基于音频指纹的两步固定音频检索①

乔立能[1], 夏秀渝[1], 叶于林[2]

[1](四川大学 电子信息学院, 成都 610064)
[2](中国人民解放军 78438 部队, 成都 610066)

摘 要: 提出了一种基于过零率和音频指纹的两步固定音频检索算法. 在基于过零率直方图的初步检索中, 采用直方图的迭代计算和动态的观测窗滑动步长来减少计算量并加快搜索速度, 快速筛选出相似度较高的候选音频片段; 接着基于降维 Philips 音频指纹对候选音频进行精检索, 进一步提高检索精度. 实验结果表明, 该音频检索算法在保证较好的检索准确性基础上, 大幅度提高了检索速度, 且具有较好的鲁棒性.

关键词: 音频检索; 过零率; 直方图; 音频指纹

Two-Stage Specific Audio Retrieval Based on Audio Fingerprinting

QIAO Li-Neng[1], XIA Xiu-Yu[1], YE Yu-Lin[2]

[1](College of Electronics and Information, Sichuan University, Chengdu 610064, China)
[2](78438 Troops of the Chinese People's Liberation Army, Chengdu 610066, China)

Abstract: This paper proposes a two-step fixed audio retrieval algorithm based on zero crossing rate and audio fingerprinting. The iterative calculation of the histogram and the sliding step of the observation time window are used in preliminary retrieval based on the zero crossing rate histogram to reduce the amount of calculation and speed up the search, fast filtering out candidate audio segments with high similarity; Then based on the dimension reduction Philips audio fingerprint, accurate retrieval of the candidate audio is carried out, further improving the retrieval accuracy. The experimental results show that the audio retrieval algorithm can improve the retrieval speed greatly and has good robustness, ensuring good retrieval accuracy.

Key words: audio retrieval; zero crossing rate; histogram; audio fingerprinting

1 概述

随着现代信息技术、多媒体技术和网络技术的发展, 多媒体信息的数据量急剧增多. 人们对如何在海量的多媒体库中快速找到感兴趣或有用的信息产生了越来越大的需求[1]. 在多媒体检索中, 音频检索是一个受人们关注且富有挑战性的研究课题[2,3]. 目前, 音频检索主要分为两大类: 一类是基于特征相似度的固定音频检索, 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]; 另一类是基于内容的音频检索技术, 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[6,7].

相对来说, 基于内容的音频检索数据复杂、技术难度大, 而基于相似度的固定音频检索实现简单灵活, 检索正确率高, 是实际常用的音频检索方法. 固定音频检索目前主要有基于距离的方法、基于特征直方图的方法[8,9]及上述 2 种方法的结合[10]. 基于特征直方图的方法本质上是属于概率统计的方法, 避免了复杂的空间距离计算, 检索速度较快, 但是检索精度低. 基于距离的方法将待检音频和模板音频按相同时间间隔划分成帧系列, 通过计算两者帧序列之间距离的累加和判断音频的相似度. 该方法的检索精度很高, 但是检索速度较慢. 文献[11]提出了基于模板子空间的固

① 收稿时间:2016-09-03;收到修改稿时间:2016-11-14 [doi: 10.15888/j.cnki.csa.005819]

定音频检索, 即根据模板间的相似性划分模板子空间, 确定各模板所属的子空间. 文献[12]利用文本搜索引擎中的倒排索引方法, 为音频建立音频字典和倒排索引, 提出基于倒排索引的音频检索方法. 建立音频索引是解决大规模静态音频数据库快速检索的有效手段, 但实际中也经常遇到未建立音频索引的情况, 如广播电台、电话等动态音频的实时监测, 事先未建立音频索引的动态音频库检索等. 这时无索引文件可用, 音频检索必须从原始音频数据分析做起, 实现快速准确的音频检索难度更大, 对检索鲁棒性的要求也更高.

针对实际中无索引文件可用的动态音频库检索, 本文提出了一种基于过零率和音频指纹的二步固定音频检索方法, 第一步利用过零率直方图从待检音频数据中初步筛选出相似度较高的音频片段, 第二步利用音频指纹对匹配出的音频片段进行精确检索, 进一步提高检索精度. 由于利用了迭代法计算直方图、动态滑动观测时间窗以及音频指纹的鲁棒性, 算法减少了计算量, 提高了筛选速度和音频检索的鲁棒性.

