ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 5,Sep. 2016,pp. 1020－1027
DOI:10.16337/j.1004-9037.2016.05.020

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种基于指纹因子的鲁棒音频检索方法

林　静[1]　杨继臣[2]　张雪源[2]　李新超[2]

(1.茂名职业技术学院机电信息系,茂名,525000; 2.华南理工大学电子与信息学院,广州,510641)

摘　要: 针对基于内容的音频检索中由于噪声造成的查找失败问题,本文提出了一种对噪声鲁棒的基于音频指纹因子的音频特征提取算法和一种半监督的音频字典训练算法,以提高噪声下音频检索的精度。本文方法从Mel谱中提取音频指纹,利用非负矩阵分解算法将指纹分解为对噪声鲁棒的频率因子和时间因子作为特征。同时通过提出的半监督音频字典训练算法进行音频字典训练,本文方法使用音效集计算基本音效的分布空间作为初始字典,在量化数据的同时动态更新字典以实现对数据的准确描述。实验结果表明,在低信噪比条件下本文提出的算法的平均查询精度明显高于其他算法。

关键词: 音频检索;音频指纹;非负矩阵分解;音频字典;倒排索引

中图分类号: TN912.3　　**文献标志码:** A

Robust Audio Retrieval Method Based on Fingerprint Factors

Lin Jing[1], Yang Jichen[2], Zhang Xueyuan[2], Li Xinchao[2]

(1. Department of Mechanical and Electrical Information, Maoming Vocational and Technical College, Maoming, 525000, China; 2. School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China)

Abstract: A noise-robust fingerprint-factor-based audio feature and a semi-supervised audio dictionary training algorithm are proposed to fill up the deficiency caused by noise in content-based audio retrieval. The proposed method extracts audio fingerprint from Mel spectra and utilizes non-negative matrix factorization to factorize fingerprint into noise-robust spectral factor and temporal factor as features. Also an semi-supervised audio dictionary training algorithm is proposed. It uses an audio effect set to calculate the distribution of basic sound effects as initialized dictionary. The quantization is conducted while the dictionary is dynamically updated at the same time to better characterize data. The experimental results show that under low signal-to-noise ratio (SNR), the proposed method significantly improves the average precision compared with other algorithms.

Key words: audio retrieval; audio fingerprint; non-negative matrix factorization; audio dictionary; inverted index

引　　言

随着互联网上多媒体数据的爆炸式增长,基于内容的多媒体检索在数据库管理、数据查找和侵权检

基金项目: 国家自然科学基金(61301300)资助项目。
收稿日期: 2015-11-27;**修订日期:** 2016-01-20

测有广泛的应用前景,因此吸引了大量工程、学术机构对其进行研究[1]。由 NIST(National institute of standards and technology)主办的 TRECVID 视频检索测评[2]自 2008 年起将基于内容的检索加入了测评中。目前的多媒体检索研究主要针对视频和图像[3],但音频作为多媒体重要的组成部分,近些年相关研究显著增多[4-12]。由于实际应用中用户获取的音频片段容易受到环境噪声和传输噪声的干扰,噪声的存在可能会导致查询样例和原始数据的失配造成检索失败。针对该问题,近些年的研究主要从鲁棒音频指纹和索引结构[5-12]两方面结合以提高检索匹配的准确率。由 Philips Research 机构提出的噪声鲁棒音频指纹提取算法[5]中,以二进制编码音频帧内相邻子带的能量大小关系作为音频指纹以减弱噪声的影响。Shi 等[6]通过统计纯净音频和噪声污染音频语谱图中的标志点之间的距离,利用最近相邻点估计方法去除噪声造成的标志点偏移。Malekesmaeili 等[7]提出了一种基于色度(Chroma)的音频指纹提取方法,他们使用图像分析方法分析音乐的语谱图,利用针对失真图像的循环位移和尺度变换等算法来弥补音乐中噪声造成的音高偏移和节奏改变。提高检索速度相关的方法主要可分为两类:加速顺序搜索速度和建立索引结构。顺序搜索的代表工作是日本 NTT[8,9]利用音频信号响度直方图以动态步长减少匹配次数。基于索引结构的研究工作主要将文本检索和视频检索的算法应用到音频上,典型算法通常将音频指纹或者特征向量量化后建立倒排索引[10],或者利用哈希映射[11]实现快速查找。张雪源[12]等提出了一种基于倒排文档的检索算法,该算法利用多层次的音频分割方法,将连续音频流切割成内容相似的短时音频段,并利用事先训练好的音频字典将音频段内的音频特征超向量量化为音频字,并通过建立倒排索引表实现快速检索。上述方法均存在以下不足:(1)音频指纹或者短时特征均针对某一音频帧,不能反映音频的时序特性,在检索阶段进行帧一级的匹配时会产生大量的虚警,影响检索效率;(2)矢量量化使用的字典是通过聚类得到的均值向量,描述的是特征空间的一个点,不能够描述噪声造成的偏差;(3)矢量量化的码书固定,哈希映射使用的哈希方程和桶的大小也事先决定好,这些码书和映射结构一旦生成将不能更改,不能根据数据特点进行扩充和更新;(4)码书和哈希方程的训练通常利用数据库中的全部数据,希望找到针对该数据库的全局最优解,但当数据量较大,该训练过程将非常耗时,如刘巍[13]利用并行运算函数库 MPICH 针对 10 000 幅高分辨率图片训练出的 5 万维的视觉字典耗时长达 3 个月。

