基于功率谱特征的音频指纹实现

鲁明明 张 晖 沈庆宏
(南京大学电子科学与工程学院 南京 210046)

摘 要：音频指纹是指一段可以表示音频信号重要声学特征的摘要，在音频信号的识别、安全验证、完整性校验等方面有广泛的用途。提出了一种基于信号功率谱特征的音频指纹编码方法，通过对音频信号做 STFT 计算功率谱，提取信号时频域的功率极值点作为特征并利用其空间结构做哈希计算得到音频指纹。经过实验验证，该算法对时长为 2 s 以上的音频样本可以达到 95%以上的识别率，对 AWGN 噪声具有较强的抑制能力，具有算法简单、识别快速准确、鲁棒性高、抗噪声能力强等特点。

关键词：音频指纹；短时傅里叶变换；峰值滤波
中图分类号：TP391.42 **文献标识码**：A **国家标准学科分类代码**：510.4040

Realization of audio fingerprint based on power spectrum feature

Lu Mingming Zhang Hui Shen Qinghong
(School of Electronics Science and Engineering, Nanjing University, Nanjing 210046, China)

Abstract: Audio fingerprint is a digital digest of the signal which represents the most significant characteristics and is widely used in identification、security validation and integrity verification of audio signal. This paper presents a method of generating audio fingerprint based on the power spectrum characteristics of the signal. After calculating STFT of the input sample and extracting the peak of the power spectrum, we use the space structure of the peaks as input of the hash function of the audio fingerprint algorithm. As experiment shows, it could achieve a recognition accuracy of over 95% when the audio sample lasts more than 2 s and shows good resistance to AWGN noise. It is easy to implement、robust enough and exhibits fast and accurate identification ability.

Keywords: audio fingerprint; short-time Fourier transform; peak filter

1 引 言

随着信号压缩技术的进步以及数字信号本身便于存储、传输和加工的优点，数字音频内容大量出现，与此同时，对海量数字音频内容进行检索和识别也变得十分困难。由于音频指纹(audio fingerprinting)可以快速高效地识别和检索音频内容，该技术成为国内外学者研究的热点问题[1]，目前已经出现了一些商业产品。2004 年美国 Gracenote 公司结合其“波形指纹信息数据库”和 Philips 公司的音频指纹识别技术[2]推出了可通过手机使用的乐曲识别软件“Gracenote Mobile”，我国北京酷我科技有限公司也基于其开发的音频指纹技术建立了一套大型指纹数据库系统供广大互联网用户使用。

大多数音频指纹提取算法都是将音频信号分成互相重叠的帧，利用每一帧的特征构造音频指纹。常用的特征有傅里叶系数、迈尔倒谱系数、频谱平滑度、频谱尖锐度、小波系数、频带正规化矩等。这些方法强调通过对音频信号做 DSP 处理并应用现有的信号分析算法得到其指纹特征，存在算法复杂度大，实时性差等问题。

音频指纹要反映音频信号的时域和频域特征，因此本文尝试从短时傅里叶变换(STFT)入手，利用功率谱特征构造音频指纹。通过分析信号的时频域和空间结构特性，提出一种新的音频指纹生成和音频信号识别算法，并研究了样本长度、噪声干扰对识别时间、识别准确率的影响，给出了实验结果。

本文提出的音频指纹算法具有噪声不敏感，识别准确快速，算法通用易于实现等优点，具有较强的理论和工程价值，可以方便的应用在 PC 端和手机端实现音频内容识别和检索。

2 音频指纹算法

音频指纹提取算法主要包括前端处理和指纹建模两个

收稿日期：2015-12

部分。前端处理用于提取音频信号的特征值，指纹建模用来生成最终保存在数据库中的指纹片段，本文音频指纹提取算法的基本流程如图 1 所示。

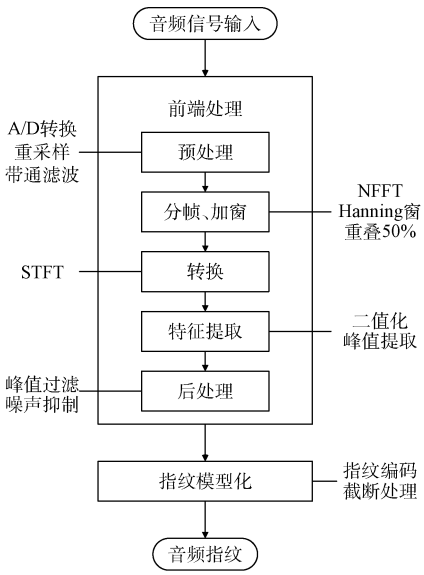

图 1　音频指纹算法流程

2.1　预处理

预处理用来获得解码后的数字音频序列。如果输入的是模拟信号，还需要进行 A/D 转换。预处理过程包含经典的音频信号数字化加工过程。

2.2　STFT

STFT 是一种简单、直观的信号时频表示，基本思想是用一个随时间平移的窗函数 $g(\tau)$对原信号加窗，然后逐段计算其频谱。短时傅里叶变换是能量对时间和频率的二维函数[3]。

