语音识别Homework 1

王蔚达 2151300

本报告阐述了MFCC（梅尔频率倒谱系数）声学特征从语音段中的提取过程。我们将依次探讨以下步骤：预加重、窗口化、短时傅立叶变换、梅尔滤波器组、取对数、离散余弦变换、动态特征提取和特征转换，并附带相关图形输出，如频谱图和MFCC。

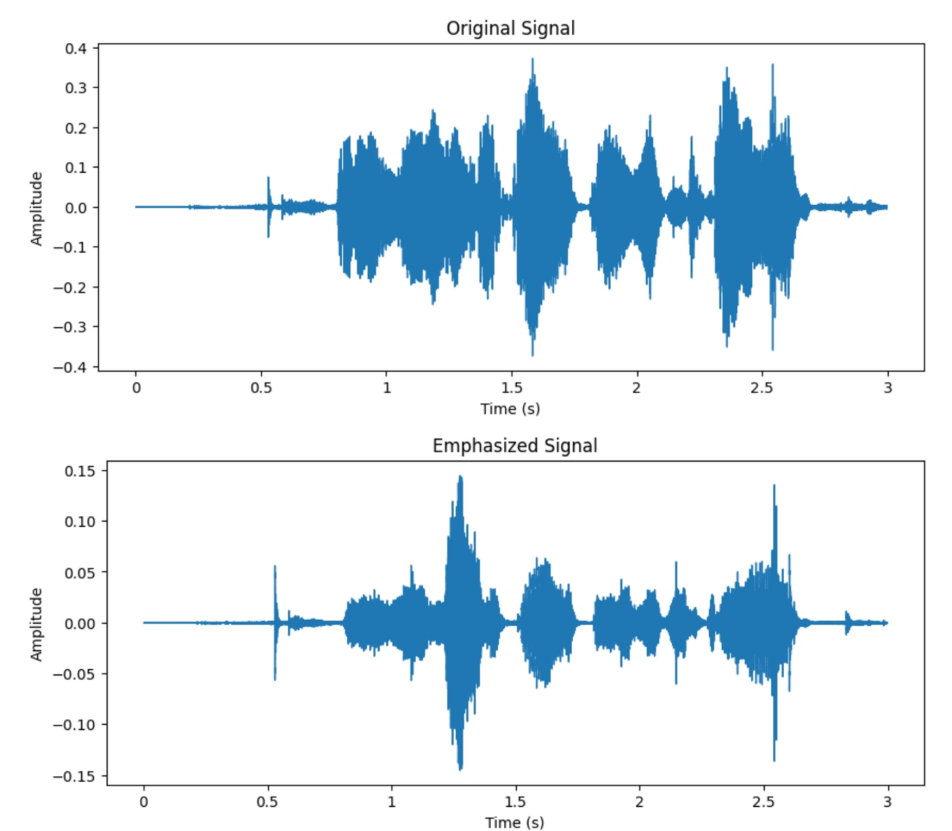
1. **声音的录制**

本实验提供了一段代码在record.py文件中，通过调用pyaudio.PyAudio()函数以进行声音采集，所采集的声音音频可以用于接下来的实验分析。

1. **预加重**

预加重是一个简单的高通滤波器，用于增加语音信号的高频能量，从而使语音信号的高频部分与低频部分具有相似的幅度。这是为了弥补语音信号中的高频损失，并提高MFCC的性能。

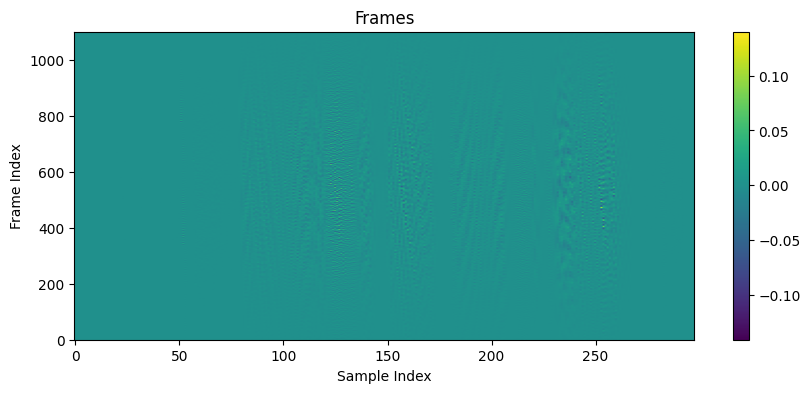
下图即展示了预加重前后声音信号的变化。



1. **窗口化**

在窗口化阶段，连续的语音信号被划分为短的帧，每一帧都乘以一个窗口函数（如汉明窗）。这是为了确保帧的开始和结束处的连续性，并减少DFT（离散傅里叶变换）引入的频谱泄漏。