针对以上问题,本文提出了一种对噪声鲁棒的音频指纹因子,该指纹因子通过对长时 Mel 频谱指纹进行因子分解,得到音频片段内的显著频率因子和时间因子,分别描述了该长时片段内最显著的频率分布信息和时序信息。由于显著频率成分能量通常大于噪声的能量,并且有用声音信号的频率分布较噪声更具规律性,因此得到的因子主要描述了有用声音信号的信息,从而对噪声具有鲁棒性。另一方面,为了训练一个可以对音频特征进行量化的音频字典,本文提出了一种基于多维高斯分布的音频字典和半监督的字典训练算法。基于高维高斯分布的音频字典较传统量化字典既包含了均值信息,同时利用方差信息描述了由于噪声造成的波动。半监督的音频字典训练算法利用一个完备的音效库初始化各类音效的分布空间,在量化时实时更新字典使得音频字典可以描述任意复杂音频的特性,使得该字典具有良好的扩充性和适应性。

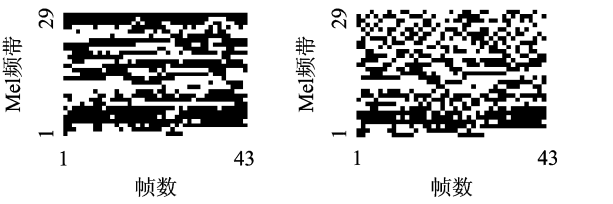
1　音频指纹因子

音频指纹是指从音频中提取的音频频带能量的紧致表示,可用于音频分类、相似音频检索等[5,7]。传统的音频指纹 F 在 Mel 频域以 1 和 0 编码同一帧内相邻频带能量的大小关系,即有

$$F(s,n)=\begin{cases}1 & m(s+1,n)>m(s,n)\\ 0 & \text{其他}\end{cases} \tag{1}$$

式中:$F(s,n)$和 $m(s,n)$分别为第 n 帧第 s 个频带的音频指纹和频带能量,且 $s\in\{1,2,\cdots,S\}$,$n\in\{1,2,\cdots,N\}$。其中以 1 和 0 编码是为了增加音频指纹对噪声的鲁棒性,使得噪声能量较小时不会影响指纹

的分布。但是当原始音频混入噪声能量较大时，噪声对各频带都产生明显影响，音频指纹将变得不可辨认，如图 1(a)所示是 500 ms 音频的原始指纹，图 1(b)为加入信噪比为 0.2 dB 的白噪声后音频的指纹，两者已完全失配。另一方面，这种音频编码方式以短时帧为基本单位，丢弃了时序信息[14]。



(a) 原始音频的指纹 (a) Original audio print (b) 带噪声频的指纹 (b) Audio print of speech with noise

图 1 电影《杀死比尔》音频片段混合噪声前后的音频指纹

Fig. 1 Audio print of speech segment from movie Kill Bill before and after mixed with noise

针对音频指纹的上述不足，本节提出音频指纹因子。指纹因子是指将音频指纹通过分解算法分解为因子的乘积，由于二维的音频指纹纵坐标为频带，横坐标为帧数，因此分解后的因子分别为频率因子和时间因子，这两个因子分别描述了音频指纹的频率分布信息和时间信息。由于在进行音频指纹分解时利用了整个音频指纹的长时信息，因此能够找到在整个长时片段中的稳定信息。这种基于整个音频指纹的长时分析方法相对于对每一帧的短时方法能够找到在噪声影响下依然稳定的频率分布特点。因此该分解方法得到的因子对噪声更具有鲁棒性。由于音频指纹的值均为 0 或 1，因此是非负矩阵，将非负矩阵分解为因子通常采用非负矩阵分解(Non-negative matrix factorization, NMF)算法[15]，分解形式如下

