计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(8):165−172 [doi: 10.15888/j.cnki.csa.007577]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于压缩感知和音频指纹的固定音频检索方法①

赵文兵, 贾懋珅, 王　琪

(北京工业大学 信息学部, 北京 100124)
通讯作者: 贾懋珅, E-mail: jiamaoshen@bjut.edu.cn

摘　要: 针对现有音频检索中样本音频特征库数据量较大且检索速率慢问题, 本文提出一种基于压缩感知和音频指纹降维的固定音频检索方法. 在音频检索的训练阶段, 首先, 对样本音频信号进行稀疏化处理, 并通过压缩感知算法对稀疏化后的音频数据进行压缩; 其次, 提取压缩信号的音频指纹; 再次, 引入音频指纹离散基尼系数通过计算音频指纹各维度的离散基尼系数对指纹实施降维, 最终得到检索特征库. 在音频检索阶段用和训练阶段相同的算法提取待检音频的特征与音频特征库数据匹配得出检索结论. 实验结果表明, 所提音频检索方法在确保较好的检索准确率的基础上, 大幅度减小了样本音频数据库的存储量, 提高了音频的检索速率.

关键词: 音频检索; 压缩感知; 离散基尼系数; 音频指纹

引用格式: 赵文兵,贾懋珅,王琪.基于压缩感知和音频指纹的固定音频检索方法.计算机系统应用,2020,29(8):165–172. http://www.c-s-a.org.cn/1003-3254/7577.html

Specific Audio Retrieval Method Based on Compressed Sensing and Audio Fingerprint

ZHAO Wen-Bing, JIA Mao-Shen, WANG Qi

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: In order to solve the problem of large amount of data and slow retrieval speed in the existing audio retrieval, a fixed audio retrieval method is proposed in this study based on compressed sensing and audio fingerprint dimensionality reduction. In the training stage of audio retrieval, the sample audio signal is sparse processed, and the sparse audio data is compressed by the compression sensing algorithm, then the audio fingerprint is extracted, and then the audio fingerprint discrete Gini coefficient is introduced to reduce the dimension of the fingerprint by calculating the discrete Gini coefficient of each dimension of the audio fingerprint. In the recognition stage of audio retrieval, we use the same algorithm as in the training stage to process the audio to be tested and match with the sample audio fingerprint. The experimental results show that the proposed audio retrieval method greatly reduces the storage of the sample audio database and improves the audio retrieval speed on the basis of ensuring a better retrieval accuracy.

Key words: audio retrieval; compressed sensing; discrete Gini coefficient; audio fingerprinting

随着数字化信息的快速发展, 各种以音频为载体的作品也越来越多, 海量的音频信息丰富了人们的生活同时也给人们带来了麻烦, 如何准确、快速的从数据库中获取自己想要的信息, 已经成为信息时代人们迫切需要同时也是音频检索领域的重要研究问题之一. 目前, 音频检索主要分为两大类: 一类是基于特征相似度匹配的固定音频检索, 其基本原理是对给定的待查询音频片段, 在样本音频库中检索与其相同或同源的片段[1,2];

① 基金项目: 国家自然科学基金 (61971015)
Foundation item: National Natural Science Foundation of China (61971015)
收稿时间: 2020-01-19; 修改时间: 2020-02-02; 采用时间: 2020-03-13; csa 在线出版时间: 2020-07-29

另一类是基于内容的音频检索技术[3], 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音频信息检索.

相对来说, 基于内容的音频检索技术较难, 该类方法需依据生物语言特征和声韵等信息去识别音频的内容, 算法比较复杂主要用于人机交互领域. 而基于特征相似度匹配的固定音频检索相对较为简单, 算法复杂度较低, 它不需要识别出待检音频的内容只需要根据其音频特征与样本音频特征库数据进行相似度比较来确定待检音频是否为目标音频, 此音频检索技术适用范围较广, 常用于音乐搜索、音频版权保护以及广告监测等领域.

