变换域音频指纹算法研究

专　　业：通信与信息系统

硕 士 生：张廷贤

指导教师：陆哲明 教授

摘 要

随着计算机网络技术与多媒体技术的高速发展，尤其是数字音频压缩技术的成熟，使得数字音频的传播更加容易与广泛，从而引起了版权保护等一系列安全问题。音频水印技术的发展提供了解决这一问题的新思路，而鉴于音频水印技术自身的局限性，人们提出了音频指纹技术。音频指纹（Audio Fingerprinting）是基于内容的紧凑的签名，概括了音频片断固有的本质特征。由于音频指纹技术可以在独立于音频格式且无需元数据或者水印嵌入等额外信息的条件下进行音频识别，其已经引起了研究者的广泛关注。

本文通过对国内外音频指纹算法的分析，提出了两种鲁棒的变换域音频指纹算法，并将其应用于音频检索中。首先，阐述了选题背景及研究意义，并对现有音频指纹算法进行总结综述。其次，介绍本论文中所涉及的基础理论知识。接着提出以下两种鲁棒的变换域音频指纹算法：一、改进了基于短时傅里叶变换的频率域音频指纹算法。该算法引入了每帧音频信号的能量，利用频谱带能量（SBE，Spectral Band Energy）替换频率子带能量进行指纹提取；二、提出了一种基于Daubechies 小波变换的时频域音频指纹算法，通过对音频信号进行 8 层小波分解得到 1 个逼近分量和 8 个细节分量，根据每个分量小波系数的方差之间的关系提取音频指纹。最后，阐述了两种算法在音频检索中的应用。

实验结果表明，本文所提出的两种音频指纹算法对常见的保留信号内容的攻击处理及加性高斯白噪声具有很好的鲁棒性，降低了指纹存储空间、减少了指纹提取运算时间。此外，基于 Daubechies 小波变换的时频域音频指纹算法对线性速度变化攻击也具有良好的鲁棒性，其指纹存储空间较改进的频率域音频指纹算法大大减少。

关键词：音频指纹，变换域，音频检索

Audio Fingerprinting Algorithms Based on Transform Domains

Major : Communication and Information Systems

Name : Ting-Xian Zhang

Supervisor: Prof. Zhe-Ming Lu

Abstract

With the rapid development of computer network and multimedia technologies, especially the audio compression technology, digital audio can be transmitted more conveniently and widely. As a result, the copyright protection and security problems become more and more urgent. Digital audio watermarking technology provides a novel way to solve this problem. Nevertheless, audio fingerprinting is proposed due to the self-limitation of audio watermarking. Audio fingerprint is a compact content-based signature that summarizes the essence of an audio clip. It has been paid much attention since it can implement audio identification regardless of audio data format and without meta-data or watermark embedding.

In this paper, two robust audio fingerprinting algorithms in transform domains are proposed after analyzing the existent audio fingerprinting algorithms at home and abroad. Their performance in audio retrieval is also explored. Firstly, the background and significance of this topic are introduced, followed by an overview of the latest audio fingerprinting algorithms. Secondly, related basic theories are summarized briefly. Then, two robust audio fingerprinting algorithms in transform domains are proposed. One extracts frequency-domain features by using the Short-Time Fourier Transform. Energy of every audio frame is introduced and the spectral band energy is used in audio fingerprint extraction instead of the sub-band energy. The other applies the Daubechies wavelet transform to extract robust time-frequency features. We perform the Daubechies wavelet transform on each audio frame directly using 8 decomposition levels to get one approximate component and eight detail components.

Then the audio fingerprint is extracted according to the relationship among the variance of each sub-band's coefficients in different frames. Finally, the performance of both algorithms is verified in audio retrieval application.

Experimental results show that the proposed algorithms do not only have good robustness to content-preserving operations and additive white Gaussian noise but also reduce storage space and computation costs. In addition, the scheme based on the Daubechies wavelet transform shows highly robust to linear speed change attack and the fingerprint storage space is greatly reduced.

Keywords: Audio Fingerprinting, Transform Domain, Audio Retrieval

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：张廷贤

日期：2009年 5 月 22日

学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅，有权将学位论文的内容编入有关数据库进行检索，可以采用复印、缩印或其他方法保存学位论文。

学位论文作者签名：张廷贤

日期：2009年5 月22日

导师签名：

日期：2009年 5月 22日

第一章 绪论

1.1 选题背景及研究意义

随着计算机网络技术与多媒体技术的高速发展，使得图像、视频和音频等多媒体信息的传播越来越方便快捷，尤其是数字音频压缩技术的成熟使得数字音频的传播更加容易与广泛，从而引起了版权保护等一系列安全问题。在这种背景下，数字水印（Digital Watermarking）技术应运而生，得到了快速的发展。音频水印技术就是在不影响原始音频质量的条件下向其中嵌入具有特定意义且易于提取的信息的过程。其应用主要包括版权保护、盗版跟踪以及认证三个方面。虽然音频水印技术已取得诸多进展，但仍有许多挑战性的研究难题等待解决[1]。因此，人们提出了音频指纹技术。与音频水印技术相比，它具有以下特点：首先，音频指纹是对音频内容特征的概括，对攻击和失真具有鲁棒性；其次，音频指纹不需要对音频内容进行嵌入，而是提取音频内容的特征；同时，音频指纹依赖于音频的具体内容[2]。

音频指纹作为音频内容的一个标识，概括了音频听觉上的相关信息。音频指纹技术可应用在音频识别、完整性认证、水印支持及基于内容的音频检索等多个领域[2]。近年来，它已成为国内外研究的热点问题之一，引起了研究者的广泛关注，例如我们所熟悉的 Philips、Google 以及 Intel 等公司，此外也出现了许多利用音频指纹实现基于内容的音频检索系统[3, 4]。

音频指纹技术有着广泛的应用前景[2]：（1）可根据实际的需要在发布端、传输信道上或者消费端任一阶段实现对音频内容的监督与跟踪，防止未授权者对受保护的多媒体信息的使用或者合法使用者对其的错误使用，起到版权保护的作用；（2）可应用于增值服务中，开发基于内容的音频检索系统；（3）可用于完整性认证系统中，实现音频内容的篡改检测，保护授权者的合法利益；（4）可实现音频的分类与统计。因此，本课题提出的变换域音频指纹算法的研究具有重要的意义，应用前景广阔。

1.2 音频指纹技术综述

1.2.1 音频指纹技术及特点

一、音频指纹的定义

音频指纹（Audio Fingerprinting）是基于内容的紧凑的签名，概括了音频片断固有的本质特征[2]。音频指纹技术可以在独立于音频格式且无需元数据或者水印嵌入等额外信息的条件下进行音频识别，引起了研究者的广泛关注。

二、音频指纹的特点[2]

（1）音频指纹是对音频内容感知的概括，最大程度上保留了听觉上相关的信息，具有区别不同音频的辨别能力。

（2）音频指纹具有对内容保留的音频信号处理的不变性，如压缩、重采样、量化以及滤波等，这就是音频指纹的鲁棒性。这与在内容完整性认证应用中所要求的脆弱性是相反的，完整性认证能够检测音频内容是否被恶意的篡改。

（3）音频指纹具备紧凑性。简短的音频指纹，能够节约存储空间、降低匹配的计算复杂度。同时会对正确性、可靠性和鲁棒性造成一定影响。

（4）音频指纹需具备低的计算复杂度，从而降低指纹提取与匹配时系统的时间开销。

1.2.2 音频指纹技术的应用及难点

一、音频指纹技术的应用[2]

（1）音频识别

音频识别的框架如图 1-1 所示，主要包括数据库生成和识别两部分。通过收集音频文件，采用特定的提取算法提取音频的指纹，根据一定的逻辑结构把音频指纹存储起来，构成音频指纹数据库。在识别时，首先采用相同的算法提取待查询音频片段的音频指纹，再与数据库中的指纹进行匹配，返回相应的识别结果。

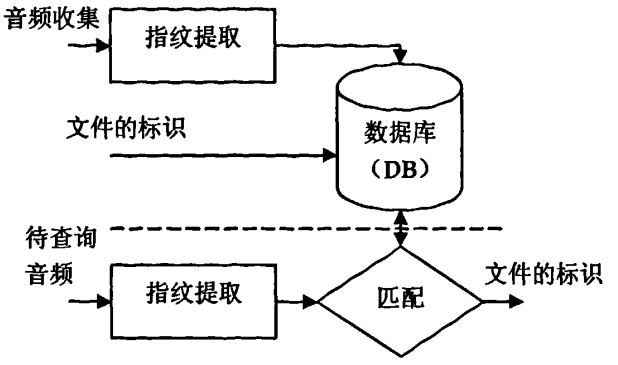


图 1-1　基于内容的音频识别框图

（2）完整性认证

完整性认证是为了实现音频内容的篡改检测，其总体框图如图 1-2 所示。首先，提取原始音频的指纹并保存。在认证阶段，将待检测的音频信号的指纹与原始指纹相比较得到认证结果。完整性认证不仅能够实现音频内容的篡改检测，还能检测出篡改的类型以及位置。

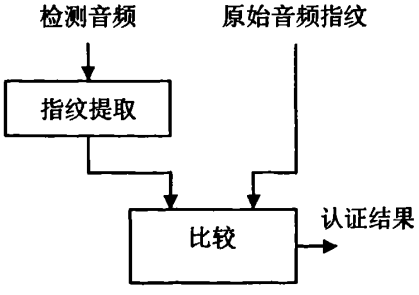


图 1-2　完整性认证框图

（3）水印支持

音频指纹能够用于生成基于内容的密钥（也称为音频哈希）作为水印，再将其嵌入到对应的音频内容中。文献[5]将音频指纹作为水印嵌入到音频信号中，通过将重建的原始指纹与观察信号所提取的音频指纹作比较得到匹配结果，这也属于完整性认证的范畴。

（4）基于内容的音频检索

从复杂的多媒体信息中提取紧凑的签名信息是多媒体信息检索的关键步骤，而音频指纹能够提取音频信号从低级的描绘算子到高级的描绘算子不同层次的

音频特征。因此，通过音频指纹的相似度匹配计算，能够实现基于内容的音频检索。

二、音频指纹技术的难点

（1）音频的数据量大；

（2）在音频指纹提取时，如何保证感知上相同的音频数据应该具有相同或者相似（即低于判决门限值）的音频指纹；

（3）如何根据提取的指纹设计相应的检索算法，从根本上降低计算复杂度，实现实时监督或检索；

（4）检索匹配门限值 T 的确定。

1.2.3 国内外研究现状及分析

目前，对于音频指纹的分类还没有形成统一的共识。文献[6]根据所提取的指纹特征与频率带的关系分为单频率带音频指纹、多频率带音频指纹和最优频率带与帧结合的音频指纹三类，而文献[7]将音频指纹分为语义特征和非语义特征两类。本文根据音频指纹所提取的特征属性将音频指纹分为时间域音频指纹算法、变换域音频指纹算法和压缩域音频指纹算法三类，以下从这三个方面对国内外的研究现状进行分析。

一、基于时间域的音频指纹

音频信号典型的时域特征包括短时能量、短时过零率、短时自相关系数和短时平均幅度差等。

文献[8]采用短时能量、短时过零率及短时基频（Short-time Fundamental Frequency）对音频进行分割与分类。文献[9]采用短时能量和短时过零率对语音、静音和谐音等进行分类。而文献[10]则采用改进的增强过零率（High Zero-Crossing Rate Ratio）、短时低能量比率（Low Short-Time Energy Ratio）分别取代短时过零率、短时能量实现对语音和非语音的分类。文献[11]提出了一种基于信息熵生成的直方图作为音频指纹的算法，其具备抵抗有损压缩和低通滤波的能力。

二、基于变换域的音频指纹

（1）基于频域的特征提取

人类对音频信号的感知过程与人类听觉系统（HAS, Human Auditory System）具有频谱分析功能是紧密相关的。因此，对音频信号进行频谱分析，是认识音频信号和处理音频信号的重要方法，常用的有离散傅立叶变换（DFT, Discrete Fourier Transform）和离散余弦变换（DCT, Discrete Cosine Transform）等。

文献[12]提出了一种以频谱极值点参数为特征的指纹提取及相应的匹配算法，其指纹提取的主要思想如下：（1）把整个频带划分为57个子带；（2）取一帧 N 点的样本（采样率为44.1KHz），作DFT变换，计算其绝对值，去除不重要的极值点再乘以延伸和移位系数 f_i，得到 $S+1$ 列的值；（3）对应于不同的 f_i，将极值点赋予每个子带，若所在子带无极值点，则赋0；（4）取具有最大值的 L（取值范围为17到25）个极值点所在子带序号，构成代表向量；（5）重复2-4，生成第 M 帧代表向量；（6）利用相应的算法压缩存储指纹。此外，文中还提出了一种递归自适应的DFT算法，大大提高了运算速度。

Haitsma 在文献[13, 14]中提出了一种高鲁棒的音频指纹系统模型。音频指纹的提取过程如下：（1）预处理，将输入的长度为 3s 的音频信号下采样为 5kHz 的单声道信号；（2）分帧与交叠，采用 0.37s 的汉宁窗，交叠因子为 31/32；（3）对每一帧采用 DFT 变换，得到其频谱值；（4）将与人类听觉系统 HAS 紧密相关的频谱范围 300Hz~2000Hz 等分为 33 个对数子带，即 Bark 域；（5）根据下式提取音频指纹 $F(n,m)$，

$$F(n,m)=\begin{cases}1, & E(n,m)-E(n,m+1)-(E(n-1,m)-E(n-1,m+1))>0\\ 0, & E(n,m)-E(n,m+1)-(E(n-1,m)-E(n-1,m+1))\le 0\end{cases} \quad (1\text{-}1)$$

式中 $E(n,m)$ 为第 n 帧子带 m 的能量，$F(n,m)$ 为第 n 帧的音频指纹的第 m 比特位。因此，每3.3s的音频信号经过处理后提取 256×32 bits的音频指纹块。实验表明，所提取的音频指纹能够抵抗MP3编解码、滤波、压缩、重采样、量化以及时间尺度拉伸等多种失真。文献[15]则把300Hz~2000Hz等分为512个对数子带，结合自相关平移不变性进行指纹提取。实验表明，改进的算法能抵抗高达 ±6% 的线性

速度变换攻击。而文献[16-18]对Haitsma所提出的算法进行了系统建模以及理论分析。文献[19]对文献[14]中所提出的检索算法作了改进，通过计算音频指纹之间的互相关系数，取其前S个极值点作为候选同步点，再计算其与待查询音频指纹的归一化汉明距，得到相应的匹配结果。实验表明，在加性高斯白噪声下，当$S=10,b=16,T=0.35$时能取到很好的匹配效果。文献[20]在文献[14]和[19]的基础上，对基音平移（Pitch-Shifted）进行分析，通过对提取的音频指纹进行滤波处理并设计相应的检索算法，能抵抗±8%的基音平移。

文献[21]结合神经网络，提出了一种采用OPCA（Oriented Principal Components Analysis）进行降维的失真判别分析（DDA, Distortion Discriminant Analysis）方法。文献[22]中采用短时傅立叶变换（STFT, Short-Time Fourier Transform）提取音频的频域特征参数，构成特征矩阵，再用高斯混合模型（GMM, Gaussian Mixture Modeling）进行建模，进而分析各特征参数在音频识别中的性能。文献[23]提出了一种基于正弦曲线模型的指纹提取算法，与经典的提取子边带参数模型相比，具有更强的抗加性噪声能力（尤其是伪随机加性噪声）以及检测更短的音频片段（1s）。

文献[24]提出了一种基于归一化频谱子频带质心（SSC, Spectral Sub-band Centroids）的指纹提取算法，文献[25]在此基础上增加了归一化频谱子带二阶距的分析与实验结果。作为对[24, 25]的改进，文献[26]在SSC基础上提出了基于Boosting学习算法的二值音频指纹，而文献[27]引入了SSC瞬时的动态特性，增强了音频指纹的鲁棒性。

（2）基于时频域的特征提取

针对频谱随时间变化的确定信号和非平稳随机信号，近年来出现了信号的时频域表示方法。其目的是将一维的时间信号或频域信号映射成时间频率平面上的二维信号，常用的有Gabor变换和小波变换。

文献[28]采用一维连续小波变换提取音频特征，构建了分别用于认证和识别的音频指纹。文献[29]提出了基于平衡多小波（BMW, Balanced Multiwavelets）的音频哈希算法。文献[30]结合计算机视觉，将音频信号的频谱图当作二维的图像进行处理。文献[31, 32]将计算机视觉技术应用于数据流处理中，运用Haar小

波对音频数据流的频谱图进行分解，提取小波系数，再利用 Min Hash 技术建模得到音频指纹，最后采用位置敏感哈希（LSH, Locality Sensitive Hashing）技术实现音频指纹检索。此外，分析了算法计算复杂度、音频指纹存储空间和识别率之间的关系。文献[33]在此基础上对系统的参数选择进行分析与验证，并将实验结果与文献[29]进行比较。

三、基于压缩域的音频指纹

文献[34]提出了一种基于心理声学模型提取压缩域参数作为音频指纹的算法，利用压缩域的 MDCT（Modified Discrete Cosine Transform）系数计算子频带能量再经过建模提取音频指纹。文献[35]将音频指纹技术应用于电视视频检索。根据对数子频带 MDCT 系数之和求得每一帧相应对数子频带调制谱的幅度再经过滤波、平滑和量化处理，生成音频指纹块，通过音频指纹块检索相应的视频。

