基于距离相关图的音频相似性度量方法

李 超 熊 璋 朱成军

(北京航空航天大学 计算机学院, 北京 100083)

摘 要: 相似性度量是基于内容的音频分析中的关键环节之一,目前往往沿用传统的统计分析思想实现,缺乏有效的呈现手段和对语义层信息的准确表达. 提出了一种基于图像分割技术的相似性度量方法,通过特征提取构造特征空间,进而绘制特征向量间的距离相关图,通过对该图的可视化分析判断最大相似方向,并分别计算出局部相似性和全局相似性. 给出了实验验证过程,该方法适用于数字广播等音频流中的片段查找等应用.

关 键 词: 音频; 相似性; 图像分割; 特征提取

中图分类号: TP 391

文献标识码: A **文 章 编 号**: 1001-5965(2006)02-0224-04

Audio similarity measure based on distance correlation image

Li Chao Xiong Zhang Zhu Chengjun

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: Similarity measure is one of the keys of content-based audio analysis. Traditionally, statistic methods are used for accomplishing the aim. But it is hard to represent the results in a visual way and it shows very weak relationship with semantic layer. An image segmentation-based method was introduced to partly solve the problem. Features were extracted respectively to form a so-called feature space. The distance correlation images were calculated by comparing the feature vectors. A maximal similar direction was estimated by segmentation on the image to calculate local similarity and global similarity. Experiment was also provided. The method was proved to be suitable for applications like clip searching in digital broadcasting stream.

Key words: audio; similarity; image segmentation; feature extraction

基于内容的多媒体分析是近年来引起研究者重点关注的方向,如基于内容的图像处理、基于内容的音频查询与检索等等[1]. 相似性度量是基于内容多媒体技术的关键步骤之一. 对声音相似性度量方法而言,其面临的一个主要困难是呈现手段的匮乏,这与声音的主观性特点有着密切的关系. 同时,基于帧为单位提取的单个特征向量并不能完全反映声音片段之间的相似性关系,因此,可靠的音频相似性度量方法必然应在一定语义分割的基础上进行.

1 相关研究

多媒体数据往往存在数据量大、时空相关度高和描述二义性等问题,其相似性度量带有模糊性的特点. 传统音频相似性度量方法多直接采用统计学中的研究成果,如基于欧氏距离、余弦距离、相关距离、马氏距离等等. 这些方法均是从统计分析角度来处理问题,因而缺乏媒体适应性,导致其结果往往在语义上难以解释. 图像是最直观的表现手段,声音的波形图、声谱图等都是图像技术同声音信号处理相结合的产物. Subramanya 等[2]直接针对音频样本的二值图像进行分割处理,这种方法简单直观,但由于采样率、量化位和时长等因素,在基本的采样率(8 kHz)和量化位(16 bit)情况下,一秒钟未处理图像大小即可达到 8 000 像素×65 536 像素,故在处理一般音频时这种方法并不具备实用性. Foote[3]也曾利用可视

收稿日期: 2005-01-17

作者简介: 李 超(1974-),男,四川乐山人,讲师, licc@buaa.edu.cn.

化方法对音频的时间结构进行分析,以寻找音乐中的自相似特点,如提取鼓点的节奏,发现旋律重复的特点. 首先提取 MFCC(Mel-Frecquency Cepstrum Coefficients)系数生成自相似图像,鼓点、节奏等信息可体现于自相似图像的纹理分布中. 作者则在此基础上进一步提出了一种基于特征提取和图像分割的声音相似性度量方法.

2　基本距离度量方法

特征距离(feature distance)是度量音频样本间相似性最直接的方法,它们一般能够反映出样本间的分离程度,常用的距离度量方法见表 1.

表 1　基本距离度量方法

名称	表达式
闵氏距离	$d(\boldsymbol{X},\boldsymbol{Y}) = \left(\sum_{i=1}^{n} \lvert x_i - y_i \rvert^{\lambda}\right)^{\frac{1}{\lambda}}$
马氏距离	$d(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{(\boldsymbol{X}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{X}-\boldsymbol{Y})}$
余弦距离	$d(\boldsymbol{X},\boldsymbol{Y}) = \dfrac{\sum_{i=1}^{n}\boldsymbol{X}(i)\boldsymbol{Y}(i)}{\sqrt{\sum_{i=1}^{n}\boldsymbol{X}(i)^2 \times \sum_{i=1}^{n}\boldsymbol{Y}(i)^2}}$
相关距离	$d(\boldsymbol{X},\boldsymbol{Y}) = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

注:表中 $\boldsymbol{C}$ 为协方差矩阵

闵可夫斯基距离是若干经典距离的通式,如棋盘格距离($\lambda=1$)、欧式距离($\lambda=2$)等. 欧式距离是特征空间距离度量的最自然的方法,也是在音频特征距离度量中应用最多的方法. 其特点在于自然、直观、易懂,而缺陷在于没有考虑方差的因素,尤其在进行多元分析时与期望往往存在较大的偏离;另一个缺陷是其距离与分量的单位有关,距离需经过规范化处理才能作为相似性适用.

