音频分类与音频分段的研究

蒋丹宁 蔡莲红

清华大学计算机系人机交互与媒体集成研究所

摘要：随着计算机速度、容量的提高，以及互联网上音频数据的急剧膨胀，发展音频信息的基于内容检索技术已经成为迫切的需要。音频信息的基于内容检索系统包括音频信息数据库与音频信息的查询、浏览系统两个部分。在建立音频信息数据库时，首先要对原始的音频数据进行处理，将它们分类，再用一定的方法建立索引；在检索的时候，也首先要确定需要检索的音频的类别。可见，音频信息的分类是建立基于内容检索系统的基础。对于较短的、只包含某一类音频信息的音频文件来说，只需直接进行分类即可；而对于较长的、包括若干不同类别的音频段落的音频文件来说，则还要将它们按照类别的不同进行分段。本文对音频的分类 和分段问题做了初步的研究，并建立了一个新闻广播的音频分类与分段系统。

1、引言

随着计算机速度和存储容量的增加以及计算机网络的飞速发展，音频数据越来越多地出现在计算机和互联网上。然而，目前对于音频数据的存储还仅仅是以字节流的方式，再加上一些基本的属性，如名字、采样率、每个采样点所占的比特数、长度等等来进行的，这种方式对于音频文件的内容来说是不透明的。对于需要访问音频数据库的管理者和使用者来说，检索音频数据中的有用信息是一件困难的事情。传统的基于关键字的检索方法往往满足不了实际的需要，因为模糊的、具有很强主观色彩的音频数据不能够由关键字来完美地代表。为了能够快捷、准确、方便地在音频数据中查找到有用的信息，就需要音频的基于内容的检索技术。

为了实现音频信息的基于内容检索，我们首先要对音频信息进行分类。对于那些比较短的音频文件来说，由于在文件中音频的类别始终保持一致，可以利用音频文件整体上的信息来提取特征，进行分类。但是，对于很多长的音频文件，它们包含很多不同类别的段落，这时问题就不那么简单了。我们必须将原始的音频文件划分为若干段落，同时确定每一段的类别。

分类与分段是相辅相成的。由于不能利用整体信息，我们必须从局部分析出发，来找到音频类别发生变化的地方。分类可以在分段之前，这时首先要进行音频的短时分析，确定每一帧的类别，得到一个类别序列，并在此基础上进行分段；分类也可以在分段之后，这时要首先找到音频信息的特征发生急剧变化的地方，并把它们作为段落的边界，然后在此基础上判断每一段中数据的类别。

目前国内对于基于内容检索技术的研究刚刚起步，音频分类也只局限于短的音频文件内。为了打破这个局限，本文研究了长的音频文件的分类与分段，并以“新闻联播”的音频数据为实验数据，建立了一个音频分类与分段的实验系统。在这个实验系统中，音频共被分成了四类：女声、男声、音乐背景下的语音以及音乐。

2、音频信息的分类与分段

2.1、基本方法

在本实验系统中，我们采用了先分类，后分段的方法。首先，以 20ms 的长度为一帧，分别确定每一帧所在的类别，得到一个帧类别序列，在此基础之上，相同类别的帧就可以合并为一个段落。当然，在实际中，会有一些分类错误的帧与段落，这样，我们还需要加入平滑过程，去掉这些夹杂在正确结果中的错误分类。可见，这种方法的关键在于尽量准确地得到每一帧音频数据的类别。