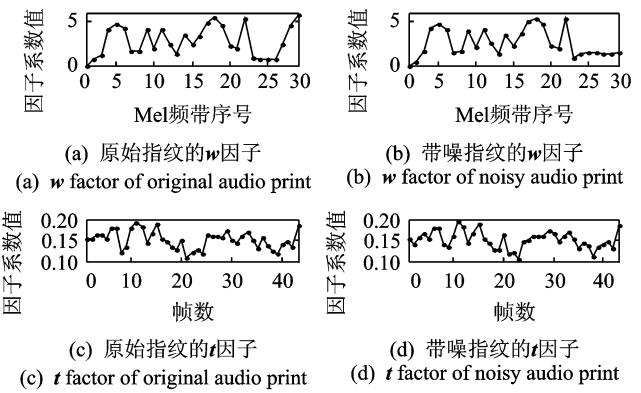
$$\boldsymbol{F}=\boldsymbol{AB} \tag{2}$$

式中：$\boldsymbol{F}\in\mathbf{R}_+^{i\times j}$ 为待分解的非负矩阵；$\boldsymbol{A}\in\mathbf{R}_+^{i\times r}$ 和 $\boldsymbol{B}\in\mathbf{R}_+^{r\times j}$ 分别称为基矩阵和编码矩阵，通常 $r\ll i$ 且 $r\ll j$，该方法广泛用于图像分类、语义识别、DNA 阵列中基因识别和盲信号分离等领域。由于在一个音频片段内由于声源不会迅速变化，因此其频谱分布通常稳定，此外由于噪声的存在，为了减少噪声对因子的影响，本文只从指纹中分解出能量最大的一组因子，即 $r=1$，式(2)退化为

$$\boldsymbol{F}=\boldsymbol{wt}^{\mathrm{T}} \tag{3}$$

式中：$\boldsymbol{w}\in\mathbf{R}_+^{S}$，$\boldsymbol{t}\in\mathbf{R}_+^{N}$ 分别为频率因子和时间因子。频率因子作为基向量描述了该音频片段内稳定的能量分布特性，时间因子作为编码向量描述了在时间片段内频率因子的能量变化规律。

对图 1(a,b)进行因子分解后的因子如图 2(a～d)所示，实心点标示了因子元素的值，图 2(a,c)所示的分别是对图 1(a)分解后的频率因子和时间因子，图 2(b,d)所示的是对图 1(b)分解后的频率因子和时间因子。对比图 2(a,b)可见，两个 $\boldsymbol{w}$ 因子整体上十分相似，只有在最高频的 3 个频点有明显不同，说明该分解方法得到的频率因子对噪声具有良好的鲁棒性。对比图 2(c,d)可见，两个 $\boldsymbol{t}$ 因子整体走势十分相近，只有在第 2 个时间点和第 22 个和第 23 个时间点略有不同。通过对音频指纹进行分解后，从视觉上因子对信号的描述准确性远远超过了原始指纹，噪声对因子的影响只出现在了个别的频点和时间点，因此该方法能够鲁棒地描述信号，将音频特征向量表示为拼接后的因子，即$[\boldsymbol{w}^{\mathrm{T}}\ \boldsymbol{t}^{\mathrm{T}}]$。



(a) 原始指纹的$\boldsymbol{w}$因子 (a) $\boldsymbol{w}$ factor of original audio print (b) 带噪指纹的$\boldsymbol{w}$因子 (b) $\boldsymbol{w}$ factor of noisy audio print

(c) 原始指纹的$\boldsymbol{t}$因子 (c) $\boldsymbol{t}$ factor of original audio print (d) 带噪指纹的$\boldsymbol{t}$因子 (d) $\boldsymbol{t}$ factor of noisy audio print

