数字音频认证研究综述

李 伟[1] 汪竹蓉[1] 李晓强[2] 刘亚多[1]
(复旦大学计算机科学与技术学院 上海 200433)[1] (上海大学计算机工程与科学学院 上海 200072)[2]

摘 要 现代音频信号处理技术使得对音频内容的篡改、替换,对时间序列的调换顺序等恶意操作可以以极低的代价进行,对音频完整性和真实性进行认证变得日益重要。对人类听觉系统来讲,音频认证技术需要保护的是音频内容而不是比特流本身,因此它应该能够容忍一些保持音频听觉质量或者语义的正常信号处理操作而不触发检测器。介绍了音频内容认证技术的产生背景、典型应用场合、需满足的必要性质、硬认证与软认证的特点、保持内容操作和恶意操作的划分,综述了典型的音频内容认证算法,最后总结并讨论了该研究领域的技术特点并提出了可能的解决方案。

关键词 音频内容认证,保持内容操作,恶意操作,重同步

Review on Digital Audio Authentication

LI Wei[1] WANG Zhu-rong[1] LI Xiao-qiang[2] LIU Ya-duo[1]
(School of Computer Science, Fudan University, Shanghai 200433, China)[1]
(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)[2]

Abstract Modern audio processing techniques have made it pretty easy to make modifications like tampering, replacement, rearranging etc to audio content and time sequence. It is more and more important to ascertain the integrity and authenticity of audio data. In terms of human auditory system, audio authentication aims to protect audio content rather than bit stream itself, this way, it should sustain some common signal manipulations that can maintain audio perceptual quality or semantic meaning while leaving authentication detector untriggered. This paper gave a vision on the background, representative application scenarios and properties of audio content authentication, characterized hard authentication and soft authentication, differentiated content-preserving and malicious manipulations, and summarized most state-of-the-art audio content authentication algorithms published in the literature. Several technical characteristics in this research field were concluded and discussed, possible solutions were also pointed out for future research.

Keywords Audio content authentication, Content-preserving processing, Malicious processing, Resynchronization

1 研究背景

近年来,多媒体压缩技术的成熟及互联网的迅猛发展使得图像、视频和音频等多媒体数字作品的创作、存储和传输都变得极其便利,以MP3为代表的海量音乐信息在互联网上得以广泛传播。现代音频编辑和处理技术对音乐、语音等音频数据的高质量篡改和伪造可以以极低的代价进行,多媒体数据的可信度经常受到怀疑或没有法律效应,原因就在于数字产品的可编辑性[1]。例如,如图1所示,音频数据的语义可以通过简单地重排或去掉几个小的片断进行改变,因此只依靠人的听觉测试来判断音频数据的完整性/真实性是完全不够的。为了得到安全的多媒体应用,有效检测对媒体内容的恶意篡改变得日益重要。

音频内容认证技术就是一个实现对音乐、语音等音频数据完整性/真实性进行保护的有效技术手段,它可以保证接收到的音频数据在传送过程中没有经过第三方的恶意编辑和篡改,即在人类感知系统的意义上与原始音频是完全相同的。该技术在政府部门、国家安全、法庭辩护、商业机密、新闻、录音讲话、音乐录制发行、军事等许多领域都有广泛的应用及巨大的社会效益。目前基于音频数据的多媒体认证技术在国内外研究界尚处于起始阶段,已发表的文献很少,未查到已申请成功的专利,成熟的商业化软件亦未出现。该研究属于多媒体信息安全领域的前沿性研究,涉及的知识面大,技术难度高,具有很大挑战性。

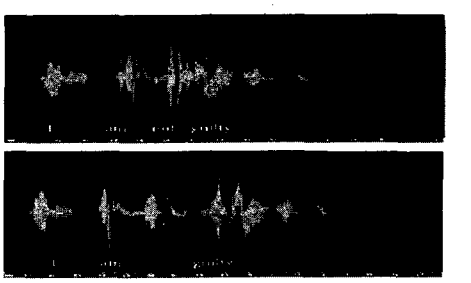

图1 数字音频可以轻易地被修改

到稿日期:2008-12-04 返修日期:2009-02-09 本文受国家自然科学基金(60873255)资助。

李 伟(1970—),男,博士,副教授,主要研究领域为信息隐藏与数字水印、音频信息检索,E-mail:weili-fudan@fudan.edu.cn;汪竹蓉(1986—),女,硕士生,主要研究领域为音频识别与认证;李晓强(1973—),男,博士,副教授,主要研究领域为信息隐藏与数字水印、数字图像处理与模式识别、数字版权管理;刘亚多(1983—),男,硕士生,主要研究领域为音频信息检索。

