基于 SVM 的流行音乐中人声的识别

石自强,李海峰,孙佳音
SHI Zi-qiang,LI Hai-feng,SUN Jia-yin

哈尔滨工业大学 计算机科学与技术学院,哈尔滨 150001
Department of Computer Science and Technology,Harbin Institute of Technology,Harbin 150001,China
E-mail:zqiangshi@126.com

SHI Zi-qiang,LI Hai-feng,SUN Jia-yin.Vocal discrimination in pop music based on SVM.Computer Engineering and Applications,2008,44(25):126-128.

Abstract: Facing the problem of vocal discrimination in pop music,the authors propose applying MFCC parameters as features, and Support Vector Machine (SVM) as classifier.Due to the continuity of audio signal features,the authors consider low-pass filtering to the classification results as post-processing.Experiment results show that at frame level,a quite promising classification accuracy of 85.76% can be obtained.It is also revealed that singers with different languages have large vocal differences in pronunciation, especially in MFCC feature statistics.The classification results may be used as a similarity measure for music structure analysis in the future work.

Key words: pop music structure analysis;Vocal/Non-vocal discrimination;Support Vector Machine(SVM)

摘 要:针对流行音乐中人声的发现问题,使用 SVM 分类器针对 MFCC 特征进行训练和分类。依据音频特征的连续性,后期对分类结果进行低通滤波。实验结果表明,该方法在帧层面上的识别率可以达到 85.76%。实验中也发现不同语种的演唱者在发音上,特别是在 MFCC 特征上存在很大的统计差异性。实验中对歌曲分类的结果可以作为近一步实现音乐相似性度量的依据之一。

关键词:流行音乐结构分析;人声发现;支持向量机
DOI:10.3778/j.issn.1002-8331.2008.25.038 **文章编号:**1002-8331(2008)25-0126-03 **文献标识码:**A **中图分类号:**TP391.4

1 引言

流行音乐中人声的识别对于音乐的文摘、检索、翻译以及曲风分类都有很大的帮助,另外还可以作为歌词识别的一个前端处理。流行音乐一般由 intro、verse、chorus、bridge 以及 outro 这五部分经过组合而成。Intro 是一首歌曲的引子部分,verse 是歌曲的叙事部分,chorus 是高潮部分,一般人们哼唱的是 verse 以及 chorus(多数情况),两者可以作为音乐的文摘,bridge 是过渡段,outro 是歌曲逐渐结束的尾声。一首典型流行音乐的结构如图 1 所示。歌曲的 intro、bridge 和 outro 部分基本都是纯音乐组成,因此完成纯音乐和人声的区分对于流行音乐的结构分析有着很大的辅助作用。歌曲的结构分析又是音乐检索以及相似性研究的前期工作,可见研究人声识别的问题很有意义。目前,虽然语音处理领域取得了巨大进展,但由于歌唱和话音的产生以及被人耳感知的机理有很大的不同,另外乐器的频谱包络曲线和歌唱的基音的包络曲线很相似,并且在流行音乐中,歌唱的和音结构常常是和键盘或者弦乐器的和音结构重合在一起的[1],所以这个问题还是一个很困难的问题。N.C. Maddaged[2]等提出了一种二次迭代复合傅立叶变换 (Twice-Iterated Composite Fourier Transform)来检测歌唱的开始和结束的技术,他们用这种方法来刻划一帧的和音结构,从而用于识别人声,但是帧识别率只达到 80%。Fujihara[3]等猜测颤音(vibrato)可以作为人声识别的一个重要线索。New,T.L.[4]等指出由颤音引起的语音特征可以很好的刻画不同歌手发音方式。

本文尝试采用 MFCC[5]作为识别的特征,使用 SVM 分类器来进行人声分类,并同在很多问题上都取得了很好效果的人工神经网络 ANN、混合高斯模型 GMM 和隐马尔科夫模型 HMM 等分类器进行了对比。由于音频数据的连续性,在后期对分类结果进行低通滤波纠错,从而进一步提高了识别率。

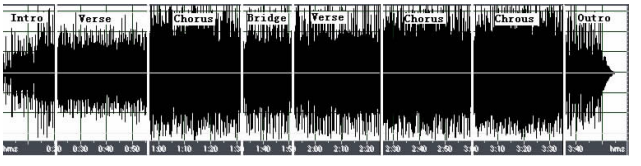

图 1 歌曲《青花瓷》的结构

2 基于 SVM 的纯音乐和人声的区分

支持向量机是一种模式分类方法,它依赖于对数据的预处理,即,在更高维的空间表达模式,并且通常比原来的特征空间

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z197)。
作者简介:石自强(1983-),男,硕士在读,主要研究音乐结构自动分析;李海峰(1969-),男,博士,教授,主要研究音频信息检索与处理、情感语音处理、多模式人机界面、人工神经网络、智能信息处理与数据挖掘、智能化测量技术等;孙佳音(1982-),男,博士在读,主要研究音乐信息检索。
收稿日期:2008-03-19 **修回日期:**2008-05-19