固定音频检索技术目前主要在匹配方法上进行了研究, 有基于特征直方图的方法、基于距离的方法[4,5]及上述两种方法的结合[6]. 且这两种方法所用的音频特征都是传统音频特征, 基于特征直方图的方法简单、快速, 但是检索准确率不高, 基于距离的方法其计算复杂度较高. 这两种算法的不同之处在于检索阶段采取的特征相似度[7] 判别方式不同, 但是在检索之前, 都需要预先对样本模板和待检音频进行特征提取和矢量量化, 而正是这些预处理耗费时间, 并在很大程度上决定检索的准确度[8]. 另外, 在检索过程中样本音频特征数据库的存储量是决定检索速率的关键因素之一. 而音频指纹具有数据量较小、抗噪性能较高、特征参数提取相对简单等优点深受该领域学者青睐, 其中 Philips 算法[9] 是其中比较经典的一种, 一经提出便受到广泛关注. Philips 算法在各种信号畸变情况下具有良好性能, 并且速度方面有很大的优势, 但是当信号有较快线性速度改变时性能不够理想. 近年来, 也有学者提出利用人工智能识别音频片段的指纹检索技术[10]. 将小波包系数的奇异值熵以及样本熵相融合作为音频片段信号的特征参数, 提取出音频指纹, 但是, 此方法需要神经网络训练, 算法复杂度较高. 也有学者利用采样子指纹和计数匹配进行音频检索[11], 该方法是提取一段音频的多个子指纹并标记, 在指纹匹配时进行子指纹计数并匹配, 该方法检索准确率较好, 由于需要多次计算子指纹使得该方法的检索速率不太理想. 另外, 有国内学者提出基于压缩感知梅尔倒谱的检索[12] 算法 (Compressed Sensing Mel Frequency Cepstrum Coefficient, CS-MFCC) 和国外学者提出[13] 基于子指纹掩码 (Sub-fingerprint Masking, SM) 的音频指纹检索算法具有很好的检索效果.

针对实际中固定音频检索样本音频特征数据库存储量大的问题, 本文提出一种基于压缩感知和音频指纹降维的音频检索方法, 该方法在构建样本音频特征库时利用压缩感知算法先对样本音频进行压缩处理再提取音频指纹特征, 然后, 对提取的音频指纹引入离散基尼系数进行指纹特征降维. 由于, 该方法对样本音频采取先压缩再进行特征降维, 这就使得在同量的样本音频下该方法构建的样本音频特征库的数据量较小, 算法减少了计算量, 提高了筛选速度和音频检索的鲁棒性.

1 基于压缩感知的音频特征库构建

1.1 声音预处理

由于音频信号具有短时平稳性, 且音频数据的首末段以及中间段有不含信息的音频段, 为了更高效的压缩样本音频, 需要对样本音频进行预处理, 分为带通滤波、预加重、分帧、加窗和静音帧判别.

1.2 音频信号的压缩处理

考虑到音频信号数据较大, 直接提取特征会使得构建的特征库数据量大, 变相增加了检索工作量. 为此, 本文在特征提取前对音频信号进行压缩感知, 来解决特征库数据量大的问题. 压缩感知算法是由 Donoho 等[14] 在 2006 年提出的概念, 是对信号压缩的同时进行采样, 不同于传统的 Nyquist 采样定理, 在压缩感知的理论框架下, 采样速率不再取决于信号的带宽, 而是取决于信息在信号中的结构和内容[15]. 当信号为稀疏信号时, 压缩感知可以以远小于采样定理要求的采样数, 通过重构算法重构原始信号[16].

为验证音频信号在频域的稀疏性, 本文选用爱荷华大学音乐乐器样本库 (University of Iowa Music Instrument Samples, Iowa-MIS)[17] 中的数据作为样本进行分析, 统计了 6 类音频信号 (采样率为 16 kHz) 的帧能量保留比与时频成分保留个数间关系[18], 如图 1 所示. 其中, 纵坐标表示各帧保留的时频点个数 (按照频率成分幅度由大到小的顺序保留时频点); 横坐标表示保留相应数量的时频成分时, 所保留的时频成分能量占该帧信号总能量的百分比. 时频变换选用 1 024 点的离散余弦变换 (Discrete Cosine Transform, DCT), 帧能量保留比从 98% 到 80% 均匀变化时, 统计分析相应的时频保留个数.

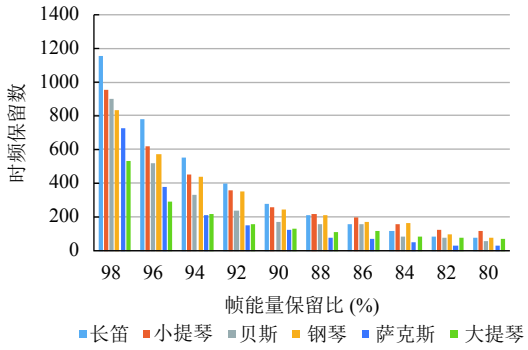


图 1 不同帧能量保留比下 6 类音频信号时频保留数统计

从图 1 中可以看出, 6 类音频信号的时频保留数随着帧能量保留比的均匀下降以非线性方式下降. 可见, 音频信号在频域的能量呈非均匀分布, 大部分能量集中在少数时频系数中. 以钢琴为例, 在帧能量保留比为 90% 时, 时频保留数为 256 个为总数 2048 的 1/8, 同样, 贝斯在帧能量保留比为 92% 时, 时频保留数为 256, 说明关键的 256 个时频系数可以包含一帧音频 92% 的信息能量. 因此可知, 音频信号在频域呈现明显的能量集中性, 即其在频域具有稀疏特性. 基于此, 本文将压缩感知理论引入音频检索领域并对其理论进行改进.

