第37卷 第19期 计 算 机 工 程 2011年10月
Vol.37 No.19 Computer Engineering October 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)19—0160—03 文献标识码：A 中图分类号：TP18

基于K-L距离的两步固定音频检索方法

齐晓倩，陈鸿昶，黄 海

(解放军信息工程大学信息工程学院，郑州 450002)

摘 要：根据音频文件数据量大、数据间存在一定相关性的特点，提出一种基于K-L距离的两步固定音频检索方法。该方法采用基于可变门限的直方图检索方法快速筛选出相似度较高的语音文件，利用特征矩阵的K-L距离对剩余语音进行精确比较，取得较好的效果。实验结果证明，该方法能使检索准确率达到90%左右。

关键词：固定音频检索；过零率；直方图；美尔频率倒谱系数；K-L距离

Two-stage Specific Audio Retrieval Method Based on K-L Distance

QI Xiao-qian, CHEN Hong-chang, HUANG Hai

(School of Information Engineering, PLA Information Engineering University, Zhengzhou 450002, China)

【Abstract】Due to the huge amount of audio data, and some relation among them, this paper proposes a two-stage specific audio retrieval method based on K-L Distance. The method uses histogram retrieval method based on variable threshold to choose audio file of high similarity, compares precisely with residual audio using K-L distance of feature matrix, and obtains good effect. Experimental results show that the retrieval accuracy is over 90%.

【Key words】specific audio retrieval; Zero Crossing Rate(ZCR); histogram; Mel Frequency Cepstral Coefficient(MFCC); K-L distance

DOI: 10.3969/j.issn.1000-3428.2011.19.052

1 概述

虽然随着通信技术的快速发展，信息交流的形式越来越多样，但语音交流仍是人们沟通的主要方式。由于业务分析的需要和存储技术等的进步，人们经常需要对海量音频数据进行处理。如今，传统的单纯依靠文字标注的音频检索方式早已不能满足需要。因此，如何在大量的音频数据快速准确地找到有用信息成为了人们日益关注的问题，音频检索技术也应运而生[1]。音频检索是指通过对音频的特征分析，人们能够从大规模的音频数据库中找到自己感兴趣的内容。目前，音频检索主要分为两大类：一类是基于内容的音频检索技术(Content-based Audio Information Retrieval)，该技术主要研究如何利用音频的幅度、频谱等物理特征，响度、音高、音色等听觉特征，词字、旋律等语义特征实现基于内容的音频信息检索，音频进行分类和识别[2-3]；另一类是基于特征相似度的固定音频检索(Specific Audio Information Retrieval)，它是指给定一个查询音频段，在待检音频库中检索与其相同或同源的片段[4-5]。固定音频检索的思想最早由日本的Kashino提出[6-7]，他设计的直方图计算方法至今仍是固定音频检索领域使用最主要的方法。与传统的逐帧提取特征逐帧比较的方法相比，直方图方法在检索速度上有着绝对的优势，但由于所使用的特征非常单一，因此检索精度较低。

本文提出一种基于K-L距离的两步固定音频检索方法，用直方图方法对数据库中数据进行预处理，从大量音频数据库中初步筛选出相似度较高的音频文件，利用可变门限进一步提高筛选速度，并标记相似度最高部分的数据位置，采用基于K-L距离的自相关矩阵相似度量,对数据进行精确检索，提高检索精度。

2 基于直方图的初步筛选

直方图检索法的主要思想是：首先提取音频信号的某一个或一组特征标量，对提取的特征标量量化后建立直方图，然后采用“直方图相交法”计算直方图的相似度，并根据事先确定好的门限对计算结果做出判决。其算法思想如图1所示。

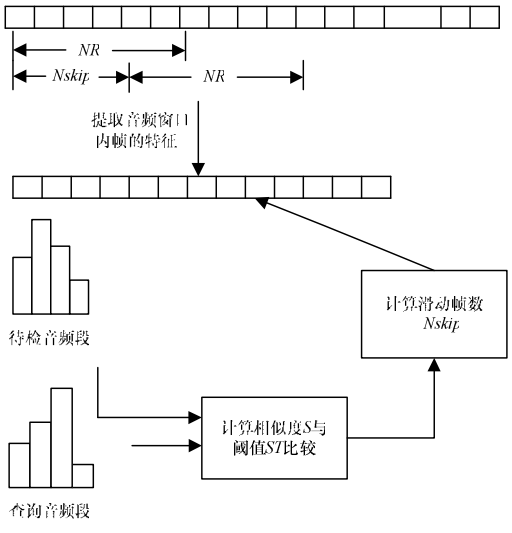

图1 直方图检索方法

分别用 $h^R=(h_1^R,h_2^R,\cdots,h_B^R)$ 以及 $h^T(n_{now})=(h_1^T,h_2^T,\cdots,h_B^T)$ 表示查询音频段和待检音频段的直方图，其中，n_{now} 为当前滑动窗口起始帧的位置；B 为直方图的柱数；h_i 代表第 i 个直方柱的频数，2个直方图的相似度定义为：