的维数高很多。支持向量机中关键是核函数的选择,这里为了取得比较好的分类结果,选择计算量相对较大的径向基函数作为核函数:

$$K(x_i,x_j)=\exp\left(-\frac{(x_i-x_j)^2}{\sigma^2}\right) \quad (1)$$

这样在高维空间训练一个支持向量机的目标是找到一个具有最大间隔的分隔平面;如果间隔越大,得到的分类器也越好。

支持向量是最接近超平面的。支持向量是那些定义最优分隔超平面的训练样本,也就是最难被分类的模式。非形式地说,它们就是对求解分类任务的最富有信息的模式。

之所以选择支持向量机方法,是因为其一个重要的优点是所获得的分类器的复杂度可以采用支持向量的个数,而不是变换空间的维数来刻划。因此支持向量机方法往往不像一些别的方法一样发生过拟合的现象[6]。

将训练音频数据分帧,每一帧提取特征向量,这里使用语音处理中常用的 MFCC 特征及其的一价和二阶差分,将这些帧分为两类,分别为纯音乐以及人声。然后用这些数据对 SVM 进行训练,得到支持向量以及最优分类超平面。

对于未知的音频信号,首先对其分帧,设一共得到 L 帧,然后利用前面训练好的 SVM 进行分类,设分类结果为 $w(t)\in\{-1,+1\}$,$t=1,2,\cdots,L$。由于音乐音频信号的连续性,歌唱的发声以及音乐的产生不是骤然停止或者开始,而是具有一定的延续性,人声帧的周围有很大可能还是人声帧,音乐帧的周围很有可能还是音乐帧,所以考虑在 SVM 的分类结果的基础上进行低通滤波[7],滤掉高频的骤变的分类结果,从而得到连续的分类结果。式(2)是对于原始的分类结果进行的低通滤波。低通滤波器容许低频信号通过,但减弱(或减少)频率高于截止频率的信号的通过。当使用在音频应用时,它有时被称为高频剪切滤波器,或高音消除滤波器。通过剔除短期波动、保留长期发展趋势提供了信号的平滑形式。

$$LPw(t)=\frac{1}{2*N+1}\left(\sum_{i=-N}^{N}w(t+i)\right) \quad (2)$$

其中在实验中 N 取为 50。然后通过公式(3)将低通滤波结果分为两类。

$$SLPw(t)=\begin{cases}1, LPw(t)>0\\-1, LPw(t)\leqslant 0\end{cases} \quad (3)$$

这样就得到了最终的帧分类结果 $SLPw(t)$。

图 2 是对歌曲《青花瓷》的人工标注,可以清楚地看到歌曲中的 intro、bridge 和 outro 部分。图 3 是 SVM 的区分结果,可以

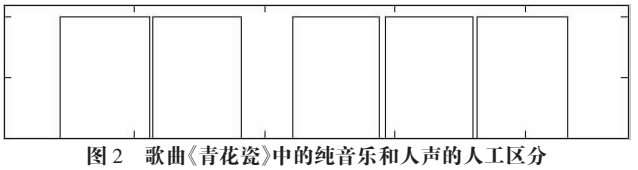

图 2　歌曲《青花瓷》中的纯音乐和人声的人工区分

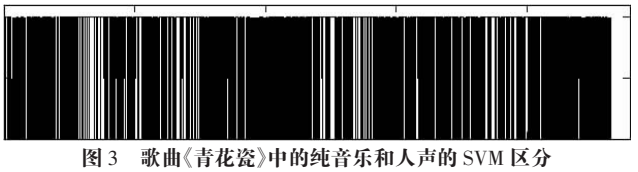

图 3　歌曲《青花瓷》中的纯音乐和人声的 SVM 区分

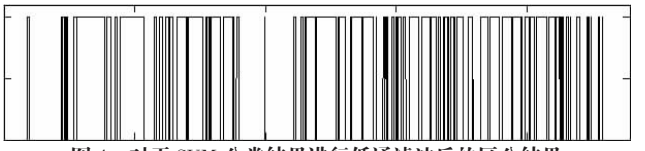

图 4　对于 SVM 分类结果进行低通滤波后的区分结果

看出这样的区分几乎没有什么使用价值。图 4 是对图 3 的分类结果进行低通滤波之后的结果,可以清楚地看到 intro 和 outro 部分,这对于音乐的结构分析是很有帮助的。

