基于隐马尔可夫模型的音频自动分类*

卢 坚, 陈毅松, 孙正兴, 张福炎

(南京大学 计算机科学与技术系,江苏 南京 210093);
(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)
E-mail: jlu@graphics.nju.edu.cn
http://www.nju.edu.cn

摘要: 音频的自动分类,尤其是语音和音乐的分类,是提取音频结构和内容语义的重要手段之一,它在基于内容的音频检索、视频的检索和摘要以及语音文档检索等领域都有重大的应用价值.由于隐马尔可夫模型能够很好地刻画音频信号的时间统计特性,因此,提出一种基于隐马尔可夫模型的音频分类算法,用于语音、音乐以及它们的混合声音的分类.实验结果表明,隐马尔可夫模型的音频分类性能较好,最优分类精度达到 90.28%.

关 键 词: 基于内容的音频分类;隐马尔可夫模型;向量量化;MFCC(mel-frequency cepstral coefficient)
中图法分类号: TP391 文献标识码: A

音频压缩和 Internet 媒体流(media streaming)技术的发展,推动着各种基于 Internet 的音频应用逐步走向实用.但是,由于原始音频数据除了含有采样频率、量化精度、编码方法等有限的注册信息外,本身仅仅是一种非语义符号表示和非结构化的二进制流,缺乏内容语义的描述和结构化的组织,因而音频的检索和内容过滤等应用都受到极大的限制.如何提取音频中的结构化信息和内容语义,使得无序的音频数据变得有序,是基于内容的音频检索技术能否得以实用的关键所在.

音频自动分类的早期研究工作以文献[1,2]为代表.文献[1]训练一种神经元网络直接将声音类别映射到所标注的文本.文献[2]使用自组织映射(self-organizing mapping,简称 SOM)聚类算法对具有相似感觉特征的声音进行聚类.真正意义上的基于内容的音频自动分类工作是由美国 Muscle Fish 公司 Erling Wold 等人完成的[3],他们详细分析了音频的区别性特征,包括响度(loudness)、音调(pitch)、亮度(brightness)、谐度(harmonicity)等,并且根据最近邻准则(nearest neighbor,简称 NN)和 Mahalanobis 距离设计音频的分类器,所用的数据集包括笑声、铃声、电话声等 16 类共 409 个样本数据.在文献[3]提供的 Muscle Fish 数据集上,文献[4~6]采用不同的特征和分类器实现音频的分类.其中,文献[4]采用 12 阶的 MFCC 系数和能量作为音频的特征表示,根据极大互信息准则(maximum mutual information,简称 MMI)训练决策树量化特征空间为离散的区域,并且根据最近邻准则对音频作分类,文献[5,6]分别采用最近特征线(nearest feature line,简称 NFL)和支持向量机(support vector machine,简称 SVM)作为分类器.

近年来,音频的自动分类在视频的检索和摘要、基于内容的语音检索等相关领域也日益引起了人们的重视.在视频的检索和摘要中,人们发现简单的视觉特征,例如颜色、纹理、运动向量等并不能很好地反映视频的内容和结构语义,而更高级的视觉语义特征的提取则相当困难,因此,文献[7~9]尝试在视频的检索和摘要中结合音频(语音、音乐)、文本(字幕、标题)等信息,以克服单纯的视觉特征语义表达能力较弱这一缺点.文献[10,11]

* 收稿日期: 2001-02-13; 修改日期: 2001-05-22
基金项目: 国家自然科学基金资助项目(69903006,60073030)
作者简介: 卢坚(1974－),男,浙江东阳人,博士,主要研究领域为音频的分割,分类和检索;陈毅松(1973－),男,四川资阳人,博士,主要研究领域为图像压缩;孙正兴(1964－),男,江苏苏州人,博士,副教授,主要研究领域为 CAD/CAM,数字图书馆;张福炎(1939－),男,浙江绍兴人,教授,博士生导师,主要研究领域为多媒体技术,数字图书馆.

