采用Shazam算法的声纹识别系统

司向军1，李冰2

（1.东南大学 无锡分校，无锡，214135；2.东南大学 集成电路学院，南京，210096）

**摘 要：**Shazam算法是一种相对简单成熟的算法，广泛应用于音乐识别系统中。经对比分析和实验论证，其对语音信号的处理也可应用于声纹识别系统。该算法最显著的优点是复杂度低、识别率较高。在声纹识别系统的实际应用环境中，环境噪声所引起的畸变严重影响着声纹识别的性能。因此有必要对语音进行消噪来提高信噪比，提高声纹识别的性能。针对基于Shazam算法的声纹识别主要应用于移动端，考虑到实际应用中背景噪声的影响，本文采用自适应门限端点检测算法进行语音信号增强，实现了基于Shazam算法的声纹识别系统，并对系统进行了实验测试分析。

**关键词：**声纹识别；Shazam算法；端点检测

**Using Shazam algorithm for voiceprint recognition system**

Si Xiangjun1, Li Bing2

(1. Wuxi Branch of Southeast University, Wuxi, 214135;

2. School of Integrated Circuits, Southeast University, Nanjing, 210096)

**Abstract:** Shazam algorithm is a relatively simple and mature algorithm, widely used in music recognition system. The comparison of the speech signal can also be applied to the voiceprint recognition system. The most significant advantage of this algorithm is that the complexity is low and the recognition rate is high. In the actual application environment of the voiceprint recognition system, the distortion caused by the environmental noise seriously affects the performance of the voiceprint recognition. It is necessary to de-noise the voice to improve the signal to noise ratio, improve the performance of voiceprint recognition. Based on the Shazam algorithm, the speechprint recognition is mainly applied to the mobile end. Considering the influence of the background noise in the practical application, this paper uses the improved the adaptive threshold endpoint detection algorithm to enhance the speech signal, and realizes the soundprint based on the Shazam algorithm Identification system, and the system of experimental test analysis.

**Key words:**Voiceprint recognition；Shazam algorithm；Endpoint detection

1 算法概述

* 1. 特征向量提取

Shazam算法[1]提取语音特征的方法比较简单，且抗噪能力强。Shazam算法首先把语音信号从时域变换到频域，然后从频谱中提取能量峰值点，同时记录出现能量峰值时的频率位置，一般情况下，一帧信号经过FFT(Fast Fourier Transformation)后，从频谱信号中提取的能量峰值点所形成的的形状图。

对一段语音进行FFT变换之后看到的只是一系列的频率和幅值，但无法知道这些频率具体在该片段中的位置，同时这些位置信息很重要，故本文中采用另一种技术：滑动窗口。

**作者简介：**司向军，（1991-），男，硕士研究生，E-mail: 1049428980@qq.com；李冰，（1968-），男，教授，博士生导师，E-mail: [bernie\_seu@sina.com](mailto:bernie_seu@sina.com).

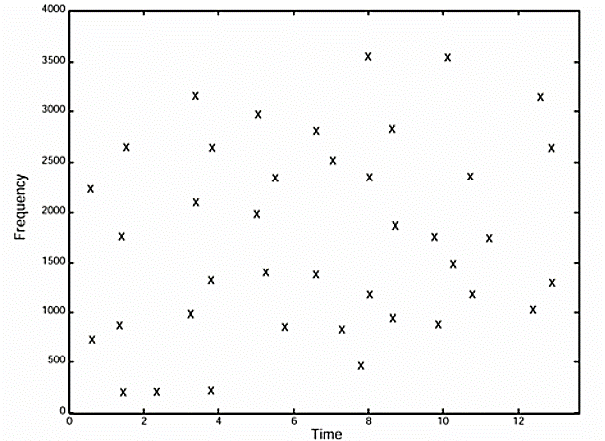


图1(a) 提取的能量峰值点形成的星状图

本文中为满足系统要求，采用双声道、16-bit精度、44100Hz采样。此时1s的数据大小为44100\*2\*2bit，约为176kB。本文中选择4kB当做数据块大小，则每秒钟需要对44个数据块进行FFT变换，这种切分密度已足够满足本文系统要求。本文中构造声纹特征采用每一帧中的最大幅值作为构造音频指纹的基础。由于幅值较大的频率跨度很广，为避免分析整个频谱，本文中将频谱分成多个子带，从每个子带中选择一个频率峰值。子带的范围为：低音子带为30-40Hz、40-80Hz和80-120Hz，中音和高音子带分别为120-180Hz和180-300Hz。每个子带的最大频率就构成了这一帧信号的签名，而这个签名又是整段语音信号指纹的一部分。当提取完语音特征，存储在数据库中作为模板使用，进行声纹识别时，对原始语音进行同样的操作，提取该段语音指纹，然后和模板库中语音指纹作比较，得出识别结果[3]。

