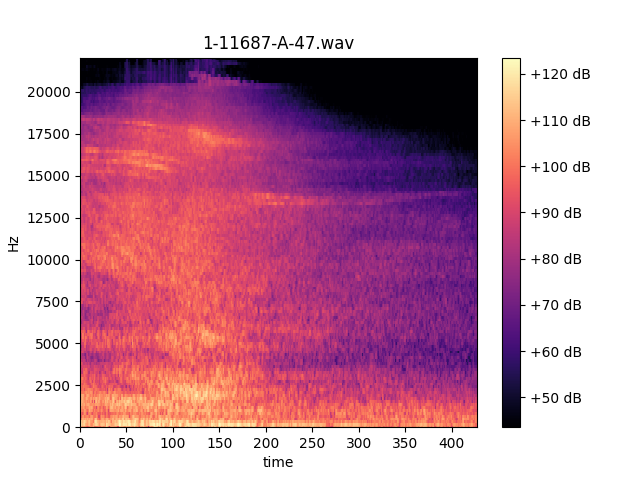
ESC-50 包含50个场景的2000个环境语音。大致分5类，动物，自然，人（非语言），家庭室内，室外。

语音采用WAV格式，每段语音长5秒，采样44100 Hz，mono格式1字节记录。

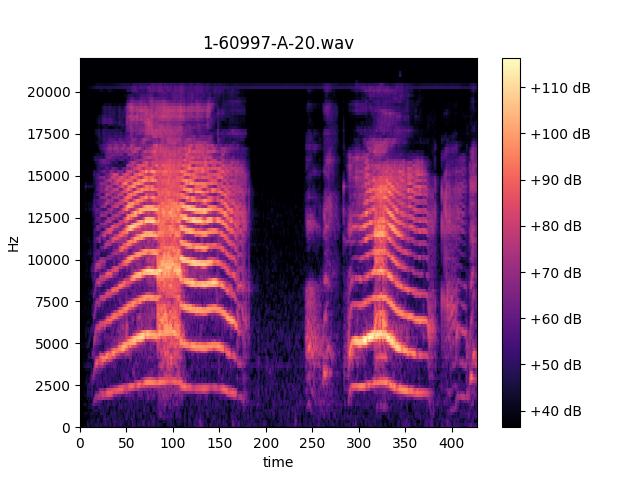
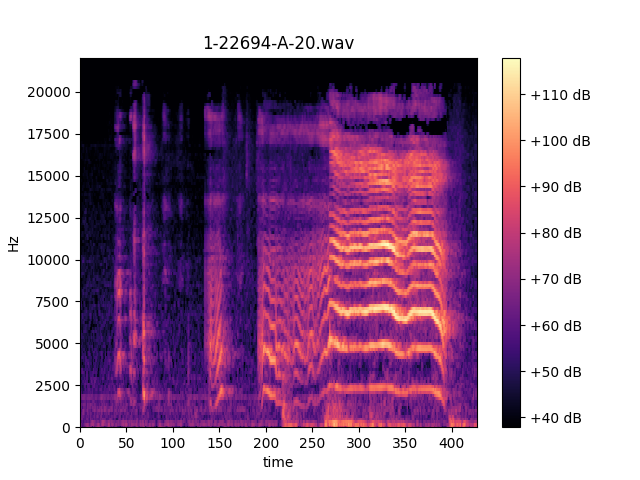
地址*https://github.com/karoldvl/ESC-50*