图 2 原始指纹和带噪指纹分解后的指纹因子

Fig. 2 Factors decomposed from original and noisy audio print

2 音频字及音频字典训练

直接使用高维的音频特征进行匹配会由

于“维数灾难”产生巨大的计算量。音频字可以将高维的音频特征进行量化,为了获得音频字的码本,传统方法通过 K-means 对所有音频特征进行聚类,每一个聚类中心当作一个码字,距离通常采用欧氏距离。本文使用高斯概率密度函数作为音频字,描述音频特征的统计分布,利用方差增强对噪声的鲁棒性。为使音频字典能实现对数据的准确描述,且具有良好的扩充性和适应性,本文提出一种半监督的音频字典训练方法。音效是复杂音频事件的基本构成单元,本文使用音效集初始化音频字典,在后续更新阶段,当现有字典不能够准确描述某复杂音频事件时,新建字典条目扩充字典描述范围。

2.1　音频字典

音频字使用 D 维高斯概率密度函数表示为 $\boldsymbol{W}=\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$,其中 $\boldsymbol{\mu}\in\mathbf{R}^D$ 为音频字的均值向量,$\boldsymbol{\Sigma}\in\mathbf{R}^{D\times D}$ 为协方差矩阵。音频特征 $\boldsymbol{x}$ 和音频字 W 的相似度通过后验概率计算,有

$$p(\boldsymbol{x}\mid\boldsymbol{W})=\frac{1}{(2\pi)^{D/2}\mid\boldsymbol{\Sigma}\mid^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\tag{4}$$

音频字典为所有音频字的集合,一个包含有 K 个音频字的字典 $\mathrm{Dic}=\{W_1,W_2,\cdots,W_K\}$。

2.2　音频字典训练

音频字典的初始化和更新的流程图如图 3 所示。首先为每一种音效类型训练高斯混合模型(Gaussian mixture model, GMM),其模型混合度通过贝叶斯准则决定,并将所有 GMM 中的每个单高斯分量保存为单个音频字,合并相似的分布后作为初始音频字典;在索引阶段,通过计算音频字对音频特征的描述效果,动态更新字典。

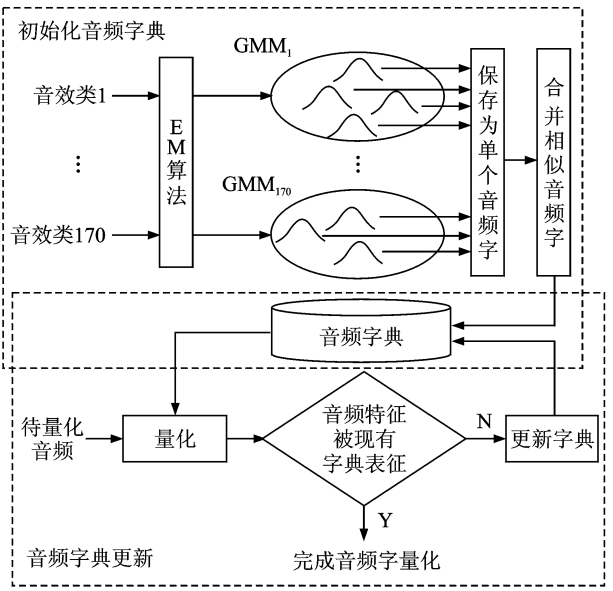


图 3　音频字典初始化和更新流程图

Fig. 3　Flowchart of audio dictionary initialization and updating

2.2.1　初始化

初始化字典所使用的音效库为 Digital Juice Sound FX Library I[16],该音效库是电视台、影视公司和广告公司等专用的专业影视后期音效素材库,涵盖了 170 种常见音效,包含 11 500 个样本。将每一个音效类内所有样本利用最大期望(Expectation maximization, EM)算法训练出一个 GMM,最佳的高斯混合数通过最大化贝叶斯分数(Bayesian information criterion, BIC)得到,该贝叶斯分数同时权衡了模型和数据的相符程度以及模型的复杂度[17],即有

$$\mathrm{BIC}=\log p(X\mid\lambda)-\frac{n_p}{2}\log(N)\tag{5}$$

式中:N 为训练集 X 中训练样本数;λ 为 GMM;n_p 为模型中自由参数的数量。一个 GMM 有 C 个高斯分量,每个分量维数为 D 时,$n_p=(C-1)+2CD$。BIC 值包括两部分,前一部分度量了当前模型对观测值的拟合程度,后一部分为模型的复杂度惩罚因子。BIC 值越大,表示模型拟合度越高同时模型复杂度越低。通过分别计算混合度 C 为 2,4,8,…,128 时的不同 BIC 值,选取最大 BIC 值对应的混合度为模型最优混合度。每个 GMM 模型可表示为