2 多媒体认证的概念及分类

传统认证方法，即使用数字签名系统[2,3]在密码学中已得到完备的研究。在一个签名系统中，通过使用密码 Hash 函数得到消息摘要，产生一个数字签名后绑定到原始数据上。即使数据只有一个比特被改变，那么计算出来的签名也无法与原始签名相匹配，这样可以发现数据的任何改变。但是数字签名必须和数据一起传输，比如存放在文件头中。如果数据转换为另一种格式，签名就会丢失，认证也就不再有效。

与其它数据不同，一个多媒体信号能够以多种不同格式等价地表示。例如，一段 WAV 格式的原始音乐被转换为 MP3，WMA，RM 等压缩格式音乐后仍然表达出同样的听觉信息，在比特率较高时用户几乎感觉不到任何差异。因此，多媒体认证追求的是内容认证而不是像传统认证方法那样简单地保护比特流。

多媒体认证，尤其是图像认证在近年来是一个非常活跃的研究领域。根据采用的技术手段，所有的认证方法可以分为两类：鲁棒数字签名法和数字水印法。数字签名法提取反映媒体特性的紧致数字签名（也可称为数字指纹），一般不对媒体内容进行修改，可以以头文件的形式附加到媒体外部，或存储在数据库中，或采用水印方法自嵌入到媒体（如图 2 所示）。基于水印的方法将认证信息嵌入到媒体中，又可分为完全脆弱水印法和半脆弱水印法。特征提取系统需要传输信道中的所有节点参与传递特征数据，这有时并不方便甚至并不可能，而脆弱/半脆弱音频水印技术则完全不需要各节点的参与。根据内容完整性标准，多媒体认证又可以分为硬认证和软认证[4]。硬认证拒绝对多媒体内容的任何修改，唯一接受的操作是保持视觉像素值和音频样本值的无损压缩和格式转换，其它所有信号处理都将触发检测器，使验证失败。此类算法主要基于完全脆弱水印技术，类似于经典认证，只是这些无损操作也被经典认证所拒绝。软认证可以通过某些内容修改，称为可容许的操作，并拒绝其它恶意处理。软认证进一步分为基于质量的认证和基于语义的认证，前者拒绝任何使感知质量下降到低于某一可接受水平的处理，后者则拒绝任何改变媒体语义的操作。此类算法一般采用鲁棒数字签名法或半脆弱水印法。软认证通常以某种度量方式测量接收信号的数字签名/水印和原始信号相应签名/水印间的差别，并与一个预定的门限进行比较，来确定接收信号是否通过认证。通常，在认证通过和认证失败的信号之间没有明显的界限。在许多应用中，可接受和不可接受操作的分类依赖于具体情况。区分由容许操作和恶意操作引起的失真通常是很困难的，这种内在的模糊性使得软认证的设计在大多数情况下非常困难并具有挑战性。有些软认证系统给出一个认证可信度而不是明确的是/否二值输出。基于质量的认证系统还需要提高对容许操作的鲁棒性，并保持对恶意处理的敏感性。基于语义的认证的主要挑战则是如何定义从人类知觉出发能够准确、唯一表征多媒体内容的特征矢量。目前提出的此类算法都使用启发式的特征来表征多媒体内容，例如块直方图、均值、低阶矩、边缘、小波变换域重要点等图像特征。此外，基于质量和语义的软认证方法还需要提高对抗能够保持感知质量或媒体语义的几何/时间域同步变换处理的能力。