2 基于直方图的快速筛选

2.1 基本算法

直方图计算方法由于不用逐帧比较, 在检索速度上有着绝对的优势, 至今仍是固定音频检索领域使用常用的方法.

图 1 给出了直方图匹配算法的示意图. 首先, 计算查询音频和待检音频片段的特征矢量. 然后用一个等长的观测时间窗来观测查询音频和待检音频, 对观测窗内特征矢量进行量化后建立直方图. 接着比较查询音频和待检音频片段之间直方图的相似度. 当计算的相似度大于给定的门限值时, 认为初步搜索到指定音频, 记录待检音频中对应的时刻信息. 否则, 观测时间窗继续向前滑动进行下一步搜索.

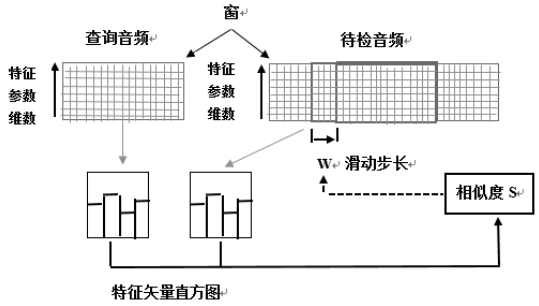

图 1 直方图算法示意图

直方图法作为本文音频检索的第一步, 目的是从大量音频数据中快速筛选出与待检音频相似度高的音频片段, 从时间消耗的角度当然希望采用计算量小的音频特征. 常用的音频特征有过零率、Mel 频率倒谱系数(MFCC)、感知线性预测(PLP)等, 其中 MFCC、PLP 计算复杂, 时间消耗大. 而过零率的计算简单, 且能较好区分不同声音. 为提高初步检索效率, 本文采用过零率来建立直方图.

根据查询音频过零率的取值范围, 划分出若干个等间隔的取值区间, 然后统计在每一个取值区间的过零率的频率, 这样就生成了直方图. 生成的直方图 h 可以表示为:

$$h=(h_1,\ h_2,...,h_l,...,h_L) \tag{1}$$

这里 L 是直方图的直方柱总数, h_i 是第 i 个直方柱的过零率的频率.

查询音频和待检音频片段之间的直方图相似度通常用直方图交集法进行测量. 其相似度定义为:

$$\mathrm{S}(h^Q,h^S)=\sum_{i=1}^{L}\min(h_i^Q,h_i^S) \tag{2}$$

其中, h^Q和h^S 分别是查询音频和待检音频片段的直方图.

2.2 观测窗滑动步长及直方图的迭代计算

采用直方图交集法的相似度具有一定的时间连续性, 因此不必逐帧进行直方图搜索匹配. 可以根据某一时间位置直方图的相似度, 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. 因为待检音频的观测时间窗是按照时间的先后顺序向前滑动的, 当观测时间窗从第 l_1 帧向前移动到第 l_2 帧时, 移动了(l_2 - l_1)帧, 在第l_2 帧, 直方图各取值区间的过零频数最多增加(l_2 - l_1)个, 假设时间窗内的总帧数是 N, 所以, 待检音频在第 l_2 帧的直方图的每个直方的最大值是第 l_1 帧的直方图的每个直方加(l_2 - l_1)/N, 因此, 当计算出第 l_1 帧待检音频和查询音频的相似度, 就可以知道在第 l_2 帧的相似度的上界.

$$S^u(h^Q,h^R(l_2))=S(h^Q,h^R(l_1))+(l_2-l_1)/N \tag{3}$$

其中, $h^R(l_1)$ 和 $h^R(l_2)$ 分别是待检音频窗函数在l_1 和 l_2 帧生成的直方图. 利用公式(3)和给定的门限值 S_T 可以给出窗函数向前滑动的步长:

$$w=\begin{cases}\mathrm{N}[\mathrm{S}_T-S(h^Q,h^R(l_1)]+1, & \text{if}\quad S(h^Q,h^R(l_1))<S_T\\ 1, & \text{otherwise}\end{cases} \tag{4}$$

其中, w 是滑动步长. 当计算的相似度超过给定的门限值时, 检索结束; 否则按照 w 的大小向前滑动窗函数, 继续进行检索. 由于观测窗动态滑动步长的引入, 使得检索速度大大提高.