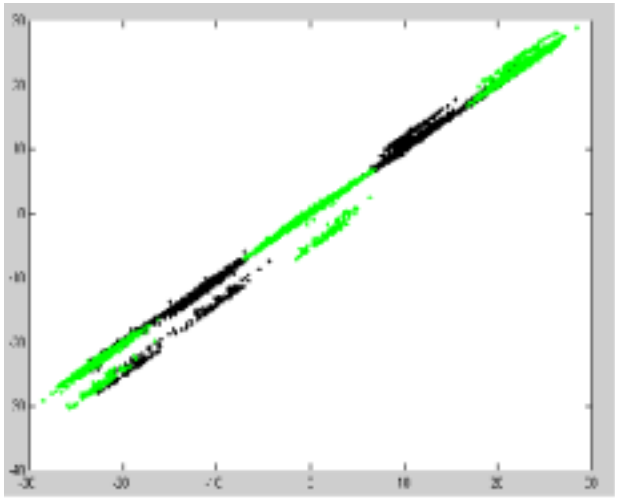
为了达到这个目的，我们采用了统计模式识别里面的贝叶斯分类方法来判别每一帧的类别。采用这种方法的基础在于，每一类音频数据的特征在空间里分布都是不同的。我们只要分别统计出它们的概率模型，就可以得到需要分类的帧对于各个类别的后验概率。选择后验概率最大的类作为该帧数据的类别即可。

注意：概率模型不但包括每类数据的类分布概率密度函数，还应包括它们的先验概率。作为分类标准的概率也不是先验概率或者类分布概率密度，而是先验概率与类分布概率密度的乘积——它与后验概率成正比。

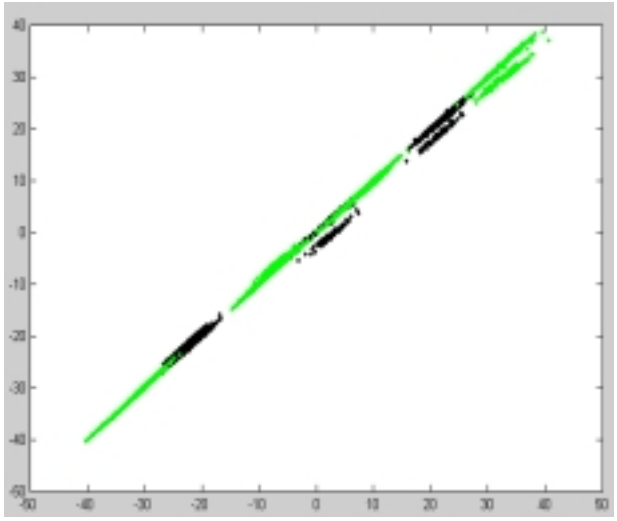
2．2、特征的选择

在本实验中所采用的音频特征为，MFCCs+MFCCs 的一阶差分。MFCC 的全称是 Mel 频率倒谱系数，它是进行音频分类的重要特征。与一般倒谱不同的是，MFCCs 是原始频谱经过一系列在 Mel 刻度上宽度相等的带通滤波器后，再变换到倒谱域上的倒谱。因此，MFCCs 更符合人的听觉特性，更能够更好地反映出不同类别的音频在人的听感中的差别。

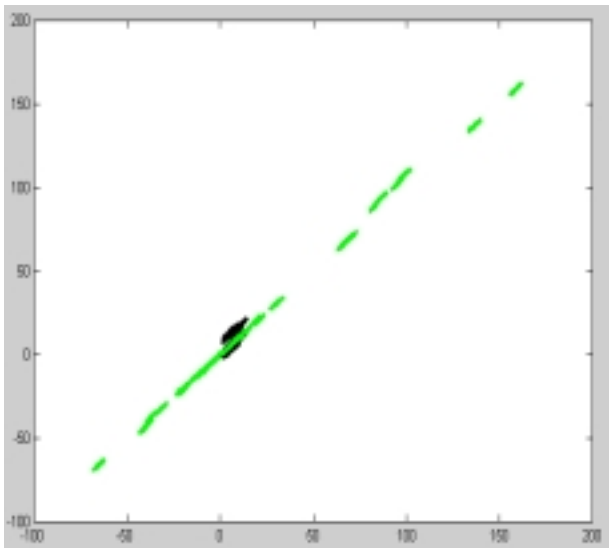
利用 K-L 变换，将上述特征降成 2 维，并在平面上表示出来。我们可以直观地看到，不同类别的音频，其特征在空间中的分布是不同的。