© 中国科学院软件研究所 http://www.jos.org.cn

根据音频特征分别训练 OCON(one-class-in-one-network)神经元网络和隐马尔可夫模型(hidden Markov model,简称 HMM)对电视节目作 5 种视频场景的分类:天气预报、新闻、广告、足球和篮球.文献[12]采用相位补偿 gamma 滤波器组提取音频特征,并用于音频的分割、音乐内容的分析、暴力镜头的检测等方面.基于内容的语音检索机制有关键词发现(keyword spotting)、子词格(sub-word lattice)索引和大词汇量连续语音识别 3 种形式[13],但是它们的开销都比较大,而且检索效果依赖于具体的语音环境,如果首先经过音频的自动分类以确定所处的语音环境,则可以提高识别精度和效率.

语音和音乐是两类最重要的音频信息.文献[7,14~16]采用基于简单决策树的语音/音乐多步层次分类方法,即每一步根据一种或者几种音频特征及其阈值判定音频所属的类别.但是,层次分类模型只能表示均值、方差等统计特性,而音频信号特征通常具有时间统计特性,例如,音乐中一般都存在揭示主题的韵律或者鼓点,在语音中,清音和浊音往往交替出现.隐马尔可夫模型本质上是一种双随机过程的有限状态自动机,它具有刻画信号的时间统计特性的能力[17].因此,本文提出一种基于 HMM 的音频分类算法,分类对象是语音(speech)、音乐(music)以及语音和音乐的混合(speech+music)共 3 类数据,并根据极大似然准则判定它们的类别.

1 音频特征提取方法

音频首先分割为 580ms(12 800 个采样)的 clip,相邻 clip 间有 290ms(6 400 个采样)的重叠部分,对每一 clip 加 23ms(512 个采样)的 Hamming 窗形成帧,相邻帧间有 11.5ms(256 个采样)的重叠部分,最后计算每一帧的傅立叶变换系数 $F(w)$和频域能量: $E=\int_0^{\overline{w}}|F(w)|^2\mathrm{d}w$,其中 $\overline{w}=fs/2$, fs为采样频率,如果某一帧的频域能量 E 小于阈值,则将该帧标记为静音帧,否则为非静音帧.根据非静音帧计算以下基于 clip 的音频特征:

(1) 静音比例(silence ratio):语音中一般都会存在停顿的部分,故其静音比例比较高;而在音乐中静音比例相当低.其定义为:静音比例=clip 中静音帧的数目/clip 中帧的总数.

(2) 子带能量比(sub-band energy ratio)均值:频域划分为 4 个子带区间 $sb_i(i=0...3)$:分别为 $[0,\overline{w}/16]$, $[\overline{w}/16,\overline{w}/8]$, $[\overline{w}/8,\overline{w}/4]$ 和 $[\overline{w}/4,\overline{w}]$,并计算各子带能量 SW_0,SW_1,SW_2,SW_3, $SW_i=\int_{w\in sb_i}|F(w)|^2\mathrm{d}w$,子带能量比定义为 $SWR_i=SW_i/E$.即各子带能量与频域总能量的比值.

(3) 带宽(bandwidth)均值:Clip 中各帧带宽的均值.带宽是衡量音频频域范围的指标,其定义为 $BW=\sqrt{\int_0^{\overline{w}}(w-FC)^2|F(w)|^2\mathrm{d}w\Big/E}$,其中 FC(frequency centroid)为频率中心,它是度量声音亮度(brightness)的指标,其定义为 $FC=\int_0^{\overline{w}}w|F(w)|^2\mathrm{d}w\Big/E$.一般地,语音的带宽范围在 300HZ~3.4KHZ 左右,而音乐的带宽范围比较宽,可以在 22.05kHZ 左右.

(4) 基音频率标准方差:基音频率是衡量音调高低的单位.本文中基音频率的检测采用中心削波(系数为 0.68)短时自相关函数波峰检测算法,当波峰大于某一阈值时,返回基音频率;否则返回 0.

(5) 谐成分比例:某一 clip 中基音频率不等于 0 的帧数所占的比例.

(6) 平滑基音比(smooth pitch ratio):若第 i 帧的基音频率不等于 0,并且其与第 i–1 帧的基音频率差值小于一定的阈值,则第 i 帧为基音平滑帧,某一 clip 中平滑帧的数目与其中基音频率大于 0 的总帧数之比为平滑基音比.