设 $x=[x_n(1),x_n(2),\cdots,x_n(N)]$ 为预处理后的第 n 帧音频信号, 根据稀疏编码模型音频信号 $x_n(p)$ 在 DCT 域的频域系数 α 可用式 (2) 表示:

$$\alpha=\psi x \tag{1}$$

其中, ψ 为 $N\times N$ 的 DCT 基矩阵, $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$.

根据上述实验可知音频信号在频域具有稀疏特性, $\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$ 中最大的 Q (Q<<N) 个系数集中了 N 个系数的绝大部分能量, 即 α 具有类 Q-稀疏特性. 基于此, 仅保留 α 中最大的 Q 个系数, 并将其余系数置零, 构成新向量 $\alpha'=[\alpha'_1,\alpha'_2,\cdots,\alpha'_N]^{\mathrm{T}}$, 即, α' 中仅有 Q 个非零元素. 由此, 构建第 n 帧频域稀疏化后的音频信号 $\bar{x}$:

$$\bar{x}=\psi^{\mathrm{T}}\alpha' \tag{2}$$

其中, $\bar{X}=[\bar{x}_n(1),\bar{x}_n(2),\cdots,\bar{x}_n(N)]$, ψ^{T} 为 ψ 的转置矩阵. 设置的 Q 值反映了音频信号 $\bar{X}$ 在 DCT 域稀疏化程度, Q 越小 DCT 域的稀疏性越好.

此时, 完成音频信号稀疏化处理, 得到满足压缩条件的时域稀疏信号 $\bar{X}$. 要实现对稀疏音频信号 $\bar{X}$ 的压缩, 需要通过观测矩阵将稀疏信号投影到低维空间. 为保证音频检索过程中观测矩阵在训练和识别阶段一致, 选择一个稳定的观测矩阵至关重要. 考虑到音频信号具有短时平稳性, 即, 相邻若干样点变化平缓, 故本文选行阶梯矩阵[19] 为观测矩阵. 通过此观测矩阵将稀疏音频信号相邻的几个采样点合成一个采样值, 这样既压缩了音频信号又保持了音频信号的短时平稳性, 便于后续二次分帧处理.

将上述 Q-稀疏化后的第 n 帧信号 $\bar{X}$ 通过行阶梯观测矩阵 Φ 投影得到 M 维的观测序列信号:

$$Y=\Phi\bar{X} \tag{3}$$

其中, $Y=[y_n(1),y_n(2),\cdots,y_n(R)]$, Φ 为 $H\times N$ 的观测矩阵 (H<N). 当选定的压缩比 N/H=3 时, 观测矩阵为:

$$\Phi=\begin{bmatrix} 111000000000\cdots0 \\ 000111000000\cdots0 \\ 000000111000\cdots0 \\ \vdots \\ 000000000\cdots0111 \end{bmatrix}_{H\times N} \tag{4}$$

故 $N\times 1$ 稀疏音频信号 $\bar{X}$ 经过观测矩阵 Φ 压缩后得到维度为 $H\times 1$ 的观测信号 Y 减小了音频序列数据量.

1.3 稀疏音频指纹特征提取

在音频的众多特征中, 音频指纹是近年来最受欢迎的一种, 音频指纹是指可以代表一段音频重要声学特征的基于内容的紧致数字签名, 其主要目的是用少量的数字信息代表大量音频数据信息. 它相对于传统的音频特征具有 3 个优点, 因为音频指纹数据量较小, 可以减小特征数据库的存储量从而提高音频特征匹配速度; 指纹的抗噪性能较高, 可以减小音频识别过程中的干扰因素; 音频指纹特征提取流程相对简单, 因此可以减少特征提取的时间增加音频减速速率.