1.2匹配算法

在执行匹配之前，需要先从录取的语音样本中提取所有的语音指纹和时间偏移信息，这两点是检索匹配算法的关键之处。语音样本中提取的语音指纹都会和数据库中的语音指纹进行匹配，每一个匹配的语音指纹都会产生一个时间对：样本中的时间和数据库中的时间。根据对应的语音ID将时间进行分类。本文中对语音特征的检索先构建倒排索引，来快速确定语音指纹的位置。倒排索引的关键字是语音指纹，即是从语音片段中提取的峰值点对，对应存储的值是语音指纹所在语音的ID以及语音指纹的第一个锚点(峰值点)所在原始语音片段中的位置，即对应所在语音中的时间偏移。采用特征点对的方法可以增加关键字的数目，从而减少碰撞率。如图1(a)所示，为该检索的倒排索引结构，从该检索结构中可以看出，一个语音指纹中所包含的信息主要有语音ID和所提取的指纹的偏移时间点。

当从语音样本提取的指纹都完成上述匹配后，通过扫描时间对获得正确的语音。每个语音片段对应的所有时间对构成了一个散点图，如果语音和样本匹配，则匹配上的特征会有相近的时间偏移，也即是样本中提取的指纹和数据库中的指纹具有相同的相对时间（这个相对时间就是样本在原始语音中

提取的指纹和数据库中的指纹具有相同的相对时间（这个相对时间就是样本在原始语音中的起始位置）。这样就可以将搜索问题简化为散点图中寻找明显对角线的问题，考虑到问题的特殊性，可以设计一个复杂度为的算法，其中N表示散点图中点的个数。不失一般性，假设对角线的斜率为1.0（录音速度和原始语音速度一致），则样本和正确语音中匹配指纹的时间满足下面的关系式：

 (1.1)

其中表示原始语音中匹配指纹的时间，表示录

图1(b)倒排索引结构

音样本中对应指纹的时间。对于散点图中的每一个点，我们计算

 (1.2)

当录音速度和原始语音速度不一致时，一般情况下录音速度为原始语音速度的0.5—2倍之间，设此因数为，此时样本语音和语音库中匹配指纹的时间应满足下面的关系式：

 (1.3)

然后计算的直方图并扫描获得直方图的最大值即是将时间对的两个时间相减，然后统计相同时间差个数的最大值，求最大值可以通过扫描完成，由于匹配对的原因，散点图通常会非常稀疏。同时每段语音对应的时间差个数通常很少，所以扫描过程会在几微秒内完成，每段语音对应的分即是直方图的最大值，如果直方图最大值很大，那就意味着该语音片段就是要识别的正确语音[4]。

1.3自适应门限端点检测

在声纹识别中，一个关键问题就是如何利用端点检测将语音信号精确的检测出来，为获得准确的识别提供前提。针对背景噪声对传统端点检测算法的影响，在传统的短时平均能量和短时平均过零率的基础上，考虑到本文应用采集的语音段短，可近似认为此段时间内背景噪声为平稳特性，提出一种自适应门限的端点检测算法，利用语音信号前后的静音段获取背景噪声特性作为门限改善在噪声背景下的语音端点检测效果[7]。

由于语音信号是一种非平稳信号，用移动的有限长的窗口进行分帧，并认为在一帧内它是平稳的，一帧内的信号能量值和过零率分别被称为短时能量和短时平均过零率，窗长的选择一般包含1~7个基音周期，一般选取10~30ms的时间作为窗长，设语音信号为短时平均能量定义为

 (1.4)

短时平均过零率定义为

 (1.5)

其中为符号函数，窗函数为

 (1.6)