余弦距离和相关距离则消除了分量单位的影响,并且其计算结果可以很方便地转换为相似性,如相关分析的输出甚至可直接用作相似性,相关又可分为自相关和互相关,本文生成的距离相关图实际就是一种互相关分析的结果.

马氏距离考虑了样本的统计特性,排除了样本间的相关影响. 其关键是协方差的计算. 尤其是当数据维较大时,协方差的计算相当耗时. 世界上第一个基于内容的音频检索查询商业软件 Muscle Fish[4] 中就使用马氏距离作为相似性度量方法. 微软研究院的 Lu 等在文献[5]中提出了一种用于音频特征距离度量的最近特征线方法,该方法实际是对解析几何中点到线最短距离的应用;而 Compaq 研究院的 Beth Logan 等[6]则将图像处理中的 EMD(Earth Mover Distance)方法应用于音乐相似性度量,该方法计算较复杂,要经过大量的协方差计算,并且仍然缺乏对语义理解的足够支持.

3　基于图像分割的相似性度量方法

基于图像分割的相似性度量方法包括预处理、特征提取、距离相关图生成等步骤,如图 1,并可分为前台处理和后台处理两部分,其中后台处理主要针对音频数据库和大型音频流,需事先进行大量特征的提取和生成特征检索库,这正是 MPEG-7 指导框架建议和目前大多数音频检索原型系统所乐于采用的.

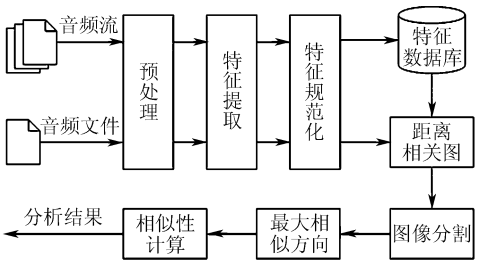


图 1　基于图像分割的相似性度量方法

3.1　音频预处理

预处理包括两个方面:①将来源不同的音频样本处理为具有相同的输入采样率和量化位数,保证特征提取在相同的基础上进行,如统一预处理为 44.1 kHz 采样率和 16 bit 量化 PCM(Pulse Code Modulation)音频. ②音频流分割,其目的在于两点,用较小的片段降低距离相关图的尺寸,同时通过增加标注信息为进一步的处理提供语义参考,具体方法另文发表.

3.2　音频特征提取

基于分割结果提取特征系数,可提取时域特征、频域特征、心理声学特征以及倒谱系数等,假设每帧提取 k 个特征系数,则对参与比较的两个片段分别得到如下的特征向量集合:

$$\left.\begin{aligned}\boldsymbol{X} &= \{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\cdots,\boldsymbol{X}_m\}\\ \boldsymbol{Y} &= \{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\boldsymbol{Y}_3,\cdots,\boldsymbol{Y}_n\}\end{aligned}\right\} \tag{1}$$

其中 $\boldsymbol{X},\boldsymbol{Y}$ 分别表示两个待比较的特征集合,$\boldsymbol{X}_i$、$\boldsymbol{Y}_j$ 为 k 维特征向量($i=1,\cdots,m,j=1,\cdots,n$).

对于语音类应用,提取特征以 MFCC 等倒谱系数为主;对音乐类应用,以带宽、子带能量比、基频和谐波分析系数等物理特征或心理声学特征为主. 本文取材主要来自机场广播语音库,因此采用的特征为 MFCC 系数(2~13)及其一阶导数.

3.3　特征规范化

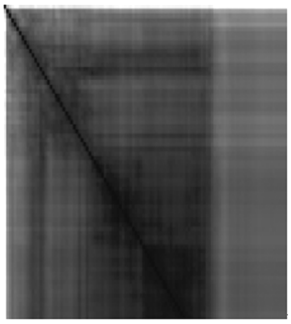
对提取的特征向量集合进行整体规范化,具

体包括设置特征分量屏蔽和规范化处理两个步骤,被屏蔽的特征维在距离计算中按0对待. 而对每个参与计算的特征维分别计算其整体均值μ和标准差σ,规范化方法见式2.