低通滤波可以明显地去掉突兀的分类结果,得到整块的人声部分以及整块的音乐部分。研究过程中曾经尝试合并一些人声段,从而形成一个整体的人声块。但是由于有些歌曲比较抒情,演唱的特点是其中有很多 1~3 秒左右的停顿,这 1~3 秒就是纯音乐。所以如果这样,帧识别率可能降低。Intro、bridge 和 outro 段几乎是纯音乐,可以通过设置阈值从而形成一个整体的纯音乐块,从而作为歌曲的一部分结构信息。

3　音频数据库及系统整体框架

3.1　音频数据库

本音频数据库共有 100 段音乐,每段 15 秒,包括流行、摇滚、经典、说唱、爵士和乡村等,其中按顺序由多到少分类选取,但是除流行以外其他种类都不少于 5 段。中文歌曲 40 段,英文歌曲 60 段。对每段音乐分帧,20 ms 为一帧,对于每一段都配有一个.lab 文件,给出每一帧的标记。标记有两种,分别是纯音乐(-1)和演唱部分(+1)。提取的特征是 MFCC 特征及其一阶差分和二阶差分。在选择分类器阶段,利用 60 段英文歌曲,其中 40 段作为训练语料,20 段作为测试语料,分别使用 ANN、GMM、HMM 和 SVM 来进行实验,结果是 SVM 有明显的优势。在对歌曲进行实验时,利用数据库中的标记好的 40 段中文歌曲进行训练,然后以 15 首完整的流行歌曲作为测试。在挑选这 15 首歌曲时尽量使曲风多种多样,包括流行、摇滚、说唱和民谣等等。

3.2　系统整体框架

系统的整体框架如图 5 所示。首先利用训练语料对分类器进行训练,然后对歌曲做歌唱和纯音乐的区分,之后对结果进行低通滤波。

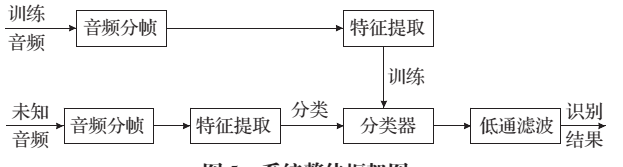

图 5　系统整体框架图

4　实验结果及分析

实验过程中,做了两个独立的实验。

实验 1 是 SVM 和其他三种分类器的对比实验。选用 13 维 mel 倒谱系数(MFCC)及其一阶和二阶差分。分类器采用 SVM,另外同 ANN、GMM 还有 HMM 进行对比。其中 ANN 分三层,输入层 39 个神经元,隐含层 40 个神经元,输出层 1 个神经元。GMM 采用 39 维的高斯,由 39 个混合而成。HMM 有两个状态,每个状态的输出由混合高斯决定。

实验是使用英文歌曲进行的,其中 40 段作为训练数据,20 段作为测试数据。这里只是基于帧,没有进行低通滤波。

基于帧的识别率的计算如下：

$$帧识别率=\frac{正确识别帧数}{总帧数}\times 100\% \quad (4)$$

表 1 给出了在相同条件下 ANN、HMM、GMM 以及 SVM 的分类效果，其中 SVM 的分类效果最好。

表 1 各种识别方法的识别率对比

	ANN	GMM	HMM	SVM
平均识别率	60.79%	63.9%	60%	67.18%

实验 2 是对第三章提出方法的有效性检验。其中训练数据是数据库 40 段中文歌曲片段，测试数据是 15 首中文流行歌曲，其中《Shall we talk》、《One day》都是中文歌曲。

表 2 给出了用中文歌曲音频数据训练的 SVM 对于 15 首中文歌曲的分类结果以及低通滤波后的识别率对比。第 2 列是 SVM 基于帧的正确识别率。第 3 列是 SVM 分类结果经过平滑滤波后的识别率，识别率平均提高 11.9%。第 4 列是歌曲中人声所占的比例，其中除了《宠姬》、《黑色柳丁》之外，其他的都比滤波后识别率低，说明滤波确实是有效用的。第一行中的数字是所在列数据的平均。