对于一给定信号 $x(t)$，其 STFT 可以通过下面公式计算：

$$X(t,f)=\int_{-\infty}^{\infty}x(\tau)g\cdot(\tau-t)\mathrm{e}^{-\mathrm{j}2\pi f\tau}\mathrm{d}\tau \tag{1}$$

对于离散时间信号，同样有如下公式：

$$X(m,\omega)=\sum_{-\infty}^{\infty}x(n)g[n-m]\mathrm{e}^{-\mathrm{j}\omega n} \tag{2}$$

对经过预处理的信号做 STFT，实际上包含了图 1 中的分帧、加窗和转换过程，窗函数移动时就将音频信号分为帧，而加窗则可以消除相邻两帧两端造成的信号不连续性[4]。本文选用了汉明窗来做 STFT，相比矩形窗可以较好的克服频谱泄露[5-6]，相邻两帧的重叠部分设置为帧长度的 50%。根据人的语音音调周期值的变化，帧长度一般取 5～20 ms 比较合适，时间太短时由于信号的能量按照信号波形的细微情况起伏较快，而时间太长时则不能较好反映波形变化的细节，本文提出的方法选择 $NFFT=4\ 096$，对应的帧长度计算公式为：

$$F_t=NFFT/F_s \tag{3}$$

式中：F_s 为音频信号的采样频率，当 $F_s=44\ 100$ Hz 时，可以计算得出 F_t 等于 9.288 ms。

2.3　特征提取

音频指纹首先需要具有足够好的抗干扰能力，即使待识别的信号中混入了较多噪声，也应该能准确的进行识别。音频指纹需要满足听觉相似的声音信号产生基本相同的输出结果，因此提取的特征必须能够反映音频信号的声学特征，目前主要的特征提取方法集中在时域、频域或时频域[7]。通过 STFT 得到的信号功率谱同时包含了信号时域和频域的特征，为了最大限度的消除噪声干扰和信号畸变，选择功率谱中的功率峰值点作为特征生成音频指纹，提取特征点的算法如图 2 所示。

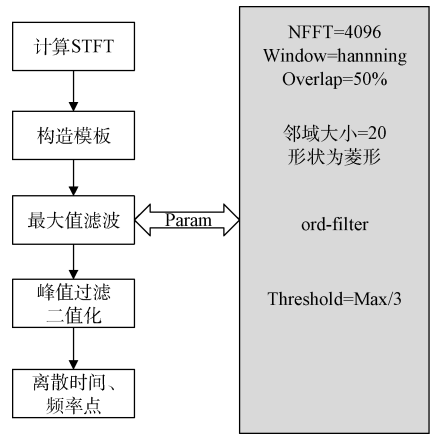

图 2　功率谱特征点提取流程

提取功率谱峰值的过程用到了最大值滤波器，滤波模板（即结构元素）选用了菱形，因为认为每个峰值功率点为时间、频率两个维度邻域的最大值。适当增大模板可以产生更少的指纹，但匹配效果会变差。部分功率点峰值可能产生于一个幅度均较小的邻域（可以认为是噪声，或者不作为音频信号的声学特征），因此必须设定一个阈值，这样有助于提高算法的鲁棒性。阈值较低时产生较多的指纹输出，匹配效果更佳，但指纹存储体积增大，本文中取功率阈值为信号最大功率分贝值的 1/3，即：

$$P_{\mathrm{th}}=P_{\max}/3 \tag{4}$$

图 3 给出了一段音频信号的功率谱和提取的峰值特征点。

不难看出，音频指纹的数量几乎同信号的频带宽度成正比，如果仅对 4 kHz 以下的信号功率谱进行处理，则可以在几乎不影响匹配效果的情况下将音频指纹的数据量降低约 70%～80%，从而节省存储空间并提高检索速度。

一旦提取出功率谱特征点，其算术值将不再重要，因为音频指纹是否相等并不取决于信号在某一时刻某一频率的功率取值是否相等，我们关心的是音频信号在时域和频域的结构特性，功率谱峰值点的取值并不参与指纹的计算过