其中*N*为矩形窗的长。

在传统的语音端点检测中，短时平均能量和短时平均过零率是一个固定的门限值，由于环境的背景噪声不同和外界的干扰差异，使固定的门限在端点检测过程中不能真正的检测到语音的起始点。自适应门限就是为了克服这个问题提出来的。一般而言，有效语音前后都包含一些体现背景噪声和干扰的语音段，把第一帧和最后一帧的短时平均能量和短时平均过零率加上某个经验值来作为检测起点和终点的门限。

1. 基于Shazam算法的声纹识别系统实现方案
   1. 总体设计

本文的系统软件流程图。首先进行对待识别语音进行录音，首先对录取到的音频进行端点检测，主要是为了减小音频中出现的背景噪声，如果是注册，则采用基于Shazam算法的声纹识别系统进行处理，对处理的结果存储在数据库中，作为待识别模板保存，等待下一次音频输入时进行检测；如果是识别，同样的先进行端点检测，再用基于Shazam算法的声纹识别对该段音频进行识别，得出结论。

其中采用Shazam算法的对预处理后的音频进行声纹识别的具体步骤为：

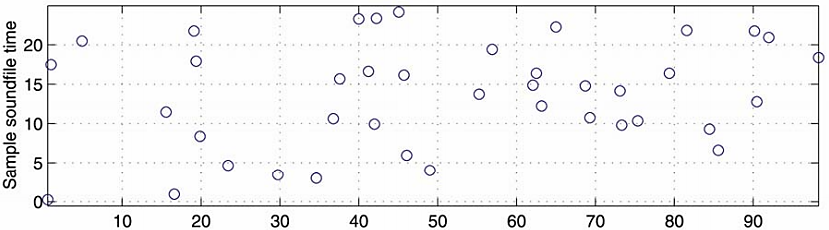
1. 采用滑动窗口对一块语音进行FFT变换，得到语音信号的频域信息；块的大小要满足系统的精度要求。
2. 分别求出低音带30-40Hz、40-80Hz和80-120Hz，中音和高音子带120Hz-180Hz和180Hz-300Hz中最高的频率作为一帧信号的声纹特征点，
3. 存储得到的声纹特征点，并记录出现的时间，用Hash值做为记录，结果存储在数据库中。



图2 系统流程图

1. 实验结果和结论
   1. 实验设计

本文给出基于Shazam算法的声纹识别错误匹配和正确匹配的图解。



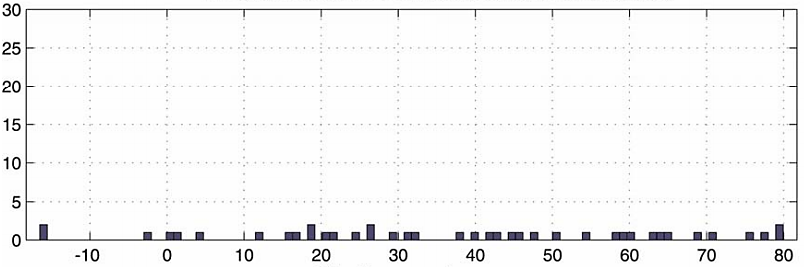
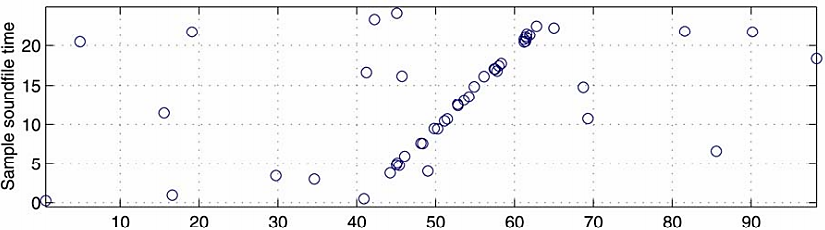