图一 女声与男声的特征对照图，绿色代表男声，
黑色代表女声



图二女声与音乐背景下的语音的特征对照图，绿色代表音乐背景下的语音，黑色表示女声



图三 女声与音乐的特征对照图，绿色代表音乐，黑色表示女声

2．3、概率模型的建立

概率模型的建立包括两个方面。其中，先验概率的估计可以通过统计原始音频数据中各个类别的音频占整体的比例而得到。然而，类分布概率密度函数的估计就不是那么简单的了。在这里，我们采用了混合高斯模型(GMM)来近似类分布概率密度函数。

简单地说，混合高斯模型就是首先将样本按照一定的聚类算法分成若干个部分，然后将每一部分的概率分布用一个高斯模型来近似，最后将各个部分的概率密度混合起来，就得到了整个样本集概率密度的估计。

建立混合高斯模型的关键在于聚类。为了体现概率统计的特性，我们采用了基于似然函数的 EM 动态聚类算法。实验证明，EM 动态聚类的迭代过程具有很好的收敛性，得到的混合高斯模型合理地反映了特征的分布情况。

3、 实验结果

实验结果的评价是从两个方面考虑的。第一，所划分段落的边界准确程度，即同一段落中是否含有不同类别的音频；第二，音频分类的正确率，这一点是只对那些边界准确的段落而言的。另外，所采用的测试数据绝大部分都在训练集外。

经测试发现，划分出来的段落边界准确，其准确率在 98%左右。经统计，音频分类正确率约在 80%~85%之间。在分类错误中，“女声错分为男声”所占的比重最大，为 60%。这主要是因为从特征图上来看，男声与女声的特征分布的确有一些部分是重叠的。另外，在女声模型的训练过程中，EM 迭代出现的震荡相对比较多，收敛情况相对较差；而男声模型的迭代过程则具有很好的收敛性质。这样，最后得到的女声模型可能不是充分收敛的模型。从这一点上可以理解，为什么“男声错分为女声”的错误相对要小得多。此外，存在的分类错误还有，“音乐背景下的语音”与男声、女声之间的分类错误，“音乐背景下的语音”与音乐之间的分类错误等。

测试结果显示，该实验系统的音频分类与音频分段性能基本令人满意。

4、结束语

本文提出了一种处理长时音频数据分类与分段问题的方法。虽然实验系统选取了“新闻联播”的音频作为实验数据，但本系统对于其他音频数据仍然有效，唯一的差别就在于音频类别及训练数据的选择不同。另外，本系统仅仅对于音频数据进行了初步的分类，而在此基础上，可以对不同类别的音频进行更细致的分类。如：对女声、男声类，可以进一步进行说话人识别；对于音乐，则可以进一步识别它的强弱、节奏等性质。

参考文献：

[1] 杨行峻、迟惠生等，《语音信号数字处理》，电子工业出版社，1995 年 8 月

[2] 林学闾，《模式识别基本教程》，清华大学计算机系校内讲义

[3] Erling Wold 、Thom Blum、 Douglas Keislar、 James Wheaton, 《Content-Based Classification，Search， and Retrieval of Audio》，IEEE MultiMedia 1996

[4] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland & S.J. Young, Segment Generation and Clustering in the HTK Broadcast News Transcription System》，http://www.informedia.cs.cmu.edu/html/enter.html

[5] 张尧庭、方开泰，《多元统计分析引论》，科学出版社

[6] C. Fraley and A. E. Raftery，《How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis》，Technical Report No.329，Department of Statistics ，University of Washington，http://www.stat.washington.edu/fraley/mclust/reps.shtml

[7] Chuanhai Liu and Don X. Sun，《Acceleration of EM Algorithm for Mixture Models Using ECME》，Statistics and Information Analysis Research，Bell Labs， Lucent Technologies