(7) MFCC+ΔMFCC 均值:MFCC(mel-frequency cepstral coefficient)即基于 Mel 频率的倒谱系数,它一般采用三角滤波器组对傅立叶变换能量系数滤波,并对其频域进行 Mel 比例变换,以更符合人类的听觉特性[18].虽然 MFCC 系数最早是为语音识别或者说话者的识别应用而提出的,但是,文献[4~6]的研究结果均表明,MFCC 系数可以用作音频分类特征,并且文献[5]的结果表明 MFCC 系数可以提高音频分类的精度.在本文的实验中分别计算 8,12,16,24,64 和 80 阶的 MFCC 及其差分系数ΔMFCC,ΔMFCC 可以比较好地反映 MFCC 的动态变化特性.

由此构造两类音频特征集合:感觉特征集(PercMFCC0)和感觉特征加 MFCC 集(PercMFCCL,L 为 MFCC 系数的阶数).感觉特征(Perc)为上述(1)~(6)类共 10 维特征,MFCC+ΔMFCC 特征的维数为 2*L.特征集合需要进行归一化处理,Perc 特征归一化处理如下: $\overline{X}_i=(X_i-\mu)/\delta$,其中$\mu$为均值,$\delta$为方差,PercMFCCL 集最终可以构造为

© 中国科学院软件研究所 http://www.jos.org.cn

$PercMFCCL = \frac{Perc}{s_1} \oplus \frac{MFCC}{s_2}$,其中 s_1 为 Percc 特征的总体方差,s_1=10*1,s_2 为 MFCC 系数的总体方差,$s_2 = \sum_{i=1}^{2L} \delta_i$,$L$ 为 MFCC 系数的阶数,δ_i 为第 i 维 MFCC 系数的方差.

2 隐马尔可夫模型理论和分类器设计

HMM 本质上是一种双重随机过程有限状态自动机(stochastic finite-state automata),其中的双重随机过程是指满足 Markov 分布的状态转换 Markov 链以及每一状态的观察输出概率密度函数,共两个随机过程.HMM 可以用 3 元组来表示: $\lambda = (A, B, \pi)$,其中 A 是状态 S_i 到 S_j 的转换概率矩阵,B 是状态的观察输出概率密度,π是状态的初始分布概率.HMM 需要研究的 3 个基本问题是:(1) 已知 HMM 模型λ的各参数,求某一观察序列 O 在该模型下的极大似然,即 $P(O \mid \lambda)$,$O=O_1 \ldots O_T$,T 为观察序列长度;(2) 在给定的 HMM 模型λ的条件下,求观察序列 O 最有可能历经的状态序列 S;(3) 在已知样本集合的条件下,如何根据样本集合训练模型并获得模型参数.问题(1)可以由前向(forward)或者后向(backward)算法解决,问题(2)是典型的状态空间搜索问题,经典的算法有基于动态规划的 Viterbi 算法、Beam Search 和 A*算法,问题(3)是统计学习过程,其学习算法有 Baum-Welch 算法、梯度算法等.Baum-Welch 算法能够在理论上证明经过有限次迭代就能收敛,但它和梯度算法一样都会陷入局部极值点,而不能得到全局最优的结果.

本文为 Speech,Music 和 Speech+Music 共 3 类数据分别训练它们各自的 left-right DHMM(discrete HMM),记为$\lambda_1,\lambda_2,\lambda_3$.Left-Right DHMM 具有计算代价小、迭代次数少和训练过程中收敛较快的优点,比较适合在线的音频分类应用.在训练 DHMM 分类器之前首先需要对样本数据进行向量量化,本文采用的量化算法是 K-Means 算法.在实验中,模型的参数 A,B 和 π的初值都是随机生成的,训练算法采用多观察序列的 Baum-Welch 算法,并且对α,β,ξ分别进行定标处理,以解决计算过程中的浮点溢出问题.分类的准则是极大似然判别,即给定一观察序列 O,分别计算 $P(O|\lambda_i)(i=1,2,3)$,并选取似然最大的模型为观察序列 O 的类别,即 $j = \underset{i}{\operatorname{argmax}}(p(o \mid \lambda_i))$.分类器结构如图 1 所示.