表 2 SVM 对于 15 首中文歌曲的纯音乐和人声的区分识别率

	SVM 分类识别率（73.83%）	滤波后识别率（85.76%）	歌曲中人声比例（77.53%）
昆明湖	77.44%	93.22%	87.90%
青花瓷	74.34%	86.72%	70.80%
Shall we talk	76.60%	88.70%	86.8%
宠姬	66.89%	77.38%	79.26%
痒	73.30%	82.97%	69.02%
一了百了	74.17%	85.20%	83.04%
记忆中的向日葵	71.85%	81.86%	67.23%
烟味	72.19%	82.02%	68.12%
一直很安静	72.11%	83.32%	60.47%
可爱女人	80.40%	87.36%	84.92%
One Day	76.48%	93.37%	79.94%
甜蜜蜜	64.54%	77.65%	76.42%
崇拜	75.64%	89.21%	81.34%
黑色柳丁	71.58%	83.56%	85.01%
给安娜	80.37%	94.00%	82.68%

表 2 中《昆明湖》、《给安娜》和《One Day》的识别率最高，前面两首是说唱风格比较重的歌曲，因此有很多的说话特征在里面，所以分类效果比较好。但和传统流行音乐相比，它的音乐性较差，其面向的听众群是年轻人。《One Day》是一首抒情摇滚风格的歌曲，演唱特点比较高亢，从而在发音特点上有所表征，而这种表征是进一步详细研究的问题。《宠姬》和《甜蜜蜜》分类效果较差，前一首可能是因为其是粤语歌曲，所以在发音特点上和普通话有比较大的不同，后一首是著名歌手邓丽君演唱的，她歌唱技巧很高，音乐性较强，分类效果较差。其他的歌曲识别率都在 80%~90%之间，大部分属于流行的风格，而市场上大部分是这样的歌曲，所以本方法是比较有效的。上述结论对于英文歌曲同样适用。

实验中发现当使用本语种歌曲训练的分类器去区分本语种的歌曲时，比区分其他语种的歌曲时有更好的效果，因此可以利用混合的各语种歌曲训练的分类器。当使用英文歌曲训练的分类器去区分中文歌曲时，分类效果（甚至包括滤波后）都只有在 50%左右，几乎和猜测差不多。这说明在音乐上，不同语种之间的歌唱在发音方法和习惯上有很大不同，至少在 MFCC 特征上有着比较大的区别。另外实验对于一些说唱、摇滚等一些音乐性不是很强的歌曲上有比较好的区分效果，对于一些浅吟低唱的音乐性较强的区分比较差，这些歌曲人声的和音结构和伴奏的音乐的和音结构非常相似，所以一定程度上分类结果可以作为歌曲音乐性的一个近似度量。

5 结论

本文将 SVM 方法应用到流行音乐中人声的识别研究中，其中采用 MFCC 特征，在后期对分类结果采用低通滤波，从而在帧的基础上识别率达到了 85.76%。MFCC 作为语音处理中常用的特征，在本问题中同样给出了比较好的分类结果。从实验结果来看，SVM 相对于其他分类器包括 ANN、GMM 和 HMM 具有比较好的泛化能力。对初始分类结果进行低通滤波能够使识别率平均提高 11.9%。从整体来看，基于帧的识别率达到 85.67%对于这个问题来讲与优秀工作相当。未来考虑从歌唱声音产生以及人耳对音乐以及歌唱感知的机理角度，以及考虑音乐的节奏来更合理地给音乐分帧以及提取特征。由于音乐节奏的发现也是一个很困难的问题，所以未来还有很多问题需要解决。另外利用流行歌曲的人声发现可以进行 intro、bridge 和 outro 的发现，从而有利于未来进行音乐的结构分析。

参考文献：

[1] Goto M.A real-time music scene-description system:predominant-F0 estimation for detecting melody and bass lines in real-world audio signals[J].Speech Communication,2004,43(4):311-329.

[2] Maddage N C,Xu C,Wang Y.Singing voice detection using twice-iterated composite fourier transform[C]//ICME,2004.

[3] Fujihara H,Kitahara T,Goto M,et al.F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search[C]//Proc IEEE Int Conf Acoustics,Speech, and Signal Processing,2006,5:253-256.

[4] Nwe T L,Li H.Exploring vibrato-motivated acoustic features for singer identification[J].IEEE Transactions,Audio,Speech and Language Processing,2007,15(2).

[5] 韩纪庆.语音信号处理[M].北京:清华大学出版社,2004.

[6] Duda R O,Hart P E,Stork D G.模式分类[M].2 版.李宏东，姚天翔，译.北京:机械工业出版社，中信出版社，2003:211-216.

[7] Richard G L.数字信号处理[M].2 版.朱光明，程建远，刘保童，译.北京:机械工业出版社，2006.