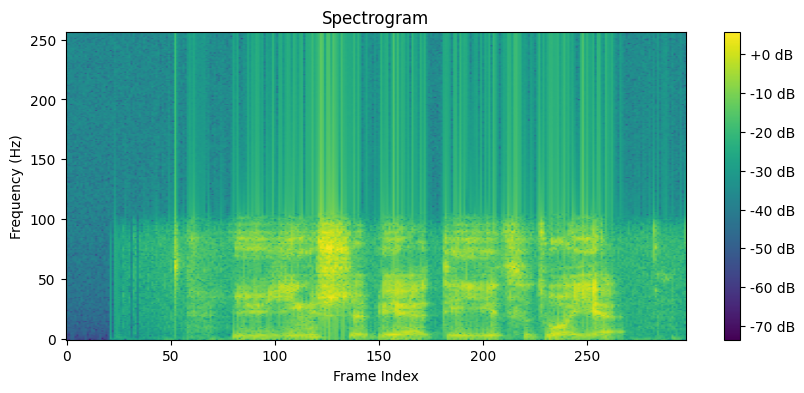
原先的音频是[1,132096]，实验中，我选取的帧长为25ms，步长为10ms，声音的频率为44.1 kHz（CD质量音频的标准采样率），将不及一帧的部分用0补全，最后得到[298,1102]的帧信息（共298帧，1102个采样点），做出下图。



1. **短时傅立叶变换 (STFT)**

短时傅立叶变换（STFT）是一个从时域到频域的转换工具，用于分析音频信号的频率组成。通过将连续的信号分割成短的帧并分别对它们应用傅立叶变换，我们可以得到每个帧的频谱。这个频谱揭示了在特定时间段内哪些频率是活跃的。对于语音，这意味着我们能够追踪其随时间变化的频率内容，使STFT成为一个极其有价值的工具。

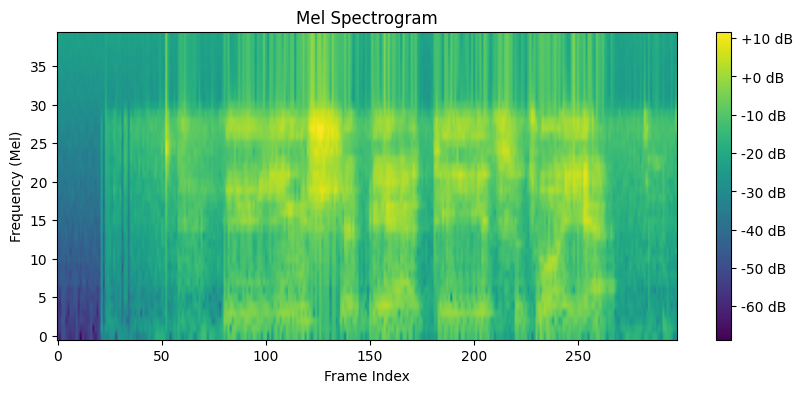
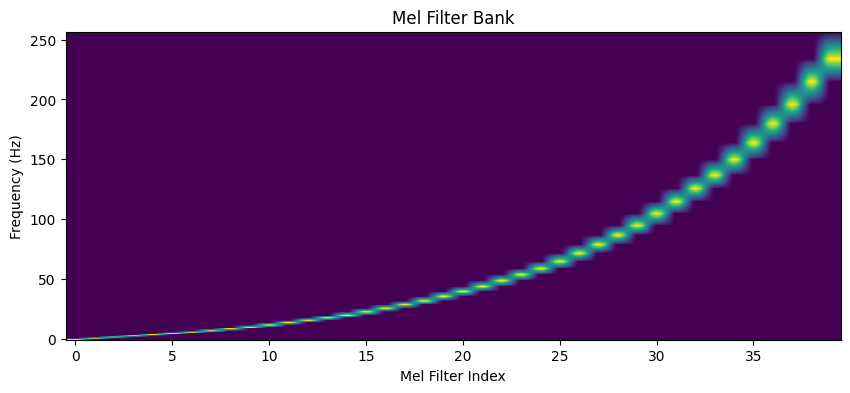
下图为实验过程中得到的频谱图。



1. **梅尔滤波器组**

梅尔滤波器组用于模拟人类听觉系统在频率识别上的非线性特性，特别是在较低频率下的更高的分辨率。这组滤波器在梅尔频率尺度上均匀分布，与线性频率尺度相比，梅尔尺度更接近于人类对频率的主观感知。每个滤波器在其中心频率有最大的响应，响应随着与中心频率的距离增加而减少。这模拟了我们的听觉系统对接近的频率更为敏感，而对远离的频率则不那么敏感的特性。通过这种方式，梅尔滤波器组提供了一个更接近人类听觉感知的音频信号的频率表示。