音频指纹技术，目前还处于探索研究过程中，很多技术还不够成熟，各种算法都有各自的优缺点。而基于变换域的音频指纹算法，具有如下优点：通常具有更好的鲁棒性，对音频信号处理操作（如重采样、量化和编码等）和背景噪声都具有一定的抵抗力；不同的变换域，能保留音频信号不同的听觉信息特征，能抵抗特定的攻击，如文献[31-33]利用 Haar 小波变换提取时频谱特征作为音频指纹对时间尺度拉伸（TSM, Time Scale Modification）攻击有很好的效果。同时，存在部分指纹算法计算复杂度高，以及不能很好的满足基于内容的音频检索系统进行实时的检索。

1.3 本论文研究的主要内容及结构安排

本论文通过对国内外音频指纹算法的分析，提出了基于 STFT 变换的频率域音频指纹算法以及基于 Daubechies 小波变换的时频域音频指纹算法，并将两种算法应用于音频检索中。论文结构如下：

第一章为绪论，简要阐述了选题背景及研究意义，并对现有音频指纹算法进行总结综述。

第二章介绍了论文中所涉及的基础理论知识。首先，概述音频信号的相关特

征。其次，简要介绍了 STFT 变换及其应用。最后，对小波变换进行了简要介绍。

第三章给出了一种基于 STFT 变换的频率域音频指纹改进算法。该算法引入了每帧音频信号的能量，利用频率带能量（SBE）替换频率子带能量，对每 3.3s 的音频信号提取128×16bits 的音频指纹块。实验结果表明，改进的频率域音频指纹算法对常见的保留信号内容的攻击处理及加性高斯白噪声具有很好的鲁棒性，此外降低了指纹存储空间、减少了指纹提取运算时间。

第四章提出了一种基于 Daubechies 小波变换的时频域音频指纹算法，通过对音频信号进行 8 层小波分解得到 1 个逼近分量和 8 个细节分量，根据每个分量小波系数的方差之间的关系对每 3.3s 的音频信号提取 64×7 bits 或64×8 bits 的音频指纹块。实验结果表明，该算法不仅对常见的保留信号内容的攻击处理及加性高斯白噪声具有很好的鲁棒性，对线性速度变化攻击也具有良好的鲁棒性，其指纹存储空间较频率域音频指纹大大减少。

第五章阐述了两种算法在音频检索中的应用。首先阐述音频检索中的匹配算法。最后通过实验结果对两种音频指纹算法在音频检索中的性能进行对比总结。

第六章对全文进行了总结并对音频指纹技术今后的发展进行了展望。

第二章 基础理论知识

2.1 音频基础知识

2.1.1 音频及其短时处理技术

音频是指人耳所能听到的所有声音，其频率范围是20Hz-20KHz，其中语音的频率范围为300Hz-3.4KHz，而音乐和其它自然声响则是分布于整个频率范围。在日常生活中，人耳所听到的音频信号都是时间和幅度连续变化的模拟信号，为了便于利用计算机进行处理，必须把模拟的信号进行数字化处理，转换为时间和幅度都是离散的数字音频信号。数字化处理主要包括采样、量化和编码三个部分。本论文中所采用的音频格式为wav格式，采样率为44100Hz，采用16bit PCM编码。

经过数字化处理的音频信号实际上是一个非平稳的时变信号，而传统的信号分析方法主要适用于平稳信号的分析。因此，对音频信号进行处理时，需要通过加窗操作得到短时的音频信号，即对音频信号进行分帧，而在几十毫秒的短时间内，可以将缓慢变化的音频信号当作平稳信号来处理。分帧可以连续，也可以采用交叠分帧的方法。常用的窗函数有矩形窗、汉明窗（Hamming）和汉宁窗（Hanning）等，窗口大小一般为几毫秒到几十毫秒。

矩形窗函数如公式（2-1）所示。其单位冲激响应如公式（2-2）所示。

$$w(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & \text{其它}\end{cases} \tag{2-1}$$

$$H(e^{jwT})=\sum_{n=0}^{N-1}e^{-jwnT}=\frac{\sin(wNT/2)}{\sin(wT/2)}e^{-jwT(N-1)/2} \tag{2-2}$$

式中N为窗长。矩形窗具有线性相位-频率特性，其频率响应的第一个零值所对应的频率为：

$$f_0 = \frac{f_s}{N} = \frac{1}{NT_s} \tag{2-3}$$

式中 f_s 为采样率，$T_s = 1/f_s$ 为采样周期。

汉宁窗（Hanning）的窗函数如公式（2-4）所示。

$$w(n) = \begin{cases} 0.5 - 0.5\cos[2\pi n/(N-1)], & 0 \le n \le N-1 \\ 0, & \text{其它} \end{cases} \tag{2-4}$$

汉明窗（Hamming）的窗函数如公式（2-5）所示。

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \le n \le N-1 \\ 0, & \text{其它} \end{cases} \tag{2-5}$$

图 2-1 为窗长 $N = 64$ 的矩形窗和汉宁窗（Hanning）的幅频响应。从图中可知，汉宁窗的带宽约为矩形窗带宽的两倍，同时在通带外，汉宁窗衰减较快。

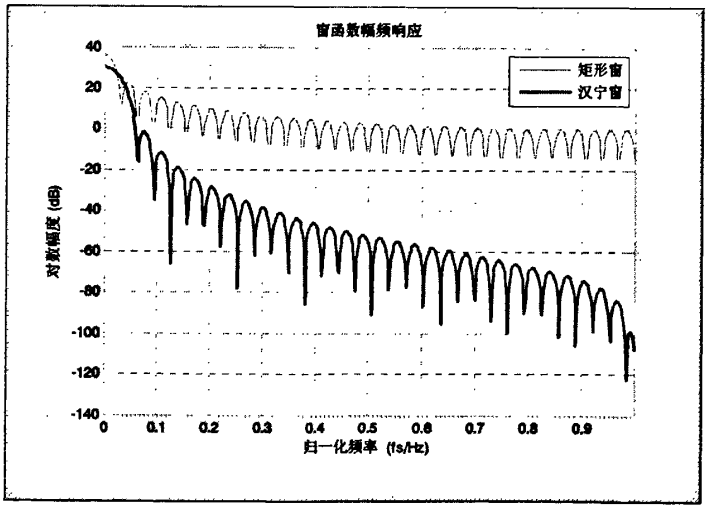


图 2-1 窗函数幅频响应

窗函数及窗长的选择将影响到音频信号短时分析的结果。矩形窗的谱比较平滑，但是波形细节丢失，并且会产生高频干扰和频谱泄漏；而汉宁窗可以有效的克服泄漏现象，应用范围广泛。如果窗长 N 很大，则等效于很窄的低通滤波器，音频信号的高频成分将受到严重衰减，导致信号短时能量变化缓慢，不能充分反映信号的变化；反之，如果窗长 N 很小，则使低通滤波器通带变宽，信号短时能量变化剧烈，不能得到平滑的短时能量信号，因此，必须选择合适的窗长 N [36]。

2.1.2 音频信号的时域分析

时域分析是以时间为变量对信号直接进行分析，音频信号典型的时域特征包括短时能量、短时平均过零率、短时自相关系数和短时平均幅度差等[36]。音频信号的短时平均能量和短时平均过零率及其改进特征可用于对音频信号进行分类[8-10]。

一、短时能量

对于音频信号 $x(\tau)$，加窗分帧处理后得到第 n 帧音频信号为 $x_n(m)$，其短时能量的定义如下：

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{2-6}$$

式中，N 为窗长。

短时能量能够反映信号幅度大小的变化，然而由于其引入了信号幅度的平方运算，因此它对高电平非常敏感。为了克服这一缺陷，引入了短时平均幅值，其定义如下：

$$M_n = \sum_{m=0}^{N-1} |x_n(m)| \tag{2-7}$$

短时平均幅值 M_n 用信号幅度的绝对值取代其平方和，简化了运算，同时解决了对急剧变化的信号的幅值进行平方运算所引入的较大差异。

短时能量和短时平均幅值的主要用途有：（1）可以区分浊音段与清音段，因为浊音时 E_n 比清音时大得多。（2）可以用来区分声母和韵母、无声与有声以及连字的分界等。（3）作为音频特征，用于语音识别[37]。

二、短时平均过零率

短时平均过零率是指每帧信号内波形通过零值（或设定阈值 T）的次数。其定义如下：

$$z_n = \frac{1}{2}\sum_{m=0}^{N-1} |\mathrm{sgn}[x_n(m)] - \mathrm{sgn}[x_n(m-1)]| \tag{2-8}$$

式中，$\mathrm{sgn}[x_n(m)]$ 为符号函数，定义如下：

$$\mathrm{sgn}[x_n(m)]=\begin{cases}1, & x_n(m)\geq 0\\ -1, & x_n(m)<0\end{cases} \tag{2-9}$$

可以将短时平均过零率和短时能量结合进行端点检测，在背景噪声较大时，采用短时平均过零率比较准确；反之则采用短时能量[36]。

三、短时自相关系数和短时平均幅度差

短时自相关函数主要用于研究信号$x_n(m)$本身的同步性和周期性，其定义如下：

$$R_n(k)=\sum_{m=0}^{N-1-k}x_n(m)x_n(m+k) \tag{2-10}$$

式中，k为延迟点数。

短时自相关函数具有以下性质：（1）若$x_n(m)$为周期信号，则其自相关函数同样为周期信号，且具有相同的周期 T；（2）自相关函数是偶函数，即$R_n(k)=R_n(-k)$；（3）当$k=0$时，自相关函数具有最大值，此时$R_n(0)$为音频信号$x_n(m)$的能量[36]。

由于乘法运算计算量较大，短时自相关函数计算时间较长，因此，常常采用具有类似作用的短时平均幅度差函数替换短时自相关函数。其定义如下：

$$r_n(k)=\sum_{m=0}^{N-1-k}\left|x_n(m)-x_n(m+k)\right| \tag{2-11}$$

若$x_n(m)$为周期信号，则$r_n(k)$同样为周期信号。与$R_n(k)$相反的是，在周期整数倍点上$r_n(k)$为谷值，并非峰值。由此可见，短时自相关函数和短时平均幅度差函数均能用于基音周期检测，且短时平均幅度差函数计算更加简单。

2.2 傅里叶变换

音频信号的频域包含了音频信号最重要的感知特征，而人类听觉系统 HAS 具有频谱分析功能。因此，对音频信号进行频谱分析，是认识音频信号和处理音频信号的重要方法，常用的有离散傅立叶变换和离散余弦变换等[38]。

2.2.1 连续傅里叶变换

对于连续时间信号 $x(t)$，其连续傅里叶变换（CFT, Continuous Fourier Transform）为：

$$X(w)=\int_{-\infty}^{+\infty}x(t)e^{-jwt}dt \tag{2-12}$$

式中，w 为模拟角频率，即 $w=2\pi f$。其逆变换（ICFT, Inverse CFT）定义如下：

$$x(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}X(w)e^{jwt}dw \tag{2-13}$$

通常，将 $x(t)$ 和 $X(w)$ 称为一个变换对，记为：$x(t)\leftrightarrow X(w)$。

2.2.2 离散傅里叶变换

对于离散时间信号 $x(\tau)$，其离散傅里叶变换 DFT 为：

$$X(e^{j\Omega})=\sum_{\tau=-\infty}^{+\infty}x(\tau)e^{-j\tau\Omega} \tag{2-14}$$

式中，Ω 为数字角频率，即 $\Omega=wT_s=2\pi f T_s$。

从公式（2-14）可知，信号 $x(\tau)$ 为无限长序列，而在现实中，所处理的信号往往有限长序列，对于序列长度为 N 的有限长序列 $x(n)$，其离散傅里叶变换定义如下：

$$X(e^{j\Omega_k})=\sum_{n=0}^{N-1}x(n)e^{-jn\Omega_k} \tag{2-15}$$

式中，$\Omega_k=\frac{2\pi}{N}k$，$0\le k<N$，将 Ω_k 代入公式（2-15）中，得到

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N},\quad 0\le k<N \tag{2-16}$$

其逆变换（IDFT, Inverse DFT）为

$$x(n)=\frac{1}{N}\sum_{n=0}^{N-1}X(k)e^{j2\pi nk/N},\quad 0\le k<N \tag{2-17}$$

在傅里叶分析过程中，为了避免频谱混叠现象，由香农定理可知采样频率 f_s 必须符合下列条件：

$$f_s \geq 2f_{max} \tag{2-18}$$

式中，f_{max} 为信号的最高频率。

而对于 N 点傅里叶变换，其频率分辨率，即在频率轴上所能分辨的最小频率间隔为

$$\Delta f = \frac{f_s}{N} \tag{2-19}$$

由上式可知，若确定了信号的最高频率 f_{max} 和采样频率 f_s，则只能通过增加傅里叶变换的样本数 N 来提高其频率分辨率。因此，傅里叶分析属于固定分辨率分析方法，适用于频谱不随时间变化的确定信号及平稳随机信号的分析。

2.2.3 短时傅里叶变换

对于信号 $x(n)$，其短时傅里叶变换 STFT 定义为：

$$X_{STFT}(n,\Omega_k) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\Omega_k m} = w(n) * x(n)e^{-j\Omega_k n} \tag{2-20}$$

式中，$w(n)$ 为窗函数，$\Omega_k = \frac{2\pi}{N}k$，$0 \leq k < N$。

其逆变换为

$$x_{ISTFT}(n) = \sum_{k=0}^{N-1} X_{STFT}(n,\Omega_k)e^{j\Omega_k n} \tag{2-21}$$

图 2-2.（a）和图 2-2.（b）为采样频率为 5KHz，样本数 $N = 512$ 的音频信号 $x(n)$ 的波形图和其短时傅里叶变换频谱图；图 2-2.（c）和图 2-2.（d）为音频信号 $x(n)$ 经过汉宁加窗操作后的波形图和其短时傅里叶变换频谱图。

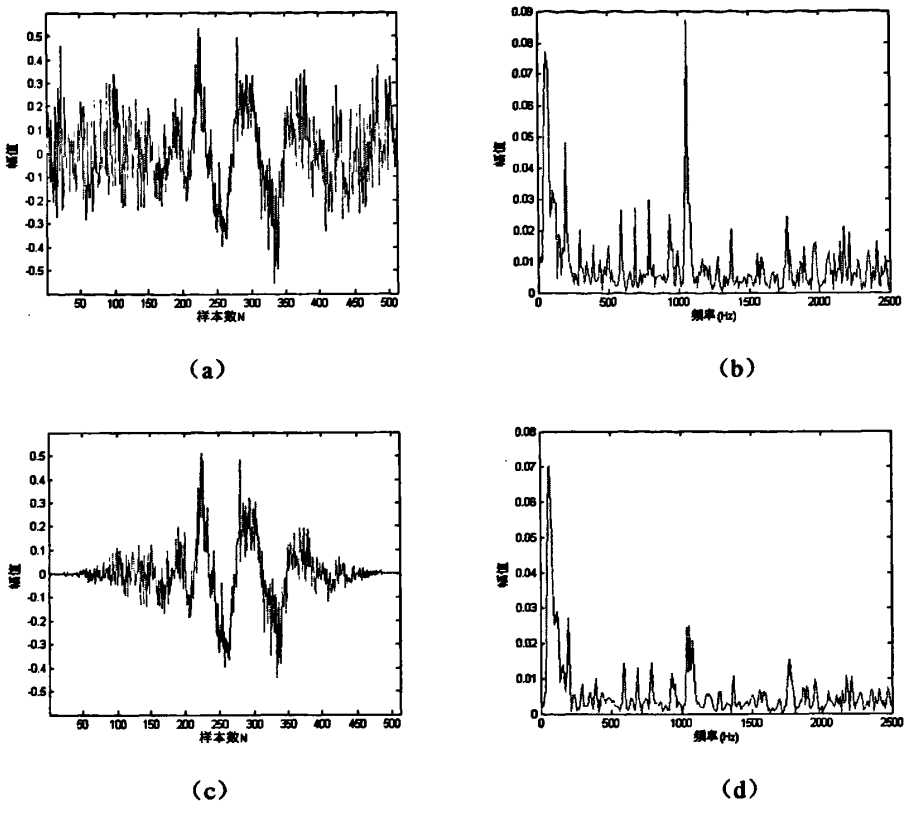


图 2-2 音频信号 $x(n)$ 的波形图及其短时傅里叶变换频谱图

2.3 小波变换

近些年来，小波分析成为信号处理中的研究热点，不仅仅在理论上取得了很多突破性的进展，而且在图像处理、语音信号处理以及数据压缩处理等许多领域中得到了极其广泛的应用。

小波变换是一种时频分析方法，克服了傅里叶变换没有任何局部化特性和短时傅里叶变换固定分辨率的缺陷，具有多分辨率分析信号的特点，而且在时域和频域内都具有表征信号局部特征的能力，时间窗和频率窗都可以根据信号的具体形态动态调整。它可以用长的时间间隔来获得更加精确的低频率的信号信息，用短的时间间隔来获得高频率的信号信息。小波分析的主要优点之一就是能够提供局部细化与分析的功能[39]。

小波定义：令函数$\psi(t)\in L^2(R)$（$L^2(R)$表示平方可积的实数空间，即能量有限的信号空间），若$\psi(t)$满足以下条件

$$C_\psi = \int_R \frac{\left|\bar{\psi}(w)\right|^2}{|w|} dw < \infty \tag{2-22}$$