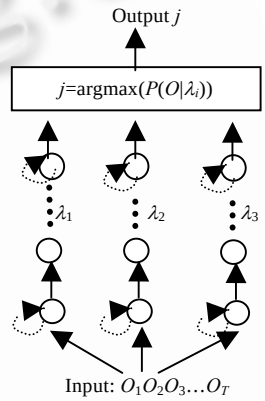

Fig.1 Classifier structure
图 1 分类器结构示意图

3 实验结果分析

音频数据来源于中央人民广播电台、南京文艺台、江苏音乐台、上海音乐台等多家广播电台的节目,内容包括新闻、人物访谈、娱乐、广告以及交响乐和中国名乐等.音频数据的采样频率是 22.050kHZ,精度为 16 位.数据总量为 350MB,总共 2 小时 20 分钟,其中 Speech 为 33.7 分钟,Music 为 72 分钟,Speech+Music 为 34.3 分钟,特征提取后共得样本序列总数为 998,其中 Speech:248,Music:497,Speech+Music:253.

由于不能在理论上确定最优的 HMM 状态和观察符号(即量化精度)的数目,本文比较了状态和观察符号数目的多种可能组合的 HMM,其中状态数目分别为 3,4 和 5,观察符号数目分别为 16,24 和 32,而对于 PercMFCC0 数据集,为进一步分析量化精度对分类性能的影响,还额外考察了量化精度为 48,64 和 72 时的分类结果.测试方法为留一交叉检验(leave-one-out cross validation),即假设样本总容量为 L,轮流将 L–1 份数据用作训练样本,而将剩下的 1 份作为测试样本.分类的准确度采用以下指标来衡量:各类别 C_i 的分类精度(precision)和平均分类精度.为与文献[15,16]保持一致,其定义如下:

- C_i 的分类精度=预测为 C_i 并且真实为 C_i 的数目/预测为 C_i 的数目;
- 平均分类精度=Σ_{Ci} 预测为 C_i 并且真实为 C_i 的数目/测试样本总容量.

通过分析表 1 中的实验结果数据,可以得出以下结论:

(1) 对于 PercMFCCL 特征集,分类精度随着 MFCC 阶数 L 的增大而相应地提高.经分析,其原因可能是随着阶数的增高,MFCC 系数能够更好地刻画音频中的谐度特性,而谐度是区分音乐和语音的重要特征,因此分类精

© 中国科学院软件研究所 http://www.jos.org.cn

度会有所提高.

(2) 随着量化精度的提高,分类精度也随之有所提高,这与预期的结果相符.可以想见,在训练样本充分的前提下,如果将本文采用的 DHMM 分类器改为 CHMM 分类器,分类精度应该还会有所提高.

(3) 文献[7]中的音频分类的目的是用于视频镜头的聚类,所以其中没有给出音频分类的结果.文献[15,16]只对纯语音和音乐进行分类,其中文献[15]的分类精度为 Speech:81%,Music:70%,平均:75%,而文献[16]的分类精度为 Speech:75%,Music:89%,平均:82%.文献[14]将音频分为 Speech,Music 和环境声音共 3 类,并根据谐度对环境声音作进一步分类,其分类精度为 90%左右.本文的最优分类结果是 Music:95.96%,Speech:88.01%和 Speech+Music:81.03%,而平均分类精度为 90.28%.实验结果表明,本文的算法分类精度优于文献[15,16],而与文献[14]相当,但是,文献[7,14~16]的层次式分类算法都需要选取阈值,而在本文的算法中不需要阈值的设置,从而可以避免阈值选取对结果的影响.

(4) 从表 1 可知,Music 的分类精度在 88%以上,Speech 的分类精度在 84%~92%,两者基本一致,而 Speech+Music 的分类精度在 80%以下,与 Music 和 Speech 的分类精度的差距很大,这使得平均分类精度急剧下降.经分析,本文实验中的 Speech+Music 的样本数据集中存在相当多的类似情况,即某一类的特征(如 Speech)非常强,而其他特征(如 Music)因相对较弱而几乎被完全掩盖.