在众多的音频指纹中, Philips 音频指纹模型因具有较高的鲁棒性且算法较为简单, 本文以此指纹模型为基础进行音频指纹提取. 首先, 对上述压缩后的音频数据 Y 进行二次分帧; 其次, 对分帧后信号进行离散傅里叶变换并对频域信号进行频谱子带划分, 从频谱中选取 M 个非重叠的频带, 频带之间是等对数间隔的. 再次, 计算每帧音频的各个子带能量, 分别求其上述选取的 M 个非重叠频带的能量. 最后, 根据子带能量的判别生成每帧音频的子指纹, 上述每帧所求的 M 个子带能量比特差分判别公式如下:

$$F(n,m)=\begin{cases} 1, \text{ if } t(n,m)-t(n-1,m)>0 \\ 0, \text{ if } t(n,m)-t(n-1,m)\le 0 \end{cases} \tag{5}$$

其中, $E(n,m)$ 表示音频第 n 帧的第 m 子带能量, $t(n,m)=E(n,m)-E(n,m+1)$ 表示第 n 帧的第 m 子带和 m+1 子带的能量差, $F(n,m)$ 为对应的二进制比特音频指纹信息. 最终, 每帧音频最后生成一个 M–1 维的二进制子指纹信息.

1.4 音频指纹降维

对于一段音频来说, 所含的音频指纹信息是由多个二进制子指纹信息构成, 其指纹信息数据量仍然很大, 在实际应用中, 希望进一步降低音频指纹维数从而有效减少指纹数据量. 为此, 本文提出基于离散基尼系数计算的音频指纹降维方法. 求取音频指纹的每一维度离散基尼系数, 各维度指纹的离散基尼系数反映了音频指纹该维度数据的离散程度, 即音频指纹该维度数据的差异性大小. 音频指纹某维的离散基尼系数越大, 不同音频在该维的差异就越大, 说明该维数据的区分性越好, 反之区分性差. 本文通过保留音频指纹中区分性较好维的信息, 去掉区分性较差维的信息, 从而实现降低指纹维数的目的.

音频指纹各维度的离散基尼系数计算过程如下:

(1) 求取音频指纹的离散洛伦兹曲线, 离散洛伦兹曲线是求离散基尼系数的关键曲线, 是由累积指纹数据占比矢量 $\vec{W}^j$ 的各个元素构成, j 表示音频指纹的维度序号, 取值范围 j=1,2,⋯, M–1 求取累积指纹数据占比矢量 $\vec{W}^j$ 的计算过程如下:

将音频指纹库中的各类音频指纹按帧处理, 音频指纹每 50 帧指纹数据为一组共分成 L 组, 构建第 j 维累积指纹数据矢量:

$$\vec{C}^j=[c_1^j,c_2^j,\cdots,c_L^j] \tag{6}$$

其中, $c_i^j=\sum_{b=1}^{50}F(50\times(i-1)+b,j)$, i=1,2,⋯, L 为组编号且有 $c_1^j\le c_2^j\le c_3^j\le\cdots\le c_L^j$, 构建第 j 维累积指纹数据占比矢量 $\vec{W}^j=[w_1^j,w_2^j,\cdots,w_L^j]$, 矢量中各元素定义为:

$$w_1^j=\frac{c_1^j}{\left\|\vec{C}^j\right\|_1},\ w_2^j=\frac{c_1^j+c_2^j}{\left\|\vec{C}^j\right\|_1},\cdots,w_L^j=\frac{c_1^j+c_2^j+\cdots+c_L^j}{\left\|\vec{C}^j\right\|_1}=1 \tag{7}$$

占比矢量 $\vec{W}^j$ 各元素构成的曲线为离散洛伦兹曲线, 如图 2 所示的曲线.

(2) 以上述所求的离散洛伦兹曲线为分界线, 可得音频指纹第 j 维度的基尼系数公式如下:

$$G^j=\frac{S_a}{S_a+S_b} \tag{8}$$

如图 2 所示, 其中, S_a为坐标对角线段 OA 与离散洛伦兹曲线围成的闭合面积, 点 O 的坐标为 (0,0) 点 A 的坐标 (1,1), S_b为坐标线段 OB、BA 与离散洛伦兹曲线围成的闭合面积, 点 B 的坐标为 (1, 0), G^j为音频指纹第 j 维度的基尼系数.

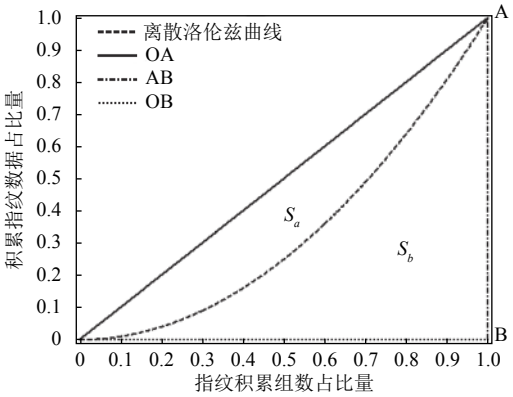


图 2 音频指纹离散基尼系数示意图

由上述可知, S_a+S_b的和为对角线段 OA 与线段 OB、BA 所围成的闭合面积, 即: $S_a+S_b=1/2$, 因为音频指纹是离散的, 故本文将上述公式离散化为:

$$\tilde{G}^j=1-\frac{1}{L}\left(2\times\sum_{i=1}^{L-1}w_i^j+1\right) \tag{9}$$

由此, 得到音频指纹第 j 维离散基尼系数$\tilde{G}^j$, 其中 i 为组编号, w_i^j为音频指纹第 j 维度累积第 i 组指纹数据占比量.