式中，$\bar{\psi}(w)$为$\psi(t)$的频域表示形式，称$\psi(t)$为小波母函数或基本小波。若将小波母函数进行伸缩和平移，则得到一个小波基函数，即

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{2-23}$$

式中，a为尺度因子，且$a>0$，b为位移因子。其对应的频域表示如下：

$$\bar{\psi}_{(a,b)}(w) = \sqrt{a}e^{-jw}\bar{\psi}(aw) \tag{2-24}$$

通过选择合适的参数对(a,b)，实现对函数和信号进行任意点处任意精度的分析，这也决定了小波分析在对非平稳信号进行时频分析时具有时频同时局部化的能力。

2.3.1 常用小波函数

常用的小波函数有 Haar 小波、Daubechies 小波、Mexico Hat 小波和 Morlet 小波。下面简要介绍一下 Haar 小波和 Daubechies 小波。

Haar 小波是在小波分析中最简单、最紧支撑的小波函数，其定义为

$$\psi_H = \begin{cases} 1, & 0 \le x \le 0.5 \\ -1, & 0.5 \le x < 1 \\ 0, & \text{其他} \end{cases} \tag{2-25}$$

Daubechies 小波的数学表达式为

$$\psi_D(x) = \sum_{k=0}^{N-1} C_k^{N-1+k} x^k \tag{2-26}$$

式中，C_k^{N-1+k}为二项式系数，那么

$$|m_0(w)|^2 = (\cos^2\frac{w}{2})^N \psi(\sin^2\frac{w}{2}) \tag{2-27}$$

式中，$m_0(w) = \frac{1}{\sqrt{2}}\sum_{k=0}^{2N-1} h_k e^{-jkw}$。

Daubechies 小波族具有紧支撑性，简写为 db$N$，其中 N 表示阶数，Haar 小波实际为 db1，即阶数 $N=1$ 的 Daubechies 小波。

2.3.2 连续小波变换

对于连续时间信号 $x(t)$，给定一个基本小波函数，则其连续小波变换（CWT, Continuous Wavelet Transform）为

$$CWT_x(a,b) = \int x(t)\psi_{(a,b)}^*(t) = \frac{1}{\sqrt{a}}\int x(t)\psi^*(\frac{t-b}{a})dt \tag{2-28}$$

连续小波变换具有叠加性、平移不变性和尺度不变性。若 $\bar{\psi}^*(w)$ 满足以下条件

$$C_\psi = \int_0^{+\infty} \frac{|\bar{\psi}^*(w)|^2}{|w|}dw < \infty \tag{2-29}$$

时，才能通过 $CWT_x(a,b)$ 重构得到原来信号 $x(t)$，即

$$\begin{aligned} x(t) &= \frac{1}{C_\psi}\int_0^{+\infty}\frac{da}{a^2}\int_{-\infty}^{+\infty} CWT_x(a,b)\psi_{(a,b)}^*(t)db \\ &= \frac{1}{C_\psi}\int_0^{+\infty}\frac{da}{a^2}\int_{-\infty}^{+\infty} CWT_x(a,b)\psi^*(\frac{t-b}{a})db \end{aligned} \tag{2-30}$$

2.3.3 离散小波变换

将连续小波的尺度因子 a 和位移因子 b 按照幂级数进行离散化，得到离散的基本小波函数，即

$$\psi_{(j,k)}^*(t) = \frac{1}{\sqrt{a_0^j}}\psi^*(\frac{t-ka_0^j b_0}{a_0^j}) = a_0^{-\frac{j}{2}}\psi^*(a_0^{-j}t - kb_0) \tag{2-31}$$

通常情况下，取$a_0=2$，此时离散小波变换（DWT, Discrete Wavelet Transform）为

$$DWT_x(j,k)=\int x(t)\psi_{(j,k)}^{*}(t)dt=2^{-\frac{j}{2}}\int x(t)\psi_{(j,k)}^{*}(2^{-j}t-kb_0)dt \qquad (2\text{-}32)$$

2.3.4 多分辨率分析及小波分解

在小波分析过程中，常常通过改变尺度因子a的大小对信号的局部特性进行分析。当a取较大值时，相当于频率分辨率较低，能够概述信号的变化趋势；当a取较小值时，相当于频率分辨率较高，便于分析信号的高频分量，观察信号的细节变化。但是，在不同a值下分析的品质因数却保持不变。这种由粗略到精细，对信号进行多角度观察的分析方法被称为多分辨率分析[40]。

离散小波变换主要用在信号处理中，一般采用 mallat 算法实现，也称为快速小波变换算法（FWT, Fast Wavelet Transform），首先对较大尺度的信号进行小波分解，得到细节分量（即高频分量）和逼近分量（即低频分量），接着对逼近分量再进行小波分解，从而实现多分辨率分析。2 层小波分解原理如图 2-3 所示。

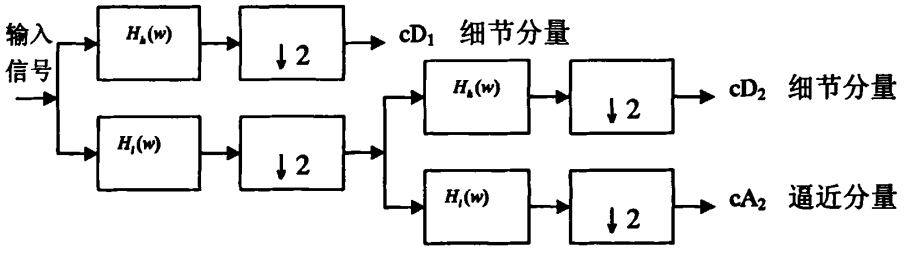


图 2-3 2 层小波分解原理图

图 2-3 中，$H_h(w)$、$H_l(w)$分别为高通滤波器函数和低通滤波器函数。2 层小波分解得到两个细节分量（cD_1、cD_2）和一个逼近分量（cA_2），其对应的归一化频率分别为$\frac{\pi}{2}\sim\pi$、$\frac{\pi}{4}\sim\frac{\pi}{2}$和$0\sim\frac{\pi}{4}$。采用小波基 db6 对音频信号$x(n)$进行 2 层小波分解，其结果如图 2-4 所示。

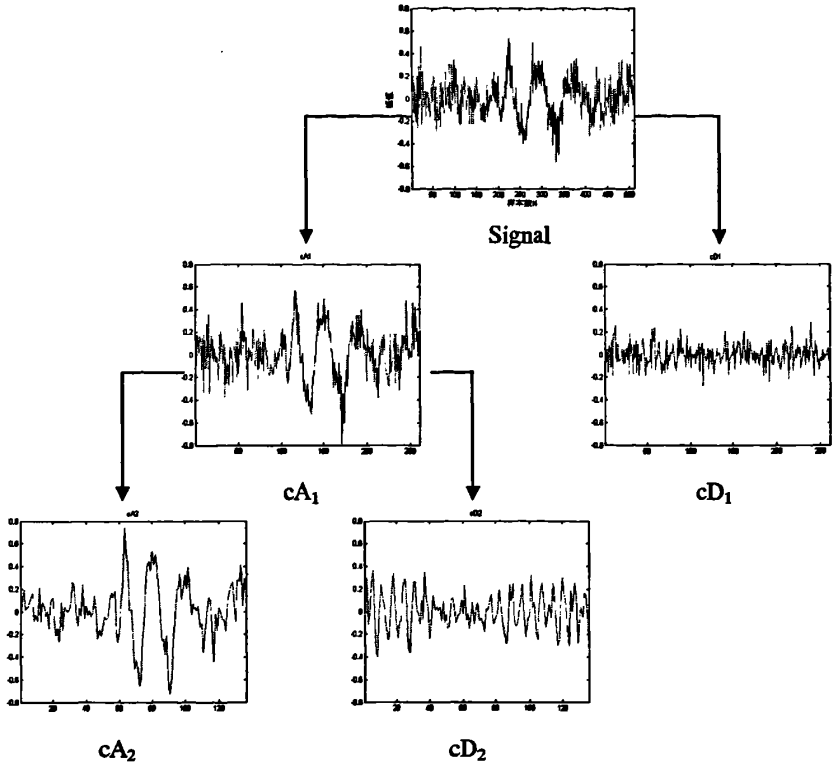


图 2-4 音频信号 $x(n)$ 的 2 层小波分解示意图

2.4 小结

本章简要介绍了音频相关的基础理论知识及傅里叶变换、小波变换的原理，分析了傅里叶变换和小波变换在音频信号处理中的应用，为本论文的进一步研究奠定了基础。

第三章 基于 STFT 变换的频率域音频指纹算法

3.1 引言

Haitsma 等人提出了一种高鲁棒的音频指纹系统模型，对分帧（0.37s）、加窗（Hanning 窗）与交叠（31/32）的音频信号进行 DFT 变换，将频谱范围 300Hz-2000Hz 的频率段均匀划分为 33 个对数频带，通过相邻帧 33 个对数频带能量之间的关系提取 32 bits 的子指纹（Sub-fingerprint）。由于单个子指纹未能携带足够的用于音频识别的信息，因此，对 3.3s 的音频片段提取 256×32 bits 的音频块（Fingerprint-block）[14]。而 Bellettini 对此算法进行了改进，将频谱范围 300Hz-2000Hz 的频率段均匀划分为 17 个对数频带提取 16 bits 的子指纹；在检索时，通过计算待查询音频指纹与源音频指纹之间的互相关系数，取其前 S 个极值点作为候选同步点，再计算其与源音频指纹的归一化汉明距，得到相应的匹配结果。实验结果表明，改进的算法对加性高斯白噪声具有很好的鲁棒性[19]。文献[22]采用 STFT 变换提取音频的频域特征参数香农熵（Shannon Entropy）、瑞利熵（Renyi Entropy）、频谱质心（Spectral Centroid）、频谱带宽（Spectral Bandwidth）、频谱带能量 SBE、频谱平稳度（Spectral Flatness Measure）、频谱极值因子（Spectral Crest Factor）和 MFCC（Mel-frequency Cepstral Coefficients），接着对于每种特征参数，构造其特征矩阵，再用高斯混合模型 GMM 进行建模，进而分析各特征参数在音频识别中的性能。

本章在 Haitsma 及 Bellettini 提出的系统模型的基础上，将频谱范围 300Hz-2000Hz 的频率段均匀划分为 17 个对数频带，引入每帧音频信号的能量，利用频谱带能量 SBE 替换频率子带能量，并通过调整系统的交叠因子，对 Haitsma 算法进行改进。实验结果表明，改进的算法 SBE 对常见的保留信号内容的攻击处理及加性高斯白噪声均具有很好的鲁棒性，同时降低了指纹存储空间。

3.2 算法的实现

频率域音频指纹算法的原理框图如图 3-1 所示。

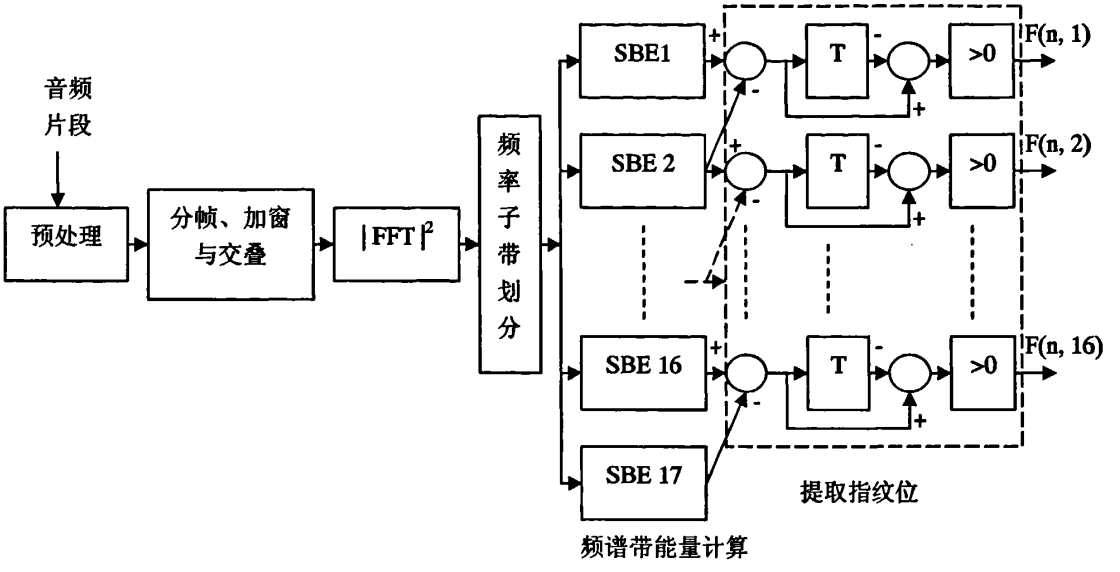


图 3-1 频率域音频指纹算法的原理框图

假设系统输入为 44.1KHz, 16bits PCM 的音频信号，结合 Bellettini 的改进算法，对每帧音频信号提取 16bits 的音频子指纹，具体提取过程如下：

（1）预处理，将输入音频信号下采样为 5KHz 的单声道信号。此处采用求平均将立体声信号转换为单声道信号。

（2）分帧、加窗与交叠，帧长为 0.37s，采用汉宁窗，交叠因子为 P。加窗得到了短时平稳的音频信号，同时减小了相邻帧之间在频率域上的跳跃，使信号的频谱更加平稳过渡。交叠改善了由于音频信号不同步所造成的影响，保证了在不同步情况下所提取的音频指纹与源音频指纹具有较高的相似度，从而达到抵抗随机起始点攻击的作用。

（3）对每一帧采用 STFT 变换，得到其频谱值。由于人类听觉系统 HAS 对相位失真不敏感，在指纹提取过程中只采用 STFT 变换后系数的模值进行计算。

（4）SBE 计算，将与人类听觉系统 HAS 紧密相关的频谱范围 300Hz~2000Hz 等分为 17 个互不交叠的对数频率子带（类似于 Bark 域），如表 3-1 所示，利用公式（3-1）计算对应频率子带的 SBE。

$$S(n,m)=\frac{\sum_{u=l_b}^{u_b}|f_n(u)|^2}{\sum_{u=300}^{2000}|f_n(u)|^2} \tag{3-1}$$

式中$S(n,m)$表示第n帧第m子带的SBE，u_b、l_b分别为第m子带的上、下限频率。

表3-1 300Hz~2000Hz频率段等分为17个互不交叠的对数频率子带结果

（表中，l_b、u_b为子带的上、下限频率，单位：Hz）

子带序号	l_b	u_b	带宽	子带序号	l_b	u_b	带宽
1	300	335	35	10	819	916	97
2	335	375	40	11	916	1024	108
3	375	419	44	12	1024	1145	121
4	419	469	50	13	1145	1280	135
5	469	524	55	14	1280	1431	151
6	524	586	62	15	1431	1600	169
7	586	655	69	16	1600	1789	189
8	655	733	78	17	1789	2000	211
9	733	819	86				

（5）利用公式（3-2），对每帧音频信号提取16bits的音频指纹。

$$F(n,m)=\begin{cases}1, & S(n,m)-S(n,m+1)-(S(n+1,m)-S(n+1,m+1))>0\\ 0, & S(n,m)-S(n,m+1)-(S(n+1,m)-S(n+1,m+1))\le 0\end{cases} \tag{3-2}$$

式中$F(n,m)$表示第n帧第m比特的指纹。

从音频“O Fortuna”中截取3.3s的音频片段，采用改进的算法SBE，取交叠因子P=30/32，提取得到128×16bits的音频指纹块，结果如图3-2所示。图中，黑像素点代表‘1’，白像素点代表‘0’。图3-2.（a）为源音频指纹块，图3-2.（b）为128K MP3压缩处理后提取的音频指纹，图3-2.（c）为处理后的音频指纹与源音频指纹的误码率（BER, Bit Error Rate）图，其BER=0.1982。

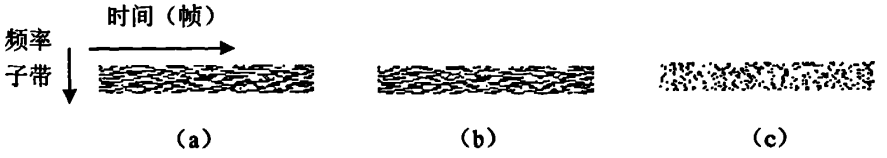


图 3-2 音频指纹及误码率图

3.3 实验结果及分析

为了验证改进算法的鲁棒性，实验中采用了 Haitsma 论文中所使用的 4 首测试音频，分别为“O Fortuna”、“Success has made a failure of our home”、“Say what you want”和“A whole lot of Rosie”。所有的音频片段都经过如下攻击处理：

（1）128Kbps 和 32Kbps MP3 压缩。

（2）二阶巴特沃斯（Butterworth）带通滤波（BPF, Band Pass Filter）：通带频率为 100Hz-6000Hz。

（3）幅度压缩（Compression），具体设置如下：当幅度$|A| \geq -28.6$dB 时，压缩比为 8.94:1；当$-46.4\text{dB} < |A| < -28.6\text{dB}$时，压缩比为 1.73:1；当$|A| \leq -46.4$dB 时，压缩比为 1:1.61。