Table 1 Experimental results

表 1 实验结果数据

Feature set①	Code/state②	Precision③			
		Music	Speech	Speech+Music	Average
PercMFCC0	16/5	0.913 481	0.866 935	0.837 945	0.882 766
	24/3	0.931 590	0.854 839	0.810 277	0.881 764
	32/5	0.951 710	0.895 161	0.790 514	0.896 794
	48/3	0.939 638	0.879 032	0.822 134	0.894 790
	64/4	0.955 734	0.883 065	0.794 466	0.896 794
	72/4	0.959 759	0.883 065	0.810 277	**0.902 806**
PercMFCC8	16/4	0.905 433	0.842 742	0.664 032	0.828 657
	24/5	0.908 722	0.853 361	0.653 323	0.834 657
	32/4	0.913 481	0.887 097	0.679 842	**0.847 695**
PercMFCC12	16/5	0.899 396	0.838 710	0.703 557	0.834 669
	24/4	0.891 348	0.903 226	0.735 178	0.854 709
	32/5	0.887 324	0.919 355	0.790 514	**0.870 741**
PercMFCC16	16/5	0.925 553	0.850 806	0.715 415	0.853 707
	24/5	0.903 421	0.895 161	0.731 225	0.857 715
	32/3	0.907 445	0.943 548	0.794 466	**0.887 776**
PercMFCC24	16/5	0.933 602	0.907 258	0.758 893	0.882 766
	24/4	0.921 529	0.911 290	0.798 419	0.887 776
	32/4	0.917 505	0.907 258	0.810 277	**0.887 776**
PercMFCC64	16/4	0.931 590	0.903 226	0.786 561	0.887 776
	24/4	0.915 493	0.947 581	0.770 751	0.886 774
	32/5	0.945 674	0.919 355	0.786 561	**0.898 798**
PercMFCC80	16/4	0.931 590	0.911 290	0.774 704	0.886 774
	24/3	0.945 674	0.927 419	0.762 846	0.894 790
	32/4	0.947 686	0.931 452	0.778 656	**0.900 802**

①特征集,②码本/状态数目,③分类精度.

4 总 结

语音和音乐是最重要的两类音频信息,语音、音乐等音频的自动分类在基于内容的音频检索、视频的检索和摘要以及语音文档的检索等领域都有重要的应用价值.本文分析了语音和音乐的区别性特征,例如,音调、音色、谐度等感觉特征以及 MFCC 系数等.HMM 可以表示音频特征的时间统计特性,从而能够揭示不同类型音频的时间统计特性.据此,本文提出一种 left-right DHMM 的分类器,用于 Speech,Music 以及 Speech+Music 这三者的分类.实验结果表明,隐马尔可夫模型对于音频的分类是有效的.

在音乐中,同一旋律或者鼓点经常会贯穿整段乐曲,因此有可能会出现状态反复的情况,left-right 结构的 HMM 由于状态转换的顺序特性而不能很好地表示音乐中的状态反复,同时,向量量化会引入量化误差,这两者都会在一定程度上影响分类的精度.另外,Baum-Welch 算法是基于极大似然准则的,其判别能力比较弱,将来可

© 中国科学院软件研究所 http://www.jos.org.cn

以考虑采用其他准则的学习算法,例如,基于极大互信息(maximum mutual information,简称 MMI)或者最小分类错误(minimum classification error,简称 MCE)准则的学习算法,以提高分类的精度.

References:

[1] Feiten, B., Frank, R., Ungvary, T. Organization of sounds with neural nets. In: Proceedings of the 1991 International Computer Music Conference, International Computer Music Association. San Francisco, 1991. 441~444.

[2] Feiten, B., Günzel, S. Automatic indexing of a sound database using self-organizing neural nets. Computer Music Journal, 1994,18(3):53~65.

[3] Wold, E., Blum, T., Keislar, D., *et al*. Content-Based classification, search and retrieval of audio. IEEE Multimedia Magazine, 1996,3(3):27~36.