直方图计算本质上就是观测时间窗内每个取值区间过零次数的累加, 因此可以通过迭代方法由前一个时间窗内的直方图求得后一个时间窗内的直方图. 公式(1)中, 过零率分为 L 个取值区间, 各区间的概率为 h_1, h_2, …, h_l, …, h_L, 设某一帧的过零率取值为 x, 定义:

$$X(i)=\begin{cases}1 & a_i<x\le b_i\\ 0 & \text{otherwise}\end{cases}, i=1,2,3,4\ldots,L \quad (5)$$

其中, a_i 和 b_i 分别是过零率第 i 个取值区间的下限和上限, 直方图计算可采用迭代公式:

$$h_i(l_2)=h_i(l_1)+\frac{1}{N}\sum_{j=l_1+N}^{l_1+N+w-1}X(i)_j-\frac{1}{N}\sum_{j=l_1}^{l_1+w-1}X(i)_j \quad (6)$$

其中, $h_i(l_1)$ 和 $h_i(l_2)$ 分别是观测窗滑动时前后两个时刻观测窗内过零率的概率, w 为滑动步长, 采用直方图迭代方法可以大幅度减少直方图的计算量.

3 基于音频指纹的精确检索

直方图编码的缺点是忽略了时序信息, 如将一段音频信号按时间倒序重新排列后, 它的直方图将和原音频信号相同, 另过零率本身包含的信息非常有限, 因此当待检音频与查询音频属于同一类音频(如语音)时, 检索的准确性能就会大大降低. 为了进一步提高检索准确性, 本文针对直方图法初步筛选出的音频片段, 采用音频指纹进行二次精确检索.

3.1 音频指纹

一个数字音频指纹可以视为一段音频的摘要, 即一个指纹函数 F 可以把一段包含大量数据的音频 X 映射为只有有限个比特的一个指纹. 音频指纹作为内容自动识别技术的的核心算法, 已广泛应用于音乐识别, 版权内容监播, 内容库去重和电视第二屏互动等领域. 使用音频指纹而不是音频数据本身进行比较和检索具有三方面好处: 因为指纹数据量相对比较小, 可以大大减少检索过程的相似度的比较计算量; 指纹来源于音频数据听觉最重要的部分, 因此在经受信号失真时仍能进行有效比对; 指纹数据库与媒体数据库相比尺寸减小很多, 可以进行更高效的搜索.

Philips 鲁棒音频指纹模型是业界许多实际商业应用的原型和学术界不断研究的对象. 当前音频哈希指纹方法不足以满足特定音频(如广告)的实时监测问题, 与现有方法相比, Philips 鲁棒音频指纹模型在保证音频检测准确性的同时, 能实现指纹的快速提取. 本文采用 Philips 鲁棒音频指纹模型[13,14], 指纹提取过程如下:

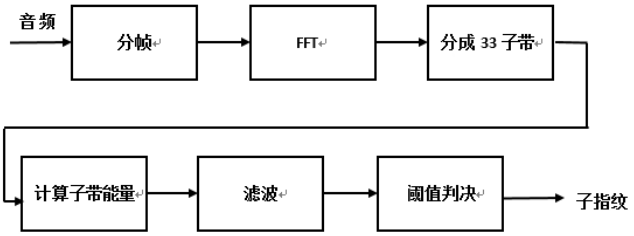

图 2　音频指纹提取算法框架

1) 分帧: 以每 0.064 秒为一帧对音频进行分帧, 帧与帧之间保持 50%的重叠率, 每一帧用相同长度的汉宁窗进行加权, 公式(7)为汉宁窗公式, 式中 N 为汉宁窗长度, 大小为一帧音频的样点数.

$$w(n)=0.5\text{-}0.5\cos(\frac{2\pi n}{N})\ n=0,1,...,N-1 \quad (7)$$

2) 傅立叶变换: 用快速傅里叶算法 FFT 对每一帧内容进行离散傅立叶变换 DFT, 一维离散傅立叶变换的定义公式如公式(8)所示, 其中 X(k)为频域信号, $x(n)$为时域信号, N 为 DFT 变换的样的长度:

$$X(k)=\sum_{n=0}^{N-1}x(n)\exp(-j\frac{2\pi}{N}nk)\quad k=0,1,\ldots,N-1 \quad (8)$$