（4）添加回声（Echo）。

（5）均衡（Equalization），采用典型的 10 频段均衡器，具体设置如表 3-2 所示。

表 3-2 典型的 10 频段均衡器参数设置

频率（Hz）	31	62	125	250	500	1K	2K	4K	8K	16K
增益（dB）	-3	+3	-3	+3	-3	+3	-3	+3	-3	+3

（6）时间尺度拉伸（Time Scale Modification），-2%、+2%、-4%、+4%、-5%和+5%。时间尺度拉伸保持基音频率不变，只对时间进行拉伸。

（7）线性速度变化（Linear Speed Change），-1%、+1%、-3%、+3%、-4%、+4%、-5%和+5%。线性速度变化对时间和基音频率都进行拉伸。

（8）添加加性高斯白噪声，信噪比分别为 20dB、15dB、10dB、5dB、3dB、2dB。

3.3.1 误码率（BER）分析

本文采用归一化汉明距（即误码率 BER）来衡量两个音频片段之间的相似度，为了能够进行正确识别，必须设定合适的阈值 T。若两个音频片段之间的相似度 d<T，则判定为相同；反之，则判定为不同。T 的大小直接影响着误检率（FPR, False Positive Rate）和漏检率（FNR, False Negative Rate），实验中取 T=0.35 [14]。下面通过对受攻击处理的音频片段与源音频片段、受攻击处理的音频片段与不同音频片段之间的误码率进行分析，验证 T=0.35 取值的合理性。

测试音频经过 128Kbps MP3 压缩、32Kbps MP3 压缩、带通滤波、幅度压缩、添加回声和均衡共六种攻击处理，对每个攻击样本随机选取 100 个起始点截取 3.3s 音频片段进行测试，并对其误码率进行统计，具体结果如图 3-3 所示。

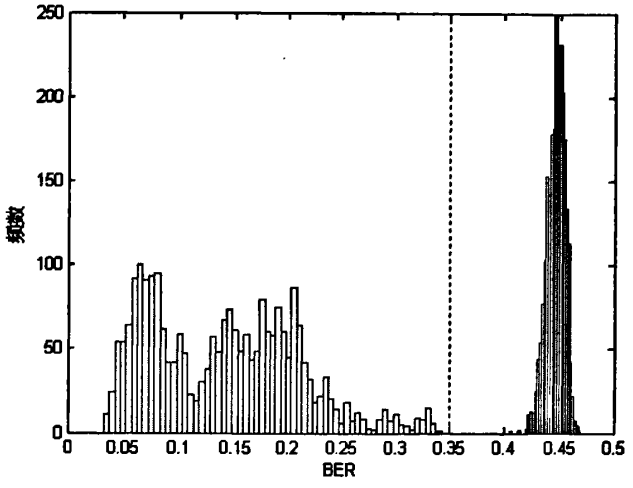


图 3-3 误码率 BER 分布图

图 3-3 虚线左边为源音频指纹与其受攻击处理的测试音频指纹之间的误码率，其分布比较分散，这是由于误码率随着攻击类型的不同而变化所造成的；右边为受攻击处理的测试音频指纹与不同音频指纹之间的误码率，呈正态分布。由图 3-3 可知，取 T=0.35 能够很好的区分相同音频指纹和不同音频指纹之间的关系。

3.3.2 鲁棒性测试

为了验证不同交叠因子 P 对系统鲁棒性的影响，选取了 P=31/32、30/32、28/32、24/32、16/32。实验中对每个攻击样本随机选取 100 个起始点截取 3.3s 音频片段进行测试。实验中，采用误码率 BER、最佳识别率（IDR, Identification Rate）和正确识别率（TPR, True Positive Rate）来衡量算法的鲁棒性。IDR 和 TPR 分别定义如下：

$$IDR = \frac{\text{相似度距离最小且匹配（忽略T）}}{\text{测试总次数}} \tag{3-3}$$

$$TPR = \frac{\text{最小相似度距离小于T且匹配}}{\text{测试总次数}} \tag{3-4}$$

一、对于常见的保留信号内容的攻击处理的鲁棒性

对于常见的保留信号内容的攻击处理，待测试音频片段指纹与源音频指纹之间的 BER 平均值如图 3-4 所示。其相应的识别结果如图 3-4 所示。

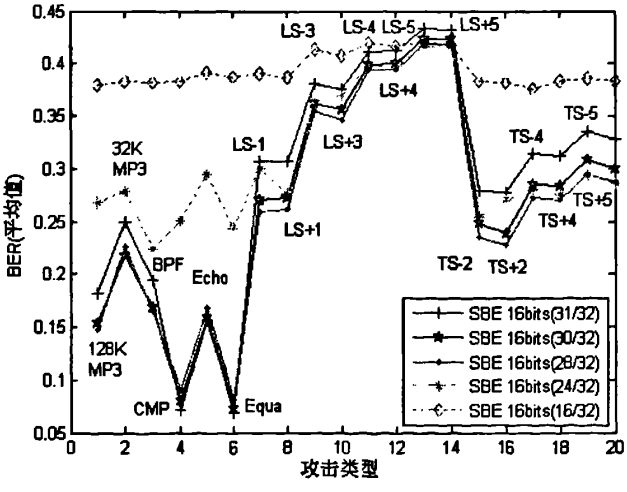
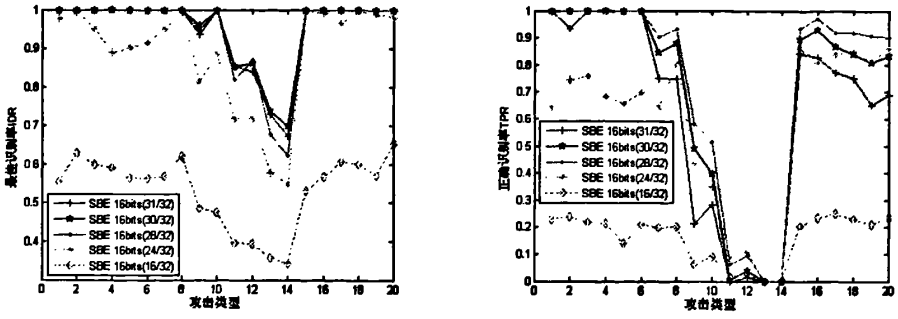


图 3-4 不同交叠因子 P 在不同攻击下的鲁棒性

由图 3-4 可知，当 P=31/32、30/32、28/32 时，BER 随着交叠因子 P 的减小而减小；当 P=24/32、16/32 时，BER 随着交叠因子 P 的减小而增大。在时间尺度拉伸和线性速度变化攻击下，P=28/32 的性能比 P=30/32 的性能稍好，在其它攻击处理下性能相当。



（a）　　（b）

图 3-5 不同交叠因子 P 在不同攻击下的识别率

由图 3-5.（a）可知，忽略了 T 取值的影响，最佳识别率 IDR 在 P=30/32 时取得最好的识别效果。而从图 3-5.（b）可知，设定了 T=0.35 后，在时间尺度拉伸和线性速度变化攻击下，P=28/32 时的正确识别率 TPR 比 P=30/32 时的正确识别率 TPR 稍好，这与其鲁棒性结果相符合。

二、对加性高斯白噪声的鲁棒性

改进的算法 SBE 在不同程度下的加性高斯白噪声的鲁棒性结果如图 3-6 所示。实验中，加性高斯白噪声的信噪比分别设置为 20 dB、15 dB、10 dB、5 dB、3 dB、2dB。其相应的识别结果如图 3-7 和图 3-8 所示。

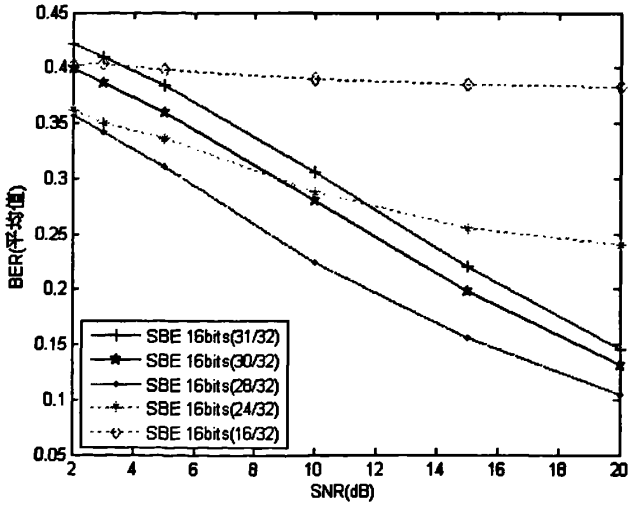
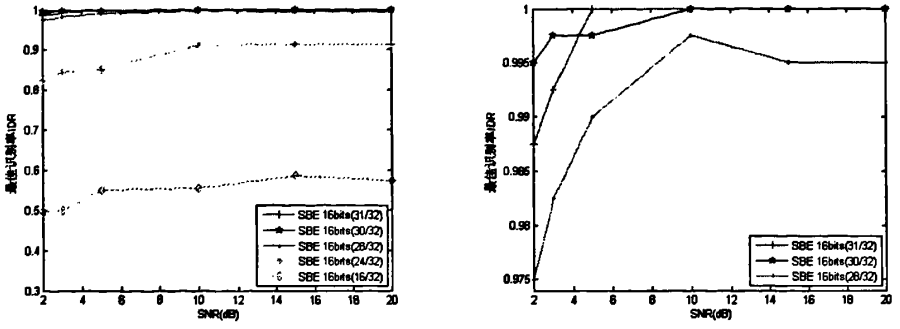


图 3-5 不同交叠因子 P 对加性高斯白噪声的鲁棒性

由图 3-6 可知，当 P=31/32、30/32、28/32 时，BER 随着交叠因子 P 的减小而减小；当 P=24/32、16/32 时，BER 随着交叠因子 P 的减小而增大。显然，当

P=28/32 时，对加性高斯白噪声的鲁棒性最好。



（a） （b）

图 3-7 不同交叠因子 P 在加性高斯白噪声下的最佳识别率 IDR

图 3-7.（b）是对图 3-7.（a）进行了局部放大处理。由图 3-7.（a）可知，当 P=28/32、24/32、16/32 时，最佳识别率 IDR 随着交叠因子 P 的减小而减小。由图 3-7.（b）可知，P=30/32 时的最佳识别率 IDR 比 P=31/32 时更稳定。显然，当 P=30/32 时，对加性高斯白噪声的最佳识别率 IDR 在总体性能最好。

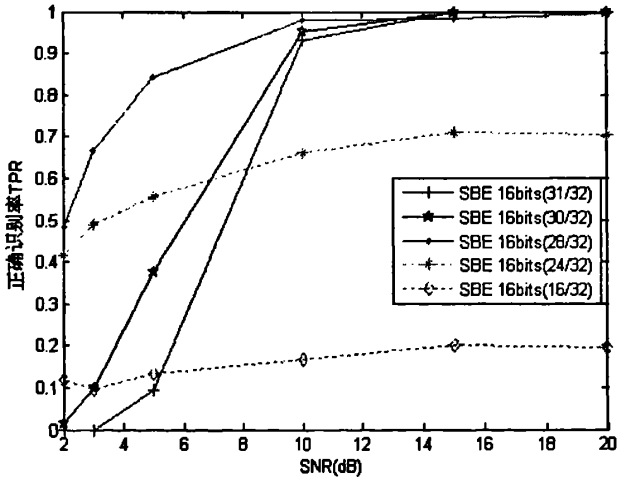


图 3-8 不同交叠因子 P 在加性高斯白噪声下的正确识别率 TPR

由图 3-8 可知，设定了 T=0.35 后，对于 P=31/32、30/32、28/32，当信噪比 SNR<10dB 时，P=28/32 正确识别率 TPR 较高，在 SNR=2dB 时正确识别率 TPR 接近 50%，当信噪比 SNR>10dB 时，P=30/32 总体性能最好。

综合上述两个实验可知，当 P=30/32 时，既能保证算法的鲁棒性，同时能提高算法的识别率。

3.4 算法比较

本节将改进的算法 SBE 与 Haitsma 和 Bellettini 算法进行对比。对常见的保留信号内容的攻击处理的鲁棒性及识别率分别如图 3-9 和图 3-10 所示。

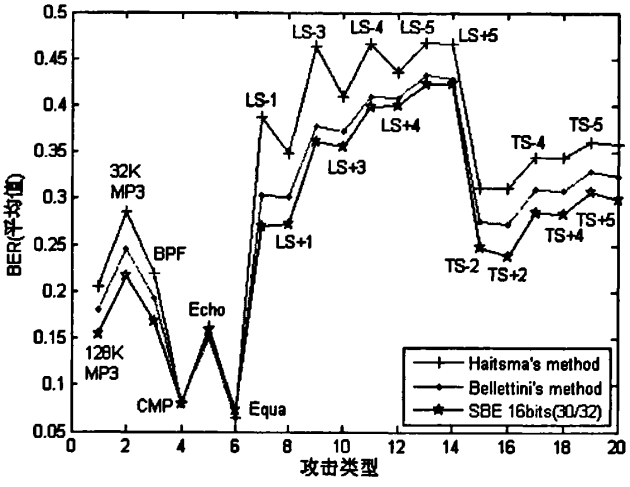
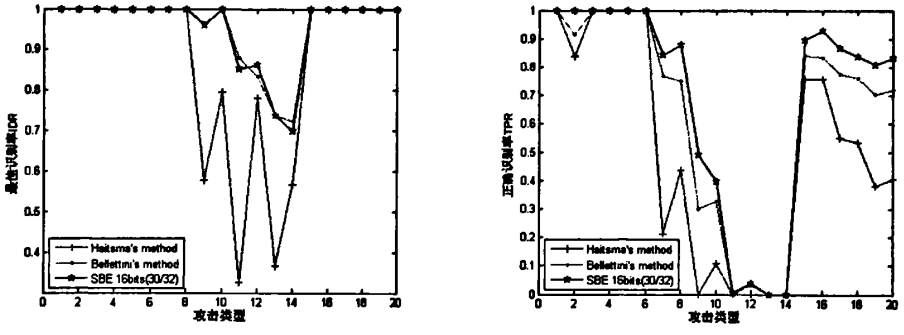


图 3-9 不同算法在不同攻击下的鲁棒性



（a） （b）

图 3-10 不同算法在不同攻击下的识别率

由图 3-9 和图 3-10 可知，改进的算法 SBE 对常见的保留信号内容的攻击处理具有更高的鲁棒性。

三种算法在不同程度加性高斯白噪声下的鲁棒性及识别率分别如图 3-11 和图 3-12 所示。

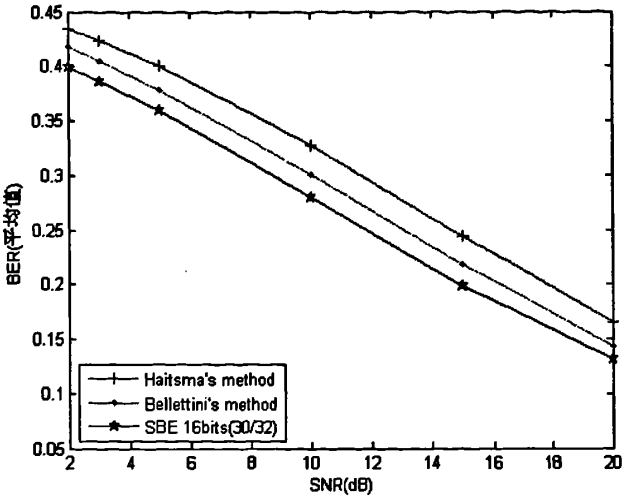
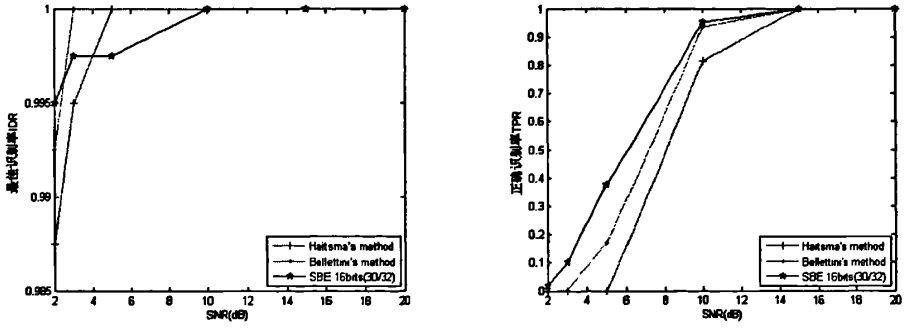


图 3-11 不同算法对加性高斯白噪声的鲁棒性



(a) (b)

图 3-12 不同算法在加性高斯白噪声下的识别率

由图 3-11 可知，改进的算法 SBE 在不同程度加性高斯白噪声下的 BER 低于其它两种算法。而从图 3-12.（a）可知，忽略 T 取值的影响，当信噪比 SNR>10dB 时，最佳识别率 IDR 达到 100%；随着信噪比 SNR 的减小，改进的算法 SBE 的最佳识别率 IDR 有所下降，但在 SNR=2dB 时，其识别率高于其它两种算法。从图 3-12.（b）可知，设定了 T=0.35 后，改进的算法 SBE 在不同程度加性高斯白噪声下的正确识别率 TPR 均高于其它两种算法。

此外，对于 3.3s 的音频片段，采用 Haitsma 算法、Bellettini 算法和改进的算法 SBE 提取的指纹块大小为分别为 256×32 bits、256×16 bits 和128×16 bits，由此可见，改进的算法 SBE 大大节省了存储空间及系统运算时间。