#### **输入：**

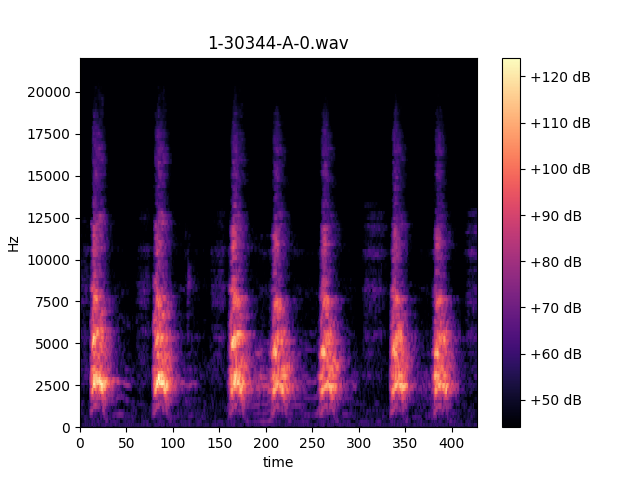
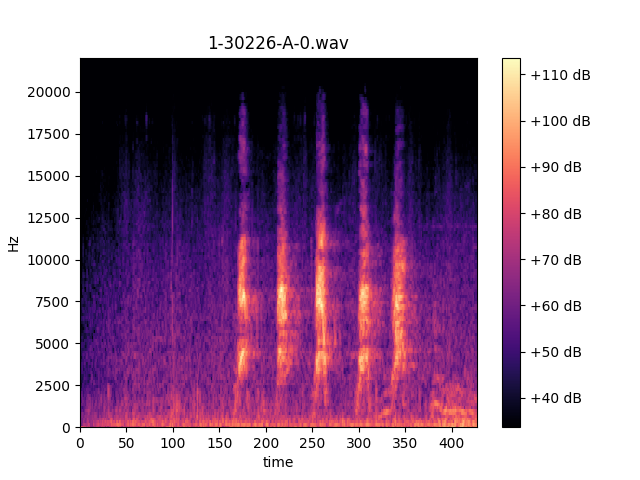
对每个语音生成梅尔能量谱，每帧2048个样本，步长512个样本。输入为128\*427\*1的图像信息。



飞机音效



哭声



狗叫

#### **模型：**

先将单色图片映射成为3色，增加一层卷积网络，卷积核数为3，核大小为1\*1。图像识别采用VGG16的CNN模型，其输入为128\*427\*3。Vgg16输出平铺后，经过映射到256维的relu层，再经过一层映射到50维的softmax层。

50

256

softmax

Relu

Vgg16

CNN

NN

128\*427\*3

128\*427\*1

**优化建议：**

1. 程序每一秒都会对过去5s的音频进行预测。可以尝试计算平均能量，来决定该段录音是否是安静的情况。
2. 目前模型参数有23,000,000个参数，模型大小100Mb，占用内存1G。可以尝试在以 一层CNN就降低输出维度，以缩减模型大小，加快效率。

## 声纹识别

声纹识别根据论文 Speaker Verification Using Adapted Gaussian Mixture Models 来实现。

地址 *<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.338&rep=rep1&type=pdf>*

#### **数据:**

TIMIT数据集，一共包含6300个句子，由来自美国八个主要方言地区的630个人每人说出给定的10个句子。目录结构如下：

**/<语料库>/<用处>/<方言种类>/<性别><说话者ID>/<句子ID>.<文件类型>**

TIMIT 的录音是NIST格式（无法用scipy.wavfile解析，参照代码timit\_reader），采样率16000Hz，每个记录占2个字节。其头信息结构如下：

**NIST\_1A**

**1024**

**database\_id -s5 TIMIT**

**database\_version -s3 1.0**

**utterance\_id -s8 cjf0\_sa1**

**channel\_count -i 1**

**sample\_count -i 46797**

**sample\_rate -i 16000**

**sample\_min -i -2191**

**sample\_max -i 2790**

**sample\_n\_bytes -i 2**

**sample\_byte\_format -s2 01**

**sample\_sig\_bits -i 16**

**end\_head**

#### **UBM-GMM**:

每段录音的MFCC特征作为训练输入（每一帧都有一个MFCC特征，可以近似看作音素特征）。

UBM统一背景模型，其本质是一个GMM混合高斯模型。按照论文阐述，设置有1024个聚类选用DR1~DR4所有语音人来训练，大约有4小时语音。

GMM对应说话人模型。是从UBM进行adaption获得。判断声纹的归属判定为



#### **GMM:**



Expectation:



Maximization:



取超参数的导数为0,得到：



对应代码（此为简便表达，需调试，正确见源码）：

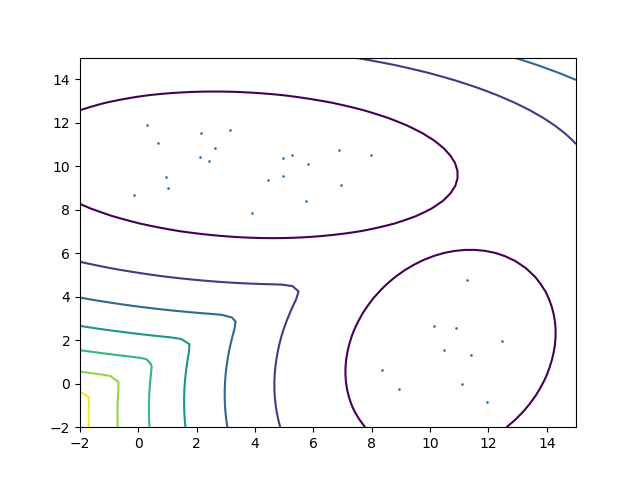
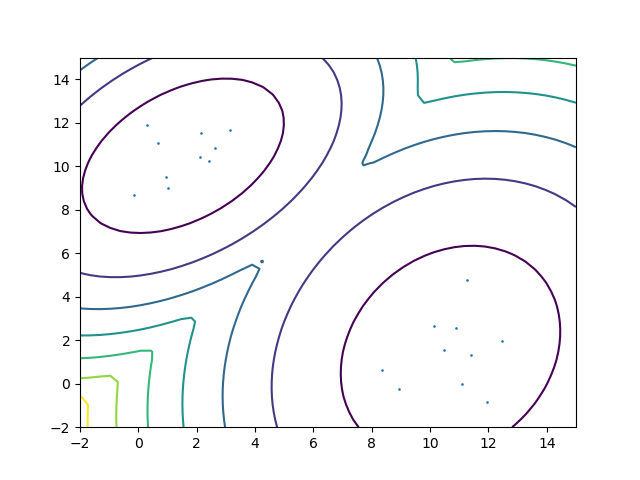
n = np.sum(pp, axis=0) #pp是Q(z|x)

w = n / m  
e1 = np.dot(pp.T, X) / n[:,np.newaxis]  
e2 = np.array([np.dot(X.T\*p, X) for p in pp.T]) / n[:,np.newaxis, np.newaxis]-

[np.dot(e[:,np.newaxis], e[np.newaxis,:].T) for e in e1]

#### **GMM Adaption:**

这部分构建说话人对应gmm。 Adaption是将原有gmm根据新来的特征进行适应的过程。如图所示：



Adaption 对gmm参数更新规则为:



按照论文，预设。详细证明请参照论文 Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains 章节1，2。

地址*http://speech.ee.ntu.edu.tw/previous\_version/Channel-MAP(SAP-1994).pdf*

#### **有待完成：**

1. 目前ubm训练了一天左右，adaption正确性有待验证。
2. 对于判定阈值的决定。
3. 真实录音下的检验。

## 代码简介

本文中的图都可以在代码中找到，对应为：

- auido/test\_mfcc 信号处理，mfcc图片

- ubm/test 声纹识别图片

- model/main 哭声识别图片

代码主要依赖于numpy，scipy，tensorflow，keras。不涉及其他语音处理库如librosa。逻辑基本可读。结构展示如下:

- audio 信号处理库

- signals 信号处理工具代码

- display 信号作图代码

- wave wav音频读取写入程序，支持ieee\_float格式

- timit\_reader NIST格式音频读取代码

- bin 录音和播放

- recorder 录音

- player 播放

- audio\_processor 对帧进行处理

- main 主程序入口

- model 哭声识别模块

- main 输出所有音频的梅尔功率谱

- predict 简单测设

- train 训练

- ubm 声纹识别模块

- core gmm的em算法以及adaption

- ubm 训练ubm以及适配ubm

- test2 程序入口