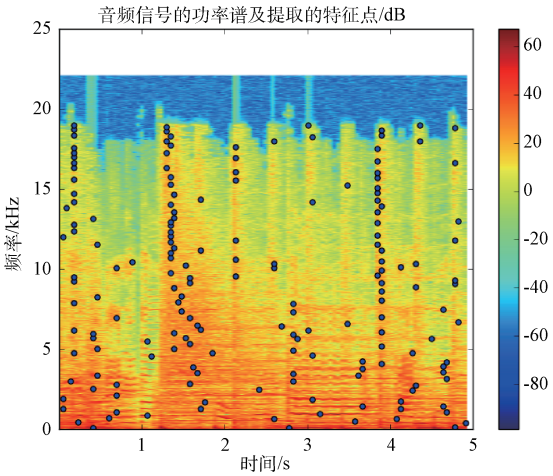

图 3 功率谱峰值点

程,因此该步骤最终的输出结果是功率谱特征点对应的(frequency,time)对。

2.4 音频指纹生成

音频指纹的生成算法很多,但大多数方法的核心都是分帧和后处理。文献[8]通过对分帧后的音频数据提取 300～2 200 Hz 的频谱特征做哈希得到指纹,构造音频指纹使用的数学方法也包括复倒谱[9]、小波分析[10]等。本文利用功率谱中提取的峰值对应的(频率、时间)对计算哈希获得音频指纹,过程如下:

步骤 1:对于序列 S(frequency, time),按照 time 进行算术排序。

步骤 2:初始化关联距离 $L=16$,$i=0$,集合 fingerprint()。

步骤 3:对于 $j=1:L$,如果 $i+j>\text{len}(S)$,算法结束,否则执行步骤 4。

步骤 4:

$F\_\text{base}=S(i)[0]$ $F\_\text{step}=S(i+j)[0]$

$T\_\text{base}=S(i)[1]$ $T\_\text{step}=S(i+j)[1]$

$T\_\text{delta}=T\_\text{step}-T\_\text{base}$

$$hash=md5(F\_\text{base},F\_\text{step},T\_\text{delta})[0:15] \quad (5)$$

将(hash,$T_$base)添加到 *fingerprint*,执行步骤 3。

步骤 5:$i=i+1$, 如果 $i<\text{len}(S)$,执行步骤 3,否则算法结束,返回 *fingerprint*。

本算法计算得到的音频指纹包括时域、频域关联距离计算得到的哈希值以及时间偏移量,在进行指纹匹配时,利用哈希值进行搜索,同时获得待识别音频片段的时间偏移量。从上述计算过程可以看出,hash 函数仅利用音频信号功率谱的空间结构信息,与具体幅值无关,因此具有很好的抗攻击能力,对不同码率编码、重采样、附加噪音、滤波干扰后的音频信号具有很好的识别能力[11]。哈希函数选取了 *md*5 哈希值的前 16 位作为音频指纹,可以有效节省存储空间。

一般需要计算音频指纹的信号可能包含多个声道的信息,依次计算每个声道的音频指纹存入一个集合即可。

3 指纹检索及识别

音频片段的识别过程对应音频指纹的检索和匹配过程。在进行音频信号的识别时,首先用同样的方法计算其音频指纹,得到的是许多(哈希值,时间偏移量)对,依次在数据库中检索每个哈希值,必然会返回许多相同的匹配结果,但其时间偏移量则多数不相同,不同的时间偏移量对应不同的歌曲。对于正确的匹配结果来说,不考虑噪声、信号畸变和干扰存在的情况下,所有匹配得到的指纹对其偏移时间的差值应该都是相等的,该偏移时间差值实际上就是待识别音频片段的开始时间。因此我们只要统计不同时间偏移量差值下可以匹配到的哈希值数目即可,无论是否存在噪声和干扰,在相同时间偏移量差值下匹配哈希值数目最多的歌曲就是识别结果。

4 实验结果

本次实验共选择了 50 首曲目建立数据库,包括纯音乐、自然声音和人声,歌手来自不同时代,不同流派,风格多样。测试平台为 Intel Core i7 3630QM, 8 GB RAM,数据库为 MySQL 5.6.28。