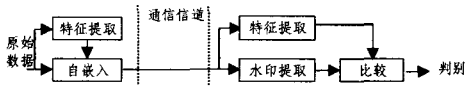

图 2 多媒体数字签名自嵌入认证方法

3 音频内容完整性认证系统

一个能够区分内容保持操作和恶意篡改的认证系统叫做内容完整性系统。不同应用会有不同的标准来区分各种操作，一个保持内容操作可能在另外一个应用中被视为恶意操作。例如，MP3 音频压缩在绝大多数应用中被视为保持内容操作，但是在录音室和唱片公司 CD 母盘制作、转录等场合中就应该被拒绝，因为原始音乐不能有任何的质量损失。

显然，硬认证具有最小的失真容忍性，而基于语义的认证则具有最大的失真容忍性，只要不改变媒体传达的语义即可。不同的应用会需要不同种类的认证。例如，录音师、作曲家、歌星等专业人员为保护高品质的原始音乐作品，一般需要只接受无损压缩和格式转换的硬认证；而在面向普通听众的电台广播、卡拉 OK 等娱乐过程中，从发送端到接收端音频数据可能会受到如码率控制、调幅、时间缩放等多种中间处理，以使在资源允许的情况下得到最大的接收质量，此种应用下需要进行对中间处理过程鲁棒的基于质量或语义的软认证。

在软认证中，尤其是基于质量的软认证中，由于保持内容操作和恶意操作缺乏明确的定义，准确地刻画它们变得非常困难，需要根据不同的应用具体分析，成为一个公开的研究难题。另一个研究难点就是如何抵抗同步攻击。对音频信号来说，抖动攻击（每隔一定数量样本进行的均匀剪切或添加）和保持音调不变的时间缩放 TSM（Time-scale modification）是两种常见的时间域同步处理。程度轻微时，它们可以保持很高的听觉质量；程度加重时，音频质量有所下降，但仍可以保持音频的语义。因此，同步处理至少可以在基于语义的认证系统中被视为可容许操作。但是，目前的绝大多数认证算法对此类处理引起的失同步没有任何的抵抗力。如何提高算法自身的重同步能力，是一个急需解决的研究难题。

如上所述，区分保持内容操作和恶意操作与特定应用有关，没有统一的区分规则。这里仅对一般听觉应用场合下保持语义的软认证系统中可能经历的音频信号处理操作和时间域同步处理进行如下分类。

（1）保持内容操作：包括音频转码、重采样、调节音量、去除噪声、滤波、均衡化、回声、无损压缩和高比特率有损压缩、A/D 和 D/A 转换等标准音频信号处理；保持音调不变的时间缩放、抖动攻击、平移等时间域同步处理。

（2）恶意操作：改变原始音频语义内容，包括局部替换、修改、删除、添加，时间轴多个片段重排序，低比特率的有损压缩等。

由于多媒体内容认证系统为了定位被恶意篡改、替换的局部区域，一般采用基于分块/分帧的技术，因此还需要抵抗一种叫做“标记转移”的专门针对分块/分帧公开水印系统设计的恶意攻击。此类攻击利用已有的带水印信号来伪造有效的水印，其中最著名的一种是 Holliman 和 Memon 提出的矢量量化攻击[5]。标记转移攻击一般如下进行：首先除去水印，然后修改信号内容，最后重新嵌入水印。对脆弱性水印，经受该攻击可能会产生虚假认证；对鲁棒性水印，此攻击能把水印

从一个受版权保护的图像拷贝到另外一个不受保护图像，而不需要知道水印或嵌入密钥。抵抗此类攻击的关键在于去除分块/分帧独立嵌入的特性，通过使水印嵌入块/帧与其它块/帧相关，伪造水印的问题会迅速变为计算不可行，使得攻击者伪造水印的可能性大大减少。例如，可以从当前块及其相邻块计算 Hash 值，或在一个大的块范围（比如 16×16）计算 Hash 值，然后嵌入一个小一些的块（如 8×8）。

一个设计良好的数字音频内容完整性认证系统应该满足如下性质：

（1）认证数据量足够小，与宿主数据无缝集成；

（2）引入的噪声不可察觉；

（3）若采用水印方法，应进行盲检测；

（4）发送端和接收端计算代价低；

（5）能抵抗在传输信道中的保持内容操作；

（6）篡改检测：能检测局部恶意修改并精确指示篡改位置，最好能对被篡改区域进行近似恢复；

（7）时间序列保护：改变音频帧的时间序列可能会改变内容，因此要加以保护；

（8）若待保护的音频序列与某个视频序列是一个整体，那么其相互间的同步关系也必须得到保护，防止被整体替换；

（9）安全性：在算法嵌入端和验证端均使用大空间的密钥，对一个具有完全算法知识的未经授权方，应该使其伪造一个有效的认证，从公开数据和知识中推断认证秘密信息，或不经检测就进行恶意处理都变得非常困难。

