# 京剧中典型唱腔和伴奏的自动分类研究

张一彬　周　杰　边肇祺
(清华大学自动化系,北京 100084)
E-mail:zyb00@mails.tsinghua.edu.cn

**摘　要**　论文使用音频分析技术和模式识别技术相结合的方法对传统京剧中的3种典型角色(生、旦、净)的唱腔和2种典型的纯伴奏形式(文场和武场)进行了基于内容的自动分类研究。实验测试数据包括266个片段,来自于许多著名京剧演员如梅兰芳、袁世海、于魁智等人的演出。实验结果表明,对于5类分类问题可以达到88.7%的平均分类正确率,对于有伴奏下的唱腔和纯伴奏之间的2类分类问题可以取得高达96.6%的平均分类正确率。这一结果对京剧的进一步研究有着重要意义。

**关键词**　京剧　音频分析　模式识别

**文章编号** 1002-8331-(2006)01-0027-04　**文献标识码** A　**中图分类号** TP391

# Content-Based Classification on Beijing Opera

**Zhang Yibin　Zhou Jie　Bian Zhaoqi**
(Department of Automation, Tsinghua University, Beijing 100084)

**Abstract**: Among all kinds of music styles, Beijing opera is very familiar in China. In this paper, we present a study on content-based classification among five kinds of typical aria and accompaniment of Beijing opera (including Sheng, Dan, Jing, Wenchang, and Wuchang), using audio analysis techniques together with pattern recognition techniques. A comparative evaluation between seven different classifiers is carried out on a testing database of 266 segments, and the results show that the BP neural network classifier (BPNNC) works best, and its average classification accuracy can achieve 88.7% in the five-class classification problem.

**Keywords**: Beijing opera, audio analysis, pattern recognition

## 1　简介

戏曲是中国传统文化中的瑰宝,是世界艺术园地中的一支奇葩。与其它艺术形式相比,中国传统戏曲显得丰富多彩,比较有影响力的剧种就有数十种之多,其中最具代表性的首推京剧。采用计算机技术对其进行分析有着重要意义。

在京剧当中,无论是唱腔还是伴奏都有着非常精细而严格的分工。京剧的角色体制大致可分为"四行",即:生、旦、净、丑。由于演员所扮演的角色的身份、年龄、性格、扮相、以及表演上唱、念、做、打的不同,在每一大行中又有比较细微的分工。各行角色的唱腔在保持了统一的京剧风格的同时又各具特色,京剧专家或爱好者能够比较容易地将其区分开来;在伴奏方面,京剧乐队是由打击乐器和管、弦乐器组成的,总称为场面或文武场。一般重唱工的文戏以管弦乐伴奏为主(虽也少不了必要的打击乐器),因此传统习惯称管弦乐为文场。而打击乐器虽然只能奏出一个有固定高低的音,但音响强烈、节奏感鲜明,一般重武打的武戏以打击乐伴奏为主,因此传统习惯称打击乐为武场。

作为一门系统艺术,京剧的各个组成部分(尤其是不同角色的唱腔)虽然特色鲜明,但同时也保持着很高的一致性和相似性,这也使得我们很难对其进行基于内容的自动分类。无论是与处理包含着各种相去甚远的音频类别(如:语音、音乐、环境声音和静音等)的数据相比,还是同处理一般的音乐类别(如:摇滚乐、钢琴曲、流行歌曲、交响乐等)的数据相比,对京剧内部的不同成份进行基于内容的自动分类无疑都具有更大的挑战性。到目前为止,还没有任何有关这方面研究工作的文献报道。

在相关工作中,Zhang等人提出了一种基于内容的音频分类、分割方法[1],他们利用4个短时特征对由下面7个类别组成的音频信号进行自动分类和分割,这7个类别包括:语音、纯音乐、歌曲、环境声音、带有音乐背景的语音、带有音乐背景的环境声音和静音。在文献[2]中,作者提出了一种方法将音频分割为语音、音乐、环境声音和静音。他们首先将音频信号分为语音信号和非语音信号两类,然后进一步将非语音信号分为音乐、环境声音和静音。在过去的工作中,我们在较大的数据库下对音频数据流的分类和分割问题也进行了较为详尽的研究[3,4],所涉及的音频数据类别包括:钢琴曲、交响乐、京剧、流行歌曲和语音。文献[5,6]针对语音类信号和音乐类信号进行分类。T. Lambrou等人利用时域和小波变换域中的一些统计特征对爵士乐、摇滚乐和钢琴曲这3类音乐数据做了自动分类[7],但是他们的测试数据库中仅仅包含了12首音乐。