$$\lambda=\{\pi_i,\mu_i,\boldsymbol{\Sigma}_i\}\qquad i=1,2,\cdots,C\tag{6}$$

式中：$\boldsymbol{\Sigma}_i$ 为协方差矩阵采用对角阵。

为每个音效类训练好 GMM 后，丢弃权重 π_i，将每一个高斯分量单独保存为一个音频字。由于某些音效之间有可能较为相似，比如小提琴和大提琴在声学特性上的重叠会造成某些分布的重叠。通过计算每一对高斯分布 $W_i=\{\mu_i,\Sigma_i\}$ 和 $W_j=\{\mu_j,\Sigma_j\}$ 之间的 KL 距离[18]，有

$$KL(W_i \,||\, W_j)=\frac{1}{2}\left(\mathrm{tr}(\Sigma_j^{-1}\Sigma_i)+(\mu_j-\mu_i)^{\mathrm{T}}\Sigma_j^{-1}(\mu_j-\mu_i)-\ln\left(\frac{\det\Sigma_i}{\det\Sigma_j}\right)\right) \tag{7}$$

将所有小于某门限的高斯合并为同一分布 $\widetilde{W}=\{\tilde{\mu},\tilde{\Sigma}\}$，有

$$\tilde{\mu}=\frac{1}{2}(\mu_i+\mu_j) \tag{8}$$

$$\tilde{\sigma}_u^2=\frac{1}{4}(\sigma_{i,u}^2+\sigma_{j,u}^2) \tag{9}$$

式中：σ_u 为协方差矩阵 $\boldsymbol{\Sigma}$ 对角线上第 u 个元素。

随后计算每个音频字的先验概率。使用初始化的字典 $\mathrm{Dic}=\{\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_K\}$ 对所有训练数据进行量化，每一个特征向量的量化结果为最大后验概率对应的音频字，即

$$id=\arg\max_{1\leqslant k\leqslant K}p(x\mid\boldsymbol{W}_k) \tag{10}$$

并且记数据库中量化为音频字 $\boldsymbol{W}_k$ 的特征向量个数为 y_k，则 $\boldsymbol{W}_k$ 的先验概率估计为

$$Pr(\boldsymbol{W}_k)=y_k\Big/\sum_{l=1}^{K}y_l \tag{11}$$

2.2.2 字典更新

首先，将音频中的某一个音频特征向量 $\boldsymbol{x}_i$ 量化为 id_i，根据贝叶斯准则并假设各特征的先验概率服从均匀分布，有

$$id_i=\arg\max_{1\leqslant k\leqslant K}p(\boldsymbol{W}_k\mid\boldsymbol{x}_i)=\arg\max_{1\leqslant k\leqslant K}\frac{p(\boldsymbol{x}\mid\boldsymbol{W}_k)\mathrm{Pr}(\boldsymbol{W}_k)}{Pr(\boldsymbol{x}_i)}=\arg\max_{1\leqslant k\leqslant K}p(\boldsymbol{x}_i\mid\boldsymbol{W}_k)Pr(\boldsymbol{W}_k) \tag{12}$$

通过计算似然比 Λ_i，即

$$\Lambda_i=\log\frac{p(\boldsymbol{x}_i\mid\boldsymbol{W}_{id_i})Pr(\boldsymbol{W}_{id_i})}{1/K\sum_{k=1}^{K}p(\boldsymbol{x}_i\mid\boldsymbol{W}_k)Pr(\boldsymbol{W}_k)} \tag{13}$$

式中：分子部分为最大的似然度，分母部分相当于一个权值全部为 $1/K$ 的全局背景模型，该模型衡量了音频特征和所有音频字的平均相似度，当 $\boldsymbol{x}_i$ 包含在音频字 W_{idi} 的分布时，Λ_i 会显著大于 $\boldsymbol{x}_i$ 不包含在任何音频字的情况。因此，通过比较 Λ_i 和一个预设门限 TH，可以确定现有音频字典能否表征当前音频特征。

该音频特征的 ID 最终确定为

$$id_i=\begin{cases}id_i & \Lambda_i\geqslant\mathrm{TH}\\ K+1 & \text{其他}\end{cases} \tag{14}$$