综合上述实验结果表明，改进的算法 SBE 优于 Haitsma 和 Bellettini 的算法。

3.5 小结

本章提出的改进算法，引入了每帧音频信号的能量，利用频率带能量 SBE 替换频率子带能量，通过选择合适的交叠因子 P=30/32，对每 3.3s 的音频片段提取128×16 bits 的音频指纹块。实验结果表明，该算法不仅对常见的保留信号内容的攻击处理具有很好的鲁棒性，对加性高斯白噪声也具有很好的鲁棒性。在保证识别率的情况下，节省了指纹块的存储空间及系统的运算时间。

第四章 基于 Daubechies 小波变换的时频域音频指纹算法

4.1 引言

小波分析是数字信号处理中非常重要的工具，在图像处理、语音信号处理以及数据压缩处理等许多领域中得到了极其广泛的应用。目前，小波分析在音频指纹技术中的应用主要分为以下两种形式：

一、利用小波变换直接对音频信号进行若干次分解，实现音频指纹的提取。文献[28]采用一维连续 Morlet 小波变换提取音频特征，构建了分别用于认证和识别的音频指纹。文献[29]提出了基于平衡多小波（BMW）的音频哈希算法。首先对每帧音频信号做 5 层平衡多小波分解，将得到的 5 个分量划分为 32 个不同的子带；接着采用大小为 5 的窗口对 32 个子带的系数进行估计量化（EQ, Estimation Quantization）；最后计算每个子带系数的方差的对数以及子带方差对数的均值，根据两者的大小关系对每帧音频信号提取 32 bits 的音频指纹。实验结果表明，该算法对低通滤波、高通滤波以及 MP3 压缩等保留音频内容的攻击处理具有较好的鲁棒性。其主要缺点是抵抗线性速度变化攻击的效果不理想以及指纹块的存储空间较大。

二、结合计算机视觉技术，将音频信号转换为时频图，再利用小波变换实现音频指纹的提取。文献[30]结合计算机视觉，将音频信号的频谱图当作二维的图像进行处理，采用小波变换对 10s 的音频片段提取 860 个描绘算子，引入 Boosting 学习方法对提取的音频指纹进行学习分类。Yan 等人已将该算法应用于实际的音频检索系统中[30]，实验结果表明，该算法在现实噪声背景下具有很高的识别率。文献[31, 32]结合计算机视觉技术与数据流处理技术，采用 Haar 小波变换提取音频指纹。首先将音频数据流转换为频谱图，运用 Haar 小波对音频数据流的频谱图进行分解，提取 t 个小波系数，接着利用 Min Hash 技术建模得到音频指纹，最

后采用位置敏感哈希（LSH）技术实现音频指纹检索。实验结果表明，该算法对时间偏置、加添回声、均衡、MP3 压缩、GSM 编码以及时间尺度拉伸（TSM）处理具有很高的识别率，但在噪声和线性速度变化攻击方面识别率相对较低。此外，分析了算法计算复杂度、音频指纹存储空间和识别率之间的关系。而文献[33]在此基础上对系统的参数选择进行分析与验证，并将实验结果与文献[30]进行比较。

为了弥补算法[29, 31]对线性速度变化攻击鲁棒性较差的不足，本章提出了一种基于 Daubechies 小波变换的时频域音频指纹算法，直接对音频信号进行 8 层小波分解，根据每个分量小波系数的方差之间的关系进行音频指纹的提取。由于小波变换具有很好的时频分辨率，符合人耳的听觉特性，因此算法具有很高的鲁棒性。实验结果表明，该算法不仅对常见的保留信号内容的攻击处理和加性高斯白噪声具有很好的鲁棒性，对线性速度变化攻击也具有很好的鲁棒性。

4.2 算法的实现

时频域音频指纹算法的原理框图如图 4-1 所示。

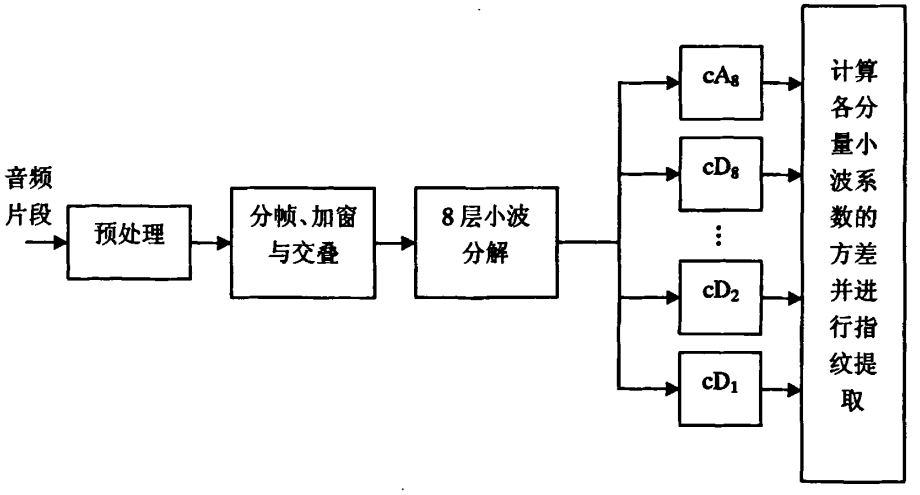


图 4-1 时频域音频指纹算法的原理框图

假设输入的音频信号采样率为 44.1KHz，采用 16bits PCM 编码格式。指纹提取的具体过程如下：

（1）预处理，将输入音频信号下采样为 5KHz 的单声道信号。

（2）分帧、加窗与交叠，帧长为 0.37s，采用汉宁窗，交叠因子为 P。

（3）采用小波基 db6 对每一帧音频信号进行 8 层小波分解，得到 1 个逼近分量 cA_8 和 8 个细节分量 cD_1-cD_8 共 9 个分量。

（4）计算每个分量小波系数的方差，用 $\sigma(n,m)$ 表示第 n 帧第 m 分量小波系数的方差。

（5）采用公式（4-1）进行音频指纹提取，对每帧经过 8 层小波分解的音频信号提取 7bits 的音频指纹。

$$F(n,m)=\begin{cases}1, & \Delta\sigma(n,m)-\Delta\sigma(n,m+1)-(\Delta\sigma(n+1,m)-\Delta\sigma(n+1,m+1))>0\\ 0, & \Delta\sigma(n,m)-\Delta\sigma(n,m+1)-(\Delta\sigma(n+1,m)-\Delta\sigma(n+1,m+1))\le 0\end{cases} \quad (4\text{-}1)$$

式中 $F(n,m)$ 表示第 n 帧第 m 比特的指纹，$\Delta\sigma(n,m)$ 为第 n 帧相邻分量小波系数方差之间的差，其定义如下式所示：

$$\Delta\sigma(n,m)=\sigma(n,m)-\sigma(n,m+1) \quad (4\text{-}2)$$

对应于 3.3s 的音频片段，采用不同的交叠因子 P=31/32、30/32、28/32、24/32、16/32，利用上述算法提取得到大小为 $L\times 7$ bits 的指纹块，其中 L 的取值为 $L=256,128,64,32,16$。在 8 层小波分解所得到的 9 个分量中，细节分量 cD_1 和 cD_2 所包含的小波系数个数最多，其方差最能反映信号的变化规律。因此，引入 cD_1 和 cD_2 分量方差之间的关系，采用公式(4-3)提取 1bit 音频指纹位作为 $F(n,8)$，与原来所提取的指纹共同构成 $L\times 8$ bits 的音频指纹块。

$$F(n,8)=\begin{cases}1, & \sigma(n,8)-\sigma(n,9)-(\sigma(n+1,8)-\sigma(n+1,9))>0\\ 0, & \sigma(n,8)-\sigma(n,9)-(\sigma(n+1,8)-\sigma(n+1,9))\le 0\end{cases} \quad (4\text{-}3)$$

式中 $\sigma(n,8)$、$\sigma(n,9)$ 分别为第 n 帧 cD_2、cD_1 分量的方差。

为了方便表述，下文采用 Wavelet1 表示提取 $L\times 7$ bits 的音频指纹算法，用 Wavelet2 表示提取 $L\times 8$ bits 的音频指纹算法。

从音频“O Fortuna”中截取 3.3s 的音频片段，采用 Wavelet1 算法和 Wavelet2 算法，取交叠因子 P=28/32，分别提取得到 64×7 bits 和 64×8 bits 的音频指纹块，结果分别如图 4-2 和图 4-3 所示。图中，黑像素点代表‘1’，白像素点代表‘0’。图 4-2.（a）和 4-3.（a）为源音频指纹块，图 4-2.（b）和 4-3.（b）为 128K MP3

压缩处理后提取的音频指纹，图 4-2.（c）和图 4-3.（c）为处理后的音频指纹与源音频指纹的误码率图，其 BER 分别为 0.2098 和 0.1992。

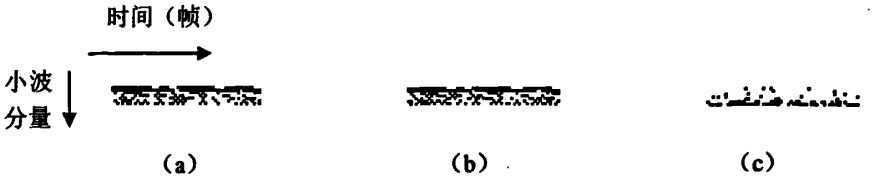


图 4-2 Wavelet1 算法音频指纹及误码率图

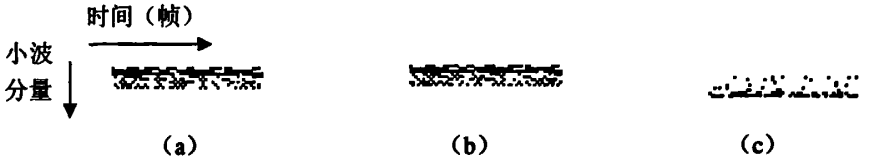


图 4-3 Wavelet2 算法音频指纹及误码率图

4.3 实验结果及分析

实验中采用了 Haitsma 论文中所使用的 4 首测试音频，分别为“O Fortuna”、“Success has made a failure of our home”、“Say what you want”和“A whole lot of Rosie”。对于所有的音频片段，除了采用与 3.3 节相同的攻击处理外，还增加了线性速度变化的比例，从 ±1% 到 ±10%，并通过实验验证本章所提出的算法对线性速度变化具有很好的鲁棒性。

4.3.1 误码率（BER）分析

本文采用归一化汉明距（即误码率 BER）来衡量两个音频片段之间的相似度，为了能够进行正确识别，必须设定合适的阈值 T。在文献[14]中取 T=0.35，而文献[29]中则取 T=0.25。下面通过对受攻击处理的音频片段与源音频片段、受攻击处理的音频片段与不同音频片段之间的误码率进行分析，确定 T 的取值。

测试音频经过 128Kbps MP3 压缩、32Kbps MP3 压缩、*LS* ±1%、*LS* +3%、和 *TS* +2% 共六种攻击处理，对每个攻击样本随机选取 100 个起始点截取 3.3s 音频片段进行测试，并对其误码率进行统计，此处选取交叠因子 P=28/32，具体结

果如图 4-4 和图 4-5 所示。

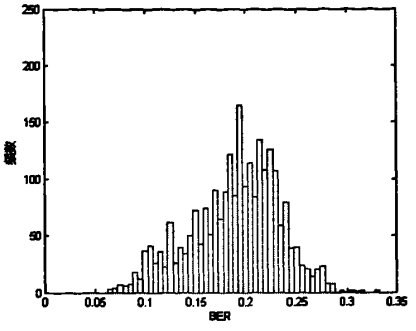


图 4-4 受攻击处理的音频片段指纹与源音频片段指纹之间的误码率 BER

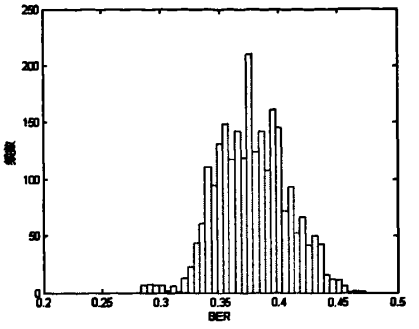


图 4-5 受攻击处理的音频片段指纹与不同音频片段之间的误码率 BER

由图 4-4 和图 4-5 可知，受攻击处理的音频片段与源音频片段之间的误码率 BER 主要集中在 0.1-0.25 之间，有部分情况超过了 0.25，主要是由于随机选取起始点所导致的信号不同步所造成；而受攻击处理的音频片段与不同音频片段之间的误码率 BER 均大于 0.25，且呈近似正态分布。因此，本章沿用文献[29]中 T 的取值，取 T=0.25。

4.3.2 鲁棒性测试

为了验证不同交叠因子 P 对 Wavelet1 算法和 Wavelet2 算法鲁棒性的影响，选取了 P=31/32、30/32、28/32、24/32、16/32，实验中对每个攻击样本随机选取 100 个起始点截取 3.3s 音频片段进行测试。

实验中，沿用 3.3.2 节中所定义的误码率 BER、最佳识别率 IDR 和正确识别率 TPR 来衡量算法的鲁棒性。

一、对于常见的保留信号内容的攻击处理的鲁棒性

采用 Wavelet1 算法和 Wavelet2 算法，对于常见的保留信号内容的攻击处理，待测试音频片段指纹与源音频指纹之间的 BER 平均值如图 4-6 和图 4-7 所示。

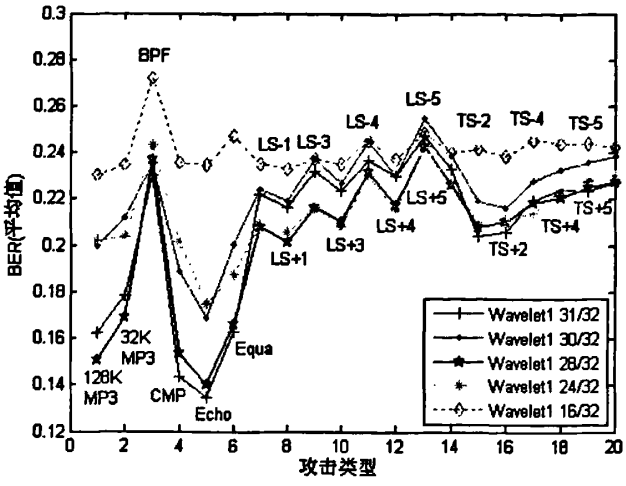


图 4-6 Wavelet1 算法在不同攻击下的鲁棒性

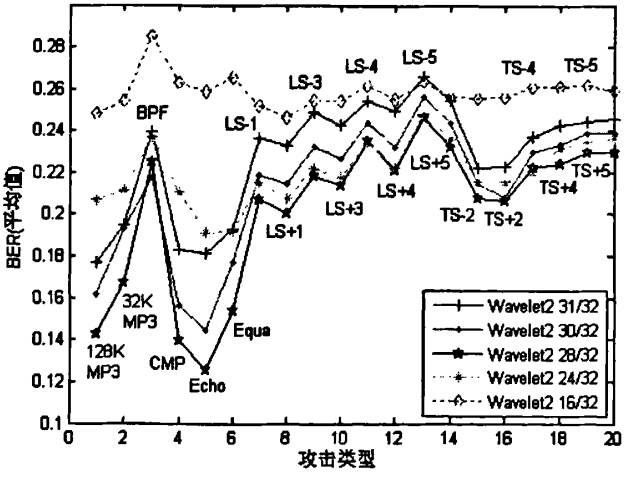
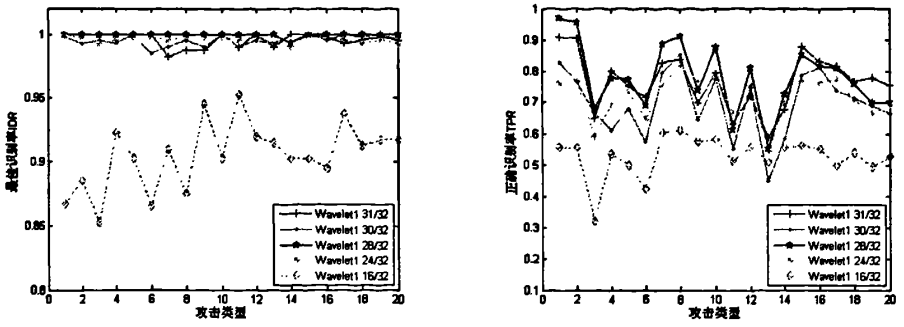


图 4-7 Wavelet2 算法在不同攻击下的鲁棒性

由图 4-6 可知，Wavelet1 算法的 BER 随着交叠因子 P 的减小而增大，且 P=28/32 为临界点，此时的 BER 要比 P=30/32 时的 BER 小。显然，当 P=28/32 时，对不同攻击情况下的 BER 效果最好，且其平均值都小于 0.25。