4 音频认证算法概述

迄今为止，绝大多数水印认证系统都集中在图像域。学术界只发表了有限的一些关于数字音频认证的文献，其中大多数又是基于语音，只有极少数关于音乐认证的工作。

4.1 基于特征的音频认证算法

Radhakrishna 等[6]根据以下原理提出了一种基于特征的音频内容认证技术，即两个听觉质量相似音频之间的掩蔽曲线几乎是一样的，即具有相当的稳定性。首先计算音频掩蔽曲线的 Hash 函数值，然后采用已知的数据隐藏方法将之作为水印嵌入到音频信号中。检测器将水印提取后与之前计算出的哈希值比较，计算其相关系数，通常该系数随着接收音频听觉质量的变化也有适度的下降，根据能够接受的听觉质量标准可以适当调整相关系数的门限值。实验结果表明，该基于内容的哈希值完全可以将 MP3 等音频信号处理与恶意篡改等操作区分开来。

Haitsma 等[7]设计了另一种基于特征的音频内容认证技术。以能量差分作为鲁棒特征，计算得到一个鲁棒哈希值对象并转换成比特串。通过比较接收音频和原始存储音频的哈希值来进行内容鉴别。实验证明，该方法对各种保持内容处理是非常鲁棒的，且具有极低的认证虚警率。

Wu 等[8,9]分别提出两种与 ITU G.723.1 和 CELP 语音编码器集成到一起的语音内容完整性校验方法，使整体计算量最小。与语义相关的语音特征被提取出来，加密后作为文件头信息附加到文件上。此方法不仅比密码的比特流完整性算法更快，而且适用于更广泛的应用。语音信号可以经历重压缩、幅度调制、转码、重采样、D/A 和 A/D 转换，以及少量白噪声而不会触发认证器。文献[9]中还使用一个低代价的同步算法来解决由于保持内容操作引起的失同步问题，语音中的静音段和有声段也用低复杂度的算法加以识别，还进行了统计分析来计算篡改检测的误检率（false positive rate）。

4.2 基于半脆弱水印的音频认证算法

Quan 等[10]提出一种新颖的小波包域量化水印方案，用于音频认证，选择最优小波包基适应心理声学模型，以使分解的子带结构近似于临界频带。该算法实现了自适应小波包分解，以使掩蔽域值比大多数现存算法都大，使得嵌入算法更加灵活。与以前方法不同，该算法将几个水印比特嵌入到一个系数中。因为有效的时间分辨率，算法能够精确定位被篡改的时频区域。此方法的篡改检测与尺度无关，而是基于掩蔽域值，这与人类知觉系统是一致的。

Steinebach 等[11]介绍了两种用于音频内容认证的算法。第一种讨论了可能的特征，以允许几种后续信号处理；第二种为得到最高安全性，对每个比特的改变进行检测，并通过引入可逆水印概念来重构原始音频。作者进而结合数字签名和数字水印，并使用密钥来产生一个可公开验证并能重建原始音频的方法。

Yan 等[12]基于线性预测系数的量化提出一种半脆弱语音水印技术，通过将参数估计错误模拟为 Laplace 分布来分析水印解码器性能。水印的脆弱性通过水印检测门限来控制，根据错误概率的要求和希望的信噪比 SNR 推断水印检测门限。实验表明，该方法对幅度伸缩具有鲁棒性，对白噪声添加具有半脆弱性，因此完全适合语音认证。

Wu 等[13,14]提出了两种用于检验内容完整性的半脆弱语音水印技术，即指数级奇偶调制技术和线性相加水印技术，并在检测局部恶意篡改、容忍保持内容操作、错误检测率和虚假检测率的统计结果等方面进行比较。这两种方法均在 DFT 域嵌入水印，不需要额外的辅助数据来进行完整性校验，并且都能够把不同的保持内容操作和恶意篡改区分开。Wu 等在文献[15]中进一步详细比较了以上两种数字音频认证方法。实验结果表明，指数级奇偶调制技术在检测局部内容篡改能力上更好，而线性相加水印技术能容忍更低比特率语音编码器，如 CELP。