4.1 不同音频片段长度与识别准确率的关系

选取时长为 0.5～5 s 的 10 组音频片段,每组共包括 50 个采样片段,从一首歌排除首尾静音部分后随机截取,图 4 是识别准确率结果。

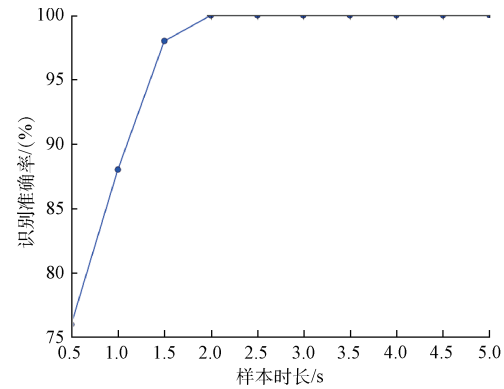

图 4 识别准确率与样本时长关系

可以看出该算法表现良好,对于时长超过 1.5 s 的音频样本,识别准确率可以达到 95%。

4.2 不同音频片段长度与识别速度的关系

待识别样本库与 4.1 节相同,每组时长包括 20 个采样片段,识别时间求平均值。

结果如图 5 所示。可以看出,识别时间和样本大小呈现出明确的线性关联,识别时间大于 1 s 时,识别准确率几乎可以达到 100%,性能良好。

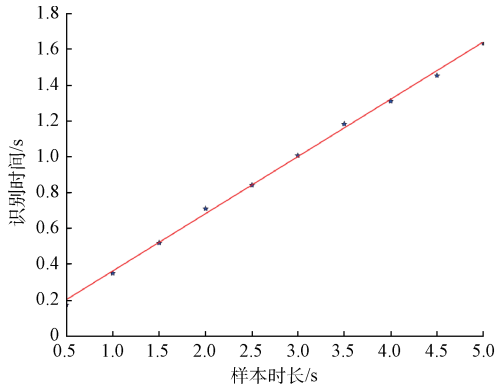

图 5　识别时间与样本时长关系

4.3　存在噪声干扰情况下识别结果

使用的音频样本与 4.2 节相同，分别添加不同信噪比的加性高斯白噪声，比较其识别准确度，结果如图 6 所示，不难看出，通过增加样本时长可以提高低信噪比时的识别准确度，样本时长大于 3 s 时基本可以保持信噪比大于 15 dB时 90%的识别准确率。实际应用场景中的信噪比足以满足这个要求。

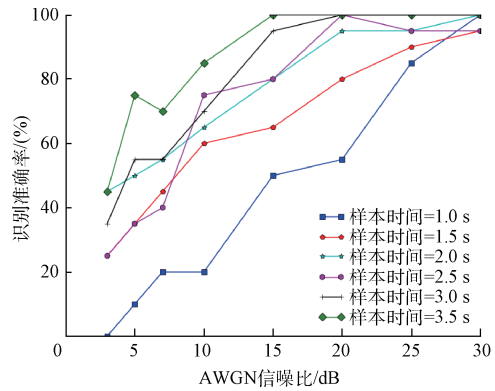

图 6　AWGN 噪声对识别准确率影响

5　结　　论

提出了一种基于功率谱特征的音频指纹实现，利用音频信号时域和频域的结构特征构造指纹，并实现了音频信号的识别。实验结果表明，该方法具有匹配速度快、精度高、抗干扰能力强的优点。

参考文献

[1]　曾柏森. 基于内容的音频检索研究[D]. 成都：西南交通大学，2009：1-11.

[2]　明建成，韩威. 基于音频指纹的压缩域音频识别方法研究[J]. 科学技术与工程，2014，14(16)：83-87.

[3]　李伟，谢华. 短时傅里叶变换在频移键控解调中的应用[J]. 电子测量技术，2011，34(7)：34-36.

[4]　牛犇，任文娟，胡东辉，等. 基于 STFT 的宽带信号时差测量方法[J]. 国外电子测量技术，2011，30(10)：17-21.

[5]　俞一鸣. 时频分析简介及应用[J]. 国外电子测量技术，2015，34(6)：12-15.

[6]　石明江，罗仁泽，付元华. 小波和能量特征提取的旋转机械故障诊断方法[J]. 电子测量与仪器学报，2015，29(8)：1114-1115.

[7]　张鸿博，蔡晓峰，鲁改凤. 基于双窗全相位 FFT 双谱线校正的电力谐波分析[J]. 仪器仪表学报，2015，36(12)：2835-2840.

[8]　张敏，欧阳建权，李泽洲，等. 一种快速的特定音频指纹提取方法[J]. 计算机工程，2010，36(2)：211-213.

[9]　周亦敏，牟同鑫. 采用负倒谱和子串匹配的音频指纹算法研究[J]. 上海理工大学学报，2010，32(3)：277-280.

[10]　龙小保. 使用提升小波和非负矩阵分解的稳健音频指纹[J]. 计算机工程与应用，2013，49(9)：197-199.

[11]　牛宪华，曾柏森，陈思利. 基于频域和时域差分的音频指纹算法研究[J]. 西华大学学报，2014，33(5)：10-15.

作者简介

鲁明明，硕士研究生，山西大同人，主要研究方向为数字信号处理，无线传感器网络。

E-mail：luminghi@hotmail.com