图3(a) 错误匹配样本图



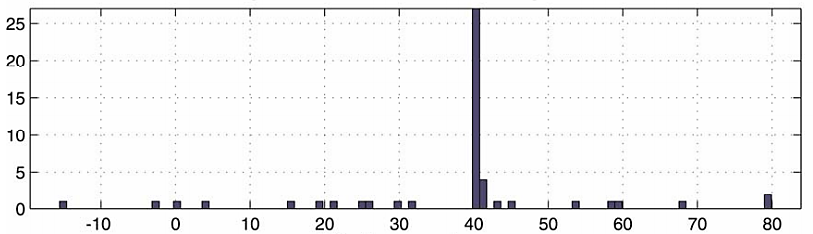


图3(b) 正确匹配样本图

图3(a)展示了一个错误匹配的例子，图3(b)展示了一个正确匹配的例子，若录音样本语音和数据库中原始语音语速不一致，此时图中的斜线的斜率应为0.5—2之间。

为了测试本文系统的性能，在办公室进行声纹识别实验。实验所在办公室内比较安静（背景为58dB（A））。实验中分别采用随机噪声和实际环境噪声对系统的抗噪声性能进行测试。其中随机噪声是用Matlab软件才生的一段高斯随机信号；实际环境噪声是在室外的背景环境噪声，此段噪声包含车辆噪声、行人噪声等。实验中分别将这两段噪声通过音箱播放出来模拟声纹识别实验的背景噪声，实验中录音的时间为5s，说话人以正常的语速说话，说话长度为8~10个字，且内容相同。在识别前，先登记5个模板，测试者分别说登记时的短句来测试系统的性能。

实现分别在安静环境下、随机噪声环境下和实录环境噪声下进行测试，无语音增强时测试结果如表1所示；语音增强下的测试结果如图2所示。

* 1. 结果与讨论

表1中“正确识别”表示能够正确识别出自己登记的模板，“错误识别”表示系统识别到了其他用户的模板，“被拒绝识别”表示没有识别到任何模板。在较为安静的情况下（58dB（A）），该系统可以取得较好的识别效果，识别率达92%；在带噪声环境下，识别效果有所下降。在110dB（A）的随机噪声条件下，使用改进型谱减算法后可以有效的抑制白噪声，提高了系统的识别率，识别率从71%提高到84%。与理想高斯白噪声不同，实际环境噪声具有突发性和不可预见性，这种噪声严重影响着识别效果。在实录噪声进行模拟实际背景噪声测试中，本文采用的语音增强技术使系统的识别率从60%提高到74%，因此从实验结果可以看出，通过采用信噪比关联谱减及自适应门限端点检测进行语音增强处理，本文设计的声纹检测系统可以有效的改善噪声背景下声纹识别的性能。

表1 无语音增强下的测试结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验条件 | 背景噪声类型 | 背景声级 | 次数（10人） | 正确识别 | 错误识别 | 被拒识别 | 识别率（%） |
| 无语音增强 | 安静 | 58/dB(A) | 100 | 92 | 2 | 6 | 92 |
| 无语音增强 | 随机噪声 | 100/dB(A) | 100 | 71 | 5 | 24 | 71 |
| 无语音增强 | 实录噪声 | 90/dB(A) | 100 | 60 | 14 | 26 | 60 |

表2 语音增强下的识别结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验条件 | 背景噪声类型 | 背景声级 | 次数（10人） | 正确识别 | 错误识别 | 被拒识别 | 识别率（%） |
| 有语音增强 | 随机噪声 | 100/dB(A) | 100 | 84 | 3 | 13 | 84 |
| 有语音增强 | 安静 | 58/dB(A) | 100 | 94 | 2 | 2 | 94 |
| 有语音增强 | 实录噪声 | 90/dB(A) | 100 | 74 | 6 | 20 | 74 |

4 结论

本文设计了采用Shazam算法和语音增强处理的声纹识别系统，噪声背景下的实验结果表明，该系统有较高的识别率，说明基于自适应门限端点检测的语音增强处理可以有效的提高声纹识别系统的性能。

# 参考文献

1. 彭诗雅.基于声纹识别的身份认证技术研究[D].南京航空航天大学，2010
2. 李韵.声纹识别系统中特征参数提取方法的对比分析研究[D].成都理工大学，2016
3. Wang A. An Industrial Strength Audio Search Algorithm[C]// Ismir 2003, International Conference on Music Information Retrieval, Baltimore, Maryland, Usa, October 27-30, 2003, Proceedings. DBLP．2003
4. Graves A, Mohamed A, Hinton G. Speech Recognition with Deep Recurrent Neural Networks[J]. 2013, 38(2003):6645-6649
5. 宋小倩，周东升.基于Android平台的应用开发研究[J].软件导刊，2011,10(2):104-106
6. 郭霖.第一行代码[M].人民邮电出版社，2014

[7] Smyth N. Android Studio Development Essentials[M]. CreateSpace Independent Publishing Platform, 2014