**基金项目**:国家自然科学基金资助项目(编号:60205002;60332010);北京市自然科学基金资助项目(编号:4042020)

**作者简介**:张一彬(1974-),男,现在清华大学自动化系攻读工学博士学位。研究兴趣包括模式识别、基于内容的音频及音乐分析、信息挖掘等。周杰,博士,清华大学自动化系教授、博士生导师,IEEE高级会员,国际学术刊物《Int.J.Robotics and Automation》编委。边肇祺,清华大学自动化系资深教授、博士生导师。

在本文中,我们选择京剧中3种角色(生、旦、净)在有伴奏下的唱腔以及2种典型的纯伴奏(文场和武场)作为我们的研究对象。由于京剧中的"丑角"更多地是通过其扮相、形体动作、道白等体现其角色特征,因而本文不将其唱腔纳入研究范围。我们收集了许多著名京剧演员如梅兰芳、袁世海、于魁智等人的演出记录作为实验数据,并通过人工分割的方法获得相应角色和纯伴奏的样本数据,然后采用音频分析技术和模式识别技术相结合的方法对这5类数据进行了自动分类研究。在实验中我们比较了7种常用的分类器,发现神经网络分类器(BPNNC)比较适合用于京剧分类。实验结果表明,在一个包含了266个片段的测试数据库上,对于5类分类问题,我们的方法可以达到88.7%的平均分类正确率;对于有伴奏下的唱腔和纯伴奏之间的2类分类问题,我们的方法可以取得高达96.6%的平均分类正确率。

论文的余下部分组成如下:第二、三小节分别简述了所使用的音频特征和分类器;第四小节给出了具体的实验结果;最后的总结在第五小节中给出。

## 2 特征提取

每个音频片段被转换成采样率为11 025/s、量化位数为8位的混和单声道WAV文件。我们从每个片段中提取一系列音频特征并组成特征向量。实验中求取音频特征时所使用的"帧"的长度为512个采样点,约46ms;相邻帧之间有112个采样点的重叠区域,约10ms。借鉴音频分析与识别领域中的前人工作,我们在本文中所使用的基本音频特征选择为:

(1)短时能量(Short-time energy(SE))[3]

音频信号的短时能量函数定义为:

$$E_n=\frac{1}{N}\sum_m[x(m)w(n-m)]^2 \tag{1}$$

其中,$\{x(m)\}$表示离散时间音频信号,$n$表示帧数,$\{w(m)\}$表示一个长度为$N$的矩形窗。

(2)低短时能量值比率(Low short-time energy ratio (LSTER))[2]

LSTER是从短时能量特征中衍生出来的一个大尺度特征,其定义为一段音频信号中短时能量值较低的帧数在这段音频信号中所占的比率,其表达式为:

$$LSTER=\frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(0.5\overline{wse}-se(n))+1] \tag{2}$$

其中,$N$是这段音频信号中所包含的音频帧的总数;$se(n)$是第$n$个音频帧的短时能量值;$\overline{wse}$是这段音频信号的平均短时能量值;sgn[·]是符号函数。

(3)过零率(Zero-crossing rate(ZCR))[3]

短时平均过零率定义为:

$$Z_n=\frac{1}{2}\sum_m|\mathrm{sgn}[x(m)]-\mathrm{sgn}[x(m-1)]|w(n-m) \tag{3}$$

其中,$\{x(m)\}$表示离散时间音频信号,$n$表示帧数,$\{w(m)\}$表示一个矩形窗。

(4)高过零率比率(High zero-crossing rate ratio(HZCRR))[2]