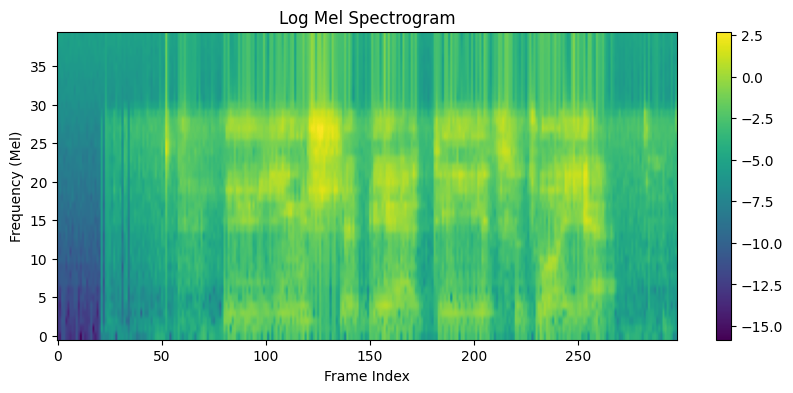
下图为梅尔滤波器组的图像与使用梅尔滤波器组过滤后的图像。



1. **取对数**

取对数操作确实是为了模拟人耳的感知特性。我们的听觉系统在对响度或振幅的感知上是对数的，意味着当声音的实际强度翻倍时，我们并不会感到响度增加了一倍。因此，通过应用对数操作，我们可以更接近地模拟人的听觉感知。对梅尔滤波器组的输出进行对数操作后，得到的结果称为log Mel spectrum，它为MFCC（梅尔频率倒谱系数）的提取提供了基础。这些MFCC系数经常用于语音和音频处理任务，因为它们为音频信号提供了一个与人类听觉更为相似的表示。

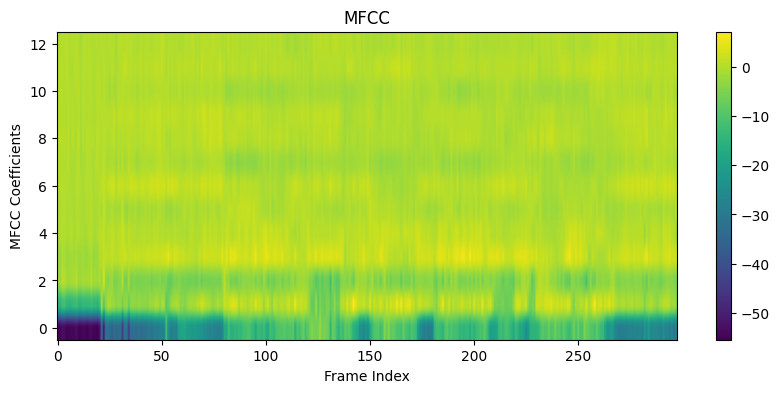
以下为取对数后得到的结果。



1. **离散余弦变换 (DCT)**

在语音信号处理中，梅尔频率倒谱系数（MFCC）已被广泛采纳为主导特征。对梅尔滤波器组的输出进行离散余弦变换（DCT）产生了这些系数。DCT旨在最小化特征之间的相关性，并有效地压缩了语音信息。实践中，仅选取DCT的前几个系数，因为它们包含了语音信号的主要能量，而高阶系数往往与微细变化和噪声相关联。这种方法不仅简化了表示，还确保了特征的鲁棒性。

