

实验一 语音信号 MFCC 特征提取

语音信号处理在语音系统中扮演着重要的角色，无论是自动语音识别（ASR）还是说话者识别等领域都发挥着重要的作用。梅尔频率倒谱系数（MFCC）是目前使用最广泛的特征之一。MFCC 特征是通过在声音频率中，对非线性梅尔刻度的对数能量频谱，进行线性变换得到的。信号通过预加重滤波器；然后将其切成（重叠的）帧，并将窗函数应用于每个帧；之后，我们在每个帧上进行傅立叶变换（或更具体地说是短时傅立叶变换），并计算功率谱；然后计算滤波器组。为了获得 MFCC，可将离散余弦变换（DCT）应用于滤波器组，以保留多个所得系数，而其余系数则被丢弃。

实验目标：

理解并掌握 MFCC 特征提取的原理和过程，并对提供的音频提取 MFCC 特征。

实验流程：

0. 语音信号准备（setup）

从主页下载提供的两个音频文件，可以尝试先听其内容了解语音信号频谱特征，并用 python 库或者 MATLAB 包载入，确定其采样频率以备后续使用。

1. 预加重（Pre-Emphasis）

第一步是在信号上施加预加重滤波器，以放大高频。预加重滤波器在几种方面有用：（1）平衡频谱，因为高频通常比低频具有较小的幅度；（2）避免在傅立叶变换操作期间出现数值问题；（3）还可改善信号噪声比（SNR）
预加重滤波器可以使用一阶滤波器应用于信号 x ：

$$y(t) = x(t) - \alpha x(t - 1)$$

其中 α 一般设置为 0.97。

预加重在现代系统中的影响不大，主要是因为除避免了不应该成为问题的傅立叶变换数值问题外，大多数预加重滤波器的动机都可以使用均值归一化来实现。

2. 成帧（Framing）

经过预加重后，我们需要将信号分成短时帧。此步骤的基本原理是信号中的频率会随时间变化，因此在大多数情况下，对整个信号进行傅立叶变换是没有意义的，因为我们会随时间丢失信号的频率轮廓。为避免这种情况，我们可以安全地假设信号的频率在很短的时间内是固定的。因此，通过在此短时

帧上进行傅立叶变换，我们可以通过串联相邻帧来获得信号频率轮廓的良好近似值。

语音处理中的典型帧大小范围为 20 ms 至 40 ms，连续帧之间有 50% (+/- 10%) 重叠。实验中的设置是帧长为 25ms，帧移为 10ms（重叠 15ms）

3. 加窗 (Window)

将信号切成帧后，我们对每个帧应用诸如汉明窗之类的窗函数。汉明窗具有以下形式：

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

其中 $0 \leq n \leq N - 1$ ， N 是窗长。

4. 傅立叶变换和功率谱 (Fourier-Transform and Power Spectrum)

对窗长 N 的每个帧进行补零填充，以形成一个扩展帧，该帧包含 256 个样本（8 和 11 kHz 采样率）和 512 个样本（16 kHz）。分别使用长度为 256 或 512 的 FFT 来计算信号的幅度谱。

信号 x 的第 i 帧 x_i 的 FFT 可以用以下公式计算：

$$\text{bin}_k = \sum_{n=0}^{NFFT-1} x_i(n) e^{-jn k \frac{2\pi}{NFFT}}, \quad k = 0, \dots, NFFT - 1$$

注意，由于对称性，只有 $\text{bin}_{0 \dots NFFT/2}$ 用于下一步处理

功率谱：

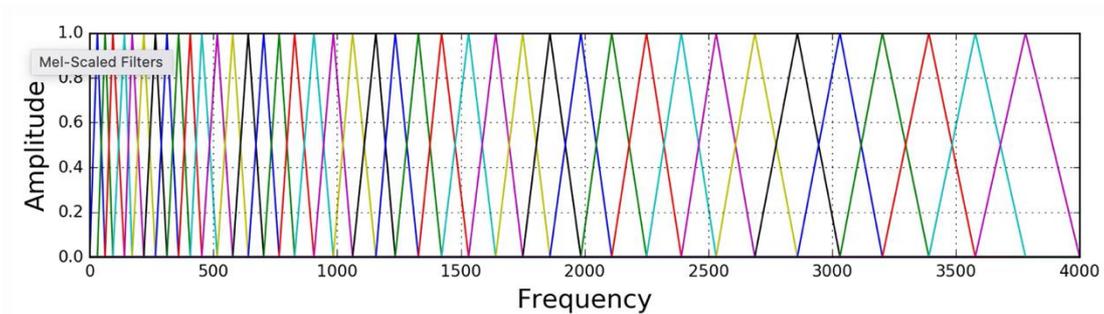
$$P = \frac{|FFT(x_i)|^2}{NFFT}$$

5. 滤波器组 (Filter Banks)

计算滤波器组是将三角滤波器在 Mel 刻度上应用于功率谱以提取频带，通常设置成 40 个滤波器。梅尔音阶的目的是模仿低频的人耳对声音的感知，方法是在较低频率下更具判别力，而在较高频率下则具有较少判别力。

Hertz(f)和Mel(m)之间的转换关系为：

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$
$$f = 700(10^{\frac{m}{2595}} - 1)$$



滤波器组中的每个滤波器都是三角形的，在中心频率处的响应为 1，并朝着 0 线性减小，直到达到响应为 0 的两个相邻滤波器的中心频率为止，可以通过以下方程式建模：

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leq k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) < k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

关键在于确定中心频率，首先根据信号的原始采样频率确定傅立叶变换对应的最高频率 (8k 采样率对于 4000)，然后将其转换到 Mel 刻度，根据滤波器数得到对应数量的 Mel 刻度中心频率 (Mel 刻度下中心频率呈线性分布)，最后转换到频域，并对应上数字信号采样点。

最终 Mel 滤波的结果为：

$$fbank_m = H_m P$$

非线性变换：

$$f_m = 10 * \log_{10}(fbank_m)$$

6. 倒谱系数 (Mel-frequency Cepstral Coefficients) (MFCCs)

事实证明，在上一步中计算出的滤波器组系数是高度相关的，这在某些机器学习算法中可能会出现。因此，我们可以应用离散余弦变换 (DCT) 对滤波器组系数进行解相关，并生成滤波器组的压缩表示形式。这里选取前 13 维作为最终的倒谱系数。

$$C_i = \sum_{j=0}^{39} f_j \cdot \cos\left(\frac{\pi \cdot i}{40} (j - 0.5)\right), \quad 0 \leq i \leq 12$$

实验要求:

1. 本次实验不限制编程语言, 可以使用 python, MATLAB 等语言。
2. 对 FFT 和 DCT 不做要求, 可以调用工具包来实现。
3. 按照之前提供的参数设置对课程网站上提供的两个音频提取 MFCC 特征。
4. 请将以下文件打包发送至 algorithm_2021@126.com

文件名如: 张三_PB14007000.zip

邮件主题:实验一_张三_PB14007000

- a) 实验报告: 至少包含用 1.wav 分析实验各个阶段结果 (傅立叶频谱图, FBANK 和 MFCC 特征); 代码分析; 2.wav 的运行耗时 (最好有截图); 并探究在不同的机器学习算法中如何选择 FBANK 和 MFCC 特征。
- b) 1.wav 的特征结果文件: 存储格式为 $N_frame * Feature_dim$, 可以是 .mat, .npy, .txt。