Chen 等[16]提出了一种使用小波包分解和最优树选取的基于质量的认证算法。以音频片段的小波包分解系数作为与质量相关的鲁棒特征，其中小波包系数的选取采用一种在最小熵意义下的最优树算法，从而达到在保证不丢失过多音频的重要信息的前提下使水印的编码量达到最小的目的。实验结果显示，除了各种比特率的 MP3 压缩，大多数的音频处理操作，如回声、均衡化、调节音量等都不能通过该系统的认证，在 PEAQ 标准下被视为“非保持质量”的操作，这与它们在一定程度上引入了听觉效果上的变化是一致的，而事实上 MP3 是一种最重要的“保持质量”的操作。此外，对于随机剪切和局部区域篡改操作的测试表明，算法对以上两类恶意操作也能作出有效的识别。

Emilia Gomez 等[17]提出一种水印与指纹混合的音频录音完整性认证的方法。在算法中，把音频信号看作是由一系列的“声音事件”（ADU）组成的一个序列，以 ADU 以及相应的时间信息作为与内容相关的指纹进行提取，指纹以水印形式自嵌入到原始音频中。对于音频信号的局部修改不会引起 ADU 序列的连续变化，因此该方法不仅能对改变内容的操作

作出有效识别，而且能根据 ADU 序列上的变化点确定篡改发生的大致位置。此外，算法能够满足实时性的要求，非常适合处理像录音之类的流数据。

大多数水印方法不能同时用于多种用途。如果需要实现多个目的，就需要同时注入多个水印。因为不同的水印担负不同的使命，隐藏的顺序十分重要。Lu 等采用鸡尾酒水印(cocktail watermarking)方法[18]将两个功能互补的鲁棒性水印和脆弱性水印同时植入原始音频[19]。第一个水印按正调制规则嵌入，递增地调制音频的 FFT 变换系数；第二个水印按负调制规则嵌入，递减地调制 FFT 变换系数，这样可以同时实现音频版权保护和内容认证的功能。该方法的关键是用不同的方法来检测鲁棒性水印和脆弱性水印，因此隐藏的顺序并不重要。由于无法获得原始音频，检测方法必须是盲检测。对音频版权保护，可得到高的鲁棒性；对音频认证，可以检测到篡改的区域。据我们所知，这是第一个可以一次性植入且不用考虑隐藏顺序的多用途音频水印方法。Cvejic 等[20]提出另一种新颖的能够将音频版权保护和内容认证结合到一起的方法。附加信息嵌入到不同信号域，其中鲁棒水印在傅立叶域使用频率跳动方法嵌入，认证数据在小波域使用 LSB 调制进行隐藏，并利用人类听觉系统 HAS 来得到高的感知透明性。对版权保护，该系统对时间域时间伸缩(Time scaling)等去同步处理得到高的鲁棒性，对 MPEG 编码效果稍差，因为压缩技术会剪掉带水印音频的某些高频谱并量化某些子带的小波系数。对内容认证，当发现认证错误的比特位时，检测系统利用小波系数的空间域信息来精确定位被篡改的音频片段。

上述各种方法虽然都具有一些自己的特色，也在某些方面取得了较好的结果，但是还都无法满足一个理想的音频内容认证系统应具有的各项要求，即对保持内容操作鲁棒、对恶意操作脆弱、精确定位并近似恢复局部篡改区域、识别时间序列是否被改变、具有极低的虚假认证率。上述要求并不互相冲突，可以通过更好的设计得以实现。