对连续被标记为 $K+1$ 的段落 $X=\{x_t,x_{t+1},\cdots,x_T\}$，为其建立新的音频字 $W_{K+1}=\{\mu_{K+1},\Sigma_{K+1}\}$，有

$$\mu_{K+1}=\frac{1}{T-t+1}\sum_{i=t}^{T}x_i \tag{15}$$

$$\Sigma_{K+1}=\frac{1}{T-t}(X-\mu_{K+1})(X-\mu_{K+1})^{\mathrm{T}} \tag{16}$$

将新条目 W_{K+1} 加入到字典中，并且更新字典的大小 $K=K+1$ 后，继续量化下一个音频特征。

3　实验结果及分析

3.1　实验条件及评价指标

为了验证本文算法的有效性,音频特征经过量化后,作为与文本检索中词项相似的一维单元,采用文献[12]的方法建立倒排索引并进行检索,与经典的音频指纹提取算法[5],以及文献[12]的倒排索引检索算法作比较。

本文使用 TRECVID 2007/8/9[2] 多媒体数据库中 385 h 的多媒体视频,只保存其中的音频部分为待检索的数据库。采样率统一为 22 kHz,短时帧长 23 ms,帧移为 11.5 ms,音频指纹的长度为 500 ms,从每个音频指纹提取的音频特征维数为 72 维。查询的条目通过从任意音频中任意位置截取不同长度的音频段产生。为了检测算法对噪声的鲁棒性,并模拟现实中用户可能的编辑手段,对查询条目分别混合 3 种类型噪声:白噪声、语音和音乐。这 3 种噪声分别来自语音文件编辑软件 CoolEdit 产生的 50 段白噪声,语音识别数据库 TIMIT[19] 中随机选取的 50 条语音,Digital Juice MusicBox Collection 1[16] 中随机选取的 50 首歌曲。本文所使用服务器有 8 核 CPU (2.13 GHz),32.0 GB 内存,64 位 Windows 7 操作系统,以 Matlab R2012a 为仿真平台。

当检索结果和查询的重叠超过 90%时,认为该结果是"匹配结果"。评价指标如下:(1)排名第 1 的检索结果即为匹配结果的查询数占所有查询数的比例(Precision in top 1,P1);(2)排名前 10 的检索结果中包含匹配结果的查询数占所有查询数的比例(Precision in top 10,P10)。P1 或 P10 的值越高表示系统能准确找到匹配结果并且排在靠前位置[12]。

3.2　实验结果及分析

本文对各种算法分别提交 3 000 次长度为 10 s 的查询,结果如图 4 所示。由图可见,当没有混入噪声时,本文方法和文献[12]的方法均能实现 80%以上的 P1,略高于文献[5]的方法。随着信噪比的降低,各种方法的性能均有不同程度的降低。文献[12]的方法性能下降最明显,当信噪比为 0 dB 时各噪