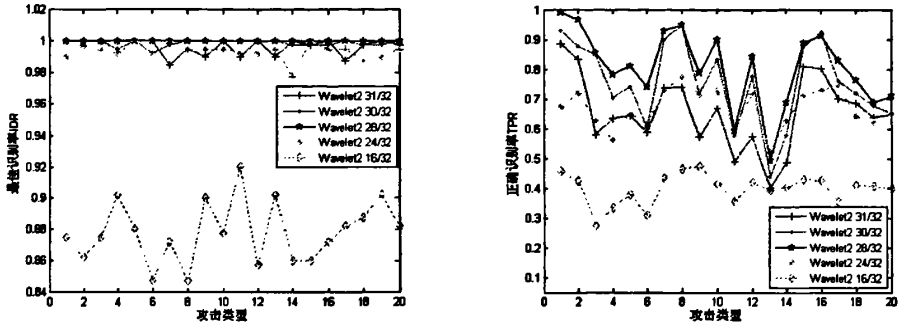
由图 4-7 可知，当 P=31/32、30/32、28/32 时，Wavelet2 算法的 BER 随着交叠因子 P 的减小而减小；当 P=24/32、16/32 时，Wavelet2 算法的 BER 随着交叠因子 P 的减小而增大；显然，对于 Wavelet2 算法，当 P=28/32 时，对不同攻击情况下的 BER 效果最好，其平均值都小于 0.25。

Wavelet1 算法和 Wavelet2 算法对于常见的保留信号内容的攻击处理的识别率分别如图 4-8 和图 4-9 所示。



(a) (b)

图 4-8 Wavelet1 算法在不同攻击下的识别率



(a) (b)

图 4-9 Wavelet2 算法在不同攻击下的识别率

由图 4-8 和图 4-9 可知，当 P=28/32 时，Wavelet1 算法和 Wavelet2 算法的最佳识别率 IDR（除 Wavelet1 算法在 LS+5 攻击处理外）均达到 100%，此时正确识别率 TPR 都达到最大值，这也与 Wavelet1 算法和 Wavelet2 算法在不同攻击下的鲁棒性实验结果相符合。

二、对加性高斯白噪声的鲁棒性

Wavelet1 算法和 Wavelet2 算法在不同程度下的加性高斯白噪声的鲁棒性结果如图 4-10 和图 4-11 所示。实验中，加性高斯白噪声的信噪比分别设置为 20 dB、15 dB、10 dB、5 dB、3 dB、2dB。

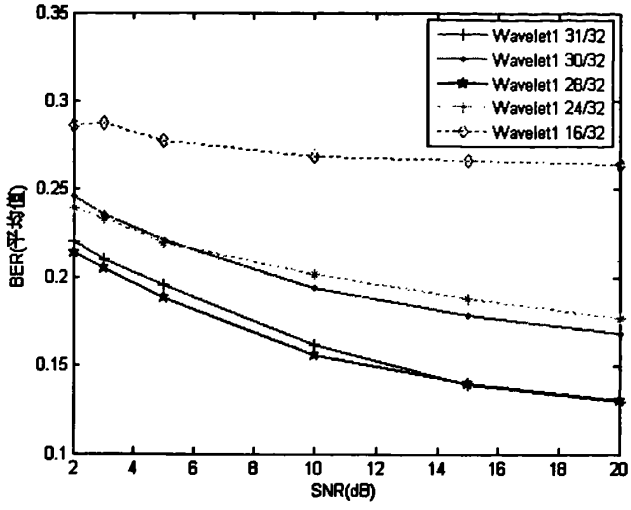


图 4-10 Wavelet1 算法对加性高斯白噪声的鲁棒性

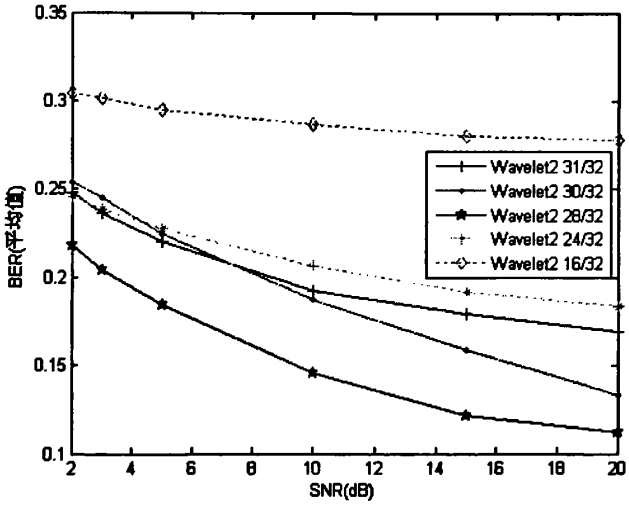
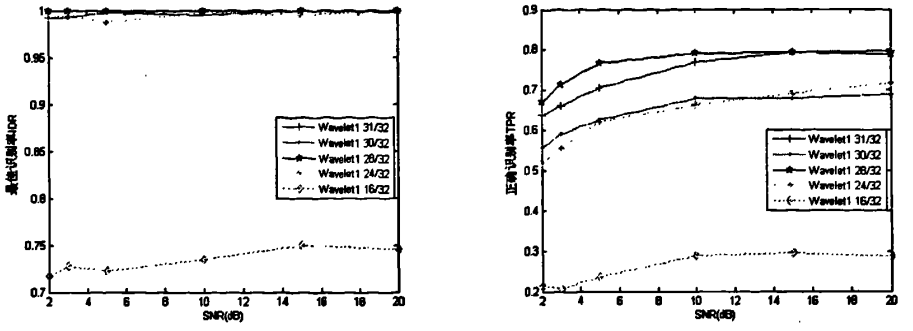


图 4-11 Wavelet2 算法对加性高斯白噪声的鲁棒性

由图 4-10 可知，Wavelet1 算法的 BER 随着交叠因子 P 和信噪比 SNR 的减小而增大，且 P=28/32 为临界点，此时的 BER 要比 P=30/32 时的 BER 小。显然，当 P=28/32 时，对加性高斯白噪声的鲁棒性最好，其平均值都小于 0.25。

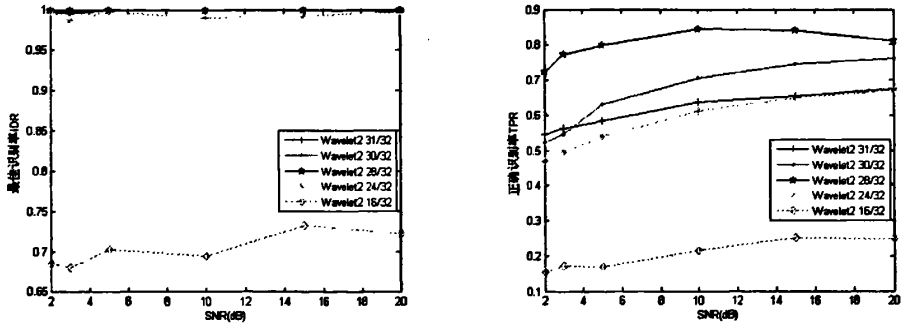
由图 4-11 可知，当 P=31/32、30/32、28/32 时，Wavelet2 算法的 BER 随着交叠因子 P 的减小和信噪比 SNR 的增大而减小；当 P=24/32、16/32 时，Wavelet2 算法的 BER 随着交叠因子 P 和信噪比 SNR 的减小而增大；显然，对于 Wavelet2 算法，当 P=28/32 时，对加性高斯白噪声的鲁棒性最好，其平均值都小于 0.25。

Wavelet1 算法和 Wavelet2 算法在不同程度下的加性高斯白噪声的识别率分别如图 4-12 和图 4-13 所示。



（a） （b）

图 4-12 Wavelet1 算法在不同程度下的加性高斯白噪声的识别率



（a） （b）

图 4-13 Wavelet2 算法在不同程度下的加性高斯白噪声的识别率

由图 4-12.（a）和图 4-13.（a）可知，当 P=28/32 时，Wavelet1 算法和 Wavelet2 算法的最佳识别率 IDR 均达到了 100%。由图 4-12.（b）和图 4-13.（b）可知，当设定了 T=0.25 后，正确识别率 TPR 的变化趋势与其鲁棒性测试结果相似，当 P=28/32 时，识别效果最好。显然，这也与 Wavelet1 算法和 Wavelet2 算法在不同程度下的加性高斯白噪声的鲁棒性实验结果相符合。

三、对线性速度变化攻击处理的鲁棒性

Wavelet1 算法和 Wavelet2 算法对线性速度变化攻击处理的鲁棒性结果如图 4-14 和图 4-15 所示。实验中，设置线性速度变化的比例从 ±1% ~ ±10% 共 20 种。

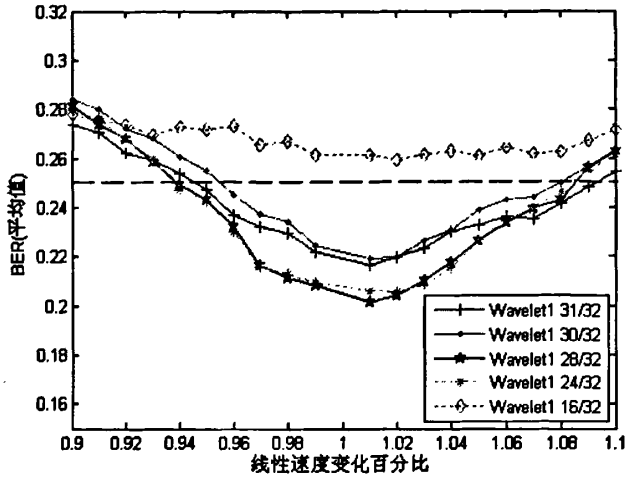


图 4-14 Wavelet1 算法对线性速度变化攻击处理的鲁棒性

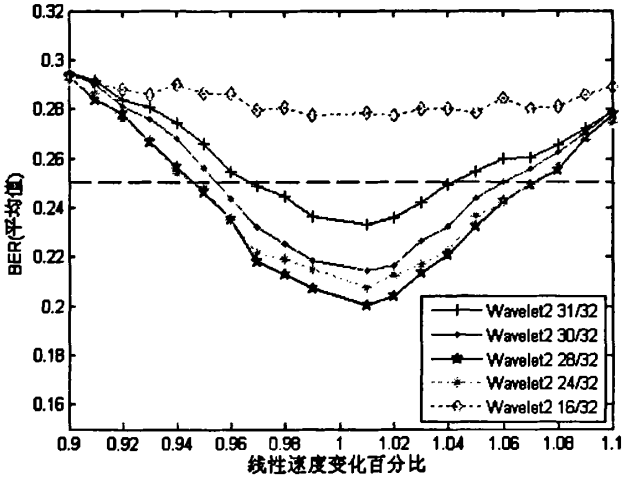
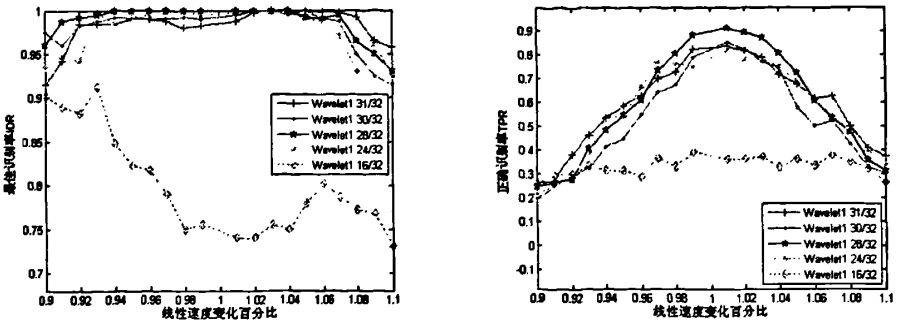


图 4-15 Wavelet2 算法对线性速度变化攻击处理的鲁棒性

由图 4-14 可知，Wavelet1 算法的 BER 随着交叠因子 P 的减小而增大，且 P=28/32 为临界点，此时对线性速度变化攻击处理的鲁棒性在总体性能上最好。

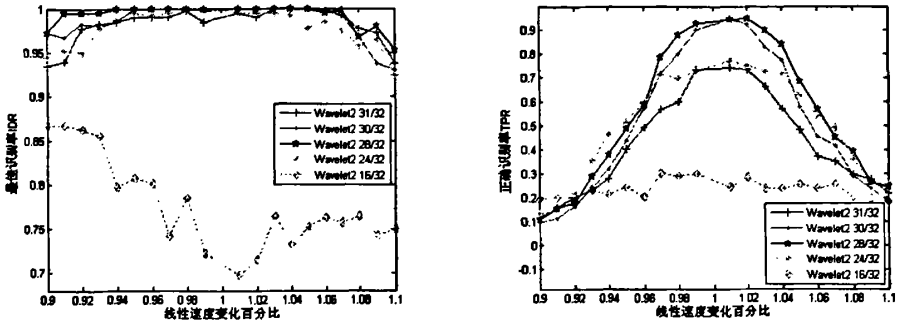
由图 4-15 可知，当 P=31/32、30/32、28/32 时，Wavelet2 算法的 BER 随着交叠因子 P 的减小而减小；当 P=24/32、16/32 时，Wavelet2 算法的 BER 随着交叠因子 P 的减小而增大；显然，对于 Wavelet2 算法，当 P=28/32 时，对线性速度变化攻击处理的鲁棒性最好。

Wavelet1 算法和 Wavelet2 算法对线性速度变化攻击处理的识别率分别如图 4-16 和图 4-17 所示。



(a) (b)

图 4-16 Wavelet1 算法对线性速度变化攻击处理的识别率



(a) (b)

图 4-17 Wavelet2 算法对线性速度变化攻击处理的识别率

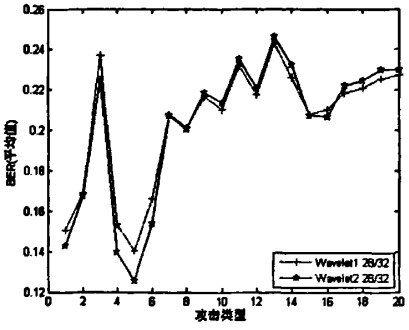
由图 4-16 和图 4-17 可知，当 P=28/32 时，Wavelet1 算法和 Wavelet2 算法最佳识别率 IDR 和正确识别率 TPR 在总体性能上都达到最优，这也与 Wavelet1 算法和 Wavelet2 算法对线性速度变化攻击处理的鲁棒性实验结果相符合。

综合上述三个实验结果可知，Wavelet1 算法和 Wavelet2 算法在交叠因子 P=28/32 时，其鲁棒性和识别率达到最佳。

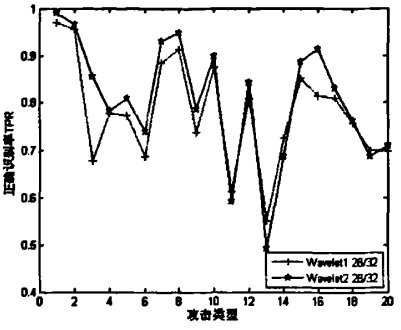
4.4 算法比较

Wavelet1 和 Wavelet2 算法在交叠因子 P=28/32 时，对常见的保留信号内容的攻击处理的鲁棒性及识别率如图 4-18（a）和图 4-18（b）所示，在不同程度下的加性高斯白噪声的鲁棒性及识别率如图 4-18（c）和图 4-18（d）所示，鉴于 Wavelet1 和 Wavelet2 算法在 P=28/32 时对不同类型攻击的最佳识别率 IDR(除

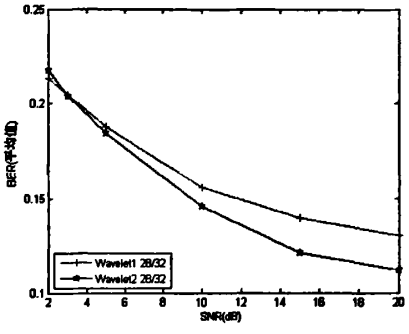
Wavelet1 算法在 LS+5 攻击处理外）均达到 100%，此处不再对最佳识别率 IDR 进行比较。对线性速度变化攻击处理的鲁棒性及识别率如图 4-19 和图 4-20 所示。



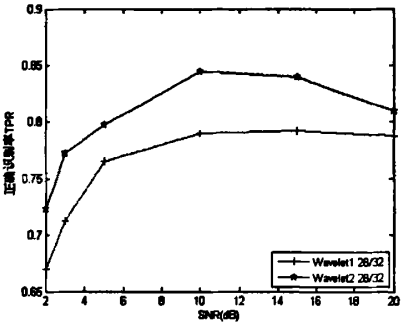
（a）对不同攻击的鲁棒性



（b）对不同攻击的识别率



（c）对加性高斯白噪声的鲁棒性



（d）对加性高斯白噪声的识别率

图 4-18 Wavelet1 算法和 Wavelet2 算法比较

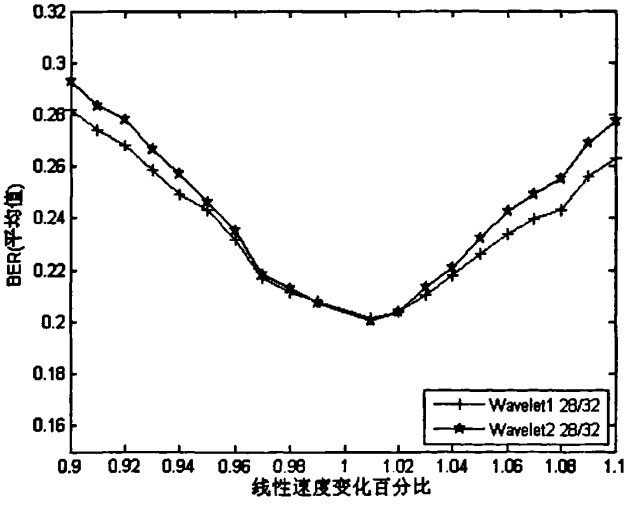
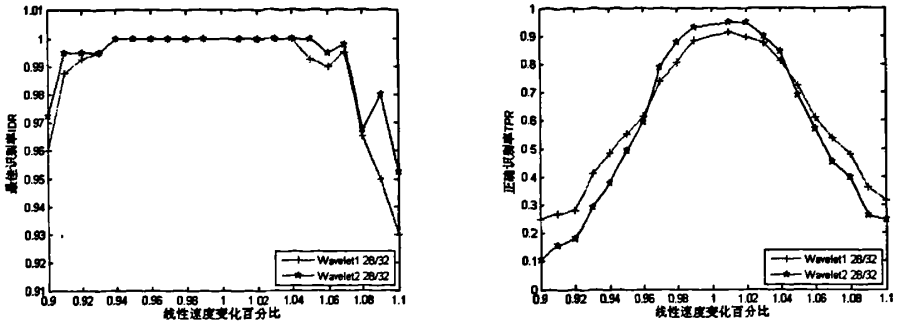