HZCRR是从过零率特征中衍生出来的一个大尺度特征,其定义为一段音频信号中过零率值较高的帧数在这段音频信号中所占的比率,其表达式为:

$$HZCRR=\frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(zcr(n)-1.5\overline{wzcr})+1] \tag{4}$$

其中,$N$是这段音频信号中所包含的音频帧的总数;$zcr(n)$是第$n$个音频帧的过零率值;$\overline{wzcr}$是这段音频信号的平均过零率值;sgn[·]是符号函数。

(5)Mel频率(Mel frequency(MF))[3]

首先我们通过Yule-Walker自回归谱估计法得到每个音频帧的频谱。在本文中我们选择了40阶的自回归模型[1],其参数可通过对音频信号的自相关估计出来。由于通过这种方法得到的频谱曲线非常光滑,使得我们可以很容易地确定其中的波峰和波谷。然后根据频谱中各个波峰的形态和间隔,我们就可以确定这帧音频信号的频谱曲线中是否存在共振峰。如果存在共振峰的话,我们就把第一共振峰所对应的频率确定为这帧音频信号的基频,并进一步将其转换为Mel频率。否则我们认为这帧音频信号属于非乐音,并将其Mel频率值记为零。

(6)和谐度(Harmonious degree(HD))[3]

一段音频信号中乐音所占的比率称为这段音频信号的和谐度。其表达式为:

$$HD=\frac{N_{har}}{N_{all}} \tag{5}$$

其中$N_{har}$表示这段音频信号中具有非零的Mel频率值的帧数,$N_{all}$表示总的帧数。

(7)谱矩(Spectral centroid(SC))[8]

谱矩是在短时傅立叶变换的基础上定义的一个频域特征,第$i$帧音频信号的谱矩表达式为:

$$SC_i=\frac{\sum_{u=0}^{M}u|f_i(u)|^2}{\sum_{u=0}^{M}|f_i(u)|^2} \tag{6}$$

其中$F_i=\{f_i(u)\}_{u=0}^{M}$表示第$i$帧音频信号的短时傅立叶变换,$M$为最高的频带号。

(8)带宽(Bandwidth(BW))[8]

接着谱矩的定义,第$i$帧音频信号的带宽表达式为:

$$BW_i^2=\frac{\sum_{u=0}^{M}(u-c_i)^2|f_i(u)|^2}{\sum_{u=0}^{M}|f_i(u)|^2} \tag{7}$$

其中$c_i$为第$i$个频带的中心频率。

(9)频谱滚动频率(Spectral rolloff frequency)[8]

频谱滚动频率的定义式为:

$$SRF_i=\max\left(h\left|\sum_{u=0}^{h}f_i(u)<TH\cdot\sum_{u=0}^{M}f_i(u)\right.\right) \tag{8}$$

其中$TH$是一个取值范围在0~1之间的阈值,在我们的实验当中$TH$的值为0.9。

(10)谱通量(Spectrum flux(SF))[2]

谱通量定义为一段音频信号当中相邻两帧之间的平均频谱变化量。在我们的实验中,用于计算谱通量的窗宽和步长分别为0.2s和0.1s。谱通量的表达式为:

$$SF=\frac{\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log(A(n,k)+\delta)-\log(A(n-1,k)+\delta)]^2}{(N-1)(K-1)} \tag{9}$$

其中$A(n,k)$为第$n$帧音频信号的离散傅立叶变换结果,即:

$$A(n,k)=\left|\sum_{m=-\infty}^{\infty}x(m)w(nL-m)e^{-j(2\pi/L)km}\right| \tag{10}$$

这里 $x(m)$ 是原始音频数据，$w(m)$ 是窗函数，$L$ 是窗宽，$K$ 是离散傅立叶变换的阶数，$N$ 是窗宽范围内的总帧数，$\delta$ 是一个极小值用于避免计算溢出。

有了基本的音频特征后，我们将从每个样本中提取出一个17维的特征向量，并用这个特征向量来代表这个音频片段。这个17维的特征向量中包括短时能量序列、过零率序列、Mel频率序列、谱矩序列、带宽序列、谱通量序列以及频谱滚动频率序列的均值和标准差再加上低短时能量值比率、高过零率比率和和谐度。值得注意的是在计算Mel频率序列的均值和标准差时将不考虑Mel频率值置为0的那些帧。