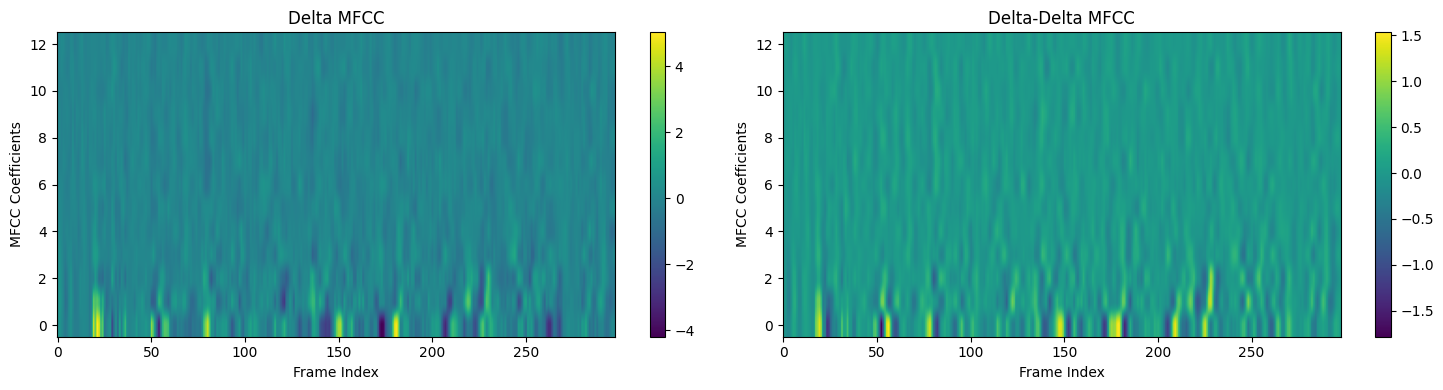
下图为生成的MFCC可视化展示结果。



1. **动态特征提取**

动态特征提取是音频信号处理中常用的一种方法，主要是为了捕获音频信号的时间动态特性。基于时间的一阶和二阶导数是提取这些动态特性的常用手段，这些导数也称为 delta（一阶导数）和 delta-delta（二阶导数）特征。

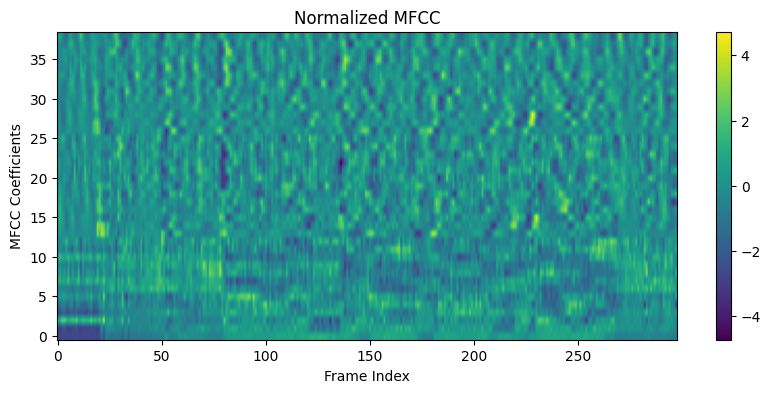
下图即为两个动态特征的可视化展示。



1. **特征转换**

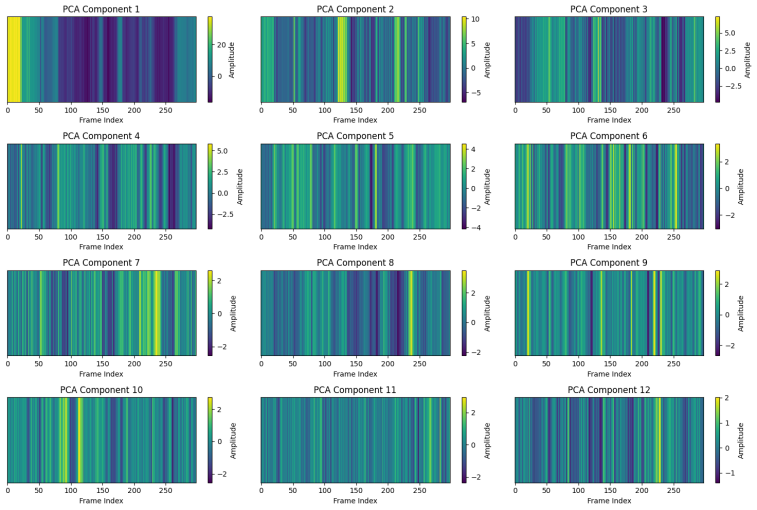
特征归一化：归一化的基本目标是转换特征，以减少训练和测试之间的不匹配。这意味着我们希望我们的模型在不同的数据集上都能够表现得很好，而不仅仅是它在训练过程中看到的数据。CMN/CVN：通常，我们可能希望在相同的说话者/通道上计算并应用均值和方差统计数据，以确保同一说话者或同一通道的数据具有相似的分布。

使用CMN/CVN进行归一化得到的结果可视化如下。



主成分分析：特征转换是用于将MFCC特征投影到一个新的、可能更低维的空间。这通常通过主成分分析（PCA）或线性判别分析（LDA）来实现。

下图中将13x3个归一化后的特征合并，做PCA，得到12维的特征可视化。



**总结**

通过上述步骤，我成功地从语音段中提取了MFCC声学特征。这些特征为语音提供了一种紧凑的表示，广泛应用于语音识别、说话者识别和其他语音处理任务。在此实验中，我们还展示了如何使用现有的工具和函数（FFT、DCT等）来辅助MFCC的计算，同时生成的可视化结果，让我更加直观地理解了整个过程。