结束语 能够进行音频真实性、完整性验证的音频认证技术是学术界研究的难点之一，目前只有少量的研究成果发表。如何根据不同应用定义在软认证中保持听觉质量或保持语义的操作，具有本质性的困难。精确定位篡改区域的需求决定了此类算法应该是逐帧进行的。当面对轻微的同步攻击时，无法借鉴强鲁棒性局部化算法中使用锚点的重同步机制，因此嵌入的水印或计算出的签名必须具有一定的统计意义，依靠自身的力量抵抗此类处理。为了提高篡改检测的准确率，每帧应嵌入多比特水印或计算多比特签名，防止出现单比特信息极易引起的虚假认证情况(发生局部篡改，但对应水印或签名信息无法反映)。对于音乐和语音这两种最主要的音频数据，因为其本身的特性各不相同，如语音通常具有有限的带宽、交替的有声段和无声段、有限范围的声调，而音乐则是连续的并具有高得多的带宽和音调范围，因此对语音和音乐应采用不同的方法嵌入水印或计算鲁棒签名。

参 考 文 献

[1] Yeung M, Mintzer F. Invisible watermarking for image verification[J]. Journal of Electronic Imaging, 1998, 7(3): 578-591

[2] Walton S. Information authentication for a slippery new age[J]. Dr. Dobbs Journa, 1995, 20(4): 18-26

[3] Stinson D. Cryptography Theory and Practice[M]. Boca Raton: CRC Press, 1995

[4] Zhu B B, Swanson M D, Tewfik A H. When seeing isn't believing[J]. IEEE Signal Processing Magazine, 2004, 21(2): 40-49

[5] Holliman M, Memon N. Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes [J]. IEEE Transactions on Image Processing, 2000, 9(3)

[6] Radhakrishnan R, Memon N. Audio content authentication based on psycho-acoustic model[C]// Proceedings of the Security and Watermarking of Multimedia Contents. San Jose, CA, February 2002

[7] Haitsma J, Kalker T, Oostveen J. Robust audio hashing for content identification[OL]. http:// www. extra. research. philips. com/natlab/download/audiofp/cbmi01audiohashv1. 0. pdf

[8] Wu C P, Kuo C C. Speech content integrity verification integrated with ITU G. 723. 1 speech coding[C]// IEEE International Conference on Information Technology: Coding and Computing (ITCC2001). 2001: 680-684

[9] Wu C P, Kuo C C. Speech content authentication integrated with CELP speech codes [C] // IEEE International Conference on Multimedia and Expo(ICME). 2001

[10] Quan X, Zhang H. Perceptual criterion based fragile audio watermarking using adaptive wavelet packets[C]// Proceedings of the 17th International Conference on Pattern Recognition (ICPR). 2004

[11] Steinebach M, Dittmann J. Watermarking-based digital audio data authentication[J]. EURASIP Journal on applied signal processing, 2003, 10: 1001-1015

[12] Yan B, Lu Z M, Sun S H, et al. Speech authentication by semi-fragile watermarking[C]// KES 2005, LNAI 3683. 2005: 497-504

[13] Wu C P, Kuo C C. Fragile speech watermarking based on exponential scale quantization for tamper detection[C]// Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing. Orlando, Florida, May 2002

[14] Wu C P, Kuo C C. Fragile speech watermarking for content integrity verification[C]// Proceedings of the IEEE International Symposium on Circuits and Systems. Scottsdale, Arizona, May 2002

[15] Wu C P, Kuo C-C J. Comparison of two speech content authentication approaches[C]// Proceedings of SPIE vol. 4675—Security and Watermarking of Multimedia Contents IV. 2002: 158-169

[16] Chen F, Li W, Li X Q. Audio quality-based authentication using wavelet packet decomposition and best tree selection[C]// Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP) '08 International Conference. Aug. 2008: 1265-1268

[17] Gomez E, Cano P, C T de L Gomes, et al. Mixed Watermarking-Fingerprinting Approach for Integrity Verification of Audio Recordings[C]// International Telecommunications Symposium(ITS). Natal, Brazil, 2002

[18] Lu C S, Liao H Y M. Multipurpose watermarking for image authentication and protection [J]. IEEE Transactions on Image Processing, 2001, 10(10): 1579-1592

[19] Lu C S, Liao H Y M, Chen L H. Multipurpose audio watermarking[C]// Proc. 15th Int. Conf. Pattern Recognition. Barcelona, Spain, 2000: 286-289

[20] Cvejic N, Seppänen T. Fusing digital audio watermarking and authentication in diverse signal domains[C]// Proc. European Signal Processing Conference. 2005: 84-87