## 3 分类器选择

在实验中，我们比较了7种常见的分类器，下面分别简单介绍一下这几个分类器：

(1)帕森分类器(PC)[4,9]

首先用帕森窗法估计出每类的概率密度分布 $\hat{p}(x|\omega_i)$，然后利用Bayes准则将未知类别的样本分到具有最大后验概率的类别中去，即：

$$\text{if } \hat{p}(\omega_i|x)=\max_{j=1}^{c} P(\omega_j|x),\text{then } x\in\omega_i \tag{11}$$

其中，$\boldsymbol{x}$ 表示未知样本的特征向量，$\omega_i$ 表示第 $i$ 个类别，$c$ 表示类别总数。

(2)K近邻法分类器(k-NNC)[3,4,9]

这个方法就是取未知样本 $x$ 的 $k$ 个近邻，看这 $k$ 个近邻中多数属于哪一类，就把 $x$ 归为哪一类。具体说就是在 $N$ 个已知样本中，找出未知样本 $x$ 的 $k$ 个近邻。若 $k_i$ 是 $k$ 个近邻中属于 $\omega_i$ 类的样本数，则定义决策规则为：

$$\text{if } k_j=\max_i k_i,\text{then } x\in\omega_j \tag{12}$$

在我们的实验中采用欧式距离，并且 $k$=3。

(3)最小Mahalanobis距离分类器(MMDC)[4,9]

该分类器采用未知样本 $x$ 同各个类别之间的Mahalanobis距离作为分类准则，并将该样本分到与之Mahalanobis距离最近的类别中去。Mahalanobis距离的定义式为：

$$m_i(x)=\sqrt{(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)} \tag{13}$$

其中，$\boldsymbol{x}$ 表示未知样本的特征向量，$\boldsymbol{\mu}_i$ 表示第 $i$ 个类别的均值向量，$\sum_i$ 表示第 $i$ 个类别的协方差矩阵，$m_i(x)$ 表示未知样本 $x$ 与第 $i$ 个类别之间的Mahalanobis距离。

(4)支持向量机(SVM)[3,4,9,10]

SVM分类器是基于结构风险最小化的，它的泛化推广能力在理论上要优于基于经验风险最小化的分类器。假设 $\{x_i\}$ 为包含着 $n$ 维空间中属于两个不同类别数据的集合，则它的最优分类面可以表示为：

$$f(x)=sign\left(\sum_i a_i y_i K(x_i,x)+b\right) \tag{14}$$

其中，$y_i=\pm1$ 表示样本 $x_i$ 的类别归属，$K(x,y)$ 是一个正定对称的函数，$a_i$ 和 $b$ 是两个参数，它们可以从训练集中估计出来。决策面主要由被称为核函数的 $K(x,y)$ 所决定，比较常见的核函数类型有：多项式核函数、指数核函数、高斯核函数、sigmoid核函数和距离核函数。在我们的实验中，2阶的距离核函数相对其它核函数而言具有最佳的分类效果。

(5)二次分类器(QC)[4,9]

基于正态分布假设的二次分类器，其决策规则为：

$$\begin{cases} g_i(x)=x^T W_i x+w_i^T x+w_{i0} \\ \text{if } g_i(x)>0,\text{then } x\in\omega_i \end{cases} \tag{15}$$

其中，$\boldsymbol{x}$ 表示未知样本的 $d$ 维特征向量，$\boldsymbol{W}_i$ 表示一个 $d\times d$ 维的实对称矩阵，$\boldsymbol{w}_i$ 和 $\boldsymbol{w}_{i0}$ 为两个 $d$ 维向量。

(6)Fisher准则分类器(FC)[4,9]

这是一个基于Fisher准则的线性分类器。首先通过训练集中的数据计算出最优的投影方向，然后将测试集中的数据沿着这个方向做投影，最后通过一个合适的阈值将测试集中的数据进行分类。

(7)BP神经网络分类器(BPNNC)[3,4,11]

BP神经网络分类器是一个采用变步长BP训练算法的3层前馈结构神经网络，其输入层节点数等于样本特征向量的维数，输出层节点数等于类别总数，中间层包含11个节点，其网络结构示意图见图1。节点函数采用Sigmoid函数。