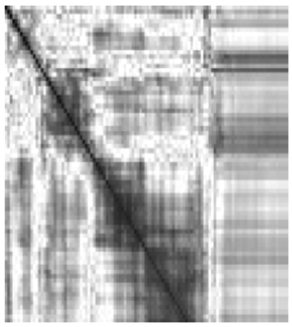
$$\tilde{X}_i = \frac{X_i - \mu}{\sigma} \tag{2}$$

3.4 距离相关图(**distance correlation image**)

针对进行比较的两组特征,计算特征距离并生成距离相关图$G_{m\times n}$,为方便起见,一般可以较长的特征集合作为横坐标方向,较短的作为纵坐标方向,如图2所示,但不限于此. 在常用的各种度量方法中,马氏距离由于协方差计算量颇大,耗时较多,因此文中主要采用欧氏距离和余弦距离作为度量基础.



a 欧氏距离



b 余弦距离

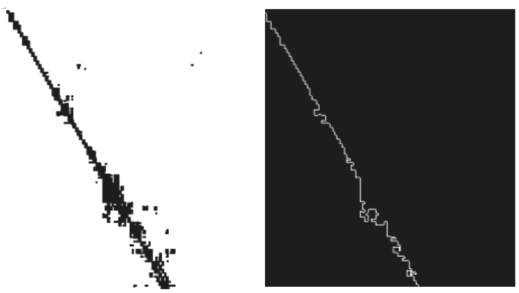
图2 对中文"一"和"一号"得到的距离相关图

3.5 最大相似方向提取

寻找具有最大相似性的分布方向$\boldsymbol{D}$,由于相关图的生成特点,若以图像左下角为原点,则该方向应该在135°. 如图2在黑线分布的方向可取得最大值1.0. 该例表示样本完全相同或部分相同的情况下的相似距离分布,即当样本中存在部分相似时距离相关图中会在某个范围内出现类似的黑线,黑线的长短说明相似片段的长度,而黑线的灰度说明相似的程度,其起止位置和方向则说明了相似片段在样本中的具体位置,具体还可采用Hough变换等方法寻找. 更多的情况下,由于窗口选择问题和样本采集问题,黑线模糊为一个或多个灰度较大的区域,此时可对其采用扩展邻域的方法(如5×5或3×3)进行搜索,可对其中黑色点包括中心的灰色点进行加权求和处理,从计算量考虑,论文采用3×3邻域.

3.6 计算分片相似性

从对图2的分析可知,余弦距离生成距离相关图中,其最大相似方向与周围部分具有较高的对比度,这点有利于进一步的边缘提取和分割,因此以下示例均以余弦距离相关图为对象,具体处理包括二值化、腐蚀和边缘提取和直线拟合等四个步骤,即寻找一条斜率为-1的直线,使得图3b中所有非0点到该直线的距离为最短,称该直线为最大相似方向$\boldsymbol{D}$.



a 经过二值化和腐蚀 b 边缘提取

图3 对中文"一"和"一号"相似距离相关图的处理

在最大相似方向上计算相似性,首先将相似线段投影于x、y轴中较长的方向,设长度为l. 则分片相似性可采用如下两种方法计算.

1) 对最大相似方向$\boldsymbol{D}$上所有像素点按其灰度值G_D求均值;

$$P_g = \frac{1}{l}\sum G_D \tag{3}$$

2) 考虑在最大相似方向上像素点某邻域内(如$k\times k$)的所有像素点,可按最大值规则、最小值规则或均值规则执行,即先对每个像素点按式(4)取其一求融合灰度值,再按式(3)计算;

$$\bar{G}_D = \begin{cases} \dfrac{1}{k^2}\sum G_D^{k\times k} & \text{均值规则} \\ \min\{G_D^{k\times k}\} & \text{最小值规则} \\ \max\{G_D^{k\times k}\} & \text{最大值规则} \end{cases} \tag{4}$$

其中,$\bar{G}_D$表示最大相似性方向上某点按$k\times k$邻域在不同规则下计算所得的灰度值,$G_D^{k\times k}$表该点$k\times k$邻域内所有像素点的灰度值集合.