[4] Foote, J.T. Content-Based retrieval of music and audio. Multimedia Storage and Archiving Systems II, 1997,32(29):138~147.

[5] Li, S.Z. Content-Based classification and retrieval of audio using the nearest feature line method. IEEE Transactions on Speech and Audio Processing, 2000,8(5):619~625.

[6] Li, S.Z., Guo, Guo-dong. Content-Based audio classification and retrieval using SVM learning. In: Proceedings of the 1st IEEE Pacific-Rim Conference on Multimedia. 2000.

[7] Jiang, Hao, Lin, Tony, Zhang, Hong-jiang. Video segmentation with the support of audio segmentation and classification. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2000), Vol 3. NY: IEEE, 2000. 1507~1510.

[8] He, Li-wei, Sanocki, E., Gupta, A., *et al*. Auto-Summarization of audio-video presentations. In: Proceedings of the 7th ACM International Conference on Multimedia. Orlando: ACM Press, 1999. 489~498.

[9] Patel, N., Sethi, I. Audio characterization for video indexing. In: Proceedings of the SPIE on Storage and Retrieval for Still Image and Video Databases, Vol 2670. 1996. 373~384.

[10] Liu, Zhu, Huang, J., Wang, Y. Classification of TV programs based on audio information using hidden Markov model. In: Proceedings of the IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing. IEEE, 1998. 27~32.

[11] Liu, Zhu, Wang, Y., Chen, T. Audio feature extraction and analysis for scene segmentation and classification. Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, 1998,20(1/2):61~79.

[12] Pfeiffer, S., Ficher, S., Effelsberg, W. Automatic audio content analysis. In: Proceedings of the 4th ACM International Conference on Multimedia. Boston, MA: ACM Press, 1996. 21~30.

[13] Foote, J.T. An overview of audio information retrieval. Multimedia Systems, 1999,7(1):2~10.

[14] Zhang, Tong, Kuo, C.C.J. Heuristic approach for generic audio data segmentation and annotation. In: Proceedings of the 7th ACM International Conference on Multimedia. Orlando: ACM Press, 1999. 67~76.

[15] Srinivasan, S., Petkovic, D., Ponceleon, D. Towards robust features for classifying audio in the cudevideo system. In: Proceedings of the 7th ACM International Conference on Multimedia. Orlando: ACM Press, 1999. 393~400.

[16] Lu, Guo-jun, Templar, H. A technique towards automatic audio classification and retrieval. In: Proceedings of the 4th International Conference on Signal Processing, ICSP, Vol 2. 1998. 1142~1145.

[17] Rabiner, L., Juang, B-H. Fundamentals of Speech Recognition. Prentice-Hall International, Inc., 1993.

[18] Vergin, R., O'Shaughnessay, D. Generalized mel-frequency cepstral coefficients for large-vocabulary speaker-independent continuous speech recognition. IEEE Transactions on Speech and Audio Processing, 1999,7(5):525~532.

Automatic Audio Classification by Using Hidden Markov Model*

LU Jian, CHEN Yi-song, SUN Zheng-xing, ZHANG Fu-yan

(*Department of Computer Science and Technology, Nanjing University, Nanjing* 210093, *China*);

(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*)

E-mail: jlu@graphics.nju.edu.cn

http://www.nju.edu.cn

Abstract: As one of the key methods to extract content semantics and structure from audio, automatic audio classification, especially for a speech and a music, is valuable for content-based audio retrieval, video summary and retrieval, and spoken document retrieval, etc. Because hidden Markov model (HMM) can well model audio signal's time statistical properties, a left-right discrete HMM is proposed to classify a speech, a music and their mixed audio. The experimental results show that HMM is excellent for audio classification, and the optimal classification accuracy is up to 90.28%.

Key words: content-based audio classification; hidden Markov model (HMM); vector quantisation; mel-frequency cepstral coefficients (MFCC)

* Received February 13, 2001; accepted May 22, 2001
Supported by the National Natural Science Foundation of China under Grant Nos.69903006, 60073030

© 中国科学院软件研究所 http://www.jos.org.cn