最终, 通过统计音频指纹各维度的离散基尼系数, 去掉区分性较差维的信息得到降维指纹$F'(n,r)$其中, r=1,2,⋯,R ($R<M$–1), 进而构建音频特征库.

2 音频特征检索

本文采用比特误码率作为匹配相似度判定, 具体过程如下:

(1) 选取待测音频经上述预处理、稀疏化处理以及压缩处理得到待测观测序列信号$\bar{Y}$.

(2) 将上述压缩处理后的待测观测序列信号$\bar{Y}$经指纹特征提取、指纹特征降维得到待测音频指纹$F_d(n,r)$, 其中, $F_d(n,r)$表示待测音频信号序列第 n 帧音频指纹的第 r 位.

(3) 将得到的待测音频指纹与样本音频指纹库中

的音频指纹进行相似度匹配, 本文选取比特误差率 (Bit Error Rate, BER) 作为匹配算法比较两个音频片段之间的相似度, 其计算公式如下:

$$BER = \frac{\sum_{n=1}^{T}\sum_{r=1}^{R} F_d(n,r) \oplus F'(n,r)}{T \times R} \tag{10}$$

其中, $\oplus$为异或操作, $F'(n,r), F_d(n,r)$分别代表降维后的样本音频和待检音频第 n 帧音频指纹的第 r 位, T 为音频总帧数, R 为音频指纹位数.

(4) 设置比特误差率的阈值, 求其 BER 的值, 若其值小于设定的阈值, 则表示待检音频与样本音频库中的音频相似度较高, 反之, 待检音频与样本音频库中的音频相似度较低, 从而得出检测结果.

3 实验结果与分析

3.1 性能评价指标

为了验证算法的有效性, 本文选用音频检索中常用的查全率与查准率作为性能评价标准; 查全率与查准率的定义如下:

查全率=从检索源中检出的正确目标数/应检索出的目标数

查准率=从检索源中检出的正确目标数/实际检索出的目标数

3.2 实验结果分析

本文实验主要在不同信噪比的数据集进行检索, 以验证本文算法的检索性能. 所用数据采样率为 8 kHz, 特征提取处理帧长为 0.256 s, 帧移为 0.032 s, 对于压缩后的音频数据每帧分为 33 个子带, 即 M=33.

数据库 1: 包含 5000 个音频文件, 每个音频文件长 3 s~5 min, 主要为课题所在实验室的采集语音数据及从互联网采集的音频数据, 总大小约为 12.3 GB, 总时长为 230 h, 音频文件为 8 kHz 采样 16 bit 编码的 PCM 格式.

数据库 2: 针对数据库 1, 添加白噪声形成信噪比为 40 dB 的数据集.

数据库 3: 针对数据库 1, 添加白噪声形成信噪比为 30 dB 的数据集.

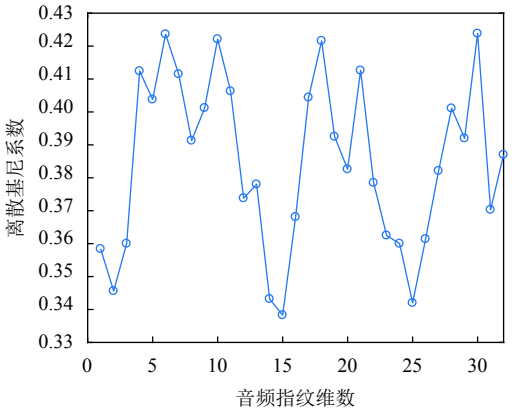
数据库 4: 针对数据库 1, 添加白噪声形成信噪比为 20 dB 的数据集.

数据库 5: 针对数据库 1, 添加白噪声形成信噪比为 10 dB 的数据集.

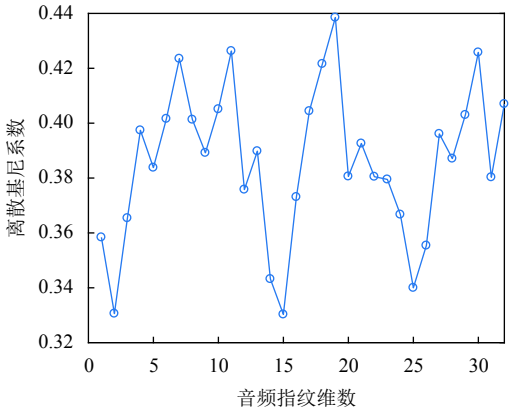
数据库 6: 从数据库 1 中任意选取 1000 个音频文件, 从其中随机位置截取一段时长为 3 s 的音频数据作为检索片段.