图 4-19 对线性速度变化攻击处理的鲁棒性



(a) (b)

图 4-20 对线性速度变化攻击处理的识别率

由图 4-18~图 4-20 可知，Wavelet2 算法的性能总体优于 Wavelet1 算法。

4.5 小结

本章提出了一种基于 Daubechies 小波变换的时频域音频指纹算法，直接对音频信号进行 8 层小波分解，根据每个分量小波系数的方差之间的关系设计了 Wavelet1 算法，对每 3.3s 的音频信号提取 64×7 bits 的音频指纹块。在 Wavelet1 算法的基础上，引入了细节分量 cD_1 和 cD_2 方差之间的关系，设计了 Wavelet2 算法，对每 3.3s 的音频信号提取 64×8 bits 的音频指纹块。

实验结果表明，Wavelet1 算法和 Wavelet2 算法对常见的保留信号内容的攻击处理和加性高斯白噪声具有很好的鲁棒性，尤其是在抵抗线性速度变化攻击上效果明显，能达到±10%，对应的最佳识别率 IDR 均高于 93%。此外，通过实验发现，Wavelet2 算法总体性能优于 Wavelet1 算法，这是由于引入的细节分量 cD_1 和 cD_2 方差之间的关系具有更高的鲁棒性的结果。

第五章 基于音频指纹的音频检索

5.1 引言

早在20世纪90年代开始，人们就开始了对音频检索的研究，其主要内容是利用音频信息的时域等物理特征，实现基于内容的音频检索。而音频指纹作为音频内容的一个标识，概括了音频听觉上的相关信息，是基于音频内容的紧凑的签名。因此，将音频指纹作为特征进行音频检索得到了广泛的应用[14, 30, 33]。

下面对现有的部分音频检索算法作简要的概述。

文献[41]提出了一种基于树结构的音频指纹高维二值空间邻域检索算法，而常用的多维邻域搜索主要是基于树结构，主要有kd-trees和vp-trees（即Vantage Point）两种方法。文献[42]提出了一种称为时间序列动态搜索（TAS, Time-Series Active Search）的直方图快速检索算法。首先利用矢量量化（VQ, Vector Quantization）对提取的音频特征向量进行矢量量化，接着计算固定窗口大小的特征向量中不同VQ码字的数量，分别生成源音频特征向量直方图及待查询音频特征向量直方图，最后通过两者的相似度比较得到检索结果。文献[43]在此基础上引入了聚类学习方法，大大提高了算法的性能。为了解决感知上相似的两个多媒体信号由于局部特征的不同所导致的相似度距离很大这一问题，文献[44]提出了一种基于局部匹配函数（PMF, Partial Match Function）的检索算法。文献[45]采用位置敏感哈希LSH和局部序列比较（PSC, Partial Sequence Comparison），提出了一个快速有效的基于内容的音频检索框架。文献[33]采用LSH技术实现了快速的音频检索。

以上算法检索速度快，检索正确率高，但是实现比较复杂。因此，本章采用文献[19]提出的检索算法，首先通过互相关运算选取候选同步点，再计算其归一化汉明距，最终得到检索结果。该算法实现简单，同时保证了检索正确率。

5.2 算法的实现

音频检索主要包括数据库生成和检索两部分，其原理框图如图 5-1 所示。

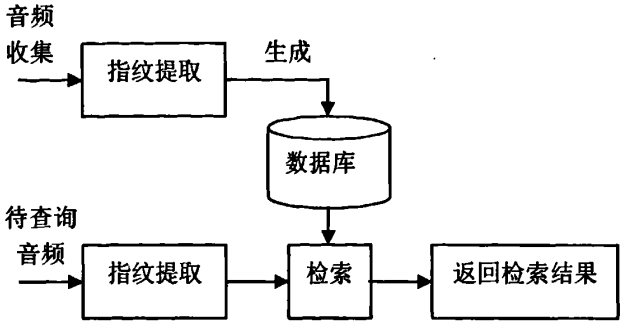


图 5-1 音频检索原理框图

本章采用第三、四章所设计的音频指纹算法生成音频检索数据库，接下来主要讨论音频检索算法的实现。

众所周知，基于音频指纹的音频检索最直接的方法就是将待查询音频指纹与数据库音频指纹两两比较得到检索结果，这种方法也称为穷举搜索（Exhaustive Searching），其原理如图 5-2 所示。

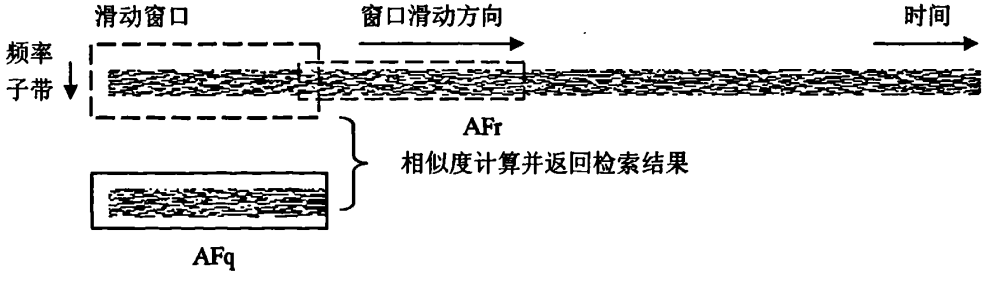


图 5-2 穷举搜索原理框图

图 5-2 中，AFq 为待查询音频指纹，大小为 $W \times b$；AFr 为数据库源音频指纹，大小为 $L \times b$，滑动窗口大小与待查询音频指纹块大小相同，即 $W \times b$。

检索过程如下：首先用大小为 $W \times b$ 的滑动窗口对数据库中的每一个源音频指纹 AFr 划分为若干部分，接着计算划分后的每一个指纹块与待查询音频指纹 AFq 的相似度距离，选取最小的相似度距离作为该源音频指纹与待查询音频指纹 AFq 之间的距离，最后重复上述步骤将数据库中所有音频指纹与待查询音频指纹

AFq 作比较并返回最终结果。穷举搜索的缺点是其检索速度较慢。

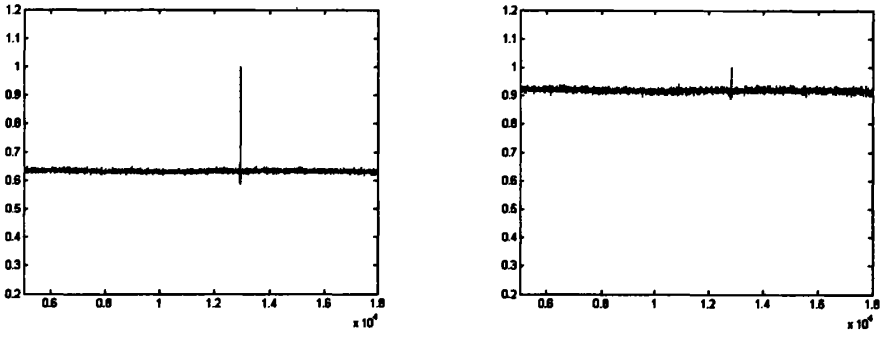
因此，本章采用文献[19]的检索算法，与文献[14]相比，在保证鲁棒性的同时实现简单。算法的主要思想如下：

首先，选取同步点。计算待查询音频指纹 AFq 与数据库源音频指纹 AFr 时间轴上的归一化互相关系数平均值，选取平均值最大的 s 个点作为候选同步点。

其次，对于选定的 s 个同步点，分别计算其与待查询音频指纹 AFq 的相似度距离 D，选取距离最小值 D_{min} 作为待查询音频指纹 AFq 与该源音频指纹 AFr 的距离。

最后，对于数据库中所有的源音频指纹，重复上述步骤，返回最终检索结果。

图 5-3 为待查询音频指纹 AFq 与数据库源音频指纹 AFr 时间轴上的归一化互相关系数平均值示意图。



（a）　　（b）

图 5-3 归一化互相关系数平均值示意图

图 5-3 中，图（a）为未受攻击处理的音频指纹与源音频指纹在时间轴上的归一化互相关系数平均值，图（b）为受到线性速度变化（-1%）攻击处理的音频指纹与源音频指纹在时间轴上的归一化互相关系数平均值。由图 5-3 可知，即使音频片段受到攻击处理，其音频指纹与源音频指纹在时间轴上的归一化互相关系数平均值也存在明显的极值点。为了保证在检索过程中能够选取到正确的同步点，文献[19]认为选取 s=10 已经足够。为了确保实验的正确性，在本章的所有实验中，s 的取值为 50。

5.3 实验结果及分析

本章实验中所采用的数据库包含200首长度为30s的音频片段，格式为wav格式。分别采用本文第三、四章中所提出的音频指纹算法生成数据库 DB_{SBE}、$DB_{Wavelet1}$、$DB_{Wavelet2}$。为了将实验结果与现有算法作比较，实验中还采用文献[14]和文献[19]中的音频指纹算法生成数据库 $DB_{Haitsma}$ 和数据库 $DB_{Bellettini}$。数据库中的所有音频片段经过攻击处理后形成待查询音频片段，攻击处理如下：

（1）128Kbps 和 32Kbps MP3 压缩。

（2）二阶巴特沃斯（Butterworth）带通滤波（BPF）：通带频率为 100Hz-6000Hz。

（3）幅度压缩（Compression），具体设置如下：当幅度 $|A| \geq -28.6$dB 时，压缩比为 8.94:1；当 $-46.4\text{dB} < |A| < -28.6\text{dB}$ 时，压缩比为 1.73:1；当 $|A| \leq -46.4$dB 时，压缩比为 1:1.61。

（4）添加回声（Echo）。

（5）均衡（Equalization），采用典型的 10 频段均衡器，具体设置如 3.3 节表 3-2 所示。

（6）时间尺度拉伸（Time Scale Modification），-2%、+2%、-4%、+4%、-5%和+5%。时间尺度拉伸保持基音频率不变，只对时间进行拉伸。

（7）线性速度变化（Linear Speed Change），-1%、+1%、-3%、+3%、-4%、+4%、-5%和+5%。线性速度变化对时间和基音频率都进行拉伸。

（8）添加加性高斯白噪声，信噪比分别为 20dB、15dB、10dB、5dB、3dB、2dB。

所有的音频片段均经过如上共26种攻击处理，生成5200首待查询音频片度。对于每一首经过攻击处理的待查询音频片段，我们随机选取 5 个起始点截取 3.3 秒的音频片段进行检索，因此，对于每一种类型的攻击处理，共进行 1000 次检索。在检索过程中，采用与生成数据库相同的五种音频指纹算法提取待查询音频片段的音频指纹，同时，忽略了 T 取值的影响。

本章所有实验数据中所涉及的检索正确率用最佳识别率 IDR 和检索成功率（SR, Success Rate）表示。检索成功率 SR 的目的是为了验证算法的正确性，用于判断距离最小的三个音频是否与待查询音频片段匹配。检索成功率 SR 的定义

如公式（5-1）所示。

$$SR = \frac{\text{距离最小三个中任一正确匹配样本数之和}}{\text{检索总次数}} \quad (5\text{-}1)$$

5.3.1 对于常见的保留信号内容的攻击处理的检索结果

对于常见的保留信号内容的攻击处理的检索结果如表 5-1 和表 5-2 所示。

表 5-1 不同算法在不同攻击下的检索结果（IDR）

攻击处理类型	Haitsma	Bellettini	SBE	Wavelet1	Wavelet2
128K MP3	100	100	99.9	95.3	98.8
32K MP3	100	99.9	99.8	94.6	98.6
BPF	100	100	100	77.3	93.2
Compression	98.7	99.1	99.6	79.4	87.6
Echo	98.8	99.2	99.3	83.3	88.8
Equalization	98.9	99.2	99.5	77	87.4
LS-1	99.7	100	99.6	93.5	99
LS+1	99.6	99.9	99.8	91.8	98.1
LS-3	6	97.3	96.6	87.5	95.6
LS+3	96.1	95.3	93.8	85.9	94.6
LS-4	2.3	85	79.9	85.5	94.8
LS+4	88	77.8	74.4	81.4	92.9
LS-5	1.5	44.1	37.1	79.1	90.9
LS+5	8.8	45.3	38.4	74.9	88.9
TS-2	99.9	99.9	99.9	93.5	98.3
TS+2	100	100	99.8	93.3	98.6
TS-4	99.9	99.9	99.8	90.2	96.7
TS+4	99.9	99.8	99.8	90.8	97
TS-5	100	99.9	99.7	87.6	94.8
TS+5	99.9	99.8	99.7	90.9	96

由表 5-1 可知，采用 Haitsma、Bellettini 和改进的算法 SBE 提取音频指纹，在 MP3 压缩、带通滤波、幅度压缩、添加回声、均衡和时间尺度拉伸攻击处理下的最佳识别率 IDR 比 Wavelet1 和 Wavelet2 算法要高，但是在线性速度攻击处理下情况恰好相反。而在各种类型攻击处理下，采用算法 Wavelet2 提取音频指纹进行检索的最佳识别率 IDR 均要比算法 Wavelet1 高出 3.5%（128K MP3 攻击）~ 15.9%（BPF 攻击），这也表明了算法 Wavelet2 提取的音频指纹具有更高的鲁棒性。

表 5-2 不同算法在不同攻击下的检索结果（SR）

攻击处理类型	Haitsma	Bellettini	SBE	Wavelet1	Wavelet2
128K MP3	100	100	99.9	96.2	98.9
32K MP3	100	100	99.8	95.5	98.8
BPF	100	100	100	82.7	95.8
Compression	98.8	99.1	99.6	83.7	90.3
Echo	99	99.2	99.4	86.4	91
Equalization	99.2	99.2	99.5	80.3	89.3
LS-1	99.8	100	99.7	96.2	100
LS+1	99.8	99.9	99.8	95.2	98.7
LS-3	9.5	97.6	97.2	92.3	97.6
LS+3	96.7	96.7	95.8	89.8	96.9
LS-4	4.2	89.9	85.4	91	97.1
LS+4	90.3	86	81.3	87.7	95.5
LS-5	2.9	55.8	46	87.1	94.8
LS+5	13.4	56.5	48.5	83.5	93.3
TS-2	99.9	99.9	99.9	96.1	98.6
TS+2	100	100	99.8	95.2	99.3
TS-4	99.9	99.9	99.8	94.3	98
TS+4	99.9	99.8	99.8	94.6	98.3
TS-5	100	99.9	99.8	91.8	96.2
TS+5	99.9	99.9	99.9	93.3	97.9

综合表 5-1 和表 5-2 可知，相同的算法在同样的攻击处理类型下 SR 比 IDR 要高，尤其是在线性速度变化（LS-5）攻击时，采用 Bellettini 算法提取音频指纹进行检索的正确率改善了 11.7%。而在各种类型攻击处理下，Wavelet1 和 Wavelet2 算法改善的分别是 0.9%（128K MP3 攻击）~ 8.6%（LS+5 攻击）和 0.1%（128K MP3 攻击）~ 4.4%（LS+5 攻击）。显然，Wavelet1 和 Wavelet2 算法所提取的音频指纹更加稳定，尤其是算法 Wavelet2。

5.3.2 在不同程度加性高斯白噪声情况下的检索结果

在不同程度加性高斯白噪声情况下的检索结果如表 5-3 和表 5-4 所示。

表 5-3 不同算法在不同程度加性高斯白噪声下的检索结果（IDR）

信噪比 SNR	Haitsma	Bellettini	SBE	Wavelet1	Wavelet2
20dB	98.7	99.3	98.9	83.6	91.8
15 dB	98.7	99.1	98.5	84.9	92.5
10 dB	98.2	98.5	98.4	85.7	91.9
5 dB	95.4	96.1	96.1	84.7	90.5
3 dB	93.3	93.5	92.5	83.2	88.3
2 dB	89.5	90.9	89.6	82	88.2

由表 5-3 可知，采用 Haitsma、Bellettini 和改进的算法 SBE 提取音频指纹，在不同程度加性高斯白噪声下的 IDR 比 Wavelet1 和 Wavelet2 算法要高，而采用算法 Wavelet2 提取音频指纹进行检索的最佳识别率 IDR 比算法 Wavelet1 要高。