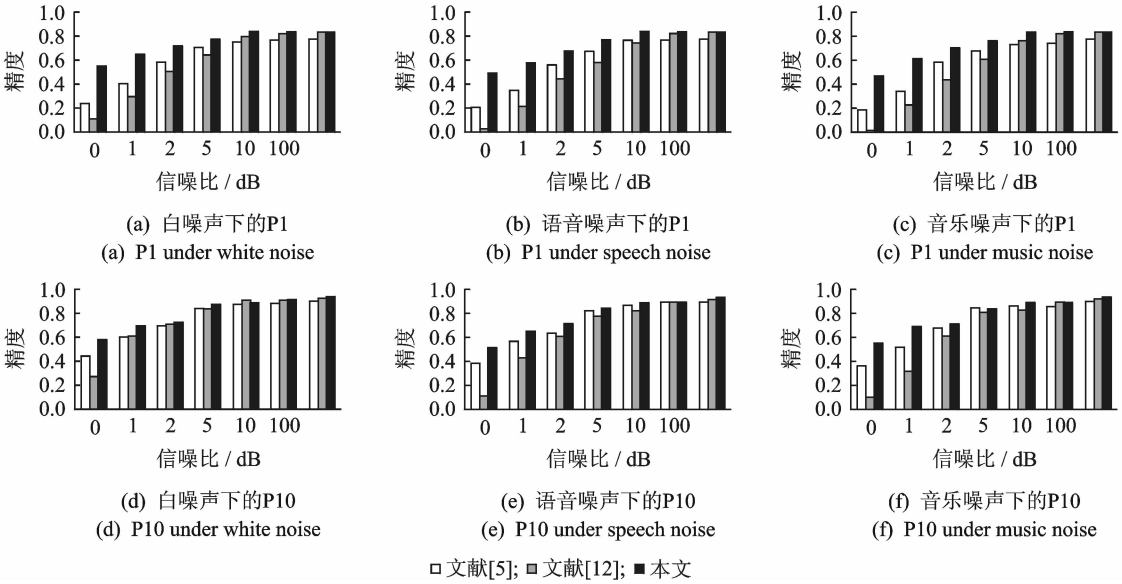


图 4　检索算法精度

Fig. 4　Precision of retrieval algorithm

声条件下的平均 P1 已降至 10%以下,此时,文献[5]的方法的平均 P1 为 20.42%,本文提出的方法性能最好,在各噪声条件下的平均 P1 为 51.79%,比文献[5]的方法性能上有显著提高。其原因在于文献[12]中提出的方法适合于完全一致的音频检索,所提特征不具备噪声鲁棒性,因此随着噪声能量增加性能显著下降,而本文所提出的方法兼具鲁棒性和时序性,因而实现了最优的性能。P10 和 P1 类似,随着信噪比降低,各种方法性能不同程度降低。但是即使在 0 dB 的条件下,本文所提出的方法仍能实现平均 55.09%的 P10,比文献[5]和文献[12]分别提升 16.31%和 39.44%。此外,不同种类的噪声对性能造成的影响不同,对于本文提出的系统,白噪声造成的影响最小,而语音和音乐造成的影响较大,其原因在于白噪声的频谱分散在各个频域,而语音和音乐的频谱相对集中,当信噪比为 0 dB(原始音频和噪声功率相同)时,语音或音乐可能产生与原始音频谱峰能量相当的噪声谱峰。由于本文所提出的音频指纹依赖于对谱分布的估计,谱能量集中的噪声类型对本文提出的方法影响较大。

除精确性外,本文还评估了不同查询方法的速度,分别使用 5 s 和 15 s 两种查询长度评估在信噪比为 0,1 和 10 dB 时的检索用时。对每种方法,提交 5 000 次查询,检索系统响应时间如表 1 所示。由表 1 可以看出,查询长度越长,系统响应时间越久。当查询长度为 5 s 时,本文所提出的方法查询所需时间略高于其他方法,这是由于本文方法使用的提取特征和量化时的计算量高于其他方法,本文所提出的方法平均需要 1.073 2 s 完成一次查询,超过文献[5]和文献[12]方法的 20.58%和49.65%。当查询长度为 15 s 时,3 种方法的查询时间几乎相同,分别为 2.891 7,2.482 8 和 2.645 6 s,这是由于当查询长度增加时,主要耗时在查询索引表上,因此本文提取特征和量化时较为耗时的不足不会对整体查询用时造成显著影响。

表 1 不同检索算法耗时

Tab. 1 Time-consuming of different retrieval algorithms

查询长度/s	信噪比/dB	文献[5]/s	文献[12]/s	本文/s
5	0	0.800 2	0.726 2	1.078 7
	1	0.907 6	0.724 8	1.126 2
	10	0.962 3	0.700 4	1.014 6
15	0	2.787 6	2.570 2	2.563 8
	1	2.928 9	2.213 2	2.595 7
	10	2.958 8	2.665 2	2.777 3

4 结束语

以基于内容的音频检索为应用背景,本文提出了一种基于音频指纹因子的特征提取算法,从 Mel 谱中提取音频指纹,并将其分解为对噪声鲁棒的频率因子和时间因子作为特征。通过本文提出的半监督的音频字典训练算法,使用音效集计算基本音效的分布空间作为初始字典,量化数据的同时动态更新字典,可以实现对数据的准确描述。采用 TRECVID 多媒体数据库的数据进行了改进算法性能实验,结果表明:本文的音频指纹因子特征对白噪声、语音和音乐噪声均具有鲁棒性。在可接受的检索时间内,所提出的检索方法在低信噪比的条件下,其检索精确度明显高于经典的音频指纹算法及倒排序检索算法,在查询音频片段时长较短时,本文算法耗时稍长;在查询音频片段时长大于 15 s 时,本文算法和另外两种算法检索耗时相当,再提高检索精度的情况下几乎没有降低检索效率。本文所提出的特征主要针对低信噪比环境,只使用了一组因子,但是对于高信噪比环境下,可以提取更多组因子以更精确地描述信号特性。在接下来的工作中,希望提出一种依据信噪比自适应地提取不同数量因子组合的算法,以进一步提升在各信噪比环境下的精确度。

参考文献:

[1] Weng L, Amsaleg L, Morton A, et al. A privacy-preserving framework for large-scale content-based information retrieval [J]. Information Forensics and Security, IEEE Transactions on, 2015,10(1):152-167.