3.2.1 音频指纹降维程度分析

为确定音频指纹降维能量, 本文从音频数据库 1 中选取语音类数据和歌曲类数据, 求取所选数据音频指纹各维度的离散基尼系数, 统计音频指纹各维度的离散基尼系数. 图 3(a)、图 3(b) 分别给出了 250 段与 500 段数据的 32 维音频指纹各维度的离散基尼系数的均值.



(a) 250段数据音频指纹各维度的离散基尼系数



(b) 500段数据音频指纹各维度的离散基尼系数

图 3 语音与歌曲数据音频指纹各维度的离散基尼系数

从图 3 可以看出测试的数据量不同时 (250 段与 500 段), 得到的音频指纹各维度的离散基尼系数的均值不相同, 但是最小离散基尼系数所对应的维数是相同的. 即, 在两个不同体量的测试数据中, 得到的结果都是音频指纹在第 2、14、15、25 维的离散基尼系数相对其他维数都比较低, 说明音频指纹在这几维的信息区分度相对较低. 根据 1.4 节分析, 降维音频指纹将保留指纹离散基尼系数大的维度信息, 舍去指纹离散基尼系数小的维度信息. 因此, 可以去掉音频信号的这几维指纹

信息, 从而达到指纹降维目的. 以此类推, 若想进一步降维可以通过图 3 看出指纹离散基尼系数在第 1、3、24、26 维也相对较低, 可以尝试去除这几维的指纹信息.

3.2.2 样本压缩比与指纹降维对检索性能的影响

利用样本音频库中的各类音频, 依次选取音频数据作为待查询音频, 然后对样本特征数据库进行检索.

(1) 样本不同压缩程度对检索性能的影响

本实验选取数据库 6 中的数据集为待查询音频, 在数据库 1 进行检索. 比较不同样本压缩比下构建的特征库的检索效果. 此实验中, 构建特征数据库时不进行指纹特征降维操作. 样本压缩比 N/H 分别设置为 1、2、3、4、5 时, 音频检索性能如表 1 所示.

表 1 样本压缩程度对检索结果的影响 (%)

样本压缩比	查全率	查准率
1	100	99.8
2	99.6	99.4
3	99.2	97.5
4	95.3	91.7
5	84.7	76.2

表 1 表明, 当样本压缩比 N/H 为 2 和 3 时, 检索效果相对较好. 考虑到样本压缩比为 3 时, 既能多压缩样本数据又能取得较好的检索效果, 因此, 样本压缩比取 3 时最为合适.

(2) 指纹维数对检索性能的影响

根据图 3 所得的音频指纹各维度的离散基尼系数情况, 采取保留指纹离散基尼系数大的维度信息, 舍去指纹离散基尼系数小的维度信息的方式进行音频指纹降维. 结合图 3 结果, 本文尝试分别丢弃 0 维 (不丢弃)、4 维、6 维和 14 维离散基尼系数最小的音频指纹信息构建特征库. 即, 音频指纹降维至 32 维、28 维、26 维和 18 维. 此实验选取数据库 6 中的数据集为待查询音频, 在数据库 1 进行检索. 此实验过程样本音频不做压缩处理. 比较不同指纹维数对检索性能的影响结果如表 2 所示.

表 2 指纹维数对检索结果的影响 (%)

指纹维数	查全率	查准率
32	100	99.8
28	100	98.5
26	96.2	91.4
18	93.1	67.6

表 2 表明, 音频指纹降至 28 维与 26 维时, 查全率相对较好, 但考虑到查准率时, 音频指纹降至 28 维时既能保证降低指纹维数, 又能保证检索性能, 因此, 指纹降至 28 维较为合适.

(3) 样本压缩程度和指纹降维程度对检索性能的影响

由表 1 可以看出样本压缩比 N/H 为 2 和 3 以及 4 时检索性能较好, 由表 2 可以看出指纹维数降至 28 维和 26 维时, 检索性能较好. 结合这两个实验结果中最好的参数, 选取数据库 6 中的数据集为待查询音频, 在数据库 1 进行检索. 比较不同压缩比和音频指纹情况下所提方法的检索性能如表 3 所示.