表 5-4 不同算法在不同程度加性高斯白噪声下的检索结果（SR）

信噪比 SNR	Haitsma	Bellettini	SBE	Wavelet1	Wavelet2
20dB	98.9	99.3	98.9	87	93.5
15 dB	98.9	99.2	98.5	87.8	94.4
10 dB	98.3	99	98.5	88.9	93.8
5 dB	95.7	96.4	96.3	87.4	92.4
3 dB	94.5	94.4	93.6	86.4	90.6
2 dB	91.4	91.9	91.8	85.6	90.7

综合表 5-3 和表 5-4 可知，相同的算法在同样的加性高斯白噪声下 SR 比 IDR 要高，而以 Haitsma、Bellettini 和改进的算法 SBE 提取音频指纹，在不同程度加性高斯白噪声下检索的最佳识别率 IDR 比 Wavelet1 和 Wavelet2 算法要高，这也说明了 Wavelet1 和 Wavelet2 算法在抵抗加性高斯白噪声攻击处理方面效果相对较差。

5.3.3 算法性能比较

实验中，共进行了 1000 次正确匹配比较和199×1000次非正确匹配比较，分别计算得到正确识别率 TPR 和误检率 FPR。误检率 FPR 的定义如下：

$$FPR = \frac{\text{不同的音频被错误判断为匹配的样本数}}{\text{测试总次数}} \tag{5-2}$$

不同算法对于常见的保留信号内容的攻击处理的 ROC 曲线如图 5-4 所示，对加性高斯白噪声的 ROC 曲线如图 5-5 所示。

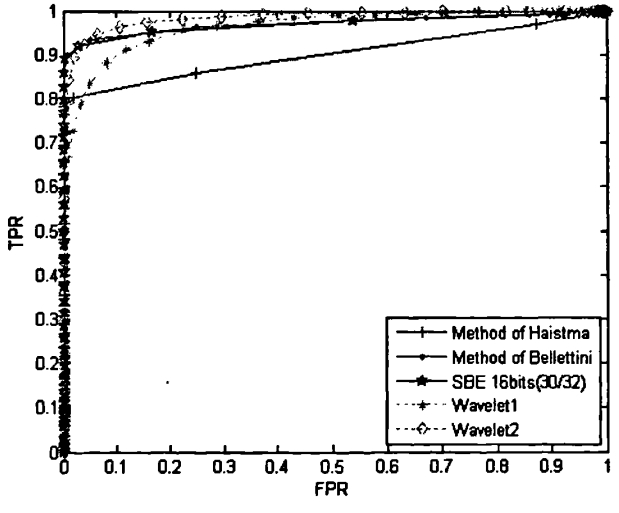


图 5-4 不同算法对于常见的保留信号内容的攻击处理的 ROC 曲线

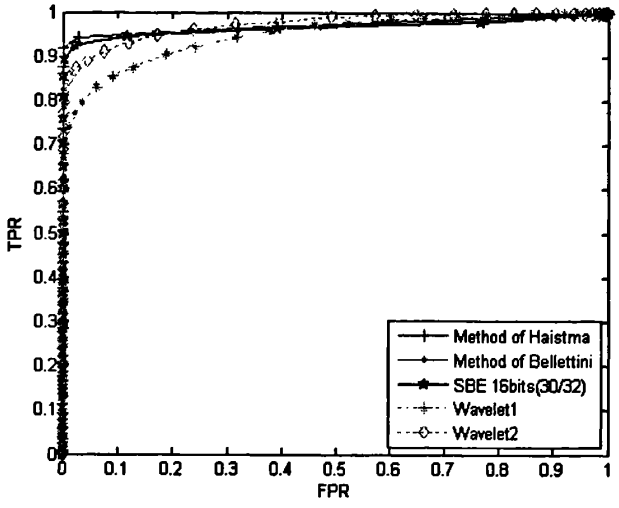


图 5-5 不同算法对加性高斯白噪声的 ROC 曲线

综合图 5-4 和图 5-5 可知， Wavelet2 算法性能最好。

5.4 小结

本章采用论文中所提出的音频指纹算法提取特征，运用文献[19]提出的检索算法进行检索。实验结果表明，论文中所提出的音频指纹算法具有很好的鲁棒性，利用改进的算法 SBE 进行音频检索的最佳识别率 IDR 高达 89.6%（除线性速度变化攻击外），而算法 Wavelet2 的性能要优于算法 Wavelet1，在线性速度变化攻击情况下能达到 88.9%的最佳识别率，不足之处是 Wavelet1 和 Wavelet2 算法对于其它攻击处理的最佳识别率略低于改进的算法 SBE。

结 语

一、全文总结

音频指纹是基于内容的紧凑的签名，概括了音频片断固有的本质特征。目前，国内外对音频指纹的研究仍处于探索阶段。本论文通过对国内外音频指纹算法的分析，提出了两种鲁棒的变换域音频指纹算法，并将其应用于音频检索中。现将本文的主要工作内容总结如下：

（1）对 Haitsma 和 Bellettini 所提出的算法进行研究，提出了基于 STFT 变换的频率域音频指纹的改进算法。该算法引入了每帧音频信号的能量，利用频谱带能量 SBE 替换频率子带能量，通过选择合适的交叠因子 P=30/32，对每 3.3s 的音频信号提取128×16bits 的音频指纹块。实验结果表明，改进的算法 SBE 不仅对常见的保留信号内容的攻击处理具有很好的鲁棒性，对加性高斯白噪声也具有很好的鲁棒性。在保证识别率的情况下，节省了指纹块的存储空间及系统的运算时间。

（2）提出了一种基于 Daubechies 小波变换的时频域音频指纹算法，直接对音频信号进行 8 层小波分解得到 1 个逼近分量和 8 个细节分量，根据每个分量小波系数的方差之间的关系设计了 Wavelet1 和 Wavelet2 算法。实验结果表明，Wavelet1 和 Wavelet2 算法对常见的保留信号内容的攻击处理和加性高斯白噪声具有很好的鲁棒性，尤其是在抵抗线性速度变化攻击上效果明显，能达到±10%，对应的最佳识别率 IDR 均高于 93%。此外，通过实验发现，Wavelet2 算法总体性能优于 Wavelet1 算法，这是由于引入的细节分量 cD_1 和 cD_2 方差之间的关系具有更高的鲁棒性的结果。

（3）将本文提出的音频指纹算法应用于音频检索中。实验结果表明，本文所提出的音频指纹算法具有很好的鲁棒性，利用改进的算法 SBE 进行音频检索的最佳识别率 IDR 高达 89.6%（除线性速度变化攻击外），而算法 Wavelet2 的性能要优于算法 Wavelet1，在线性速度变化攻击情况下能达到 88.9%的最佳识别率。

二、不足与展望

由于时间关系，本文的研究还存在着许多不足之处，主要体现在以下三个方面：

（1）改进的算法 SBE 对线性速度变化攻击处理鲁棒性较差，检索识别率较低。为了解决这一问题，可以在音频指纹数据库生成阶段添加受线性速度变化攻击的音频样本。

（2）基于 Daubechies 小波变换的时频域音频指纹算法虽然对线性速度变化攻击具有很好的鲁棒性，但是对均衡、幅度压缩等处理效果相对较差，这是由于此类处理改变了信号的变化趋势所导致的。可尝试将小波变换与傅里叶变换结合起来提取音频指纹。

（3）未能根据音频指纹而设计相应的检索算法；同时，用于音频检索仿真的数据库有待进一步完善。

参考文献

[1] 李伟，袁一群，李晓强，薛向阳，陆佩忠. 数字音频水印技术综述. 通信学报， February 2005, Vol. 26(Issue 2): pp. 100-111.

[2] P. Cano, E. Batlle, T. Kalker, J. Haitsma. A Review of Audio Fingerprinting. Journal of VLSI Signal Processing Systems, 2005, Vol. 41(Issue 3): 271-284.

[3] Foosic, libFooID Free Fingerprinting Library <http://foosic.org/ >. 2008.

[4] MusicBrainz, Picard, the Next-Generation Musicbrainz Tagger, <http://musicbrainz.org/doc/PicardTagger>. 2008.

[5] E. Gomez, P. Cano, L. D. C. T. Gomes, E. Batlle and M. Bonnet. Mixed Watermarking-Fingerprinting Approach for Integrity Verification of Audio Recordings. Proceedings of the International Telecommunications Symposium. Natal. Brazil. September 2002.

[6] P. J. O. Doets and R. L. Lagendijk. Distortion Estimation in Compressed Music Using Only Audio Fingerprints. IEEE Transactions on Audio, Speech, and Language Processing, February 2008, Vol. 16(Issue 2): 302-317.

[7] 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, Vol. 29(No. 11): 2124-2130.

[8] T. Zhang and C. C. J. Kuo. Hierarchical Classification of Audio Data for Archiving and Retrieving. International Conference on Acoustics, Speech and Signal Processing. 1999. Vol. 6. pp. 3001-3004.

[9] 许刚. 基于内容的音频检索方法研究. 硕士论文. 电子科技大学. 2006.

[10] L. Lu, H. Jiang and H. J. Zhang. A Robust Audio Classification and Segmentation Method. ACM International Conference on Multimedia. Ottawa. Canada. 2001. Vol. 9. pp. 203-211.

[11] A. C. Ibarrola and E. Chavez. A Robust Entropy-Based Audio-Fingerprint. IEEE International Conference on Multimedia and Expo. July 2006. pp. 1729-1732.

[12] D. Fragoulis, G. Rousopoulos, T. Panagopoulos, C. Alexiou and C. papaodysseus. On the Automated Recognition of Seriously Distorted Musical Recordings. IEEE Transactions on

Signal Processing, April 2001, Vol. 49(Issue 4): 898-908.

[13] J. Haitsma, T. Kalker and J. Oostveen. Robust Audio Hashing for Content Identification. Content Based Multimedia Indexing. Brescia. Italy. September 2001.

[14] J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System. Proceedings of International Conference on Music Information Retrieval. 2002.

[15] J. Haitsma and T. Kalker. Speed-Change Resistant Audio Fingerprinting Using Auto-Correlation. International Conference on Acoustics, Speech and Signal Processing. 2003. Vol. 4. pp. 728-731.

[16] P. J .O. Doets and R. L. Lagendijk. Theoretical Modeling of a Robust Audio Fingerprinting System. IEEE Benelux Signal Processing Symposium. 2004. pp. 101-104.

[17] F. Balado, N. J. Hurley, E. P. McCarthy and G. C. M . Silvestre. Performance of Philips Audio Fingerprinting under Additive Noise. International Conference on Acoustics, Speech and Signal Processing. 2007. Vol. 2. pp. 209-212.

[18] F. Balado, J. Hurley, P. McCarthy and C. M. Silvestre. Performance Analysis of Robust Audio Hashing. IEEE Transactions on Information Forensics and Security, June 2007, Vol. 2(Issue 2): 254-266.

[19] C. Bellettini and G. Mazzini. On Audio Recognition Performance via Robust Hashing. Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems. Xiamen. China. 2007. pp. 20-23.

[20] C. Bellettini and G. Mazzini. Reliable Automatic Recognition for Pitch-Shifted Audio. International Conference on Computer Communications and Networks. 2008. pp. 1-6.

[21] C. Burges, J. Platt and S. Jana. Distortion Discriminant Analysis for Audio Fingerprinting. IEEE Transactions on Speech and Audio Processing, May 2003, Vol. 11(Issue 3): 165-174.

[22] A. Ramalingam and S. Krishnan. Gaussian Mixture Modeling of Short Time Fourier Transform Features for Audio Fingerprinting. IEEE Transactions on Information Forensics and Security, December 2006, Vol. 1(Issue 4): 457-463.

[23] M. Betser, P. Collen and J. Rault. Audio Identification Using Sinusoidal Modeling and Application to Jingle Detection. Austrian Computer Society. 2007.

[24] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo. Audio Fingerprinting Based on Normalized Spectral Subband Centroids. International Conference on Acoustics, Speech and

Signal Processing. 2005. Vol. 3. pp. 213-216.

[25] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo. Audio Fingerprinting Based on Normalized Spectral Subband Moments. IEEE Signal Processing Letters, April 2006, Vol. 13(Issue 4): 209-212.

[26] S. Kim and C. D. Yoo. Boosted Binary Audio Fingerprinting Based on Spectral Subband Moments. International Conference on Acoustics, Speech and Signal Processing. 2007. Vol. 1. pp. 241-244.

[27] M. Jin and C. D. Yoo. Temporal Dynamics for Spectral Sub-band Centroid Audio Fingerprints. IEEE International Conference on Multimedia and Expo. 2007. pp. 180-183.

[28] C. S. Lu. Audio Fingerprinting Based on Analyzing Time-Frequency Localization of Signals. Multimedia Signal Processing. 2002. pp. 174-177.

[29] L. Ghouti and A. Bouridane. A Robust Perceptual Audio Hashing Using Balanced Multiwavelets. International Conference on Acoustics, Speech and Signal Processing. 2006. Vol. 5. pp. 209-212.

[30] Y. Ke, D. Hoiem, R. Sukthankar. Computer Vision for Music Identification. In Proceedings of Computer Vision and Pattern Recognition. 2005. pp. 597-604.

[31] S. Baluja and M. Covell. Content Fingerprinting Using Wavelets. Conference on Visual Media Production. 2006. pp. 198-207.

[32] S. Baluja and M. Covell. Audio Fingerprinting Combining Computer Vision and Data Stream Processing. International Conference on Acoustics, Speech and Signal Processing. 2007. Vol. 2. pp. 213-216.

[33] S. Baluja and M. Covell. Waveprint: Efficient Wavelet-based Audio Fingerprinting. Pattern Recognition, November 2008, Vol. 41(Issue 11): 3467- 3480.

[34] Y. Jiao, B. Yang, M. Li and X. Niu. MDCT-Based Perceptual Hashing for Compressed Audio Content Identification. IEEE Workshop on Multimedia Signal Processing. 2007. pp. 381-384.

[35] H. G. Kim, J. Y. Kim and T. Park. Video Bookmark Based on Soundtrack Identification and Two-Stage Search for Interactive-Television. IEEE Transactions on Consumer Electronics, 2007, Vol. 53(Issue 4): 1712- 1717.

[36] 韩纪庆, 冯涛, 郑贵滨, 马翼平. 音频信息处理技术. 北京：清华大学出版社, 2007. pp. 28-44.

[37] 赵力. 语音信号处理. 北京：机械工业出版社, 2003. pp. 31-45.

[38] A. Spanias, T. Painter and V. Atti. Audio Signal Processing and Coding. USA: Wiley-Interscience, 2007. pp. 13-25.

[39] 薛年喜. Matlab 在数字信号处理中的应用(第 2 版). 北京：清华大学出版社, 2008. pp. 313-340.

[40] 丁玉美, 阔永红, 高新波. 数字信号处理. 西安：西安电子科技大学出版社，2002. pp. 230-235.

[41] M. L. Miller, M. A. Rodriguez and I. J. Cox. Audio Fingerprinting: Nearest Neighbor Search in High Dimensional Binary Spaces. IEEE Workshop on Multimedia Signal Processing. December 2002. pp. 182-185.

[42] K. Kashino, G. Smith and H. Murase. Time-Series Active Search for Quick Retrieval of Audio and Video. International Conference on Acoustics, Speech and Signal Processing. March 1999. Vol. 6. pp. 2993-2996.

[43] A. Kimura, K. Kashino, T. Kurozumi and H. Murase. Very Quick Audio Searching: Introducing Global Pruning to the Time-Series Active Search. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2001. Vol. 3. pp. 1429-1432.

[44] A. Qamra and E. Y. Chang. Scalable Indexing for Perceptual Data. International Workshop on Multimedia Content Analysis and Mining. 2007. pp. 24-32.

[45] Y. Yu, M. Takata and K. Joe. Similarity Searching Techniques in Content-Based Audio Retrieval via Hashing. International Multimedia Modeling Conference. 2007. Part 1. pp. 397-407.

攻读硕士学位期间发表学术论文情况

[1] Ting-Xian Zhang, Wei-Min Zheng, Zhe-Ming Lu and Bei-Bei Liu. Comments on "A Semi-blind Digital Watermarking Scheme Based on Singular Value Decomposition". The Eighth International Conference on Intelligent Systems Design and Applications. Kaohsiung, Taiwan, China. November 2008. Vol. 2. pp. 123-126. ISTP 收录.

[2] Ting-Xian Zhang, Ji-Xin Liu and Zhe-Ming Lu. Robust Audio Fingerprinting Based on Daubechies Wavelets. Submitted to International Journal of Information Analysis and Processing.

致 谢

值此论文完成之际，谨向指导、关心和帮助我的老师、同学、朋友和亲人致以衷心的感谢。

感谢我的导师陆哲明教授，从论文的选题、资料的搜集，算法研究到论文的撰写，陆老师都给予了悉心的指导。陆老师渊博的专业知识，严谨的治学态度，精益求精的科研精神对我硕士期间的研究工作有着重大的影响，也让我终生受益。在此谨向陆老师致以诚挚的谢意和崇高的敬意。

感谢实验室郑慧诚老师为论文研究工作的开展提供了宝贵的实验室资源。感谢实验室的全体博士、硕士研究生，尤其是刘继新师兄。与他们的交流探讨促使我进步与成长，正是他们的帮助和支持，使我的求学过程充满了乐趣，使我能克服一个个困难和疑惑。

感谢我的父母、家人和朋友，他们的鼓励、理解与支持，是我一直前进的动力和信心。

最后，由于本人能力有限，论文中难免有疏漏之处，恳请各位评委和同学批评指正。同时，衷心感谢各位评审老师以及各位评委对本文所提出的建议与指正。