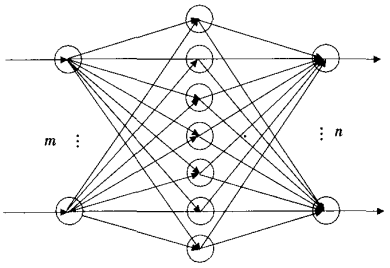


**图1 BP神经网络分类器网络结构示意图**

## 4 实验结果

我们采用基于样本训练的方法实现基于内容的京剧唱腔及伴奏的自动分类。首先我们收集了十多位著名传统京剧演员的演出记录50余段；然后采用手工分割的方法获得相应的有伴奏下的角色唱腔和纯伴奏的样本数据，其中“旦”137段、“净”98段、“生”127段、“文场”109段、“武场”62段，每个样本片段的长度为4s至9s不等；最后我们将这533个属于各个类别的样本片段平均分配到训练集和测试集中。对于训练集中的每个样本，我们将从中提取出一个17维的特征向量用于训练分类器。有了所需的分类器后，我们就可以对来自测试集的样本特征向量进行分类。实验中，我们比较了在第三小节中所介绍的各个分类器。5类情况下，各个分类器在测试集上所得到的平均分类正确率见表1；我们进一步将“生”、“旦”、“净”合并为“唱腔”类，将“文场”、“武场”合并为“场面”类。则各个分类器在测试集上所得到的平均分类正确率见表2。

**表1 在5类情况下，不同分类器在测试集上所得到的平均分类正确率**

| PC | K-NNC | MMDC | SVM |
|---|---|---|---|
| 69.2% | 59.4% | 86.8% | 75.2% |
| QC | FC | BPNNC | … |
| 88.0% | 82.7% | 88.7% | … |

从表1和表2中我们可以看出，在区分京剧内部不同角色唱腔和纯伴奏的分类问题中，BPNNC的分类效果要优于其它分类器，这一点与我们在文献[3,4]中所得出的结论是相同的。BPNNC在5类分类和2类分类时所得到的详细分类结果分别

见表 3 和表 4。

表 2　在 2 类情况下，不同分类器在测试集上所得到的平均分类正确率

| PC | K-NNC | MMDC | SVM |
|---|---|---|---|
| 85.0% | 83.1% | 89.5% | 90.2% |
| QC | FC | BPNNC | … |
| 90.6% | 90.6% | 96.6% | … |

表 3　在 5 类分类情况下，BPNNC 在测试集上所取得的详细分类结果

| | 旦 | 净 | 生 | 文场 | 武场 |
|---|---|---|---|---|---|
| 旦 | 92.6% | 0 | 2.9% | 2.9% | 1.5% |
| 净 | 6.1% | 89.8% | 2.0% | 0 | 2.0% |
| 生 | 6.3% | 4.7% | 82.8% | 6.3% | 0 |
| 文场 | 7.4% | 0 | 3.7% | 88.9% | 0 |
| 武场 | 3.2% | 0 | 0 | 6.5% | 90.3% |

注：其平均分类正确率为 88.7%，其中第 $i$ 行第 $j$ 列上的数字表示属于第 $i$ 类的数据被分类器分到第 $j$ 类中的百分比

表 4　在 2 类分类情况下，BPNNC 在测试集上所取得的详细分类结果

| | 唱腔 | 场面 |
|---|---|---|
| 唱腔 | 97.8% | 2.2% |
| 场面 | 5.9% | 94.1% |

注：其平均分类正确率为 96.6%，其中第 $i$ 行第 $j$ 列上的数字表示属于第 $i$ 类的数据被分类器分到第 $j$ 类中的百分比

## 5　结论

本文在收集了大量实验数据的基础上，对京剧中的 5 种典型的角色唱腔和伴奏进行了自动分类研究。实验结果表明通过采用基于样本的方法，可以比较有效地解决京剧中不同成份之间的自动分类问题。在一个包含了 266 个片段的测试数据库上，对于 5 类分类问题，我们的方法可以达到 88.7%的平均分类正确率；对于有伴奏下的唱腔和纯伴奏之间的 2 类分类问题，我们的方法可以取得高达 96.6%的平均分类正确率。将来我们会继续研究京剧中不同角色唱腔和伴奏所具有的特征，希望通过学习与规则相结合的方法来进一步提高算法的效果。

