5) 对非静音部分分类

分层算法中的第二层处理，即对非静音部分分类，而在这一部分中又将分为两层进行分类。

第一层将非静音部分分为含音乐部分（With Music）和不含音乐部分（Without Music）这两大类，使用韵律特征。

**第二层将第一层分类得到的含音乐部分进一步分为Music、Speech with Music、Advertisement和Song四小类，使用频域特征，若四类均不属于则划分到Unknown类。**不含音乐部分则划分为Speech类。

由于韵律特征并非对所有音频都适用，即并非对所有音频数据都能提取韵律特征，因此会有一些非静音部分没有经过第一层分类，需要在第二层的时候直接对其进行处理，同样，若五类均不属于则划分到Unknown类。

6) 分割

音频的语义内容具有时间连续性，基于这一特性可以设计分割准则，对非静音部分的Clip序列进行平滑，对前面的分类结果进行修正，然后将具有相同类别的Clip合并成音频事件片段。

本文采用的分割合并准则有如下四个：

准则1：设C1、C2和C3为三个相邻的Clip，如果它们的类别各不相同，则认为C2的类别应与C1相同[3]。―――与前面一致

准则2：设C1、C2和C3为三个相邻的Clip，如果C1和C3属于同一类别，而C2与C1、C3的类别不同，则认为C2的类别判断有误，C2应与C1、C3的类别相同[3]。

准则3：设C1、C2和C3为三个相邻的Clip，如果C1为一个非静音部分的第一个Clip，C2和C3属于同一类别，而C1与C2、C3的类别不同，则认为C1的类别判断有误，C1应与C2、C3的类别相同。

准则4：设C1、C2和C3为三个相邻的Clip，如果C3为一个非静音部分的最后一个Clip，C1和C2属于同一类别，而C3与C1、C2的类别不同，则认为C3的类别判断有误，C3应与C1、C2的类别相同。

经过分割后得到的结果即若干个分属于不同类别的音频事件片段。

使用Praat软件提取基频和共振峰时，对每一个输入音频文件根据设定好的窗长(frame)，提取所有的基频和共振峰。

然后对于每个包含75帧的Clip（时长约1.3s）计算统计值，包括平均值（Mean）、中位值（Median）、标准差（Standard Deviation）、最小值（Minimum）、最大值（Maximum）和变化范围（Range），相邻两个Clip有20帧的重叠（时长约0.5s）。

输出文件中每个Clip对应一个统计值的特征向量，其中基频的特征向量为6维，共振峰的特征向量为24维（四个共振峰，各6维），共30维（5\*6）。

由于并非所有音频都可以提取到基频这一特征，在第一层分类结束后仍会有部分非静音部分并未分到With Music类或Without Music类。因此在第二层分类时，还需对这些第一层分类遗留下来的非静音部分进行分类，使用分类器直接将其分成五类，即本文所涉及到的全部五种音频事件类别。

为和多层分类结果进行对比，还将进行单层分类的实验。单层分类实际上是减少了第一层分类，即只使用频域特征，在将静音部分切分之后，对所有非静音部分使用分类器，将其分成五类。

使用GMM的时候，混合数的选取比较重要。本文选取不同混合数，经过多次测试，最后选择了整体性能较好的一个值，。

2) 在基本VAD算法结束后，对静音段端点附近使用分类器，可以进一步提高静音和非静音的切分准确度。本文在这里使用Silence和Voice两个GMM。