3) 分成 33 子带: 将每一帧频谱图 300Hz-2000Hz 的内容按对数空间映射成 33 个不重叠的子带, 第 m 子带的起始频率也即第 m-1 子带的终止频率 $f(m)$可表示为式(9), 其中 F_{min} 为映射下限, 此处为 300Hz, F_{max} 为映射上限, 此处为 2000Hz, M 为子带个数, 此处为 33.

$$f(m)=\exp(\log F_{min}+(m\text{-}1)\frac{\log F_{max}\text{-}\log F_{min}}{M})\ \ m=1,2,...,M+1 \quad (9)$$

4) 计算能量: 计算每个子带所包含的能量, 设第 m 子带起始频率为 $f(m)$,终止频率为 $f(m+1)$, DFT 之后的频域信号为 $X(k)$, 则下式给出子带 m 的能量计算表达式:

$$E(m)=\sum_{f(m)}^{f(m+1)}|x(k)|^2 \quad (10)$$

5) 生成指纹: 假定第 n 帧的第 m 子带的能量为 $E(n,m)$,其对应的二进制指纹比特为 $F(n,m)$, 则音频指

纹的每个比特定义为:

$$F(n,m)=\begin{cases}1 & if \quad E(n,m-1)-E(n,m)-[E(n,m)-E(n,m+1)]>0\\ 0 & if \quad E(n,m-1)-E(n,m)-[E(n,m)-E(n,m+1)]\leqslant 0\end{cases} \tag{11}$$

所以每一帧数据最后生成 31 比特的二进制指纹信息.

3.2 指纹降维

上述音频指纹提取, 每一帧数据最后生成 31 比特的二进制指纹信息. 实际应用中, 希望进一步降低指纹维数从而有效的减少数据量. 本文提出基于音频指纹每一位方差大小来降低指纹维数的方法. 利用音频库数据, 我们统计了音频指纹每一位的方差, 由于随机变量的方差描述的是它的离散程度, 也就是该变量离其期望值的距离. 音频指纹某位方差越大, 不同音频在该位的差异越大, 说明该位的区分性越好, 反之区分性差. 所以保留区分性好的位, 而去掉区分性差的位, 可以将 31 维音频指纹转换为较低的维数从而有效的减少数据量.

3.3 精确检索方法

本文对直方图法初步筛选出的结果基于音频指纹进行二次精确检索. 首先提取查询音频段以及直方图相似度大于门限值的待检音频段的数字音频指纹. 然后对查询音频段和待检音频段的数字音频指纹进行比对, 这里就需要一个简单有效的检索匹配算法. 本文采用比特误差率(Bit Error Rate, BER)比较两个音频片段数字音频指纹之间的相似度, 其计算如下:

$$\mathrm{BER}=\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}F(n,m)\oplus F'(n,m)}{N\times M} \tag{12}$$

$F(n,m)$, $F'(n,m)$分别代表查询音频和待检音频第 n 帧音频指纹的第 m 位, N 为总帧数, M 为指纹位数. 当搜索到低于预设门限的比特误差率时, 则表明找到了匹配的音频文件.

4 实验分析

4.1 性能评测指标

为了对算法结果进行有效的评价, 本文采用了信息检索领域常用的评价标准: 查全率和查准率, 对检索结果进行评价, 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.

4.2 实验结果

本文实验所用数据采集于成都人民广播电台播放的节目, 包括新闻、音乐、广播剧、广告等, 音频数据总时为 20h, 数据均为单声道, 采样率为 8 kHz, 量化精度为 8 bit. 在提取声学特征参数时, 帧长为 0.064s, 帧移为 0.032s.

1) 音频指纹性能分析

首先考察信号幅度变化对音频指纹的影响. 设 $y(t)=ax(t)$, $x(t)$为原始音频, a 为放大系数, $y(t)$为幅度发生变化后的音频. 实验结果显示, 任意改变信号幅度(a 值随机选取), 同一音频所提取出来的音频指纹都是一样的, 即音频指纹不受幅度变化的影响.