（收稿日期：2005 年 10 月）

## 参考文献


1.T Zhang，C J Kuo.Audio Content Analysis for Online Audiovisual Data Segmentation and Classifica-tion[J].IEEE Trans Speech and Audio Processing，2000；9(4)：441~457

2.L Lu，H J Zhang，H Jiang.Content Analysis for Audio Classification and Segmentation[J].IEEE Trans.Speech and Audio Processing，2002；10(7)：504~516

3.Y B Zhang，J Zhou.A study on content-based music classification[C].In：IEEE Proc Seventh International Symposium on Signal Processing and Its Applications，France，2003；2：113~116

4.Y B Zhang，J Zhou.Audio Segmentation Based on Multi-Scale Audio Classification[C].In：IEEE Proc ICASSP，2004；4：349~352

5.W Chou，L Gu.Robust Singing Detection in Speech/Music Discriminator Design[C].In：IEEE Proc ICASSP，Salt Lake City，USA，2001；2：865~868

6.J Ajmera，I A Mccowan，H Bourlard.Robust HMM-Based Speech/Music Segmentation[C].In：IEEE Proc ICASSP，Orlando，USA，2002；1：297~300

7.T Lambrou，P Kudumakis，R Speller et al.Classification of Audio Signals using Statistical Features on Time and Wavelet Transform Domains[C].In：IEEE Proc ICASSP，Seattle，USA，1998；6：3621~3624

8.D G Li，I K Sethi，N Dimitrova et al.Classification of General Audio Data for Content-Based Retrieval[J].Pattern Recognition Letters，2001；22(5)：533~544

9.边肇祺，张学工.模式识别[M].第 2 版，清华大学出版社，2000-01

10.G D Guo，S Z Li.Content-Based Audio Classification and Retrieval by Support Vector Machines[J].IEEE Trans Neural Networks，2003；14：209~215

11.B D Ripley.Pattern Recognition and Neural Networks[M].Cambridge University Press，1996


（上接 26 页）

算法简单、健壮性较好；并引入了重叠网格主从关系图有效解决多层次复杂重叠情形；建立基于网格的可调节 kd 树提高找重效率。实验结果表明该技术在大网格量、复杂重叠区域时仍能得到较好的“挖洞”结果和较高的效率。在实际流场计算中，我们发现重叠区域的主从网格尺度差异和重叠网格数对计算的稳定性和计算结果有一定的影响，如何合理使用重叠网格技术是下一步研究的重点。（收稿日期：2005 年 11 月）

## 参考文献


1.Robert L Meakin.Object X-Rays for Cutting Holes in Composite Overset Structured Grids[C].In：15th AIAA computational fluid dynamics conference，Anaheim，CA，AIAA，2001：2001~2537

2.庞宇飞，洪俊武.交点判别法在重叠网格中的应用[C].见：CARDC 计算空气动力学研究所.CARDC 计算空气动力学研究所 2004 青年科技报告会，中国绵阳，2004

3.李亭鹤，阎超.一种新的分区重叠洞点搜索方法——感染免疫法[J].空气动力学报，2001；19(2)：156~160

4.李亭鹤.重叠网格自动生成方法研究[D].北京航空航天大学，2004

5.Norman E Suhs，Stuart E Rogers.PEGASUS 5：An Automated Pre-Processor For Overset-Grid CFD[J].AIAA Journal，2003；41(6)：1037~1045

6.CFD Research Corporation.CFD-FASTRAN Theory Manual Version[M].2002

7.C Y Lee.An Algorithm For Path Connections And Its Applications[J].IRE Trans on Electronic Computers，1961；EC_10(3)；346~365

8.Jonathan B Rosenberg.Geographical Data Structures Compared：A Study of Data Structures Supporting Region Queries[J].IEEE Transactions on Computer-aided Design，1985；CAD-4(1)：53~68