基金项目：国家“863”计划基金资助项目(2008AA011002)

作者简介：齐晓倩(1984－)，女，硕士，主研方向：固定音频检索；陈鸿昶，教授；黄 海，博士

收稿日期：2011-03-30 **E-mail：**kathy.qq@163.com

$$S(h^R,h^T(n_{now}))=\sum_{i=1}^{B}\min(h_i^R,h_i^T(n_{now})) \tag{1}$$

为了使算法在提取音频特征的步骤中不至于消耗过多时间，总是希望可以使用计算量小的音频特征。常用的 MFCC、PLP 等，虽然能较好地区分音频，但是计算复杂，不适用于音频数据库大量数据的初步筛选。Kedem 早在 20 世纪 80 年代末就已在其文章中指出，音频信号及其差分信号的过零率可以用来区分声音[8]，除此之外，过零率的计算量非常小，因此可以有效地提高检索效率。基于此，本文的研究和分析将采用过零率来建立直方图。i 阶差分信号的过零率定义为：

$$Z_i=\sum_{n=1}^{N}\frac{|\operatorname{sgn}(s_i(n))-\operatorname{sgn}(s_i(n-1))|}{2} \tag{2}$$

其中，N 代表抽样点个数；$s_i(n)=s_i(n)-s_i(n-1)$ 是第 i 阶差分信号。

假设查询音频段共有 N_R 帧，某个待检音频段共有 N_T 帧（$N_T>N_R$，否则不予处理）。首先，滑动窗口从音频库中某个待检音频文件的头部开始，取 N_R 帧，生成直方图进行对比，对比结束后滑动窗口向后滑动 N_{skip} 帧，直至该音频的结尾，然后对下一个音频文件再进行同样的动作，直至将库中的音频文件全部检索完毕。该算法检索速度之所以快，一方面是因为过零率和相似度计算的计算量都量非常小；另一个关键之处在于 N_{skip} 并不总是等于 1，直方图可以根据某一个位置的相似度，预测出之后若干位置的相似度上界，如果这些位置的相似度上界小于预设门限，则可以直接“跳过”，从而进一步提高检索速度。设 S_T 为相似度门限，则滑动窗口下一次应滑动的帧数 N_{skip} 为：

$$N_{skip}=\begin{cases}N_R[S_T-S(h^R,h^T(n_{now}))] & S<S_T\\ 1 & S\geqslant S_T\end{cases} \tag{3}$$

由此可见，S_T 越大，窗口滑动得越快，检索速度越快，因此 S_T 的设置对检索速度和精度的影响都非常大，选择合适的 S_T 十分重要。为了解决这个问题，本文采用了一种可变门限法。首先以检索准确度为标准为 S_T 设置一个合理值，对比过程中，一旦有某个位置 n 的直方图相似度 $S(h^R,h^T(n))>S_T$，则令 $S_T=S$，并标记 n，作为下一步精确检索部分所需数据的起始位置，因为在精确检索中，需要的仅是超过阈值的最多部分的数据，而不是整个文件的数据。由式(3)可知，S_T 的增大会带来 N_{skip} 的增大，这样检索速度就可以得到进一步的提高。

3　基于 K-L 距离的精确检索

过零率最开始主要用于对不同性质的音频进行分类，如进行语音、非语音和音乐的分类等[9]。因此，用过零率建立的直方图可以轻松地将不同的种类区分清楚。但是，若希望再进行细致的划分，或者当待检音频与查询音频都属于同一个种类(如语音)时，基于直方图检索的准确性能就会大大降低。究其原因，是因为过零率本身包含信息非常有限，而且建立直方图的过程仅是一个先标量量化后累加的过程，影响到了检索的效果。

为了进一步提高检索准确性，本文提出采用基于特征矩阵相似度的检索方法，对采用直方图筛选所得结果进行二次精确检索。

3.1　改进检索算法

概括来说，相似度的检索都是基于特征间距离的。因此，矩阵相似度的检索方法首先需要对数据库中的音频文件提取某些特征生成特征矩阵，这些选出的特征应该能够充分表示音频的重要区分特征，并具有一定的鲁棒性。生成特征矩阵后，再根据特征矩阵的特性，选择合适的距离度量方法计算和目标的距离，并最终判决是否满足阈值。