接着, 我们考察了噪声对音频指纹的影响. 从数据库中随机选取了一段 30s 长的音频, 然后分别叠加不同信噪比的高斯白噪声生成带噪音频数据, 我们统计了不同信噪比下带噪音频和无噪音频的音频指纹误码率, 实验结果如图 3 所示.

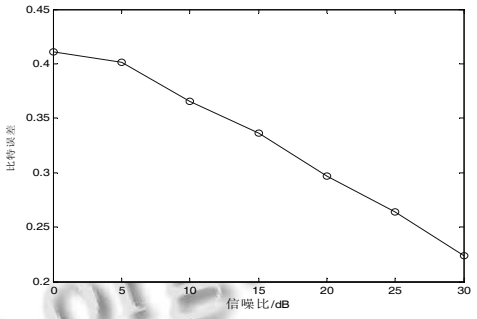

图 3 音频指纹距离曲线图

从图 3 可以看出, 音频指纹具有一定的抗噪性, 但还不算太好. 于是对提取音频指纹作如下改进:

$$F(n,m)=\begin{cases}1 & if \quad E(n,m-1)-E(n,m)-[E(n,m)-E(n,m+1)]>T\\ 0 & if \quad E(n,m-1)-E(n,m)-[E(n,m)-E(n,m+1)]\leqslant T\end{cases} \tag{13}$$

门限值 T 的取值以各帧信号子带能量的均值为基准, 并乘以不同系数 c 进行动态选取. 改进后的音频指纹抗噪性能如图 4 所示.

图 4 显示系数 c 取得越大, 音频指纹抗噪性能越好, 但音频指纹区分不同音频的能力也会下降. 我们反复实验显示, 当门限值取各帧信号子带能量均值的 0.1 倍时, 既能很好地提高音频指纹抗噪性能, 又能有效区分不同类型的音频. 后续实验均基于改进音频指纹完成.

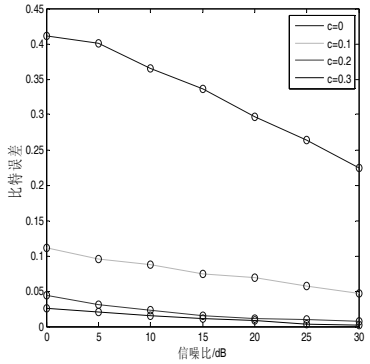

图 4　不同门限值对音频指纹的影响

我们从音频库中挑选出不同种类适量的音频数据, 提取其 Philips 音频指纹, 然后统计了音频指纹每一位的方差, 为音频指纹降维做准备. 31 位 Philips 音频指纹每一位的方差统计结果如图 5 所示.

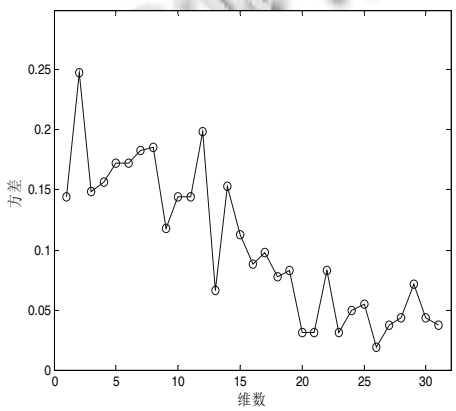

图 5　31 维音频指纹方差

根据 3.2 节的分析, 降维音频指纹将采取保留方差大的位, 去掉方差小的位来降低指纹维数.

2) 检索性能分析

利用采集的音频数据库, 每次随机从数据库中选择时长 2s 的音频作为查询音频, 然后对数据库进行检索, 每类实验重复进行 100 次实验.

① 不同维数音频指纹的检索性能

根据实验统计的音频指纹各位方差情况(图 5), 采取保留方差值大的位进行音频指纹降维. 分别选取了 31 维, 15 维, 7 维的音频指纹进行对比实验, 不同维数音频指纹的检索结果如表 1 所示.

表 1　音频指纹维数对检索结果的影响

音频指纹维数	查全率(%)	查准率(%)
31	99	96
15	96	94
7	95	73

表 1 表明, 音频指纹取 15 维时, 既能达到有效减少数据量的效果, 又能取得较好的检索性能, 所以最终我们选取了 15 维的音频指纹进行精检索.