则最终的局部相似性和整体相似性可计算为

① 局部相似性

$$S(\boldsymbol{X},\boldsymbol{Y}) = \frac{P_g}{255} \tag{5}$$

② 整体相似性

$$S(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{255}\frac{l \cdot P_g}{\max(m,n)} \tag{6}$$

其中 m,n 分别表示图像的宽度与高度.

4　实验结果

对图 2 示例采用图像分割法(包括无邻域和 3×3 邻域法)进行相似性计算,同时对 3×3 邻域的情况分别考虑最大值规则、最小值规则和融合规则,并与采用欧拉距离、余弦距离和相关距离结合类间连接规则的相似性计算结果进行比较,如表 2,其中主观分析为根据本例中样本的生成方法的主观判断结果,作为语义验证基础. 显然,根据分析结果,本文所提出的相似性计算方法比直接距离方法偏差更小,且更符合语义理解.

表 2　根据图 2 的相似性计算比较

项目	整体相似性		局部相似性	
	相似性	偏差	相似性	偏差
主观分析	0.640	0.000	1.000	0.000
直接欧拉距离	0.781	0.141	0.802	0.198
直接余弦距离	0.441	0.199	0.510	0.490
直接相关距离	0.527	0.113	0.558	0.442
图像分割法无邻域	0.604	0.036	0.933	0.067
图像分割法最大值规则(3×3 邻域)	0.623	0.017	0.976	0.024
图像分割法最小值规则(3×3 邻域)	0.442	0.198	0.691	0.301
图像分割法融合规则(3×3 邻域)	0.540	0.100	0.844	0.156

通过对更多样本(500 段共 20 min 广播语音流)和多种实验方法(更多的距离计算方法和邻域规则)的综合分析还可得出以下结论:

1) 考虑邻域(3×3)的最大值规则在多数样本中具有最好的相似性判断表现,而考虑邻域过大(7×7),一方面引入太多的计算耗费,同时模糊的区域划分也导致计算结果的可信度降低;

2) 在各种距离度量方法中,余弦距离虽然直接用于计算相似性效果较差,但以此生成的距离相关图分割效果却可达最好,因其所生成图像中相似区域与其它区域之间对比度较大所致,这一点正有利于二值化过程中阈值的自适应设定;

3) 欧拉距离在几种直接距离度量方法中具有相对较高的稳定性,与之有密切关系的闵科夫斯基距离和棋盘格距离也有着较好的表现.

5　结 束 语

基于图像分割的相似性度量方法弥补了针对特征向量直接进行距离计算难以避免的分割界限选择困难等缺陷,并且滤除了与比较无关的对整体相似性计算的干扰. 此外,从最大相似方向线段的长度和位置可以确定相似样本在音频流中的具体位置,这在进行音频流片段搜索时非常有效,如客户提交一个音频片段,在数据库或一个较大的音频流中进行搜索,因此该法可用于广播音频的片段查找和检索、音频文件的相似性匹配、基于示例的音频检索[7] 等用例,可大大提高查找的速度,前提是需如 MPEG-7 所要求,事先对音频流进行特征集的提取,建立特征索引库.

参考文献 (References)

[1] 龚声蓉. 基于内容的图像检索方法研究[D]. 北京:北京航空航天大学计算机科学与工程系,2001
Gong Shengrong. Research for content-based retrieval methods [D]. Beijing:Dept of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, 2001 (in Chinese)

[2] Subramanya S, Abdou Y. Segmentation of audio data based on the binary images of the audio samples[A]. In: Proc of Inter Conference on Intelligent Systems[C]. Denver: ISCA, 1999

[3] Foote J. Automatic audio segmentation using a measure of audio novelty[A]. In: Proc of ICME 2000[C]. NY: IEEE, 2000. 452 ~ 455

[4] Erling W, Thom B, Douglas K, *et al*. Content-based classification, search, and retrieval of audio[J]. IEEE Multimedia, 1996, 3(3):27 ~ 36

[5] Lu L, Jiang H, Zhang H J. A robust audio classification and segmentation method[A]. In: Proc of the 9th ACM International Conference on Multimedia[C]. Ottawa: ACM, 2001. 203 ~ 211

[6] Beth L, Ariel S. A content-based music similarity function[R]. CRL2001/2, 2001

[7] Liu Z, Huang Q. Content-based indexing and retrieval-by-example in audio[A]. In: Proc of IEEE Inter Conf on Multimedia and Expo[C]. NY: IEEE, 2000. 877 ~ 880