3.2　特征提取

查询音频需要提取整个音频文件的特征值作为比较的目标对象。假设音频例子一共分为 N 帧，每帧提取 $m(m<N)$ 个特征值，生成一个 $m\times N$ 的特征值矩阵；对于经直方图法筛选出的某个待测音频，并不需要考虑整个文件，而只需要提取超过阈值最多的滑动窗口内帧的特征值即可，结果也是一个 $m\times N$ 的矩阵，这样可以极大降低计算量。

在音频检索中，所选出的特征应该能够充分表示音频的重要区分特性，并具有一定的鲁棒性。在表述音频的多种特征中，Mel 倒谱系数是一种十分重要的特征参数，它采用一种非线性的 Mel 频率单位来模拟人耳的听觉系统，充分考虑到了人耳听觉的非线性特性，去除了因激励影响而引起的音频频谱峰值的波动，在音频检索中取得了非常好的效果[10]，是目前非常流行的音频特征参量。近年很多人利用改进后的 MFCC 对音频进行规整和处理也取得了理想的实验效果[11]。

根据人耳对低频声音感知能力强，在 1 KHz 以下约成线性关系，而对高频声音感知能力较弱，在 1 KHz 以上约成对数关系的非线性特性，Mel 频率值大致与实际频率成对数关系，Mel 频率与实际频率的转换关系如下：

$$\operatorname{Mel}(f)=2\,595\times\lg(1+f/700) \tag{4}$$

其中，f 为实际频率，单位为 Hz。

本文采样量化的步骤在第 1 步已经完成，故可以略去。由于人耳对音频的动态特性更为敏感，而标准的倒谱参数只反映了音频参数的静态特征，因此如果在音频特征中增加 Mel 差分倒谱，则理论上可以进一步提高检索准确度。Mel 倒谱系数提取流程如图 2 所示。

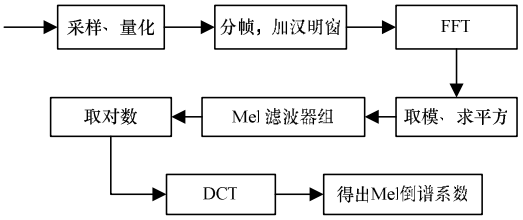

图 2　Mel 倒谱系数提取流程

3.3　精确检索方法

特征矩阵的相似性度量是语音精确检索的关键过程。人们往往利用特征矩阵的统计值，如均值、频的重要区分特征，并具有一定的鲁棒性。在描述音频的多种特征中，Mel 倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)是一种十分重要的特征参数，它采用了一种非线性的 Mel 频方差等，生成特征向量以表征音频文件，然后利用欧式距离或其他距离衡量查询音频和某个待测音频间的距离，从而根据距离阈值判断该待测音频文件是否满足条件。

这种方法在图像检索中用得十分广泛，但是在音频检索中却不适宜。因为音频帧的特征值通常都会有很大的变化，单纯依靠均值和方差会掩盖许多细节的变化，难以准确描述，所以需要使用更多的信息来使音频得到更准确的表示。本文使用自相关矩阵来表示音频，因为同类矩阵有着相似或者代数相近的不变量。

由 3.2 节可知，已得到了查询音频段和满足条件的待检音频段的特征值矩阵。首先对其进行标准化，使得该矩阵的均值为 0、方差为单位 1，设标准化后的特征值矩阵为：

$$\boldsymbol{X}_{m\times N}=(x_1,x_2,\cdots,x_N)$$

其中，$\boldsymbol{x}_i$ 是第 i 帧的音频特征向量。下面利用 $\boldsymbol{X}$ 的自相关矩阵 $\boldsymbol{C}$ 计算他的特征向量。$m\times m$ 维的自相关矩阵 $\boldsymbol{C}$ 可用式(5)求得：

$$\boldsymbol{C}_{m\times m}=\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \tag{5}$$

最后通过计算特征矩阵的 K-L 距离可得到语音片段之间的相似度，计算公式如下：

$$D_{kl}(C_1,C_2)=\frac{1}{2}(\overline{\mu_2}-\overline{\mu_1})^{\mathrm{T}}(C_2^{-1}-C_1^{-1})(\overline{\mu_2}-\overline{\mu_1})+\frac{1}{2}tr(C_1^{-1}C_2+C_2^{-1}C_1-2I) \tag{6}$$

4 实验分析

与音频分类不同，本文更偏重于同类音频中与查询音频高度相似的音频文件检索。本文实验所用数据文件全部是从电话信道中采集的语音所形成的语音数据文件，文件长度不限。数据均为单声道，采样率为 8 kHz，量化精度为 8 bit。首先把每个文件都去除静音帧，然后分成帧长是 32 ms 的帧序列，帧移 16 ms，滑动窗口步长也同样取 16 ms。