表 3 样本压缩结合指纹降维对检索结果的影响

样本压缩比	指纹维数	查全率(%)	查准率(%)
2	28	99.8	98.6
2	26	95.4	89.5
3	28	98.7	97.8
3	26	94.5	86.3
4	28	94.8	90.6
4	26	90.1	75.8

表 3 表明, 在综合考虑到减小样本音频特征库数据量与保证检索准确率的情况下, 样本压缩比 N/H=3 及音频指纹降至 28 维时, 既能减小样本音频特征库数据量又能保证检索准确率. 因此, 选取样本压缩比为 3 以及音频指纹为 28 维进行音频检索最为合适.

3.2.3 不同信噪比下不同算法的音频检索性能对比

为了验证本文算法的优劣性, 特将本文算法与其他同类型检索方法进行性能比较. 考虑到基于压缩感知梅尔倒谱的检索算法[12](Compressed Sensing Mel Frequency Cepstrum Coefficient, CS-MFCC) 和基于子指纹掩码 (Sub-fingerprint Masking, SM) 的音频指纹检索算法[13] 具有很好的检索效果. 本文选用这两个方法为参考方法, 简写为 CS-MFCC 算法和 SM 算法.

在本次对比试验中, 本文方法依据上述的综合讨论取样本压缩比为 3 以及音频指纹为 28 维的指纹特征作为实验特征参数. 在音频检索阶段时, 添加不同信噪比的高斯白噪声作为干扰, 选取数据库 6 中的数据集为待查询音频, 分别在数据库 1、2、3、4、5 进行检索. 3 种方法的检索性能如表 4 所示.

由表 4 可以看出, 在信噪比相同的情况下, 本文算法的查全率与查准率相对较高, 说明在相同环境下本文的算法法案优于 CS-MFCC 算法和 SM 算法. 另外, 在不同信噪比下, 3 种算法的查全率与查准率都发生不同程度的改变. 本文算法、CS-MFCC 算法和 SM 算法

的查全率变化趋势如图 4 所示, 查准率的变化趋势如图 5 所示.

表 4 不同信噪比下不同算法的音频检索性能 (%)

SNR	本文算法		CS-MFCC		SM	
(dB)	查全率	查准率	查全率	查准率	查全率	查准率
clean	98.7	97.8	94.3	92.5	97.8	94.6
40	98.3	97.5	91.8	90.5	94.8	93.6
30	97.2	96.8	89.5	88.7	93.4	92.5
20	94.6	93.8	83.4	82.6	92.3	90.1
10	87.3	86.2	77.5	75.4	85.8	85.6

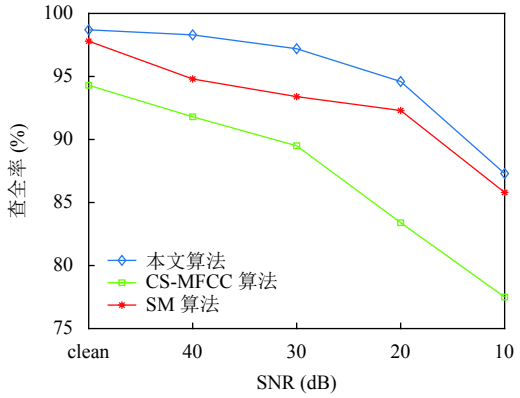


图 4 3 种算法的查全率趋势图

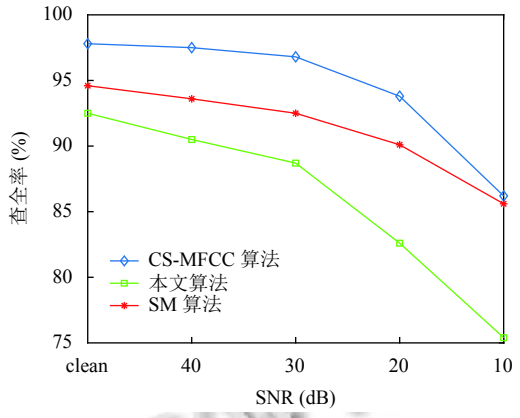


图 5 3 种算法的查准率趋势图

由图 4 和图 5 可以看出本文算法、CS-MFCC 算法和 SM 算法的查全率与查准率虽然都随着信噪比的降低而减小. 但是, 减小的幅度与快慢不同, 说明 3 种算法的鲁棒性能不同. 在信噪比为 20 dB 以上时, 本文算法与 SM 算法的鲁棒性相差不大, CS-MFCC 算法的鲁棒性相对较差. 在信噪比低于 20 dB 后, SM 算法的鲁棒性比本文算法的鲁棒性较为好点, CS-MFCC 算法的鲁棒性在信噪比低于 30 dB 后就开始急速变差. 综上所述, 3 种算法中本文算法的检索性能与鲁棒性都相对较好, 因此, 可以知本文方法具有良好的检索性能.