② 初检索与精检索对比

本文基于两步法进行固定音频检索, 第一步采用过零率直方图法进行初检索, 选取的门限主要要保证足够高的查全率, 尽量不漏掉目标; 第二步依靠精检索来确保高的查准率. 每次随机从数据库中选择 2s 音频作为查询音频, 然后对数据库进行检索, 重复进行 100 次实验. 初检索和精检索实验结果如表 2 所示.

表 2　初检索和精精索性能对比

	查全率(%)	查准率(%)
初检索	99	63
精检索	98	96

从实验结果可以看出, 初检索在保证基本不漏检的情况下, 精检索可以大幅度提高检索准确性.

③ 检索鲁棒性实验

我们还在待检音频中加入噪声, 进行了不同信噪比情况下的仿真实验, 实验结果如表 3.

表 3　信噪比对检索结果的影响

信噪比/dB	查全率(%)	查准率(5)
30	99	95
25	98	93
20	96	85
15	98	81
10	93	73
5	92	69

由实验结果可以看出, 查全率几乎不受噪声的影响; 当信噪比降低时, 检索准确性有不同程度的下降. 由音频指纹的性能分析可知, 抗噪性能与提取音频指纹的门限值有关, 当门限值越大, 检索准确性越好, 但音频指纹区分不同音频的能力也会下降. 而本文方法主要适用于录音片段在线查询等应用, 实际中录音机的信噪比应在 40dB 以上, 实验结果表明该方法能够满足实际应用需求.

5 结语

本文提出了一种基于过零率和音频指纹的二步固定音频检索方法. 首先利用过零率直方图从待检音频数据中快速筛选出相似度较高的音频片段, 采取直方图迭代计算, 动态的观测窗滑动步长等措施减少计算量并加快了搜索速度. 然后利用降维 Philips 音频指纹对匹配出的音频片段进行精确检索, 基于降维音频指纹的简洁性、区分性及鲁棒性, 精检索不仅提高了检索精度, 而且检索匹配速度快, 具有良好鲁棒性. 实验结果给出该音频检索算法良好性能的证明. 本文重点针对无索引文件可用的动态音频检索问题, 提出了一系列简化计算、加快搜索速度的措施, 适用于录音片段在线查询等应用, 后续我们将针对大规模静态音频数据库建立音频索引开展进一步的研究应用.

参考文献

1 Wang Y, Liu Z, Huang JC. Multimedia content analysis using both audio and visual clues. IEEE Signal Processing Magazine, 2000, 17(6): 12–36.

2 Foote J. An overview of audio information retrieval. Multi-Media Systems, 1999, 7(1): 2–10.

3 杨继臣,王伟凝.一种基于随机段的固定音频检索方法.计算机应用,2010,30(1):230–232.

4 张卫强,刘加.网络音频数据库检索技术.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.自然科学进展,2008,18(7):808–813.

6 Hanesn JHL, Huang RQ. Speech find: Advances in spoken document retrieval for a national gallery of the spoken Word. IEEE Trans. on Speech and Audio Processing, 2005, 13(5): 712–730.

7 Chechil G, Le E, Rehn M, et al. Large scale content based audio retrieval from text queries. Proc. of the 1st ACM International Conference on Multimedia Information Retrieval. New York, USA. ACM Press. 2008. 105–112.

8 Kashino K, Kurozumi T, Murase H. A quick search method for audio and video signals based on histogram pruning.IEEE Trans. on Multimedia, 2003, 5(3): 348–357.

9 Kim KM, Kim SY, Jeon JK, et al. Quick audio retrieval Using multiple feature vectors. IEEE Trans. on Consumer Electronics, 2006, 52(1): 200–205.

10 齐晓倩,陈鸿昶.基于 K-L 距离的两步固定音频检索方法.计算机工程,2011,37(19):160–162.

11 谈会星,陈福才,李邵梅.基于模板子空间的快速固定音频检索方法.计算机工程,2012,38(20):260–263.

12 张雪源,贺前华.一种基于倒排索引的音频检索方法.电子与信息学报,2012,34(11):2561–2567.

13 郭杰,王之禹.应用于快速音乐检索系统中的音乐指纹提取算法.中国声学学会 2007 年青年学术会议.2007.135–136.

14 李伟,李晓强,陈芳,王淞昕.数字音频指纹技术综述.小型微型计算机系统,2008,29(11):2124–2130.