在实验中，查询音频段理论上可以取任意长度。但是由于发音具有一定的延续性，一般至少取 2 s 的音频段效果会比较好，不过也不宜过长。在查询时，如果数据库中某个待检音频文件的长度小于查询音频段，则直接跳过，不予处理。

本文用查全率和查准率 2 个指标来评价本文方法的检索性能。查全率即从检索源中正确检出的目标数和目标总数的比值；查准率即从检索源中正确检出的目标数和检索出的目标数的比值。

针对此数据库，任意选择数据，共进行了 8 次实验，实验结果统计如图 3、图 4 所示。其中，每组柱图左边为采用原始的直方图法进行检索的结果，右边为增加了可变阈值和 K-L 距离精确检索后的结果。信噪比对检索准确性的影响如表 1 所示。

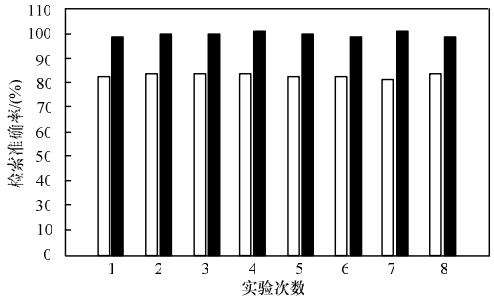

图 3 检索准确率

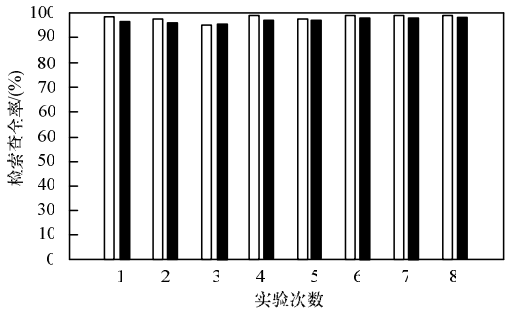

图 4 检索查全率

表 1 信噪比对检索准确性的影响

信噪比/dB	检索准确度/(%)
30	90.4
25	82.3
20	68.7

由实验结果可以看出：

(1)增加 K-L 相似度检索虽然查全率有 1%的轻微下降，但是查准率却得到了很大的提高，因此，整体检索性能也大幅度提高。

(2)本文实验利用数据间的相似性来检索包含查询音频段内容的待测音频文件，并不关心具体的内容。

(3)在实验中发现，滑动步长的大小影响实验结果，当滑动步长比较小时，查全率高，但所需间比较长；当滑动步长较大时，计算量比较小，需要的时间比较短，但查全率低和查准率的矛盾相对突出，因此，可根据需要进行调整。

5 结束语

本文根据大规模的音频数据检索问题，对传统的直方图算法进行了改进，可变的阈值使得直方图检索速度得到进一步提高。通过利用 K-L 距离计算特征矩阵的距离，提高了检索精度，并经实验验证了本文算法能够有效提高音频检索的精度，而且耗时较少。

参考文献

[1] 韩纪庆, 冯 涛, 郑贵滨, 等. 音频信息处理技术[M]. 北京: 清华大学出版社, 2007.

[2] Hanesn J H L, Huang Rongqing. Speech Find: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(5): 712-730.

[3] Chechil G, Le E, Rehn M, et al. Large Scale Content Based Audio Retrieval from Text Queries[C]//Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. New York, USA: ACM Press, 2008: 105-112.

[4] 张卫强, 刘 加. 网络音频数据库检索技术[J]. 通信学报, 2007, 28(12): 152-155.

[5] 张卫强, 刘 加. 一种基于仿生模式识别思想的固定音频检索方法[J]. 自然科学进展, 2008, 18(7): 808-813.

[6] Smith G, Murase H, Kashino K. Quick Audio Retrieval Using Active Search[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. New York, USA: IEEE Press, 1998: 3777-3780.

[7] Kashino K, Kurozumi T, Murase H. A Quick Search Method for Audio and Video Signals Based on Histogram Pruning[J]. IEEE Transactions on Multimedia, 2003, 5(3): 384-357.

[8] Kedem B. Spectral Analysis and Discrimination by Zero-crossings[J]. Proceedings of the IEEE, 1986, 74(11): 1477-1493.

[9] Saunders J. Real-time Discrimination of Broadcast Speech Music[C]//Proceedings of IEEE ICASSP'96. [S. 1.]: IEEE Press, 1996: 993-996.

[10] Li S Z. Content-based Classification and Retrieval of Audio Using the Nearest Feature Line Method[J]. IEEE Trans. on Speech Audio Processing, 2000, 8(5): 619-625.

[11] 江星华, 李 应. 基于LPCMCC的音频数据检索方法[J]. 计算机工程, 2009, 35(11): 246-247, 253.

编辑 索书志