[2] Awad G, Michel M, Joy D, et al. Evaluation campaigns and TRECVid[EB/OL]. http://trecvid.nist.gov/, 2015-05-01.
[3] Wang Y, Mohammed B, Bashar T. Near-duplicate video retrieval based on clustering by multiple sequence alignment[C]// Proceedings of the 20th ACM International Conference on Multimedia. Nara, Japan: ACM, 2012:941-944.
[4] Huurnink B, Snoek M, de Rijke M, et al. Content-based analysis improves audiovisual archive retrieval[J]. IEEE Transactions on Multimedia, 2012,14(4):1166-1178.
[5] Haitsma J, Kalker T. A highly robust audio fingerprinting system[C]// 3rd International Conference on Music Information Retrieval. Paris, France: IRCAM, 2002:107-115.
[6] Shi Jianhua, Yu Xiaoqing, Wang Yunhui, et al. Noise reduction based on nearest neighbor estimation for audio feature extraction[C]// International Conference on Audio, Language and Image Processing. Shanghai, China: the Institute of Electrical and Electronics Engineers Press, 2012:768-771.
[7] Malekesmaeili M, Ward K. A novel local audio fingerprinting algorithm[C]// 14th International Workshop on Multimedia Signal Processing. Banff, Canada: the Institute of Electrical and Electronics Engineers Press, 2012:136-140.
[8] Kimura A, Kashino K, Kurozumi T, et al. A quick search method for audio signals based on a piecewise linear representation of feature trajectories[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008,16(2):396-407.
[9] 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):348-357.
[10] Zhao L, Wu X, Ngo W. On the annotation of web videos by efficient near-duplicate search[J]. IEEE Transactions on Multimedia, 2010,12(5):448-461.
[11] Li M, Sun R, Han J. Fast audio retrieval using symbolized LSH address based on p-stable distribution[J]. Journal of Information & Computational Science, 2012,9(5):1265-1272.
[12] 张雪源, 贺前华, 李艳雄,等. 一种基于倒排索引的音频检索方法[J]. 电子与信息学报, 2012,34(11):2561-2567.
Zhang Xueyuan, He Qianhua, Li Yanxiong, et al. An inverted index based audio retrieval method[J]. Journal of Electronics and Information Technology, 2012,34(11):2561-2567.
[13] 刘巍. 基于内容的同源音频和视频检索[D]. 北京:北京邮电大学, 2011.
Liu Wei. Content-based video and copy detection[D]. Beijing: Beijing University of Posts and Telecommunications, 2011.
[14] Bruce C, Donald M, Trevor S. Search engines: Information retrieval in practice[M]. Upper Saddle River: Addison-Wesley, 2010:22-23.
[15] 孙健,张雄伟,曹铁勇,等. 基于卷积非负矩阵分解的语音转换方法[J]. 数据采集与处理,2013,28(2):141-148.
Sun Jian, Zhang Xiongwei, Cao Tieyong, et al. Voice conversion based on convolutive nonnegative matrix factorization[J]. Journal of Data Acquisition and Processing, 2013,28(2):141-148.
[16] Digital Juice, Inc. Audio effect library[EB/OL]. http://www.digitaljuice.com/, 2012-01-05/2014-12-06.
[17] Lee Y, Lee Y, Lee J. The estimating optimal number of Gaussian mixtures based on incremental k-means for speaker identification[J]. International Journal of Information Technology, 2006,12(7):13-21.
[18] Shlens J. Notes on Kullback-Leibler divergence and likelihood[J]. Eprint ArXiv, 2014,14(4):2000-2004.
[19] Garofolo John. TIMIT: Acoustic-phonetic continuous speech corpus[EB/OL]. https// www.ldc.upenn.edu/, 1993-09-02/2015-07-02.

作者简介:

林静(1982-),女,讲师,研究方向:音频信号处理,E-mail: Linjing80615 @ 163.com。

杨继臣(1980-),男,副研究员,研究方向:音视频信号处理。

张雪源(1987-),男,博士研究生,研究方向:音视频信号处理。

李新超(1980-)男,博士,研究方向:智能优化与信号处理。