4 结束语

本文针对现有音频检索中样本音频特征库数据量较大且检索速率慢问题, 提出一种基于压缩感知和音频指纹降维的固定音频检索方法, 该方法利用压缩感知算法对样本音频进行先压缩再提取音频指纹特征随后引入离散基尼系数对音频指纹进行降维, 使得样本音频特征库的数据量减小. 该方法的特征匹配算法简单, 而且匹配速率较快, 实验表明, 该方法在选取合适的样本音频压缩比与音频指纹维数时具有较好的检索性能.

参考文献

1 张卫强, 刘加. 网络音频数据检索技术. 通信学报, 2007, 28(12): 152–155. [doi: 10.3321/j.issn:1000-436x.2007.12.026]

2 张卫强, 刘加, 陈恩庆. 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. [doi: 10.3321/j.issn:1002-008X.2008.07.013]

3 Doidge AN, Evans LH, Herron JE, *et al*. Separating content-specific retrieval from post-retrieval processing. Cortex, 2017, 86: 1–10. [doi: 10.1016/j.cortex.2016.10.003]

4 Kashino K, Kurozumi T, Murase H. A quick search method for audio and video signals based on histogram pruning. IEEE Transactions on Multimedia, 2003, 5(3): 348–357. [doi: 10.1109/TMM.2003.813281]

5 Kim KM, Kim SY, Jeon JK, *et al*. Quick audio retrieval Using multiple feature vectors. IEEE Transactions on Consumer Electronics, 2006, 52(1): 200–205. [doi: 10.1109/TCE.2006.1605048]

6 齐晓倩, 陈鸿昶, 黄海. 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. [doi: 10.3969/j.issn.1000-3428.2011.19.052]

7 Tzanetakis G, Cook P. Music analysis and retrieval systems for audio signals. Journal of the American Society for Information Science and Technology, 2004, 55(12): 1077–1083. [doi: 10.1002/asi.20060]

8 Tian L, Song QH, Lu XS. Information technology and an audio retrieval method based on a novel audience rating system. Advanced Materials Research, 2014, 886: 664–667. [doi: 10.4028/www.scientific.net/AMR.886.664]

9 Haitsma J, Kalker T. A highly robust audio fingerprinting system. Proceedings of the 3rd International Conference on Music Information Retrieval. Paris, France. 2002. 107–115.

[doi: 10.1076/jnmr.32.2.211.16746]
10 王晖楠, 魏娇. 基于人工智能识别的音乐片段指纹检索技术研究. 自动化与仪器仪表, 2019, (5): 119–122, 126.
11 Yao SS, Niu BN, Liu JQ. Audio identification by sampling sub-fingerprints and counting matches. IEEE Transactions on Multimedia, 2017, 19(9): 1984–1995. [doi: 10.1109/TMM.2017.2723846]
12 于云, 周伟栋. 基于压缩感知的鲁棒性说话人识别参数研究. 计算机技术与发展, 2016, 26(3): 18–22. [doi: 10.3969/j.issn.1673-629X.2016.03.005]
13 Son W, Cho HT, Yoon K, *et al*. Sub-fingerprint masking for a robust audio fingerprinting system in a real-noise environment for portable consumer devices. IEEE Transactions on Consumer Electronics, 2010, 56(1): 156–160. [doi: 10.1109/TCE.2010.5439139]
14 Donoho DL. Compressed sensing. IEEE Transactions on Information Theory, 2006, 52(4): 1289–1306. [doi: 10.1109/TIT.2006.871582]
15 李秀梅, 吕军. 基于压缩感知的信号时频表示重构. 计算机系统应用, 2016, 25(7): 176–181. [doi: 10.15888/j.cnki.csa.005239]
16 王蓉芳, 焦李成, 刘芳, 等. 利用纹理信息的图像分块自适应压缩感知. 电子学报, 2013, 41(8): 1506–1514. [doi: 10.3969/j.issn.0372-2112.2013.08.009]
17 University of Iowa Electronic Music Studios. University of Iowa musical instrument samples. http://theremin.music.uiowa.edu/MIS.html.
18 Jia MS, Yang ZY, Bao CC, *et al*. Encoding multiple audio objects using intra-object sparsity. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(6): 1082–1095. [doi: 10.1109/TASLP.2015.2419980]
19 叶蕾, 杨震, 王天荆, 等. 行阶梯观测矩阵、对偶仿射尺度内点重构算法下的语音压缩感知. 电子学报, 2012, 40(3): 429–434. [doi: 10.3969/j.issn.0372-2112.2012.03.003]